**Analysis of Cohort Default Rates (CDR) in U.S. Post-Secondary Education Institutions**

A Thesis

Submitted to Graduate Faculty of the

National University, School of Professional Studies

In partial fulfillment of the requirements for the degree of

Master of Science in Data Science

Prepared By:

Prince Gyamfi

Marketne O. Noel

Dillon Orr

National University

May 2020

**MASTERS THESIS APPROVAL FORM**

We certify that we have read the project of Prince Gyamfi, Marketne O. Noel, and Dillon Orr entitled ANALYSIS OF COHORT DEFAULT RATES (CDR) IN U.S. POST-SECONDARY EDUCATION INSTITUTIONS and that, in our opinion, it is satisfactory in scope and quality as the thesis for the degree of Master of Science in Data Analytics at National University.

Approved:

John Vivian, PhD Capstone Project Advisor Date

Bioinformatics Scientist

Atreca, Inc.

Ebrahim Tarshizi, PhD, MBA Capstone Project Instructor and Advisor Date

Adjunct Professor, School of Professional Studies

National University

# Abstract

This research measures the impact that public, non-profit, and proprietary (for-profit) postsecondary institutions have on cohort default rate (CDR). The information is used to show the link between the concentration of non-white populations in counties and the presence of proprietary institutions. To statistically illustrate these connections, Machine Learning models are used for regression tasks and classification instances. For the regression task, the continuous variable CDR2016 is set as the dependent variable, and for classification instances, the dependent variable is set as proprietary. In addition to predicting school type, schools were distributed into high CDR and low CDR categories to determine if other contributing factors link an institute to either category. The researcher used Linear Regression, Support Vector Machine, naïve Bayes, K-Nearest Neighbors, and Logistic Regression to create predictive models. In conclusion, it is found that CDR, school type, and level of CDR can be predicted successfully by the Linear Regression and Logistic Regression models. The key finding from the research is that there is a more significant number of Proprietary institutions in counties that have less than the nation’s median white percentage.

*Keywords:* cohort default rate, financial aid, post-secondary education, machine learning

Table of Contents

[Abstract iii](#_Toc39254765)

[List of Tables ix](#_Toc39254766)

[List of Figures x](#_Toc39254767)

[Chapter 1: Introduction 1](#_Toc39254768)

[1.1 Background 1](#_Toc39254769)

[1.2 What Is Cohort Default Rate? 1](#_Toc39254770)

[1.3 Calculating Cohort Default Rate 2](#_Toc39254771)

[1.4 Cohort Default Rate Criteria for Sanctions 2](#_Toc39254772)

[1.5 Cohort Default Rate Benefits 3](#_Toc39254773)

[1.6 Problem Statement 3](#_Toc39254774)

[1.7 Research Hypothesis 4](#_Toc39254775)

[1.10 Definition of Terms 6](#_Toc39254776)

[1.11 Summary 7](#_Toc39254777)

[Chapter 2: Literature Review 9](#_Toc39254778)

[2.1 Official CDR and Trends 9](#_Toc39254779)

[2.2 Pre-College Contributions to CDR 11](#_Toc39254780)

[2.3 In-College Experience Attributed to CDR 12](#_Toc39254781)

[2.3.1 School Dropout 12](#_Toc39254782)

[2.3.2 College Major 12](#_Toc39254783)

[2.3.3 Number of Years in School 13](#_Toc39254784)

[2.4 Post-College Attributes to CDR 14](#_Toc39254785)

[2.4.1 Unemployment 14](#_Toc39254786)

[2.4.2 Low Wages 14](#_Toc39254787)

[2.4.3 Disposable Income and Student Attitude 15](#_Toc39254788)

[2.4.4 Personal and Family Experience 16](#_Toc39254789)

[2.4.5 Knowledge of Repayment Obligations 17](#_Toc39254790)

[2.5 Cohort Default Rate by Institution Sectors and Ethnicity 17](#_Toc39254791)

[2.5.1 School Sectors or Type and CDR 17](#_Toc39254792)

[2.5.2 Ethnicity and CDR 19](#_Toc39254793)

[2.6 Other Causes of High CDR 20](#_Toc39254794)

[2.7 Consequences of Defaulting on Student Borrowers 21](#_Toc39254795)

[2.8 Recovery from Default 21](#_Toc39254796)

[2.9 Impact of Student Debt on Retirement and Longevity 22](#_Toc39254797)

[2.10 Effectiveness of Federal Student Loan Accountability Regulations 23](#_Toc39254798)

[2.11 What Should Regulations Look Like? 24](#_Toc39254799)

[2.12 Policy Recommendations for Facilitating Repayment 24](#_Toc39254800)

[2.13 Student Loan Tips 25](#_Toc39254801)

[Chapter 3: Methodology 27](#_Toc39254802)

[3.1 Instruments 27](#_Toc39254803)

[3.2 Procedure 28](#_Toc39254804)

[3.3 Data Analysis 30](#_Toc39254805)

[3.4 Introduction to Variables 32](#_Toc39254806)

[3.5 Data Limitations 34](#_Toc39254807)

[3.6 Continuous Variables 35](#_Toc39254808)

[3.6.1 Objective 35](#_Toc39254809)

[3.6.2 Numerical Variable Identification 36](#_Toc39254810)

[3.6.3 Analyze Basic matrices 38](#_Toc39254811)

[3.6.4 Non-Graphical Analysis 38](#_Toc39254812)

[3.6.5 Graphical Analysis 40](#_Toc39254813)

[3.6.6 Variable Transformation 41](#_Toc39254814)

[3.6.7 Outlier Handling 49](#_Toc39254815)

[3.6.8 Correlation Analysis 51](#_Toc39254816)

[3.7 Categorical Variables 53](#_Toc39254817)

[3.7.1 One-Hot Encoding (Dummy Variables) 53](#_Toc39254818)

[3.8 Feature Selection 56](#_Toc39254819)

[3.8.1 Approach 57](#_Toc39254820)

[3.8.2 Choice of Method 61](#_Toc39254821)

[3.9 Data Splitting 62](#_Toc39254822)

[3.9.1 Train/Test/Validation Split 62](#_Toc39254823)

[3.9.2 K-Folds Cross-Validation 62](#_Toc39254824)

[3.10 Modeling 63](#_Toc39254825)

[3.10.1 Linear Regression 63](#_Toc39254826)

[3.10.2 Logistic Regression 63](#_Toc39254827)

[3.10.3 K-Nearest Neighbors (KNN) 64](#_Toc39254828)

[3.10.4 Naïve Bayes 64](#_Toc39254829)

[3.10.5 Decision Trees 65](#_Toc39254830)

[3.10.6 Random Forest 66](#_Toc39254831)

[3.10.7 Support Vector Machines (SVM) 66](#_Toc39254832)

[3.10.8 XG Boost 67](#_Toc39254833)

[Chapter 4: Results 68](#_Toc39254834)

[4.1 Regression Models 68](#_Toc39254835)

[4.2 Classification Models 75](#_Toc39254836)

[4.2.1 Introduction to Classification 75](#_Toc39254837)

[4.2.2 Classification Algorithms Chosen 76](#_Toc39254838)

[4.2.3 Classification Performance Measures 76](#_Toc39254839)

[4.2.4 Interpreting Results of Classification Models 78](#_Toc39254840)

[4.3 CDR Dashboard 80](#_Toc39254841)

[4.4 Comparison of Total Proprietary Institutions in High Percentage White Counties 82](#_Toc39254842)

[Chapter 5: Conclusion 84](#_Toc39254843)

[5.1 Overview of the Research 84](#_Toc39254844)

[5.2. Interpretation of the Results 84](#_Toc39254845)

[5.2.1 Cohort Default Rates (CDR 2014-CDR 2015) 85](#_Toc39254846)

[5.2.2 Household Income and Marital Status 85](#_Toc39254847)

[5.2.3 Institution and Program Length 86](#_Toc39254848)

[5.2.4 Ethnicity 86](#_Toc39254849)

[5.3 Findings 87](#_Toc39254850)

[5.4 Research Limitations 88](#_Toc39254851)

[5.5 Conclusions Drawn 89](#_Toc39254852)

[5.6 Recommendations 90](#_Toc39254853)

[5.6.1 Policy Reforms 90](#_Toc39254854)

[5.6.2 Future Research 91](#_Toc39254855)

[Appendix A 92](#_Toc39254856)

[Appendix B 94](#_Toc39254857)

[References 95](#_Toc39254858)

# List of Tables

[Table 1 39](#_Toc39243634)

[Table 2 44](#_Toc39243635)

[Table 3 45](#_Toc39243636)

[Table 4 46](#_Toc39243637)

[Table 5 56](#_Toc39243638)

[Table 6 58](#_Toc39243639)

[Table 7 62](#_Toc39243640)

[Table 8 63](#_Toc39243641)

[Table 9 63](#_Toc39243642)

[Table 10 76](#_Toc39243643)

[Table 11 77](#_Toc39243644)

[Table 12 81](#_Toc39243645)

[Table 13 82](#_Toc39243646)

[Table 14 84](#_Toc39243647)

[Table 15 85](#_Toc39243648)

[Table 16 85](#_Toc39243649)

[Table 17 88](#_Toc39243650)

[Table 18 89](#_Toc39243651)

[Table 19 90](#_Toc39243652)

[Table 20 92](#_Toc39243653)

# List of Figures

[Figure 1 43](#_Toc39243654)

[Figure 2 49](#_Toc39243655)

[Figure 3 53](#_Toc39243656)

[Figure 4 57](#_Toc39243657)

[Figure 5 59](#_Toc39243658)

[Figure 6 66](#_Toc39243659)

[Figure 7 67](#_Toc39243660)

[Figure 8 68](#_Toc39243661)

[Figure 9 73](#_Toc39243662)

[Figure 10 74](#_Toc39243663)

[Figure 11 78](#_Toc39243664)

[Figure 12 79](#_Toc39243665)

[Figure 13 80](#_Toc39243666)

[Figure 14 81](#_Toc39243667)

[Figure 15 86](#_Toc39243668)

[Figure 16 87](#_Toc39243669)

[Figure 17 88](#_Toc39243670)

[Figure 18 90](#_Toc39243671)

# Chapter 1: Introduction

## 1.1 Background

In the 1980s, the United States experienced a rise in Post-Secondary Institutes in cities with high numbers of minority and impoverished residents (OECD, 2008). It was believed that these institutes encouraged students to enroll in programs they were academically unqualified for and to take loans that they would most likely end in default. After a comprehensive review of different measures, Secretary of Education William J. Bennett responded by proposing the Cohort Default Rate (CDR) metric and included sanctions on the policy. This was adopted in the Omnibus Budget Reconciliation Act of 1990 and established a two-year CDR as a formal accountability mechanism. Under the new Act (Title IV of the amended Higher Education Act of 1965), any institution whose cohort default rate exceeded 35 percent in the fiscal years 19991 or 1992 and 30 percent for any other fiscal years after 1992 would lose eligibility to participate in federal aid programs.

## 1.2 What Is Cohort Default Rate?

CDR is the metric used to identify Post-Secondary Institutes whose borrowers have defaulted or met “other specified conditions” on their student loans within three years of entering repayment. The statement “other specified conditions” refers to incidents in which an entity besides the borrower made payments to prevent the loan from going into default. These incidents are classified as the borrower defaulting. The Department of Education calculates the CDR with the Federal Stafford, Direct Stafford, and Ford loans, and does not include PLUS, Perkins, or FISL loans. A former college student is considered to have defaulted on the federal loan if, after 360 days of leaving college, he or she fails to enter a payment plan. It is important to note that the CDR is not calculated based on the number of loans entering repayment, but rather the number of borrowers entering repayment.

## 1.3 Calculating Cohort Default Rate

There are two methods by which it is computed, and they are highly dependent on the number of an institution’s borrowers who enter repayment in the fiscal year. The first method is known as the Average Rate Formula (ARF) (1), known as the unofficial CDR. It is used to calculate CDRs of institutes where less than 29 borrowers are in repayment during the fiscal year. In comparison, the Non-Average Rate Formula (NARF) (2), the official CDR, is calculated for institutes where 30 or more borrowers are in repayment during the fiscal year. There are two noteworthy differences between the two formulas. One, the ARF computes the CDR by applying the two previous fiscal years to the denominator, as shown in formula 1. In contrast, the NARF uses only the current fiscal year in its computation of a CDR, as shown in formula 2. Secondly, the NARF is never used as a sanctioning criterion to penalize institutes.

(1)

(2)

## 1.4 Cohort Default Rate Criteria for Sanctions

An institute can be sanctioned if it fulfills any of the following two requirements: one, if it maintains a CDR of more than 30% but less than 40% for three years, then the offending institute will be sanctioned, two, if an institute maintains a CDR of 40% or higher for one year, then the offending institute will be sanctioned. Institutes that are sanctioned effectively lose the right to use federal assistance such as Direct Loans and partake in the Pell Grant program. These sanctions are an effort of the U.S. Department of Education to save taxpayers money and help maintain an affordable tuition rate.

## 1.5 Cohort Default Rate Benefits

Though it may seem that CDR is a metric that penalizes institutes, it is a metric used to encourage institutes to improve the quality of their admissions and educational practices. There are two benefits to institutes that can maintain a low CDR. Firstly, if an institute's most recent CDR is below 5% and it maintains a study abroad program then, students’ loans may be distributed in a single installment without there being a delay in distribution to students studying abroad. Secondly, if an institute CDR is below 15% for the past three recent years. That institute will be allowed to distribute its loans with a single installment and must distribute it within 30 days for first-year and first-time undergraduate students.

CDR data is significant for students in terms of selecting the college to send applications too. A student will have all the necessary information about an institution in terms of the different financial aid he or she is likely to get to assist with the tuition fees and any other program needs. With this information, the students are able to make better decisions and choices of institutes to join.

## 1.6 Problem Statement

Though CDR has been decreasing on an annual basis (U.S. Department of Education, 2019), the underlying reasons for which the Secretary of Education adopted the CDR still persists to date. Students have continued to default on their loans with the vast majority of the U.S. cities with high concentrations of minority populations registering the highest default rates. Additionally, a majority of proprietary institutions and Historically Black Colleges and Universities within the U.S. continued to register higher default rates (Mitchell & Fuller, 2019).

## 1.7 Research Hypothesis

The research hypothesis of the study (H1) is that Post-Secondary Institutions with high cohort default rates are located more frequently in U.S. cities with large minority populations. Based on this hypothesis, the team suggests that those cities that have a higher concentration of African Americans, American Indians, and Alaska Natives, and the Asians are likely to register higher default rates than those cities with concentrations of white populations.  **1.8 Objective**

The research team aims to show that proprietary institutions, on average, have a significantly higher CDR as compared to their public or private, non-profit counterparts. Another goal is to show that minority (non-white borrowers) are more susceptible to defaulting on loans compared to white students. Overall, roughly half of Black borrowers default within 12 years of enrolling in college. Studies have shown that “entering a for-profit (institute) is associated with a 10-point higher rate of default…” for Black students (Fain, 2018).

After demonstrating that for-profit colleges have a significantly higher default rate among minority students, the research team will attempt to show that these for-profit colleges are more often located in minority rich neighborhoods. In 2013, Johns Hopkins University studied youth in Baltimore neighborhoods, finding that 53% of participants chose to sign up for occupational certifications at for-profit institutions. Only 31% of those students completed their program (Johns Hopkins University, 2016).

The end goal is to implement an online, cloud-based dashboard that can be accessed free of charge by parents, schools, and, most importantly, future students. This dashboard will be able to show the distribution of CDR by college type, as well as geographical locations of all schools in the U.S. Users will be able to view information about their selections’ CDR, in-regards to the school type. Implementing a predictive model will allow the forecast of Default Rate by institutes, allowing users to see which colleges may be at risk of causing a default. Increasing the amount of information that is accessible by the youth and their stakeholders is key to improving understanding of the differences between types of colleges. Registering for college can be an exciting time, and the hope is that this research will help students make this decision with confidence.  **1.9 Limitations of the Study**

CDR is a measure of whether a college should be eligible to receive federal student aid. Though, it has been called a “low bar for performance” (Sen. Lamar Alexander, 2019) because it does not include the “one-third of borrowers who are not yet in default but do not make payments on time” (Community College Daily, 2019). It implies that there are many more students who are struggling to make payments on their loans, and simply have not yet been labeled as having defaulted. Thus, the target population, who are students who have been saddled with debt they cannot repay, is not fully defined by CDR. While only 12% of loans extended to students are actively in a default status, nearly 26% are in forbearance or deferment, and 9% of borrowers are delinquent (late on making a payment). In total, roughly 47% of students are struggling to make their loan payments (Office of Federal Student Aid, 2019).

Many factors can influence a student’s inability to pay back their loans outside of the quality of an educational institute. Parent socioeconomic status has shown to have an impact on the amount of debt that a student incurs (Houle, 2013). Household income is a variable that will be controlled in the model.

Due to different policy regulations and privacy rights, this study is not able to include the student level demographics or data. This means that the study will mainly use the state-level demographics of different institutes to conduct the analysis and make assumptions in certain areas if the need arises. This may lead to predictions that may be questionable in terms of accuracy and presentation of facts.

Natural disasters and economic crises in certain areas can create severe consequences for students. “Between 2000 and 2012, worldwide natural disasters affected nearly three billion people and cost $1.7 trillion in damages” (McWhirter, n.d.). From 2008 to 2017, national student debt more than doubled (Federal Reserve Bank of New York, 2017). Many experts believe this was partially due to the “Great Recession.” At this point, the role that Natural Disasters and Economic downturns play in the increase in student loan debt between groups of borrowers has not been accounted.

