CONTRIBUTING FACTORS OF DUI RECIDIVISM AMONG FIRST-TIME OFFENDERS IN NORTH DAKOTA

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yun Zhou

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Transportation, Logistics, and Finance

October 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Contributing Factors of DUI Recidivism Among First-Time Offenders in North Dakota

By

Yun Zhou

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Joseph Szmerekovsky
Co-Chair

Dr. Kimberly Vachal
Co-Chair

Dr. Diomo Motuba

Dr. Kambiz Farahmand

Approved:

11/29/2022
Date

Dr. Tim Peterson
Department Chair

ABSTRACT

This study explored utilizing tree-based machine learning models to identify associations in a range of 107 factors and DUI recidivism among first-time DUI offenders. Three tree-based machine learning models, Decision Tree, Random Forest, and Gradient Boosting were performed on 12,879 first-time DUI offenders during 2013-2017 using a three-year following period, to classify repeat DUI offenders. Study cohorts include 11,651 drivers without recidivism and 1,228 drivers with recidivism. The models tested 107 variables/predictors, including the driver's demographic factors, drinking behaviors, traffic violations, crash histories, DUI-related violations, social-economic factors, and health and safety factors based on the driver's residence. oversampling technique was used to balance two classes in the training data in all three models. The top 15-20 predictors were selected from the feature impact analyses of these predictions. Lastly, multiple logistic regression analyses were performed to quantify the effects of selected factors/predictors on the outcome.

Among the three models, Gradient Boosting achieved the best predictions on both the original and oversampled datasets. Oversample techniques did improve prediction performances by roughly 10% on the F1 score for Gradient Boosting. Results coalesced around two findings. First, male drivers with higher BAC values, younger age at first DUI citation, whose first DUI citation took place during the weekday, had at least one low-risk citation within three years before first DUI citation, and lived in counties with lower income inequality ratio and higher violent crime rate were more likely to commit a subsequent DUI offense. Second, male drivers who complied with a BAC test upon arrest, whose first DUI citation took place on a weekday, had at least one low-risk citation within three years before the first DUI citation, lived in a county with a lower income inequality ratio, and higher violent crime rate were more likely to

commit a subsequent DUI offense. Findings can be used by stakeholders in implementing and improving DUI prevention strategies. The study is limited to a single state, but the comparison of techniques and their shared findings suggest that a multitude and variety of approaches may be appropriate in future impaired driving prevention research.

ACKNOWLEDGMENTS

This dissertation was supported by my advisor, Dr. Joseph Szmerekovsky and Dr. Kimberly Vachal who provided insights and expertise that greatly assisted the dissertation. Their utmost professionalism, considerable efforts, patient, wisdom and continuous encouragement have kept me moving forward during the study. I would also like to show my gratitude to Dr. Diomo Motuba and Dr. Kambiz Farahmand for their time and feedback to improve the quality of this dissertation.

I thank the North Dakota Department of Transportation (NDDOT) for providing the data and funding for this research. I thank the Department of Transportation, Logistics and Finance for providing me high quality coursework and teaching experience. I thank the Upper Great Plains Transportation Institute (UGPTI) for providing me workstation and stipend for my entire doctoral study.

This dissertation would not have been done without my family behind me, my husband Tong Lin, my mother Meiying Zhu and my father Zhuliang Zhou. I am indebted for the mental and physical support they have provided that has spurred me on through tough times during not only my dissertation but also my entire doctoral study. I am thankful for their love and prayers. I also thank the Red River Valley Chinese Christian Church to help me with spiritual growth and always providing me support in my daily life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

1. INTRODUCTION

1.1. General Introduction

Alcohol-impaired driving or driving under the influence (DUI) of alcohol is a serious problem in the United States. In 2018, there were 10,511 fatalities in motor vehicle traffic crashes in which at least one driver was alcohol-impaired with a blood alcohol concentration (BAC) of 0.08 g/dL or higher, representing 29 percent of all traffic fatalities for the year (NHTSA, 2019). Every day, 29 people in the United States die in motor vehicle crashes that involve an alcohol-impaired driver. This is one death every 50 minutes (CDC, n.d). The annual cost of alcohol-related crashes totals more than $44 billion (CDC, n.d.). The effects of alcohol on drivers include but are not limited to impaired judgment, deteriorated reaction time, poor muscle coordination, impaired vision, interference with concentration, and a false sense of confidence (CDC, .n.d.).

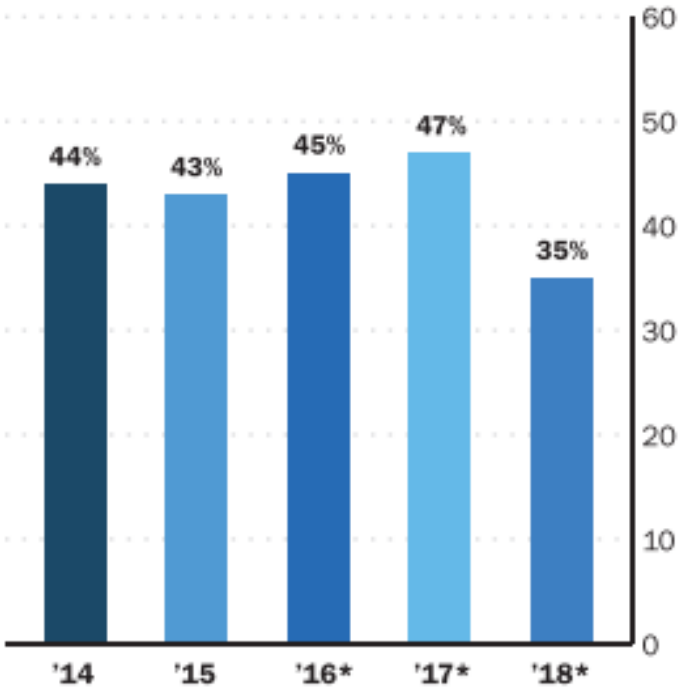

Figure 1.1. North Dakota percent alcohol-related fatal motor vehicle crashes, 2014-2018. Data source: NDDOT (2018).

As shown in Figure 1.1, alcohol contributes to about 43% of fatal crashes in North Dakota annually (NDDOT, 2018). Alcohol-related crashes are 100% preventable. Many lives would be saved each year if every driver consistently chose to be a sober driver. Reducing the number of alcohol-impaired drivers in the state is one of the North Dakota Department of Transportation (NDDOT) priorities.

The impaired driving trends showed the Midwest region had the highest rate of impaired driving, with 643 episodes per 1,000 population (Bergen, Shults, and Rudd 2011). North Dakota had the highest self-reported impaired driving rate in the Midwest region Dakota (Bergen, Shults, and Rudd, 2011). A recent survey of North Dakota drivers also shows great propensity, with 35.2% of the population reporting they had operated a vehicle within two hours of consuming one or two alcoholic beverages (Vachal, Benson, and Kubas 2019). Besides, as shown in Tables 1.1 and 1.2, although the numbers of alcohol-related fatal crashes, fatalities, and injuries decreased in 2018, the numbers of alcohol-related fatal crashes and fatalities for 2019 increased again. More work is needed to prevent alcohol-impaired driving. It is more important to recognize repeat offenders as a high-risk sub-population because they are more likely to be involved in fatal motor vehicle crashes (NHTSA, 2008; Dickson, 2013).

Due to the limited number of law enforcement agencies, it is impossible to catch every DUI offense. Thus, DUI first-time offenders on record are very likely to be repeat offenders who just haven’t been caught by the law enforcement agency yet (Voas & Lacey, 1990; Beitel, Sharp, & Glauz, 2000; Wickens et al., 2018). Therefore, the actual number of convicted DUI offenders may be underestimated. The change in the recidivism rate of convicted DUI offenders is one way to evaluate the effects of countermeasures on deferring impaired driving (Kubas and Vachal, 2019). The number of alcohol-related crashes may be affected by a lot of factors other than the

countermeasures, such as reduced traffic volume. The recidivism rate is relatively more independent of outside factors besides the efforts of law enforcement and actual recidivism.

Table 1.1. North Dakota alcohol-related motor vehicle crashes, fatalities, and injuries, 2014-2018. Data source: NDDOT (2018).

Year	Total fatal crashes	Fatalities	Injures
2014	53	63	564
2015	48	57	496
2016	46	54	446
2017	50	57	472
2018	33	34	407

Table 1.2. Fatal crash statistics-yearly totals. Data source: Vision Zero.

	2017		2018		2019		2020 To Date As of 8/04/2020	
	#	%	#	%	#	%	#	%
Fatal crashes with Operator Positive BAC and/or LE Reported	51	48.1%	30	35.0%	37	40.7%	12	23.1%
Fatal crashes w/ Investigation Pending							14	26.9%
Fatalities from Alcohol Crashes	55	47.4%	31	29.5%	42	42.0%	13	24.5%

1.2. Problem Statement

This study intent to investigate the associated factors to DUI recidivism among first-time DUI offenders. DUI repeat offenders scored higher on the risk of DUI recidivism and are more likely to be involved in fatal motor vehicle crashes (NHTSA, 2008; Dickson, 2013; Wickens, 2018). The ability to identify the contributing factors to a subsequent DUI recidivism will be an invaluable aid in determining appropriate judicial and administrative sanctions and countermeasures for all DUI offenders. Current literature mainly focuses on exploring driver profiles and their offense histories, results from age, gender, racial, geographic factors, drinking

behaviors, traffic violation histories, crash histories, criminal histories, mental health, and environmental factors in relation to recidivism occurrence using a traditional statistical method such as Chi-square analysis, t-tests, logistic regression, multivariate regression, survival analysis, etc. (Marowitz, 1998; Cavaiola et al., 2007; Portman et al., 2010; Chaudhary et al., 2011; Møller et al., 2015; MacLeod et al., 2017; Wickens et al., 2018; etc.).

However, a research gap is evident in a cohesive study integrating these factors into one analysis. A possible reason might be that collecting all the factors into one dataset is difficult. Thus, the resulting dataset can have a complicated data structure that is unsuitable for analysis by traditional statistical methods. However, the results could be inaccurate without integrating these factors into one analysis. For example, the recidivism rate in an area could be higher than in other regions because more law enforcement agencies were patrolling in the area. In recent decades, less constrained tree-based machine learning models alleviate assumptions common in investigations aimed at target outcomes and predictors.

This study intends to integrate drivers' demographic factors, drinking behaviors, past traffic violations, crash histories, past DUI-related violations, social-economic factors, and health factors based on the driver's residence in one analysis to evaluate their influences on DUI recidivism using a mixed model approach.

1.3. Significance of the Study

This study integrates factors from different aspects to analyze DUI recidivism comprehensively. From the limited administrative record provided by NDDOT, a list of variables representing driver profiles, conviction and crash histories, law enforcement indexes, behavior treatment interventions, and seasonal factors were selected. Secondary data were collected to

measure health and social-economic factors associated with the driver's residence county. To the author's knowledge, it is the first study that integrates all those factors into one study.

Considering the complicated data structure in this study, three tree-based machine learning models- Decision Tree, Random Forest, and Gradient Boosting - were used to identify the associations in a range of 110 factors and DUI recidivism among first-time DUI offenders. Unlike traditional statistical methods, these three nonparametric models require no statistical assumptions and no underlying relationship between dependent and independent variables, so they are more suitable for the target data (Mitchell, 1997; Friedman, 2002; Wijenayake et al., 2018;). However, machine learning models are often known as "black-box" models that are sufficiently complex and have low interpretability. To improve interpretability, this study performed multiple logistic regression analyses to quantify the effects of selected factors/predictors on the outcome. To the author's knowledge, it is the first study that integrates machine learning models and statistical methods in the impaired driving literature.

The sample data in this study was imbalanced, with repeat offenders only accounting for 9.5% of the sample. Machine learning predictions made on such datasets might be less likely to favor the minority group – the repeat offenders and lead to misclassification. Thus, Synthetic Minority Oversampling Technique (SMOTE) - Tomek Links technique was used to oversample the minority groups so that the machine learning predictions were made on a balanced sample. SMOTE -Tomek Links technique did improve Gradient Boosting prediction. To the author's knowledge, it is the first study that applied advanced oversampling techniques on sample data to improve machine learning predictions in the impaired driving literature.

2. LITERATURE REVIEW

2.1. Background

In North Dakota, the same as most states in the U.S., drivers with a blood alcohol concentration (BAC) of 0.08% or higher are considered alcohol-impaired by law, and sanctions will be applied to such drivers. However, drivers can be convicted of drunk driving even when their BAC is under that limit, e.g., a noticeable impairment. The average BAC among North Dakota DUI offenders is .17 - one of the highest in the country and is more than twice the legal limit of .08 (NDDOT, n.d.).

Strategies implemented in North Dakota to reduce alcohol-impaired driving includes but are not limited to normal patrol, sobriety checkpoints, saturation, and roving patrols, a 24/7 Sobriety Pilot Program, and administrative licensing sanctions. Sobriety checkpoints deter impaired driving, not increase arrests (Goodwin et al., 2015). It is a concentrated enforcement effort to identify and arrest impaired drivers. Law enforcement agencies stop vehicles at a preselected, highly visible location to check whether the driver is impaired. These checkpoints are selected based on high alcohol or drug-related incidences and will be established and published before each operation. Law enforcement agencies either stop every vehicle or stop vehicles at some regular interval, such as every third or tenth vehicle (Goodwin et al., 2015).

A saturation patrol is a large number of law enforcement agencies patrolling a selected area during a selected period to increase enforcement visibility (Goodwin et al., 2015). Saturation patrol agencies mainly search for impaired-driving behaviors, such as problems maintaining proper lane position, driving without lights at night, failure to signal, aggressive driving, speeding, and following too closely (Goodwin et al., 2015; Richard et al., 2017). The primary purpose of saturation patrols, like sobriety checkpoints, is to deter alcohol-impaired

driving by increasing the perceived risk of arrest. Thus, saturation patrols are usually publicized extensively and conducted regularly (Goodwin et al., 2015; Richard et al., 2017). The advantages of saturation patrols compared to sobriety checkpoints include increased effectiveness, reduced staffing, and comparative ease of operation (Goodwin et al., 2015).

The 24/7 Sobriety Program is used to monitor offenders at high risk for probation violations and notify offenders that there will be an immediate penalty after every probation violation. Consequently, these individuals remain sober to keep roadways safe from hazardous drivers (Kubas and Vachal, 2019). The program mandated offenders are tested for alcohol twice daily for breath testing, wearing an ankle bracelet to monitor alcohol electronically, and using a drug patch or urine testing (Kubas and Vachal, 2019). The project has strict enforcement to keep participants sober. If the offenders fail an alcohol screening test or do not show up to take it, then they will be sent directly to jail (Kubas and Vachal, 2019).

House Bill 1302 mandated enrollment for repeat offenders. As part of the legislation enacted in 2013, second-time offenders now have a mandatory 12-month enrollment in the 24/7 Sobriety Program. Third-time offenders also have a mandatory 12-month enrollment in the program but are further subjected to supervised probation. Fourth-and-subsequent offenders are required by law to be enrolled in the program for 24 months in addition to being placed on supervised probation. This law went into effect on August 1, 2013.

2.2. DUI Recidivism

DUI is a major public health and safety problem worldwide. Research from various perspectives (e.g., public health, legal, behavior science, road safety, etc.) reveled that repeat DUI offenders are a heterogeneous group, and only one or two characteristics are unlikely to account for the behavior of DUI offenders (Nochajski and Stasiewicz, 2006). Instead, multifactor

analyses are needed to help explain the complexity interplay of factors from various perspectives to predict or prevent future DUI recidivism (Nochajski and Stasiewicz, 2006). While studies regarding DUI recidivism have been done on various data sources, this study limited the literature review to past research that was conducted with at least one official record to fit the scope of this study.

2.2.1. DUI Recidivism Defination

There are several ways to define DUI recidivism in the literature. Factors associated with DUI recidivism can be different when the definitions were different. The broadest definition is driving under influence of any amount of alcohol, given that one drink might put some individuals at significantly higher risk for a crash than if they had not consumed any alcohol drink (Nochajski and Stasiewicz, 2006). According to this definition, the recidivism rate calculated by official driving records may underestimate the "true" recidivism rate. The estimate of the number of DUI occurrences that happen before an arrest has ranged from 50 trips to 1,000 trips (Voas & Lacey, 1990; Beitel, et al., 2000). In this case, self-report information may provide more accurate estimates of DUI recidivism than the official driving record (Nochajski and Stasiewicz, 2001; Nochajski and Stasiewicz, 2006).

In DUI literature, the most common definition of recidivism is a subsequent DUI arrest on official records. However, based on this definition, the chance of a driver being arrested and identified as a DUI repeat offender depends on the level of law enforcement in the community and the amount of time that a conviction remains on the driver's official driving record (Nochajski and Stasiewicz, 2006). Therefore, it is important to take these two factors into account when examining DUI recidivism through official records. Unless otherwise noted, the literature reviewed in this study accorded this definition.

In recent decades, the number of drug-impaired offenses or driving under influence of drugs increased dramatically (Nochajski and Stasiewicz, 2006). Legislations have included illicit drugs and controlled medicinal drugs in the DUI law (Impinen et al., 2009). Unless otherwise noted, the literature reviewed in this study focused on alcohol-related DUI recidivism.

2.2.2. Characteristics of Repeat DUI Offenders

2.2.2.1. Demographic Characteristics

In the literature, analyses of demographic characteristics revealed significant associations between DUI recidivism and gender, age, education, ethnicity, employment status, income, and marital status (Nochajski and Stasiewicz, 2006). Generally, repeated DUI offenders tend to be young – age under 34, unmarried, males, who consume more drinks, and often reside in rural areas where fewer alternative transportation choices are available (McMillen et al. 1992a; McMillen et al. 1992b; Reynolds et al., 1991; Chang et al., 1996; C'de Baca et al., 2001; Cavaiola et al., 2007; Impinen et al., 2009; Robertson et al., 2016; Greene et al., 2018; Weisheit, 2020).

Among all the demographic factors, the most consistent insight is that males are much more likely to commit a subsequent DUI offense than females. In terms of recidivism rate, male drivers tend to be 1.2 to 1.7 times as female drivers (Chang et al., 1996; C'de Baca et al., 2001; Impinen et al., 2009; Robertson et al., 2016). Though no research provided a decent explanation for this insight, results from Hubicka et al. (2010) might provide a possible answer by examining personality traits and mental health among severe DUI offenders in Sweden. Male offenders scored low on openness to experience domain than female offenders and normal populations. This insight indicated that male offenders had less intellectual curiosity, receptivity to the inner

world of fantasy and imagination, appreciation of art and beauty, openness to inner emotions, values, and active experiences, and can be resistant to rehabilitate (Hubicka et al. 2010).

Age is a second significant factor in the DUI recidivism literature (C'de Baca et al., 2001; Impinen et al., 2009; Dugosh et al.; 2013; Robertson et al., 2016). Young drivers with impaired driving skills might also be detected more easily from traffic because inexperienced drivers are affected more by impaired substances (Vaez and Laflamme 2005; Impinen et al., 2009). Dugosh et al. (2013) reveled that early age at the time of first arrest for any criminal action, early age at the time of first DUI conviction, and early age of onset of substance abuse were all significant indicators for DUI recidivism. Although Dugosh et al. (2013) didn't define a range of "early age", two other studies concluded more decent findings regarding this factor. For high-risk recidivism, 28-year-olds and younger were concluded by Baca et al. (2001), and 33-year-old and younger were concluded by Robertson et al. (2016).

Ethnicity, education, employment status, income, and marital status are also commonly used factors in cross-sectional studies (Nochajski and Stasiewicz, 2006). Ethnicity and its relationship with repeat offender status vary in different regions of the country (Nochajski and Stasiewicz, 2006). The majority of repeat offenders tend to be White in the Northeast, Midwest, Northwest, and South regions, whereas the majority of repeat offenders tend to be Hispanic, African American, or Native American in the Southwest region (Chang et al., 1996; C'de Baca et al., 2001, Nochajski and Stasiewicz, 2006; Robertson et al., 2016).

Less-educated drivers have a higher risk of recidivism (C'de Baca et al., 2001; Robertson et al., 2016), and they greatly benefit from remedial interventions/education programs that help offenders to improve their knowledge of and intentions to avoid drink-driving (Wickens et al., 2018). While income level and employment status are highly related to education level, it is

understandable that drivers with lower income or unemployed are more likely to commit a subsequent DUI (Wieczorek & Nochajski, 2005; Nochajski and Stasiewicz, 2006). Finally, the marital status shows an association with repeat DUI offender status. Those who have never married or who have been divorced, separated, or widowed are more likely to receive a subsequent DUI than those who are married (C'de Baca et al., 2001; Nochajski & Wieczorek, 2000; Wieczorek & Nochajski, 2005; Nochajski & Wieczorek, 2006).

2.2.2.2. Alcohol-related Variables

BAC level at arrestment is often used as an essential factor of DUI recidivism, and a consistent finding is that a higher BAC level leads to a higher chance of recidivism (McMillen et al. 1992a; Marowitz, 1998; C'de Baca et al., 2001; Impinen et al., 2009; Dugosh et al., 2013; Roma et al., 2019). Though BAC has been identified as a robust predictor for future DUI in the literature, other factors should be considered meanwhile when determining the risk of recidivism, so that appropriate treatment and/or intervention can be ordered for offenders to rehabilitate (Dugosh et al. 2013).

Marowitz (1998) investigated the effect of the BAC at arrest, driving history, and other demographic factors on the one-year post-arrest probability of recidivism for DUI offenders through logistic regression models. Results indicated that high BAC at the time of arrest and prior 2-year traffic convictions contributed significantly to DUI recidivism (Marowitz 1998). The recidivism rate was increased during the BAC range of 0.09 g/dL to 0.29 g/dL. A high recidivism rate at high BACs indicated that DUI offenders might have a high dependency on alcohol daily (Marowitz 1998).

Roma et al. (2019) investigated DUI recidivism with BAC value on the license suspension report after DUI and the psycho-diagnostics tool Minnesota Multiphasic Personality

Inventory-2 (MMPI-2). Results showed that, compared to non-repeat offenders, repeat offenders had higher BAC at the time of their first conviction and more problematic MMPI-2 profiles, despite the presence of social desirability responding (Roma et al., 2019). The best prediction of recidivism was made with BAC and the scales of Lie (L), Correction (K), Psychopathic Deviate (4-Pd), Hypomania (9-Ma), and Low Self-Esteem (LSE) (Roma et al., 2019).

Besides BAC values, the time of drinking and substance use mixture are two important factors that differentiate between first-time offenders and repeat DUI offenders. Impinen et al., (2009) examined the DUI rearrest rate concerning other substance use and drinking patterns through a 15-year record. Two Cox proportional hazards models were constructed. Model 1 examined the difference among three subgroups: (1) alcohol only, (2) drug only, and (3) drugs combined with alcohol. Model 2 further examined the effects of different drug - alcohol combinations and the effects of alcohol only on the recidivism rate, with drug-drug mixtures excluded. Results of both models showed that young, males, with high BAC, DUI from Monday to Friday and from noon to midnight had the highest chance of recidivism (Impinen et al., 2009). It is important to note that the DUI recidivism defined here included drug-related DUI. In addition, to the best of the author's knowledge, this is the first study and only study that examined the time of day and the day of the week related to DUI recidivism.

2.2.2.3. Traffic Violations and Criminal History

Previous violations, crimes, and crashes also differentiate repeat DUI offenders from first-time offenders. A study based on a 12-year follow-up period of first-time DUI offenders showed that a driving history of traffic violations and crashes before the first DUI offense was a predictor of later recidivism (Cavaiola et al., 2007). Non-traffic-related violations or crimes could also be predictors, such as having a prior summary of an alcohol- or drug-related offense,

having a prior misdemeanor offense, having a misdemeanor arrest for a crime against persons, having a prior treatment episode, or loss of employment or expulsion from school because of drug or alcohol use (Marowitz 1998; Nochajski et al., 2000; Schell et al., 2006; Bouchard et al., 2012; Dugosh et al., 2013; Robertson et al., 2016).

However, traffic violations or crime history often being viewed as one variable or index in the literature. Only Cavaiola et al. (2007) noted reckless driving violations as a subgroup of traffic violations that was a significant predictor of DUI recidivism. Reckless driving behaviors maybe an indicator of a poor decision-making lifestyle rather than alcohol abuse (Cavaiola et al.2007). Examining traffic violations or crime history in smaller categories may provide more information about the motions of drink and driving, and further help the court determine the more appropriate education or treatment programs that are mandated.

2.2.2.4. Personality and Mental Health

Personality can predict DUI recidivism, and Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Hathaway & McKinley,1951; Graham, 1990) can be used to identify high-risk offenders in terms of DUI recidivism (Cavaiola et al., 2007). Iowa Gambling Task (IGT) can also be used to measure decision-making, and DUI reoffenders tend to have more disadvantageous decision-making (Bouchard et al., 2012). Antisocial attitudes can also be predictors of DUI (Jornet-Gibert et al., 2013), and the Jesness Inventory-Revised (JI-R; Jesness, 1996) was used to assess attitudes toward antisocial behavior.

2.2.3. Analytic Strategies

When it comes to analytic strategies, all reviewed literature regarding DUI recidivism adopted traditional statistical methods, including Analysis of Variance (ANOVA), chi-square test, T-test, multivariate analysis of variance (MANOVA), logistic regression, and survival

analysis (e.g., C'de Baca et al., 2001; Cavaiola et al., 2007; Impinen et al., 2009; Hubicka et.al., 2010; Bouchare et al., 2012; Robertson et al., 2016; Wickens et.al., 2018; Roma et al. 2019). These statistics methods were performed with well-developed softwares, such as SAS, SPSS, Stata, and R. These softwares are user-friendly and are easily operated by non-programmers, so they are popular among researchers from all practical research areas.

However, each of these statistical methods has as least one underlaying assumption, and all assumptions should be verified before applying these methods. If one assumption is violated, then researchers should apply techniques to fix the data or alternative methods that are suitable for the data type. For example, one type of survival analysis, the Cox proportional hazards model, assumes that the hazard ratio is constant over time. If this assumption is violated, then stratification should be performed on the data to reduce the time-dependent feature of the dataset first, and the stratified Cox's proportional hazards model should be applied for analysis then after (Lee and Wang, 2003; Hosmer et al., 2008). Fail to verify the underlying assumptions in statistical models or ignoring the violation of assumptions would lead to inaccurate results from the analysis (Lee and Wang, 2003; Hosmer et al., 2008).

Unfortunately, assumptions verification might be sometimes ignored by researchers. In the DUI recidivism literature reviewed here, none of the studies stated that they have checked assumptions, nor provided any test statistics that indicated they have done assumption verifications. It is also possible that this information is not necessary for the main reader group of DUI recidivism research. Calculation cost can be another issue for assumption verification and modeling when evaluating too many factors at one study.

In addition, among the statistical methods used in the DUI literature, the regression models are predictive analytics. Logistic regression and the survival analysis such as Cox

proportional hazards regression, are used to predict a likelihood. In the DUI literature, they can be used to predict the likelihood of DUI recisivism. However, researchers usually use logistic regression or survival analysis for explanation purpose, and no literature have demonstrated the predictions made from these regression models. A possible reason for this phenomenon is that the regression model may have high error rate and low prediction power. The regression algorithm only select strong predictors to enter in the model based on a significance level. These unselected weak predictors are not strong predictors when they are used individually, but the weak predictors can dramatically improve prediction accuracy when they are integrated together (Berk and Bleich, 2013). A prediction model with a lot weak predictors can be very difficult to interpret, so there is a trade-off between the interpretability and prediction accuracy (Berk and Bleich, 2013). Though weak predictors can somewhat improve the model prediction, easily interpretable functional forms are usually more popular.

2.2.4. Summary

The above subsections summarized insights from existing literature regarding the definition of DUI recidivism, the characteristics of repeat DUI offenders, and analytic strategies applied in these studies. There are three areas seems to be understudied: (1) evaluating traffic violations and crime in the subgroup, (2) evaluating the effects of environmental factors, (3) and a comprehensive study that integrates factors from multi-dimensions. Future studies can focus on these three areas.

2.3. Machine Learning Applications in Recidivism Prediction

A large database became available with the improved computer technologies in computing and storage in 1980. Some non-statistical ML methods were developed to process and analyze large complex datasets. These ML methods usually require no assumptions that are

usually needed by statistical methods and treat the data mechanism as unknown (Breiman, 2001). Generally, ML methods have two advantages over traditional statistical methods. Many of them can address non-linear relations between predicters and the response variable and automatically find interaction effects (Tollenaar & Van Der Heijden, 2019). In addition, ML methods can analyze complex data, such as noise data, with many correlated or irrelevant predictors (Tollenaar & Van Der Heijden, 2019). Traditional statistical methods usually can't handle such complex data. Because of these advantages, ML methods are expected to improve the predictive performance when datasets contents complex interactions between variables or non-parametric variables (Tollenaar & Van Der Heijden, 2019). There is no machine learning application in DUI recidivism literature, so the review of this new method was based on crime recidivism in general.

2.3.1. Machine Learning Prediction in Crime Recidivism

In past decades, different ML methods have been tried to predict recidivism risks. In an early study of ML in recidivism, Liu et al. (2011) compared logistic regression (LR), classification and regression trees (CART), and neural networks (NN) in the prediction of violent recidivism using a sample of 1225 male prisoners in the United Kingdom, followed up for approximately three years after release. The violent recidivism rate was 28.0% in this cohort, namely, 343 prisoners were reconvicted for repeat violent offenses, whilst 882 prisoners had at most non-violent recidivism, including 499 prisoners with no convictions. Twenty items in Historical Clinical Risk Management–20 (HCR-20; Webster et al., 1997) were chosen as predictors. Although NN slightly outperformed LR and CART, this result did not reach significance. The overall accuracy of the three models varied between 59% and 67%.

