Predictive Analysis on The Relationship Between Student Loan Default Rate, University Statistical and Demographic Data

MIS-690 Capstone Project Thesis

Submitted to Grand Canyon University

Graduate Faculty of the Colangelo College of Business

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Business Analytics

by

Samson Feyisetan, Chiti Nkhuwa, Tahiro Bano, Koffi Serge Ndri

Phoenix, Arizona

July 20th, 2022

Approved by:

Professor Date

Table of Contents

List of Tables

List of Figures

Business Problem Identification

In the United States, educational institutions have opportunities to get funding from the Federal Government. The criterion might vary based on the type of funds and the location of the universities but some of the funding is based on academic performance. According to Kolodner (2019), schools can lose eligibility for federal financial aid if they have a default rate of 30 percent or higher for three years in a row or if their rate hits 40 percent for one year then a loss can be triggered.

Background

Founded in January 2020, STCK is a data consulting firm situated in Phoenix, Arizona. As a consulting company, STCK has helped different organizations and small companies solve their issues using analytical strategies. STCK's data scientists have more than 5 years' experience which makes them capable of conducting in-depth analysis. This makes STCK service quite attractive to institutions and companies with no analytics-based team. In June, the company was contracted by an association of universities to help improve their funding by identifying factors that contribute towards their default rates. Institutions require STCK's services because when an institution has a high default rate it becomes at risk of losing their funding. Therefore, they want us to analyze their statistical data and see what affects student loan default rate.

This problem has been previously encountered. Schools risk losing eligibility for federal financial aid if they have a default rate of 30 percent or higher three years in a row or if their rate hits 40 percent for one year then a loss can be triggered (Kolodner, 2019). If the institution crosses these thresholds, they are placed on the U.S. Department of Education's at-risk list, at

which point the institution could lose eligibility to participate in the Federal student loan program and Pell Grant program. Furthermore, if an institution finds itself on the at-risk list, then it must form a default prevention squad to identify factors leading to default among its former students. Due to the costly nature of these squads, not every institution has the tangible and intangible resources to collect data and conduct in-depth analysis that will result in solid recommendations for a plan of action. Therefore, outsourcing investigations to a data consulting company like STCK becomes a more cost-effective option.

The problem has existed for several years. According to Kolodner (2019), the U.S. Department of Education published its annual report looking at student loan default rates in 2019. According to this list, 15 higher education institutions were singled out for rates so high that they were at risk of losing their access to federal student aid. Clearly this is not a rare occurrence and as such an institution's capacity to have an idea of what contributes towards default rate has become exceedingly cardinal. Furthermore, considering that the federal pause on student loan repayments could be lifted soon, new numbers might emerge that will put some institutions at risk. Not knowing the factors that could negatively impact the default rate, the institutions are operating blindly. Through STCK, they will be able to learn about the factors that affect default rate and lower it with a plan of action.

To get information about the issue, STCK plans to interview college administrators that have knowledge of the institution’s demographic data. Since administrators such as department deans and directors manage, lead and work closely with employees that are consistently involved with students and alumni, they are good prospects for STCK to get valuable information. The company will also get information from universities that are currently affected by the high

default rates because they might have prematurely been taking actions to avoid being on the at-risk list and losing the funding. The problem the universities are experiencing can be solved using analytical techniques. STCK formulated the problem so that a model can be created to provide results and create the recommendations for the group of universities.

The necessary data for the problem has been obtained from public data repositories, lendedu.com and kaggle.com. The first dataset that will be utilized for this business problem has been obtained from Kaggle.com, an online community platform for data scientists and machine learning enthusiasts. This first data set lists University rankings and default rates.
The second dataset has been obtained from the US Department of Education. It is known as a college scorecard that serves to provide various data elements that identify and describe unique characteristics about a university, student body demographics and alumni information.

The nature of the analysis is to determine the relationship between default rate and a university's demographic data. The main benefit to the organization will be that STCK will provide the statistical factors that impact the default rate. STCK's model can predict default rates and thus put measures in place to keep it under the danger mark of 30 percent for three consecutive years and 40 percent for one year.

Business Problem Statement

We are predicting default rates using the various institutions' statistical data. We seek to identify factors that are in the institution's control that they can use to potentially combat high default rates and keep them below the at-risk threshold.

Analytics Assumptions

After the universities explained the matter to STCK, the analytics-based assumption is that student loan default rates and a university's statistical data possess a relationship that can be used to predict the former. The next assumption was that knowing the significant factors that impact the default rate can help the institutions to reduce their default rate better to avoid being at-risk.

Through the implementation of the analytics-based solution, the specific outputs that would result from implementation are the institution's continued access to federal student aid such as the Pell Grant. The cost of not implementing the analytics-based solution is an institution's inability to predict default rates among its former students which in turn could affect its eligibility for federal student aid. The direct consequence of this situation is the institution finding itself on the U.S. department's list of at-risk schools. On this list the institution must put costly measures in place to identify the factors that lead to default rate.

There are specific inputs needed to carry out the analysis. The inputs needed are the demographic information, loan, and grant information. The analytics focus area is the use of institutional statistical data, demographics, student financial reports to determine how impactful some factors can be for the default rate. In addition, STCK focuses on analyzing the mentioned variables from these institutions to predict their default rate in the coming years. Institutional data received comes in many different forms that need transformation. There were categorical data then needed some form of transformation to make it usable for analysis.

For this analysis, we will focus on using loan records of students in respective institutions, grants, and other financial aid records for the purpose of this project. Other

important information considered for this analysis include demographic variables of the institution as well as national ranking of the institution. The first dataset was obtained on Kaggle, and the second data was acquired from the US Department of Education. The second dataset had a lot of information about the universities and their students too. As an illustration, a variable such as the number of alumni in a specific income range.

Analytics Problem Statement

The nature of the analysis is to build a model to predict default rate using multiple regression analysis.

Data Understanding, Acquisition, and Preprocessing

We employed a quantitative research approach by collecting relevant financial, and demographic data for this project. The financial data consists of loan information, repayment information, grants and other forms of aid received by students of respective institutions. The demographic information covers the population, location, school type, number of students in different categories of those affected by the financial variables such as loans, and grants mentioned earlier. The data collected contains distinct types of variables which range from discrete, continuous, textual, and binary. Crucial to our analysis is the preparation of the data to make it suitable for the type of analysis we identified, Multiple linear regression. We dedicated a lot of time to data cleaning to ensure that it meets the requirement of a multiple linear regression. The raw data contained a lot of null, privacy suppressed, we decided to use average values of the corresponding variables for the nulls in cases where the null values are less than 10%. We decided to do away with variables having nulls more than 10% in of the data to have a result without much noise.

Sample of data included in the screenshot below.

OPEID6	Institution Name	Ranking	Default Ra	City	State	Zip Code	Number o	CONTROL	Region	Number o	Total Male Stud	Total females Stud	OPENADM	NPT4_PUE	NPT4_PRI
1009	Auburn University	147	2.6	Auburn	AL	36849	1	1	5	24147	7099	7436	2	23696	25000
1579	Augusta University	165	3.5	Augusta	GA	30912	1	1	5	5146	2265	4407	2	13438	25000
3545	Baylor University	153	2.2	Waco	TX	76798	1	2	6	14159	4490	6867	2	10000	32601
2836	Binghamton University	191	1.9	Vestal	NY	13850	1	1	2	13990	6083	6242	2	18958	25000
2128	Boston College	141	0.9	Chestnut	MA	2467	1	2	1	9639	2709	3157	2	10000	33562
2130	Boston University	35	1.4	Boston	MA	2215	1	2	1	17238	4450	8423	2	10000	30729
2133	Brandeis University	90	1.2	Waltham	MA	2454	1	2	1	3627	975	1331	2	10000	37688
3670	Brigham Young Universit	140	1.3	Provo	UT	84602	1	2	7	31441	14255	11894	2	10000	13181
3401	Brown University	36	0.8	Providenc	RI	2912	1	2	1	6752	1597	1769	2	10000	31685

The variables and names of the respective variables are given in later sections of this project.

Collection of Initial Data

The biggest challenge in obtaining this data set is the Privacy Act and the timeline for this project. The initial dataset was collected from the US Department of Education and Kaggle.com. The financial dataset was collected from the US Department of Education, this dataset contains the loan information, grant and other aids information of students, while the dataset obtained from kaggle.com has the default rate, national ranking, and other demographic variables of the respective institutions. We merged these two datasets using the “VLOOKUP” function in excel.

The data obtained from the department of education was a zipped folder containing financial loans and other financial aid information of student from 1996 up to 2021, data collected form kaggle.com only contained information from 2017 to 2018, after searching several resources for the most recent records for default rate which we are considering predicting for the next two years, only 2017 to 2018 records were found. Through further research, we found out that the most recent default rate record available is 2018. We merged this data with other corresponding data for the analysis. Below in table 1, we have listed each variable that was used in the analysis as well as its name, category, and type.

Table 1.1

Variables	Names	Category	Type
OPEID6	Office Of Postsecondary Education ID	School	String
Institution Name	Institution Name	School	Character
Ranking	University Ranking	School	Integer
Default Rate	Historical Default Rates	School	Integer
City	City	School	Character
State	State	School	String
Zip code	Institution zip code	School	String
Number of Campus	Number of branch campuses	School	Integer
Control	Control of institution(Ownership)	School	Integer
Region	Region ID	School	Integer
Number of UGD Stud	Enrollment of all undergraduate students	Student	Integer
Total Male Stud	Total number of male students at each institution	Student	Integer
Total females Stud	Total number of female students at each institution	Student	Integer
OPENADMP	Open admissions policy indicator	School	Integer
NPT4_PUB	Average net price for Title IV institutions (public institutions)	Cost	Integer

NPT4_PRIV	Average net price for Title IV institutions (private for-profit and nonprofit institutions)	Cost	Integer
NUM4_PUB	Number of Title IV students (public institutions)	Cost	Integer
NUM4_PRIV	Number of Title IV students (private for-profit and nonprofit institutions)	Cost	Integer
PCTFLOAN	Percent of all undergraduate students receiving a federal student loan	Aid	Float
LOAN_YR4_N	Number of loan students in overall 4-year completion cohort	Completion	Integer
GRAD_DEBT_MDN	The median debt for students who have completed	Aid	Float
WDRAW_DEBT_MDN	The median debt for students who have not completed	Aid	Float
DEP_DEBT_MDN	The median debt for dependent students	Aid	Float
IND_DEBT_MDN	The median debt for independent students	Aid	Float
PELL_DEBT_MDN	The median debt for Pell students	Aid	Float
NOPELL_DEBT_MDN	The median debt for no-Pell students	Aid	Float

FEMALE_DEBT_MDN	The median debt for female students	Aid	Float
MALE_DEBT_MDN	The median debt for male students	Aid	Float
DEP_DEBT_N	The number of students in the median debt dependent student's cohort	Aid	Integer
IND_DEBT_N	The number of students in the median debt independent student's cohort	Aid	Integer
PELL_DEBT_N	The number of students in the median debt Pell student's cohort	Aid	Integer
NOPELL_DEBT_N	The number of students in the median debt no-Pell students' cohort	Aid	Integer
FEMALE_DEBT_N	The number of students in the median debt female student's cohort	Aid	Integer
MALE_DEBT_N	The number of students in the median debt male student's cohort	Aid	Integer
D_PCTPELL_PCTFLOAN	Number of undergraduate students (denominator percent receiving a Pell grant or federal student loan)	Student	Integer
PLUS_DEBT_INST_MD	Median PLUS loan debt disbursed at this institution	Aid	Integer

PLUS_DEBT_INST_MALE_N	Student recipient count for median PLUS loan debt disbursed to males at this institution	Aid	Integer
PLUS_DEBT_INST_NOMALE_N	Student recipient count for median PLUS loan debt disbursed to non-males at this institution	Aid	Integer
PLUS_DEBT_ALL_NOMALE_N	Student recipient count for median PLUS loan debt disbursed to non-males at all institutions	Aid	Integer
PLUS_DEBT_ALL_PELL_N	Student recipient count for median PLUS loan debt disbursed to Pell recipients at all institutions	Aid	Integer
PLUS_DEBT_INST_NOPELL_N	Student recipient count for median PLUS loan debt disbursed to non-Pell-recipients at this institution	Aid	Integer
PLUS_DEBT_ALL_NOPELL_MD	Median PLUS loan debt disbursed to non-Pell-recipients at all institutions	Aid	Integer
YEAR	Years(2016, 2017 and 2018)	Interval	Interval

Description of Data

The dataset contains forty-three variables across the board, all of which fall under certain categories. School variables that are used to describe the institution and unique quantitative

figures related to that institution such as historical default rates. In addition, we obtained student variables which are qualitative and quantitative descriptions of the student body. Furthermore, we have cost variables in the dataset that quantify how much it costs to attend each private and each public institution. Moreover, we also obtained aid variables that refer to the number of students who received federal student aid such as Pell grants. Finally, the completion variable in the dataset refers to the number of loan students in an overall 4-year completion cohort.