Perhaps most shockingly, there is a growing culture around intentional student loan default. Some students believe that choosing to default on their loans will urge nationwide social and political changes (Farrington, 2018). The data does not provide information on which of these borrowers are choosing to default on their loans. It is not clear on the exact manner of how borrowers are engaging this practice in the United States.

## 1.10 Definition of Terms

Cohort Default Rate (CDR): CDR can be defined as the overall percentage of the portfolio of

lenders who have student borrowers entering repayment on Federal Family Education Loan Program (FFELP) loans in a given fiscal year and then default on those loans in the same

fiscal year or the following fiscal year.

Minority Population: This is a term that is used in the study to refer to the African Americans,

Native Indians, Asians, Hispanic, and other populations in the United States other than Whites.

Default: Used to define the failure of a student to pay their loans after graduation from an

institution or school (usually 360-days after graduation).

Average Rate Formula (ARF): used to calculate the Cohort Default Rate for a school with 29 or

fewer borrowers entering repayment during a cohort fiscal year.

Non-Average Rate Formula (NARF): used to calculate the Cohort Default Rate for a school with

30 or more borrowers entering repayment during a cohort fiscal year

## 1.11 Summary

This study aims to show that proprietary institutions, on average have a higher Cohort Default Rate (CDR) than public or private (non-profit) institutions. Additionally, the team wishes to illustrate that proprietary institutions are focused more often in U.S. counties that have high minority populations. Machine learning models will be used to predict CDR and school type in order to illustrate the relationship between institution-related variables and default rate. Microsoft Excel, Python, and Tableau will be the primary instruments used in the entire process.

Significant research has been performed on the relationship between CDR and student demographic information. Most often, studies are focused on one county or city area, due to the fact that many studies are funded by these areas’ local governments. While there has been work done to help students and institutions avoid high default rates, there is significant information to be learned by completing a nation-wide analysis of CDR on a county-by-county basis.

# Chapter 2: Literature Review

## 2.1 Official CDR and Trends

The Cohort Default Rate (CDR) is the percentage of borrowers who are placed into or enrolled in any institute repayment system during a Fiscal Year (FY). If they default on their loans by the end of the following year, they are in default. In order to calculate the CDR for FY 2018, one needs the total number of student borrowers who entered repayment in 2018 and failed to repay by the end of the FY 2019 divided by the total number of students who entered the program in FY 2018. Based on the three-year CDR result, which is released by the United States Department of Education, the Federal Student Loan CDR is estimated to be around 10.1%, which represents a total decrease from the previous year by 0.7% (U.S. Department of Education, 2019). Many pieces of literature have analyzed the trends in CDR, in an attempt to understand how to reduce the likelihood of student default nationwide. However, to eliminate students' risk of default is a hard task, and the United States is still far from realizing its objective (Fuller, 2019).

While overall CDR is decreasing, the rate of default has been rising each year for African American and Hispanic borrowers since the FY 2000 (“Black Students’ Default,” 2019). While African American and Hispanic borrowers experience a rise in their CDR trend, the national CDR has a negative slope in its growth. CDR by the school sector in the FY 2000 to FY 2002 declined from 4.8 % to 4% for public four-year institutes, while public two-year institutes experienced a decline from 9.2% to 8.5% (“Black Students’ Default,” 2019). For the same fiscal years, it is shown that not-for-profit private four-year institutes had a decline from 3.8% to 3.1% CDR, and private for-profit or proprietary institutes had declined from 9.4% to 8.7%.

By taking a closer look into the 1996 entry cohort by Scott-Clayton and Li (2016), it is revealed that the cumulative default rates of all borrowers continue to go up between 12 and 20 years after the first entry. When the two authors applied the trends, they had observed to the 2004 entry cohort, they projected that approximately 40% of the borrowers would default on their loans by 2023 (Scott-Clayton, & Li, 2016). Nevertheless, not all groups of borrowers are projected to experience the same default rates due to the fact that the CDR difference between minority and white students has been on the rise. Scott-Clayton and Li (2016) noted in their study that the debt and default rates among African American institutes’ students are currently at a crisis level, to the extent that a bachelor’s degree (BA) does not guarantee security. The study found that African American BA holders are five times more likely to default on their loans than their white counterparts; this means that 21 African Americans default on their loans for every four White students (Scott-Clayton, & Li, 2016).

To date, most of the research conducted on CDR uses individual and student-level traits as the primary determinant for whether a student defaults on their loan. Most of the studies regarding student default are placed into three major categories: pre-college, in-college, and post-college experiences. These categories include the characteristics of different institutions, socioeconomic status, financial aid, and student loan debts, as well as the borrowers’ attitude and awareness towards their debt.

Lundgren (2017) recently undertook quantitative research in which he interviewed several students, staff, and faculty to test the hypothesis that students’ attitude contributes significantly to the increase of default rates. In his research, Lundgren found that different students possess specific characteristics that are independent of the institutions they attend; these characteristics significantly contribute to their capacity to repay their loans. He noted that a trait in their attitude towards debt and default is their dissatisfaction with the institution (Lundgren, 2017). In recent years, the focus of similar studies has been on the effects of economic factors and post-secondary institutes on CDR. This literature review looks at the most applicable and relevant studies of the CDR in post-secondary education in order to determine which factors have the most impact on CDR in U.S. student borrower populations.

## 2.2 Pre-College Contributions to CDR

Pre-college experiences are often captured best by looking at the family demographics or the background traits a student had before joining a college. These traits are often independent of the institution a student attends and may sometimes include the student’s age, ethnicity, gender, parent’s income, family education background, borrower aptitude, pre-college or high school curriculum, and achievement. All these factors are considered to have a significant contribution to a student’s likelihood of default. In many cases, the factors that are mostly linked with higher default rates include coming from families with low income and low level of education (Volkwein, et al., 2016). Lundgren (2017) stated that many students who lack safety nets are likely to default on their loans. Students whose families face financial challenges have a higher chance of defaulting. This implies that financial factors related to families besides demographics are essential to determine the eligibility of a student to enroll in post-secondary education, need-based or merit-based financial aid, such as grants, federal financial aid, or loans (Volkwein et al., 2016). There is research suggesting that the more a student borrows, the lower the risk of default. This is often because more loans reflect that they have more school years, or that they are attending institutions that offer high-quality education. Both of these show that the student will attain a high-quality education, which will eventually lead to program completion, and thus obtain a position that will enable them to repay their loans (Volkwein et al., 2016).

## 2.3 In-College Experience Attributed to CDR

### 2.3.1 School Dropout

Most researchers summarize students' college-experience as a prediction of student loan default and their ability to take loans without repaying (Macy, & Terry, 2017). For instance, it is argued that approximately 25% of all student borrowers who drop out before obtaining their certificates or degrees have outstanding student debt and borrowers who do not finish college face high chances of defaulting (Podgursky et al., 2017). Basis Points (BPS) show that about 40% of students who defaulted from 2011 to 2012 did not earn an educational credential, and 55% of these students were African American borrowers (Macy & Terry, 2017). The best way to prevent at-risk students from defaulting is for institutions to take necessary measures that enable a payment plan which favors all parties involved. This means that institutions have to use student data to determine whether the student has a higher risk of defaulting before creating a payment scheme (Webber, 2017).

### 2.3.2 College Major

Studies have shown the significance of institute-related factors that impact cohort default rates (CDR). These studies outline the main factors that impact students during their college life. The choice of major is one key factor that can help predict the likelihood of a student defaulting. Research has shown that college majors have a moderate impact on the default rate of each student. Students who enroll in a more generalized major have a higher rate of defaulting on their debts as compared to those with specialized majors, such as medicine, law, or nursing (Volkwein et al., 2016). Institutions have been shown to have a large variance in the default rate when comparing majors. According to Volkwein (2016), a student with a major in engineering, agriculture, or a scientific discipline, often has the lowest probability of defaulting on their debt. These fields of study often have a probability of 4% when compared to other four-year and two-year university borrowers who choose different majors (Volkwein et al., 2016). Volkwein et al. (2016) explain that student borrowers who change their majors once or twice are at a lower risk of defaulting on their loans as compared to borrowers who change their majors more than twice.

Many students who choose a second major often have a lower default rate as compared to those who have a single, very common major (Looney, 2019). This is primarily due to employment scarcity in the field of their related major. The larger the gap between students' majors and their current field of employment, the higher their default risk (Looney, 2019).

Students’ default rates are significantly affected by their attendance. It has been shown that students’ default rates are more likely to decrease with an increase in their average school attendance. Many defaulters often attend classes between one to four semesters. However, students who attend classes about eight times a semester, which constitutes more than 111 hours have a lower default rate (Looney, 2019).

### 2.3.3 Number of Years in School

Similar to the effect the number of semesters attended has on students’ default rate, the number of years that a student enrolls in an institution plays a significant part in determining their default rate. Many students who left school after undertaking three to five years of classes defaulted at a relatively lower rate when compared to those who left school after one or two years of classes. Those student borrowers who took one or two years of schooling defaulted at a rate that was 14% higher than students who spent more time in school (Perna, Kvaal, & Ruiz, 2017). Students who attended school for five years were less likely to default on student debts, but, if a student has six or more years of enrollment, they are at a higher risk of defaulting (Perna et al., 2017).

Undergraduate borrowers who attend classes for six or more years between their enrollment and their graduation date have a higher default rate as compared to their counterparts who graduated after four or five years. This means that school years have a significant relation to default rates. It is essential to note that student borrowers who have never attended classes during their summer semester years, default on their loans at an 8.9% higher rate than those who attended classes during their summer semesters. Students who have attended two or more summer semesters of classes default at a rate of 2.9% (Perna et al., 2017). When students manage to secure employment while in college, their risk of default is reduced by nearly 7.9%. However, this is only significant for non-white borrowers (Perna et al., 2017).

## 2.4 Post-College Attributes to CDR

### 2.4.1 Unemployment

All the post-college experiences can be reflected best by looking at the state of the economy. Unemployment rates have a significant impact on a student’s ability to repay their student loans. For example, Woo (2017), states that all students who finished their studies and remained unemployed have the highest probability of default. The same article explains that the probability that these odds will increase to 80% over the original probability of defaulting when the economy goes higher, which means that there is a direct relationship between unemployment, economy, and CDR. According to Woo (2017), the most influential post-college variable that is linked to default rates is filing for unemployment. It accounts for 80% of the default rates and is increased by 83% when a former student files for unemployment (Woo, 2017). According to a study undertaken by Kesterman, (2016), nationally, students revealed that the main reason for defaulting on their loans was unemployment.

### 2.4.2 Low Wages

Besides unemployment, other borrowers stated that having low wages that could barely cater for their needs was closely associated with their default rates. It revealed that not only unemployment but low-income levels while employed led to high default rates among post-secondary educated borrowers (Kesterman, 2016). To reiterate this point, Kesterman cites studies on borrowers who left postsecondary education from 1976 to 1985. The study follows a selection of borrowers who were interviewed to determine the relevance of various known factors. The author, Kesterman, states that most of the borrowers mentioned that being unemployed or having low-income employment were the primary reasons for their loan default.

### 2.4.3 Disposable Income and Student Attitude

In many instances, students report that having some disposable income may not necessarily mean that one can offset their student loans. Disposable income does not mean that the student has enough income. Most of the borrowers with disposable income may not be able to pay their loans, which is justified by their urgent needs or wants. It is observed that 11.6% of the borrowers with disposable income were unwilling to pay. In comparison, 83% who had a disposable income were in repayment (Kesterman, 2016). It led Kesterman to the conclusion that despite income being a contributing factor to student default rates, the attitude and the behavior of the borrowers’ matters.

Within many pieces of student debt literature, the areas of focus for many researchers have been on financial aid policies, size of institutions, research expenditure, admission policies, and federal loan policies are the main factors impacting student loan default rates. Macy and Terry (2017) found that one of the factors that led to high rates of student loan defaults during post-college is high tuition fees. Public universities with four-year programs are shown to have higher average debt upon student graduation than private universities, which is believed to be attributed to the enrollment of lower income-students.

To expand on the causes of student loan-defaults, studies conducted by Macy and Terry, (2017), focused on the student background characteristics, and institution characteristics. The study pointed out that those students whose families have a high amount of education were more likely to honor their commitment to paying back debts. Additionally, institutions that demand less in terms of tuition fees are more likely to have a lower rate of defaulters since many students will be able to pay from out of pocket or borrow less to pay for the lessons (Kesterman, 2016). Some institutions carried out studies like the Tennessee Higher Education Commission to find out what the constant predictors of default rates were. The study led to the realization that higher graduation rates and instructional spending were some of the constant default rate predictors. The same study focused on other causes of student default rates, such as family background traits in addition to income level and wealth, level of education, in addition to the number of degrees attained by an individual student (Kesterman, 2016).

Analysis of cohort default rates (CDR) in Post-Secondary Education goes beyond just looking at the contributing factors to student default. It covers different areas, such as the default rates by school types, ethnicity, and state. It includes the differences in variation between different ethnic groups and makes a significant conclusion on the type of population. While many researchers focus on the reasons for variations, this study goes beyond looking at these reasons to include other areas of cohort default in the United States (Kesterman, 2016).

### 2.4.4 Personal and Family Experience

According to Perna et al., (2017), individuals who are separated, divorced, or widowed have an increased probability of defaulting on their loans by 7%, having children raises the probability of default by 4.5%. The authors state that any person who has children and is separated, divorced, or widowed may have their probability of defaulting raised above 40%. In this section, the factors that can be applied to lower the rate of defaulting are universal across all ethnic groups within the United States. However, the impact on the non-whites may still be higher than the impact on the white population (Perna et al., 2017).

### 2.4.5 Knowledge of Repayment Obligations

Even though most borrowers understand that they are obligated to pay back their loans graduating, some may be confused about their payment options, the repayment processes, and the urgency of their payments. When this happens, the borrower ends up defaulting because of their inability to proceed efficiently with their payments (Volkwein et al., 2016).

## 2.5 Cohort Default Rate by Institution Sectors and Ethnicity

### 2.5.1 School Sectors or Type and CDR

One of the facts worth mentioning on different reports and literature is the ten striking patterns that have emerged in how default rates vary by institution types and ethnicity (Black, White, Hispanic, and Asians). Institutions are grouped in the following categories; public, private for-profit, private-not-for-profit two-years, and private-not-for-profit four-years. According to Scott-Clayton and Li (2016), by the time black college graduates earn their bachelor's degree, they would owe nearly $7,000 more than their white peers. It is estimated that these students often owe nearly $23,000 when compared to their white peers $16,000. The gap between the default rates of African American students and White students will triple over the next few years to approximately $25,000 (Scott-Clayton & Li, 2016).

Most of the time, the difference between graduate school borrowing and its interest accruals often leaves the students in debts amounting to about $54, 000 within four-years after their graduation. According to Kesterman (2016), African American student debt is often double that of white students. The reports issued by the Center for American Progress, explains that the racial gaps in the debt default rate is worse and growing each year. This may attribute to the fact that trends in the economy have become more troubling and unpredictable than before (Miller, 2019).

Analysis of default rates among post-secondary students in the United States on institutions reveals that these trends are most alarming among for-profit colleges (Scott-Clayton, 2018). Scott-Clayton states that per 100 students who went to a private for-profit institution, 23 of them defaulted within the 12 years of beginning their college studies in 1996, and 43 students defaulted within the 12 years of 2004. These numbers are compared against the 8 and 11 students who defaulted after attending a not-for-profit institution. The high CDR for-profit institutions are not the only alarming factor for private-for-profit institutions. They are concerned about the race gap, which seems to be growing at a worrying rate (Miller, 2019). At the private for-profit organizations or institutions, African American borrowers have a rate that is triple that of the white borrowers and double that of Latino or Hispanic borrowers (Miller, 2019).

According to Macy & Terry (2017), student borrowers who attend doctoral-granting institutions have a low default rate compared to borrowers who attend proprietary (for-profit) institutions whose default rates have been high. Furthermore, it is thought that different student policies and legislation are substantially based on the belief that colleges or universities have a significant impact on the actions of the students. Volkwein et al. (2016) found out that there is little data to prove this. The authors stated in their literature that the difference in default rates by school type is more linked to the nature of the borrowers and their achievement and less linked to the nature of the institution they attended because different institutions have different students. Woo mentions in his article that the fact that many borrowers who attend short-term (two-year) programs have higher default rates as compared to the borrowers who attend long-term (four-year) programs can be explained based on the types of students enrolled in the programs and not on factors associated with the programs or institutions (Woo, 2017). The student number within an institution does not play any significant role in the default rates. If closer monitoring of students and personal relationship with them reduced the probability of default, then smaller institutions should have the lowest default rates in the United States, but according to different research, bigger institutions compared to smaller ones have the lowest default rates in the United States; thus, the size of an institution and the default rates have no significant relation (Woo, 2017).

### 2.5.2 Ethnicity and CDR

Miller’s study was based on BPS data. In his study, he found that students who defaulted had a trend of borrowing small sums of money, while those who honored their debts or part of their debts borrowed very large sums of money. The median borrowers had taken up to $6,750 in loans, which were lower than the maximum allowable loans for students in their first year of college (Miller, 2019). The major ethnic attributes that are linked to lower default rates include being a student of an Asian American or White descent, having parents who have attained a college certificate, and coming from a family whose income is over $30,000 (Miller, 2019). On the other hand, the ethnic attributes linked to higher default rates include being a student of an African American or American Indian descent, coming from a family in which the parents hardly obtained a certificate or high school diploma, with extremely lower income levels (Scott-Clayton, 2018). African Americans and Hispanics are mostly associated with a relatively lower level of degree attainment, and twice the number of children as that of their white counterparts. Additionally, African American and Hispanic families are often associated with twice the rate of separations and divorce as compared to Whites. These are the main predictors of default rates rather than ethnicity (Volkwein et al., 2016).