In Tollenaar and van der Heijden (2013), prediction results of LR and linear discriminant analysis about three response variables - general, violent, and sexual recidivism - were compared

with results from several ML methods based on available datasets of several offender databases in the Netherlands. The ML methods included in the comparison are recursive partitioning, adaptive boosting, logitBoost, NN, linear support vector networks, and k-nearest-neighbors classification. Overall, they found that the prediction accuracy of each method varied with datasets and predictors in the datasets and the predicted response variable, from 67% to 72.9%. They concluded that ML approaches to predicting criminal recidivism generally were not superior to traditional regression-based approaches. However, Tollenaar and van der Heijden (2013) didn't state the distribution of classes in the sample, nor was any sampling method used to keep class distribution balanced. The general recidivism data for the population in Tollenaar and van der Heijden (2013), was 1.2% for general, 1.9% for violent recidivism, and 1.4% for sexual recidivism, so the original sample was likely to be imbalanced.

However, Tollenaar and van der Heijden (2013) didn't include two important tree ensemble methods, Gradient Boosting (GB) and Random Forest (RF). Therefore, the conclusions were considered premature (Berk & Bleich, 2013; Tollenaar & Van Der Heijden, 2019). Berk and Bleich (2013) concluded that tree-based ML methods should be included in the comparison since they have several advantages over LR: the ability to predict response variables with more than two classes.

Berk and Bleich (2013) performed stochastic Gradient Boosting (SGB), RF, and LR on a dataset of 25,000 observations to predict rearrest for a serious crime within two years of release on probation, with eight predictors. The recidivism rate in this sample was approximately 13%, so this sample was class-imbalanced data. Berk and Bleich (2013) concluded that RF performed better than LG and SGB, while SGB performed as well as LR based on model error - class predicted incorrectly divided by the total number in that class. The author calculated accuracy,

precision, recall, and F1 score based on the confusion matrixes provided by Berk and Bleich (2013). For RF, the abovementioned indices were 71.0%, 26.3%, 62.8%, and 37.1%; for LG, these indices were 66.3%, 21.4%, 55.6%, and 30.9%; and for SGB, the abovementioned indices were 63.4%, 22.5%, 58.25 and 32.5%. Based on these indices, RF performed the best among the three algorithms.

Hamilton et al. (2015) compared the predictive accuracy of the Washington State Static Risk (RSA) Assessment using traditional LR, NN, and RF methods in a large sample of all corrections clients (felony, drug, violent, sex) who were repeat offenders in the state of Washington (N = 297,600). LR and ML approaches demonstrate comparable performance. Since Hamilton et al. (2015) attempted to predict multiple offenses with each offense having multiple levels, the class distributions in this sample were not clearly described.

Duwe and Kim (2017) further examined the performance of newer ML approaches relative to traditional methods in predicting recidivism. They compared the predictive accuracy of 12 supervised learning algorithms. The data set used in the study was derived from that used to develop the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR; Duwe, 2014) and comprised 24,917 male offenders released from prisons in Minnesota. There were 4,497 repeat offenders, which account for 18% of the sample. Results suggested that newer ML approaches such as LogitBoost (accuracy = 84.2%, precision=44.9%, recall=26.8%), RFs (accuracy = 83.7%, precision=51.7%, recall=24.1%), and MulitBoosting (accuracy = 84.1%, precision=44.8%, recall=26.8%) were found to yield better results for general recidivism.

A few more recent studies reported some promising findings. Ozkan et al. (2020) examined 336 predictors in a sample of 3,061 juveniles in Florida about sexual offense recidivism. In this sample, 317 juveniles recidivated with a new sex offense, representing 10.4%

of the sample. Ozkan et al. (2020) also found that RF models yielded strong findings with areas under the ROC curve (AUCs) of 0.71 for an “all-predictors model” and 0.65 for a “legal factors” model.

Ghasemi et al. (2021) applied decision trees, random forests, and support vector machines to two datasets, with 72,725 records in dataset one and 26,450 records in dataset two. Both data were provided by the Ontario Ministry of Community Safety and Correctional Services (MCSCS). Recidivism for both data sets was defined as any criminal offense that led an individual returning to the MCSCS system on a reconviction, sentenced to either incarceration or community supervision (Ghasemi et al. 2021). The overall recidivism rate for the two datasets combined is 31.98%, indicating a class-imbalance issue in this sample. Accuracy for decision trees, random forests, and support vector machines were reported as a performance measure and were 69.5%, 73.6%, and 70.4% respectively. Again, random forest performed the best.

2.3.2. Summary

In the last decade, machine learning applications have made significant contributions to health care, business, and entertainment. Researchers in criminal justice decision-making also embraced these applications to assist with risk assessment. The performance of these applications varied with different data structures and the outcome of predictions. Among all machine learning models, the random forest seems to yield the best predictions.

Crime recidivism is usually a low-chance event, which leads the dataset with a class-imbalanced structure. Predictions made on class-imbalanced data may be less accurate than predictions made on balanced data. For example, for a sample that repeat offenders only account for 10%, a model can predict all records in test data as non-repeat offenders and still achieve 90% accuracy. In this case, the model completely fails to detect recidivism, and the performance

measure, accuracy, fails to reflect this fact. Unfortunately, none of the crime recidivism literature reviewed here noted the class-imbalanced issue nor attempted to deal with it.

In addition, since the prediction of recidivism is predicting human behaviors, there is one important rule to correctly understand the prediction: accurate predictions require the future be substantially like the past (Berk and Bleich, 2013). However, it is not always the case in the reality. People's mind and behavioral patterns can change over time. A method that can make accurate prediction in the short period may not produce accurate forecast during a long time. The duration of recidivism research varies from two years to more than ten years, and it is likely that offenders' behavioral patterns changed during this period.

2.4. Class-Imbalanced Data Handling

In real-world machine learning applications, the data imbalance imposes challenges to performing data analytics and producing accurate results. The raw dataset often suffers from the skewed data distribution of one class over the other class (Kaur et al. 2019). The performance of classifier**s** on such datasets leans towards the majority class. Thus, the solutions lean toward better accuracy in the majority class and result in poor accuracy in the minority class. The problem of imbalanced data distribution is a common problem in machine learning applications that try to detect rare events, such as fraud detection, crash detection, tumor detection, and so on. handling imbalanced datasets has been intensively studied by researchers (Wang, et al. 2020). Systematic reviews on data imbalanced issues and their solutions can be found in Haixiang et al. (2017), Spelmen and Porkodi (2018), and Kaur et al. (2019). Given the DUI recidivism is a rare event, techniques are needed to address the data imbalance issue.

Among all the approaches proposed to handle imbalanced data by Kaur et al. (2019), the sampling method is the most widely used approach (Haixiang et al. 2017; Spelmen and Porkodi,

2018; Santos et al. 2018). There are three sampling method categories: oversampling, undersampling, and hybrid-sampling. Oversampling strategically replicates the minority classes, while undersampling strategically removes the majority classes. Hybrid-sampling is a mix application of the former two. Because oversampling approaches can keep the variations among the minority class (Haixiang et al. 2017; Spelmen and Porkodi, 2018), it is the most appropriate approach for data with a small sample size.

It is important to note that oversampling techniques should be applied cautiously in a joint application with cross-validation (Santos et al. 2018). Cross-validation is a standard procedure to evaluate classification performance and select optimal hyperparameters. Incorrectly applying oversampling while performing cross-validation may derive from two main issues: overoptimism and overfitting (Santos et al. 2018). Oversampling should be performed in the training sets at each iteration of a cross-validation process. Meanwhile, the test sets at each iteration of a cross-validation process should be kept original during the whole process. Incorrectly oversampling the entire dataset would lead to a structure change in the test data. In this case, the size of the minority group also increased in the test data, and performance indices calculated based on this data structure fail to reflect the true evaluation of the original data structure. (Santos et al. 2018).

Santos et al. (2018) compared 12 well-established oversampling algorithms based on data complexity analysis. The best oversampling techniques shared three key characteristics: the use of cleaning procedures, cluster-based example syncretization, and adaptive weighting of minority examples (Santos et al. 2018). Among all algorithms, the Synthetic Minority Oversampling Technique coupled with Tomek Links (SMOTE - Tomek Links) and Majority Weighted Minority Oversampling Technique (MWMOTE) achieved the best results. Both algorithms

changed the overlapping areas in the data and increased the discriminative power of data (Santos et al. 2018).

There is a well-developed software program designed to achieve oversampling techniques. Lemaître et al. (2017) presented an integrated Python library called *Imbalanced-learn* for data-level resampling for imbalanced classification. It can be treated as an extension of *Scikit-learn*, a Python library that integrates a wide range of state-of-the-art machine-learning algorithms and provides elementary methods to deal with class-imbalanced issues (Pedregosa et al. 2011). *Imbalanced-learn* library integrated many oversampling techniques including the SMOTE - Tomek Links, but MWMOTE is not available in the library. This library largely reduced the amount of coding to oversample the data and greatly benefited researchers with less programming skills to improve the predictions with a more balanced dataset.

3. DATA

The scope of this study mainly focused on drivers in North Dakota. The state agency approved records from North Dakota Driver License (NDDL) administrative system for limited use within studies. This study was reviewed by the NDSU Institutional Review Board (IRB) with minimal risk to subjects.

The data source from NDDL consists of six types of driver records: master files, conviction and crashes records, licensing files, arrests records, DUI/BAC record, and driving training record. Each driver has a unique record ID which will be used as the primary key to connect information from different records. These datasets were available from 2011 to 2020 for research purposes. Two separate data files regarding drivers' total DUI count came from the administrative hearing results, and their intent to be organ donors was also provided by NDDOT. In addition, US Census County demographic and US Department of Justice law enforcement officer counts by county were collected (US Census Bureau 2018; US Department of Justice and Federal Bureau of Investigation 2019). By connecting the offender's residence county with the ND county profile, the offender's demographic information can be obtained, and law enforcement indexes can be estimated at the county level.

This study selected DUI first-time offenders from 2013-2017 as a study group, with a follow-up period of 3 years (1095 days) from the dates of their first DUI citations. A subsequent DUI citation within the follow-up period was considered DUI recidivism, and drivers in this group were defined as repeat offenders (ROs). Drivers without a DUI recidivism within the follow-up period were defined as non-repeat offenders (NROs). In the total of 12,879 records analyzed in this study, there are 11,651 offenders without recidivism and 1,228 offenders with recidivism. A binary response variable was created to define recidivism: 1 represents the record

with DUI recidivism - repeat offenders (ROs), while a value of 0 represents the record without DUI recidivism – non-repeat offenders (NROs).

Health and social-economic factors were shown to associate with DUI (e.g., Room, 2005; Impinen et al., 2011; Wieczorek,2013). To measure community health and social-economic factors at the driver's county of residence, the 2017 County Health Rankings for North Dakota were used. This data was provided by the University of Wisconsin Population Health Institute (UWPHI), aiming to build awareness of the multiple factors that influence health and support leaders in growing community power to improve health equity. This data contained more than 60 measures of health factors of nearly every county in all 50 states. It provided indices for health behaviors (30%) and clinical care (20%). Social and economic factors (40%) and physical environment (10%) were modeled from more than 20 national data sources (UWPHI, 2017). In this data source, 27 indices for all 53 counties in North Dakota were selected and linked to the driver's county of residence for analysis. Table 3.1 described the list of variables examined in this study.

The data preparation started with filtering the first- and second-time offenders in this dataset. Meanwhile, the conviction and crash records in was merged into the master file of that year for each year from 2013-2020, then the yearly datasets were aggregated into one aggregated dataset to include 7-year conviction and crash record. After that, the aggregated 7-year conviction and crash record was linked to the subset of the DUI count file by Record_id, to verify whether the total DUI count that came from the administrative hearing result matched the total number of DUI convictions in the system. North Dakota only kept DUI records for seven years in the system, and the offense that occurred in the eighth year will be re-counted as the first offense. The data was transformed to contain all the DUI convictions in one row for each ID.

For the first-time offender, the records with only one DUI conviction in 2013 - 2017, and no second DUI offense within 1095 days were selected. To select second-time offenders, three filters were applied: (1) only two DUI convictions in 2013-2020, (2) the first DUI conviction happened in 2013- 2017, and (3) the second DUI conviction happened within 1095 days of the first DUI conviction.

To link the traffic violation and crash history, the conviction and crash records were aggregated again from 2011 to 2017 and linked to the filtered DUI offender list. Time interval markers were created to count the record of each citation category in Table 3.1 and crash in four monitoring time intervals: (1) 0 to 60 days prior, (2) 61 to 365 days prior, (3) 365 to 730 days prior; (4) 731 to 1095 days prior. The count of each citation category or crash in 3-year before the first DUI was the sum of the total counts in four monitoring time intervals. Then the binary version of each of these variables was created. Due to the data availability, the offenders who had their first DUI offense in 2013 had historical records of less than three years.

Offenders' BAC records came from the DUI/BAC subset. Because nearly one-third of the sample refused the BAC test on their first DUI date, data imputation was performed to fill in the missing values with the sample mean in that field. Machine learning models were performed in this imputed dataset, while the two logistic regression models were performed on the original data to test BAC values and BAC refusal status separately since the logistic regression models needed a much smaller calculation capacity.

Health and social-economic factors were then linked to the data by the county code of the driver's residence. US Census County demographic and US Department of Justice law enforcement officer counts were also linked to the data by the county code. As discussed in Section 2.2.1 that the level of law enforcement should be considered when evaluating DUI

recidivism based on official records (Nochajski & Wieczorek, 2006), so a law enforcement density index was created with the number of officers in each county divided by the population in that county.

Table 3.1. List of Variables.

Variable Name	**Variable Description**
Second DUI citation	Target variable: yes=1, no=0
RecID	Driver identification variable
Profile	
Gender	Driver's gender:1=male, 0= female
Age	Age of first offense: date of birth - conviction date of first DUI Age was categorized into six groups:18-24; 25-34; 35-44; 45-54; 55-64; 64+
County	Residence county
Court region	ND District Court Region
Population density	Census population 2017 per square mile in the residence county
Registered organ donor	Yes=1, no=0
Status1N	Driver license status code at first DUI citation. 1= LISPR; license suspended for 24/7 program 2= LIS; license suspended for other reasons 3=RO; license revoked 4=other; all other licenses, with 90% being licensed
Law enforcement index (LEI)	
Law enforcement density	Number of officers in county/population in the county
Behavioral treatment interventions	
Intervention	The 24/7 Sobriety Program participant =1, no= 0
Seasonal factors	
Weekend	The first DUI conviction date was on the weekend: Yes=1, no=0
Holiday	The first DUI conviction date was on holiday: Yes=1, no=0
Season	Winter: Nov, Dec, Jan, Feb; 1 for winter Summer: Jun, Jul, Aug; 2 for summer Other: Mar, Apr, May, Sep, Oct; 3 for other

(Table 3.1 List of variables continuing to the next page)

Table 3.1. List of variables (Continued)

Variable Name	Variable Description
Traffic Violation and Crash History	
High-risk DUI-related citation binary*	High-risk DUI-related citation before or during the study period: yes=1, no=0 for each time interval. **
High-risk DUI-related reckless driving citation binary*	High-risk DUI-related reckless driving citation before or during the study period: yes=1, no=0 for each time interval. **
High-risk non-DUI-related citations binary*	High-risk non-DUI-related citations before or during the study period: yes=1, no=0 for each time interval. **
High-risk non-DUI-related improper driver action citations binary*	High-risk non-DUI-related citations before or during the study period: yes=1, no=0 for each time interval. **
High-risk non-DUI-related careless driving citations binary*	High-risk non-DUI-related related improper driver action citations before or during the study period: yes=1, no=0 for each time interval.**
High-risk non-DUI-related restriction violation citations binary*	High-risk non-DUI-related restriction violation citations before or during the study period: yes=1, no=0 for each time interval.**
High-risk non-DUI-speeding citations binary*	High-risk non-DUI-related speeding citations before or during the study period: yes=1, no=0 for each time interval.**
Low-risk citation binary*	Low-risk citation before or during the study period: yes=1, no=0 for each time interval. **
Low-risk improper driver action citation binary*	Low-risk improper driver action citation before or during the study period: yes=1, no=0 for each time interval. **
Low-risk seatbelt citation_binary*	Low-risk seatbelt citation before or during the study period: yes=1, no=0 for each time interval. **
Low-risk speeding citation binary*	Low-risk speeding citation before or during the study period: yes=1, no=0 for each time interval. **
Low-risk road sign violation citation binary*	Low-risk road sign violation citation before or during the study period: yes=1, no=0 for each time interval. **
Crash binary	Crash before or during the study period: yes=1, no=0 for each time interval. **
BAC highest	The highest blood alcohol concentration (BAC) on record
BAC mean	The average blood alcohol concentration (BAC) on record
BAC median	The median blood alcohol concentration (BAC) on record
BAC refusal	BAC test was refused by the driver: yes=1, no=0
*High-risk citations: citations with three or more points deduction on driver's license; *Low-risk citations: citations with two or fewer points deduction on driver's license **Monitoring Time Intervals around first DUI citation: Pre-DUI citation: (1) 0 to 60 days prior, (2) 61 to 365 days prior, (3) 365 to 730 days prior; (4) 731 to 1095 days prior; (5) prior 3-year	

(Table 3.1 List of variables continuing to the next page)

Table 3.1. List of variables (Continued)

Variable Name	Variable Description
Health and Social-economic factors at the county level	
Poor or fair health	Percentage of adults reporting fair or poor health (age-adjusted)
Frequent physical distress	Percentage of adults reporting more than 14 days physically unhealthy days in the past 30 days (age-adjusted)
Frequent mental distress	Percentage of adults reporting more than 14 days mentally unhealthy days in the past 30 days (age-adjusted)
Adult smoking	Percentage of adults who are current smokers
Adult obesity	Percentage of adults that report a BMI of 30 or more
Food environment index	Index of factors that contribute to a healthy food environment, 0 (worst) to 10 (best)
Physical inactivity	Percentage of adults aged 20 and over reporting no leisure-time physical activity
Access to exercise opportunities	Percentage of population with adequate access to locations for physical activity
Excessive drinking	Percentage of adults reporting binge or heavy drinking
Alcohol-impaired driving deaths	Percentage of driving deaths with alcohol involvement
Teen births	Teen birth rate per 1,000 female population, ages 15-19
Uninsured	Percentage of population under age 65 without health insurance
Primary care physicians	The ratio of population to primary care physicians
Mental health providers	The ratio of population to mental health providers
High school graduation	Percentage of the ninth-grade cohort that graduates in four years
Some college	Percentage of adults ages 25-44 years with some post-secondary education
Unemployment	Percentage of population ages 16 and older unemployed but seeking work
Children in poverty	Percentage of children under age 18 in poverty
Income inequality	The ratio of household income at the 80th percentile to income at the 20th percentile
Children in single-parent households	Percentage of children that live in a household headed by a single parent
Social associations	Number of membership associations per 10,000 population
Violent crime	Number of reported violent crime offenses per 100,000 population
Severe housing problems	Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities
Long commute - driving alone	Among workers who commute in their car alone, the percentage that commutes more than 30 minutes
Insufficient sleep	Percentage of adults reporting not getting enough rest or sleep in the past 30 days
Median household income	Median household income
Rural	Percentage of the population lives in the rural area

4. METHODOLOGY

This study will use three tree-based machine learning algorithms to predict DUI recidivism: Decision Tree, Random Forest, and Gradient Boosting. The prediction power of the three algorithms was evaluated and compared. Decision Tree is a single tree method, while Random Forest and Gradient Boosting are ensemble methods based on a combination of multiple training Decision Trees. Thus, Random Forest and Gradient Boosting methods can be treated as the upgraded version of the Decision Tree method. All three algorithms can predict either the numerical response variable, the regression case, or the categorical variable, the classification case. Since the response variable in this study is a categorical variable with two classes, a DUI recidivism occurs on a driver and no DUI recidivism occurs, the classification version of these three algorithms was used. Detailed descriptions of these three algorithms are discussed in the subsection below.

Generally, when building a machine learning algorithm, the original dataset was usually split into training and testing subsets. The training data is used to "teach" the model the concepts that are useful for prediction, and the testing data is reserved for testing if the trained model has successfully grabbed the essence of things. The confusion matrix was used to evaluate the final model performance.

Prediction performance is mainly depending on the hyperparameters used in the model. Thus, hyperparameter optimization is necessary to select model parameters for the best performance. This study used hyperparameter optimization with five-fold cross-validations to choose model parameters for their best performance. Python Package *Scikit-learn* (Sklearn) was used to perform hyperparameter optimization, classification, and model evaluation.

The dataset was imbalanced, with repeat offenders only accounting for 9.5% of the sample. Thus, predictions made on such datasets might be less likely to favor the minority group – the repeat offenders. To address this issue, Synthetic Minority Oversampling Technique (SMOTE) - Tomek Links technique was used to oversample the minority groups in the training data in all three models. Six predictions were made: Decision Tree, Random Forest, and Gradient Boosting based on the original dataset and prediction models created with the oversampled dataset. Python Package *Imbalanced-learn* was used to oversample the minority group.

For practical research areas like impaired driving, prediction is not enough. Interpreting factors influencing the prediction is also important to help readers do the best research. Machine learning is also well-known as a black-box prediction, in which the variable interpretability is usually low. Machine learning models were used as variable selection methods to rank top performance variables to predict DUI recidivism to overcome this issue. The logistic regression model was used to quantify the effects of selected factors/predictors on the outcome. SAS 9.4 was used to perform descriptive analysis and logistic regression analysis.

4.1. Decision Tree

The Decision Tree algorithm is a non-parametric supervised learning technique for regression and classification. This algorithm has a few advantages: (1) It can perform predictions with both categorical predictors (usually referred to as features in machine learning terminology) and numerical predictors (compared to the Support Vector Machine algorithm). (2) It starts the prediction with the most relevant feature to the least relevant feature so that it can rank the importance of input features. Since the response variable in this study is a categorical variable with two classes, a DUI recidivism occurs on a driver and no DUI recidivism occurs, the classification tree algorithm will be used.

As shown in Figure 4.1, the training of a classification tree works like a flowchart that it inputs one feature at a time and applies a logic question of the feature to split the tree until it reaches a terminal node of the tree (Esposito & Esposito, 2020). Three steps complete this training process:

1. The root node has a splitter that contains a logic question of a feature. Each branch, a child node or terminal node, from a splitter is an answer to the logic question of the feature in that splitter.
2. Each child node then acts as a root node and splits again.
3. Step 2 was repeated until all branches reached the terminal node, containing a pure set of one class.

Thus, by default, each terminal node in the training step is a class of the response variable. After the training step, a portion of the dataset will be used as test data to test the model's accuracy. Once model accuracy reaches the acceptable level, the trained classification tree is ready to predict unknown outcomes using known features trained in the classification tree.

During the training stage, there are two algorithms used for classifying the features, in other words, splitting a feature in a node: Classification and Regression Trees (CART) and Iterative Dichotomiser 3 (ID3)/C4.5 (an improved classification algorithm version of ID3). Both algorithms measure the inequality of the data distribution (impurity) of the dataset split in each node. The lower the impurity, the higher the homogeneity. However, CART and ID3/C4.5 measure impurity differently. Because of the different measures in impurity, ID3/C4.5 is more susceptible than CART to outliers. As indicated in Equation 1, CART measures impurity using the Gini index, the sum of the squared frequency of feature values (Esposito & Esposito, 2020).

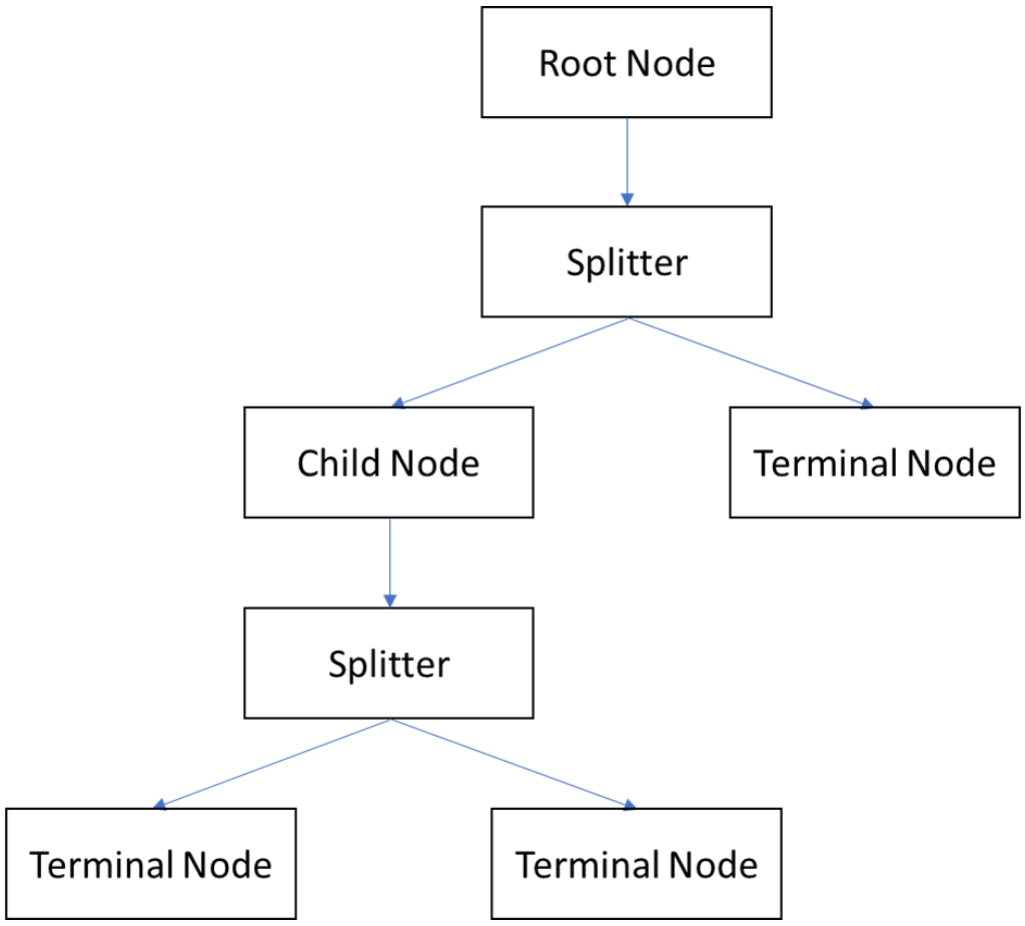

Figure 4.1. Sample Classification Tree Algorithm

$$Gini\ (S)\ =\ 1 - \sum_{i=1}^{k} P_i^2 \qquad \text{(Equation 1)}$$

Where S is the full dataset at the root node or the sub-dataset at the child node before splitting, K is the number of the total categories in a feature (if a feature is a numerical variable, then K = 2; a categorical with values less than the mean value and one with values larger or equal to the mean value), P_i is the frequency of ith category (categorical variable) or element within the given value (numerical variable) in the dataset.

ID3/C4.5 measures impurity based on the concept of entropy, which results from the sum of the frequency multiplied by the logarithm of the frequency of the feature value, as shown in Equation 2 (Esposito & Esposito, 2020).

$$H\ (S)\ =\ \sum_{i=1}^{k} -P_i * \log_2(P_i) \qquad \text{(Equation 2)}$$

Where S, K, and P_i indicate the same parameter as in Equation 1.

Besides, CART and ID3/C4.5 perform classification (splitting a feature in a node) differently. CART can reuse the same feature over multiple splits, while ID3/C4.5 stops using a feature after a split has been made (Esposito & Esposito, 2020). This difference results in CART typically producing larger trees and more chances to contain better splits. However, on the other side, CART has more chances of overfitting. Overfitting happens when a model learns the training data so well that the noise or random fluctuations in the training data is considered and learned as concepts by the model. In contrast, these concepts do not apply to new data and negatively impact the model's generalization ability. To save calculation costs, ID3/C4.5 was used in this study.

4.2. Random Forest

A single Decision Tree can make a good prediction but might not make the best prediction. For example, a Decision Tree algorithm tends to overfit training data which can give poor results when applied to the full data set. Having a forest of trees - the Random Forest algorithm - could limit overfitting without substantially reducing prediction accuracy. Random Forest is an ensemble method that takes multiple individual learning models and combines them to produce an aggregate model that is more powerful than any of its learning models alone. This is because each model might overfit a different part of the data. After combining other individual models into an ensemble, their mistakes were averaged to reduce the risk of overfitting while maintaining strong prediction performance.

The Random Forest algorithm is a bootstrap aggregating algorithm (often shortened to bagging) and was introduced by Breiman (2001). It builds Decision Trees independently and

parallelly, then returns a weighted average or majority vote of their results from the trees as the final result. As a result, the final model is more likely to be balanced between potential bias (underfit) and variance (overfit) compared to the Decision Tree.

The Random Forest algorithm has two important characteristics: (1) randomly selected features and (2) bagging (Esposito & Esposito, 2020). In the Random Forest algorithm, each training tree is constructed on a randomly selected subset of features, so the number of features used to split nodes for each Decision Tree is controlled. This characteristic can help to mitigate overfitting.

Bagging is the process of building a combination of weak learners (e.g., Decision Trees) based on bootstrapping samples and aggregating (either majority voting or averaging) the models learned on each bootstrapping sample. Bootstrapping is a statistical resampling method used to create test samples by randomly selecting observations (rows) in the original dataset with replacement. The resulting bootstrap sample has the same number of rows as the original training set, but possibly some rows from the original dataset are missing, and others occur multiple times. This characteristic can help to mitigate the imbalance of the training dataset (e.g., for a feature with two classes, one class is in 90% of the rows while the other class in only in 10% of the rows).