Data Diagnostics and Descriptive Summary

The data was obtained from the department of education, lendedu.com, and Kaggle.com, the first being a government source and the two others (lendedu.com and Kaggle.com) being trusted and well-known data repositories. Dataset obtained from these mentioned sources were merged using excel function “VLOOKUP, then we embarked on data cleaning by transforming the nulls where necessary, variables containing nulls less than 10% of the data were transformed to average value of the respective variables, enough care was taken with the transformation so that results would not produce noise or give an unreliable outcome. Irrelevant variables and variables and rows with null and privacy suppressed making up more than 10% of the variables were carefully removed as they could cause unprecedented errors in the analysis, values could not be assumed for the privacy suppressed due legal implications and consistency's sake in the data, this is done to allow for appropriate data types for the chosen model. The problem of mismatching was encountered when trying to merge the data obtained from two data sources. This was because there were varying special characters in the spelling of names, therefore, adjustments were made to the characters to ensure consistency.

Exploratory Data Analysis

We aggregated the default rate across states for the combined years depicted by the heat map in figure 1.1 shows that Maine had the highest default rate, followed by Idaho and West Virginia while New Hampshire had the lowest default rate in the country. The heat map also has the capacity to show which states had the highest default rates for each individual year instead of an aggregated figure if need be.

In 2017 we discovered that Mississippi, West Virginia, and Washington possessed the highest default rates while Arkansas, Ohio and Indiana possessed the lowest default rates as shown in figure 1.1(a). Finally, in 2018 we discovered that Tennessee, Alaska, and West Virginia, possessed the highest default rates while Arkansas, Massachusetts and Indiana possessed the lowest default rates as shown in figure 1.1(b). Finally, across both years we noticed that West Virginia was consistently in the top three states with the highest default rate, while New Indiana and Ohio were consistently in the bottom three states with the lowest default rates as shown in figure 1.1(c).

Figure 1.1(a)

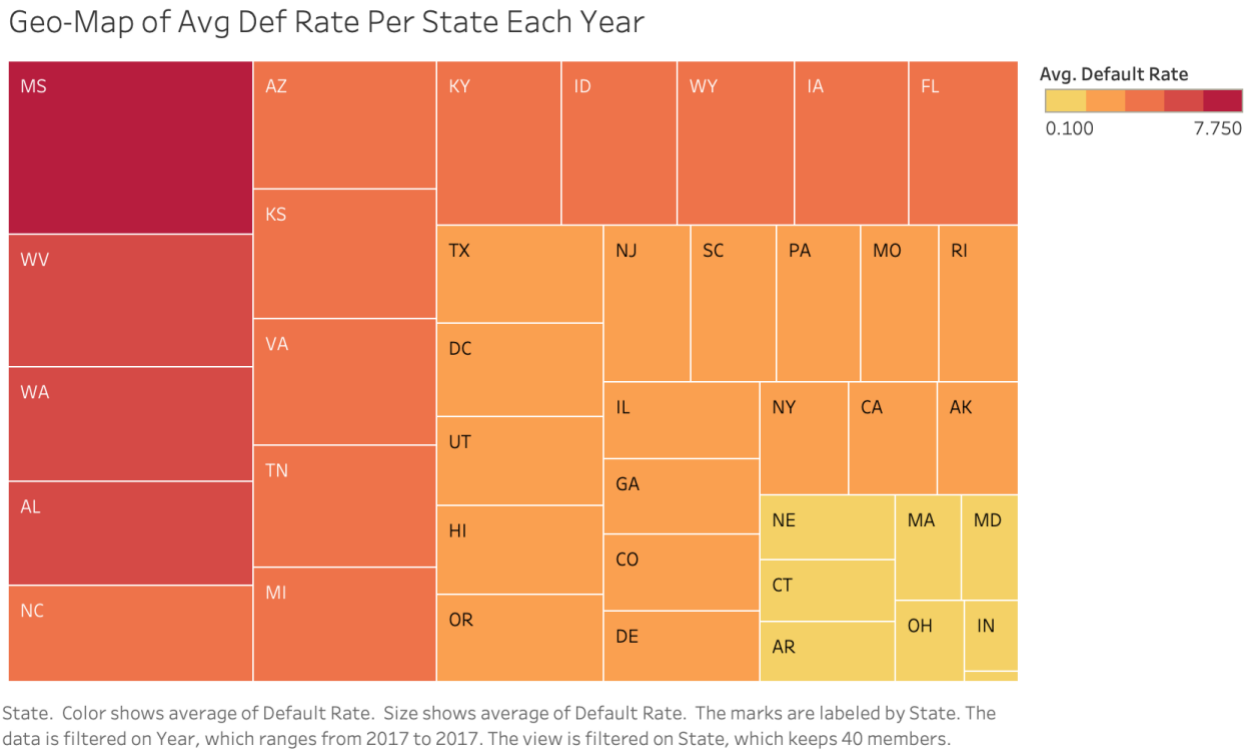

Geo-Map of Avg Def Rate Per State Each Year
MS
AZ
KY
ID
WY
IA
FL
KS
WV
TX
NJ
SC
PA
MO
RI
VA
DC
WA
IL
NY
CA
AK
TN
UT
GA
AL
NE
MA
MD
HI
CO
CT
NC
MI
OR
DE
AR
OH
IN
Avg. Default Rate
0.100
7.750
State. Color shows average of Default Rate. Size shows average of Default Rate. The marks are labeled by State. The data is filtered on Year, which ranges from 2017 to 2017. The view is filtered on State, which keeps 40 members.

Figure 1.1(b)

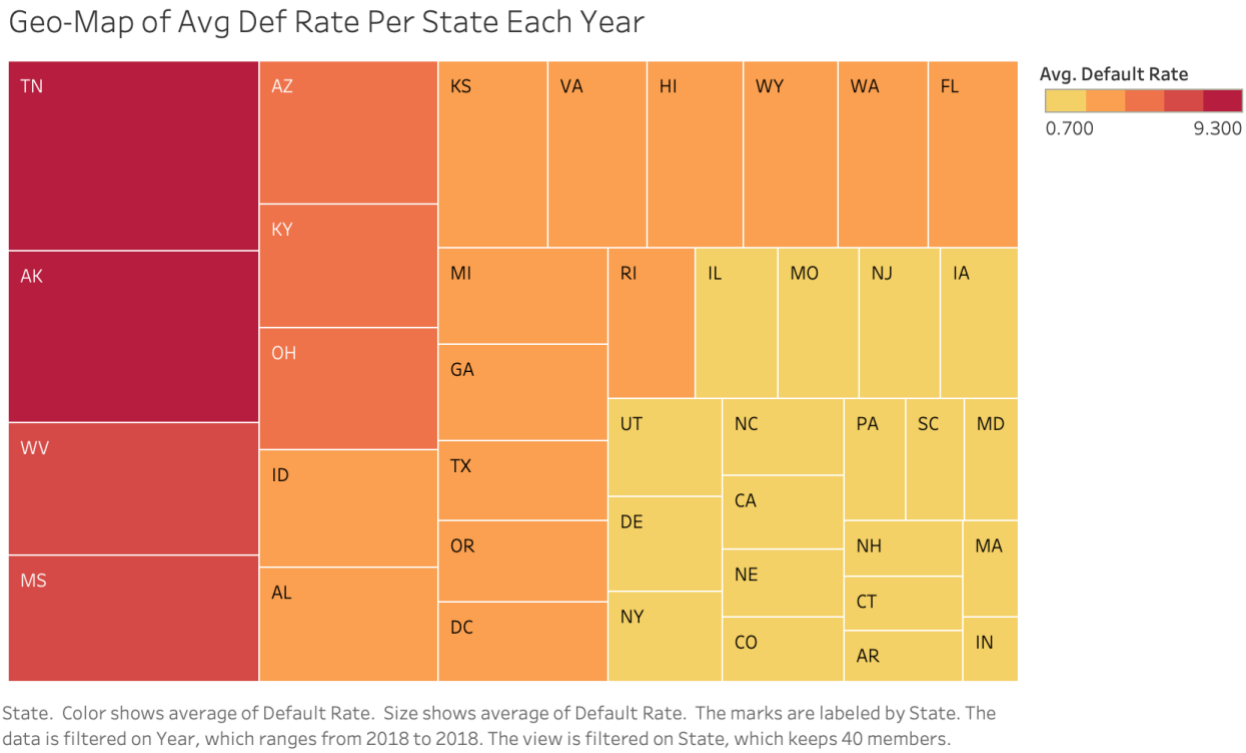
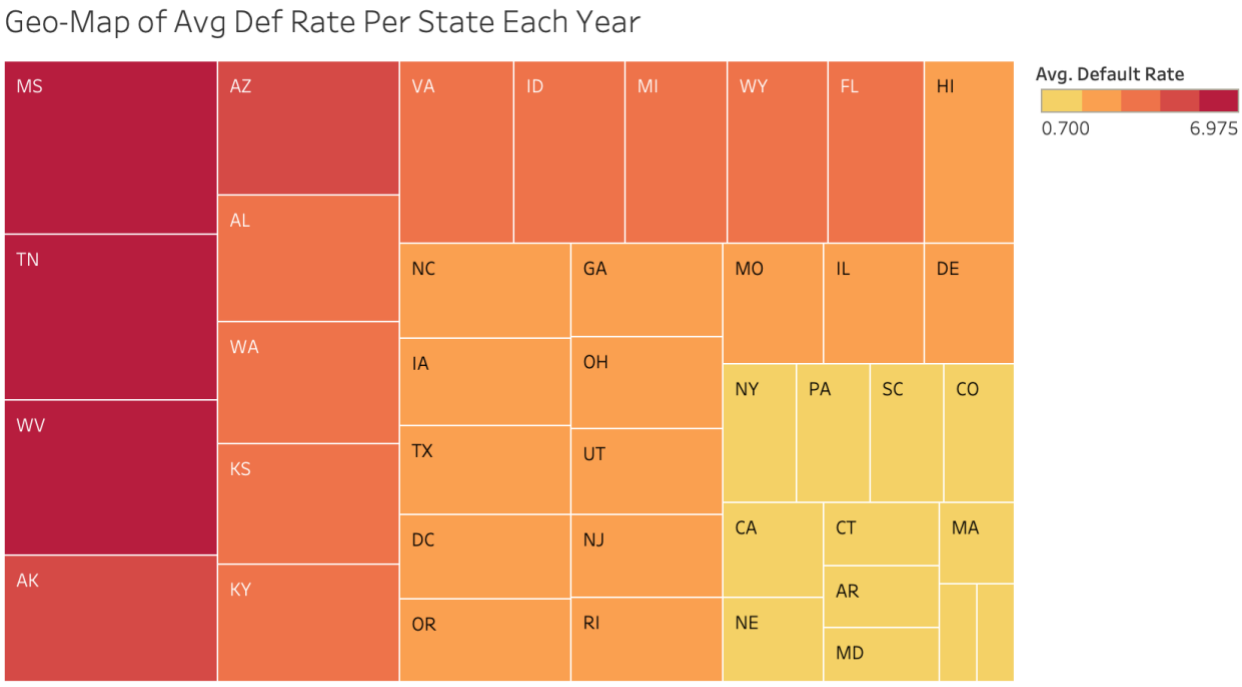

Geo-Map of Avg Def Rate Per State Each Year
TN
AZ
KS
VA
HI
WY
WA
FL
KY
AK
MI
RI
IL
MO
NJ
IA
OH
GA
UT
NC
PA
SC
MD
WV
ID
TX
CA
DE
NH
MA
MS
AL
OR
NE
CT
DC
NY
CO
AR
IN
Avg. Default Rate
0.700
9.300
State. Color shows average of Default Rate. Size shows average of Default Rate. The marks are labeled by State. The data is filtered on Year, which ranges from 2018 to 2018. The view is filtered on State, which keeps 40 members.