Nevertheless, Scott-Clayton (2018) states that most African American and Hispanic defaulters have the highest probability of being unemployed, being dissatisfied with their curriculum programs, and having personal commitments and issues that affect their repayment (Scott-Clayton, 2018). Across all ethnic groups in the United States, all borrowers owe approximately one-half of the initial loan amounts four years after they graduate. According to Miller, the factors that lower defaults are the same across all ethnic groups, but their impact on non-Whites is greater. Miller (2019), states that among all populations in the United States, being a married female lowers the probability of default, but it does so more dramatically for non-Whites than for Whites.

## 2.6 Other Causes of High CDR

Besides ethnicity and sector in which an institution belongs, the researchers have focused on loan servicing factors as another approach to understanding cohort default rates in post-secondary education. According to a study carried out by Woo (2017), all the students who had loans held by more than one servicer had the highest probability of defaulting on the loan. The additional servicer often increases the probability of defaulting on the loan by around 18% (Woo, 2017). Additionally, the number of loans taken out by a student and not the amount of the loan taken is directly linked to default. This means that a higher number of loans and not a large number of loans signal a higher risk of defaulting (Volkwein et al., 2016). The reason for this high risk is because any person who has loans with many servicers has to draft more repayment checks.

Students who take out loans from many agents often have more to pay with regards to the amount received because every loan has to be paid back with an accrued interest percentage (Woo, 2017). The default rate is often very high in these cases because of the different interest rates and the inability to pay, which is partly due to lack of employment or dropout, as well as high economic demands in society.

## 2.7 Consequences of Defaulting on Student Borrowers

Analysis of the cohort default rate includes understanding the consequences of default and the life of different students after defaulting on their debts. Miller (2019) further states that debt without a degree is a bad place for any student to be, especially if a student is unable to file for unemployment. In many instances, if a borrower does find a way of making payment arrangements with the servicers to get out of the default, the loan owed may go to collections (Miller, 2019). If a loan goes to collection, the servicer or the loan agent may charge an additional 25% of the amount owed, meaning that the defaulter lands into even more debt. For students who have not defaulted, according to Fuller (2019), defaulting may lead to lower credit scores, which makes it harder for them to access some basic needs, such as renting a home in the future. Scott-Clayton (2018) states that there are states in which default leads to complete revocation of different professional licenses. Additionally, in these states, credit histories may be evaluated as a part of the employment process. Thus, it becomes harder for defaulters to get any jobs or retain the ones they already have. For students who may need further loans to advance their studies, accessing additional student loans or other aid may be difficult (Webber, 2017).

## 2.8 Recovery from Default

Miller, (2019), states that defaulting on a loan never results in a permanent designation. Thus, students have the chance to repay their loans if circumstances allow. Miller (2019), suggests that more than half (approximately 54%) of the defaulters manage to resolve one of their default cases by the end of a 12-year follow up after the first entry into the program, and approximately 14% of the defaulters are able to go back to school after making use of the four primary approaches, namely; loan discharge, rehabilitation, paying in full and consolidation of the loans, or getting out of the default status (Miller, 2019).

However, getting out of the default status is not an easy experience, especially if the reason for defaulting is unemployment or school dropout. For unemployed defaulters, it is a matter of time before they get employed. At the same time, for dropouts, it depends on their ability and willingness to go back to school and later seek employment after graduation. In the event that a student who dropped out of school is unable to complete his or her studies, they may be branded as defaulters for the rest of their lives, which makes their condition worse, especially in cases where they are denied access to certain facilities due to their credit rating (Armona et al., 2019).

## 2.9 Impact of Student Debt on Retirement and Longevity

Defaulting on debts does not only affect the students after they leave school, but it extends to their retirement days. Currently, nearly 44 million Americans of all ages bear the heavy burden of over $1.4 trillion in student loan debt (Brady, Miller, Balmuth, D’Ambrosio, & Coughlin, 2019). In addition, the burden of defaulting spreads beyond the defaulter to their families and other taxpayers within the country. Student loan borrowers are often caught unprepared financially for life after the meager jobs, making it hard for them to provide sufficient care for their family members, let alone pay the already interest-accruing student loans (Brady et al., 2019). According to a recent study, older people in the United States hold smaller student loan proportions as compared to younger citizens. It is seen that the older adults are labeled as the fastest-growing subset of borrowers who have excessively high rates of student loan defaults. Because of their default on the loans awarded to them for education, Americans aged 65 years and above experienced a 500% increase in offsets over the past few decades in terms of their social security benefits (Brady et al., 2019).

## 2.10 Effectiveness of Federal Student Loan Accountability Regulations

Within the education sector in the United States of America, accountability regulations have played a huge role in ensuring that student outcomes have been improved effectively (Lau, 2020). In higher education, major regulations have been put in place to enhance the ability of borrowers to pay back their loans by ensuring that sanctions are placed on institutions that have more than the acceptable limit of student defaults. The question is whether the regulations put in place are effective enough to manage student default (Lau, 2020).

Looney (2019) learned in his study of the regulations governing higher education that several recent regulatory actions have been taken by the Department of Education (DoEd) that have crippled the accountability systems that were put in place to track the use of the federal financial aid. One proposal from the DoEd was to cut the ‘borrower defense” regulation that was put in place to protect and relieve students who were misled by their institutions. Another move by the Department was to repeal the Gainful Employment rules, which were put in place to tie federal funding to different programs’ ability to guide and prepare students for gainful and better employment (Looney, 2019).

Current regulations are not strong enough to protect the huge investments that taxpayers are making for students (Lau, 2020). Now, federal regulations are only focused on repayment plans and loan award processes and procedures. Though many of the federal aid given to students flows to low-quality education programs which lead to poor degrees. These degrees do not guarantee a student a quality job within the country and leave a huge debt burden on the students after graduation (Lau, 2020). Institutions vary significantly in their quality and value of output. Some of them have programs that provide little or no economic value to the students, thus leading to poor financial outcomes and skills. This means that regulations that support accountability should be implemented to limit subsidies to low-quality institutions and redirect enrollment to other institutions that can deliver results that have high economic value (Looney, 2019).

## 2.11 What Should Regulations Look Like?

The federal government needs to implement regulations that are strictly based on appropriate and robust measures to ensure that student outcomes are able to facilitate better results and improve their earning potential (Mueller, & Yannelis, 2019). Regulations will be able to solve many problems that plague federally supported students. Developing effective sanctions is crucial because it is likely to disproportionately affect for-profit programs, which are known to be the main cause of the problems that student borrowers undergo (Mueller & Yannelis, 2019). Regulations have the potential to set up a system of accountability that has the ability to improve the economic outcomes of students by addressing the quality challenges and improving or eliminating low-performing institutions (Lau, 2020).

## 2.12 Policy Recommendations for Facilitating Repayment

Student loans, over the years, have been known to be risky for student borrowers for many reasons. One reason is low completion rates in many post-secondary educational institutions. Often, there is the possibility of both short-term and long-term liquidity limitations that hinder the student from honoring their commitment to pay back their loans.

Recommendations have been provided for Federal policymakers to come up with a set of initiatives that reduce non-repayment by ensuring that the students are able to access quality education while encouraging forbearance and overall management of deferment approaches, as well as strengthening all income-driven repayment options (Perna et al., 2017).

Student loans are considered one of the most significant pillars of the United States’ post-secondary education system. Borrowing to pay for school or college is a wise investment for many students in the United States, but they still default at a very high rate, making the investment a heavy burden on taxpayers. This means that policies should be put in place to inform the student about the best ways to handle loans without defaulting. Institutions should have a repayment plan or scheme to help mitigate default risk from students. As it stands, federal regulations and repayment policies, together with institutional policies, play a greater role in the default rates. Perna et al. (2017) state that ensuring that the default rate is brought down significantly is a joint effort between policymakers, students, and the institutions themselves.

Policymakers can make laws that ensure all students who graduate from college are absorbed into the job market. Students contribute by changing their attitudes and behavior towards repayment, while institutions can come up with cost-friendly payment schemes that are not very demanding in terms of repayment premiums (Perna et al., 2017).

## 2.13 Student Loan Tips

According to Looney, if a first-time borrower should decide to take out a loan, they should take their time to complete an entrance counseling program and sign a master promissory note to ensure that they are able to access their online account for the loans. This is one way to ensure that the student tracks the loan and only goes for the option that they presume will be easier to pay in the future. Looney states that this is one way of ensuring that students are not misled by institutions and pushed into taking out loans that they cannot pay. On the other hand, Lau, (2020) states that it is essential that students cancel subsequent loans if they drop out before finishing their studies; this is only possible if the students are made aware of their obligations and the consequences of taking out too many loans that may be a problem in future. Policies must be implemented to address these loopholes, and thus, ensure that students are trained and made aware of existing loan packages, and the ways of selecting a loan that best suits their education needs (Brady et al., 2019).

Not only is training students a way of ensuring that they take out loans they can repay, but it helps them to understand different enrollment patterns they may encounter (Looney, 2019). When institutions access cross-sectional enrollment data that is often issued by the National Postsecondary Student Aid Survey (NPSAS), they are able to assess the new trends of graduate enrollment patterns. With this information, institutions can come up with a module to control different ethnic disparities and gaps in the cohort default rate. Therefore, they would be able to determine how enrollment in different types of institutions contributes to loan default, and how best to mitigate this issue.

# Chapter 3: Methodology

## 3.1 Instruments

Data for this study is collected from three datasets, the U.S. Department of Education Federal Student Aid, United States Census Bureau, and SimpleMaps for access to the U.S. Cities Database. Different instruments were used to meet all the obligations, and thus come up with a list of all the desired records and file formats. An essential preliminary tool is Microsoft Excel since all the datasets downloaded are in the Excel file format, which holds relevant information in different rows and columns. The entire study and analysis will be done using Machine Learning (ML) to come up with a pertinent prediction of the current default rates in postsecondary education in different cities within the United States. This means that the researcher must have different outcome variables that contain information on the types of institutions, and the significant factors that contribute to a student defaulting on the loans awarded to them. The required information for the analysis includes the county, cities, name of the school, state, the number of students in default for the years 2016, 2015 and 2014, number of students in payment for the three years from 2014 to 2016, the type of institution, The Office of Postsecondary Education Identification (OPEID), the region in which a school is found, the ethnic codes, the gender of the students, the programs and the program length among much other information. The aim is to cover all the information relevant to the analysis and to come up with all the predictors that can be used to determine the likelihood of default and the overall rate of default in the entire country. All the information above and any other relevant information will be organized in Excel sheets.

Information for this study on CDR will be retrieved by conducting a search on three distinct websites maintained by the state and federal departments within the U.S. All Excel files in these websites belonging to 2014, 2015, and 2016 cohorts will have to be downloaded and stored in CSV files. This is to eliminate too much research due to the short time available for the study. The analysis processes will be implemented based on the attributes presented in these databases created with the help of the CSV files, and the pre-existing data held on the websites used for this study. The researcher will make use of a third tool in the process of collecting data called power query, to look up and access the outcome variables that will be maintained on the latest CSV file. The statistical analysis required for the research will be performed in an iPython notebook environment, Jupyter Notebook. Using Python scripts and Tableau, the research team can create data visualizations. All tools used in the research are available for free and can be downloaded at any time.

## 3.2 Procedure

In this study, quantitative research was a significant approach to the entire project and data analysis because of all that was involved. Choosing the type of quantitative analysis to run is an essential task because it determines if one is going to obtain the right dependent and independent variables. If one picks a wrong approach to the process of data analysis, there is often a high chance of getting a completely wrong result. It is essential to note that regression analysis is one of the most critical quantitative approaches to data analysis that many researchers use to reject or accept their hypothesis. It plays a massive role in determining the relationship that independent variables have on the dependent variable, considered all the statistical assumptions such as the fixed nature of all other variables. The process of calculating the relationships will always result in a straight line or otherwise, and Pearson’s correlation coefficient (*r*). This project takes a similar approach to determine the correlation between the dependent and the independent variables.

Data collection in this study will be done by searching the Federal Student Aid website database that contains data on the CDR, the U.S. Census Bureau website, and the U.S. Cities Database maintained on the SimpleMaps website for data that can be used to determine cohort default rates in the U.S. The SimpleMaps website is used to keep data about city locations in the U.S. Additionally, the U.S. Census database maintains the distribution of the cohorts by gender and ethnicity or race. It is used to hold the distribution of population and schools within the different cities of the U.S. This process follows the thesis formulation and literature review, and since the dataset is pre-existing, there was no survey conducted, which means that no human participants were used as sources of information. The researcher will start by locating cohort data for the years 2014, 2015, and 2016. The data on these databases are in CSV file formats, which will be downloaded for the analysis. The information maintained is on cohort default rates by cities within the U.S. After downloading the CSV files, the researcher will have to store them on the Relational Database Management System (RDMS), awaiting further processing. Statistical data will be stored in tables. The table will hold information on different entities and their attributes, which will provide an overview of information on various records relevant to the analysis and the study. Based on the entities in different databases, the researcher will conduct an evaluation to find the outcome variable, which will be grouped into either dependent or independent variables that will be used for the process of determining the correlations.

Since the principal researcher is using data from different sources, it will be advisable to merge and establish a relationship between all these records. As a result, the researchers will use a combination of Excel and Python merge tools to discover and connect to all this data from these three sources. After connecting to these databases, the researcher will proceed to transform and join this data using the Python scripts.

## 3.3 Data Analysis

Once all the data has been prepared and stored in the different databases, the researcher will implement the analysis of data in the iPython notebook and generate an interactive visualization by using Tableau. The project researcher will have to save all the Excel files in a CSV format for an effortless loading into the iPythonnNotebook. In the software, the researcher has to load the data from the CDR database, Census database, and Sitemap database to come up with a CDR and city data frame. The researcher must conduct exploratory data analysis (EDA) processes that will be used to determine data summary, coefficient, skewness, and variance, among others. In the process of establishing an overview of the data, the principal researcher will have to evaluate the collected and organized data based on parameters such as the dimension of the entire dataset, breakdown of each instance of the class of data, types of data attributes, the level of the class attributes, and the statistical summary of all the variables used in the study. All these must be done to ensure that a clear summary of all the relevant data is established in the entire process and to give a clear visual presentation of all the merged data attributes.

The project researcher must set up different tests to harness for the 10-fold cross-validation as well as to run a skewness test on the data. The researchers will use regression models to predict the rate of cohort default within the U.S. Some of the algorithms run by the researcher to predict the likelihood of loan default in post-secondary include *K*-nearest neighbors, random forest, and logistic regression. Additionally, the algorithms selected above will be run directly on the validation sets, and the result will be summarized in a confusion matrix to determine the level of effectiveness of the algorithms used in the process of prediction. The confusion matrix will be run for all the algorithms used in the analysis.

After using Python to create the models, the researcher team will use Tableau to run the Python scripts to manage the parameters established during data visualizations and run all the processes that will include editing the queries to suit the merged data.

After the researcher loads all the necessary data to the Tableau dashboard and after developing the script, the next step will be to use the different buttons to build a new database in which all the detailed data will be loaded in the table, and the information relating to the parameters will be modified based on existing attributes. When an advanced editor prompt pops up, the researcher will have to run other functions to load all the edited records on new tables. This is the step used to clean all the data.

Once the researcher is sure that all the data has been cleaned, the next step involves conducting a preliminary data analysis. In this process, the researcher will try to determine the initial graduation rate. This default rate was taken for all schools in the U.S. Schools analyzed include public schools, private, not-for-profit schools, and proprietary schools. The box plots will be used to visually determine the relationships and make all the concepts clear in the long run. The principal researcher will use the skewness to determine how the distribution of the default rates compare by each institution within the different states in the U.S. Census data was used to create another dataset based on the gender, race, and state of the students who are either in default or in the payment plan.

To handle all the missing data, the principal researcher will delete all non-essential data. This includes data that returned null values. The researcher has to check again for all the values eliminated before proceeding to the next step. If the data is not enough for the analysis, the researcher will use the internet source to copy additional data that will be added to the data frame to fix all the missing values in the study. After adding all the data, the researcher must run a test to check if there are still missing values and then eliminate them from the set. Data frames that have missing values and those with missing values must be exported separately. After this step, the modeling process will begin, and the second and third models will have to be created.

## 3.4 Introduction to Variables

After loading the file to the application, the researcher expects to find a total of 55 variables that are divided into one dependent variable and 54 independent variables, respectively. In this case, the graduation rate is the dependent variable for all the years used in this cohort. The list of variables used in this study is in Table 1.