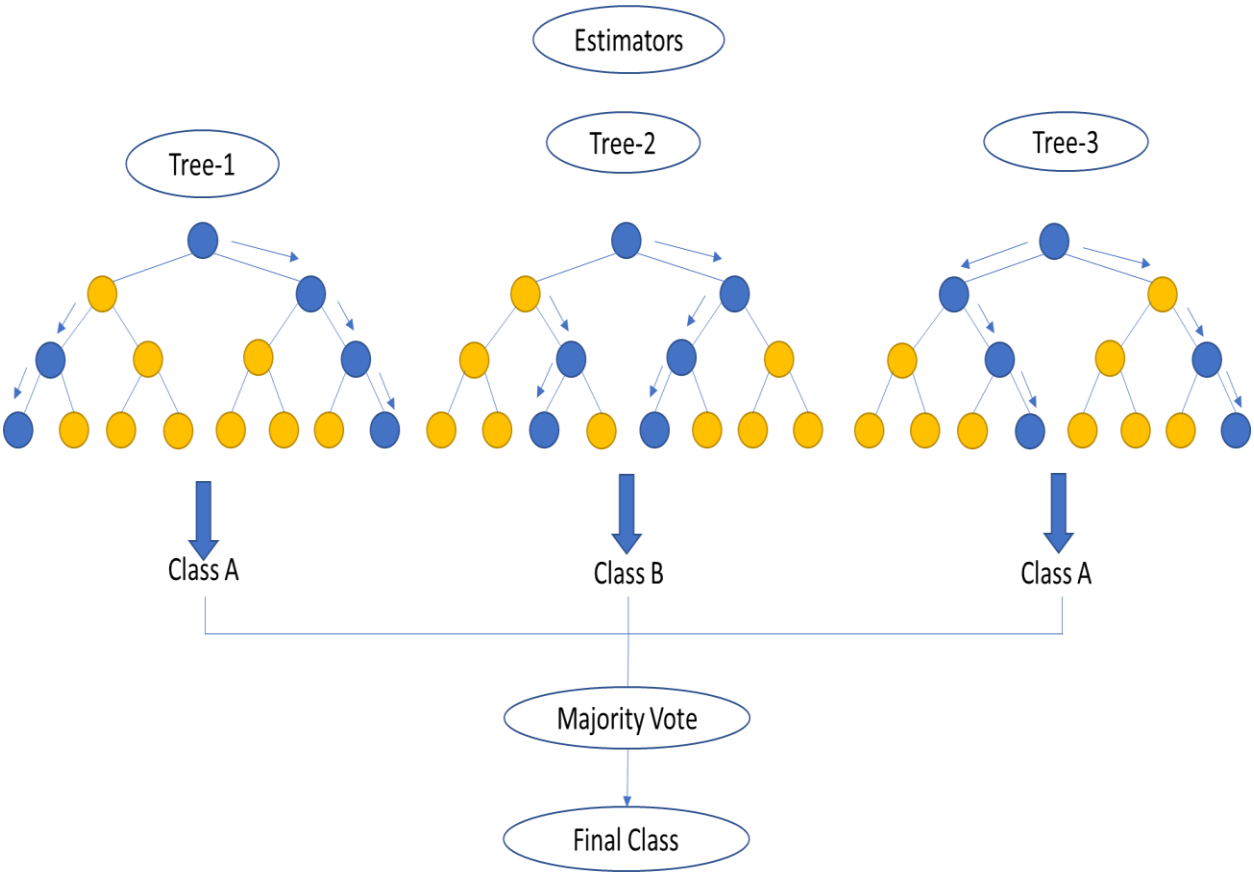

Figure 4.2. Structure of a Random Forest

Generally, the training process of a Random Forest classification algorithm follows four steps (Zhou et al., 2020): (1) The bootstrapping method was applied to randomly resample a dataset that has the same size as the original dataset to build a Decision Tree. (2) K features were randomly selected from total M features where K < < M (typically, K is chosen to be equal to the square root of M). (3) A combination of parallel Decision Trees was built by using the bootstrapping sample and chosen features from steps 1 and 2. (4) A majority vote was made on all the predictions made in steps (3). Figure 4.2 provides the structure of the Random Forest algorithm.

4.3. Gradient Boosting

The Gradient Boosting method is another tree-based machine learning algorithm, and it is proposed by Friedman (2002, 2003) at Stanford University. Thus, it contains all the advantages of tree-based algorithms mentioned in section 4.1. The Gradient Boosting method is also a type of ensemble method, while it builds trees sequentially rather than in parallel (as with the Random Forest algorithm). The subsequent tree was built based on the errors captured in the previous tree. Thus, the Gradient Boosting algorithm reduces bias, making it more accurate than the Random Forest algorithm. Other advantages of the Gradient Boosting method include handling large datasets without pre-processing, resistance to outliers, handling of missing values, robustness to complex data, and resistance to over-fitting (Friedman and Meulman 2003; Lu et al. 2020).

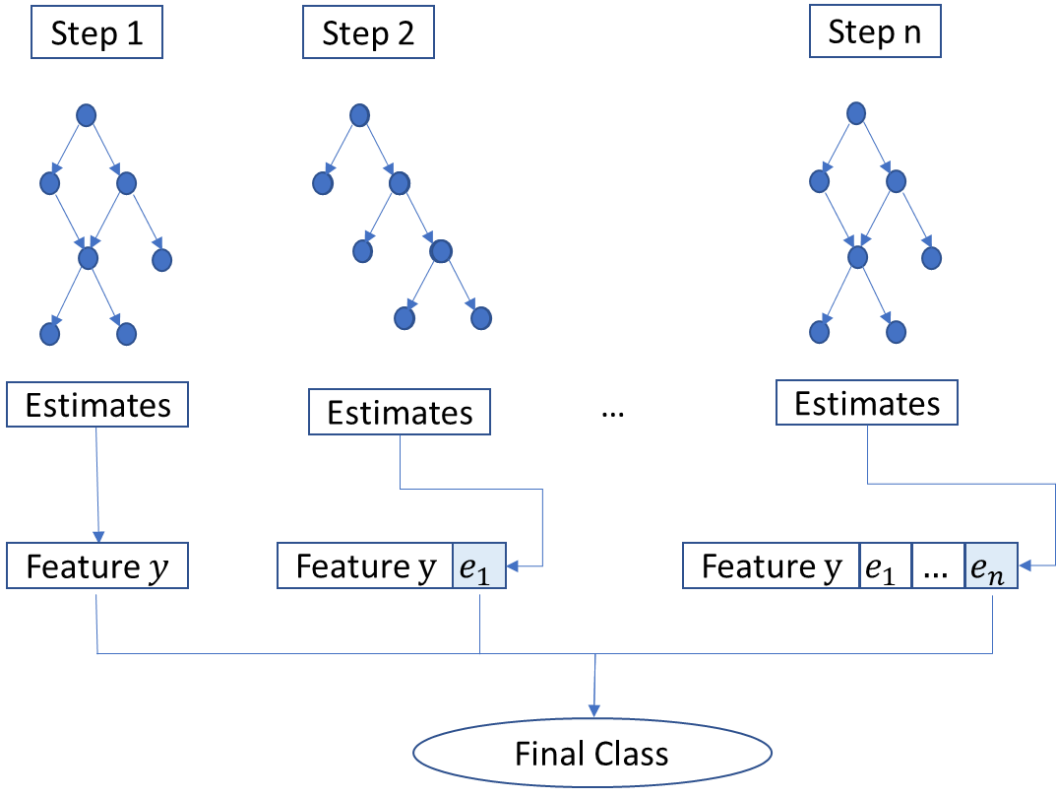

Figure 4.3. Steps of a Gradient Boosting algorithm

Generally, the training process of a Gradient Boosting algorithm can be described in three steps: (1) A Decision Tree was created and trained to fit the dataset at an acceptable level and contained a certain level of errors. (2) A subsequent Decision Tree was created and trained based on the residual errors of the previous tree. (3) Step (2) was repeated until errors were minimized or in any way acceptable for the problem at hand. Figure 4.3 provides a detailed graphical representation of the Gradient Boosting algorithm (Esposito & Esposito, 2020).

A detailed algorithm of Gradient Boosting for binary classification can be described as follows:

$$F(x) = sign\left(\sum_{i=0}^{m} f(x)\right) = sign\left(\sum_{i=0}^{m} \alpha_i f_i(x)\right) \quad \text{(Equation 3)}$$

Where $F(x)$ is the final predict output that contains only two classes, +1 and -1, $f_i(x)$ is the classifier at ith steps, α_i is the coefficients at ith steps to weight the classifier at that step (Freund & Schapire, 1999; Guestrin,2015; Esposito & Esposito, 2020; Lu et al., 2020).

4.4. SMOTE -Tomek Links

SMOTE is one of the most popular oversampling techniques developed by Chawla *et al*. (2002). SMOTE generates examples based on the distance between each data (often using Euclidean distance) and the minority class's nearest neighbors, so the created examples are distinct from the original minority class. Random oversampling only copies some random examples from the minority class.

The procedure for creating the synthetic samples is, in essence, as follows.

1. Choose random data from the minority class.
2. Calculate the Euclidean distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.

4. Repeat the procedure until the desired proportion of the minority class is met.

In contrast to the original oversampling method, this method adds new "information" to the data since the generated synthetic data are reasonably close to the feature space of the minority class.

Tomek Links is a modification from the Condensed Nearest Neighbors (CNN, not to be confused with Convolutional Neural Network) undersampling technique developed by Tomek (1976). The Tomek Links technique employs the rule to choose the pair of observations (let's say, a and b) that are fulfilled these qualities, as opposed to the CNN method, which just randomly selects the samples with its k nearest neighbors from the majority class that wants to be deleted.

1. The observation a's nearest neighbor is b.
2. The observation b's nearest neighbor is a.
3. Observations a and b belong to a different class. That is, a and b belong to the minority and majority class (or *vice versa*), respectively.

Mathematically, it can be expressed as follows.

Let d(x_i, x_j) den the Euclidean distance between x_i and x_j, where $x_i$denotes the sample belonging to the minority class, and x_j denotes the sample belonging to the majority class. If there is no sample, x_k satisfies the conditions (1) d(x_i, x_k) < d(x_i, x_j), or (1) d(x_j, x_k) < d(x_i, x_j), then the pair of (x_i, x_j) is a Tomek Link. By using this method, one can locate the desired samples of data from the majority class that have the smallest Euclidean distance from the data from the minority class (i.e., the data from the majority class that is closest to the minority class data, thus make it ambiguous to distinct), and then eliminate it.

SMOTE-Tomek Links was first introduced by Batista et al. (2003). It combines SMOTE's capacity to provide synthetic data for minority classes with Tomek Links' capacity to eliminate data from the majority class classified as Tomek links (that is, samples of data from the majority class are closest to the minority class data). The SMOTE - Tomek Links procedure is as follows:

1. (Start of SMOTE) Choose random data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Repeat steps 2–3 until the desired proportion of the minority class is met. (End of SMOTE)
5. (Start of Tomek Links) Choose random data from the majority class.
6. If the random point's nearest neighbor is the point from the minority class (i.e., create the Tomek Link), then remove the Tomek Link.

4.5. Model Assessment

The model assessment in this study was based on a confusion matrix tested on the test data part. Accuracy is a subtle indicator and might not be truly useful if taken alone. Accuracy measures the percentage of good predictions - when an offender with recidivism was recognized as a reoffender and when an offender without recidivism was recognized as an offender without recidivism. However, in the case where a class is a minority class, the accuracy rate can be useless. For example, in the case of this study, the goal is to recognize reoffenders who account for 10% of the total data entries; a 90% accuracy rate can mean that none of the reoffenders was

recognized, and offenders without recidivism were recognized as reoffenders. Therefore, a confusion matrix is needed to produce three important indicators: precision, recall, and F1 score.

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This matrix provided a holistic view of how well the classification algorithms perform and what errors they make. The confusion matrix is in a 2 x 2 matrix for a binary classification problem, as shown below in figure 4.4. The target variable has two values: positive (offender with recidivism) and negative (offender without recidivism). The columns represent the actual values of the target variable. The rows represent the predicted values of the target variable. There are four possibilities:

1. True Positive (TP): The predicted value matches the actual value. The actual value was positive, and the model predicted a positive value, meaning that an offender with recidivism is predicted as an offender with recidivism.
2. True Negative (TN): The predicted value matches the actual value. The actual value was negative, and the model predicted a negative value, meaning that an offender without recidivism is predicted as an offender without recidivism.
3. False Positive (FP) – Type 1 error: The predicted value was falsely predicted. The actual value was negative, but the model predicted a positive value, meaning that an offender without recidivism is predicted as an offender with recidivism. FP is also known as the Type 1 error.
4. False Negative (FN) – Type 2 error: The predicted value was falsely predicted. The actual value was positive, but the model predicted a negative value, meaning that an

offender with recidivism is predicted as an offender without recidivism. FN is also known as the Type 2 error.

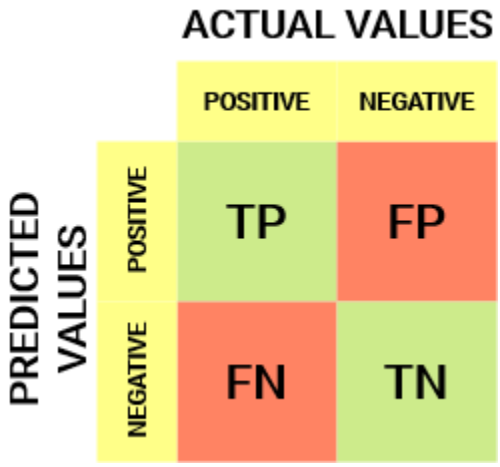

Figure 4.4. A sample 2 x 2 confusion matrix.

Starting from this confusion matrix, two important performance indicators are generated to estimate the precision of the model: recall and precision. The calculation of recall is indicated in Equation 4 below:

$$Recall = \frac{TP}{TP + FN} \quad \text{(Equation 4)}$$

Recall indicates the percentage of true positives the model predicts with respect to the total number of actual positives in the dataset. The calculation of recall is indicated in Equation 5 below:

$$Precision = \frac{TP}{TP + FP} \quad \text{(Equation 5)}$$

Precision indicates the percentage of true positives the model predicts with respect to the total number of actual positives in the dataset. In addition, the model accuracy can also be calculated using a confusion matrix, as shown in Equation 6:

$$Accuracy = \frac{TP + TN}{Total} \quad \text{(Equation 6)}$$

F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad \text{(Equation 7)}$$

Precision is important in music or video recommendation systems, e-commerce websites, etc. Wrong results could lead to customer churn and be harmful to the business. The recall is important in medical cases where it does not matter whether we raise a false alarm, but the actual positive cases should not go undetected. In this study, both indicators are important. The higher recall values, precision, and accuracy mean the models perform better. Thus, the F1 score was used as the primary evaluation metric for model performance.

4.6. Logistics Regression

Logistic regression (also known as the logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as recidivism occurring or not occurring, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds—the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and the following formulas represent this logistic function:

$$logit\ (\text{pi}) = \log(\tfrac{\text{p}}{1-\text{p}}). \quad \text{(Equation 8)}$$

$$\log\left(\frac{p}{p+1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i \quad \text{(Equation 9)}$$

Where p_i is the probability of DUI recidivism; x_i is ith predictor variable; β_0 is the intercept of the probability of the DUI recidivism; and β_i = parameter estimate of the ith predictor variable.

The predicted probability p_i for ith record in the sample can be calculated as below:

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_{1j} + \cdots + \beta_i x_{ij}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1j} + \cdots + \beta_i x_{ij}\right)} \quad \text{(Equation 10)}$$

Log odds can be difficult to make sense of within a logistic regression data analysis. As a result, exponentiating the beta estimates is common to transform the results into an odds ratio (OR), easing the interpretation of results. The OR represents the odds that an outcome will occur given a particular event, compared to the odds of the outcome occurring in the absence of that event. If the OR is greater than 1, then the event is associated with a higher odd of generating a specific outcome. Conversely, if the OR is less than 1, then the event is associated with a lower odd of that outcome occurring.

$$Odds\ ratio = \frac{Probability\ of\ a\ event\ happen}{Probability\ of\ a\ event\ not\ happen} = \frac{p}{1-p} \quad \text{(Equation 11)}$$

Logistic regression assumes that there is no severe correlation/multicollinearity among the explanatory variables. Multicollinearity occurs when two or more continuous variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the model. In the case that two variables have high correlations, one variable should be removed from modeling. Variance inflation factor (VIF) or Pearson correlation coefficient can be used to test correlations between two continuous variables. Chi-square test can be used for categorical variables. Kruskal-Wallis H Test or t-test or ANOVA can be used to test correlations between a continuous and categorical variable.

5. RESULTS ANALYSIS

This section first presented predictions made by Decision Tree, Random Forest, and Gradient Boosting on the original imbalanced sample (subsection 5.1) and the oversampled sample (subsection 5.2). Subsection 5.3 provided a summary of model predictions. Subsection 5.4 presented results from statistical models to quantify associations and causation between predictors and the response variable and provided interpretation and understanding of the predictors.

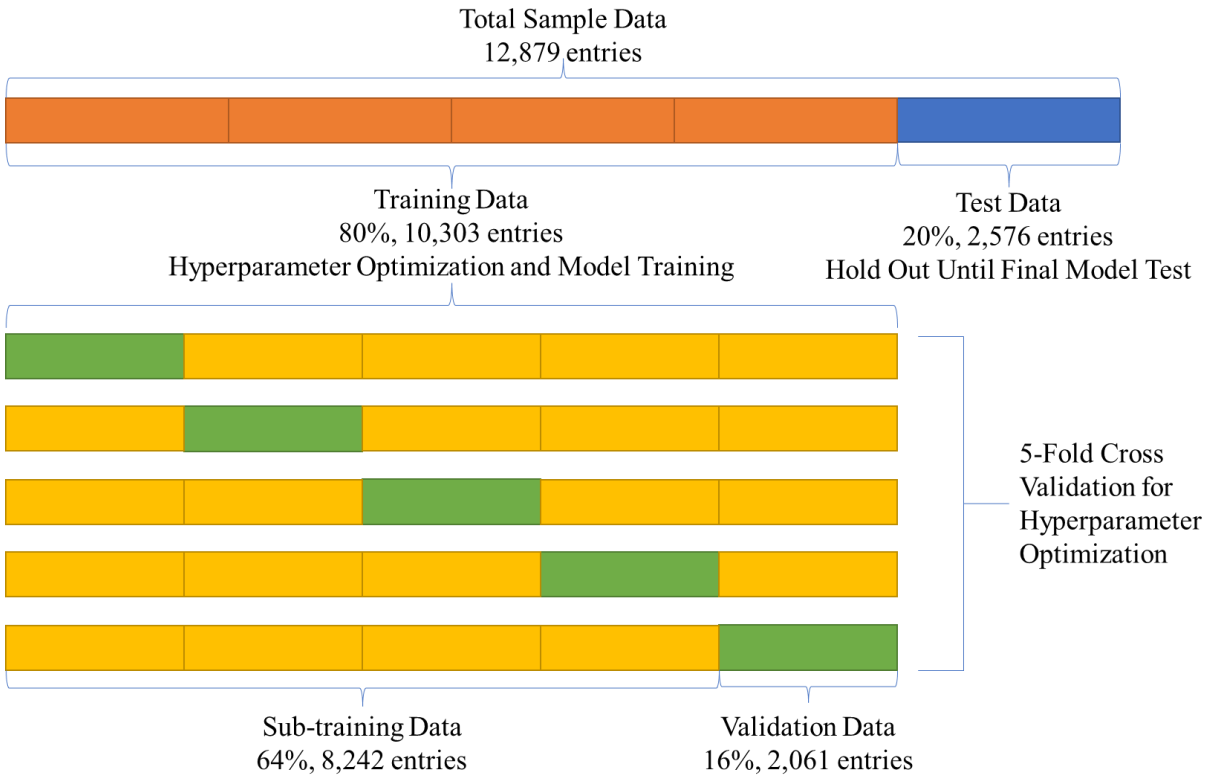

Figure 5.1. Sample Data Usage Allocation

5.1. Machine Learning Model Prediction

This subsection presented predictions made by Decision Tree, Random Forest, and Gradient Boosting on the original imbalanced sample. The sample data was fist split into two parts, training data and test data. A stratified split method was performed here so that the class

distributions were identical in training data and test data. Figure 5.1 shows the sample data usage allocation for these three predictions.

For each of these three methods, hyperparameter optimizations were performed first on the training data to find the hyperparameter(s) that yield the highest prediction power for the minority group, the F1 score. Models were trained to fix the prediction while performing the hyperparameter optimation with the random state numbers. Performance indices were reported to select the best hyperparameter(s) based on the F1 score. After choosing the best hyperparameter(s), the models were tested on the test data that was held out in the beginning for final evaluation. Predictions results made with optimized hyperparameter(s) were then reported for each method.

5.1.1. Decision Tree

In the Decision Tree model, hyperparameter optimization consisted of two parts: hyperparameter tuning and pruning. The analysis started with a hyperparameter tuning on the training data part to find the best number of maximum depth, resulting in the highest F1 score. The GridSearchCV function in the Scikit-learn package was used to conduct the hyperparameter optimization. A set of nine candidates 2,4,5,6,8,10,12,15, and 20, were input to hyperparameter tuning with training data during 5-fold cross-validation. 15 was selected for the maximum depth of the tree, and the associated F1 score mean over five folds was 25.6%, with a standard deviation of 1.66%. A complete result summary of all nine parameters of maximum depth for the Decision Tree can be found in Table 5.1.

Then, the cost complexity pruning (CCP) method was applied to prune nodes to prevent the overfitting of the tree further and reduce the overall misclassification error rate. The Minimal Cost-Complexity Pruning method in the Scikit-learn package was used to complete the pruning

process. Before applying the pruning method, a 5-fold cross-validation on the training data for maximum depth 15 without CCP was performed to collect baseline performance indices (Table 5.2). The performance indices of the baseline model on test data were 88.1% for accuracy, 21.0% for recall, 32.1% for precision, and 25.4% for F1 score.

The complexity hyperparameter α *(alpha)* was determined in three steps: (1) The F1 scores were plotted with different α values on both training data and test data, and estimating the α value that resulted in the highest F1 score on the test set from the figure (Figure 5.2). 0.0007 was estimated for α. The F1 score associated with this α value was approximately 30%. (2) A 5-fold cross-validation was performed on the training data with the selected α value 0.0007 in step (1) to test for sensitivity of F1 scores. The F1 scores ranged from roughly 26% to 32%, with a mean of 29.44% and a standard deviation of 2.1% (Table 5.3). With pruning, the precision and F1 scores were increased. On average, the F1 score increased by roughly 4 percent, and accuracy improved by approximately 25 percent. (3) On the training data, plot the mean of F1 scores with different α values using the 5-fold cross-validation, then a searching method was used to find the optimal value of α with the highest mean F1 value. In this case, the optimal value was in the interval between 0.0005 to 0.0015 (Figure 5.3), so a set of means of F1 test scores with α values was output to search for the optimal value. There were 52 data points in this interval. After sorting F1 mean values, 0.000729 was finally selected for the α value with a slightly higher F1 mean of 29.6% and a standard deviation of 2.1%. Table 5.4 shows the performance indices from the validation set during hyperparameter optimization.

Table 5.1. Hyperparameter Tuning Results for Decision Tree on Original Dataset.

Rank	Hyperparameter	Validation F1 Scores					Mean	SD
	Max Depth	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**15**	**23.19%**	**24.30%**	**27.46%**	**27.22%**	**26.01%**	**25.64%**	**1.66%**
2	12	24.10%	24.24%	26.14%	22.22%	24.92%	24.32%	1.27%
3	20	21.64%	21.86%	25.27%	25.46%	23.40%	23.53%	1.62%
4	10	24.03%	24.64%	23.74%	22.79%	20.61%	23.16%	1.41%
5	8	20.69%	26.72%	21.37%	21.34%	20.41%	22.11%	2.34%
6	5	16.95%	19.66%	21.01%	23.08%	18.26%	19.79%	2.13%
7	6	16.46%	18.88%	18.93%	18.26%	16.59%	17.83%	1.09%
8	4	10.05%	14.61%	20.34%	20.87%	16.59%	16.49%	3.98%
9	2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5.2. Baseline Performance Indices for Decision Tree with Maximum Depth 15 Without CCP from 5-Fold Cross-Validation

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	87.14%	20.41%	26.85%	23.19%
2	88.21%	19.90%	31.20%	24.30%
3	88.21%	23.47%	33.09%	27.46%
4	88.06%	23.47%	32.39%	27.22%
5	87.57%	22.96%	30.00%	26.01%

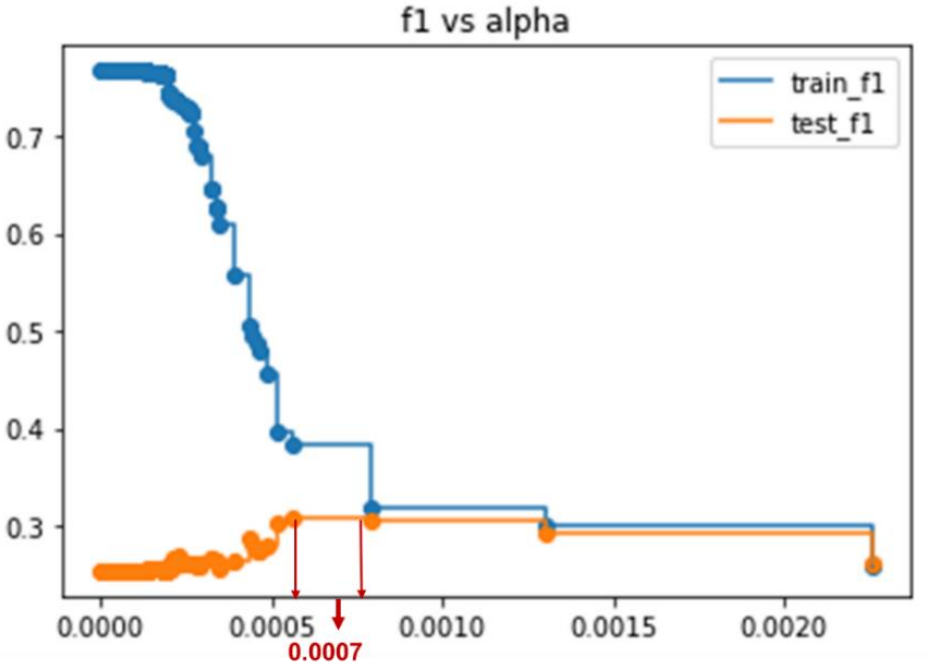

Figure 5.2. F1 Scores and Complexity hyperparameter α (Alpha) for Decision Tree on Single Test/Train Split Data Sets

Table 5.3. Performance Indices for Decision Tree Maximum Depth 15 with α of 0.0007 from 5-Fold Cross-Validation

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	90.88%	17.35%	56.67%	26.56%
2	91.21%	18.37%	63.16%	28.46%
3	91.02%	19.90%	58.21%	29.66%
4	90.25%	22.45%	47.31%	30.45%
5	90.54%	23.47%	50.55%	32.06%

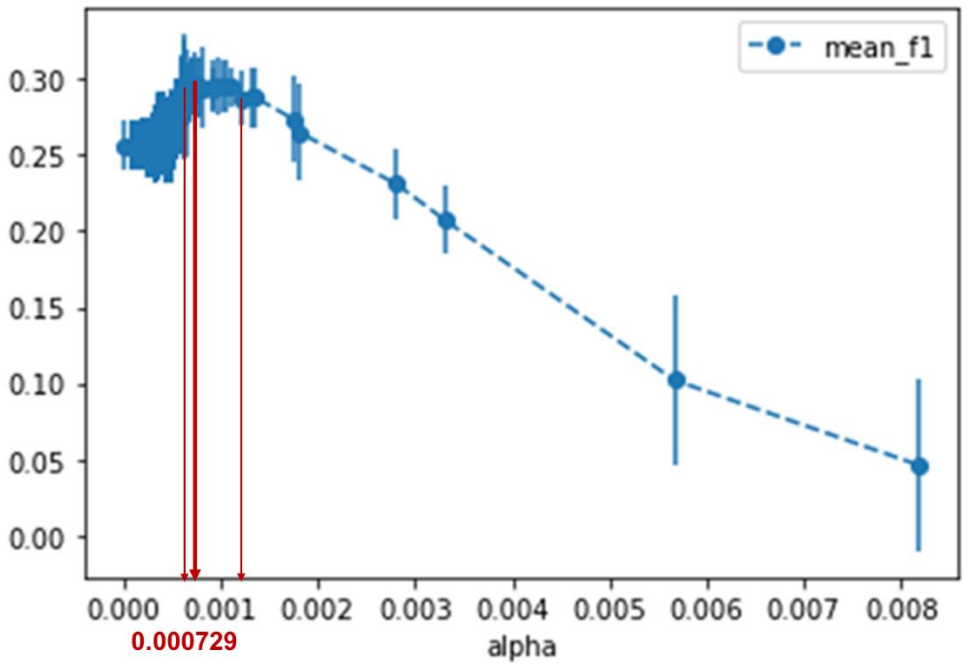

Figure 5.3. Mean F1 Score and Complexity hyperparameter α (Alpha) for Decision Tree from the 5-Fold Cross-Validation

Table 5.4. Final Model Performance Indices for Decision Tree with Maximum Depth 15 with α of 0.000729 from 5-Fold Cross-Validation on Train Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	90.93%	17.35%	57.63%	26.67%
2	91.17%	17.86%	62.50%	27.78%
3	91.21%	20.41%	61.54%	30.65%
4	90.39%	22.45%	48.89%	30.77%
5	90.64%	23.47%	51.69%	32.28%
Mean	90.87%	20.31%	56.45%	29.63%

After hyperparameter tuning and pruning, hyperparameter optimization for the Decision Tree was finished. The confusion matrix for test data was output in Figure 5.4. The performance indices on the test data were 91.5% for accuracy, 19.4% for recall, 72.7% for precision, and 30.6% for F1 score. Hyperparameter tuning and pruning did improve model performance by enhancing the F1 score, mainly on the precision score. On the other hand, the recall scores were relatively low across the optimization process, so this model was weak in predicting/identifying repeat offenders. Two hundred repeat offenders were not correctly identified (Figure 5.4).