Figure 1.1(c)



This relationship required further exploration; therefore, we analyzed default rate at regional level to see how the regions fared against each other. According to the dataset each region possessed a specific designation with the states in those regions being fed into that designation. The region locale table below shows the designation of those regions from zero to nine, with each designation possessing a specific name.

Table 2.2

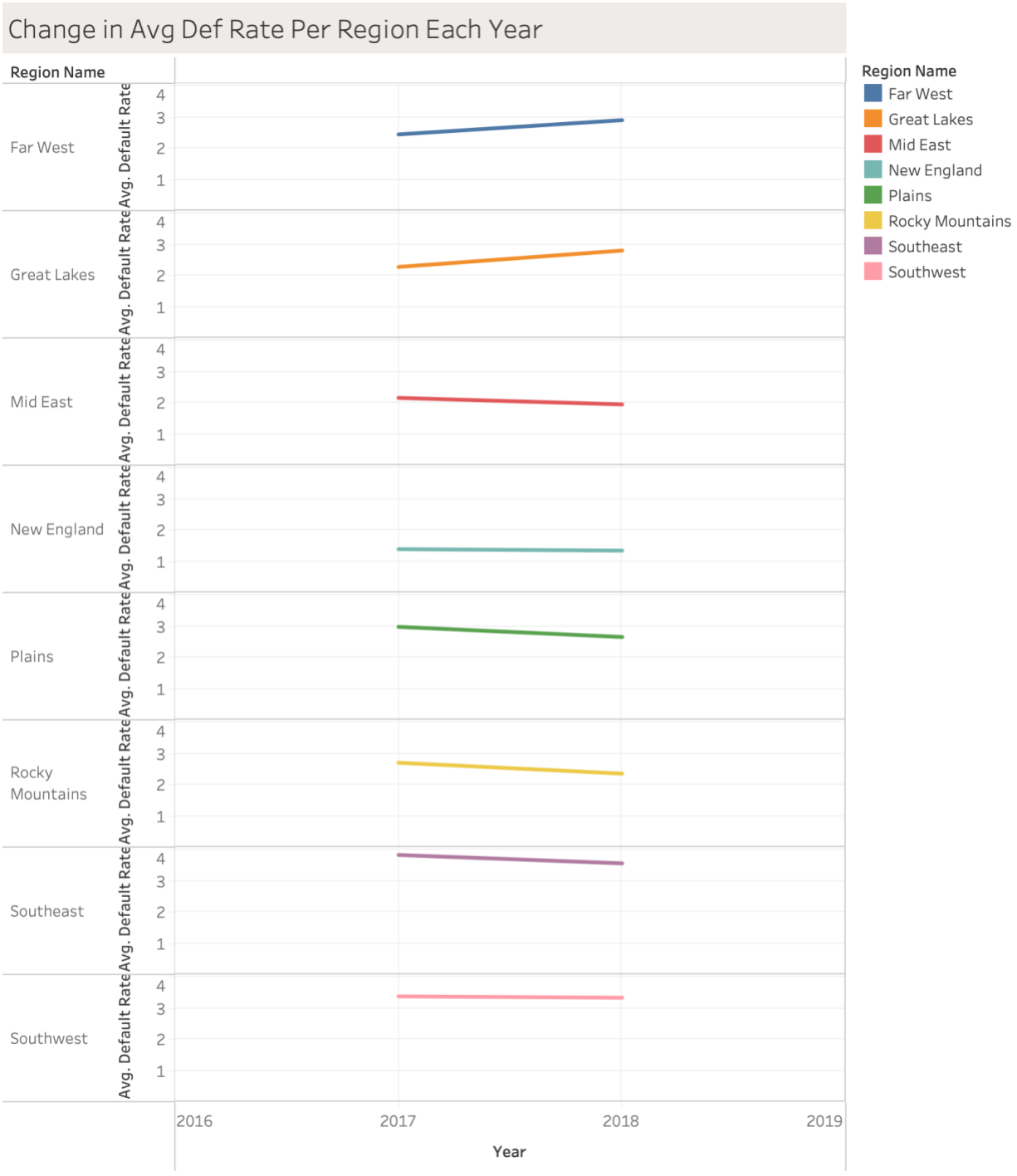
0	U.S. Service Schools
1	New England (CT, ME, MA, NH, RI, VT)
2	Mid-East (DE, DC, MD, NJ, NY, PA)
3	Great Lakes (IL, IN, MI, OH, WI)

4	Plains (IA, KS, MN, MO, NE, ND, SD)
5	Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)
6	Southwest (AZ, NM, OK, TX)
7	Rocky Mountains (CO, ID, MT, UT, WY)
8	Far West (AK, CA, HI, NV, OR, WA)
9	Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)

We then proceeded to explore the change in default rate for each region. It is important to note that the dataset we obtained only possessed universities from region 1 to region 8. Therefore, institutions in the service school's region and the outlying areas region were not part of the analysis.

As shown in figure 1.2, a change in the average of the student loan was aggregated across the 7 regions and the trend shows that the change in average default rates in the Mid-East region, Plains region, Rocky Mountains region, and Southeast region declined from 2017 to 2018, but default rates in the Far West region and Great Lakes region steadily increased from 2017 to 2018. Lastly, the New England region and Southwest region showed little to no change in average default rates at all from 2017 to 2018

Figure 1.2



The trend of average of Default Rate for Year broken down by Region Name. Color shows details about Region Name. The view is filtered on Region Name and Year. The Region Name filter keeps 8 of 8 members. The Year filter ranges from 2017 to 2018.

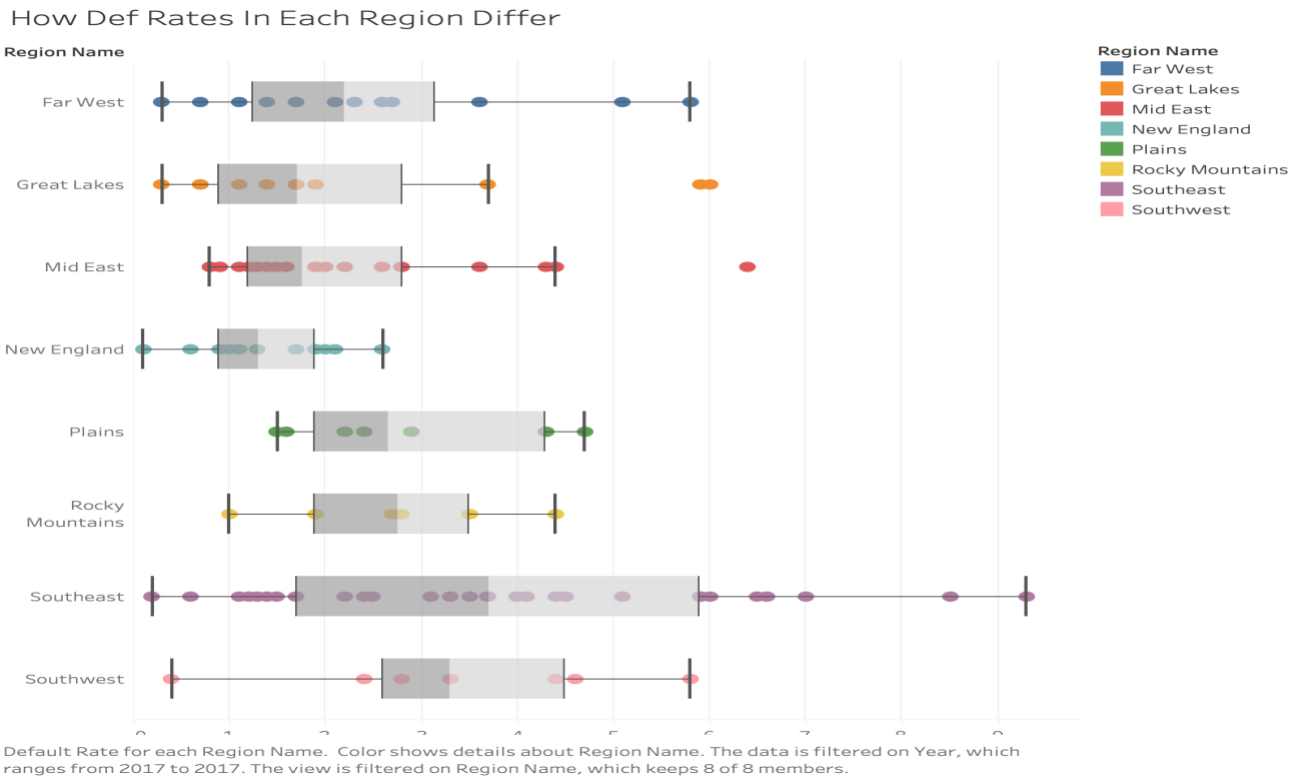
Trends Analysis

We created four box plots to analyze the trend across each year before aggregating the years across the board. As Stated by Galarnyk (2018) box plots can be defined by the median (middle quartile) which marks the midpoint of the data. This is often displayed by the line or shading that divides the box into two parts. The middle "box" which represents the middle 50% of scores for the group. The interquartile is referred to as the range of scores from lower to upper quartile. The lower quartile shows twenty-five percent of scores that fall below the lower quartile whereas seventy-five percent of the scores that fall below are referred to as the upper quartile There are four broad interpretations for box plots namely, an analysis of the median, interquartile range, outliers, and skewness. For each year we opted to apply one analysis each and for the combined box plot which ranges from 2017 to 2018 we applied each of the four analytical points.