Table 1

List of Variables

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Data Example |
| OPEID | int | 1060 21706 33614 41215 1020 1015 5697 7871 ... |
| state.abv\_x | factor | w/ 51 levels "AK","AL","AR", … |
| state | factor | w/ 51 levels "Alabama","Alaska", … |
| county | factor | w/ 890 levels "Abbeville","Ada", … |
| city | factor | w/ 1833 levels "Aberdeen","Abilene", ... |
| Name | factor | w/ 4056 levels "A – TECHNICAL …", … |
| Address | factor | w/ 4235 levels "#377 PONCE DE LEON …”, … |
| Zip.Code | int | 36507 36526 36535 36561 36265 36331 35662 ... |
| Zip.Ext | int | 2698 7055 0 3815 1602 1300 3206 2000 5232 159 ... |
| Prog.Length | int | 5 8 3 8 8 5 5 5 8 5 ... |
| School.Type | int | 1 2 3 3 1 1 1 1 2 1 ... |
| Year.1 | int | 2016 2016 2016 2016 2016 2016 2016 2016 2016 ... |
| X.Num.1 | int | 213 14 62 444 212 58 191 197 73 36 ... |
| X.Denom.1 | int | 1202 226 278 4623 2157 323 830 1382 213 297 ... |
| DRate.1 | num | 17.7 6.1 22.3 9.6 9.8 17.9 23 14.2 34.2 12.1 ... |
| PRate.1 | factor | w/ 3 levels "A","B","P |
| Ethnic.Code | int | 5 5 5 5 5 5 5 5 2 5 ... |
| Cong.Dis | factor | w/ 63 levels " ","0", … |
| Region | int | 4 4 4 4 4 4 4 4 4 4 ... |
| AverageorGreater.than.30 | int | 0 0 0 0 0 0 0 0 0 0 ... |
| Year.2 | int | 2015 2015 2015 2015 2015 2015 2015 2015 2015 ... |
| X.Num.2 | int | 250 16 59 458 292 79 148 236 111 50 ... |
| X.Denom.2 | int | 1345 203 244 4886 2454 377 675 1524 357 330 ... |
| DRate.2 | num | 18.5 7.8 24.1 9.3 11.8 20.9 21.9 15.4 31 15.1 ... |
| PRate.2 | factor | w/ 4 levels "","A","B","P" 4 2 2 2 2 2 4 2 2 2 ... |
| Year.3 | int | 2014 2014 2014 2014 2014 2014 2014 2014 2014 ... |
| X.Num.3 | int | 274 16 65 297 272 109 166 326 83 80 ... |
| X.Denom.3 | int | 1363 249 376 3890 2393 490 680 1660 298 414 ... |
| DRate.3 | num | 20.1 6.4 17.2 7.6 11.3 22.2 24.4 19.6 27.8 19.3 ... |
| PRate.3 | factor | w/ 4 levels "","A","B","P" 4 2 2 2 2 2 4 2 2 2 ... |
| lat | num | 30.9 30.6 30.4 30.3 33.8 ... |
| lng | num | -87.8 -87.9 -87.7 -87.6 -85.8 ... |
| population | int | 8532 69065 33113 6029 12612 38737 14022 3352 ... |
| density | num | 205 583 219 158 444 356 327 290 513 94 ... |
| military | logi | FALSE FALSE FALSE FALSE FALSE FALSE ... |
| age\_median | num | 32.7 37.8 50.2 53.9 26.7 37.3 41.4 52.7 37.1 42.2 ... |
| male | num | 51.6 49.9 46.4 52.4 47.5 48.5 47.8 47.4 44.3 47.3 ... |
| female | num | 48.4 50.1 53.6 47.6 52.5 51.5 52.2 52.6 55.7 52.7 ... |
| married | num | 35.2 50.9 44.5 58.4 29.7 51.8 52.8 37 29.1 51.1 ... |
| family\_size | num | 3.96 3.39 2.97 2.78 3.12 3.18 3.05 3.02 3.15 3.09 ... |
| income\_household\_median | int | 31310 65739 42468 69388 40012 54304 52201 ... |
| income\_household\_six\_figure | num | 11.7 26.4 10.1 31.9 14.8 23.4 18.6 9 8.2 10.4 ... |
| home\_ownership | num | 49.2 64.5 67.3 68.3 49.4 57.5 74.8 57.8 44.7 68.9 ... |
| home\_value | int | 124313 189560 161857 259226 147232 171948 ... |
| rent\_median | int | 697 1130 818 929 796 850 772 545 566 637 ... |
| education\_college\_or\_above | num | 15.5 41.9 26.2 43.7 32.8 29.1 27.1 12.3 17.2 9.3 ... |
| labor\_force\_participation | num | 47.7 66.7 51.5 58 56.3 63 58.3 38.9 50.2 51.3 ... |
| unemployment\_rate | num | 5 4.8 10.5 5.8 13.1 5.9 4.8 14 18.1 6.9 ... |
| race\_white | num | 54.9 80.4 80.6 96.6 71.7 72.1 83.6 92.1 17.3 92.3 ... |
| race\_black | num | 41.5 14.9 15.8 0.7 23.5 21.7 13 5.6 80.4 0.6 ... |
| race\_asian | num | 0.4 2.6 0.7 0.2 1.4 1.7 0.9 0 0.8 0 ... |
| race\_native | num | 0.4 0 0.3 0 0.1 0.5 1.1 0 0.1 0.3 ... |
| race\_pacific | num | 0 0 0 0 0 0 0 0 0.1 0 ... |
| race\_other | num | 1.5 0.8 2.3 0.1 0.5 0.9 0.4 0.4 0.4 3.3 ... |
| race\_multiple | num | 1.4 1.2 0.4 2.4 2.8 3 1.1 1.9 0.9 3.6 ... |

## 3.5 Data Limitations

The Official Cohort Default Rates (CDR) for Schools (2019) dataset does not offer a student-by-student breakdown of CDR, so the research team cannot look at a detailed distribution of each school’s student population. The researchers plan to use the county demographic data for the school’s respective headquarters to get an idea of the distributions of the school’s students. It is essential to keep in mind that many students who attend college, especially online colleges, do not live in the county in which the school is located. This can create discrepancies between the distribution of variables such as median income in a school area and the median income of students. However, in 2019, 67% of online students stated that they live within 50 miles of their institution, and 77% of online students live within 100 miles of the institution (Lederman, 2019). This allows the team to compare the demographic information of a county with the performance of a school in that county, regardless of the instruction method.

The dataset that gives the researchers most of the geographical and demographic data, the U.S. Cities Dataset, is generated by a third party, SimpleMaps. However, it is derived from open data sources, including the U.S. Census, U.S. Geological Survey, and the American Community Survey. All the data is publicly available and can be gathered independently to check the validity of the data contained in the SimpleMaps U.S. Cities Dataset. Additionally, it is periodically updated to ensure the data is accurate and up to date. The last update to the dataset was made on September 11, 2019.

## 3.6 Continuous Variables

To create an ML model, the data needs to be explored thoroughly. Therefore, it is essential that a methodical approach is taken to understand the trends and variation, which is shown within each variable. To recap, the goal of this project is to create a predictive model for the variable CDR of an institute while adjusting for race and population. Wongsuphasawat et al. (2019) call this technique Model Development.

Model Development is defined as an EDA technique used to create new metrics, ML models, data rules, and trends. It is useful when the base idea of a project is to develop new functional applications such as dashboards and reports to enable further insights (Wongsuphasawat et al., 2019). The project will use EDA techniques similar to Wongsuphasawat et al. (2019) Model Development.

### 3.6.1 Objective

EDA for numerical data will be done to determine whether the observations made within each variable are useful in the prediction of CDR. It is done to dissuade some of the biases that may be assumed when dealing with race and economic status. The data shall be explored for the following aspects; data distribution, shape, quality, extreme values, outlier, and inconsistencies. These measurements shall be used to determine the transformation, scaling, and adjustments that are necessary to confirm all numerical variables. Figure 1 illustrates the steps that need to complete the exploration of numerical data. However, the process is conducted with the intent of retaining variables and observations, if possible. This is done for data integrity, which is due to the low dimensionally within the dataset.

Figure 1

*Numerical Variables EDA Steps*

Identify numerical variables

Identify Basic matrices

Non-Graphical Analysis

Graphical Analysis

Variable Transformation

Outlier Handling

Correlation Analysis

These steps shall be done in using the high-level language Python 3 in the Jupyter Notebook version 5.4.0 application.

### 3.6.2 Numerical Variable Identification

The identification of numerical variables is essential to the further understanding of the data and thought they are identified in the previous sections; they shall be reintroduced see Table 2. This section shows an overview of the numerical variable in the dataset and purely cosmetic, which has been conducted. Overall, the merged CDR dataset contains 30 numerical variables.

Table 2

*Numerical Variables*

|  |  |  |
| --- | --- | --- |
| Variable | Column Name | Data Type |
| Program Length | proglength | Int |
| Cohort 2016 | cohort2016 | Int |
| CDR 2016 | cdr2016 | Float |
| Average or Greater Than 30 | averageorgreaterthan30 | Int |
| Cohort 2015 | cohort2015 | Float |
| CDR 2015 | cdr2015 | Float |
| Cohort 2014 | cohort2014 | Float |
| CDR 2014 | cdr2014 | Float |
| Population | population | Float |
| Density | density | Float |
| Age Median | agemedian | Float |
| Male | male | Float |
| Female | female | Float |
| Married | married | Float |
| Family Size | familysize | Float |
| Income Household Median | incomehouseholdmedian | Float |
| Income Household Six Figure | incomehouseholdsixfigure | Float |
| Home Ownership | homeownership | Float |
| Home Value | homevalue | Float |
| Rent Median | rentmedian | Float |
| Education College or Above | educationcollegeorabove | Float |
| Labor Force Participation | laborforceparticipation | Float |
| Unemployment Rate | unemploymentrate | Float |
| Race White | racewhite | Float |
| Race Black | raceblack | Float |
| Race Asian | raceasian | Float |
| Race Native | racenative | Float |
| Race Pacific | racepacific | Float |
| Race Other | raceother | Float |
| Race Multiple | racemultiple | Float |

A few improvements have been implemented to the dataset in lieu of missing value handling and data consistency. The following changes have been made:

1. Drate: Renamed to CDR of the appropriate year, i.e. (2014, 2015, and 2016).
2. Denom: Rename to cohort of the appropriate year, i.e. (2014, 2015, and 2016).
3. Year: Variable has been dropped from the dataset due to being redundant.
4. Num: Variable has been dropped from the dataset due to being a calculation of DRATE and DENOM.

### 3.6.3 Analyze Basic matrices

The identification of the metrics which will be used to explore a variable is defined in this section. The parameters selected will be used to identify critical features within the variables that will affect the accuracy of the models’ design by this dataset. These metrics will enable the achievements of the goals mentioned in Section 1.1: data distribution, shape, quality, extreme values, and inconsistencies. These primary metrics are defined in Table 3.

Table 3

*Basic Statistics*

|  |  |
| --- | --- |
| Metric | Insights which can be made |
| Count | a numerical value that gives insight into the non-null observations with a variable. |
| Mean | provides central tendency and relative spread of a variable in regard to its’ minimum and maximum values. |
| Minimum Value | lowest observed value with a variable. |
| Maximum Value | highest observed value with a variable. |
| Standard Deviation | a measure of spread within a variable in regard to its’ mean. |
| Quantile | division of a variable into its probability of occurrence. |
| Skewness | a measurement of the symmetry of a distribution (Lovric, 2011). |
| Pearson’s Correlation Coeffiencent | a value that measures the linear correlation amongst two variables (Sedgwick, 2012). |

### 3.6.4 Non-Graphical Analysis

The non-graphical analysis focuses on the significant five basic metrics; count, mean, standard deviation, minimum, and maximum value, which are explained in Table 3. The raw values of these metrics are shown in Table 4. Firstly, by observing the count metric in Table 4, it is seen that there are no missing values with each variable having a count of 4259. Secondly, the variable “Average or Greater Than 30” and “Program Length” needs further investigation in this section, due to the limited range between their minimum and maximum values.

Table 4

*Numeric Variables General Statistics*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Count | Mean | STD | MIN | MAX | Skewness |
| Age Median | 4259 | 35.45 | 5.67 | 19.6 | 70.9 | 0.23 |
| Average or Greater Than 30 | 4259 | 0.09 | 0.29 | 0 | 1 | 0 |
| CDR 2014 | 4259 | 10.83 | 8.37 | 0 | 100 | 2.22 |
| CDR 2015 | 4259 | 10.43 | 7.75 | 0 | 100 | 1.59 |
| CDR 2016 | 4259 | 10.3 | 7.42 | 0 | 66.6 | 1.23 |
| Cohort 2014 | 4259 | 1071.82 | 3311.21 | 0 | 159144 | 28.12 |
| Cohort 2015 | 4259 | 1067.87 | 3101.32 | 1 | 139070 | 23.64 |
| Cohort 2016 | 4259 | 1002.36 | 2693.86 | 3 | 111509 | 19.53 |
| Density | 4259 | 1551.46 | 1942.55 | 4 | 21116 | 4.12 |
| Education College or Above | 4259 | 33.37 | 14.53 | 2.5 | 95.3 | 0.93 |
| Family Size | 4259 | 3.19 | 0.3 | 2.09 | 5.32 | 1.08 |
| Female | 4259 | 51.29 | 2.34 | 6.5 | 66.8 | -3.09 |
| Home Ownership | 4259 | 53.97 | 12.87 | 4.7 | 100 | 0.34 |
| Home Value | 4259 | 239035.88 | 202628.85 | 29814 | 2000001 | 3.15 |
| Income Household Median | 4259 | 54762.25 | 22595.58 | 13272 | 250001 | 2.42 |
| Income Household Six Figure | 4259 | 22.83 | 12.64 | 0 | 84.1 | 1.31 |
| Labor Force Participation | 4259 | 63.12 | 6.83 | 6.9 | 82.7 | -1.12 |
| Male | 4259 | 48.71 | 2.34 | 33.2 | 93.5 | 3.09 |
| Married | 4259 | 41.01 | 9.28 | 3.5 | 77.7 | -0.12 |
| Population | 4259 | 1031715.79 | 2959840.09 | 241 | 19354922 | 4.6 |
| Program Length | 4259 | 5.82 | 2.04 | 1 | 11 | -0.11 |
| Race Asian | 4259 | 5.48 | 7.57 | 0 | 72.7 | 3.72 |
| Race Black | 4259 | 15.82 | 17.3 | 0 | 97.4 | 1.61 |
| Race Multiple | 4259 | 3.46 | 2.29 | 0 | 37.9 | 5.21 |
| Race Native | 4259 | 0.71 | 2.02 | 0 | 58.4 | 14.56 |
| Race Other | 4259 | 4.48 | 5.84 | 0 | 51.6 | 2.61 |
| Race Pacific | 4259 | 0.18 | 0.76 | 0 | 26.2 | 16.54 |
| Race White | 4259 | 69.88 | 19.31 | 1.2 | 100 | -0.73 |
| Rent Median | 4259 | 1190.93 | 511.77 | 306 | 4001 | 1.31 |
| Unemployment Rate | 4259 | 7.11 | 3.04 | 0 | 41.1 | 1.85 |

The variable “Average or Greater Than 30” is a binary representation of CDR values, which are higher than 30%. It is a binary input and, therefore, will be removed from numerical data EDA.

Program length contains values on the scale of 1 to 11 and represents the completion time of a program in an institute. Though this variable will be reclassified as a categorical variable, due to the nature of its measurement, we shall include it in the numerical variable EDA.

### 3.6.5 Graphical Analysis

Before diving into the critical aspects of data transformation and the necessary preparations model designs, a graphical representation of each variable shall be created. This section shall discuss the type of illustrations that will be shown throughout the EDA process and highlight their importance.

#### Histograms

A histogram depicts the distribution of a variable by illustrating four of the criteria, which are listed in section 3.6.3 Table 3.

#### Density Plots

The density plot will be used to depict similar information as the histogram with the added benefit of the smoother display. Effectively, it will show where values are concentrated clearer.

#### Box Plots

Box plots will be used to display the five summary statistics that will be used to understand the variability of a variable distribution, essentially showing the outliers within the variable.

#### Heatmap

The heatmap will illustrate how numerical variables correlate to each other, by using a color map to indicate the level of correlations between each variable.

### 3.6.6 Variable Transformation

#### Transformation

The secondary goal of this project is to create a linear regression model that will predict CDR; hence the transformation of variables is not necessary. However, it is essential to use other ML models to determine whether regression is optimal. Variable transformation shall be done using three techniques, Logarithmic (3) and (4) or Reciprocal (6) for a variable that are positively skewed and Square-Root (5) for those which are negatively skewed (Howell, pp. 318-324, 2007). Figure 2 shows the histograms of the transformed variables.

(3)

(4)

(5)

(6)

*x*: an instance of a numerical observation.

*C*: constant, which is added to the value, making the minimum value *C*.

Figure 2

*Histogram of Variable Transformation*

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

#### Scaling

By observing the statistics in Table 4 and reviewing the data dictionary, it is revealed that there are three measurement types being utilized (dollars, percentages, and counts). The Linear Scaling (7) will be used to scale these measurement types. Linear Scaling, often known as Range Scaling, is a technique that sets all values of a variable between zero and one by computing the difference between an observation and its maximum and minimum values. Using Linear Scaling is optimal because all are under the effect of a transformation, i.e., “log, square or reciprocal” and outlier handling. This helps prevent the clustering of observations into the smallest part of the scale (Normalization Techniques at a Glance, n.d.). Figure 3 depicts the results of linear

(7)

Figure 3

*Density Plot of Variable Scaling*

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### 3.6.7 Outlier Handling

Based on the credibility of the data sources and the evaluation of the basic statistics, it is apparent that there are no recoding errors, miscalculation, and or any irrelevant observation. Outlier handling will focus on dealing with observations that are very extreme, typically any observation in the 0.3% range of a variable. This is done by using the Empirical Rule, which explains that 99.7% of the observation in a variable is within three standard divination (Bluman, 2009). Simply put, if observation does not fall within three standard deviations, then it will be dropped. Table 5 and Figure 4 illustrates the results of outlier removal.

Table 5

*Outliers Removed per Variable*

|  |  |  |
| --- | --- | --- |
| Variable | # of Outliers Removed | Box Plots of Variables |
| Age Median | 24 | Figure 4  *Outlier Boxplot* |
| CDR 2014 | 0 |
| CDR 2015 | 1 |
| CDR 2016 | 0 |
| Cohort 2014 | 49 |
| Cohort 2015 | 13 |
| Cohort 2016 | 7 |
| Density | 47 |
| Education College or Above | 40 |
| Family Size | 40 |
| Female | 64 |
| Home Ownership | 32 |
| Home Value | 17 |
| Income HouseHold Median | 35 |
| Income Household Six Figure | 19 |
| Labor Force Participation | 40 |
| Male | 66 |
| Married | 23 |
| Population | 0 |
| Program Length | 0 |
| Race Asian | 26 |
| Race Black | 0 |
| Race Multiple | 58 |
| Race Native | 45 |
| Race Other | 2 |
| Race Pacific | 29 |
| Race White | 20 |
| Rent Median | 10 |
| Unemployment Rate | 39 |
| Total Outliers Removed | | 746 |

### 3.6.8 Correlation Analysis

Correlation analysis will determine the effect each variable has on the other; this will be done using Pearson’s Correlation Coefficient (*r*). Table 6 explains the relationship between two variables based on their absolute *r*-value; Figure 5 is a visual representation of their relation, and Table 6 shows variables that are positive and negative correlated.