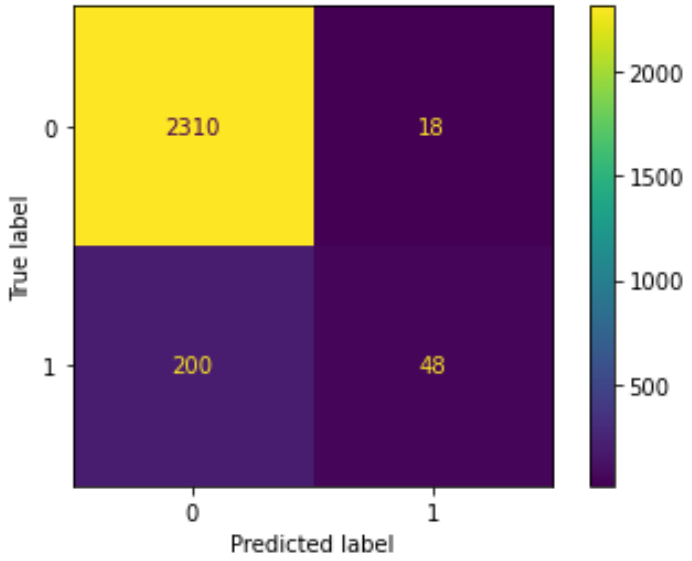

Figure 5.4. Confusion Matrix for Decision Tree on Test Data

Figure 5.5 shows the top 20 ranks and the magnitudes of the feature/variable impacts on the prediction. The variable impact magnitude was measured by the average of the absolute Shapley Additive exPlanations (SHAP) value of each variable calculated from 2,557 forecasts on the test data. SHAP values are applications of game theory and are unique, consistent, and locally accurate attribution values (Lundberg, 2019). A positive SHAP value of a variable indicates a positive impact on the prediction, and a negative SHAP value of a variable indicates a negative effect on the prediction.

The highest BAC value on record had the most significant impact on the predictions, with an impact magnitude of approximately 0.2. The mean BAC value ranked second with an impact magnitude of roughly 0.16. The binary variable of driver's license status “licensed driver and another status” and the median BAC value ranked third and fourth, with a similar impact magnitude of approximately 0.125. The binary variable of license status, “License suspended,” ranked fifth with an impact magnitude of 0.025, followed by age at the first offense with an impact magnitude of 0.02. The impact magnitude for the rest of the variables was less than 0.02. The variable impacts to predict non-repeat offenders and repeat offenders were identical. The rest variables had minor impacts on this prediction.

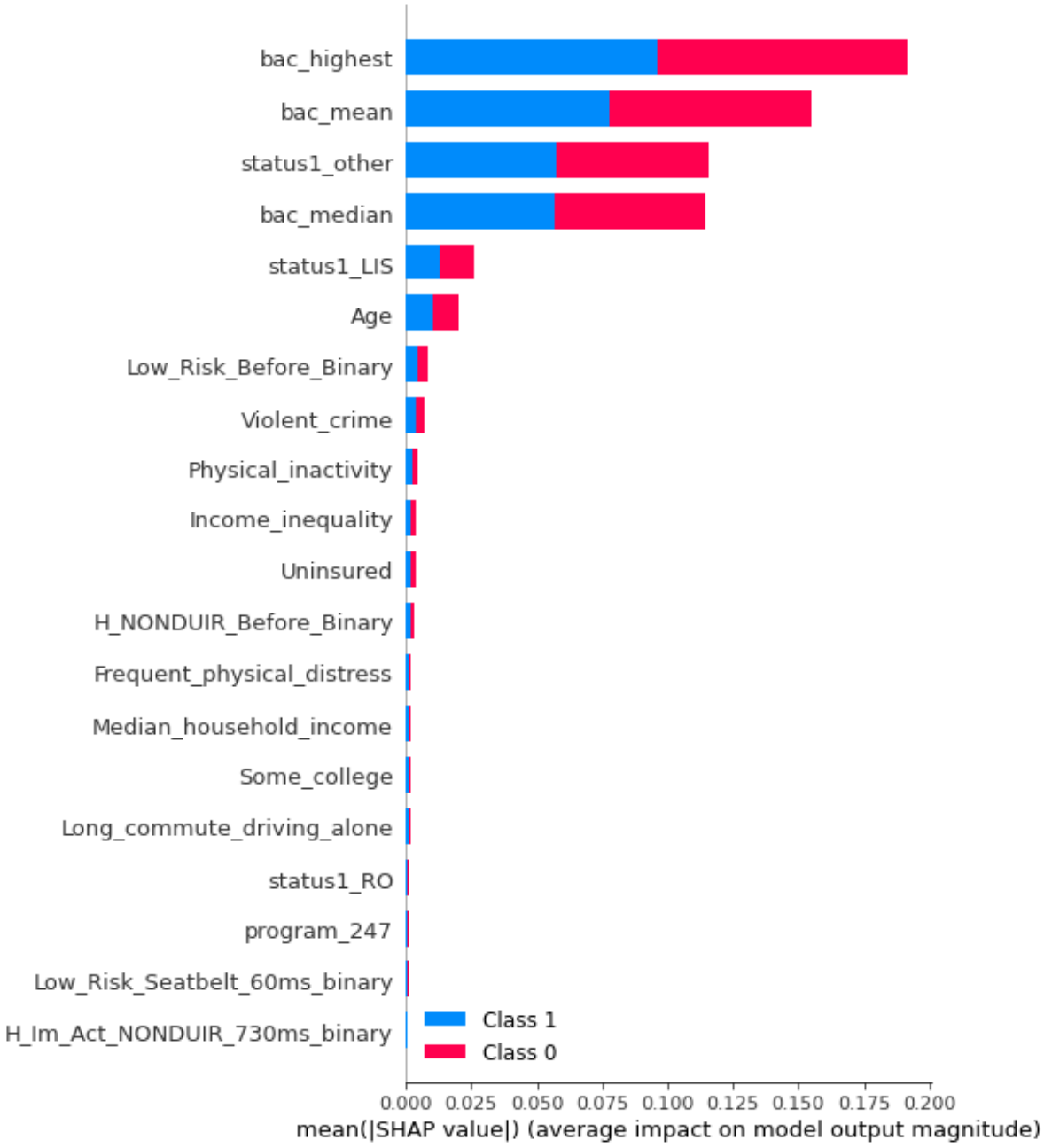

Figure 5.5. Features Importance Plot for Decision Tree Prediction on Test Data

Figure 5.6 shows the top 20 rank and directionality of the feature/variable impact on the predictions. In this figure, the x-axis stands for SHAP value, and the y-axis has all the features/variables. Each point in the figure is one SHAP value for a prediction and variable. Red color means a higher value of a variable. Blue indicates a lower value of a feature. The distribution of the red and blue dots provided the variable's directionality impact on the overall prediction. A positive SHAP value positively impacts prediction, leading the model to predict the repeat offender. A negative SHAP value means a negative impact, leading the model to predict the non-repeat offender. The directionality of the feature/variable impact should be considered cautiously when the prediction power of a model is low.

Based on Figure 5.6, the likelihood of being a repeat offender decreased when the highest BAC value on a driver's record increased. This insight contradicted findings from previous studies (e.g., Marowitz, 1998; C'de Baca et al., 2001; Roma et al., 2019). Also, note that there was a wide purple area between SHAP values 0 and 0.05. The highest BAC value was a numerical variable, and the purple areas could occur because (1) a lot of blues (low-value points) and red dots (high-value points) overlapped in this interval, and (2) the distribution of middle-value points. This distribution further implied that the directionality of the impact might be complicated. Given the recall value on the test data was low in this prediction, insights from this figure can be disregarded.

Figure 5.7 provides a visualization of the Decision Tree prediction process. This visualization helps understand the set of rules that cause the prediction. However, this visualization plot should be used cautiously when the prediction power is low. Given the recall value on the test data was low in this prediction, insights from Figure 5.7 were disregarded.

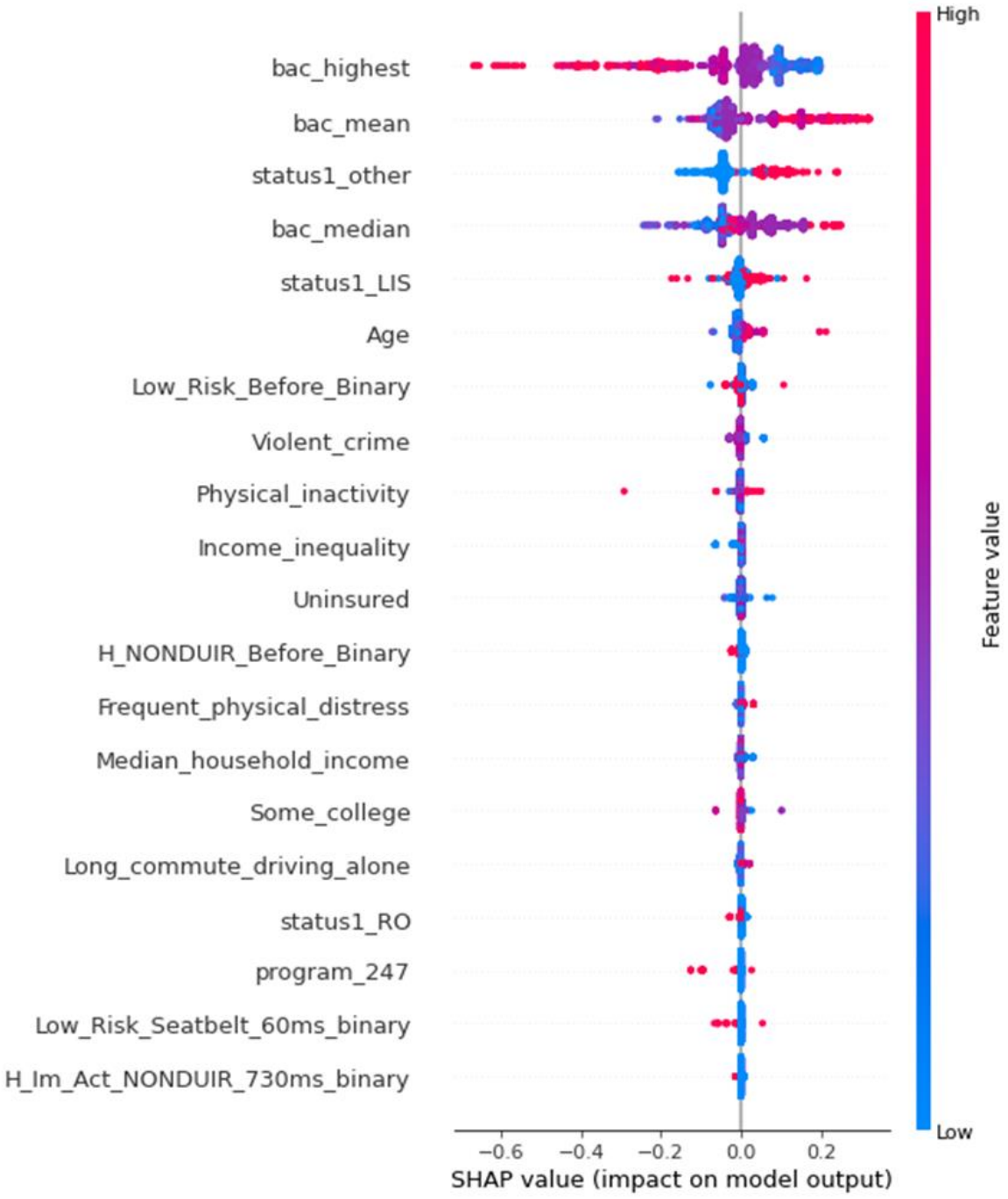

Figure 5.6. Feature Impact Directionality on Decision Tree Prediction on Test Data

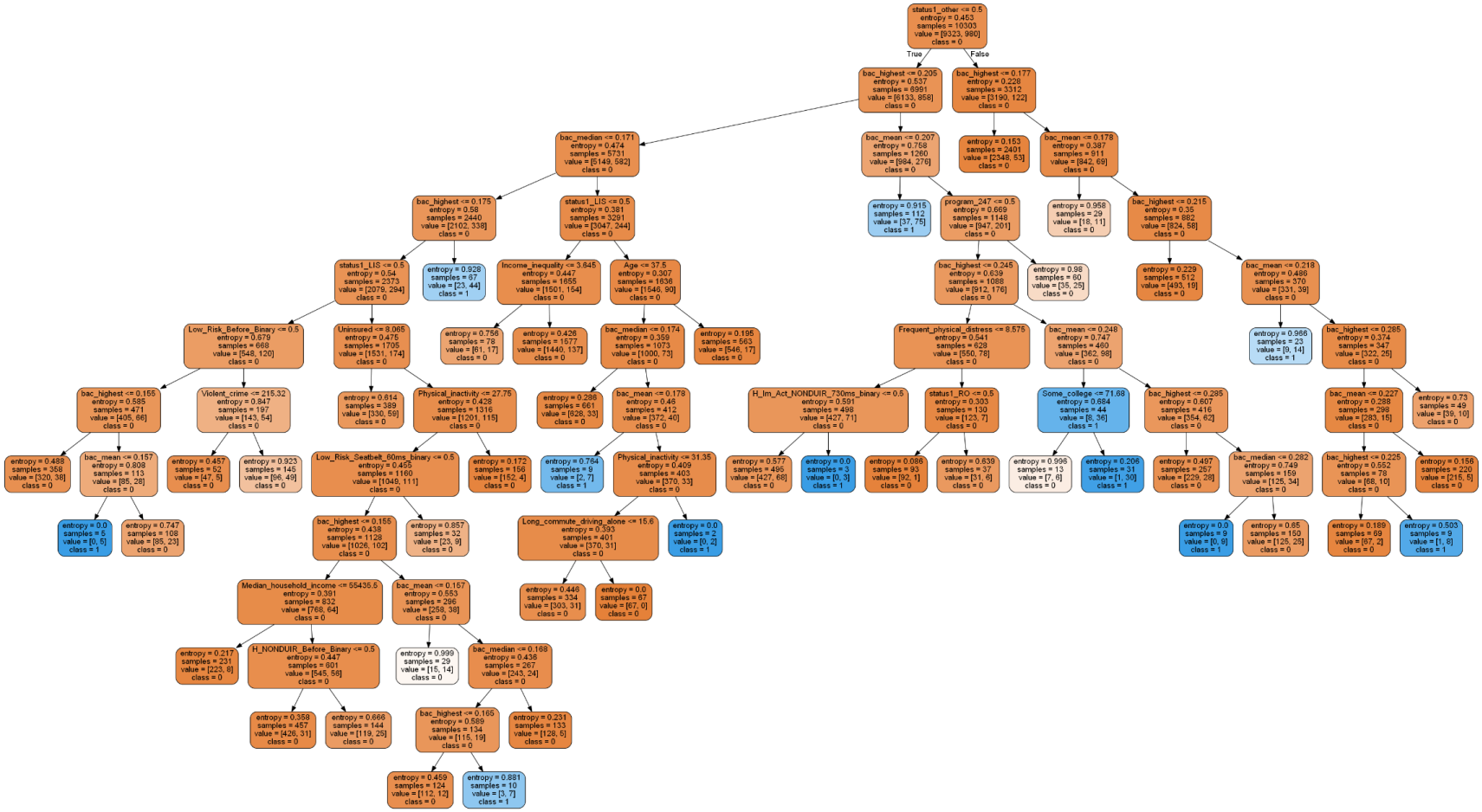

Figure 5.7. Decision Tree Visualization

5.1.2. Random Forest

In the Random Forest model, hyperparameter optimization was performed on hyperparameter tuning only. Pruning is unnecessary for optimizing ensemble models like Random Forest and Gradient Boosting since their algorithms have already been featured to reduce overfitting (Efron, 1993; Breiman, 1996; Breiman, 2001; Boehmke, 2020). Random Forest uses bootstrap aggregation (or sampling with replacement) along with a random selection of features for a split. The correlation between the trees (or weak learners) would be low. That means individual trees would have high variance, and the ensemble output will be appropriate (lower variance and lower bias) because the trees are not correlated. Gradient Boosting reduces overfitting by adding a feature at each iteration representing the prediction error (Breiman, 2001).

Tuning was performed on three hyperparameters – the number of estimators, maximum depth, and random state. The number of estimators represents the number of trees in the forest, and the random state fixes the randomness of the prediction so that the forecast can be replicated in the future. For hyperparameters, there were four candidates of a number of estimators/trees [20, 50, 100, 200], eight candidates of maximum depth [4, 5, 6, 8, 10, 12, 15, 20] and three candidates of random states [13, 16, 20]. In total, 96 combinations of hyperparameters were tested with training data with 5-fold cross-validation. The top 30 combinations with the highest F1 mean scores were reported in Table 5.5.

Hyperparameter combination with a maximum depth of 18, 20 trees, and a random state of 20 achieved the best performance in terms of F1 score. However, the associated F1 score was 6.84%, which was considered poor performance. The 5-fold cross-validation results associated with this hyperparameter combination on training data were reported in Table 5.6. The recall

scores were extremely low across all five folds, indicating a low prediction power in identifying repeat offenders.

Figure 5.8 presents the confusion matrix for test data. The performance indices on the test data were 90.4% for accuracy, 2 % for recall, 50% for precision, and 3.9% for the F1 score. Recall and F1 scores were too low to produce a good prediction. Figure 9 and Figure 10 showed the rank, magnitude, and directionality of the feature impact on the predictions for test data. However, due to the model's poor performance, no further insights were obtained from Figures 5.9 and 5.10.

Table 5.5. Top 30 of Hyperparameter Tuning Results for Random Forest on Original Dataset.

Rank	Parameters			Validation F1 Scores					Mean	SD
	Max Depth	Estimators	Random State	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**18**	**20**	**20**	**4.81%**	**8.33%**	**9.57%**	**5.03%**	**6.45%**	**6.84%**	**1.86%**
2	20	20	13	4.76%	4.74%	10.33%	6.83%	7.24%	6.78%	2.05%
3	18	20	16	6.64%	6.42%	6.83%	5.97%	7.34%	6.64%	0.45%
4	20	20	20	4.72%	9.26%	5.77%	6.73%	5.53%	6.40%	1.57%
5	18	20	13	6.60%	5.74%	5.80%	6.00%	3.62%	5.55%	1.01%
6	20	50	16	4.74%	3.86%	6.86%	4.98%	7.27%	5.54%	1.31%
7	20	20	16	3.79%	8.33%	5.80%	5.83%	3.70%	5.49%	1.70%
8	20	50	20	2.90%	7.51%	5.77%	5.00%	5.61%	5.36%	1.49%
9	18	75	20	3.88%	5.83%	5.88%	6.03%	4.69%	5.26%	0.84%
10	18	50	20	4.74%	6.76%	6.80%	5.94%	1.90%	5.23%	1.83%
11	20	75	16	4.74%	5.69%	4.95%	5.03%	5.50%	5.18%	0.36%
12	20	100	20	2.93%	5.71%	5.88%	5.03%	5.56%	5.02%	1.09%
13	20	180	16	3.86%	2.94%	4.00%	5.88%	7.34%	4.81%	1.59%
14	20	50	13	1.91%	6.73%	4.93%	4.95%	5.50%	4.81%	1.59%
15	18	100	20	3.86%	3.90%	5.91%	5.05%	4.65%	4.68%	0.76%
16	18	50	13	2.94%	7.66%	5.88%	3.08%	3.74%	4.66%	1.83%
17	20	120	20	2.93%	5.71%	4.95%	5.00%	3.76%	4.47%	1.00%
18	20	100	16	2.90%	4.81%	4.02%	5.03%	5.50%	4.45%	0.91%
19	20	75	20	1.96%	5.71%	3.96%	5.00%	5.56%	4.44%	1.38%
20	18	120	20	4.81%	3.90%	4.95%	4.06%	3.74%	4.29%	0.49%
21	20	150	16	4.85%	2.93%	4.00%	4.95%	4.65%	4.28%	0.75%
22	18	50	16	3.85%	4.81%	3.94%	4.00%	4.69%	4.26%	0.41%
23	20	120	16	3.86%	4.83%	3.02%	4.02%	5.53%	4.25%	0.86%
24	20	200	16	3.86%	1.97%	3.00%	5.91%	6.45%	4.24%	1.70%
25	20	150	20	2.96%	4.81%	3.98%	4.06%	4.67%	4.10%	0.66%
26	15	20	13	2.90%	0.98%	6.86%	4.12%	5.58%	4.09%	2.05%
27	18	75	13	2.93%	2.94%	4.00%	5.13%	4.69%	3.94%	0.90%
28	20	200	20	2.96%	4.81%	3.98%	4.06%	3.76%	3.91%	0.59%
29	18	180	13	3.90%	1.98%	3.00%	5.03%	5.61%	3.90%	1.32%
30	15	20	20	0.99%	1.93%	4.88%	6.90%	4.67%	3.87%	2.14%

Table 5.6. Performance Indices for Random Forest from 5-Fold Cross-Validation on Train Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	90.39%	2.50%	62.50%	4.81%
2	90.39%	4.57%	47.37%	8.33%
3	90.83%	5.15%	66.67%	9.57%
4	90.83%	2.66%	45.45%	5.03%
5	90.15%	3.48%	43.75%	6.45%
Mean	90.52%	3.67%	53.15%	6.84%

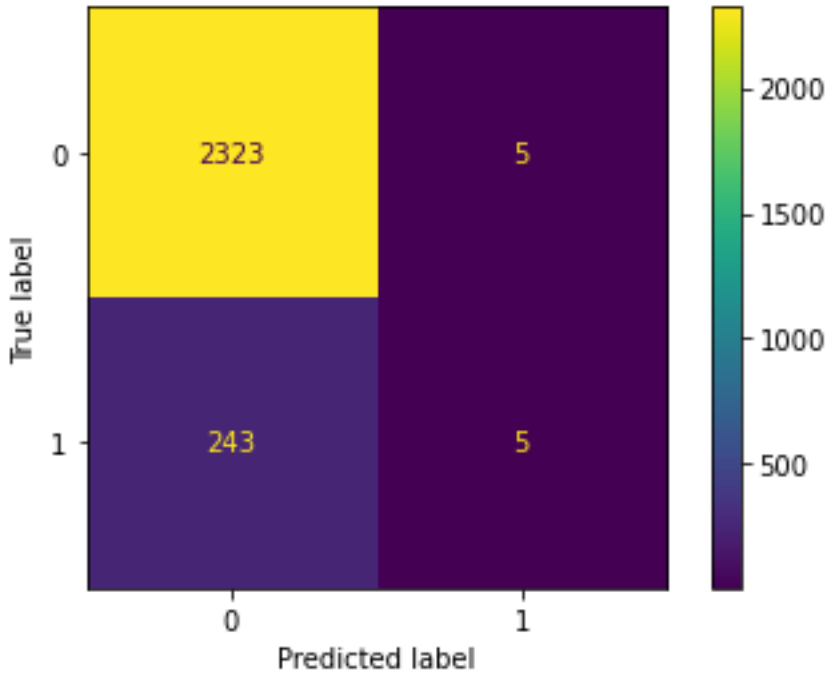

Figure 5.8. Confusion Matrix for Random Forest on Test Data

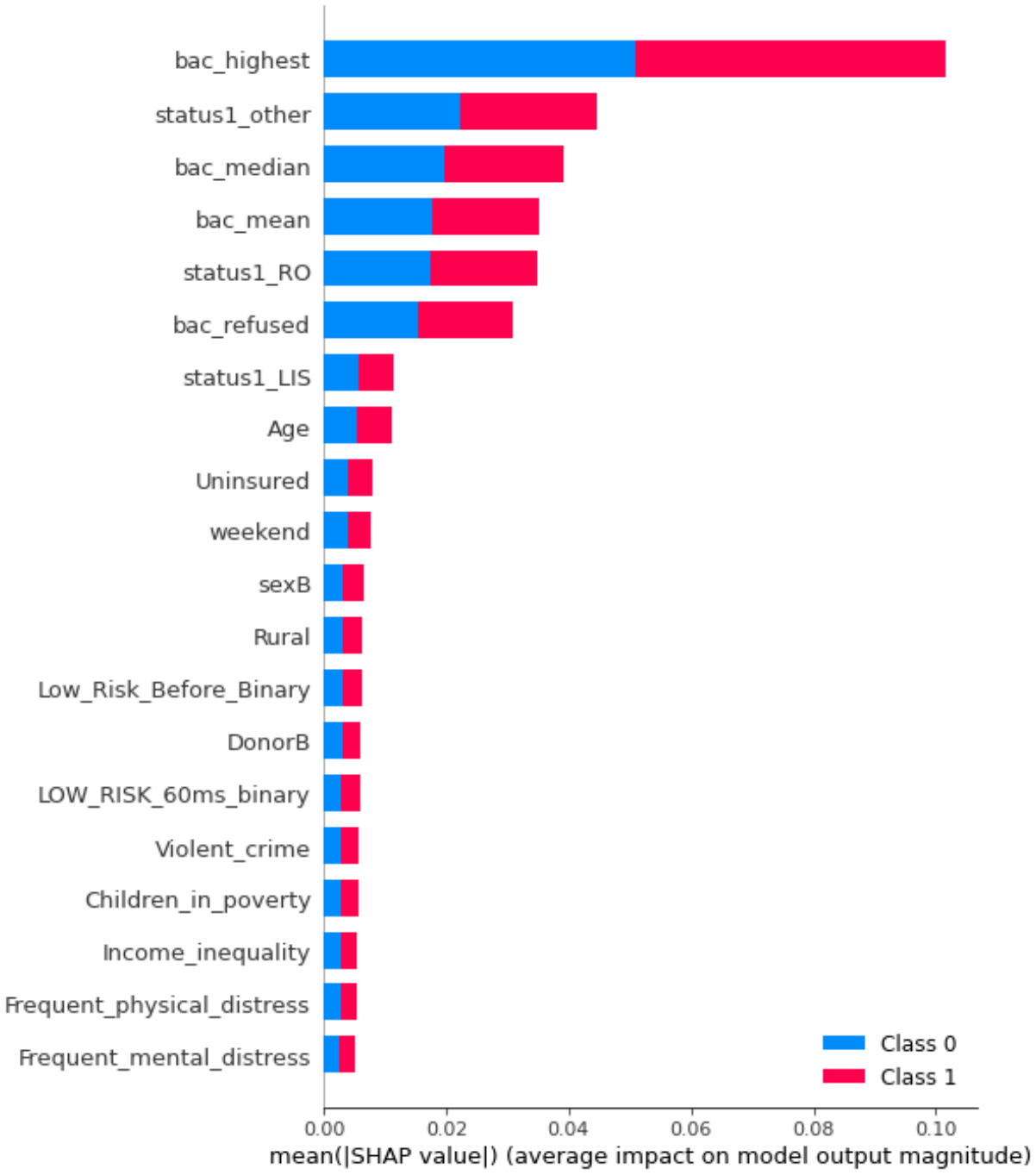

Figure 5.9. Features Importance Plot for Random Forest Prediction on Test Data

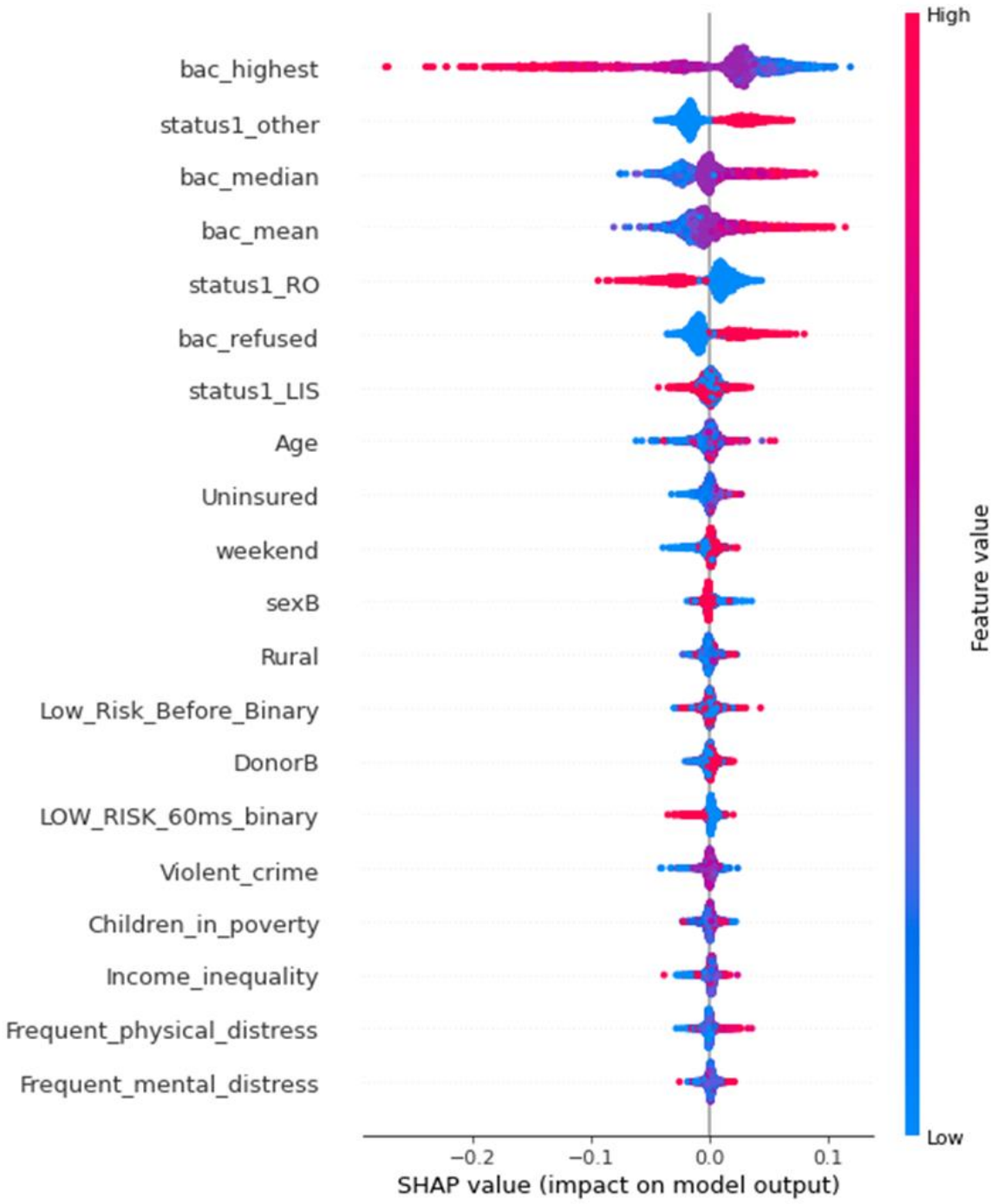

Figure 5.10. Feature Impacts on Random Forest Prediction on Test Data

5.1.3. Gradient Boosting

For the Gradient Boosting model, hyperparameter optimization was completed with hyperparameter tuning only. The tuning process was performed on four hyperparameters – learning rate, the number of estimators, maximum depth, and random state. The learning rate controls the rate at which a tree learns from the data and the magnitude of the modification to the overall model, so it weighted the effect each tree has on the final prediction and improves the prediction power in the long run (Friedman, 2001; Friedman, 2002).