A comparison between the medians of each box plot was shown. As stated by McLeod(2019), there is likely to be a difference between the two groups if the median line of a box plot lies outside of the box of a comparison box plot. As shown in figure 1.3(a), The Far West region, Great Lakes Region and Mid East region have relatively similar medians whereas the Plains region and Southeast region are vastly different. This framework was applied to all the year's box plots as a means for comparison. Furthermore, the box plot for 2017 allowed for a comparison between the interquartile ranges of each box and whisker plot. As a rule of thumb, a long box plot would indicate more dispersed the data and a small box plot would indicate less

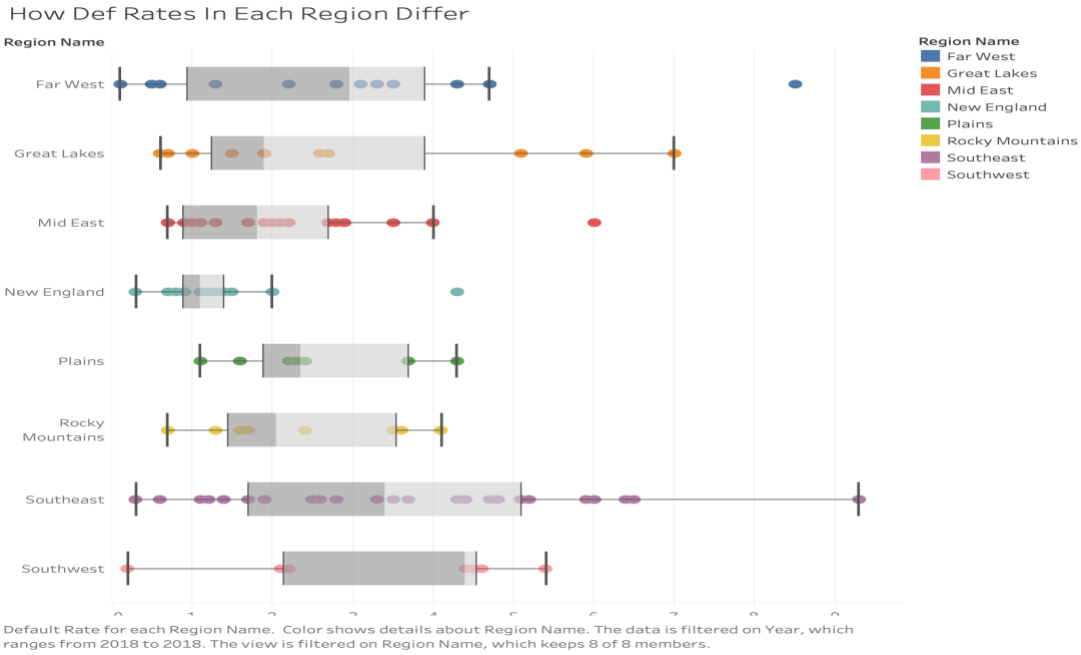
dispersed data. The data in the Southeast region is the most dispersed in comparison to the data in the New England region. Larger ranges indicate wider distribution and more scattered data. We discovered that the Southeast region, Far West region, and the Southwest region have the widest distribution of data. This framework for analysis of data distributions was applied to each year.

Figure 1.3(a)



As shown in figure 1.3(b), the box plot for 2018 we looked for potential outliers in each region. Outliers are the data points that are located outside the whiskers of the box plot. The Far West region, New England region, and Mid East region all possessed outliers.

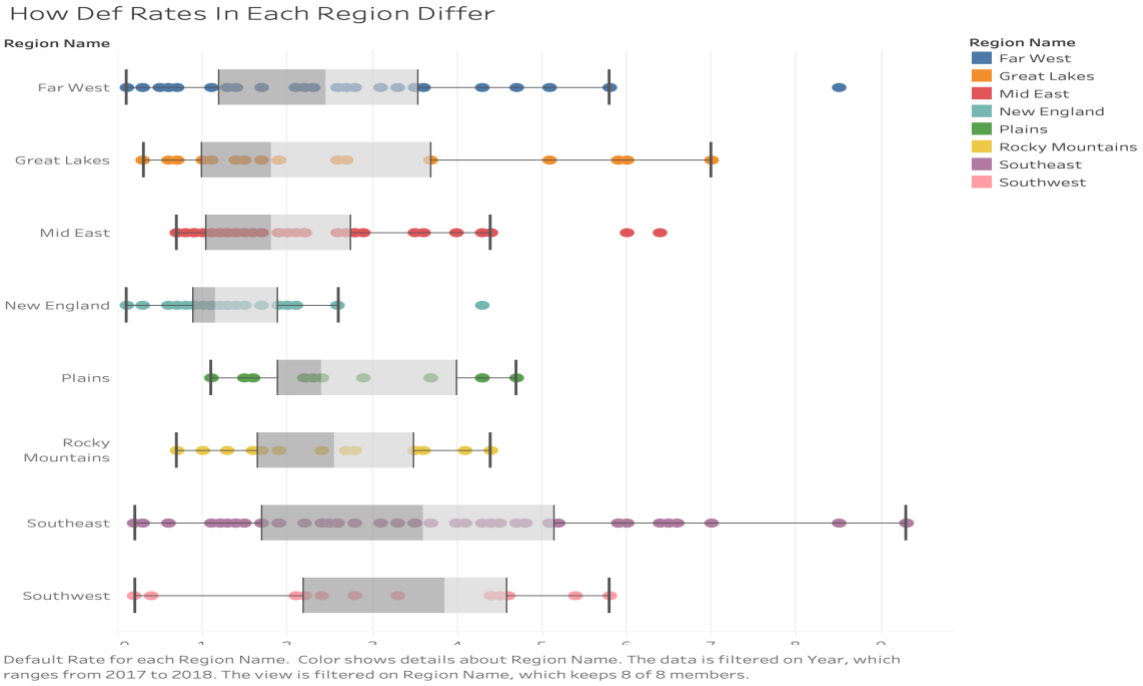
Figure 1.3(b)



As shown in figure 1.3(c), the combined box plot across both years showed that the medians varied across each region. However, a more specific comparison between two or three regions would yield a different result but for the purposes of an overview there is not much that is unusual. However, we discovered that several of the regions possessed data that is either left-skewed or right-skewed. We can evaluate the skewness of the distribution based on the position of the median value from the box plot. Zach(2021), stated that the distribution is skewed to the right when the median is closer to the bottom of the box and the whisker on the lower end of the box is shorter. However, with median appearing closer to the top of the box and a shorter whisker on the upper end of the box, we have a left skewed distribution. Finally, a symmetrical distribution is if we have the median in the middle of the box and roughly equal whiskers are on each side.

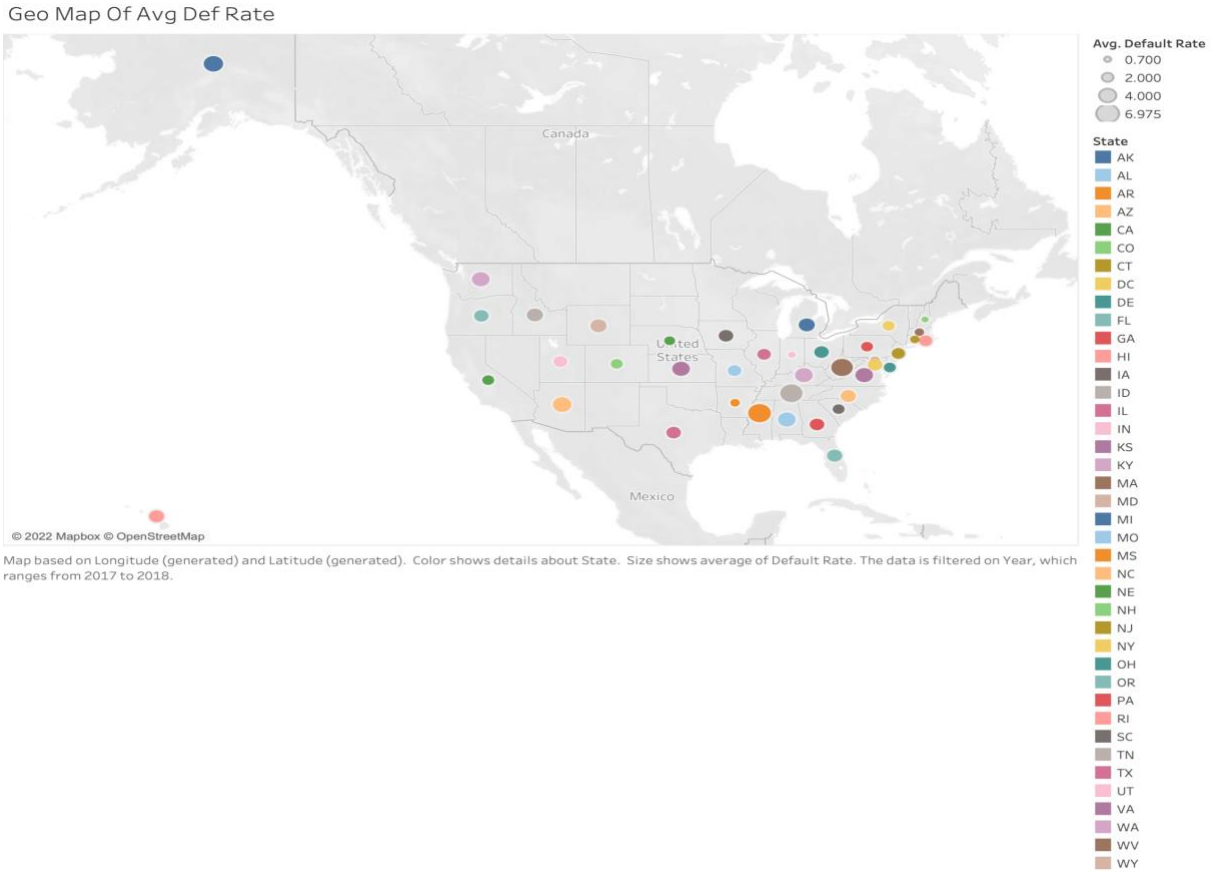
The Plains region, New England region, Great Lakes region and Plains region possessed data that was right-skewed whereas the Southwest region possessed data that was left-skewed. The distribution for the Far West region, Mid East region, Rocky Mountain region and Southeast region is symmetrical.

Figure 1.3(c)



As shown in figure 1.4, the average default rate over two years represented by the size of the bubble in the map chart we observed that most of the high default rates are clustered in the east coast region. Furthermore, the map supports the trend analysis findings that Mississippi, West Virginia and Tennessee had the highest default rates.

Figure 1.4.



Simpson Paradox.

Grigg (2018), mentioned that the Simpson's paradox can be defined as a trend or result that is present data is group in a way that it reverses or disappears when the data is combined. Therefore, to prevent this occurrence, the team analyzed trends from different classifications or groupings within the dataset. We grouped the data by institution, grouped by states, by school

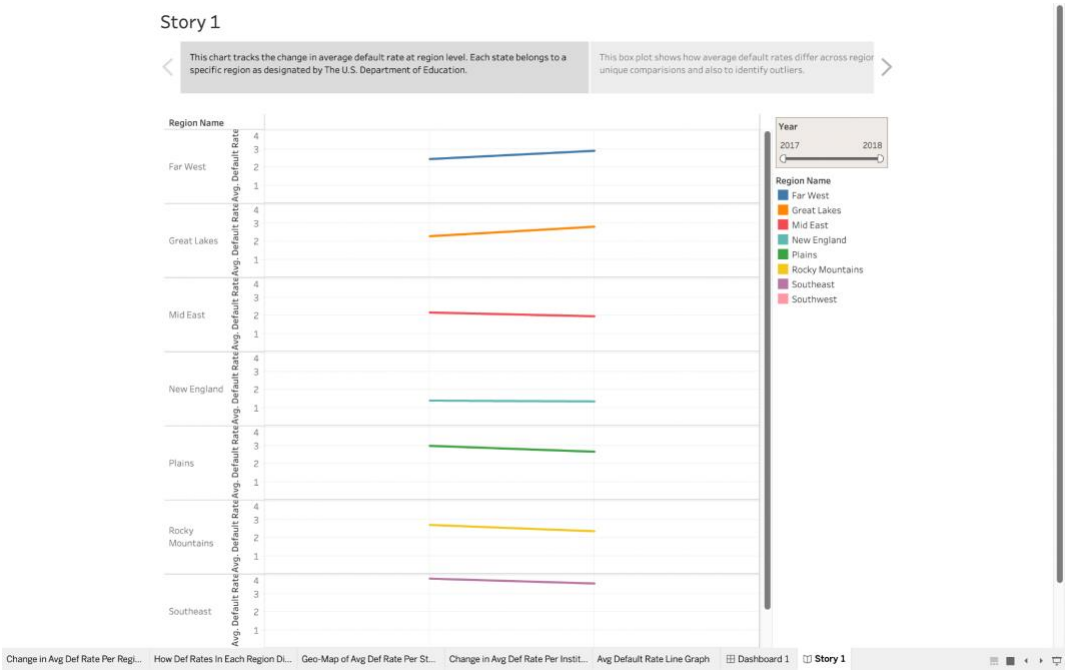
type, by ethnicity and by national ranking. Finally, the trends in each grouping were studied on an individual basis.