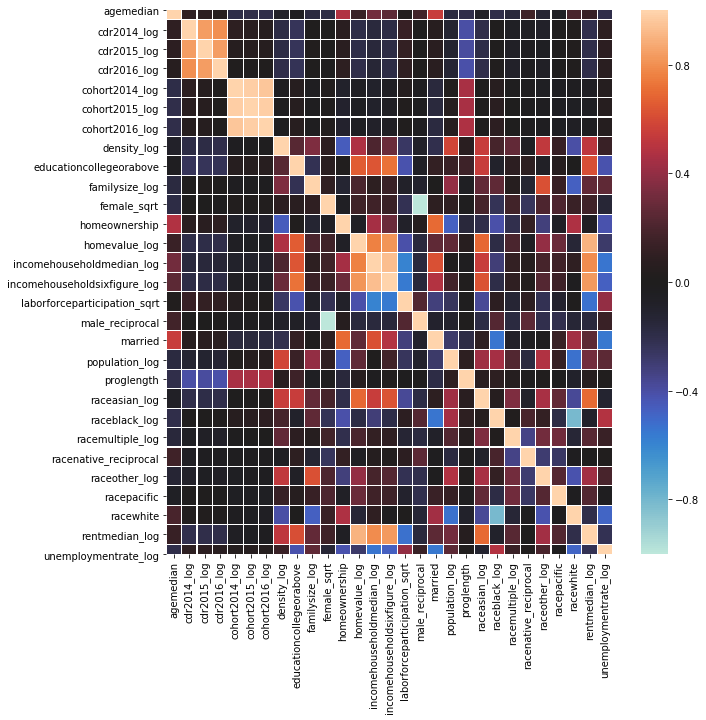
Table 6

*Pearson (r) Ranking*

|  |  |
| --- | --- |
| Absolute *r* | Relationship |
| 0.0 – 0.19 | Very Weak |
| 0.20 – 0.39 | Weak |
| 0.40 – 0.59 | Moderate |
| 0.60 – 0.79 | Strong |
| 0.80 – 1.0 | Very Strong |

Figure 5

*Correlation Heatmap*



#### Educational Variables

* CDR 2015, CDR 2015, and CDR 2016 have a very strong correlation, as shown in Figure 5. This is as expected due to the two representing a historical record of the prior.
* Cohort 2014, Cohort 2015, and Cohort 2016 have a very strong positive correlation to the following previous years. This is expected, the likelihood of an institute cohort exponentially increasing would mean that there are errors in the data or there is something else going on with institutes and their enrollment processes. What is surprising is that cohorts do not correlate to CDR.
* Program Length is shown to have a moderate negative impact on CDR, and it proves that there is a relation to the two variables, as mentioned by Perna, Kvaal, and Ruiz, 2017 study, which can be found in section 2.3.3 of this paper.

#### Socioeconomic Variables

* The correlation analysis was not able to find a correlation between Unemployment, Rent Median, and CDR. The results disprove the arguments which were made in section 2.4, but it is necessary to note that this research does not focus on the economic factors that lead to CDR.
* Figure 5 shows that Home Ownership has a small correlation to Age Median, which was expected and explained thoroughly in “Housing America's Older Adults 2018” (Joint Center for Housing Studies of Harvard University, 2018). There is a moderately negative correlation with Density and might need to be further explored.

#### Ethnicity Variables

* Based on the correlation matrix displayed in Figure 5, race does not correlate to CDR except for Race Asian, which is shown to have a weak negative relation and is discussed in section 2.5.2 with references to Miller 2019 study on ethnicity.

## 3.7 Categorical Variables

### 3.7.1 One-Hot Encoding (Dummy Variables)

While continuous data like DRate1 (Cohort Default Rate for 2016) is well suited to modeling, the datasets contain several categorical variables, such as Program Length, School Type, and Ethnic Code. In order to use these variables in the models, the research team determined that the appropriate method of translating these categories was to use a technique called One-Hot Encoding. This way, the team can create Dummy Variables that represent the presence of a specific variable value in a column.

Program Length is divided into 13 categories that refer to the length of the longest program offered by the institution:

0 – Short-Term (300-599 hours)

1 – Graduate/Professional (>= 300 hours)

2 – Non-Degree (600-899 hours)

3 – Non-Degree 1 Year (900-1799 hours)

4 – Non-Degree 2 Years (1800 – 2699 hours)

5 – Associate’s Degree

6 – Bachelor’s Degree

7 – First Professional Degree

8 – Master’s Degree of Doctor’s Degree

9 – Professional Certification

10 – Undergraduate

11 – Non-Degree 3 Plus Years (>= 2700 hours)

12 – Two-Year Transfer

School Type is divided into six categories that represent the ownership control of the institution:

1 – Public

2 – Private, Nonprofit

3 – Proprietary

5 – Foreign Public

6 – Foreign Private

7 – Foreign For-Profit

Ethnic Code refers to the code classifying the ethnic affiliation of the institution:

1 – Native American

2 – HBCU

3 – Hispanic

4 – Traditionally Black College

5 – Ethnicity Not Reported

In order to use these variables in the model, the researchers opted to use dummy variables, implemented with the use of one-hot encoding, to mark the presence of a value of a specific variable. Traditionally, one-hot encoding is used to represent the state of a finite-state machine, where each combination is represented by a “1” where the value is present. Otherwise, the instance is marked with a “0” (Harris & Harris, 2013, p. 129).

After on-hot encoding the Program Length variable, the team will have data in the format described in Table 7.

Table 7

*1-Hot Encoding of Program Length*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Associates | Bachelors | First-Professional | Grad-Professional | Masters or Doctors | Non-Degree | Non-Degree 1y | Non-Degree 2y | Non-Degree 3y |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … | … | … | … |

Similarly, the School Type variable has been expanded to the set listed in Table 8.

Table 8

*1-Hot Encoding of School Type*

|  |  |  |
| --- | --- | --- |
| private | proprietary | public |
| 1 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| … | … | … |

Lastly, the Ethnic Code variable is represented in the same way. See Table 9.

Table 9

*1-Hot Encoding of Ethnic Code*

|  |  |  |  |
| --- | --- | --- | --- |
| hbcuCollege | hispanicCollege | nativeAmericanCollege | notReportedCollege |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| … | … | … | … |

## 3.8 Feature Selection

Feature selection is the study of algorithms that play a significant role in reducing the dimensionality of data, and thus, enhance the overall performance of machine learning. Feature selection, when applied on a data set with N features and M dimensions, reduces the M dimension to M′ and M′ ≤ M. Feature selection is, therefore, an essential approach to dimensionality reduction. In this research, feature selection is the best approach to selecting the optimal subset and filtering both correlated and uncorrelated data. Feature selection has been used in this research to eliminate variables that are not central to the project and to reduce overfitting. In the statistical analysis of data, there are three types of feature selections as defined below:

* **Wrapper Method:** This approach to variable selection mainly evaluates a subset of features based on the resulting performance of the learning algorithm used. There are three major approaches in this method, namely, Forward Selection, Backward Elimination, and Recursive Feature Elimination (Raschka, n.d.).
* **Embedded Method:** This approach to feature selection uses algorithms that perform both model fitting and feature selection simultaneously by making use of the sparsity regularization, which works by reducing the weight of some features to zero (Raschka, n.d.). The various approaches used in this method include; Lasso regression (L1 regularization), Ridge regression (L2 regularization), and Elastic Net) (MathWorks, n.d.).
* **Filter Method:** This approach to feature selection mainly evaluates the importance of features, or variables based on their inherent traits, without using any learning algorithm. As compared to wrapper methods, these methods are faster and less computationally expensive. The different approaches used in this method include; ANOVA, Pearson Correlation, and Variance Thresholding (Raschka, n.d.).

### 3.8.1 Approach

To select the most significant variables for our research, various feature selection methods were run, and thus, the most cost-friendly methods were chosen. The team applied Wrapper, Embedded, and Filter methods of feature selection to lower both the computational time, and cost, and to reduce the number of input variables, to improve the general performance of our model. Using the different feature selection methods, it was possible to evaluate the relationships that existed between the input variables and the target variables in our data set; we then selected only those variables that exhibited stronger relationships with the target variables.

In the first instance, the team ran the Recursive Feature Elimination on our data set; this is considered as a type of greedy search, which only selects features by considering smaller sets of features recursively. It works by ranking features by order of their elimination. In this case, the approach returned a total of 22 variables, as indicated in Table A1 in the appendix section.

The second method was the Forward selection, which did not yield any result. The research team then ran the Backward Elimination Method, which gave a total of 13 variables, as presented in Table A2 in the appendix section. Figure 6 shows the critical features determined through Backward Elimination.

Figure 6

*Feature Importance with Backward Elimination*

A picture containing drawing

Description automatically generated

After the wrapper methods, the research team ran the embedded method of feature selection. The method that was used in this section is Lasso Regularization (L1), which works by penalizing the different parameters in the model to reduce their levels of freedom. The different penalties are then applied to the coefficients to reduce them to zero by multiplying all of them by zero. After applying the lasso method, the research team settled on 29 variables that had close relationships with the targeted variables presented in Table 3A in the appendix section. In the final prediction, the values of all variables that have been shrunk will not appear because they are reduced to zero; therefore, they do not contribute to the final prediction. Figure 7 shows a representation of the results from the Lasso method:

Figure 7

*Feature Importance with Lasso Method*

A close up of a logo

Description automatically generated

In the final approach, the team applied the Filter method of feature selection. The approach used for the selection was Pearson Correlation, and it was noted that it is impossible to correlate values that have been used to create a regression model. Feature selection using this approach yielded a total of 10 variables, as presented in Table 4A in the appendix section. Figure 8 below represents the results of the feature selection using the Pearson Correlation method:

Figure 8

*Feature Importance with Pearson Correlation*

A screenshot of a cell phone

Description automatically generated

### 3.8.2 Choice of Method

After applying the different types of feature selection methods to the data set used in this research, the team chose the em­­­­­bedded method of feature selection. Additionally, the team chose the results produced by the embedded method as the best option because it removed all the attributes or variables, which were not relevant for the prediction of the cohort default rate by using feature importance. It trains the machine learning model and derives all the essential features from the model; the derived features are often the most important when making a prediction. This means that the embedded method is the most accurate as compared to the other approaches applied in this stage of data pre-processing.

## 3.9 Data Splitting

Data splitting is a technique used to separate data into distinct sections in order to train and validate machine learning models. Two common methods for splitting data are Train/Test/Validation and Cross-Validation.

### 3.9.1 Train/Test/Validation Split

The goal of splitting data is to allow the machine learning model to train itself on a “training set” then measure its performance on an unbiased set called the “test set” that the model has never seen. This split is determined by the user. Common splits are 80% training, 20% test, 75% training, 25% test. When models allow for the manipulation of hyperparameters, a “validation set” can be introduced. This set operates as a middle step between training and testing, allowing the user to measure the effect of hyperparameter tuning on the model performance prior to confirming its performance on the test set. When using a validation set, a split may resemble 70% training, 15% validation, 15% test.

### 3.9.2 K-Folds Cross-Validation

Cross-validation resembles a Train/Test split but iterated through multiple times, selecting unique training sets and test sets, *k* times. These “folds” allow for the model to be run on each instance of the train/test split to determine the most likely outcomes over *k* iterations. The model’s performance is determined by an average of the results from each iteration.

## 3.10 Modeling

### 3.10.1 Linear Regression

Linear regression is used when one wishes to understand the relationship between two or more variables, represented by *Y* and *X* Simple linear regression estimates how much *Y* will change when *X* changes by a certain amount,” (Bruce & Bruce, 2017). In the equation below, *Y* is the dependent variable because the change in *Y* depends on the change in variable *X*, the independent variable. The intercept, *b0*, represents a constant value added to the equation. *b1* is the regression coefficient. It tells one that for each one increase in *X*, *Y* increases by *b1*.

(8)

However, this equation does not take into consideration a key issue with plotting a regression line on the data. Error, or *ei*, must be included since there is some difference between the projected regression line and the data itself.

(9)

Once a model is *fitted* (applied and allowed to output a predicted dataset), the equation changes slightly to denote the fact that this model represents predicted data instead of observed data. The symbol “^” indicates the presence of predicted value in the model.

(10)

Multiple Linear Regression expands on Simple Linear Regression by adding more predictors. For example, instead of predicting with just, one can measure the impact of , , and on (Bruce & Bruce, 2017).

(11)

### 3.10.2 Logistic Regression

Logistic Regression follows a similar style to Linear Regression and can be represented by a similar formula. The key differentiating factor, however, is the dependent variable, *p*, representing the odds that the class label is a 1. Logistic Regression is used when predicting a binary variable, as opposed to a continuous variable. To model *p*, an *inverse logit function* is applied to the linear predictor formula. This guarantees the value of *p* to be between 0 and 1 (Bruce & Bruce, 2017).

(12)

### 3.10.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a lazy machine learning algorithm that attempts to predict *x* by observing the *K* records that have similar values and then making an assumption based on these features (Bruce & Bruce, 2017). A “lazy” algorithm is labeled as such due to the fact that it makes a prediction after the data is stored and a query is made. Therefore, the *K-*NN algorithm is not actively learning and waits until the query has been issued to make a generalization. K Nearest Neighbors can be used with classification as well as continuous variable prediction. In order to determine the class of a categorical variable, each neighbor is assigned a vote, and the majority class is chosen to represent *x*. In the case of regression, a similar approach is used. However, the values of the neighbors are averaged, and that value is used to substitute for *x.*

Choosing *K* is a deceptively, simple task. There is no hard and fast rule applied to choose *K*, outside of the suggestion that *K > 1*. The true best *K* is determined by the data. Most often, an odd value of *K* is chosen due to the possibility of a tie with an even *K.*

### 3.10.4 Naïve Bayes

Bayesian classification is built on probability. “The naïve Bayes algorithm uses the probability of observing predictor values, given an outcome, to estimate the probability (*P*) of observing outcome *Y = I*, given a set of predictor values,” (Bruce & Bruce, 2017). This algorithm is “naïve” due to the fact that it makes these assumptions based on observed values without knowing the validity of the prediction. Therefore, it is essentially making educated guesses on the most likely outcome given a set of predictor values.

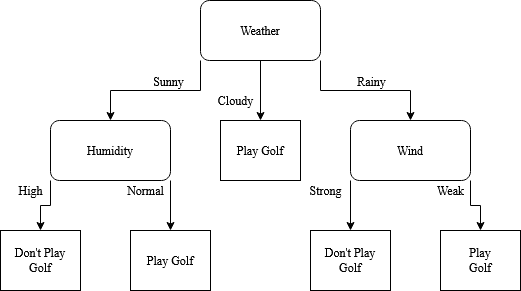
(13)

### 3.10.5 Decision Trees

Decision Trees are often heralded for their interpretability. They lend themselves well to being represented visually as well as the basis of each branching decision being based on the probability of outcomes. Starting with the root, a decision tree determines the most divisive variable, uses that to create the first partition of the data. Then, the tree continues to branch from decision to decision until a leaf, or terminal node, is reached. This represents the classification that the model has chosen for a particular series of values. For those familiar with Computer Science, this format follows the popular *if-then*-*else* rules (Bruce & Bruce, 2017). An example is reflected in Figure 9.

Figure 9

*Decision Tree for “If to play golf” (Learning Decision Trees, n.d.)*



### 3.10.6 Random Forest

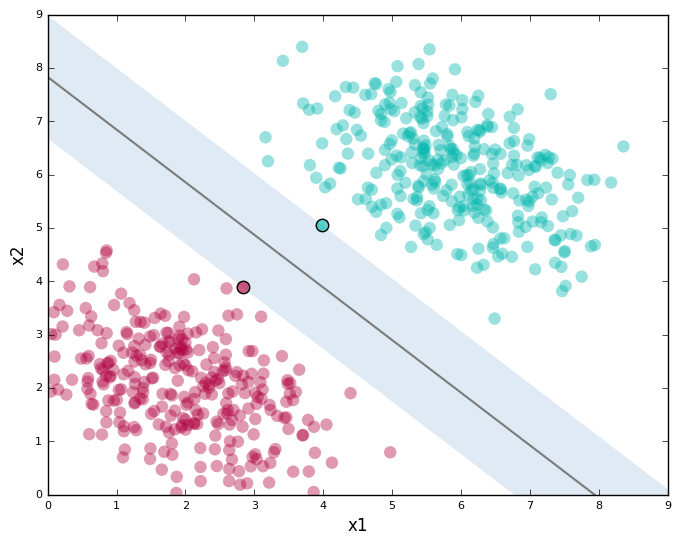
Based on the Decision Tree model, Random Forest, many decision trees, starting from randomly selected observations and building outward from that point. Then, after all, decision trees have been created, each tree is given a “vote,” similar to the voting method employed with K Nearest Neighbors. The votes are tallied, and the simple majority determines the winning class.