For hyperparameters, there were seven candidates of learning rate [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2], seven candidates of number of estimators/trees [10, 20, 50, 100, 120, 150, 200] seven candidates of maximum depth [4, 5, 8, 10, 12, 15, 20] and two candidates of random states [0, 13]. In total, 686 combinations of hyperparameters were tested with training data with 5-fold cross-validation. The total calculation time was approximately 2 hours. The top 30 combinations with the highest F1 mean scores were reported in Table 5.7.

Hyperparameter combination with a learning rate of 0.2, maximum depth of 5, 150 trees, and random state of 13 achieved the best performance in terms of F1 score. The associated mean F1 score with this combination was 32.2% on the training data and was relatively low in model prediction power. The 5-fold cross-validation results associated with this hyperparameter combination on training data were reported in Table 5.8. Compared to the Decision Tree model prediction, the mean recall scores for Gradient Boosting prediction were slightly improved from 20.31% to 22.98%, while the mean precision scores were slightly decreased from 56.45% to a mean of 54.57%. The prediction power of identifying repeat offenders, F1 score, was also improved slightly from 29.63% to 32.21%.

Table 5.7. Top 30 of Hyperparameter Tuning Results for Gradient Boosting on Original Dataset.

Rank	Parameters				Validation F1 Scores					Mean	SD
	Learning Rate	Max Depth	Estimators	Random State	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**0.2**	**5**	**150**	**13**	**34.3%**	**26.9%**	**32.9%**	**33.9%**	**33.1%**	**32.2%**	**2.72%**
2	0.15	5	200	13	35.3%	27.3%	28.3%	33.8%	35.9%	32.1%	3.61%
3	0.2	5	200	13	37.7%	25.6%	30.7%	32.1%	34.0%	32.0%	3.98%
4	0.15	5	150	13	33.7%	26.8%	29.9%	33.7%	35.5%	31.9%	3.13%
5	0.2	4	200	0	33.9%	28.5%	28.2%	37.8%	31.0%	31.9%	3.62%
6	0.2	5	200	0	37.9%	27.3%	29.0%	33.7%	31.6%	31.9%	3.70%
7	0.2	5	120	13	33.5%	27.4%	31.0%	32.4%	33.7%	31.6%	2.30%
8	0.2	5	150	0	34.7%	26.6%	29.6%	34.0%	32.4%	31.5%	3.01%
9	0.2	4	150	0	32.7%	29.3%	28.7%	35.8%	30.6%	31.4%	2.58%
10	0.15	5	200	0	36.2%	28.9%	29.6%	33.0%	28.8%	31.3%	2.90%
11	0.2	4	120	0	34.3%	28.8%	27.0%	36.0%	30.4%	31.3%	3.36%
12	0.15	8	150	0	37.5%	24.1%	28.4%	33.7%	31.8%	31.1%	4.58%
13	0.15	8	120	0	37.2%	24.8%	29.1%	31.5%	32.0%	30.9%	4.04%
14	0.15	4	200	0	35.4%	28.4%	29.5%	31.2%	30.1%	30.9%	2.43%
15	0.2	4	200	13	33.7%	27.9%	28.3%	32.6%	30.7%	30.6%	2.29%
16	0.2	5	100	13	32.6%	27.6%	30.0%	31.6%	31.4%	30.6%	1.73%
17	0.2	5	120	0	32.2%	27.0%	29.6%	32.5%	31.4%	30.5%	2.04%
18	0.1	5	200	0	31.7%	26.8%	28.4%	34.4%	31.0%	30.5%	2.65%
19	0.1	5	200	13	32.0%	28.8%	28.2%	32.0%	31.1%	30.4%	1.60%
20	0.15	4	200	13	32.6%	26.7%	28.3%	36.4%	27.8%	30.3%	3.62%
21	0.15	8	200	0	35.8%	23.4%	28.0%	33.7%	30.7%	30.3%	4.36%
22	0.2	8	200	13	35.8%	22.7%	31.0%	31.9%	29.9%	30.3%	4.26%
23	0.1	4	200	13	30.8%	29.8%	26.7%	35.3%	28.2%	30.2%	2.93%
24	0.2	5	100	0	32.6%	26.0%	29.3%	32.7%	29.5%	30.0%	2.48%
25	0.2	4	100	0	33.7%	26.2%	27.4%	33.5%	29.1%	30.0%	3.09%
26	0.15	5	120	13	30.8%	25.4%	28.7%	31.4%	33.3%	29.9%	2.71%
27	0.15	5	150	0	33.2%	27.2%	30.5%	33.0%	25.6%	29.9%	3.04%
28	0.1	4	200	0	31.4%	30.0%	27.4%	32.8%	27.7%	29.9%	2.11%
29	0.15	8	100	13	36.1%	24.1%	28.3%	29.7%	31.1%	29.9%	3.90%
30	0.15	5	120	0	32.5%	27.3%	29.1%	32.5%	27.6%	29.8%	2.26%

Table 5.8. Model Performance Indices for Gradient Boosting from 5-Fold Cross-Validation on Train Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	91.27%	23.50%	63.51%	34.31%
2	90.49%	18.27%	50.70%	26.87%
3	90.88%	23.71%	53.49%	32.86%
4	90.73%	26.06%	48.51%	33.91%
5	90.78%	23.38%	56.63%	33.10%
Mean	90.83%	22.98%	54.57%	32.21%

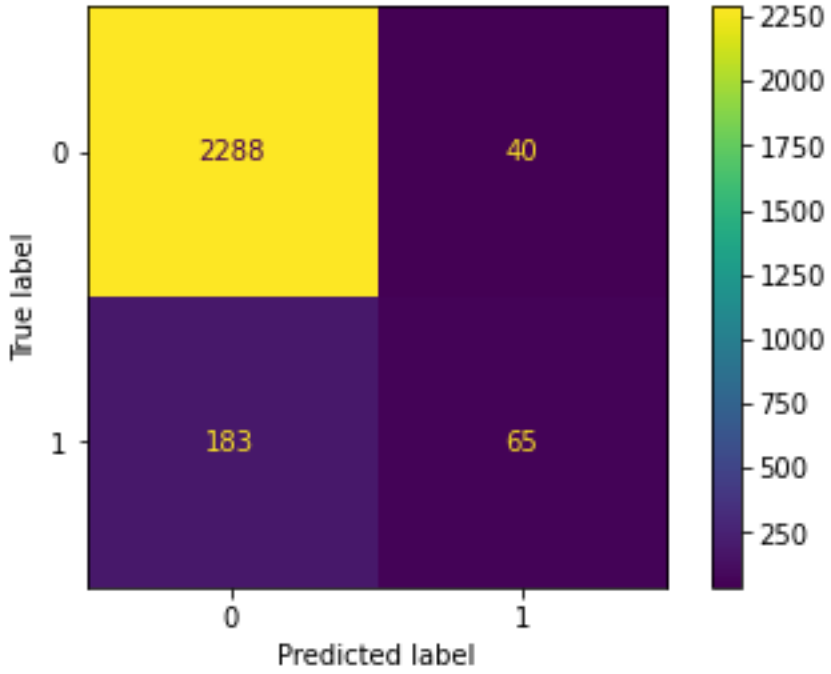

Figure 5.11. Confusion Matrix for Gradient Boosting on Test Data

Figure 5.11 presents the confusion matrix for test data. The performance indices on the test data were 91.3% for accuracy, 26.2 % for recall, 61.9% for precision, and 36.8% for F1 score. Recall and F1 scores were low to produce a good prediction. Figure 11 and Figure 12 showed the rank, magnitude, and directionality of the feature impact on the predictions for test

data. However, due to the model's poor performance, insights obtained from Figure 11 and Figure 12 should be used cautiously.

As shown in Figure 5.12, the top five predictors were the highest BAC value, the mean BAC value, the binary variable of driver license status “other,” the median BAC value, and age at first offense. In Figure 5.13, large purple areas with a SHAP value interval length of 1 were evident on the highest BAC value, the mean BAC value, and the median BAC value near the origin point. The purple areas extended to both sides of these three variables, indicating that the directionality of impacts of these three variables on the predictions can be complicated. Given the prediction power of the F1 score was low in this model, further analysis is needed to understand how the BAC values impact the likelihood of being a DUI repeat offender.

In contrast, the impact directionality of the binary variable of driver's license status “other” was clear. A higher value when driver's license status was “other” leads to a lower chance of recidivism, and a lower value when driver's license status was either “suspended” or “revoked” leads to a higher chance of recidivism. Given drivers with a driver's license status of “suspended” or “revoked” have a higher chance of breaking the rules, it is reasonable to trust this insight.

The directionality of age was also unclear. A large number of blue dots, the younger aged drivers, dominated the origin area, slightly towards the positive direction. The SHAP value interval was short on both side, meaning the variable impacts for both the positive and negative directions was small. Further analysis is needed to understand this variable.

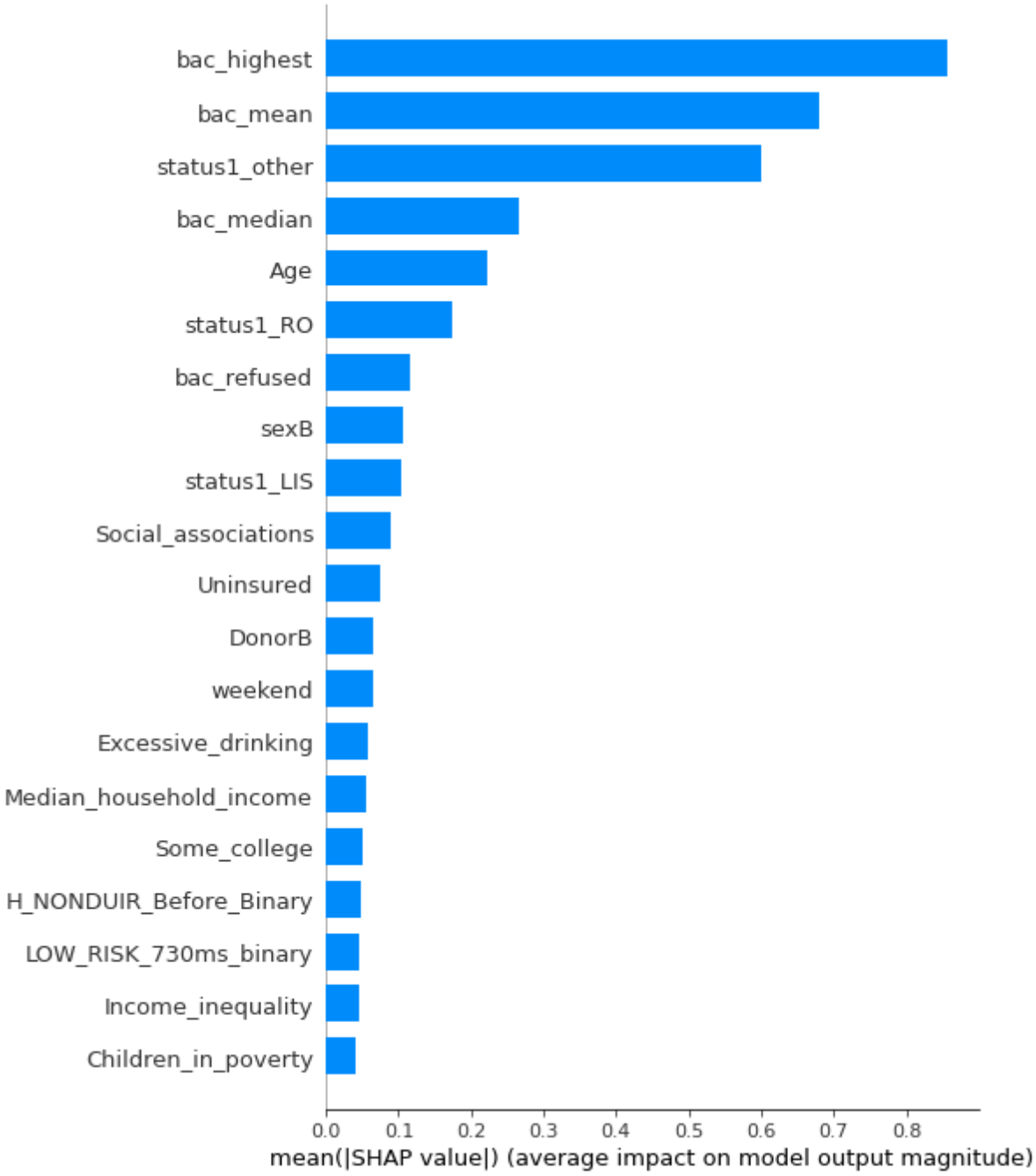

Figure 5.12. Features Importance Plot for Gradient Boosting Prediction on Test Data

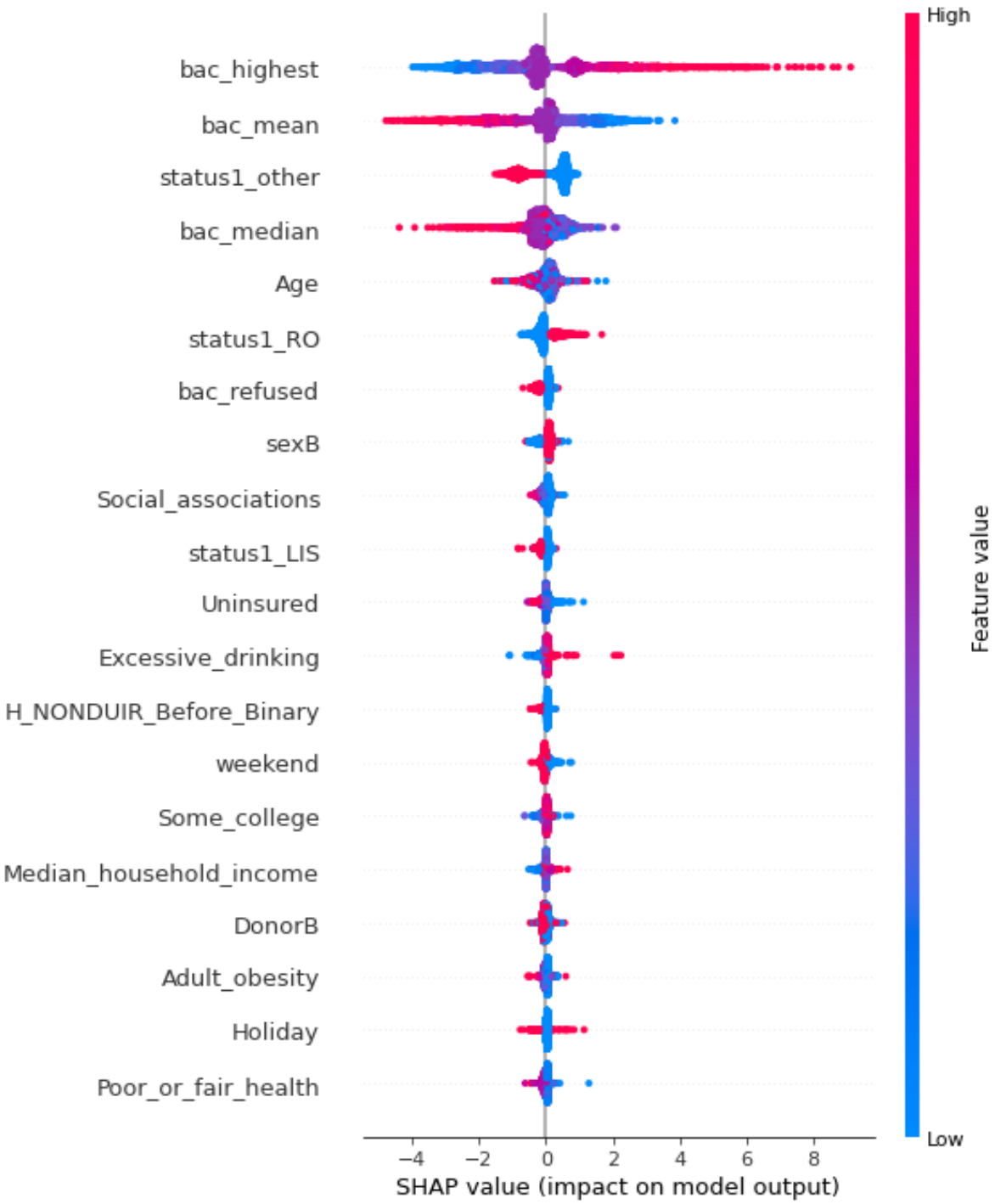

Figure 5.13. Feature Impacts on Gradient Boosting Prediction on Test Data

5.2. Model Prediction with Oversample Technique

An oversampling technique was applied to balance two classes to address issues related to data imbalance. Specifically, this technique was applied twice in the sample to improve prediction. First, during hyperparameter optimization through 5-fold cross-validation, the oversampling technique was only applied to the sub-training data. Validation data remained original throughout the optimization process (Figures 5.14 and 5.15). Secondly, after hyperparameter optimization, the tuning process implemented oversampling on the entire training data with the hyperparameter selected. The model was trained to learn the prediction rule on the oversampled data. Test data was kept original to provide the final model evaluation (Figure 5.16).

Figure 5.14 shows the overview of the oversampling implemented during hyperparameter optimization through 5-fold cross-validation. Figure 5.15 shows the details of the oversampling process in hyperparameter optimization. Fold 2 in iteration 1 contains 2,060 entries, with 1864 NROs (90.5%) and 196 ROs (9.5%). After oversampling, the Fold 2 of iteration 1 contained the balanced class structure with 1,864 NROs (50%) and 1,864 ROs (50%). Note the total sample size in each fold was increased due to oversampling. Figure 5.16 shows the oversampling implementation on the entire training data to train the model with balanced data for prediction, and then the prediction was tested on the original test data.

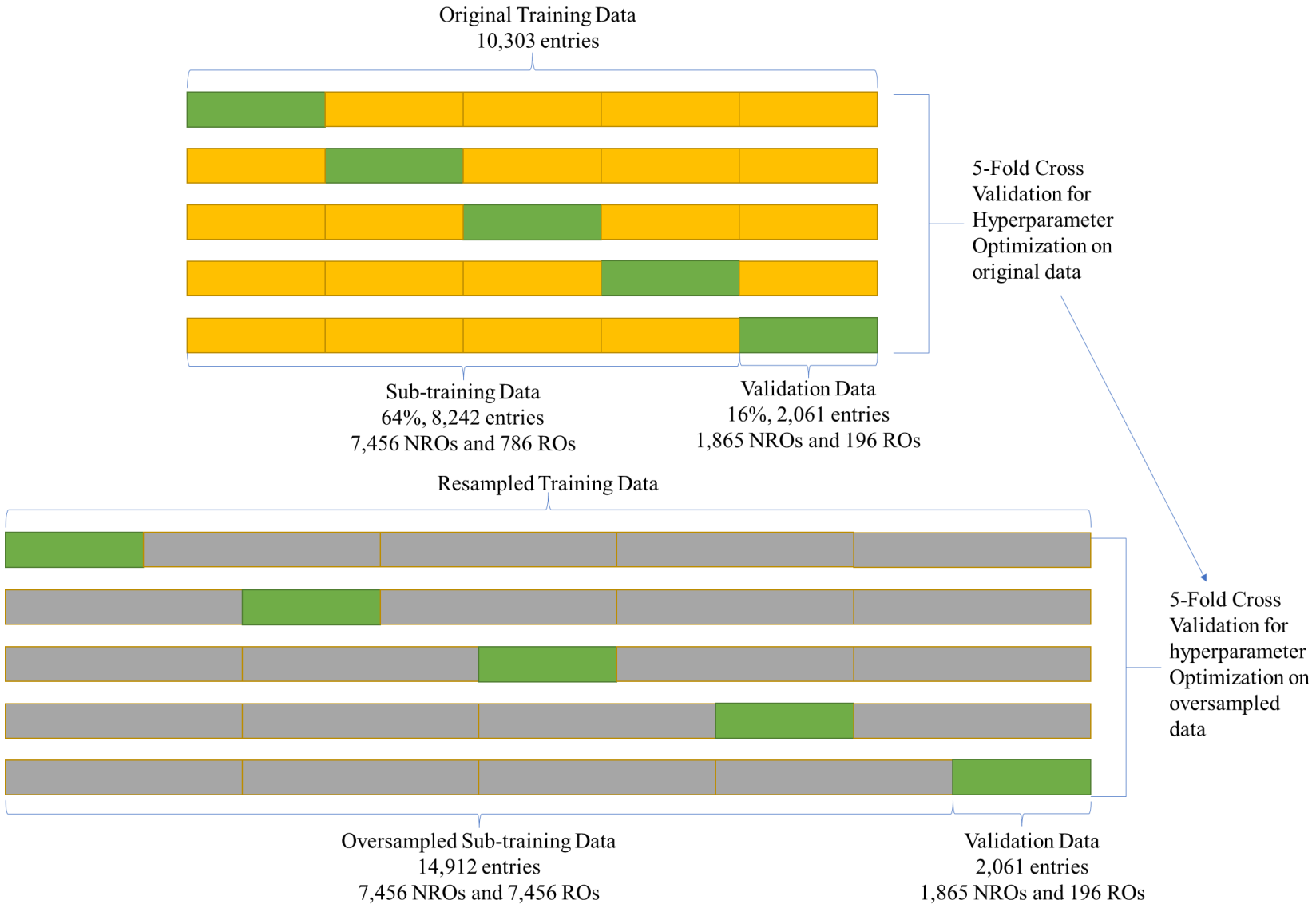

Figure 5.14. Oversampling Implementation in Hyperparameter Optimization

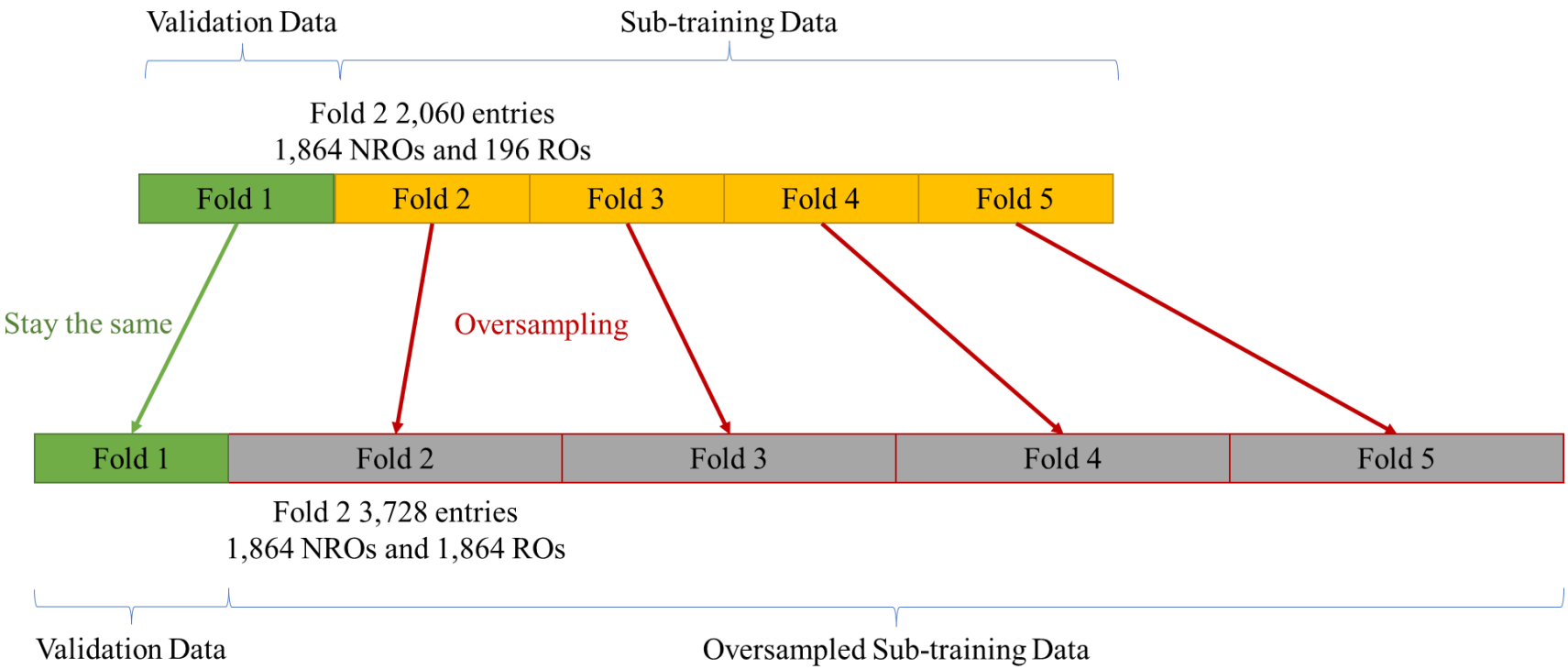

Figure 5.15. The Details of Oversampling Process in Hyperparameter Optimization

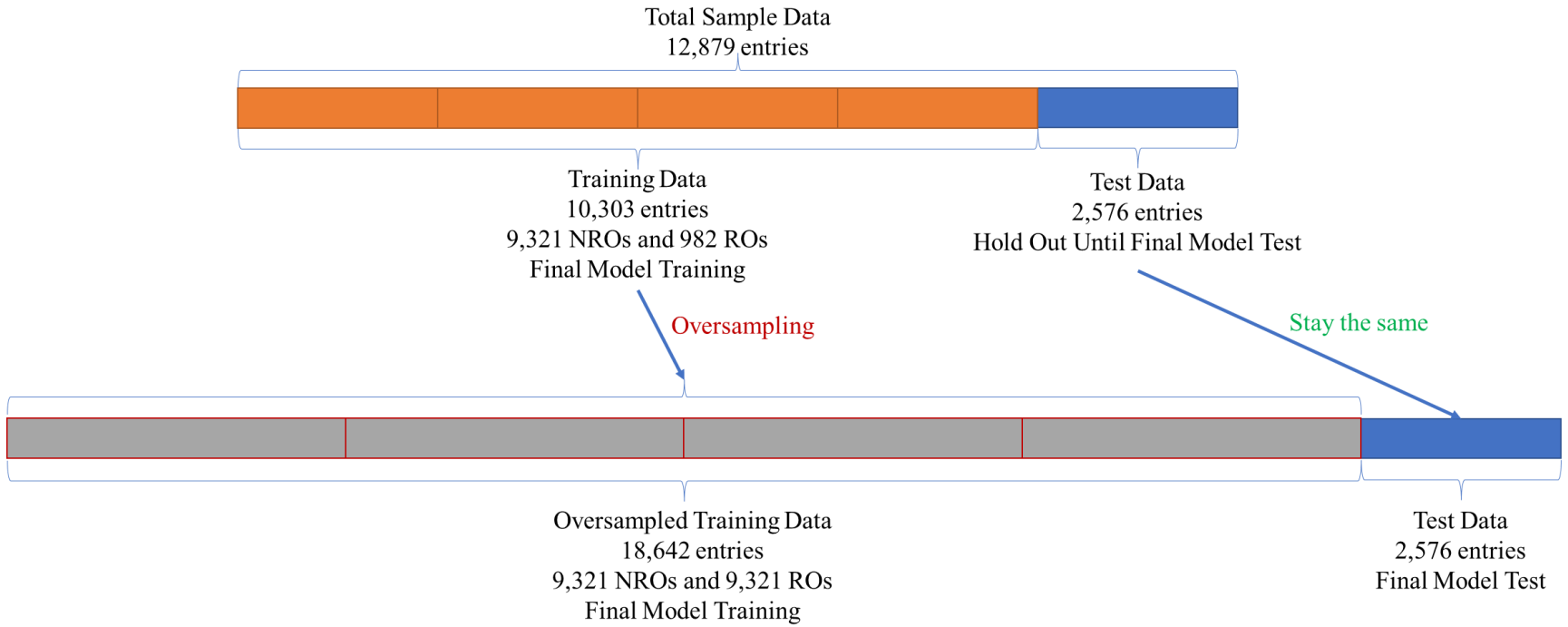

Figure 5.16. Oversampling Implementation on the Entire Training Data

5.2.1. Decision Tree

As discussed in section 5.1.1, Decision Tree prediction started with hyperparameter tuning and pruning. The same set of maximum depth [2, 4, 5, 6, 8, 10, 12, 15, 20] was tested with the oversampled dataset (illustrated in Figures 5.14, 5.15 and 5.16). Prediction with a maximum depth of 4 achieved the highest mean F1 score, so hyperparameter four was selected for further modeling. A complete result summary of all nine hyperparameters for the Decision Tree on oversampled data can be found in Table 5.9. The 5-fold cross-validation results associated with this hyperparameter on training data were reported in Table 5.10. Before CCP, the performance indices for Decision Tree with a maximum depth of 4 on the test data were 71.9% for accuracy, 51.6% for recall, 17.5% for precision, and 26.1% for F1 score.

Then, the cost complexity pruning (CCP) method was applied. Figure 5.17 shows the change of F1 values with different α values on both training data and test data. By observing the plots, the estimated α value that resulted in the highest F1 score was either at 0 or between intervals of 0.02 to 0.06. The F1 score stayed unchanged when the α value was between 0.02 and 0.06. Thus, a comparison of model performance from 5-Fold Cross-Validation with complexity parameters α of 0 and 0.02 were reported in Table 5.11. By comparing the accuracy, recall, precision, and F1 score, the model performance with α of 0 was better than the model with α of 0.02, indicating that no pruning was needed. Thus, the hyperparameter optimization ended with selecting a maximum depth of 4 and no pruning.