In each grouping we looked for the relationship between the default rate and the variables in the group, this is done to check if the default rate is determined by a particular variable and not the combination of each of the variables. We were able to observe the effect of varieties of variables on the default rates.

Descriptive Analytics

Here is a storyboard showing trends in default rates plotted against different groupings. This line graph tracks the change in average default rate at region level. Each state belongs to a specific region as designated by the U.S. Department of Education. The average change in default rate is the change in the rate from the previous year.

Figure 1.5(a)

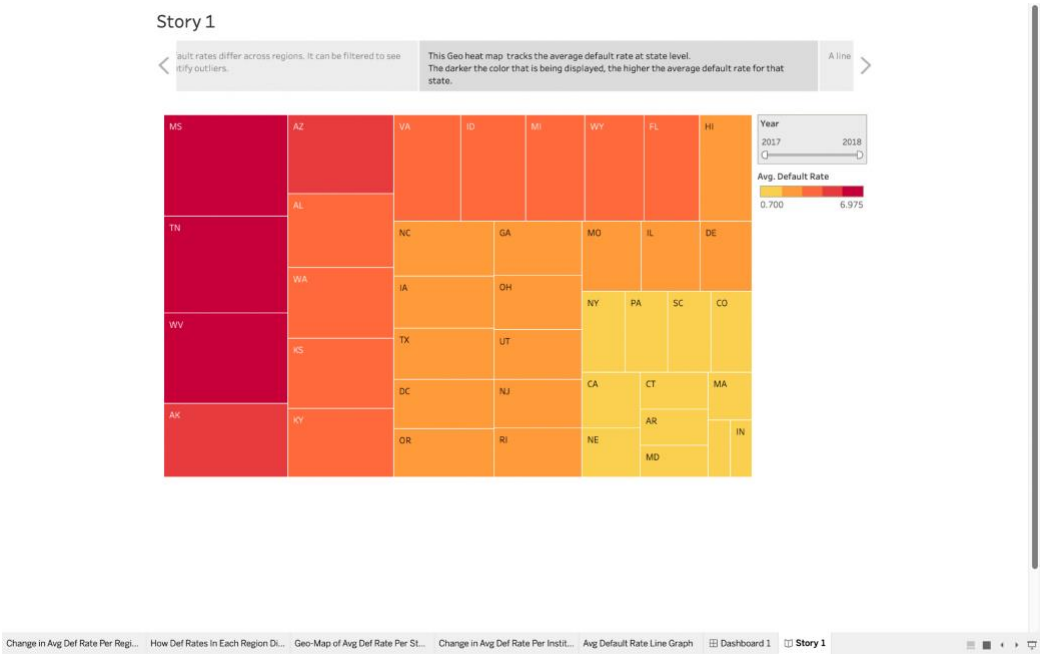


The box plot shows how average default rates differ across regions. It can be filtered to see unique comparisons and to identify outliers as shown in figure 1.5(b).

Figure 1.5(b)

This tree heat map tracks the average default rate at state level. The darker the color that is being displayed, the higher the average default rate for that state.

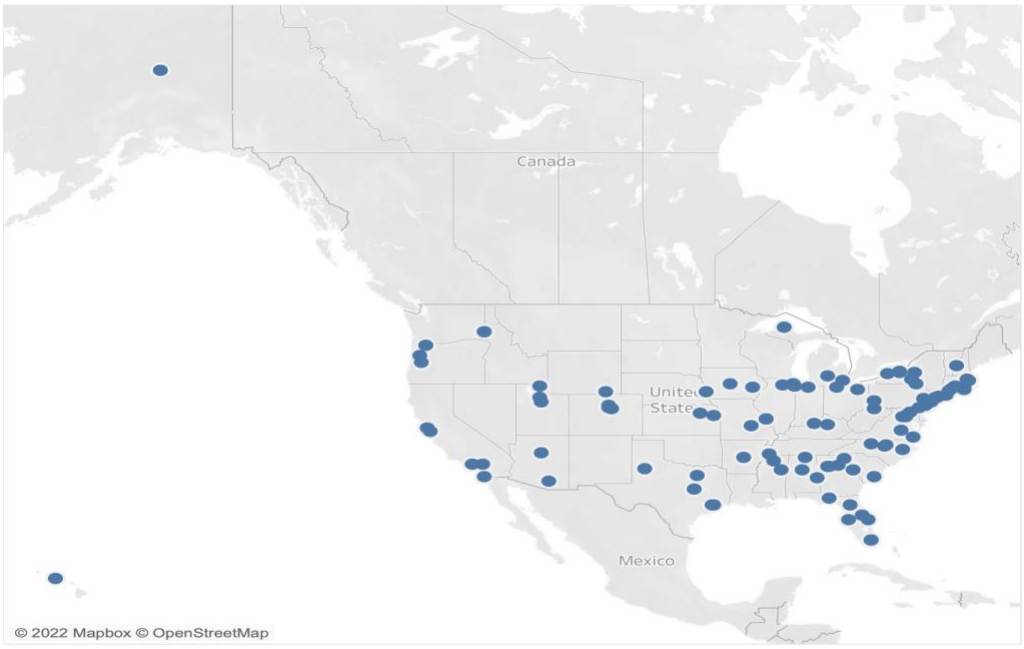
Figure 1.5(c)



The Geo-map allows the user to track changes in average default rate at institution level.

Figure 1.5 (d)

Change in Avg Def Rate Per Institution Each Year

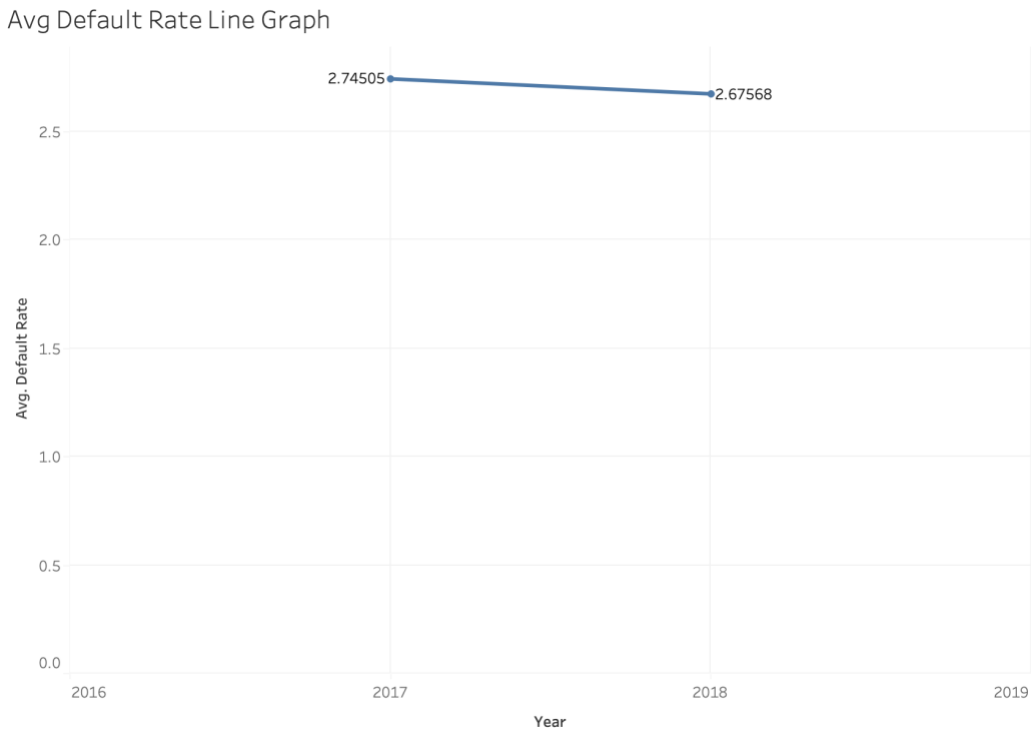


Map based on Longitude (generated) and Latitude (generated). Details are shown for Zip Code, Institution Name and State. The data is filtered on Year, which ranges from 2017 to 2018. The view is filtered on Institution Name, which keeps 111 of 111 members.

Conclusion

Finally, in figure 1.6, the data was segmented into the different years and the trend of the default rate was studied along the years. We observed a decline in default rate from 2016 to 2017 across the country. In 2017 through 2018 we observed a decrease in the default rate which can be attributed to the behavior of each region in the dataset.

Figure 1.6



Methodology Approach and Model Building

We built a Multiple Linear Regression Model to show the relationship between the independent variables identified from the data collected from the U.S. Department of Education. Using this model, future average default rates of student loans from each institution can be predicted. We used excel formulas to merge, sort, and clean the data. The model was built using codes in R Studio. Thereafter, we adopted the Re-sampling method whereby we re-estimated the model development process without of sample data.

Modeling Methods

We loaded our data from the data directory created in R, converted the data into a data frame, then we separated the dependent variable and did a normality check to confirm fulfillment of model assumptions or its violation, a normal plot was done to do this check. We proceeded to transform the dependent variable to fulfill the model assumption. We transformed the dependent variable by finding the square using the R code "sqrt()." Unwanted variables were removed from the data, these included names of institutions, zip codes, cities, and variables that have the same values in all its rows, we proceeded to convert categorical variables as factors to get our data ready for the model building.

Test Design

We tested for normality of the dependent variable by plotting a normal plot using normal QQ plot and histogram.

Model Building

To build the model in R, we used "lm" function expressed in the formula below:

defmodelall <- lm(Defaultratesqrtall~.,data=studloanall)

where:

- defmodelall is the assigned name we gave our model of the model
- Defaultratesqrtall is the transformed dependent variable(Default rate)
- Studloanall is our data already converted to a data frame

The last step of our model building was to do the summary to display the model results.

Our model summary produced the following results:

Residual standard error: 0.3275 on 178 degrees of freedom

Multiple R-squared: 0.7306, Adjusted R-squared: 0.6656

F-statistic: 11.23 on 43 and 178 DF, p-value: < 2.2e-16

Conclusion

The model supports the analytical problem statement. The model summary shows that the independent variables included in our model accounts for 73% of the variations in the dependent variable.

Model Evaluation

Evaluation Process Justification

The Evaluation of the model was guided by the principles of ordinary least squares(OLS). This is a technique for linear regression analysis that makes five overall assumptions about the data. These assumptions are linearity, absence of multicollinearity, absence of autocorrelation, homoscedasticity, normal distribution of errors.

Test For Linearity

The first assumption we tested for was whether the residual errors had linearity. This was done using a residual versus fitted plot and a normal Q-Q plot. The first plot shows the dispersion from the predicted values on the standardized scale on the y-axis which makes it easier to detect the presence of any outliers. Residuals are the difference between the actual values and the predicted values, it is also known as errors, the random behavior of these errors by showing no pattern when plotted against the predicted values shows that the model fits the data, when there is non-randomness observed in the plot, this means that the model does not fit the data properly. The second plot, otherwise referred to as a normal Q-Q plot shows the linear

relationship between the predictors (x) and the outcome (y). This allowed us to determine if the residuals are normally distributed.