### 3.10.7 Support Vector Machines (SVM)

With a series of data points, there are infinite lines (or planes) that can separate points that represent two classes. The Support Vector Machine aims to find the optimal hyperplane (a subspace of *n* – 1 dimensions, with *n* being the number of dimensions of the ambient space, that bisects the data into representations of the two-class values. In order to determine the optimal hyperplane, the algorithm looks for a plan that has the maximum *margin* between the two sets of class data points (Gandhi, 2018).

Figure 10

*Representation of 2-dimensional dataset bisected with optimal hyperplane (Kumar, 2019)*



### 3.10.8 XG Boost

XG Boost is a powerful type of boosting algorithm used to classify data. Boosting refers to the process of building a model that adds *weight*, or more importance, to previously misclassified observations. The algorithm starts with a base assumption, then, after comparing the misclassification rate, adjusts the weights with the goal of minimizing the weighted error of the model. The process is rerun, with further misclassified observations receiving more weight until an iteration counter, *m* is equal to the maximum number of models set to be fit, *M.*

# Chapter 4: Results

## 4.1 Regression Models

Two algorithms were used to predict CDR for future institutes, Support Vector Machine Regression (SVR), and Linear Regression (LM). These two models are created using the sci-kit-learn package, which is available in the iPython platform and run on Jupyter Notebook. Each model uses the same features selected by Lasso Regulation and is evaluated using the following metrics; Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Root Mean Square Error (RMSE) and R Squared (R2) for model selection. The formula for each metric is defined in Table 10 and defined further into the section.

Table 10

*Evaluation Metrics Formulas for Regression Models*

|  |  |
| --- | --- |
| Metric | Formula |
| Akaike Information Criterion (AIC) | (14)  k = number of variables  SSE = sum square error |
| Bayesian Information Criterion (BIC) | (15)  *n* = number of observations |
| Root Mean Square Error (RMSE) | (16)  f = prediction  o = observed |
| R Squared (R2) | (17)  SSR = sum square of residual  SST = sum square of total |

To improve the efficiency of each metric, the data is split into eight folds using the K Fold Cross Validation methods, where 3142 (87.5%) observations of the data are used for training, and 449 (12.5%) observations are used for testing during each iteration. There was no tuning parameter set for the LM models. The SVM algorithm is tuned using the Grid Search function listed in Table 11.

Table 11

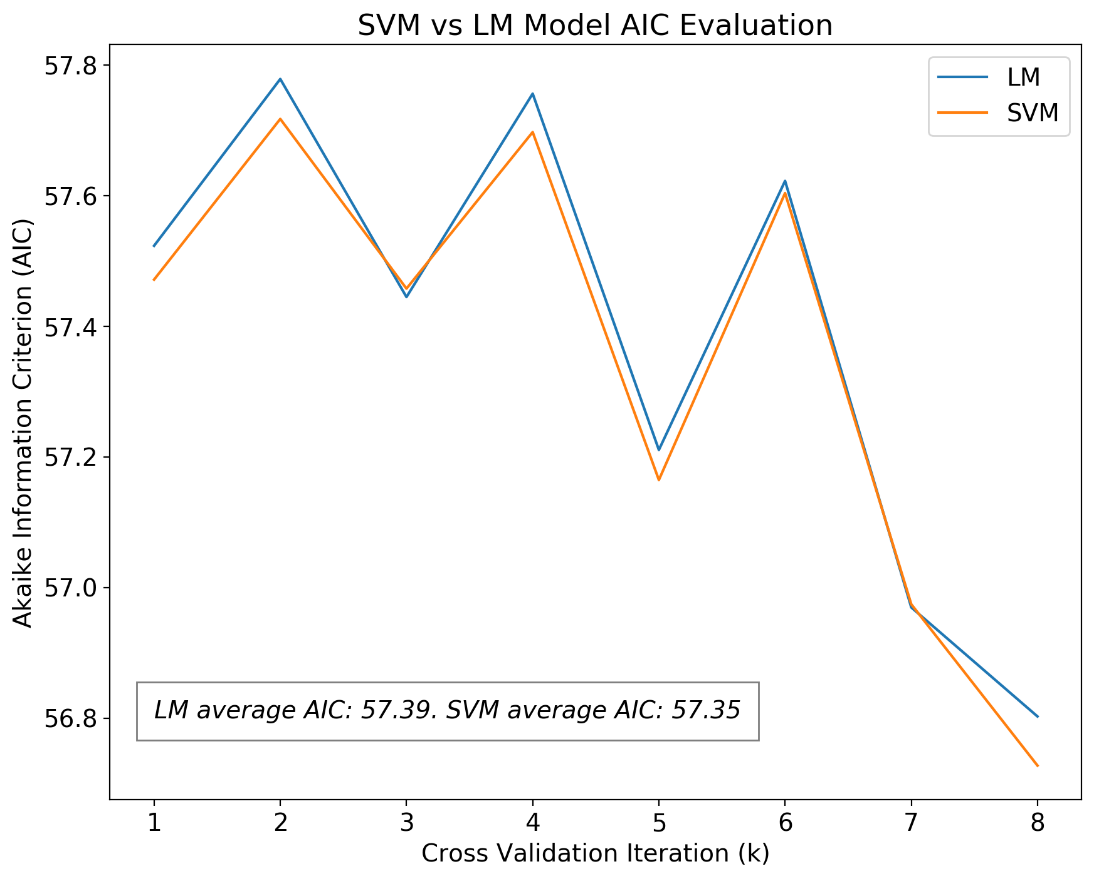
*Parameters Used to Tune Support Vector Machine*

|  |  |
| --- | --- |
| Tuning Parameter | Argument |
| Kernel | 'rbf' |
| Tolerance | 1e-6 |
| C | 1 |
| Epsilon | 0.1 |
| Shrinking | True |
| Max iteration | -1 |
| Gamma | 'auto' |

The first metric which will be evaluated is the AIC, which is an estimate of the predicted distance to the actual observation likelihood of a model. Effectively, the lower the AIC, the better a model is at predicting the actual observations (Dziak et al., 2012). Figure 11 shows that during each iteration, the SVM has a consistently lower AIC, which means that it is better at predicting CDR than the LM model.

Figure 11

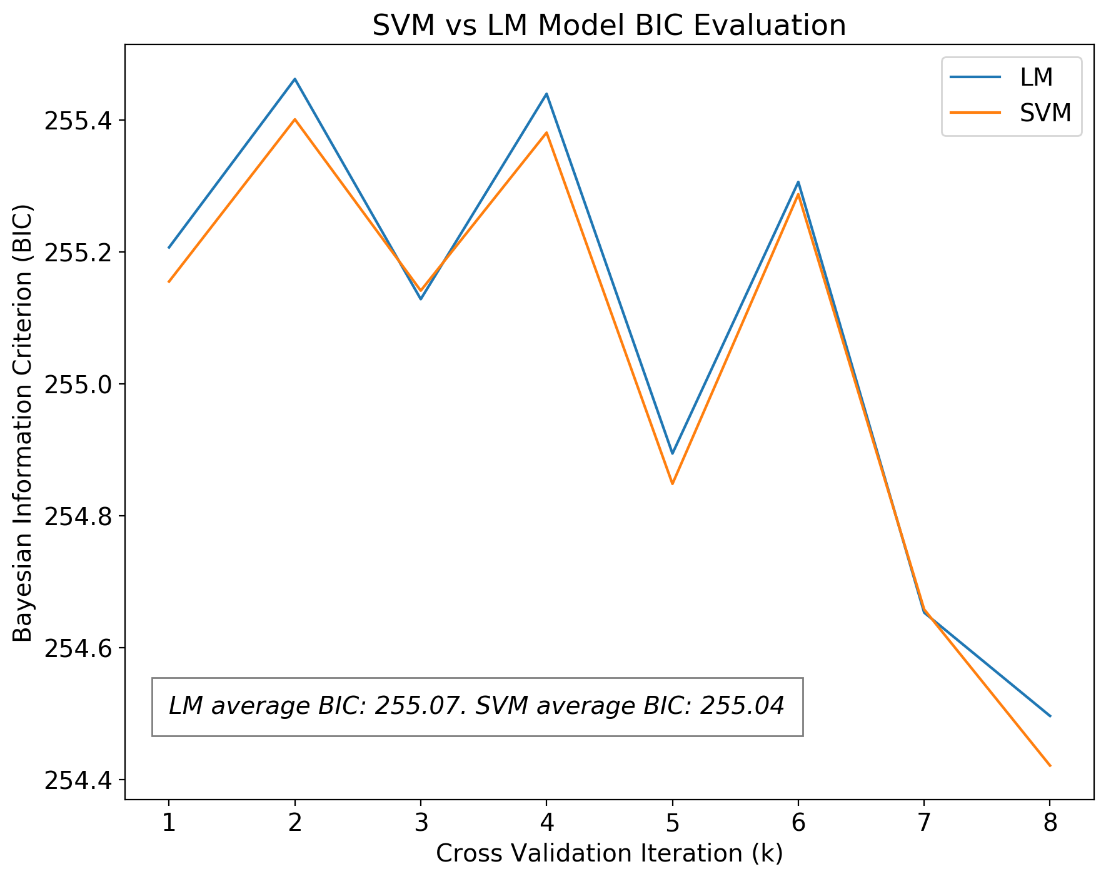
*AIC Model Evaluation*



The BIC is similar to the AIC in that they both evaluate the likelihood of the selected variables making their predictions, which are evaluated using the maximum likelihood. Only by calculating the BIC for each model and selecting the ones with the lowest value, it can be determined which model performs the best. Figure 12 shows that during each iteration, the SVM has a consistently lower BIC, which means that it is more successful at predicting CDR than the LM model.

Figure 12

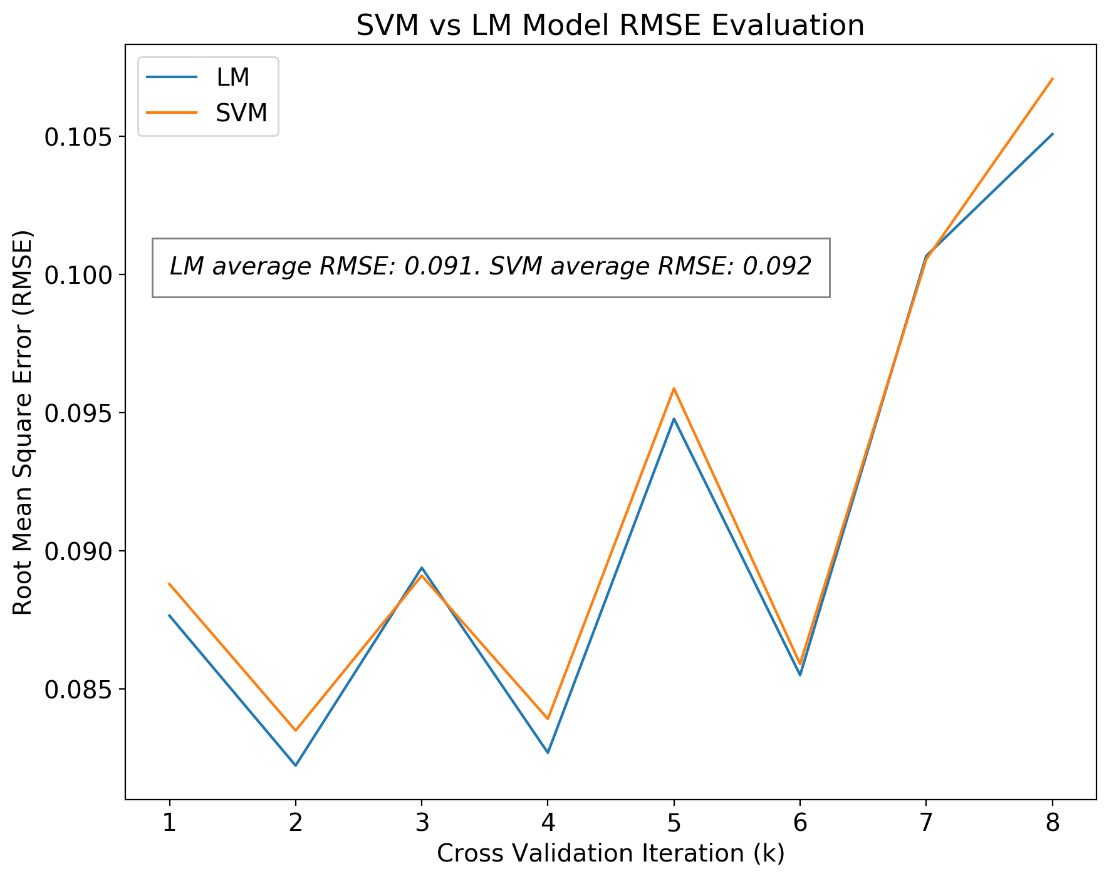
*BIC Model Evaluation*



The Root Mean Squared Error (RMSE) is computed using Formula 1 and describes how well the model fits the data by using the difference between the prediction and observations. The lower the RMSE, the better the model fit. Figure 13 depicts the LM and SVM model performance during the cross-validation training and testing instances. It is shown that the LM performance is slightly better than the SVM model throughout each instance.

Figure 13

*RMSE Model Evaluation*



The final evaluation metric is R2, which is used to measure how well the predicted data fit the regression line of the observed data. It is a percentage of how effective a model prediction is at explaining the dependent variable variation. In this text, R2 is measured from zero to one, with one representing the perfect fit between prediction and observed data. Figure 14 shows that the linear regression model consistently performs better than SVM by an average of 1%.

Figure 14

*R2 Model Evaluation*

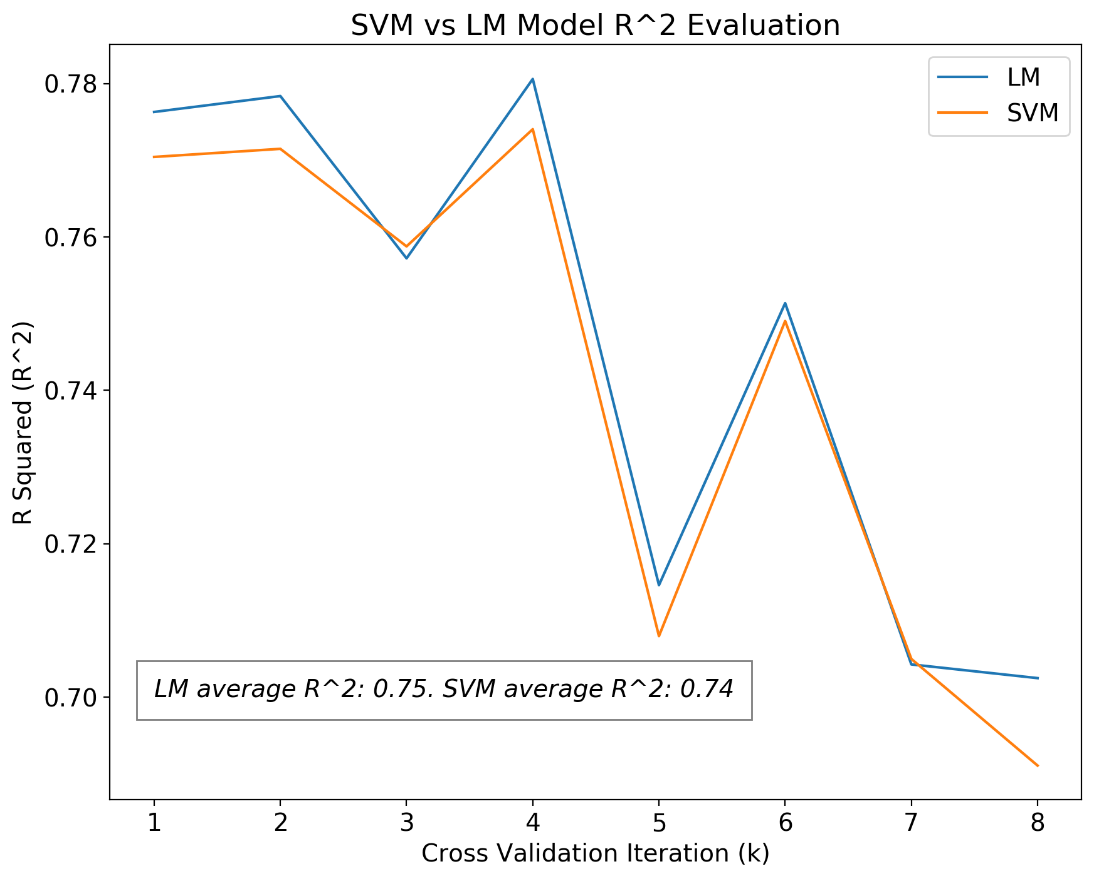


Table 12

*Models Assessment*

|  |  |  |
| --- | --- | --- |
| Metric | SVM | LM |
| RMSE | 0.092 | 0.091 |
| R2 | 0.741 | 0.746 |
| AIC | 57.352 | 57.388 |
| BIC | 255.04 | 255.07 |

Table 12 shows the coefficients and intercept used by the LM model to make its predictions of CDR. The most contributing variables are the previous CDRs, and the cohort of the fiscal year is predicted, which is as expected. By evaluating the race variables, it is shown that the White Race has 48% less CDR than the others, while areas with a high Black, Other, and Native race populace are likely to have greater than 93% higher CDR. All coefficients can be seen in Table 13. Formula 18 represents the formula for the linear regression model.