Table 5.9. Hyperparameter Tuning Results for Decision Tree on Resampled Dataset.

Rank	Parameter	Validation F1 Scores					Mean	SD
	Max Depth	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**4**	**25.36%**	**25.26%**	**27.93%**	**24.24%**	**28.85%**	**26.33%**	**1.75%**
2	2	26.18%	23.76%	24.83%	23.96%	26.38%	25.02%	1.09%
3	15	25.11%	25.27%	22.99%	23.53%	25.62%	24.50%	1.04%
4	20	25.50%	24.07%	24.27%	23.93%	23.50%	24.26%	0.67%
5	12	27.12%	22.08%	22.00%	22.41%	27.29%	24.18%	2.47%
6	10	27.25%	23.61%	17.46%	24.86%	26.69%	23.98%	3.51%
7	5	23.20%	22.61%	22.97%	24.19%	25.13%	23.62%	0.92%
8	6	16.67%	23.02%	20.95%	24.06%	24.52%	21.84%	2.86%
9	8	19.10%	21.88%	21.60%	19.10%	24.94%	21.32%	2.16%

Table 5.10. Baseline Performance Indices for Maximum Depth 4 Without CCP from 5-Fold Cross-Validation for Resampled Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	72.00%	49.00%	17.10%	25.36%
2	65.55%	60.91%	15.94%	25.26%
3	67.44%	67.01%	17.64%	27.93%
4	74.51%	44.68%	16.63%	24.24%
5	68.40%	65.67%	18.49%	28.85%
Mean	69.58%	57.45%	17.16%	26.33%

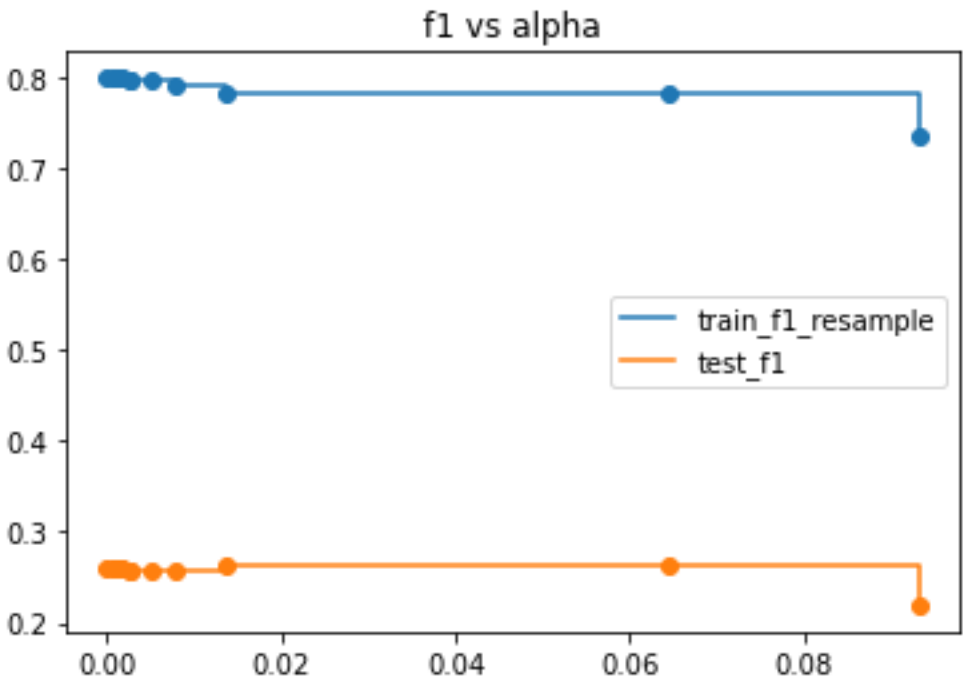

Figure 5.17. F1 Scores and Complexity hyperparameter α (Alpha) on Single Test/Train Split Data Sets

Table 5.11. Performance Indices for Maximum Depth 4 with α of 0 and 0.02 from 5-Fold Cross-Validation

Index		**Validation Accuracy**		**Validation Recall**		**Validation Precision**		**Validation F1**	
Alpha		**0**	0.02	**0**	0.02	**0**	0.02	**0**	0.02
Round	**1**	72.0%	82.6%	49.0%	25.0%	17.1%	19.4%	25.4%	21.8%
	2	65.6%	81.8%	60.9%	20.8%	15.9%	15.7%	25.3%	17.9%
	3	67.4%	57.7%	67.0%	74.2%	17.6%	14.9%	27.9%	24.8%
	4	74.5%	58.1%	44.7%	72.3%	16.6%	14.4%	24.2%	24.0%
	5	68.4%	61.3%	65.7%	71.1%	18.5%	16.2%	28.9%	26.4%
Mean		**69.6%**	68.3%	**57.5%**	52.7%	**17.2%**	16.1%	**26.3%**	23.0%
SD		**3.2%**	11.4%	**9.0%**	24.4%	**0.9%**	1.8%	**1.8%**	2.9%

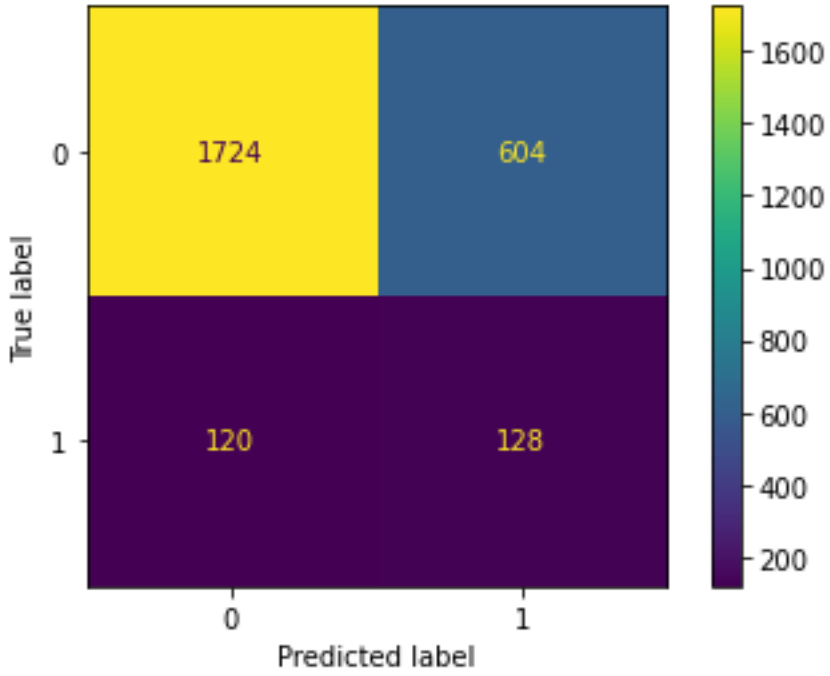

Figure 5.18. Confusion Matrix for Resampled Decision Tree Model on Test Data

The confusion matrix of the optimized model for test data was output in Figure 5.18. Performance indices were 71.9% for accuracy, 51.6% for recall, 17.5% for precision, and 26.1% for F1 score. Compared to the Decision Tree model performance on the original data, the model with oversampling didn't improve model performance. Only recall increased from 19.4% to 51.6. Accuracy decreased from 91.5% to 71.9%, precision decreased from 72.7% to 17.5%, and F1 score decreased slightly from 30.6% to 26.1%. It was unacceptable that this model misclassified 604 non-repeat offenders to repeat offenders. Due to the poor performance on this prediction, no further insights were obtained from the feature importance plot (Figure 5.19 and Figure 5.20) and Decision Tree visualization (Figure 5.21).

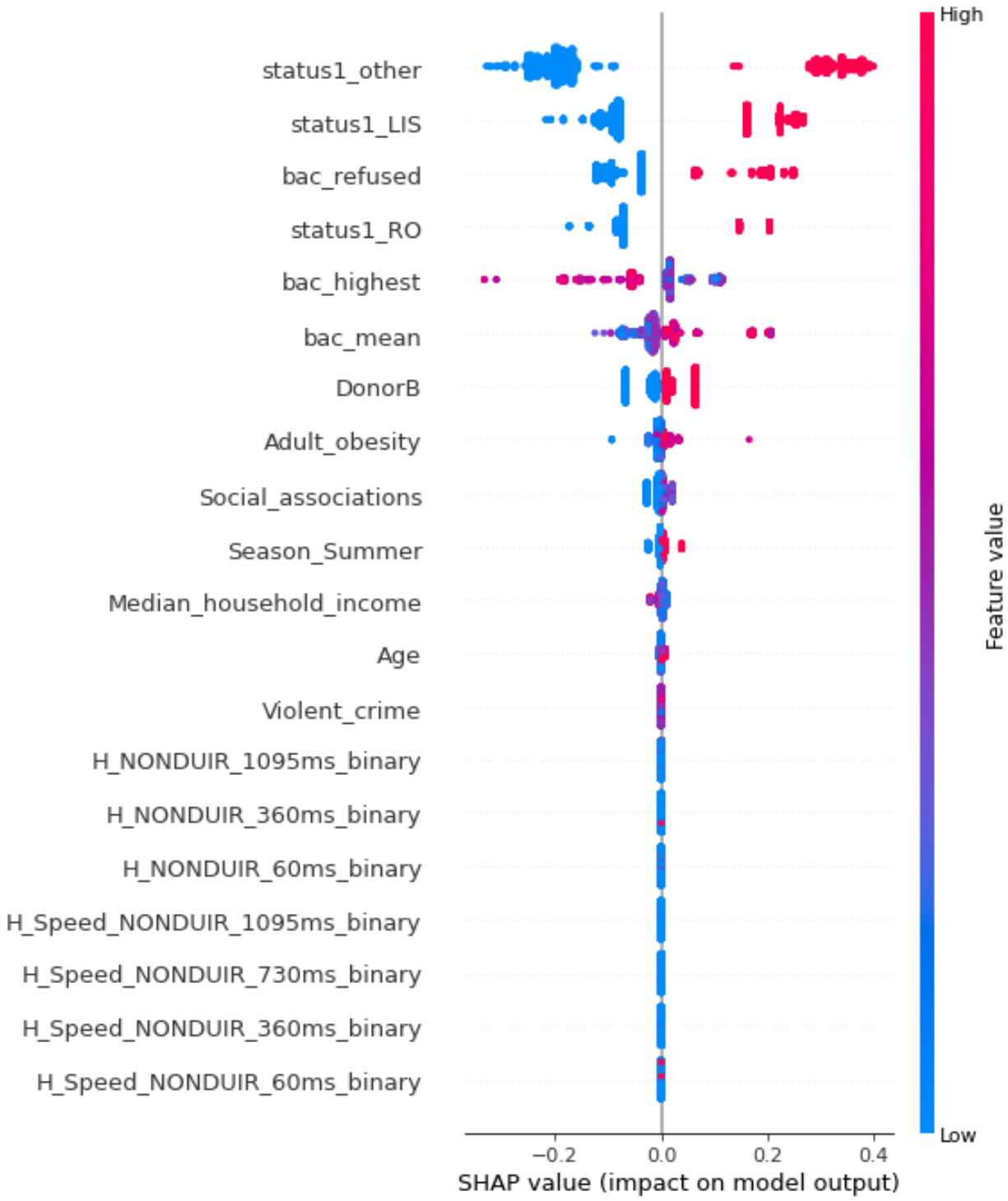

Figure 5.19. Features Importance Plot for Decision Tree Prediction on Test Data

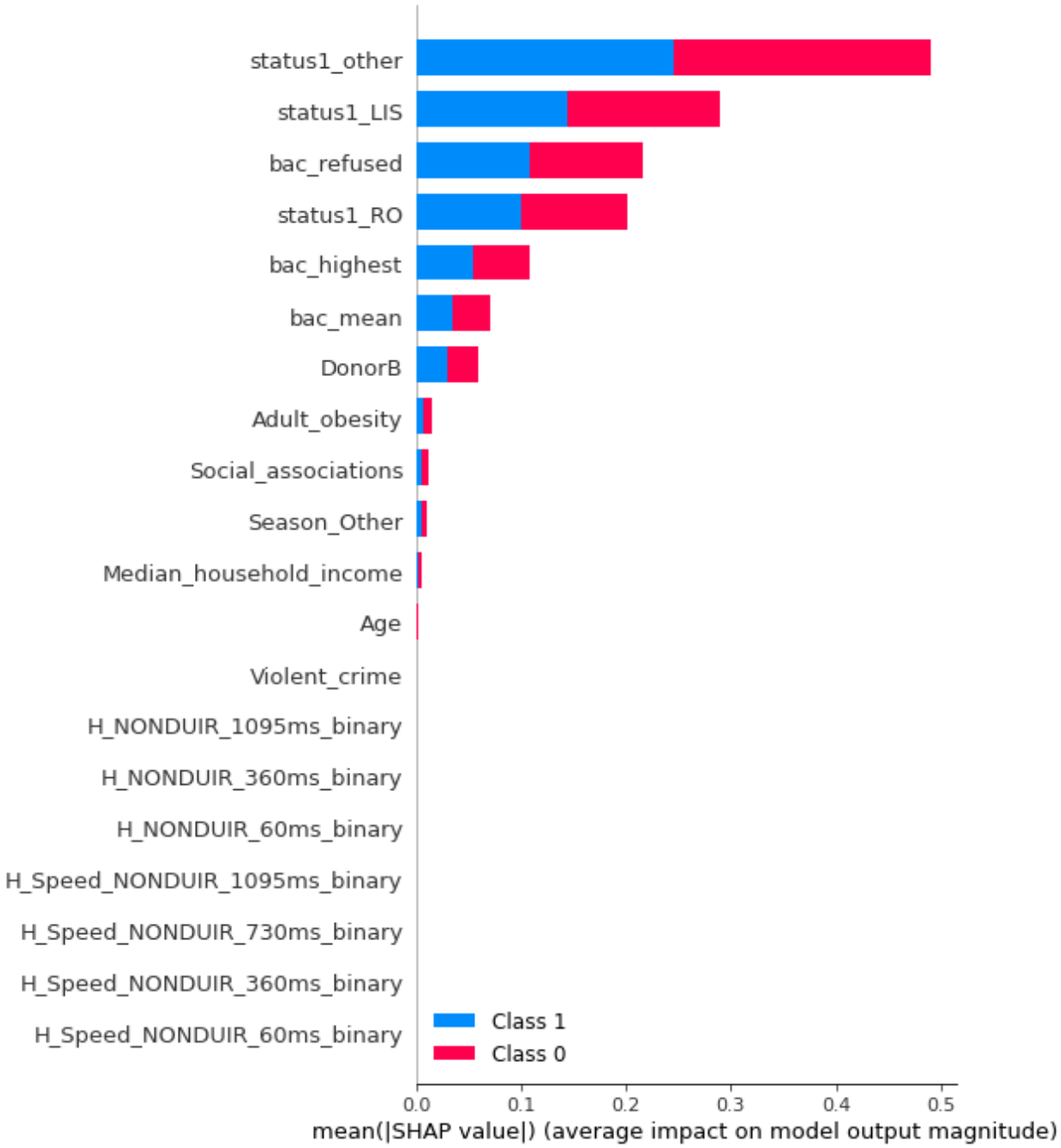

Figure 5.20. Feature Impacts on Decision Tree Prediction on Test Data

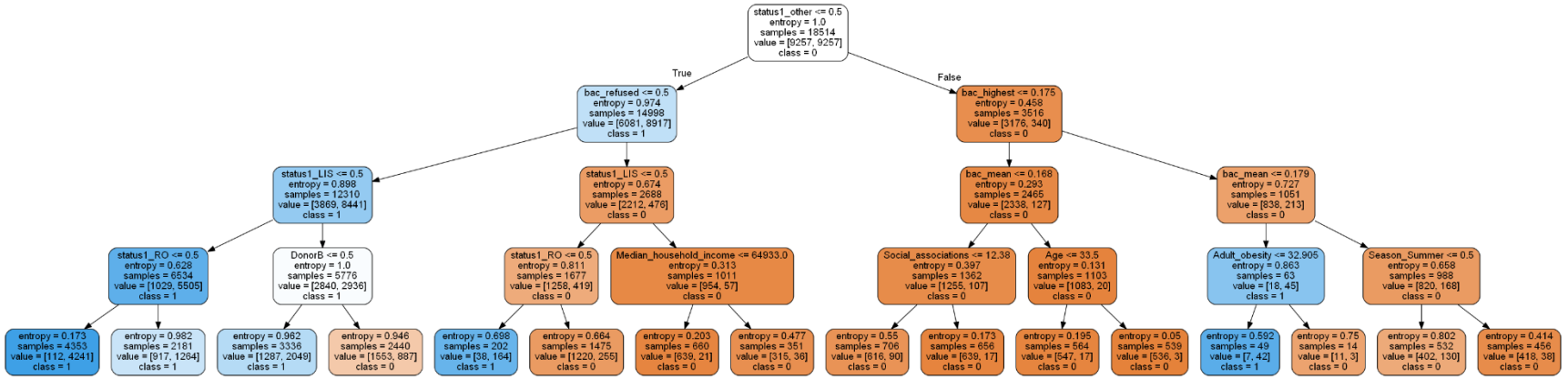

Figure 5.21. Decision Tree Visualization

5.2.2. Random Forest

In the Random Forest model, hyperparameter optimization was performed on hyperparameter tuning only. Tuning was performed on three hyperparameters – the number of estimators, maximum depth, and random state. There were four candidates of number of estimators/trees [20, 50, 100, 200], eight candidates of maximum depth [4, 5, 6, 8, 10, 12, 15, 20] and three candidates of random states [13, 16, 20]. In total, 96 combinations of hyperparameters were tested with oversampled training data with 5-fold cross-validation, as illustrated in Figures 5.14, 5.15 and 5.16. The total calculation time was approximately 90 minutes. The top 30 combinations with the highest F1 mean scores were reported in Table 5.12.

Hyperparameter combination with a maximum depth of 8, 20 trees, and a random state of 13 achieved the best performance in terms of F1 score. The associated F1 score was 22.9%. The 5-fold cross-validation results associated with this hyperparameter combination on training data were reported in Table 5.13. The recall and F1 scores improved across all five folds compared to the random forest performance on the original model. Mean recall increased from 3.67% to 34.08%, and mean F1 score increased from 6.84% to 22.89%. However, accuracy and precision decreased. Mean accuracy decreased from 90.52% to 78.14%, and mean precision decreased from 53.15 to 17.25%. Based on the model performance reported from Hyperparameter optimization, the oversampling technique did not improve the performance of Random Forest.

Table 5.12. Top 30 of Hyperparameter Tuning Results for Random Forest on Resampled Dataset.

Rank	Parameters			Validation F1 Scores					Mean	SD
	Max Depth	Estimators	Random State	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**8**	**20**	**13**	**22.4%**	**18.9%**	**23.9%**	**22.3%**	**26.9%**	**22.9%**	**2.59%**
2	5	150	16	23.4%	19.5%	21.9%	23.0%	25.9%	22.7%	2.09%
3	4	75	16	22.9%	21.1%	22.6%	22.4%	24.7%	22.7%	1.16%
4	4	180	16	23.7%	20.1%	22.1%	23.1%	24.5%	22.7%	1.53%
5	10	20	20	22.7%	18.9%	22.4%	24.0%	25.4%	22.7%	2.17%
6	8	50	13	22.5%	20.3%	22.2%	22.7%	25.6%	22.7%	1.71%
7	4	100	16	23.6%	20.7%	22.0%	22.9%	24.0%	22.6%	1.19%
8	10	20	13	22.5%	19.2%	23.2%	23.6%	24.7%	22.6%	1.85%
9	4	200	16	23.7%	20.4%	21.9%	23.0%	24.1%	22.6%	1.34%
10	4	180	13	23.2%	20.6%	21.6%	23.4%	24.1%	22.6%	1.28%
11	4	150	16	23.0%	20.4%	21.8%	23.1%	24.4%	22.5%	1.35%
12	4	200	13	23.2%	20.3%	21.9%	22.8%	24.3%	22.5%	1.34%
13	8	50	16	22.0%	16.7%	23.8%	24.7%	25.3%	22.5%	3.12%
14	4	120	13	22.5%	19.8%	22.2%	23.7%	24.2%	22.5%	1.51%
15	5	180	16	23.3%	19.0%	21.9%	23.4%	24.6%	22.5%	1.91%
16	5	200	16	23.8%	18.8%	21.7%	23.4%	24.6%	22.4%	2.06%
17	5	180	20	23.2%	19.0%	22.5%	22.5%	24.7%	22.4%	1.89%
18	4	120	16	22.9%	20.3%	21.8%	22.7%	24.1%	22.4%	1.29%
19	5	120	16	23.1%	19.0%	21.6%	22.7%	25.2%	22.3%	2.03%
20	5	100	16	23.3%	19.4%	21.2%	22.3%	25.2%	22.3%	1.93%
21	10	120	13	21.6%	17.8%	22.3%	21.9%	27.8%	22.3%	3.22%
22	4	50	16	23.0%	21.5%	20.7%	22.1%	24.0%	22.3%	1.14%
23	4	100	13	22.5%	19.8%	21.6%	23.3%	23.9%	22.2%	1.43%
24	8	100	13	24.1%	18.0%	22.3%	21.4%	25.3%	22.2%	2.50%
25	10	50	13	22.7%	16.9%	22.2%	23.3%	25.7%	22.1%	2.88%
26	5	50	16	22.9%	19.9%	21.5%	21.4%	25.0%	22.1%	1.70%
27	5	150	20	22.3%	19.3%	22.3%	22.1%	24.5%	22.1%	1.68%
28	5	200	20	22.7%	18.7%	22.1%	22.4%	24.4%	22.1%	1.85%
29	5	180	13	22.5%	19.0%	22.2%	22.8%	23.9%	22.1%	1.65%
30	4	150	13	22.2%	19.8%	22.0%	22.5%	24.0%	22.1%	1.35%

Table 5.13. Final Model Performance Indices from 5-Fold Cross-Validation on Train Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	77.15%	34.00%	16.71%	22.41%
2	77.54%	27.41%	14.44%	18.91%
3	78.07%	36.60%	17.75%	23.91%
4	78.01%	34.57%	16.46%	22.30%
5	79.95%	37.81%	20.88%	26.90%
Mean	78.14%	34.08%	17.25%	22.89%

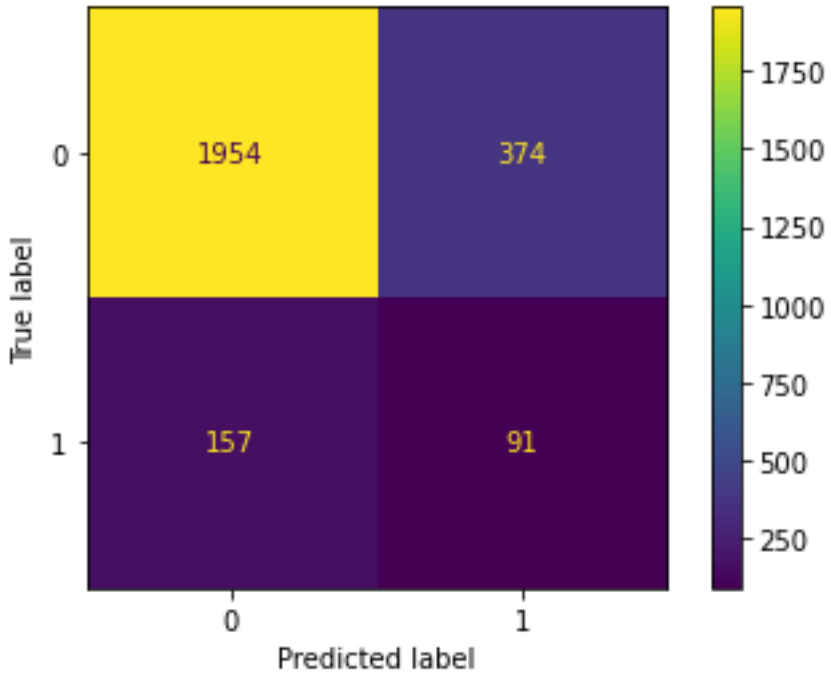

Figure 5.22. Confusion Matrix for Resampled Random Forest Model on Test Data

Figure 5.22 presents the confusion matrix for test data. The performance indices on the test data were 79.4% for accuracy, 36.7% for recall, 19.6% for precision, and 25.5% for the F1 score. Oversampling technique did not improve the model performance. Recall and precision were low. Because of low precision, 374 non-repeat offenders were misclassified to repeat offenders, which was unacceptable. Thus, due to the model's poor performance, no further insights were obtained from Figure 5.23 and Figure 5.24 about features' impact on the prediction.

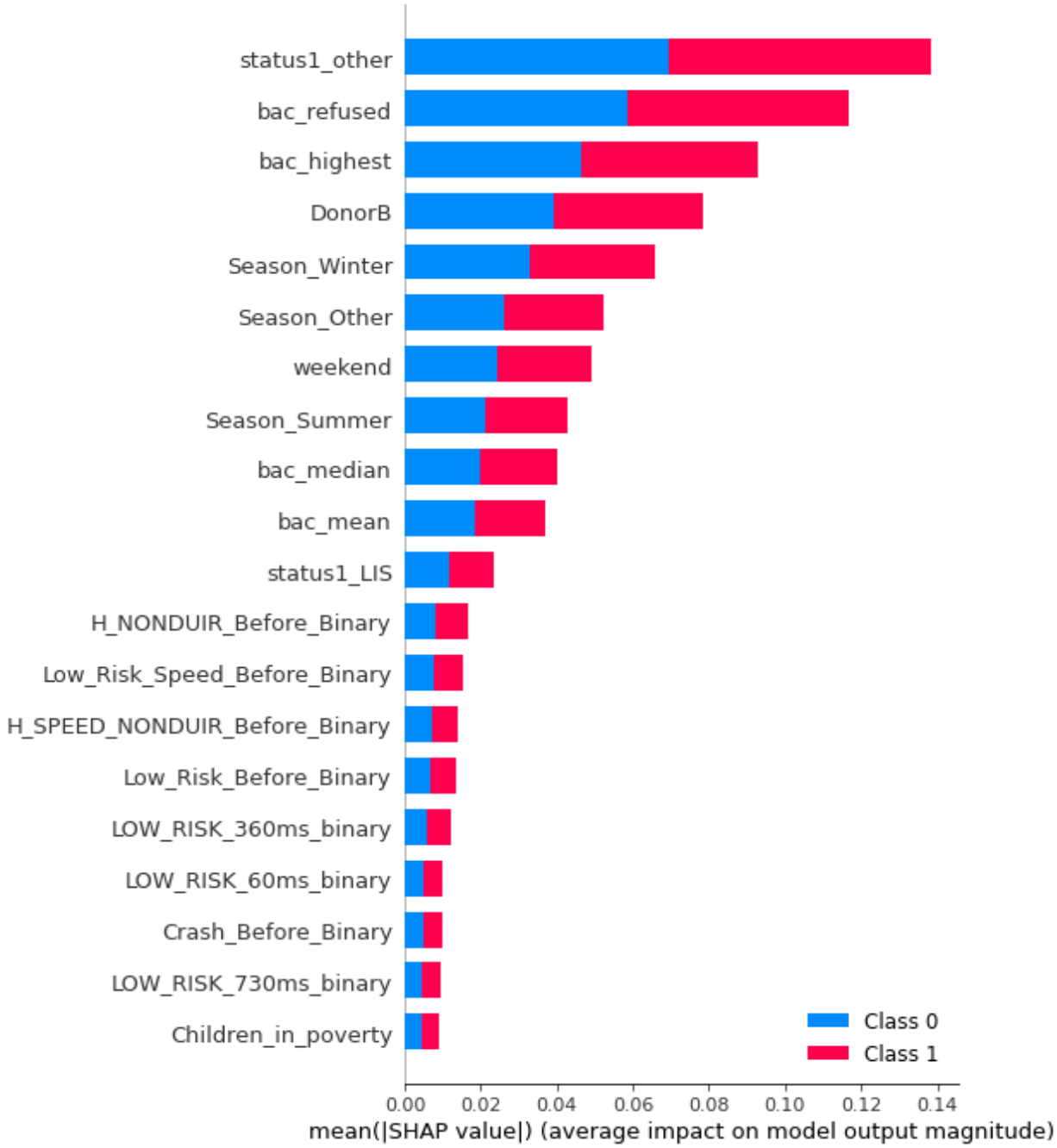

Figure 5.23. Features Importance Plot for Random Forest Prediction on Test Data

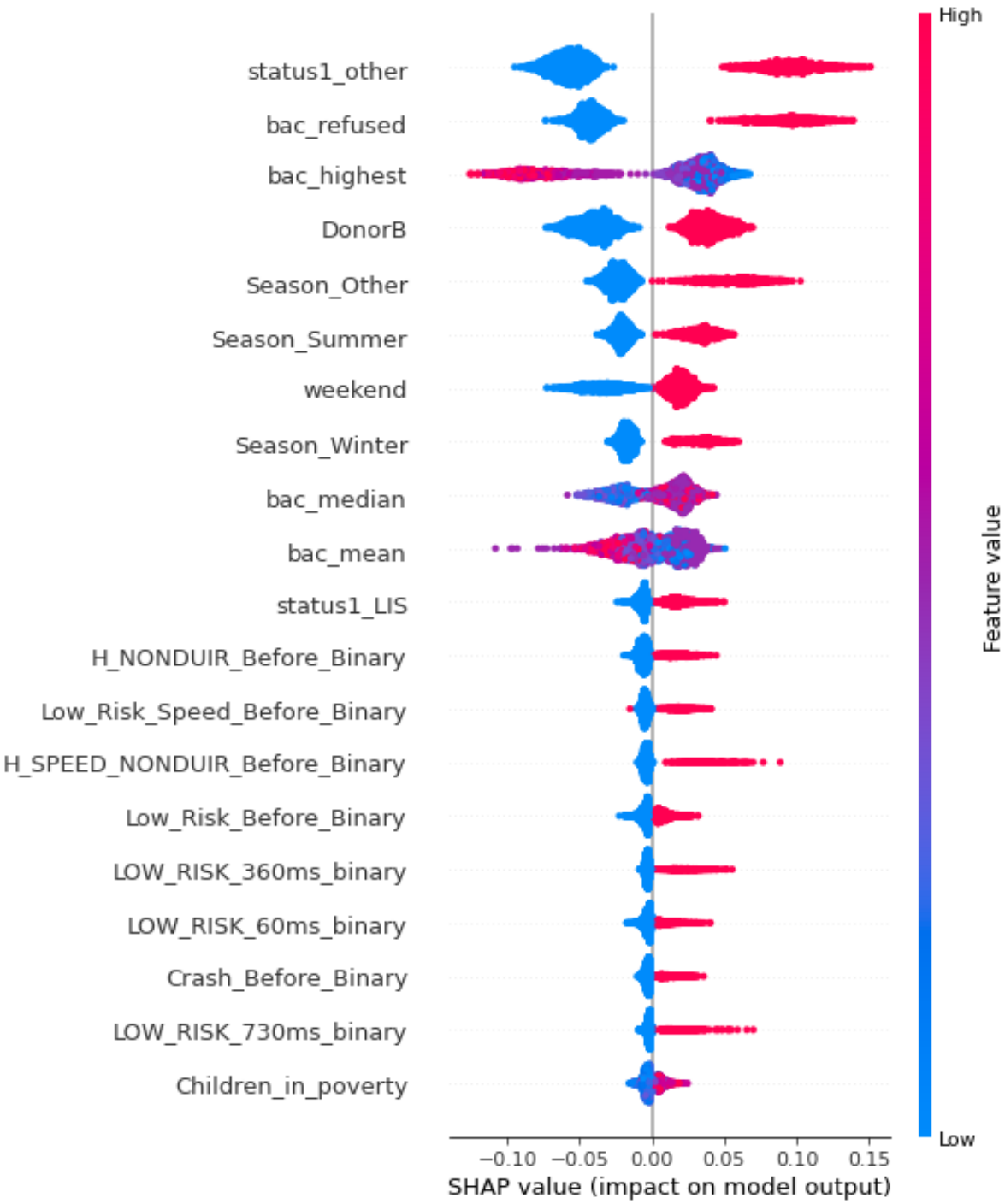

Figure 5.24. Feature Impacts on Random Forest Prediction on Test Data

5.2.3. Gradient Boosting

For the Gradient Boosting model, hyperparameter optimization was completed with hyperparameter tuning only. The tuning process is performed on four hyperparameters – learning rate, the number of estimators, maximum depth, and random state. There were seven candidates of learning rate [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2], seven candidates of number of estimators/trees [10, 20, 50, 100, 120, 150, 200] seven candidates of maximum depth [4, 5, 8, 10, 12, 15, 20] and two candidates of random states [0, 13]. In total, 686 combinations of hyperparameters were tested with oversampled training data with 5-fold cross-validation. The total calculation time was approximately 18 hours. The top 30 combinations with the highest F1 mean scores were reported in Table 5.7.