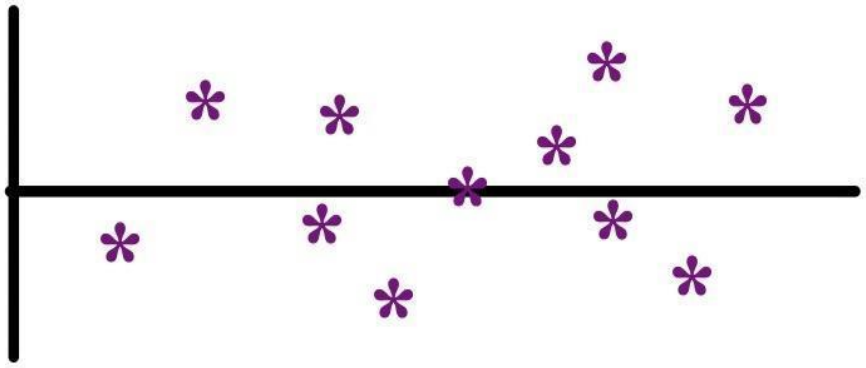
The Model was evaluated by using a scatter plot to see the behavior of the residuals(errors) when plotted against the predicted values. Firstly, the residuals should be distributed in a random way and the number of plus and minus residuals sign should be equal to the error. This symmetrical distribution is achieved when the variables are connected show a true linear relationship (Martin et al., 2017).

The steps involve:

- Recalling the model built earlier
- Standardize the residuals to see easily detect outliers
- Fit the model
- Graphically analyze Residuals by doing a scatter plot
- Note observations from the plot

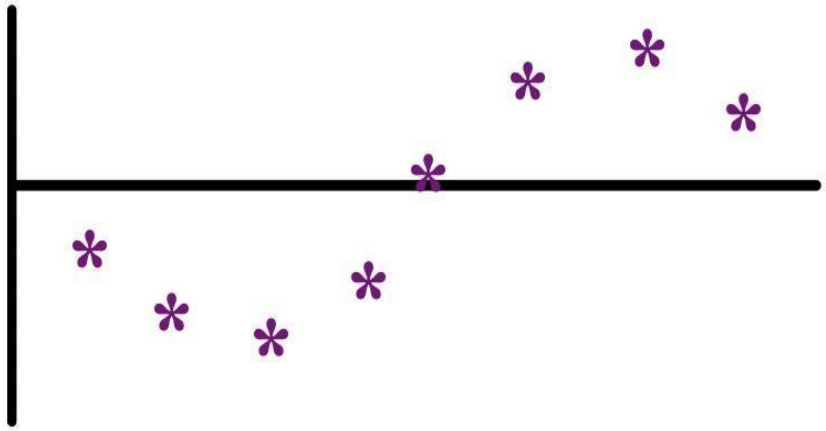
Residual versus fitted plots can be visually analyzed in three general ways. The first pattern shows a random scatter for points that form a constant width band around the identity line. This is known as an ideal residual plot. Kim (2019), mentioned that if the relationship between X and Y is truly linear, the scattered dots will not have any trend like linear or curved. In effect this means that residual and X variables are unrelated. In other words, the residuals appear randomly scattered in relation to X (Kim, 2019). Furthermore, most observations should lie close to the line, while observations far from the line are less frequent, following the characteristic of assumed normal distribution. This characteristic can be seen in figure 1.7(a).

Figure 1.7(a)



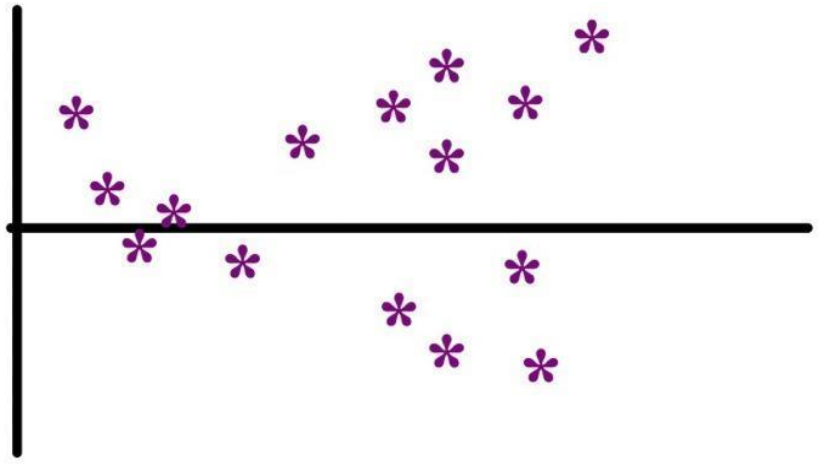
The second pattern in figure 1.7(b), otherwise known as a curved pattern, shows data points that are below residual equal to 0, then is followed by data points that are above residual equal to 0. This clearly distinguishable curved pattern is an indication that there is a nonlinear relationship in the original dataset.

Figure 1.7(b)



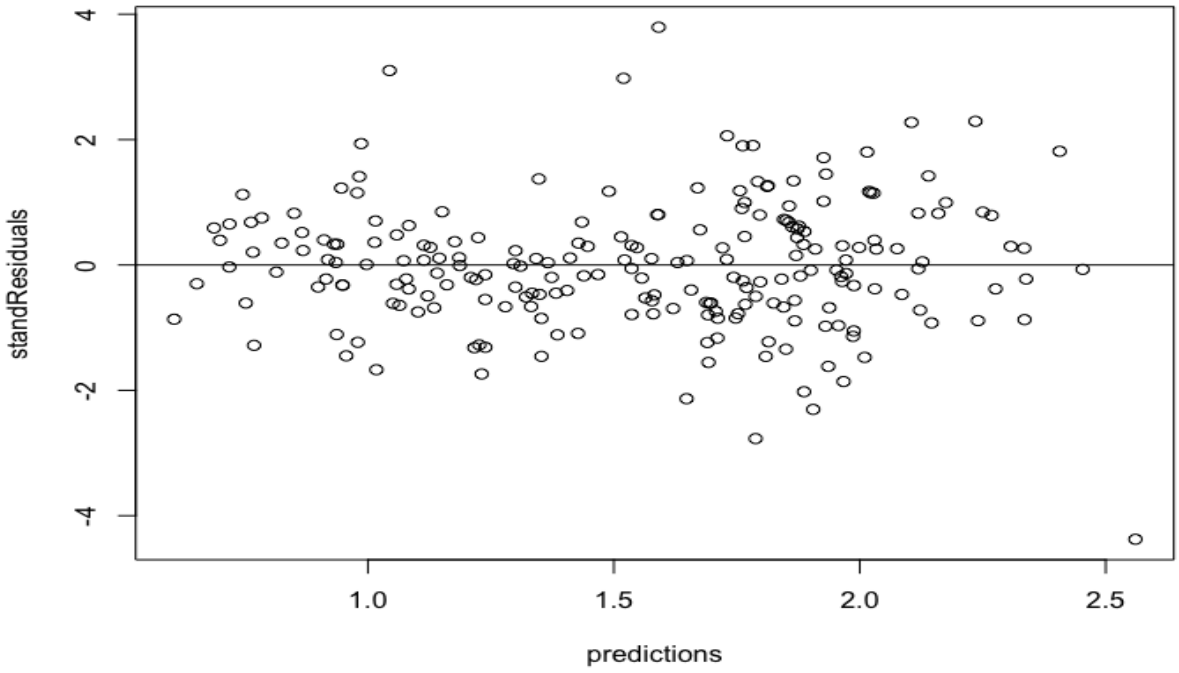
Finally, the third pattern in figure 1.7(c) is known as a megaphone pattern or cone shaped pattern and is an indication that the model will be less accurate for predicting larger values of x.

Figure 1.7(c)



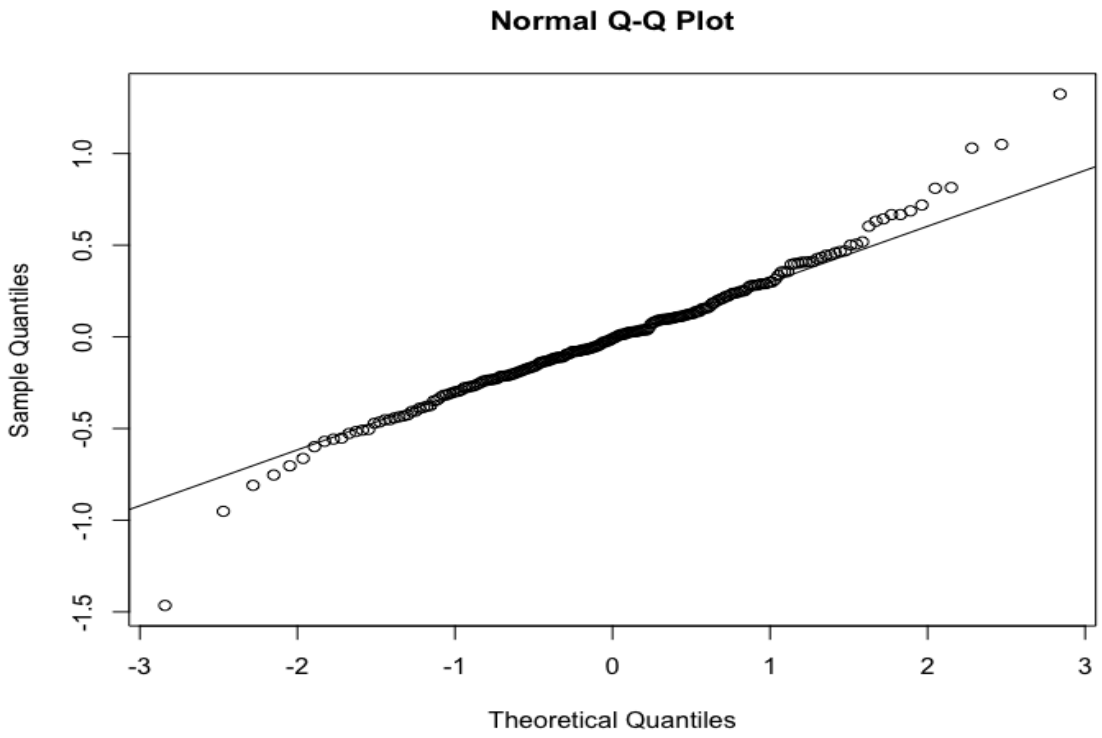
As shown in figure 1.7(d) the residual plot that was generated from our linear regression model showed homoscedasticity and no recognizable pattern. As earlier mentioned, this is an indication that most observations are located near the regression line, while observations far from the line are not as frequent, in accordance with the characteristic of assumed normal distribution.

Figure 1.7(d)



As stated, prior normal Q-Q plots are used to check if regression models meet the assumptions of linearity. As shown in figure 1.7(e), the relationship of normality is satisfied when the residuals are lined well on the straight dashed line.