(18)

Table 13

*LM Coefficients & Intercept*

|  |  |
| --- | --- |
| Variable | Coefficient |
| Intercept | 0.282 |
| CDR 2015 | 0.594 |
| CDR 2014 | 0.302 |
| Cohort 2016 | 0.252 |
| Income Household Median | 0.093 |
| Married | 0.057 |
| HBCU College | 0.029 |
| Proprietary | 0.011 |
| Non-Degree 2 Years | 0.010 |
| Home Ownership | 0.009 |
| Non-Degree 1 Year | 0.004 |
| Non-Degree | -0.001 |
| Private | -0.001 |
| Race Other | -0.005 |
| Race Native | -0.008 |
| First Professional | -0.009 |
| Not Reported College | -0.009 |
| Race Black | -0.013 |
| Bachelors | -0.018 |
| Unemployment Rate | -0.019 |
| Population | -0.021 |
| Race Multiple | -0.022 |
| Masters or Doctors | -0.023 |
| Education College or Above | -0.027 |
| Home Value | -0.043 |
| Race Asian | -0.051 |
| Family Size | -0.093 |
| Race White | -0.093 |
| Age Median | -0.122 |
| Cohort 2014 | -0.268 |

## 4.2 Classification Models

### 4.2.1 Introduction to Classification

When exploring the data, the research team realized that there was a need to develop methods to categorize our instances with the use of a predictive model. A variety of approaches were used to answer two questions. First, do the features of the data lend themselves to describing the school type? Does the data reveal that an instance is indeed a proprietary school without the inclusion of the school type variable? Secondly, if CDR can be binned into two groups, one representing a CDR that is above the median CDR across schools, and one that represents a CDR that is below the median CDR, do the data provide insights that can help classify a school as having a relatively high or low CDR? These questions can be answered by classification methods since the goal is to test group membership. Therefore, two predictive models were envisioned: one that predicts school type as proprietary, and one that predicts whether an institution falls into the high CDR category.

### 4.2.2 Classification Algorithms Chosen

In order to answer the questions posed above, several classification methods were implemented. The algorithms that were chosen were Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine. Tree-based algorithms such as Decision Tree and Random Forest were implemented in the early stages, but these approaches lead to overfitting, even after implementing cross-validation and pruning to prevent this.

### 4.2.3 Classification Performance Measures

Feature selection using Pearson Correlation lead the research group to choose variables 'associates', 'cdr2015\_log', 'density\_log', 'educationcollegeorabove', 'private', 'proglength', 'public', 'raceasian\_log', and 'rentmedian\_log' for use with the proprietary school classification problem. The same variables were used for predicting high CDR, with the addition of the ‘proprietary’ variable. The measurements chosen were accuracy, precision, recall, Cohen’s Kappa, and area under the receiver operating characteristic ROC) curve (AUC). These formulas can be seen in Table 14.

Table 14

*Evaluation Metric Formulas for Classification Models*

|  |  |
| --- | --- |
| Metric | Formula/Definition |
| Accuracy | (19) |
| Precision | (20) |
| Recall | (21) |
| Cohen’s Kappa | (22)  Where:  (Pkyes, 2020) |
| AUC | Given a ROC curve (representation of True Positive Rate [TPR] vs. False Positive Rate [FPR]), AUC is the calculation of the area under that curve. |

Below in Table 15 and 16 are performance metrics for the models. All results were sorted by AUC.

Table 15

*Performance of predict proprietary models with a reduced dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | Kappa | AUC |
| Logistic Regression | 0.9541 | 0.8745 | 1.0000 | 0.8983 | 0.9662 |
| Naïve Bayes | 0.8680 | 0.7321 | 0.9270 | 0.7168 | 0.8836 |
| Support Vector Machine | 0.6477 | 0.4566 | 0.5261 | 0.2222 | 0.6156 |
| K-Nearest Neighbors | 0.4628 | 0.1708 | 0.1757 | -0.2245 | 0.3869 |

Table 16

*Performance of predicting high CDR models with a reduced dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | Kappa | AUC |
| Support Vector Machine | 0.4188 | 0.3928 | 0.2968 | -0.1623 | 0.7350 |
| Naïve Bayes | 0.9307 | 0.9276 | 0.9343 | 0.8613 | 0.6719 |
| K-Nearest Neighbors | 0.2417 | 0.2523 | 0.2628 | -0.5166 | 0.6035 |
| Logistic Regression | 0.6667 | 0.6001 | 1.0000 | 0.3332 | 0.3878 |

### 4.2.4 Interpreting Results of Classification Models

From these results, Logistic Regression performed the best for the prediction of proprietary school type. With an AUC of .9662, this shows that 96.62% of the data lies under the ROC curve, which is the curve generated by the model’s confusion matrix. Formula 23 shows the formula for the logistic regression model.

(23)

However, the inclusion of Cohen’s Kappa shows that the kappa value is high with the predict proprietary logistic regression and naïve Bayes models, implying low inter-rater reliability and high agreement between agreement scores and total scores (Viera, 2005).

K-nearest neighbors in predict proprietary show a Kappa that is very low; however, the performance of the model is slightly more reliable than a coin flip, and therefore not very useful in application.

Figure 15

*Area Under ROC Curves for “Predict Proprietary” Models*

A close up of a map

Description automatically generated

SVM performed adequately for predicting high CDR, with an AUC of .7350. For SVM, we see that the Kappa is very low, at -0.1623, which shows that there is high inter-rater reliability. This model’s results have a low likelihood of being generated by chance. Naïve Bayes, while it had an accuracy of 93%, only showed an AUC of .6719, which demonstrates why accuracy alone is not a reliable measure of performance.

**Figure 16**

*Area Under ROC Curves for “Predict High CDR” Models*

A close up of a map

Description automatically generated

## 4.3 CDR Dashboard

To help institutes understand their CDR status and encourage preventative measures to decrease CDR, the research team has created The Official CDR Dashboard. The dashboard is intractable and focuses on the feature that is defined in Chapter 3.8 by Lasso Regression as being significant to high CDR in an institute. It is placed into three sections, Institution, Geographical, and Socioeconomic. Table 17 shows each respective section as they equate to each view of the dashboard.

Table 17

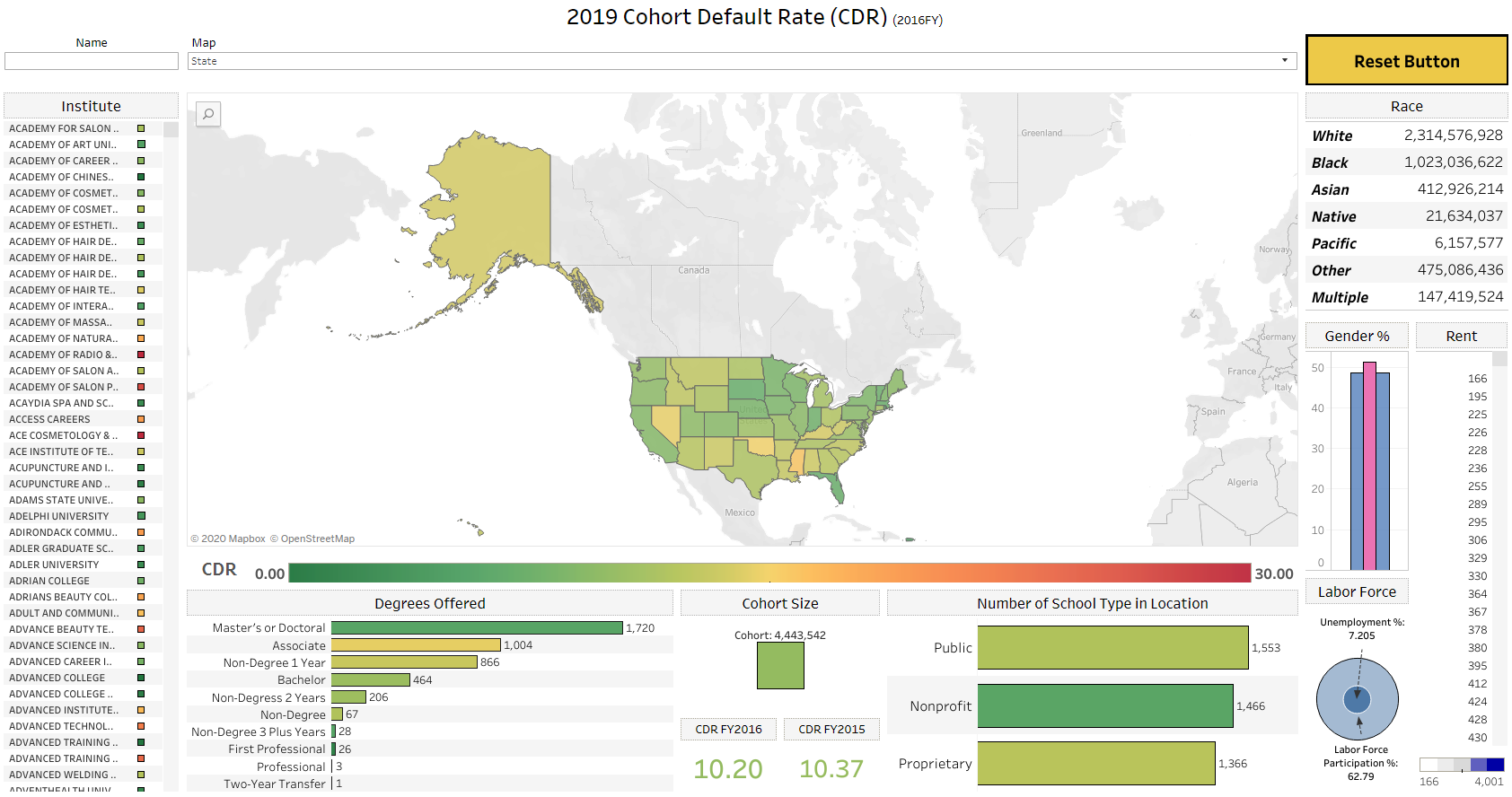
*Tableau Dashboard Sections*

|  |  |  |
| --- | --- | --- |
| Institute | Geographical | Socioeconomic |
| Institute | Map State | Race |
| Degree Offered | Map County | Gender % |
| Cohort Size | Map City | Rent |
| CDR FY |  | Labor Force |
| School Types in Location |  |  |

The Intuition and Geographical sections are colored by the averaged CDR with a color range of green (0) to red (30 or higher). Green Zone represents low CDR, and Red Zone represents areas in default. A general view of the dashboard can be seen in Figure 17.

Figure 17

*2019 Cohort Default Rate (CDR) Dashboard*



The dashboard contains three searchable fields. The first field is “Name,” which allows the user to search for a specific institution by name. The second field is “Map,” which gives the user the option to switch between State, County, and City. The third field is represented by a search icon on the map view that allows the user to search for a specific location. The top right of the dashboard contains a global reset button, which resets back to the main view. The Cohort Default Rate (CDR) Dashboard is published in Tableau Public and can be found here:

https://bit.ly/CDRbyNU

## 4.4 Comparison of Total Proprietary Institutions in High Percentage White Counties

When comparing the total number of proprietary institutions in counties, a threshold was developed to label a county as “High White” or “Not High White.” This threshold was determined by finding the median percent white value for counties. That value was calculated to be roughly 0.7308. This equates to the median “whiteness” of a county is 73.08%. Therefore, all instances with a percent white below 73.08% were labeled not high white, and all instances with a value equal to or higher than 73.08% were labeled high white. This resulted in a split of 1797 schools existing in a high white county and 1794 not existing in a high white county. Mean was not chosen due to the class imbalance created by splitting by mean.

Table 18

*Percentage High White Schools vs. Not High White Schools*

|  |  |  |
| --- | --- | --- |
| highWhite | All Schools | Percentage |
| 1 | 1797 | 50.04% |
| 0 | 1794 | 49.96% |
| Total | 3591 | |

After splitting the data, a grouping function was run on the new dataset. Of schools that were proprietary, the goal was to determine which schools were located in a high white area and which schools were not. This split was less even, demonstrating that there are 46 more proprietary schools located in less-white counties. This was a difference of 4%. See Table 19 and Figure 18 for a comparison.

Table 19

*High White Proprietary Schools vs. Not High White Proprietary Schools*

|  |  |  |
| --- | --- | --- |
| highWhite | Proprietary | Percentage |
| 1 | 552 | 48% |
| 0 | 598 | 52% |
| Total | 1150 | |

Figure 18

*Bar chart comparing the count of proprietary schools in high white vs. not high white counties*

A picture containing screenshot

Description automatically generated

# Chapter 5: Conclusion

## 5.1 Overview of the Research

This study has been conducted with the aim of examining the different issues that are attributed to the cohort default rates in the United States. The main objective of the research was to determine that the rate of defaulting on loans among the students who attend proprietary schools is higher than that of their peers attending public schools and private-not-for profit schools. The team had an objective of proving that minority groups have a higher rate of defaulting than the majority population. While the project aimed to prove the existing hypothesis that minority groups default at a higher rate and proprietary schools have the highest default rates, the team hoped to obtain adequate information on the distribution of CDR by college type, as well as geographical locations of all schools in the U.S. to help the students make better decisions and choices during the selection of colleges. The team conducted their research based on state-level data since it was not possible to obtain student-level data due to privacy reasons. They used different regression models to ensure that the results obtained were closer to the truth as possible.

## 5.2. Interpretation of the Results

The interpretation of the results is based on the hypotheses defined in the first chapter. The team’s null hypothesis (H0) stated that Post-Secondary Institutions with high Cohort Default Rates are located more frequently in U.S. Cities with large minority populations.

The alternative hypothesis (H1) states that the rate of the defaulting on loans among the students who attend proprietary schools is higher than that of the peers attending public schools and private, not-for-profit schools.

The interpretation will be based on the p-values (coefficients) of each of the significant variables selected using the Lasso Regression. The p-value in the case is considered a number between 0 and 1. Based on the CDR prediction formula, the p-value for the CDR is 0.797. The team used the following figures to determine the strength of the relationship:

Table 20

*P-Values and Significance*

|  |  |
| --- | --- |
| **P-value (coefficient)** | **Relationship** |
|  |  |
| 0.0 – 0.19 | Very Weak |
| 0.20 – 0.39 | Weak |
| 0.40 – 0.59 | Moderate |
| 0.60 – 0.79 | Strong |
| 0.80 – 1.0 | Very Strong |

### 5.2.1 Cohort Default Rates (CDR 2014-CDR 2015)

Based on the results, CDR 2014 has a p-value of 0.302, which falls under the moderate category of correlation, while the CDR 2015 has a p-value of 0.594, which falls under the category of strong correlation. This is an indication that an institution that has a high number of defaulters in the first fiscal year has a very high probability of having the same trend in the next fiscal year.

### 5.2.2 Household Income and Marital Status

The LM indicates that the Income Household Median variable has a correlation coefficient of 0.09, which falls in the very weak correlation category. Nevertheless, it shows that this variable has a significant effect on CDR. This generally shows that family income levels have an effect on the probability of students defaulting in the United States.

On the other hand, being married has a significant relation to CDR, owing to its correlation coefficient (0.057). This confirms the assertion made by Miller (2019) that being married (especially females) lowers the probability of default.

### 5.2.3 Institution and Program Length

The team established that the proprietary status of an institution has some correlations with the CDR, but the correlation is weak (0.11). In addition, the length of the program in which a student is enrolled has some form of correlation with the CDR. This is shown by the correlation coefficients of Non-Degree 2 Years and Non-Degree 1 Year programs, which are 0.010 and 0.004, respectively.

The correlation coefficient of private institutions is -0.001 as compared to 0.011 of the proprietary institutions. This is significant as it shows that students who attend private, not-for-profit institutions are less likely to default on their loans than those students who attend proprietary institutions.

### 5.2.4 Ethnicity

Based on the linear regression model, race or ethnicity does not seem to have any significant relation to the cohort default rate. This is because the coefficients for race other, race Native, race Black, race Multiple, race Asian, and race White are -0.005, -0.008, -0.013, -0.022, -0.051, and -0.093 respectively. However, based on the p-values, it is clear that students from the minority groups (Native Americans, African Americans, Asians) are more likely to default than the majority of White students.

Furthermore, colleges that are considered Historically Black Colleges and Universities (HBCU colleges) have a p-value of 0.029, which shows a significant relationship with the cohort default rate (CDR). This affirms that states or regions that have high concentrations of minority populations and HBCU colleges are likely to register high default rates.

## 5.3 Findings

The team managed to carry out different statistical analyses on the three-year cohort starting from the year 2014 to 2016. The regression analysis yielded adequate data from which different conclusions and findings can be drawn with regards to the relationship between different independent variables and the cohort default rate.

The team’s first finding based on Pearson's correlation is that the cohort default rates (CDRs) of each of the years (CDR 2014, CDR 2015, and CDR 2016) have a very strong correlation. This is more understandable and is expected since the default rate for an institution is unlikely to increase or reduce exponentially after a single fiscal year. If this were to happen, it would mean that there are flaws in the analysis technique, data used in the process, or enrollment process of the institution. However, Pearson’s correlation coefficient values show that cohorts do not correlate with CDR. The team noted that the cohort for the year under prediction (cohort 2016) has a significant correlation coefficient (0.252) like the previous years’ CDR.

The team’s second finding proves some of the literature reviewed in chapter two as being right. The team found out that program length has a mild effect on the cohort default rate (CDR). Based on Pearson’s correlation results, the length of a program that a student is enrolled in has a significant effect on his or her probability of defaulting; the longer the program, the lower the probability of default.

Using Pearson’s correlation, the team was unable to determine any correlation between unemployment, rent median, and CDR. However, this is not significant in this research because it was not aimed at determining the economic factor that contributes to CDR. The data used were state level and not student-level, which would make it difficult to determine the unemployment rates and their effect on the CDR. Additionally, the team found out that having a home slightly affects the cohort default rate.

Based on the feature selection method used by the team to eliminate all variables that do not have any significance in the prediction, the team was able to determine that race or ethnicity has very weak or no correlation to the CDR. All the race variables showed signs of weak negative correlation to the CDR. However, from the correlation coefficients presented by the Linear Regression model, the team was able to determine that minority groups have a higher probability of defaulting on their loans as compared to white students.