Hyperparameter combination with a learning rate of 0.2, maximum depth of 4, 200 trees, and random state of 13 achieved the best performance in terms of F1 score. The associated mean F1 score with this combination was 40% on the training data, which was acceptable in model prediction power. The 5-fold cross-validation results associated with this hyperparameter combination on training data were reported in Table 5.14. Compared to the Gradient Boosting prediction on the original data, the oversampling technique did improve the prediction power of this oversampled model in all four indices. Mean accuracy improved from 90.83% to 91.45%. Mean recall increased from 22.98% to 30.05%, and Mean precision increased from 54.57% to 59.95%. The mean F1 score increased from 32.21% to 40.02%.

Table 5.14. Top 30 of Hyperparameter Tuning Results for Gradient Boosting on Resampled Dataset.

Rank	Parameters				Validation F1 Scores					Mean	SD
	Learning Rate	Max Depth	Estimators	Random State	Split 1	Split 2	Split 3	Split 4	Split 5		
1	**0.2**	**4**	**200**	**13**	**45.8 %**	**35.4 %**	**38.2 %**	**36.5 %**	**44.2 %**	**40.0 %**	**4.22 %**
2	0.2	4	200	0	45.5%	34.4%	38.2%	36.2%	44.1%	39.7%	4.38%
3	0.2	4	150	13	44.6%	36.7%	36.5%	35.9%	43.1%	39.4%	3.72%
4	0.1	5	200	13	48.5%	33.2%	35.2%	36.8%	43.0%	39.3%	5.62%
5	0.2	5	150	13	47.5%	33.6%	33.4%	36.0%	45.7%	39.2%	6.11%
6	0.1	5	200	0	48.0%	33.7%	34.9%	36.6%	42.7%	39.2%	5.39%
7	0.2	4	150	0	44.8%	36.2%	36.1%	35.6%	42.9%	39.1%	3.91%
8	0.2	5	120	13	47.5%	34.8%	33.1%	35.5%	44.4%	39.1%	5.77%
9	0.2	5	100	0	45.3%	33.1%	32.4%	38.2%	46.3%	39.1%	5.87%
10	0.15	5	200	13	44.4%	33.0%	34.8%	37.7%	44.9%	39.0%	4.89%
11	0.075	5	200	13	46.6%	33.2%	36.2%	35.9%	42.9%	38.9%	4.96%
12	0.2	5	120	0	47.2%	31.8%	33.0%	36.7%	45.9%	38.9%	6.45%
13	0.1	5	150	0	45.6%	32.4%	35.5%	37.2%	43.8%	38.9%	5.02%
14	0.075	5	200	0	46.6%	33.3%	35.9%	35.9%	42.7%	38.9%	4.94%
15	0.1	5	120	0	45.9%	31.8%	35.1%	37.1%	44.3%	38.8%	5.42%
16	0.2	4	100	0	42.4%	36.3%	36.6%	36.9%	41.9%	38.8%	2.73%
17	0.1	5	120	13	45.8%	31.7%	35.0%	36.8%	44.8%	38.8%	5.56%
18	0.1	5	150	13	45.7%	32.4%	35.5%	36.5%	43.8%	38.8%	5.12%
19	0.15	4	200	13	42.2%	35.5%	36.4%	38.1%	41.5%	38.8%	2.68%
20	0.2	5	200	13	46.8%	33.7%	33.2%	35.5%	44.6%	38.7%	5.75%
21	0.2	4	100	13	42.0%	36.1%	36.7%	36.9%	41.9%	38.7%	2.66%
22	0.2	5	150	0	47.7%	32.6%	33.3%	36.1%	43.8%	38.7%	5.98%
23	0.2	5	100	13	45.7%	32.9%	33.1%	36.8%	44.8%	38.7%	5.57%
24	0.2	4	120	13	42.6%	35.5%	37.6%	36.1%	41.4%	38.6%	2.87%
25	0.15	5	150	13	43.7%	32.5%	36.4%	37.2%	43.3%	38.6%	4.29%
26	0.15	5	200	0	45.3%	32.6%	33.9%	36.3%	44.9%	38.6%	5.44%
27	0.2	4	120	0	42.9%	35.6%	37.5%	35.5%	41.4%	38.6%	3.03%
28	0.15	4	200	0	42.2%	35.7%	35.5%	37.9%	41.5%	38.6%	2.84%
29	0.15	5	150	0	43.5%	33.2%	35.3%	37.3%	43.3%	38.5%	4.20%
30	0.075	5	150	0	45.1%	33.1%	34.8%	36.7%	42.9%	38.5%	4.68%

Table 5.15. Final Model Performance Indices from 5-Fold Cross-Validation on Train Data

	Validation Accuracy	Validation Recall	Validation Precision	Validation F1
1	92.09%	34.50%	68.32%	**45.85%**
2	90.78%	26.40%	53.61%	**35.37%**
3	91.36%	28.35%	58.51%	**38.19%**
4	91.21%	27.66%	53.61%	**36.49%**
5	91.80%	33.33%	65.69%	**44.22%**
Mean	91.45%	30.05%	59.95%	40.02%

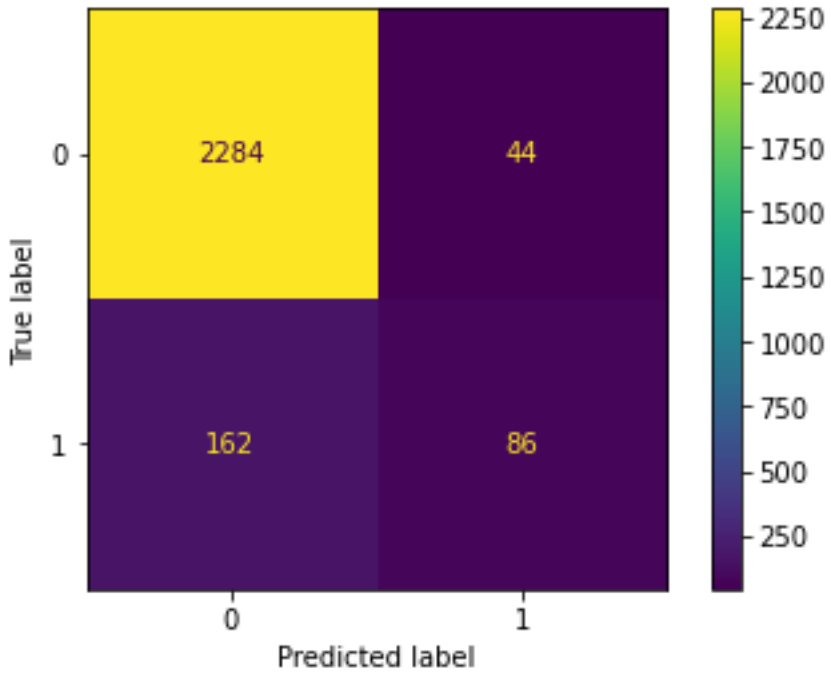

Figure 5.25. Confusion Matrix for Resampled Gradient Boosting Model on Test Data

Figure 5.25 presents the confusion matrix for test data. The performance indices on the test data were 92.0% for accuracy, 34.7% for recall, 66.2% for precision, and 45.5 % for F1 score. The Recall was still low to produce a good prediction, but precision and F1 score were acceptable. All four indices improved from the indices in the Gradient Boosting model without oversampling. This model performed the best among all six models.

Figure 5.26 and Figure 5.27 showed the rank, magnitude, and directionality of the feature impact on the predictions for test data. However, because the model performance was acceptable, insights obtained from Figure 5.26 and Figure 5.27 should be used cautiously.

As shown in Figure 5.26, the top five predictors were the highest BAC value, a binary variable of driver's license status “other,” and “license suspended,” binary variable of season “other” and “winter.” In Figure 5.27, a small purple area was evident on the highest BAC value in SHAP values -0.5 to -1. The purple tail extended to the negative side in SHAP values -2.2 to 1. These purple areas indicated that the directionality of impacts of the highest BAC value could be complicated within these two intervals. However, the blue dot was distributed mainly at an interval less than -2.2, indicating a lower value for the highest BAC on record would lead to a lower chance of recidivism; the red dot was distributed mainly at an interval more significant than 0, indicating a higher value for the highest BAC on record would lead to a higher chance of recidivism. Further analysis is needed to understand the impact of the values in the purple area. The logistic model will further explain other categorical variables with more than two classes.

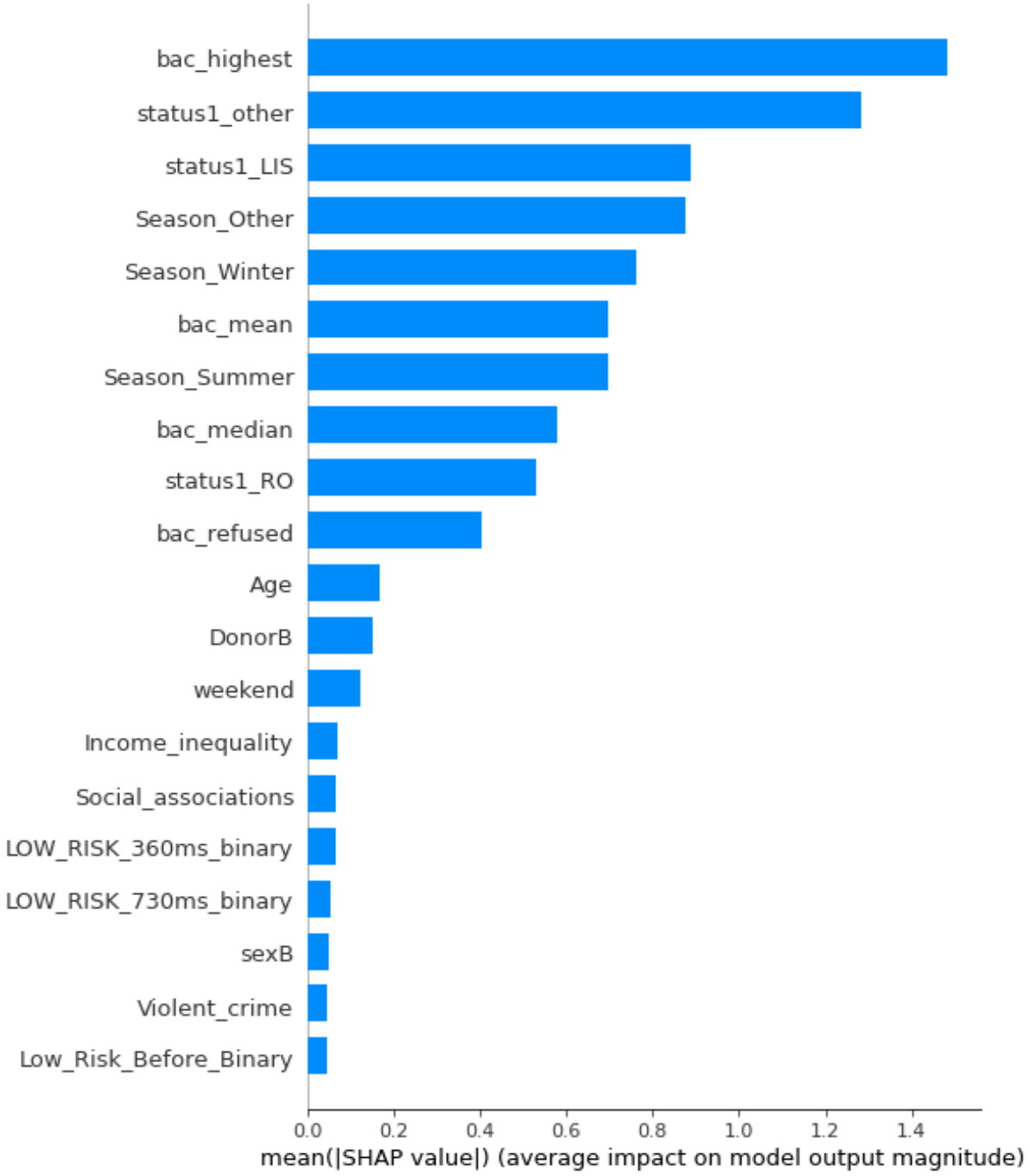

Figure 5.26. Features Importance Plot for Gradient Boosting Prediction on Test Data

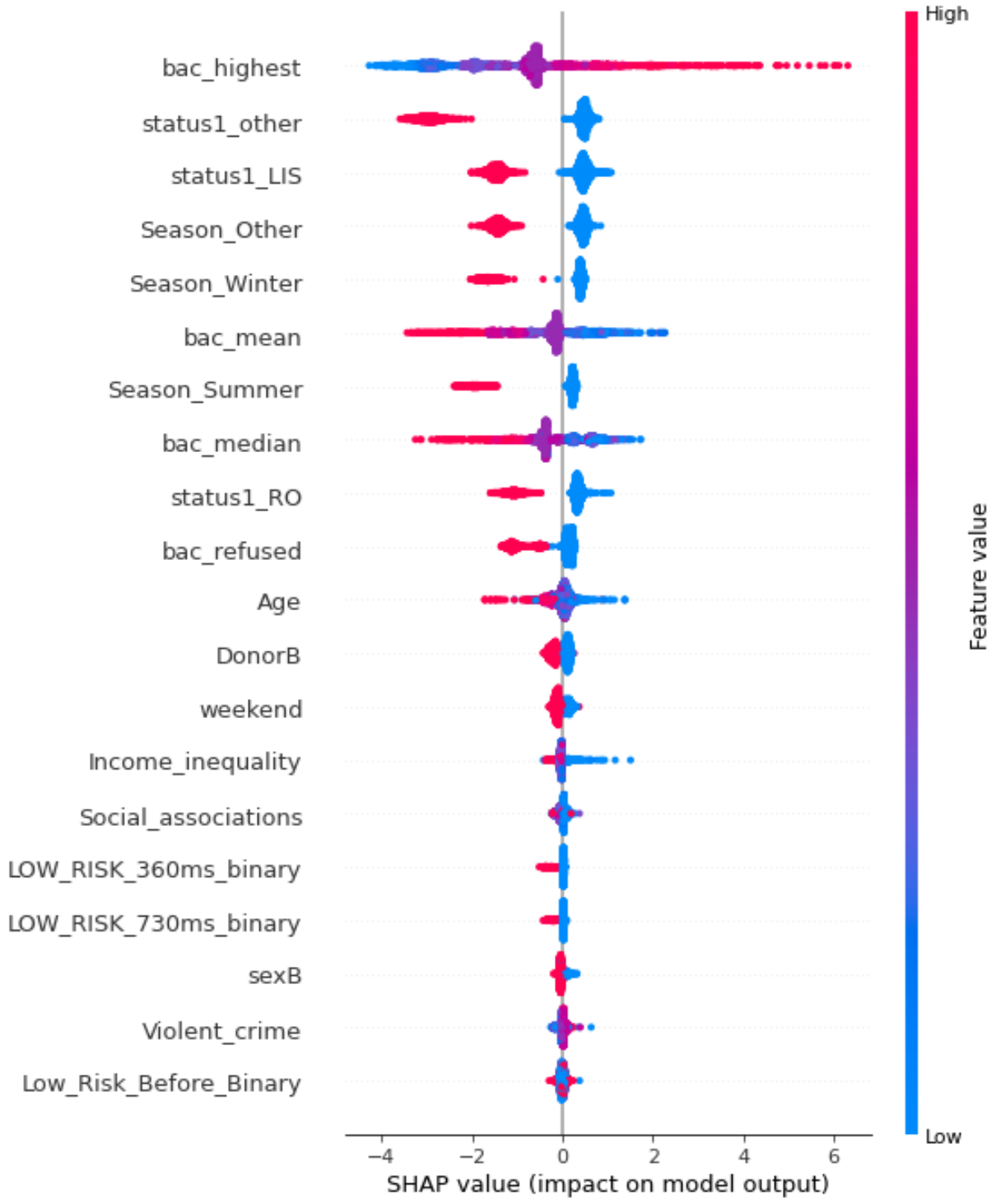

Figure 5.27. Feature Impacts on Gradient Boosting Prediction on Test Data

5.3. Performance Comparison

As shown in Table 5.16, for this imbalanced data sample, the F1 score was used to measure the prediction power. Among all six predictions, only the prediction from Gradient Boosting on the original data and the oversampled data was considered acceptable. In Berk and Bleich (2013), random forest prediction achieved 71.0% for accuracy, 62.8% for recall, 26.3% for precision, and 37.1% for F1, and gradient boosting achieved 66.3% for accuracy, 55.6% for recall, 21.4% for precision, and 30.9% for F1. The Gradient Boosting prediction in this study achieved comparable or better results. For the oversampled data, the F1 score was 45.5%, which was superior to Berk and Bleich (2013).

Based on all four indices, Random Forest predictions made on original and oversampled data were the worst. Although surprised, this insight was reasonable. A single tree would be less likely to favor the minority group, and a parallel tree would be much less likely to favor the minority group. Although Random Forest performed with bagging and bootstrapping features, there are chances that the training process was performed heavily on the majority class. Oversampling didn't significantly improve the performance for Random Forest, showing its limited power to improve prediction power. It may also indicate that the differences between two classes were small.

For imbalanced data, an algorithm such as Gradient Boosting that build learner sequentially to lower the overall error rate in each step may be more appropriate. Although oversampling techniques did improve the model performance of Gradient Boosting, the improvement was limited. Further data cleaning to remove outliers in both classes might be needed to improve prediction power.

There were a few common leading factors from two Gradient Boosting predictions. They were highest BAC values, mean BAC values, median BAC values, BAC refusal, age, first DUI on weekend, gender, social associations index in driver's county of residence, three driver license status code. Factor ranks from Random Forest and Decision Tree were disregarded due to their low prediction power.

Table 5.16. Model Prediction Summary

Index	Method					
	Decision Tree		Random Forest		Gradient Boosting	
	No Oversampling	Oversampling	No Oversampling	Oversampling	No Oversampling	Oversampling
Accuracy	91.5%	71.9%	90.4%	79.4%	91.3%	92.0%
Recall	19.4%	51.6%	2.0%	36.7%	26.2%	34.7%
Precision	72.7%	17.5%	50.0%	19.6%	61.9%	66.2%
F1 score	30.6%	26.1%	3.9%	25.5%	36.8%	45.5%

5.4. Factor Explanations

Based on feature importance ranks produced by Gradient Boosting on an oversampled dataset, 15 variables were selected for logistic regression analysis to quantify their effects on the outcome. Driver's license status code at the year of the first DUI was also chosen by the machine learning model initially. However, due to the complexity of the data recording process for this variable and uncleared causation with DUI recidivism, this variable was removed from the list to avoid misinterpretation.

The list of selected variables and their descriptive statistics is shown in Table 5.17. Univariate logistics analyses were performed between these variables and DUI recidivism. Variables that were statistically significant in univariate analysis were reported in Table 5.17. Note the dataset was further inspected on the distributions of the selected variables. Ninety-six records were removed for three filters: (1) drivers aged 21 and older with a BAC value larger than 0.4 on age over 21; (2) drivers aged 21 and older with BAC less than 0.04; (3) drivers under 21 with BAC less than 0.02. After removal, the dataset had 12,783 records, with 11,656 non-repeat offenders (90.47%) and 1,218 repeat offenders (9.53%). The rate between non-repeat offenders and repeat offenders stayed identical. The statistics in Table 5.17 reflected this change.

Among the selected variables, the associations between variables were first evaluated. Several variables were highly related to other variables, so only one of these variables was used for logistic regression modeling at a time. Prior 3-year low-risk citations included prior 1-year low-risk citation and prior second-year low-risk citation (360-730 days before first DUI), so Prior 3-year low-risk citations was tested alone in some models. In contrast, the latter two were tested together in other models.

In this sample, 31.1% of drivers refused to take the BAC test, and the remaining 68.9% of drivers took the BAC test. The blood alcohol concentration (BAC) record and BAC test refusal were mutually exclusive, so they were tested separately. Among those who took the BAC test, 11.4% of drivers have more than two BAC values on record, so the highest BAC value, the mean BAC value, and the median BAC value were the same for 88.6% of the drivers. Based on descriptive statistics in Table 5.17, the mean BAC and median BAC were almost identical, so only the mean BAC was used. Thus, the highest BAC value and the mean BAC value were tested separately in logistic regression modeling.

Correlations among the rest of the variables were also evaluated. Pearson's correlation coefficient was used for the numerical variables: mean BAC mean, income inequality ratio, social associations ratio and violent crime ratio. As discussed in last paragraph, mean BAC mean, highest BAC and median BAC values were highly related, so only the mean BAC was test with other numerical variables for correlation. Pearson's correlation coefficients showed that there was no correlation between any of the pair of these numerical variables. ANOVA was used to test correlations between numerical variables and the categorical variables here, and results showed that there was no correlation between any of the pair of these variables.

Chi-square was used to test correlations between categorical variables. Age and BAC test refusal have associations with other variables. Age group variable has associations with first DUI during weekend (p-value<0.0001, Cramer's V statistics =0.0618), prior 3-year low-risk citation(s) (p-value<0.0001, Cramer's V statistics =0.1098), gender (p-value<0.0001, Cramer's V statistics =0.0484), and BAC test refusal (p-value<0.0001, Cramer's V statistics =0.0703). In addition, BAC test refusal has association with gender (p-value=0.0199, Cramer's V statistics =0.0206) and weekend (p-value<0.0001, Cramer's V statistics =0.0624). An association is

considered as “little if any association” if Cramer’s V statistics is between 0 to 0.1, and an association is “low association” if Cramer’s V statistics is between 0.1 to 0.3 (Crewson, n.d.). Thus, all the associations between variables abovementioned were either “little if any association” or “low association”. Age and BAC test refusal were kept on the list due to their importance in the impaired driving literature and low associations with other variables. In addition, interactive terms between these associated variables were also evaluated in the models, but none of them were statistically significant.

Table 5.17. Variable/Predictor Candidates and Their Descriptive Statistics

Variables	**NROs (N=11,565)**		**ROs (N=1,218)**	
	N	**%**	**N**	**%**
BAC test refusal**				
No	7,799	67.44%	1,009	82.84%
Yes	3,766	32.56%	209	17.16%
Gender**				
Female	2,968	25.56%	267	21.92%
Male	8,597	74.34%	951	78.08%
Age at first DUI*				
18-24	3,002	25.96%	355	29.15%
25-34	3,901	33.73%	398	32.68%
35-44	2,140	18.50%	214	17.57%
45-54	1,647	14.24%	168	13.79%
55-64	727	6.29%	71	5.83%
64+	148	1.28%	12	0.99%
First DUI on the weekend**				
No	3,955	34.20%	464	38.10%
Yes	7,610	65.80%	754	61.90%
Registered for organ donor				
No	5,985	51.75%	652	53.53%
Yes	5,580	48.25%	566	46.47%
Prior 1-year low-risk citations*				
No	9,893	85.54%	1,023	83.99%
Yes	1,672	14.46%	195	16.01%
Prior second the year, low-risk citations				
No	9,692	83.80%	1,018	83.58%
Yes	1,873	16.20%	200	16.42%
Prior 3-year low-risk citations**				
No	6,417	55.49%	622	51.07%
Yes	5,148	44.51%	596	48.93%
Prior 3-year high-risk DUI citations				
No	11,303	97.73%	1,181	96.96%
Yes	262	2.27%	37	3.04%
	Mean	**S.D.**	**Mean**	**S.D.**
Income inequity*	4.351	0.692	4.289	0.62
Social association	15.518	6.828	15.243	6.429
Violent crime rate*	240.4	109.68	249.18	105.82
	N=7,799[1]		**N=1,009[1]**	
The highest BAC value**	0.171	0.052	0.192	0.059
The mean BAC value**	0.170	0.052	0.179	0.053
The median BAC value**	0.170	0.052	0.179	0.053

[1]The statistics were based on drivers who did not refuse the BAC test.
Univariate analysis: * p-value <0.05; **p-value <0.01

5.4.1. Factor Interpretations

Eight variables were statistically finally quantified in two logistic regression models (Table 5.18 and Table 5.19). Both models were tested with the Hosmer and Lemeshow Goodness-of-Fit, and there is no evidence that the models were a poor fit. Test Both BAC record and BAC test refusal were statistically significant and had the most prominent effects on the likelihood of DUI recidivism in the models. As shown in Table 5.18, the first model consisted of seven variables: the mean BAC value, the first DUI citation on the weekend, gender, prior 3-year low-risk citation, age of first DUI offense, income inequality ratio in driver's county of residence, and violent crime rate in driver's county of residence.

In Model 1, the mean BAC value had the most significant effect on predicting the likelihood of DUI recidivism. An increase of 0.01 in the driver's mean BAC was associated with a rise of 45.9% in the odds of DUI recidivism. This finding affirmed insights from previous studies that identified driver's BAC as a predictor of future DUI recidivism (e.g., Marowitz, 1998; C'de Baca et al., 2001; Roma et al., 2019)

Drivers whose first DUI took place on weekdays have a greater likelihood of recidivism than those on weekends (OR =1.172). This finding supported similar insights from Impinen et al. (2009). In addition, it is reasonable to believe that drivers who drink during weekdays might have some alcohol addiction, as they might rely on drinking alcohol to reduce negative emotions. Studer et al.(2014) found that alcohol use on weekdays was strongly related to coping motives to reduce a negative affect and obtain an internal reward, e.g., drinking to forget worries. Similar insight was found by Lau-Barraco et al.(2016) that weekday drinking was associated with tension-reduction expectancies among nonstudent emerging adults.

In Model 1, gender and past traffic violation history were also shown to be strong predictors. Males had a higher risk of recidivism than females (odds ratio (OR) =1.255). This finding was consistent with past studies (e.g., C'de Baca et al., 2001; Impinen et al., 2009; Robertson et al., 2016; Kubas et al., 2018; Kubas and Vachal, 2019). Drivers with at least one low-risk citation within three years before their first DUI had a higher risk than those without such citations (OR=1.201). Similar insights related to associations between traffic violations were found in previous studies (e.g., Marowitz, 1998; Hubicka et al., 2008; Robertson et al., 2016)

Two environmental factors, income inequality ratio and violent crime rate at driver's county of residence were associated with DUI recidivism. However, further studies might be needed to understand their effects on DUI recidivism. Drivers who lived in counties with higher income inequality ratios had lower chances of DUI recidivism. This finding may be related to higher income inequality ratios in North Dakota's urban areas, where alternative transportations were more accessible for drivers after drinking. Also, potentially more concentrated law enforcement presence. The violent crime rate in the driver's county of residence was positively related to DUI recidivism, although the effect was small. One violent crime offense per 100,000 population increased in driver's county of residence is associated with a 0.1% increase in DUI recidivism. Further studies might be needed to understand the causation between DUI recidivism and violent crime rate in driver's county of residence, as literature has shown a strong association between alcohol accessibility and violent crime (Gorman et al., 2001; Toomey et al., 2012; Trangenstein et al., 2018).