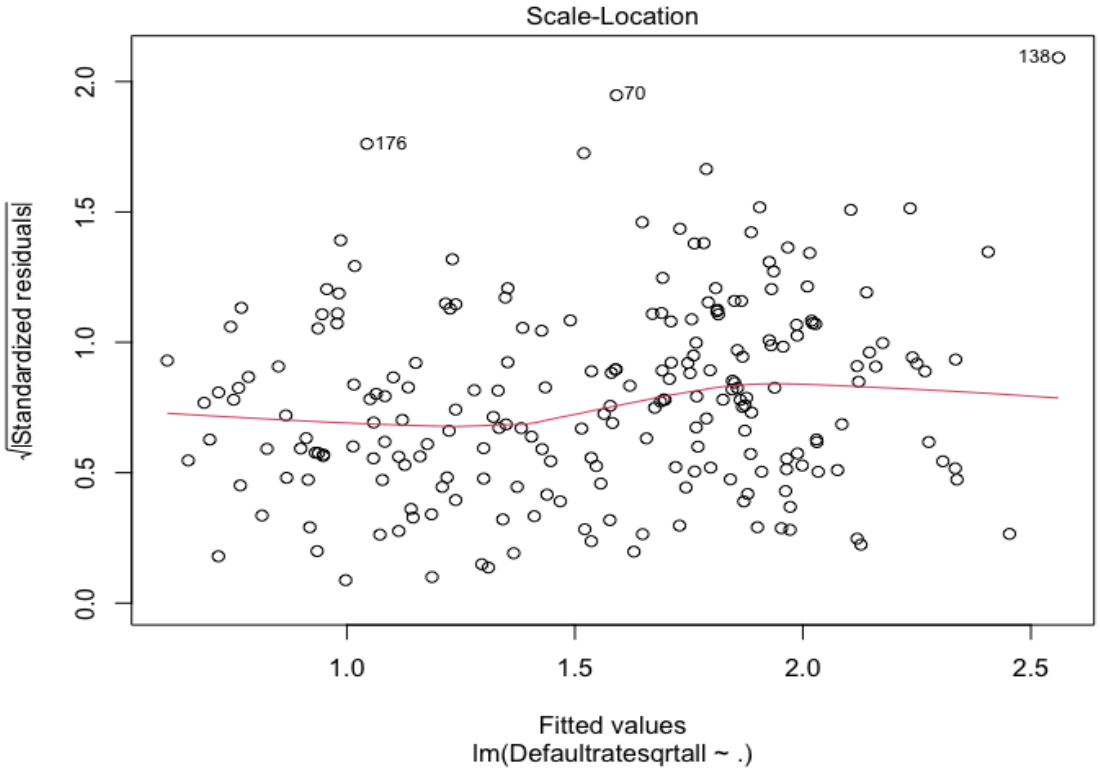
Figure 1.7(e)



Test For Constant Variance

The scale-location plot was selected because it allowed us to check the assumption of homoscedasticity. This describes a situation in which the random disturbance in the relationship between the independent variables and the dependent variable is the same across all values of the independent variables. However, the violation of this is known as heteroscedasticity and it is a situation in which the size of the error term differs across values of an independent variable. The problem brought forth when heteroscedasticity is present is that the cases with larger disturbances have more "pull" than other observations which makes the linear regression model less useful. As shown in figure 1.7(f), the model we built satisfied the test for constant variance because the residuals are normally spread, and the red line is approximately horizontal.

Figure 1.7(f)



Test For Multicollinearity

Multicollinearity is a situation where the independent variable can be predicted from another independent variable in a regression model. Wu(2020), a change in one of the highly correlated variables would cause a change to another variable, causing significant fluctuations in the model results. This results in instability of the model, causing the model to vary a lot in response to a minor change in data.

We tested for multicollinearity using the Variance Inflation Factor(VIF) for each of the independent variables. Higher the value of VIF mean correlation between the variable and

the rest. As a rule of thumb variables with a VIF higher than five require further investigation, and variables with a VIF higher than ten would indicate high correlation and thus be a cause for concern. As shown in table 3.3, the variables in our regression model were not highly correlated with each other, thus the test for multicollinearity was satisfied.

Table 3.3

```
> ols_vif_tol(defmodelall3) ## Final model
              Variables Tolerance      VIF
1               Ranking 0.7459517 1.340569
2               Region2 0.4481600 2.231346
3               Region3 0.5745870 1.740380
4               Region4 0.6274517 1.593748
5               Region5 0.3409897 2.932640
6               Region6 0.6490679 1.540671
7               Region7 0.5892393 1.697103
8               Region8 0.4879471 2.049402
9             OPENADMP2 0.8201114 1.219347
10            NPT4_PRIV 0.5505264 1.816443
11       WDRAW_DEBT_MDN 0.1779465 5.619667
12 PLUS_DEBT_ALL_PELL_N 0.6222857 1.606979
13                 YEAR 0.1912300 5.229305
>
```

Test For Autocorrelation

According to Kahn(2021), when there is a correlation between the error values that means autocorrelation is present. If a p-value greater than 0.05 is obtained, it means the null hypothesis will not be rejected. This gives us enough evidence that our independence assumption

is met. As shown in table 4.4, the p-value is greater than 0.05, therefore there is no autocorrelation in the regression model.

Table 4.4

```
> durbinWatsonTest(defmodelall3)
 lag Autocorrelation D-W Statistic p-value
   1      0.03305757      1.911337   0.482
 Alternative hypothesis: rho != 0
```

Model Summary

As shown in table 5.5, the linear regression model generated an R-squared value of 0.6164. Zach(2021), described R-squared as the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. Therefore, the model we created showed that 62% of the variables are responsible for the obtained dependent variable outcome. In addition, the model also resulted in an adjusted R-squared value of 0.5924 or 60%. (Zach, 2021), also describe adjusted R-squared as the modified R-squared that adjusts for the number of predictors in a regression model. As more predictors are added to the model the adjusted R-squared will increase.

Table 5.5(a)

```
> summary(defmodelall3)

Call:
lm(formula = Defaultratesqrtall ~ ., data = studloanall3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46500 -0.21218 -0.00878  0.19921  1.32438

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.427e+02  2.238e+02   3.319 0.001066 **
Ranking               3.190e-03  4.655e-04   6.852 8.12e-11 ***
Region2               6.853e-02  9.091e-02   0.754 0.451809
Region3               3.029e-02  1.071e-01   0.283 0.777659
Region4               1.303e-01  1.184e-01   1.100 0.272675
Region5               2.481e-01  9.356e-02   2.652 0.008628 **
Region6               2.379e-01  1.239e-01   1.920 0.056169 .
Region7               6.489e-02  1.222e-01   0.531 0.596049
Region8               1.018e-02  1.119e-01   0.091 0.927604
OPENADMP2             6.001e-01  1.358e-01   4.417 1.61e-05 ***
NPT4_PRIV            -1.144e-05  2.192e-06  -5.219 4.34e-07 ***
WDRAW_DEBT_MDN        3.067e-05  9.158e-06   3.349 0.000962 ***
PLUS_DEBT_ALL_PELL_N  1.798e-04  4.355e-05   4.129 5.28e-05 ***
YEAR                 -3.681e-01  1.110e-01  -3.317 0.001072 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3615 on 208 degrees of freedom
Multiple R-squared:  0.6164,    Adjusted R-squared:  0.5924
F-statistic: 25.71 on 13 and 208 DF,  p-value: < 2.2e-16
```

While the model generated satisfactory results, it is worth noting that the original dataset known as the college scorecard obtained from the U.S. Department of Education contained the following categories of variables.

- School- Describe the institution and its location.
- Admissions- Describe admission requirements for the institution such as standardized test scores
- Academics- Majors offered at the institutions.
- Student- Demographic information on the student body.
- Cost- Average net price for attending the institution and income brackets that families fall under.

- Completion- Graduation rate for the institutions.
- Repayment- Repayment rate in years for students who have either completed or dropped out of the institution.
- Aid- The number of students who received federal aid such as a Pell grant.
- Earnings- The average earnings of students between three to eight years after entry into the job market.

The team tried its utmost to retrieve as many variables as possible from each of the categories, especially the repayment, aid, and earnings categories as we believed those to have a profound effect on the dependent variable. However, due to information privacy suppression on quite a considerable number of variables, we were only able to retrieve a few variables from each category to fit the business problem and this can be counted as a limitation in our study.

External Model Verification and Calibration

Literature Review

Cross validation was completed on the model by partitioning the data used in building the model into training and testing data sets. Cross-validation is a statistical method that is used to estimate the skill models built using machine learning methods. According to Brownlee(2018), the procedure is used on a limited sample to estimate how well the model would perform. This is generally used to make predictions on data not introduced during the initial training of the model. As such, we divided the dataset into distinct categories of partitions, first we partitioned into 70% by 30% with the former being used to train the model while the latter was used to validate the model. Data was also partitioned into 75% and 25%, as well as 80% and 20% using the higher percentages to train the model while testing was done with the lower percentages with

both sets containing all the variables in the dataset. After building the model, it was validated by making predictions with the test set, each model built with the respective test and training set were summarized and evaluated by assessing model quality and comparing model performance(results below), then visualization of the prediction against the fitted values followed.

We validated our model by comparing the R-squared, adjusted R-squared, p-values, Residual standard errors, Root Mean Square Error, AIC, BIC, and the accuracy by checking the predictive Error rates. Model 1 with 70% by 30% partition produced better results. Making predictions with model 1 produces higher accuracy, having produced the lowest predictive error rate of 24%(an accuracy of 76%). The R-squared proves that about 60% of the variation in the default rate is explained by the predictor variables included in our model. The RMSE (Root Mean Square Error) being the average error by the model in predicting the default rate is about the same for model 1 and model 3, we would go with model 1 after considering other measurements such as AIC and BIC having lower values than in model 3.

70/30 Partition

Figure 1.8(a) - QQ Plot for Model at 70-30 Split

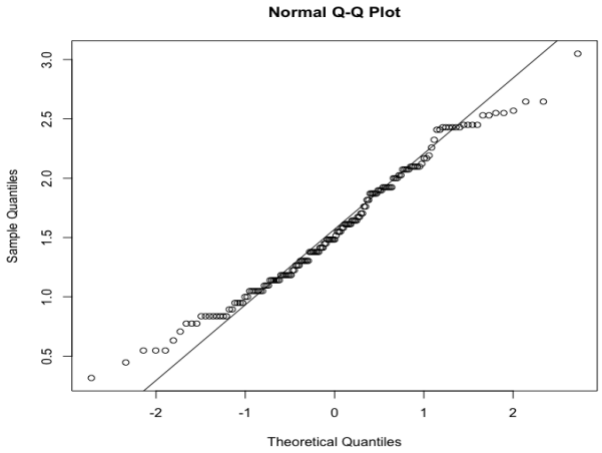


Table 5.5(b) - Linear Regression Model at 70-30 Split

```
Call:
lm(formula = defrate_train ~ ., data = train_studloan)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41628 -0.23894 -0.00254  0.24360  1.15164

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.177e+01  1.276e+02   0.719 0.473167
Ranking                  2.828e-03  7.102e-04   3.982 0.000109 ***
Region2                  1.606e-02  1.292e-01   0.124 0.901232
Region3                 -1.407e-02  1.513e-01  -0.093 0.926076
Region4                 -7.377e-02  1.696e-01  -0.435 0.664181
Region5                  7.639e-02  1.346e-01   0.568 0.571283
Region6                  1.254e-01  1.721e-01   0.729 0.467406
Region7                 -5.287e-02  1.712e-01  -0.309 0.757889
Region8                 -1.819e-01  1.470e-01  -1.237 0.218047
Total.females.Stud      -1.674e-06  8.342e-06  -0.201 0.841293
NPT4_PUB                 1.440e-05  5.225e-06   2.757 0.006619 **
GRAD_DEBT_MDN            2.076e-05  1.222e-05   1.698 0.091665 .
PLUS_DEBT_ALL_NOMALE_N   1.809e-04  7.231e-05   2.502 0.013511 *
PLUS_DEBT_ALL_NOPELL_MD -1.151e-05  4.615e-06  -2.494 0.013800 *
YEAR                    -4.506e-02  6.325e-02  -0.712 0.477375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3789 on 140 degrees of freedom
Multiple R-squared:  0.588,     Adjusted R-squared:  0.5468
F-statistic: 14.27 on 14 and 140 DF,  p-value: < 2.2e-16
```