The team found out that types of the institution had an effect on the cohort default rate. It was clear, from the correlation coefficients, that students who attend proprietary schools have a higher probability of defaulting on their loans as compared to students from public and private, not-for-profit institutions. Furthermore, students who attend schools or institutions that are considered Historically Black Colleges and Universities (HBCU) have a high probability of defaulting on their loans.

Finally, it was evident that Income Household Median has a slight relationship with CDR. This means that household incomes have some effect on the probability of a student defaulting on a loan in the first year of enrollment or the subsequent years.

## 5.4 Research Limitations

The biggest limitation of the study is that the Official Cohort Default Rates (CDR) for Schools (2019) dataset does not offer student-level information, which meant that the team had to rely on county demographics data for their prediction of the cohort default rates. Relying on this level of information to get a rough estimate of the distribution of students may not provide accurate information because there is a group of students who may not be considered (those taking online courses and do not live within the county) in the dataset. There is a high probability of getting discrepancies between variable distribution, such as the level of student income and that of the schools.

The team had to conduct numerous steps of data organization and processing because most of the variables used in the dataset were not significant in predicting the cohort default rate. Of all the data present in the dataset, only 30 variables were found to be significant. Finally, there were other factors that were not included in the dataset as predictors. For instance, disasters were not used as a predictor of the probability of a student defaulting on the loan.

## 5.5 Conclusions Drawn

The results from the analysis point to the fact that both hypotheses have to be accepted in this case. Based on the Linear Regression model, the team accepts that Post-Secondary Institutions with high Cohort Default Rates are located more frequently in U.S. Cities with large minority populations. This is based on the analysis and interpretation made under section 5.2.4. Furthermore, evaluation of the race variables points out that the white students have 48% less probability of defaulting on the loans than the other races, while Black students and other native populations have a 93% probability of defaulting on their loans.

The team accepts the second hypothesis (H1) that states that the rate of the defaulting on loans among the students who attend proprietary schools is higher than that of their peers attending public schools and private, not-for-profit schools. This conclusion is based on the interpretation made under section 5.2.3. Generally, the team concludes that the highest default rates are in the counties that have a higher saturation of minority groups, proprietary schools, and Historically Black Colleges and Universities. A combination of any of the two factors would register the higher default rate per institution.

The findings of this research can be relied on to draft a comprehensive lending plan for the institutions that are found to be highly at risk of having defaulters above the set limit. The plan would work to ensure that these institutions coordinate with their students to come up with a better repayment plan that is both beneficial to the student and the school. This will ensure that many schools escape the penalties due to the high number of defaulters.

## 5.6 Recommendations

There are numerous gaps that exist in this area of research that call for different policy adjustments, actions from different school communities, and academic researchers. If all these groups come together to work on these issues, all the existing gaps will be adequately filled.

### 5.6.1 Policy Reforms

The first policy reforms that the U.S. Department of Education could make is to develop a National Student Loan Data System that is more user-friendly to Support institutional administrators and staff who have the urge and will to understand and reduce default rates (Commisso, 2017). Through the system, the U.S. Department of Education will have the ability to offer guidance to colleges on different options for managing student debt, as well as the different techniques of preventing delinquency and default.

The Federal Government can create policy reforms that could improve the general administration of CDR challenges and appeals (Soldner, & Campbell, 2017). This will be vital as a lifeline for colleges that are on the verge of facing sanctions. This reform should be made because the current implementation of challenges and appeals appears to be discouraging many players or stakeholders from participating in the federal student loan program.

Currently, there are no policies that exist to improve the entrance and exit counseling. An improvement is needed to ensure that the entrance and exit counseling are offered to all the students who qualify for federal student loans (Hillman, 2016). Additionally, the current policy does not offer the information that is presented via loan counseling in a timely manner, which means that this is another gap that needs to be filled via policy reforms.

The Federal Government should create a policy that would redesign financial aid to offer better support for community colleges that comprises the highest number of loan defaulters. The policy should ensure that all the financial aid offered to students in community colleges are at subsidized interest rates.

### 5.6.2 Future Research

Future research should focus on a Student Default Risk Index (SDRI) that can be used for college accountability. This is an area that deserves more attention since the CDRs are not sufficient enough because they tend to exclude data on the students who do not borrow. Thus, it fails to contextualize the entire scope of the default problem in post-secondary institutions. It is a general belief that if SDRI is examined comprehensively through research and developed appropriately, it will convey an accurate student’s risk of defaulting. An SDRI will ensure that the federal aid dollars are spent wisely.

# Appendix A

**Table A1:**

*Variables from the Recursive Feature Elimination*

|  |
| --- |
| 1. familysize\_log; agemedian; incomehouseholdmedian\_log; density\_log; raceblack\_log; educationcollegeorabove; population\_log; laborforceparticipation\_sqrt; cdr2014\_log; raceasian\_log; cdr2015\_log; female\_sqrt; male\_reciprocal; incomehouseholdsixfigure\_log; proglength; homevalue\_log; homeownership; cohort2015\_log; cohort2014\_log; racemultiple\_log; married; racenative\_reciprocal; |

**Table A2:**

*Variables selected using Backward Elimination*

|  |
| --- |
| 'agemedian'; 'cdr2014\_log', 'cdr2015\_log'; 'cohort2014\_log'; 'cohort2016\_log'; 'familysize\_log'; 'homevalue\_log'; 'incomehouseholdmedian\_log'; 'married'; 'proglength'; 'raceasian\_log'; 'racewhite'; 'private'; |

**Table A3:**

*Variable selected using Lasso Regression (L1 Regularization)*

|  |
| --- |
| 'proprietary', 'agemedian', 'cdr2014\_log', 'cdr2015\_log', 'cohort2014\_log', 'cohort2016\_log', 'educationcollegeorabove', 'familysize\_log', 'homeownership', 'homevalue\_log', 'incomehouseholdmedian\_log', 'married', 'population\_log', 'raceasian\_log', 'raceblack\_log', 'racemultiple\_log', 'racenative\_reciprocal', 'raceother\_log', 'racewhite', 'unemploymentrate\_log', 'bachelors', 'firstprofessional', 'mastersordoctors', 'nondegree', 'nondegree1year', 'nondegree2years', 'private', 'hbcuCollege', 'notReportedCollege' |

**Table A4:**

*Variables selected using Pearson Correlation*

|  |
| --- |
| 'associates', 'cdr2015\_log', 'density\_log', 'educationcollegeorabove', 'private', 'proglength', 'proprietary', 'public', 'raceasian\_log', 'rentmedian\_log' |

# Appendix B

In this study, Github was used to organize iPython notebooks as well as raw data, figures, and tables. Please see the link below for access to the Github repository:

https://bit.ly/2YtFSw3

# References

Armona, L., Chakrabarti, R., & Lovenheim, M. (2019). How does for-profit college attendance

affect student loans, defaults and labor market outcomes? *Federal Student Aid Default Management.* <https://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html>

Black students’ default rate surges amid IDR options. (2019). *The College Post.*

<https://thecollegepost.com/student-loan-default-rates/>

Black Youth. *Johns Hopkins University News Releases.*

<https://releases.jhu.edu/2016/09/15/for-profit-trade-schools-prove-costly-for-disadvantaged-black-youth/>

Brady, S., Miller, J., Balmuth, A., D’Ambrosio, L., & Coughlin, J. (2019). Retiring and

caregiving in the age of student loans: the impact of student debt on retirement and longevity planning*. Innovation in Aging, 3(Suppl 1), S42.* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6840048/>

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multimodal inference: A*

*practical information-theoretic approach.* New York: Springer.

Commisso, L. (2017). Identification of Best Practices to Assist Financial Leaders of Proprietary Institutions to Comply with Cohort Default Rate Requirements. *A Phenomenological Study (Doctoral dissertation, Colorado Technical University).* [http://search.proquest.com/openview/95974ad5854b15fd2bbd03a12ffe4e6f/1?pq- origsite=gscholar&cbl=18750&diss=y](http://search.proquest.com/openview/95974ad5854b15fd2bbd03a12ffe4e6f/1?pq-%09origsite=gscholar&cbl=18750&diss=y)

Community College Daily (2019). Is the cohort default rate past its prime? *Community College*

*Daily.* <https://search-proquest-com.nuls.idm.oclc.org/docview/2206726545/fulltext/38807B405B5044E4PQ/1?accountid=25320>

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R. (2017). *Sensitivity and specificity of information*

*criteria.* doi:10.7287/peerj.preprints.1103

Elementary Statistics for the rest of us! (n.d.). Retrieved April 19, 2020, from

https://www.statisticshowto.com/

Fain (2018). Digging Deeper on Student Loan Default Rate. *Inside Higher Ed*.

<https://www.insidehighered.com/news/2018/06/22/big-racial-and-sectoral-gaps-student-loan-default-rates-mostly-persist-even-after>

Farrington (2018). The Growing Culture of Student Loan Defaulters Fighting the System with

Strategic Default. *Forbes.* <https://www.forbes.com/sites/robertfarrington/2018/09/25/student-loan-defaulters-strategic-default/#292de88f229c>

Federal Reserve Bank of New York (2017). Household Debt Surpasses its Peak Reached During

the Recession in 2008. Federal Reserve *Bank of New York Press Releases*. <https://www.newyorkfed.org/newsevents/news/research/2017/rp170517>

Fuller, J. T. W. (2019). Impact of the Financial Aid Process on College Choice for Middle-I

Income Post-Secondary Students: A Human Capital Theory Study*Doctoral dissertation, California Lutheran University.* <http://search.proquest.com/openview/2c3d15cc1c37e167314f391fe265e290/1?pq-origsite=gscholar&cbl=18750&diss=y>

Ghandi, R. (2018) Support Vector Machine – Introduction to Machine Learning Algorithms.

*Towards Data Science.* [*https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47*](https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47)

Hillman, N. W. (2016).Designing and assessing risk-sharing models for federal student aid. Working Paper*.* Wisconsin Center for Advancement of Postsecondary Education, *Madison, WI*. [https://www.luminafoundation.org/files/resources/Designing%20and%20Assessing%20R isk-Sharing%20Models.pdf](https://www.luminafoundation.org/files/resources/Designing%20and%20Assessing%20R%09isk-Sharing%20Models.pdf)

Houle (2013) Disparities in Debt: Parents’ Socioeconomic Resources and Young Adult Student

Loan Debt. *Sociology of Education.* <https://journals.sagepub.com/doi/10.1177/0038040713512213>

Johns Hopkins University (2016) For-Profit Trade Schools Prove Costly for Disadvantaged

Joint Center for Housing Studies. (2019, March 6). Housing Americas Older Adults 2018.

https://www.jchs.harvard.edu/housing-americas-older-adults-2018

Kesterman, F. (2016). Student Borrowing in America: Metrics, Demographics, Default Aversion

Strategies.*Journal of Student Financial Aid, 36(1), 34-52.*

<https://eric.ed.gov/?id=EJ965800>

Kumar, D. (2019) Demystifing Support Vector Machines. *Towards Data Science*.

<https://towardsdatascience.com/demystifying-support-vector-machines-8453b39f7368>

Lau, C. V. (2020). Are federal student loan accountability regulations effective? *Economics of*

*Education Review, 75, 101957.*

<https://www.sciencedirect.com/science/article/pii/S0272775719303796>

Learning Decision Trees (n.d.) Learning Decision Trees. *Williams College*.

http://www.cs.williams.edu/~andrea/cs108/Lectures/Learning/DecTrees.pdf

Looney, A. (2019). Accountability in higher education after deregulation*. Brookings.*

<https://www.brookings.edu/blog/up-front/2019/02/12/accountability-in-higher-education-after-deregulation/>

Lovric, M. (2011). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg:

Springer-Verlag Berlin Heidelberg.

Lundgren, J. (2017). The Effect of Changing Unemployment Rates on Student Loan Cohort

Default Rates *Doctoral dissertation, Georgetown University.* <https://repository.library.georgetown.edu/bitstream/handle/10822/558651/Lundgren_georgetown_0076M_12212.pdf?sequence=1&isAllowed=y>

Macy, A., & Terry, N. (2017). The determinants of student college debt.*Southwestern Economic*

*Review, 34, 15-25.*

<http://swer.wtamu.edu/sites/default/files/Data/15-26-49-178-1-PB.pdf>

MathWorks (n.d.) Lasso and Elastic Net. *MathWorks Help Center*.

https://www.mathworks.com/help/stats/lasso-and-elastic-net.html

McWhirter (n.d.). Recovering from a Natural Disaster in College. *Affordable Colleges Online.*

<https://www.affordablecollegesonline.org/college-resource-center/natural-disasters/>

Miller, B. (2019). The Continued Student Loan Crisis for Black Borrowers. Center for American

Progress.

<https://www.americanprogress.org/issues/education-postsecondary/reports/2019/12/02/477929/continued-student-loan-crisis-black-borrowers/>

Mitchel, J. & Fuller, A. (2019). The Student-Debt Crisis Hits Hardest at Historically Black

Colleges. *The Wall Street Journal.* <https://www.wsj.com/articles/the-student-debt-crisis-hits-hardest-at-historically-black-colleges-11555511327>

Mueller, H. M., & Yannelis, C. (2019). The rise in student loan defaults.*Journal of Financial*

*Economics, 131(1), 1-19.* <https://www.sciencedirect.com/science/article/pii/S0304405X18302009>

Mukherjee, M., Luna-Torres, M., McKinney, L., Shefman, P. K., Wade, J., & Breed, R. (2017). Redesigning Financial Aid to Better Support Community College Borrowers*. Journal of Applied Research in the Community College, 24(1), 27-41.* [https://www.ingentaconnect.com/content/montezuma/jarcc/2017/00000024/00000001/art 00004](https://www.ingentaconnect.com/content/montezuma/jarcc/2017/00000024/00000001/art%0900004)

OECD (2008) Higher Education to 2030. *Higher Education to 2030, Vol. 1.*

Office of Federal Student Aid (N.D.) Cohort Default Rate Guide – Part 2 General Information.

*Federal Student Aid.* <https://ifap.ed.gov/dm/cdrguidepart2>

Office of Federal Student Aid (2019). Federal Student Loan Portfolio. *Federal Student Aid.*

<https://studentaid.gov/data-center/student/portfolio>

Perna, L. W., Kvaal, J., & Ruiz, R. (2017). An updated look at student loan debt repayment and

default.*Penn Wharton Public Policy Initiative, 46.* <http://www.academia.edu/download/53808212/Perna__Kvaal__Ruiz_-_Penn_Wharton_PPI_-_Loans.pdf>

Podgursky, M., Ehlert, M., Monroe, R., Watson, D., & Wittstruck, J. (2017). Student loan

defaults and enrollment persistence.*Journal of Student Financial Aid, 32(3), 27-42*. <https://www.researchgate.net/profile/Michael_Podgursky/publication/228768951_Student_loan_defaults_and_enrollment_persistence/links/02e7e53c9325f227bc000000.pdf>

Pykes, K. (2020) Cohen’s Kappa. *Towards Data Science*.

<https://towardsdatascience.com/cohens-kappa-9786ceceab58>

Raschka, S. (n.d.) What is the difference between filter, wrapper, and methods for feature

selection? <https://sebastianraschka.com/faq/docs/feature_sele_categories.html>

Sedgwick, P. (2012). Pearsons correlation coefficient. *Bmj*, *345*(Jul041). doi: 10.1136/bmj.e4483

Scott-Clayton, J. E. (2018).The looming student loan crisis is worse than we thought. *Brookings*.

<https://www.brookings.edu/research/the-looming-student-loan-default-crisis-is-worse-than-we-thought/>

Scott-Clayton, J., & Li, J. (2016). Black-white disparity in student loan debt more than triples

after graduation.*Economic Studies, Volume 2 No. 3.* <https://www.brookings.edu/research/black-white-disparity-in-student-loan-debt-more-than-triples-after-graduation/>

Sen. Lamar Alexander (2019). Higher Education Accountability. *Senate Committee on Health,*

*Education, Labor, & Pensions.* <https://www.alexander.senate.gov/public/_cache/files/cfd3c3de-39b9-43dd-9075-2839970d3622/alexander-staff-accountability-white-paper.pdf>

Soldner, M., & Campbell, C. O. L. L. E. E. N. (2017). Using and improving federal student aid data systems to support policy analysis*. Resource document. Institute for Higher Education Policy.* [http://www.ihep.com/sites/default/files/uploads/postsecdata/docs/resources/using\_and\_im proving\_fsa\_data\_systems.pdf](http://www.ihep.com/sites/default/files/uploads/postsecdata/docs/resources/using_and_im%09proving_fsa_data_systems.pdf)

Support Vector Machine - Regression (SVR). (n.d.). Retrieved April 19, 2020, from

<https://www.saedsayad.com/support_vector_machine_reg.htm>

Viera, A. (2005) Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine.*

<http://web2.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf>

Volkwein, J. F., Szelest, B. P., Cabrera, A. F., & Napierski-Prancl, M. R. (2016). Factors

associated with student loan default among different racial and ethnic groups.*The*

*Journal of Higher Education, 69(2), 206-237.*

<https://www.tandfonline.com/doi/abs/10.1080/00221546.1998.11775133?journalCode=>uhej20

Webber, D. A. (2017). Evaluating the Costs and Benefits of a Federal Risk-Sharing Program.

https://www.wm.edu/sites/socialmobility/\_documents/session\_iii\_webber.pdf

Woo, J. M. H. (2017). Clearing accounts: The causes of student loan default*. EdFund.*

Zhou, Q., Chen, W., Song, S., Gardner, J., Weinberger, K., Chen, Y. (2015) A Reduction of the

Elastic Net to Support Vector Machines with an Application to GPU Computing. *Tsinghua National Laboratory for Information Science and Technology*. <http://www.cs.cornell.edu/~kilian/papers/aaai15_sven.pdf>