Surprisingly, the p-value of individual age groups did not meet the significance level of 0.05 to predict the likelihood of recidivism in the multiple logistic regression model. However, if

the significance level for the overall effects loosed to 0.1, then age became a significant predictor (0.0998) since it improved the general prediction power. It is reasonable to allow age to stay in the model for three reasons: (1) it was a significant factor in univariate analysis at a 0.05 significance level, and the odds ratios between age groups were similar to those in multiple logistic regression; (2) it had associations with first DUI on the weekend, gender, and prior 3-year low-risk citations, and age would become significant at the 0.05 level if any of these three variables were removed from the model; (3) age was identified as a strong predictor of DUI recidivism in the literature, and finding that 18-24 cohort was the highest risk group among all age group supported insights from other studies (e.g., C'de Baca et al., 2001; Impinen et al., 2009; Robertson et al., 2016). Thus, the findings from age in Model 1 were reported, although none of the age groups was statistically significant.

Among all age groups, drivers whose age of first DUI offense was 64 and older had the lowest chance of recidivism. The 18-24 cohort had the highest probability of recidivism, which was 1.462 times of odds of the 64 and above cohort. The age 25-34 cohort and the age 45-54 cohort were approximately 1.21 times of odds of the 64 and above cohort in terms of DUI recidivism. The age 35-44 cohort was 1.189 times the odds of having a DUI recidivism compared to the age 64 and above cohort.

Table 5.18. Logistics Regression Model 1

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	95% Confidence Limits	
Intercept		1	-3.1373	0.4102	58.4995	<.0001		0.043		
BAC mean		1	3.8481	0.6646	33.5302	<.0001	0.0919	46.905	12.648	171.222
First DUI on weekend	**No**	1	0.1585	0.0627	6.3955	0.0114	0.0416	1.172	1.036	1.324
First DUI on weekend	**Yes***	0	0	.	.	.	.	.		
Gender	**Male**	1	0.2274	0.0728	9.7525	0.0018	0.0545	1.255	1.09	1.45
Gender	**Female***	0	0	.	.	.	.	.		
Prior 3-year low-risk citations	**Yes**	1	0.1830	0.0610	8.9935	0.0027	0.0502	1.201	1.065	1.353
Prior 3-year low-risk citations	**No***	0	0	.	.	.	.	.		
Income inequality		1	-0.1130	0.0502	5.0609	0.0245	-0.0428	0.893	0.808	0.984
Violent crime		1	0.000634	0.000295	4.6044	0.0319	0.0373	1.001	1	1.001
Age at first DUI	**18-24**	1	0.3800	0.3069	1.5335	0.2156	0.0922	1.462	0.836	2.812
Age at first DUI	**25-34**	1	0.1906	0.3060	0.3877	0.5335	0.0496	1.210	0.693	2.324
Age at first DUI	**35-44**	1	0.1735	0.3098	0.3135	0.5756	0.0371	1.189	0.675	2.298
Age at first DUI	**45-54**	1	0.1926	0.3119	0.3812	0.5370	0.0371	1.212	0.685	2.351
Age at first DUI	**55-64**	1	0.1455	0.3259	0.1992	0.6554	0.0194	1.157	0.633	2.294
Age at first DUI	**64/+***	0	0	.	.	.	.	.		

* The reference group among categorical variables.

Table 5.19. Logistics Regression Model 2

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	95% Confidence Limits	
Intercept		1	-2.9888	0.2656	126.6657	<.0001		0.05		
BAC test refusal	**No**	1	0.8703	0.0789	121.698	<.0001	0.2221	2.388	0.358	0.488
BAC test refusal	**Yes**	0	0	.	.	.	.	.		
First DUI on weekend	**No**	1	0.2138	0.0628	11.5932	0.0007	0.0561	1.238	1.094	1.400
First DUI on weekend	**Yes**	0	0	.	.	.	.	.		
Gender	**Male**	1	0.2261	0.0729	9.6179	0.0019	0.0542	1.254	1.088	1.449
Gender	**Female***	0	0	.	.	.	.	.		
Prior 3-year low-risk citations	**Yes**	1	0.1831	0.0607	9.0954	0.0026	0.0502	1.201	1.066	1.353
Prior 3-year low-risk citations	**No**	0	0	.	.	.	.	.		
Income inequality		1	-0.1032	0.051	4.0943	0.043	-0.039	0.902	0.815	0.995
Violent crime		1	0.000775	0.000299	6.7393	0.0094	0.0456	1.001	1.000	1.001

* The reference group among categorical variables.

Model 2 comprised the BAC test refusal, the first DUI citation on the weekend, gender, prior 3-year low-risk citation, income inequality ratio in driver's county of residence, and violent crime rate in driver's county of residence (Table 5.19). Surprisingly, the drivers who did not refuse the BAC test upon arrest have 2.388 times of chance of DUI recidivism as those who refused. Notably, age at first DUI citation was not significant in this model, even when the significance level was loosed to 0.1. However, a strong association between BAC test refusal and age group was evident. Drivers aged 18-24 were least likely to refuse a BAC test upon arrest among age groups. Given this youngest group has the highest risk of DUI recidivism, it is reasonable that drivers who complied with the BAC test had a higher risk of DUI recidivism than those who refused. The coefficients of other factors were similar to those in Model 1.

5.4.2. Logistic Regression Predictions

Like other regression model, logistic regression model is a predictive model. In the case here, it can be used to predict the likelihood of DUI recidivism for DUI offenders. Predictions of 8 real cases from were described below as examples of real-world applications of this research. For each model, there were 4 prediction cases showed below, with 2 cases for the non-repeat offenders and repeat offenders.

5.4.2.1. Model 1 Case Prediction Demonstrations

Case 1: This was a non-repeat offender, a male, aged 37, with mean BAC of 0.23, first DUI conviction on weekend, no prior 3-year low-risk citation, lived at Cass County that the income inequality rate was 4.27 and violent crime rate was 307.18 per 100,000 population. His predicted likelihood of DUI recidivism was calculated with parameters in the "Estimate" field in Tabe 5.18, started with the intercept and the order of variables listed in the case description. Below is the likelihood of DUI recidivism for this driver:

$$\frac{\exp(-3.1373 + 0.2274 * 1 + 0.1735 * 1 + 0.23 * 3.8481 + 0 + 0 - 0.1130 * 4.27 + 307.18 * 0.000634)}{1 + \exp(-3.1373 + 0.2274 * 1 + 0.1735 * 1 + 0.23 * 3.8481 + 0 + 0 - 0.1130 * 4.27 + 307.18 * 0.000634)} = 0.1054$$

Case 2: This was a non-repeat offender, a male, aged 20 (under the legal drinking age), with mean BAC of 0.16, first DUI conviction on weekday, had at least one prior 3-year low-risk citation, lived at McLean County that the income inequality rate was 4.12 and violent crime rate was 97.96 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-3.1373 + 0.2274 * 1 + 0.38 * 1 + 0.16 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 4.12 + 97.96 * 0.000634)}{1 + \exp(-3.1373 + 0.2274 * 1 + 0.38 * 1 + 0.16 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 4.12 + 97.96 * 0.000634))}$$

$$= 0.1217$$

Case 3: This was a repeat offender, a male, aged 28, with mean BAC of 0.29, first DUI conviction on weekday, had at least one prior 3-year low-risk citation, lived at Stark County that the income inequality rate was 3.98 and violent crime rate was 218.32 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-3.1373 + 0.2274 * 1 + 1901 * 1 + 0.29 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 3.98 + 218.32 * 0.000634)}{1 + \exp(-3.1373 + 0.2274 * 1 + 1901 * 1 + 0.29 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 3.98 + 218.32 * 0.000634)}$$

$$= 0.1718$$

Case 4: This was a repeat offender, a male, aged 20, with mean BAC of 0.26, first DUI conviction on weekday, had at least one prior 3-year low-risk citation, lived at Cass County that the income inequality rate was 4.27 and violent crime rate was 307.18 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-3.1373 + 0.2274 * 1 + 0.1906 * 1 + 0.29 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 4.27 + 307.18 * 0.000634)}{1 + \exp(-3.1373 + 0.2274 * 1 + 0.1906 * 1 + 0.29 * 3.8481 + 0.1585 * 1 + 0.1830 * 1 - 0.1130 * 4.27 + 307.18 * 0.000634)}$$

$$= 0.1718$$

In these demonstrations, predicted likelihood for the repeat offender was only 17%. There was a large discrepancy between the predicted likelihood and the true outcome. Based on discussions in literature review, there were a few possible reasons for this phenomenon: (1) The current information was not enough to provide solid prediction. (2) The offender's mindset and behavioral patterns has changed (3) The unselected variables can have aggregated effects that largely affect the model performance.

5.4.2.2. Model 2 Case Prediction Demonstration

Case 5 (the same driver as Case 1): This was a non-repeat offender, a male, complied with BAC test, first DUI conviction on weekend, no prior 3-year low-risk citation, lived at Cass County that the income inequality rate was 4.27 and violent crime rate was 307.18 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-2.9888 + 0.2261 * 1 + 0.8703 * 1 + 0 + 0 - 0.1032 * 4.27 + 0.000775 * 307.18)}{1 + \exp(-2.9888 + 0.2261 * 1 + 0.8703 * 1 + 0 + 0 - 0.1032 * 4.27 + 0.000775 * 307.18)} = 0.1096$$

Case 6: This was a non-repeat offender, a female, refused BAC test, first DUI conviction on weekend, no least one prior 3-year low-risk citation, lived at Stutsman County that the income inequality rate was 4.21 and violent crime rate was 212.32 per 100,000 population. Her predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-2.9888 + 0 + 0 + 0 + 0 - 0.1032 * 4.12 + 0.000775 * 212.32)}{1 + \exp(-2.9888 + 0 + 0 + 0 + 0 - 0.1032 * 4.12 + 0.000775 * 212.32)} = 0.037$$

Case 7 (same driver as Case 3): This was a repeat offender, a male, complied with BAC test, first DUI conviction on weekday, had at least one prior 3-year low-risk citation, lived at Stark County that the income inequality rate was 3.98 and violent crime rate was 218.32 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-2.9888 + 0.2261 * 1 + 0.8703 * 1 + 0.2138 * 1 + 0.1831 * 1 - 0.1032 * 4.27 + 0.000775 * 307.18)}{1 + \exp(-2.9888 + 0.2261 * 1 + 0.8703 * 1 + 0.2138 * 1 + 0.1831 * 1 - 0.1032 * 4.27 + 0.000775 * 307.18)} = 0.1547$$

Case 8: This was a repeat offender, a male, refused BAC test, first DUI conviction on weekday, had at least one prior 3-year low-risk citation, lived at Ward County that the income inequality rate was 3.67 and violent crime rate was 226.71 per 100,000 population. His predicted likelihood of DUI recidivism was calculated below:

$$\frac{\exp(-2.9888 + 0.2261 * 1 + 0 + 0.2138 * 1 + 0.1831 * 1 - 0.1032 * 3.67 + 0.000775 * 226.71)}{1 + \exp(-2.9888 + 0.2261 * 1 + 0 + 0.2138 * 1 + 0.1831 * 1 - 0.1032 * 3.67 + 0.000775 * 226.71)} = 0.0712$$

In Model 2, the discrepancy between predicted probability and the true outcome was even larger, especially in Case 7. It may be because this model was simpler than Model 1 and relied on less variable for the prediction. In practice, the court can set a cut-off point as the boundary to determine the judgement for offenders. For example, if the predicted likelihood of is larger than 0.13, then this offender is considered high risk for recidivism. More assessment should be done to find out the treatment or rehabilitation program that the offender need.

6. CONCLUSIONS

This study used a multi-model approach to determine the factors that affect the likelihood of DUI re-offense among drivers in North Dakota. It explored utilizing tree-based machine learning models to identify associations in a range of 107 factors and DUI recidivism among first-time DUI offenders. Three machine learning models were applied, including Decision Tree, Random Forest, and Gradient Boosting, to predict the likelihood of DUI recidivism. To improve predictions on imbalanced sample data, oversampling technique SMOTE - Tomek Links was applied to balance two classes in the sample data. To enhance interpretability, logistic regression analyses were performed to quantify the effects of top-ranked factors that the model selected with superior prediction power.

In this study, gradient boosting performed the best when dealing with an imbalanced dataset in which the minority group is of interest. For an imbalanced dataset, the oversampling technique SMOTE - Tomek Links improved prediction by nearly 10% on the F1 score, reducing the number of false positives and false negatives. Logistic regression was a great supplement to gradient boosting in terms of interpreting associations between factors and the outcome.

Results coalesced around two findings. First, male drivers with higher BAC values, younger age at first DUI citation, whose first DUI citation took place during the weekday, had at least one low-risk citation within three years before first DUI citation, and lived in counties with lower income inequality ratio and higher violent crime rate were more likely to commit a subsequent DUI offense. Second, male drivers who complied with a BAC test upon arrest, whose first DUI citation took place on a weekday, had at least one low-risk citation within three years before the first DUI citation, lived in a county with a lower income inequality ratio, and higher violent crime rate were more likely to commit a subsequent DUI offense.

The study also demonstrated a few cases to show how the logistic regression models can be used to predict the likelihood of recidivism for a particular offender. Suggestion was given on how to determine a high-risk offenders based on the prediction model. This practice can be used to assist court for judgement.

The study is limited to a single state, but the comparison of techniques and their shared findings suggest that a multitude and variety of approaches may be appropriate in future impaired driving prevention research. In addition, only first-time DUI offenders among licensed North Dakota drivers and the single administrative record system for feature variable. Future research may broaden the population to consider other states and drivers with multiple DUI offenses. Linkages to other administrative records may present opportunity to consider variables such as court record, criminal records and social relationships that have proven valuable in other DUI prediction research. In addition, assessment tools can be used to identify offenders with specific mental health disorders, so further and more particular treatments will help to improve offenders' chances of rehabilitation.

REFERENCES

Batista, G. E. A. P. A., Bazzan, A. L. C., and Monard, M. A. (2003). Balancing Training Data for Automated Annotation of Keywords: Case Study. Proceedings of the Second Brazilian Workshop on Bioinformatics, pp. 35–43.

Beitel, G. A., Sharp, M. C., & Glauz, W. D. (2000). Probability of arrest while driving under the influence of alcohol. Injury Prevention, 6(2), 158-161.

Berge, C. M. (2019). The Effects of the DUI 24/7 Program in Cass County, North Dakota (Doctoral dissertation, North Dakota State University).

Bergen, G., R.A. Shults, and R.A. Rudd. 2011. "Vital Signs: Alcohol-Impaired Driving Among Adults – United States, 2010." Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report 60(39):1351-1356.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. Criminology & Pub. Pol'y, 12, 513.

Bishop, Nicholas J. 2011. Predicting rapid DUI recidivism using the Driver Risk Inventory in a sample of Floridian DUI offenders. Drug and Alcohol Dependence, 118: 423– 429

Boehmke, B., & Greenwell, B. (2020). Hands-on machine learning with R. Chapman and Hall/CRC. Available at https://bradleyboehmke.github.io/HOML/. Accessed on 09/27/2022.

Bouchard, S. M., Brown, T. G., & Nadeau, L. (2012). Decision-making capacities and affective reward anticipation in DWI recidivists compared to non-offenders: A preliminary study. Accident Analysis & Prevention, 45, 580-587.

Breiman, L., Bagging Predictors, Machine Learning, 24(2), pp.123-140, 1996.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brennan, T., & Oliver, W. L. (2013). Emergence of machine learning techniques in criminology: implications of complexity in our data and in research questions. Criminology & Pub. Pol'y, 12, 551.

Cavaiola, A. A., Strohmetz, D. B., & Abreo, S. D. (2007). Characteristics of DUI recidivists: A 12-year follow-up study of first time DUI offenders. *Addictive behaviors*, *32*(4), 855-861.

C'de Baca, J., Miller, W. R., & Lapham, S. (2001). A multiple risk factor approach for predicting DWI recidivism. Journal of substance abuse treatment, 21(4), 207-215.

Centers for Disease Control and Prevention (CDC) (n.d.). Impaired Driving: Get the Facts – BAC Effects. Accessed on 08/05/2020, retrieved at https://www.cdc.gov/motorvehiclesafety/impaired_driving/impaired-drv_factsheet.html

Chang, I., Lapham, S. C., & Barton, K. J. (1996). Drinking environment and sociodemographic factors among DWI offenders. Journal of Studies on Alcohol, 57(6), 659-669.

Chaudhary, N. K., Tison, J., McCartt, A. T., & Fields, M. (2011). Patterns of recidivism related to case dispositions of alcohol-impaired driving offenses. Traffic injury prevention, 12(3), 210-216.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, H., & Chen, L. (2017). Support vector machine classification of drunk driving behaviour. International journal of environmental research and public health, 14(1), 108.

Collins-Thompson, K. (n.d). Cross-validation [Coursera Lecture]. In Collins-Thompson, K., Applied Machine Learning in Python. Coursera. https://www.coursera.org/learn/python-machine-learning. Accessed on April 28, 2021.

Crewson, P.(n.d.). Applied Statistics Desktop Reference. AcaStat Software. Available at https://www.researchgate.net/profile/Philip-Crewson/publication/297394168_Applied_Statistics/links/56dec13a08aedf2bf0c9c63c/Applied-Statistics.pdf. Accessed on 09/23/2022.

DeMichele, M., & Payne, B. (2013). If I had a hammer, I would not use it to control drunk driving: Using predictive tools to respond to persistent drunk driving. Criminology & Pub. Pol'y, 12, 213.

Dickson, M. F., Wasarhaley, N. E., & Webster, J. M. (2013). A comparison of first-time and repeat rural DUI offenders. Journal of Offender Rehabilitation, 52(6), 421-437.

Dugosh, K. L., Festinger, D. S., & Marlowe, D. B. (2013). Moving beyond BAC in DUI: Identifying who is at risk of recidivating. *Criminology & public policy*, *12*(2), 181-193.

Duwe, G., & Kim, K. (2017). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. Criminal Justice Policy Review, 28(6), 570-600.

Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL

Esposito, D., & Esposito, F. (2020). Introducing Machine Learning. Microsoft Press.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society for Artificial Intelligence, 14(771-780), 1612.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. Statistics in medicine, 22(9), 1365-1381.

Ghasemi, M., Anvari, D., Atapour, M., Stephen Wormith, J., Stockdale, K. C., & Spiteri, R. J. (2021). The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidivism. Criminal Justice and Behavior, 48(4), 518-538.

Guestrin, C. (2015). Boosting [Coursera Lecture]. In Fox, E., & Guestrin, C, Machine Learning: Classification. Coursera. https://www.coursera.org/learn/ml-classification. Accessed on April 20, 2021.

Goodwin, A., Thomas, L., Kirley, B., Hall, W., O'Brien, N., & Hill, K. (2015, November). Countermeasures that work: A highway safety countermeasure guide for State highway safety offices, Eighth edition. (Report No. DOT HS 812 202). Washington, DC: National Highway Traffic Safety Administration.

Greene, K. M., Murphy, S. T., & Rossheim, M. E. (2018). Context and culture: Reasons young adults drink and drive in rural America. *Accident Analysis & Prevention*, *121*, 194-201.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications, 73, 220-239.

Henke, R., Nelson, M., Mongeon, K., Doan, A.,Harsche, L., Thurn, C., Wilson, S., Malafa, L. and Heinert, L.(2017). 2018 North Dakota Highway Safety Plan. North Dakota Department of Transportation, Bismarck, North Dakota. Accessed on 08/05/2020, retrieved at https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/north_dakota_fy2018_hsp.pdf

Hosmer, Lemeshow, S., & May, S. (2008). Applied survival analysis: regression modeling of time-to-event data (2nd ed.). Wiley-Interscience.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Hubicka, B., Laurell, H., & Bergman, H. (2008). Criminal and alcohol problems among Swedish drunk drivers—Predictors of DUI relapse. International journal of law and psychiatry, 31(6), 471-478.

Hubicka, B., Källmén, H., Hiltunen, A., & Bergman, H. (2010). Personality traits and mental health of severe drunk drivers in Sweden. Social psychiatry and psychiatric epidemiology, 45(7), 723-731.

Impinen, A., Rahkonen, O., Karjalainen, K., Lintonen, T., Lillsunde, P., & Ostamo, A. (2009). Substance use as a predictor of driving under the influence (DUI) rearrests. A 15-year retrospective study. Traffic injury prevention, 10(3), 220-226.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR), 52(4), 1-36.

Kontschieder, P., Bulo, S. R., Bischof, H., & Pelillo, M. (2011). Structured class-labels in random forests for semantic image labelling. In 2011 International Conference on Computer Vision (pp. 2190-2197). IEEE.

Kubas, Andrew, Kimberly Vachal, and Donald Malchose. The Effects of Regular Alcohol Monitoring on North Dakota Impaired Drivers, DP-300. North Dakota State University, Fargo: Upper Great Plains Transportation Institute, 2018.

Kubas, A., and Vachal, K., (2019). The 24/7 Sobriety Program's Effects on Impaired Drivers, Report No. DP-304. Upper Great Plains Transportation Institute, North Dakota State University, Fargo, ND.

Lau-Barraco, C., Braitman, A. L., Linden-Carmichael, A. N., & Stamates, A. L. (2016). Differences in weekday versus weekend drinking among nonstudent emerging adults. Experimental and Clinical Psychopharmacology, 24(2), 100–109. https://doi.org/10.1037/pha0000068 https://psycnet.apa.org/record/2016-09358-001

Lee, E. T., & Wang, J. (2003). Statistical methods for survival data analysis (Vol. 476). John Wiley & Sons.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 18(1), 559-563.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. Journal of Quantitative Criminology, 27(4), 547-573.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. R journal, 6(1).

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., & Huang, Y. (2020). A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. Journal of advanced transportation, 2020.

MacLeod, K. E., Karriker-Jaffe, K. J., Satariano, W. A., Kelley-Baker, T., Lacey, J. H., & Ragland, D. R. (2017). Drinking and driving and perceptions of arrest risk among California drivers: Relationships with DUI arrests in their city of residence. Traffic injury prevention, 18(6), 566-572.

Marowitz, L. A. (1998). Predicting DUI recidivism: Blood alcohol concentration and driver record factors. Accident Analysis & Prevention, 30(4), 545-554.

McMillen, D. L., Adams, M. S., Wells-Parker, E., Pang, M. G., & Anderson, B. J. (1992). Personality traits and behaviors of alcohol-impaired drivers: A comparison of first and multiple offenders. Addictive behaviors, 17(5), 407-414.

McMillen, D. L., Pang, M. G., Wells-Parker, E., & Anderson, B. J. (1992). Alcohol, personality traits, and high risk driving: A comparison of young, drinking driver groups. Addictive behaviors, 17(6), 525-532.

Mitchell, T. M. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 45(37), 870-877.

Møller, M., Haustein, S., & Prato, C. G. (2015). Profiling drunk driving recidivists in Denmark. Close

National Center for Education Statistics (2016). Institutional characteristics files 1993–2015. Retrieved online from: https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx.Accident Analysis & Prevention, 83, 125-131.

National Highway Traffic Safety Administration (NHTSA) (2008). Traffic safety facts: Repeat intoxicated driver laws. Report No. DOT HS 810 879. Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration (NHTSA), (2019). Traffic Safety Facts - Alcohol-Impaired Driving for 2018. Report No. DOT HS 812 864. Washington, DC: U.S. Department of Transportation.

Nelson, S. E., LaRaja, A., Juviler, J., & Williams, P. M. (2021). Evaluating the Computerized Assessment and Referral System (CARS) screener: sensitivity and specificity as a screening tool for mental health disorders among DUI offenders. Substance Use & Misuse, 56(12), 1785-1796.

Nochajski, T. H., & Wieczorek, W. F. (2000). Driver characteristics as a function of DWI history. In Proceedings International Council on Alcohol, Drugs and Traffic Safety Conference (Vol. 2000). International Council on Alcohol, Drugs and Traffic Safety.

Nochajski, T. H., & Stasiewicz, P. R. (2001). Under-reporting by DWI offenders: Implications for motivational interviewing. Alcoholism: Clinical and Experimental Research, 25(5), 277.

Nochajski, T. H., & Stasiewicz, P. R. (2006). Relapse to driving under the influence (DUI): A review. Clinical psychology review, 26(2), 179-195.

North Dakota Department of Transportation (NDDOT) (2018). 2018 North Dakota Crash summary. North Dakota Department of Transportation, Bismarck, North Dakota. Accessed on 08/05/2020, retrieved at https://visionzero.nd.gov/uploads/24/NDDOT_2018_Crash_Summary_hires_nobleed1.pdf

North Dakota Department of Transportation (NDDOT) (n.d.). North Dakota Vision Zero Plan - Strategic Highway Safety Plan Update 2018-2023. North Dakota Department of

Transportation, Bismarck, North Dakota. Accessed on 08/05/2020, retrieved at https://www.dot.nd.gov/divisions/safety/docs/FINAL_NDDOT_SHSP.pdf

Ozkan, T. (2017). Predicting recidivism through machine learning (Doctoral dissertation).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Portman, M., Penttilä, A., Haukka, J., Eriksson, P., Alho, H., & Kuoppasalmi, K. (2010). Predicting DUI recidivism of male drunken driving: a prospective study of the impact of alcohol markers and previous drunken driving. Drug and alcohol dependence, 106(2-3), 186-192.

Reynolds, J. R., Kunce, J. T., & Cope, C. S. (1991). Personality differences of first-time and repeat offenders arrested for driving while intoxicated. *Journal of Counseling Psychology*, *38*(3), 289.

Richard, C. M., Magee, K., Bacon-Abdelmoteleb, P., & Brown, J. L. (2018, April). Countermeasures that work: A highway safety countermeasure guide for State Highway Safety Offices, Ninth edition (Report No. DOT HS 812 478). Washington, DC: National Highway Traffic Safety Administration.

Robertson, A., Gardner, S., Walker, C. S., & Tatch, A. (2016). DUI recidivism by intervention adherence: A multiple risk factor approach. *The American journal of drug and alcohol abuse*, *42*(5), 597-605.

Roma, P., Mazza, C., Ferracuti, G., Cinti, M. E., Ferracuti, S., & Burla, F. (2019). Drinking and driving relapse: Data from BAC and MMPI-2. *PLoS One*, *14*(1), e0209116.

Rookey, B. D. (2012). Drunk driving in the United States: An examination of informal and formal factors to explain variation in DUI enforcement across US counties. W. Criminology Rev., 13, 37.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Computational intelligence magazine, 13(4), 59-76.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine learning, 37(3), 297-336.

Schell, T. L., Chan, K. S., & Morral, A. R. (2006). Predicting DUI recidivism: Personality, attitudinal, and behavioral risk factors. Drug and alcohol Dependence, 82(1), 33-40.

Spelmen, V. S., & Porkodi, R. (2018, March). A review on handling imbalanced data. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-11). IEEE.

Stringer, R. J. (2018). Exploring traffic safety culture and drunk driving: An examination of the community and DUI related fatal crashes in the US (1993–2015). Transportation research part F: traffic psychology and behaviour, 56, 371-380.

Studer, J., Baggio, S., Mohler-Kuo, M., Dermota, P., Daeppen, J.-B., & Gmel, G. (2014). Differential association of drinking motives with alcohol use on weekdays and weekends. Psychology of Addictive Behaviors, 28(3), 651–658. https://psycnet.apa.org/record/2014-33489-001

Tollenaar, N., & van der Heijden, P. G. (2013). Which method predicts recidivism best: a comparison of statistical, machine learning and data mining predictive models. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(2), 565-584.

Tomek, I. (1976). Two Modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, vol. 6, no. 11, pp. 769–772.

University of Wisconsin Population Health Institute (UWPHI), 2017. North Dakota Rankings Data 2017. County Health Rankings. Available at https://www.countyhealthrankings.org/app/north-dakota/2022/downloads . Accessed on 09/29/2022.

Vachal, K., L. Benson, and A. Kubas. 2019. "North Dakota Statewide Traffic Safety Survey, 2019: Traffic Safety Performance Measures for State and Federal Agencies. In press. Fargo, ND: Upper Great Plains Transportation Institute, North Dakota State University.

Vaez M, Laflamme L. (2005). Impaired Driving and Motor Vehicle Crashes among Swedish Youth: An Investigation into Drivers' Sociodemographic Characteristics. Accid. Anal. Prev., Vol. 37, pp. 605– 611.

Vision Zero. Fatal Crash Stat Board. Accessed on 08/05/2020, retrieved at https://visionzero.nd.gov/uploads/28/StatusBoardUpdateasof842020.pdf

Voas, R. B., & Lacey, J. H. (1990). Drunk driving enforcement, adjudication, and sanctions in the United States. Drinking And Driving: Advances in Research and Prevention Edited By R Jean Wilson, Robert E Mann.

Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters, 136, 190-197.

Weisheit, R. A. (2020). Preventing rural alcohol-and drug-related crime. In Rural Crime Prevention (pp. 97-108). Routledge.

Wickens, C. M., Flam-Zalcman, R., Stoduto, G., Docherty, C., Thomas, R. K., Watson, T. M., ... & Mann, R. E. (2018). Multiple “Lower BAC” offenders: Characteristics and response to remedial interventions. *Accident Analysis & Prevention*, *115*, 110-117.

Wieczorek, W. F., & Nochajski, T. H. (2005). Characteristics of Persistent Drinking Drivers: Comparisons of First, Second, and Multiple Offenders. In T-2000 International Conference on Alcohol, Drugs and Traffic Safety, Stockhol, Sweden; This chapter is a greatly expanded and updated version of a paper presented by the authors at the aforementioned conference. Nova Science Publishers.

Wijenayake, S., Graham, T., & Christen, P. (2018, June). A decision tree approach to predicting recidivism in domestic violence. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 3-15). Springer, Cham.

Wilson, D. K., & Grube, J. (1994). Role of psychosocial factors in obtaining self-reports of alcohol use in a DUI population. Psychology of Addictive Behaviors, 8(3), 139.

Zheng, Z., Lu, P., & Tolliver, D. (2016). Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis. Transportation Research Record, 2545(1), 115-122.

Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. Journal of safety research, 65, 115-124.

Zhou, X., Lu, P., Zheng, Z., Tolliver, D., & Keramati, A. (2020). Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree. Reliability Engineering & System Safety, 106931.