Figure 1.8(b) - Prediction of Test Data for 70-30 Split

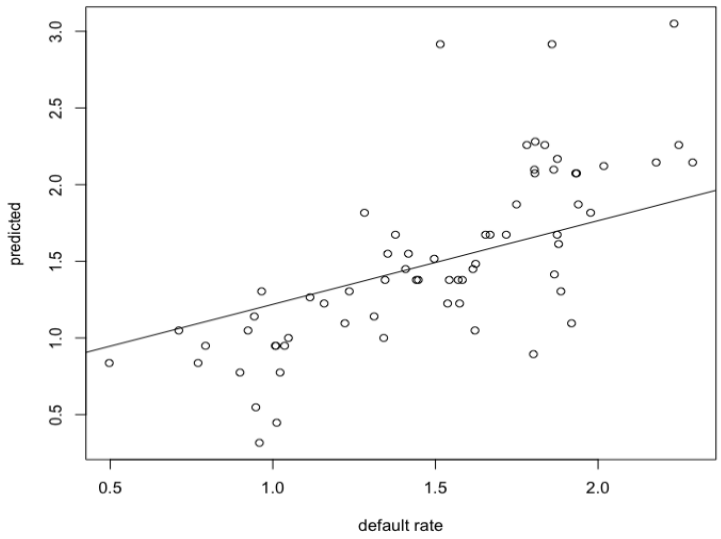
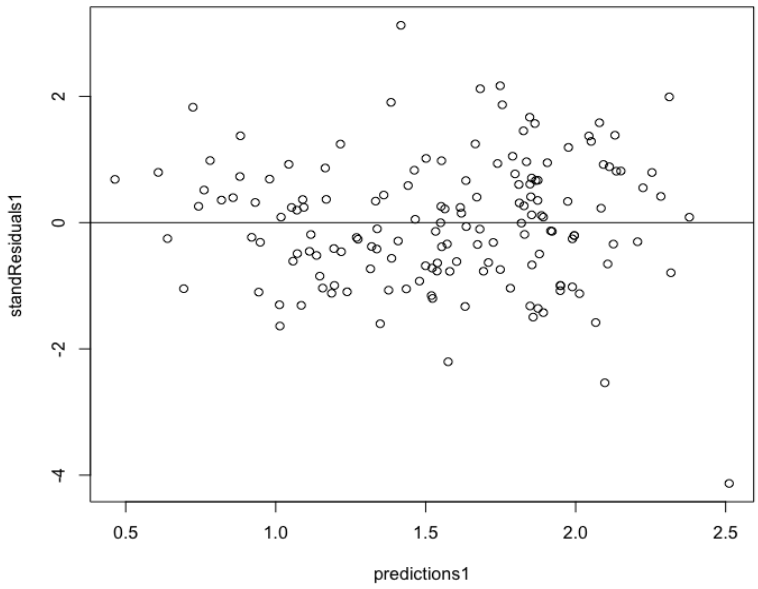


Figure 1.8(c) - Residual Plot for Model at 70-30 Split



80/20 Partition

Figure 1.9(a) - QQ Plot for Model at 80-20 Split

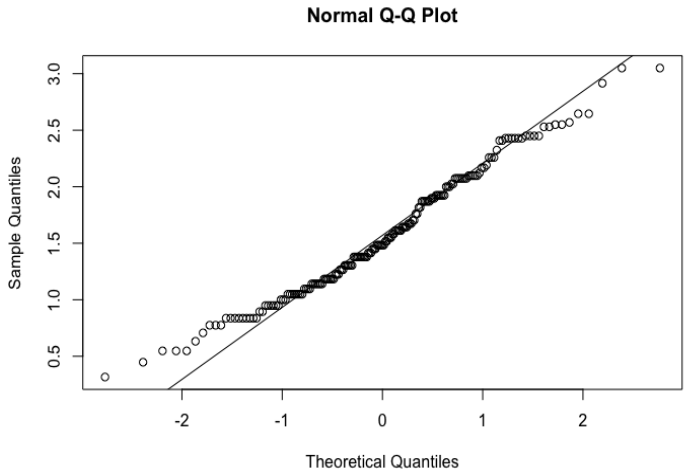


Figure 1.9(b) - Prediction of Test Data for 80-20 Split

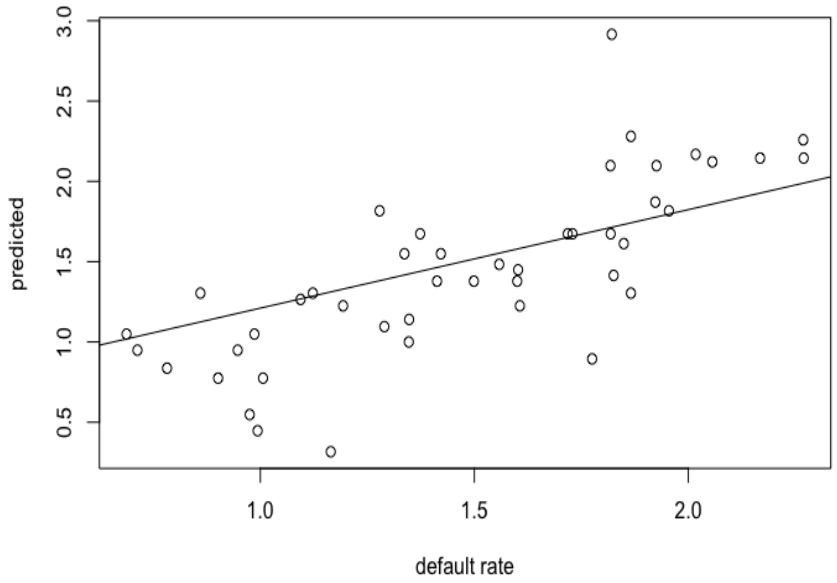


Table 5.5(c)

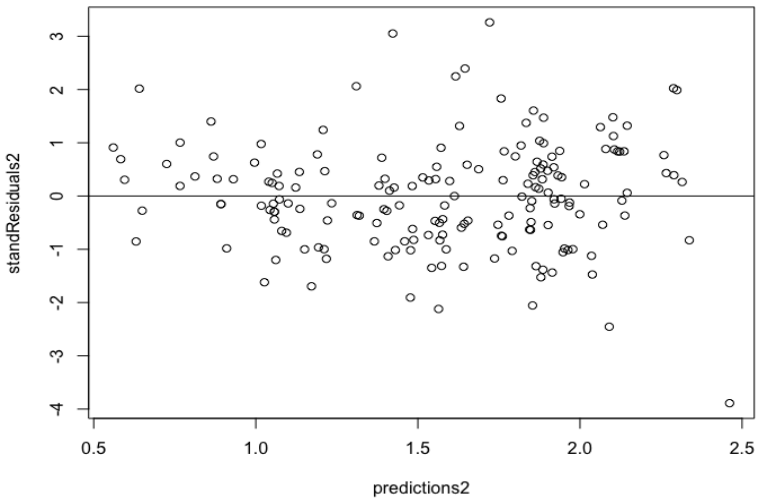
```
Call:
lm(formula = defrate_train2 ~ ., data = train_studloan2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36605 -0.23500 -0.01841  0.21125  1.19421

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.401e+00  1.192e+02   0.079 0.937254
Ranking                  2.450e-03  6.890e-04   3.556 0.000495 ***
Region2                 -2.054e-02  1.172e-01  -0.175 0.861039
Region3                 -4.397e-02  1.383e-01  -0.318 0.750907
Region4                 -7.887e-02  1.539e-01  -0.512 0.609048
Region5                  7.216e-02  1.237e-01   0.583 0.560531
Region6                  1.156e-01  1.566e-01   0.738 0.461693
Region7                 -6.765e-02  1.606e-01  -0.421 0.674170
Region8                 -5.484e-02  1.310e-01  -0.419 0.675970
Total.females.Stud      -2.884e-06  8.181e-06  -0.353 0.724872
NPT4_PUB                 1.410e-05  4.975e-06   2.835 0.005170 **
GRAD_DEBT_MDN            2.670e-05  1.166e-05   2.289 0.023362 *
PLUS_DEBT_ALL_NOMALE_N   1.674e-04  6.727e-05   2.489 0.013829 *
PLUS_DEBT_ALL_NOPELL_MD -1.502e-05  4.263e-06  -3.524 0.000552 ***
YEAR                    -4.201e-03  5.911e-02  -0.071 0.943433
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3855 on 162 degrees of freedom
Multiple R-squared:  0.5789,    Adjusted R-squared:  0.5425
F-statistic: 15.91 on 14 and 162 DF,  p-value: < 2.2e-16
```

Figure 1.9(c) - Residual Plot Model at 80-20 Split



75/25 Partition

Figure 2.1(a) - QQ Plot for Model at 75-25 Split

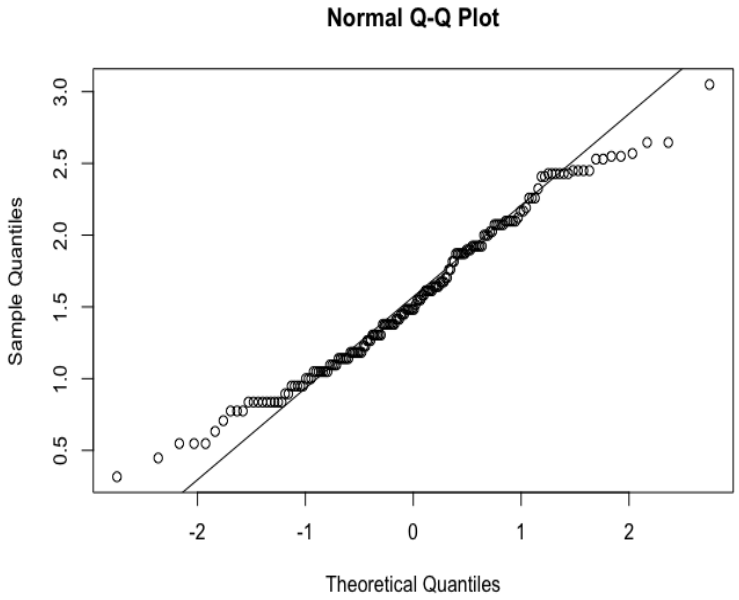


Figure 2.1(b) - Prediction of Test Data for 75-25 Split

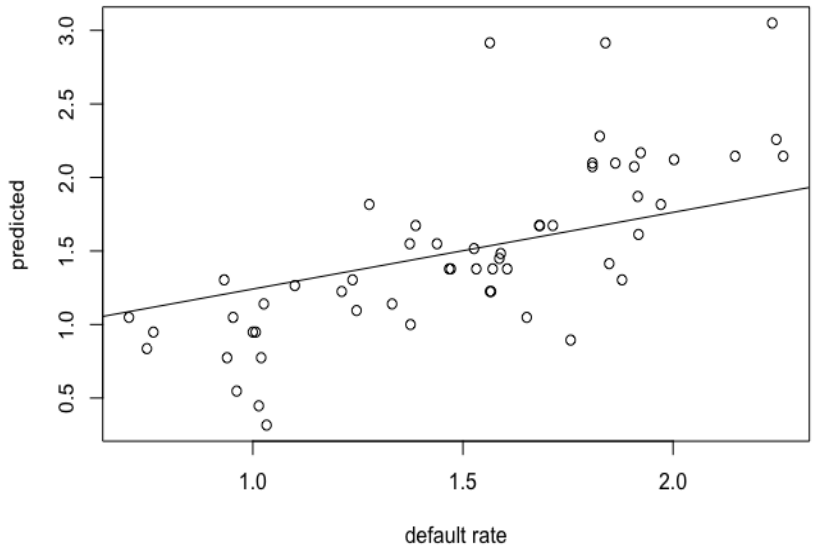


Table 5.5(d)

```
Call:
lm(formula = defrate_train3 ~ ., data = train_studloan3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42370 -0.24119  0.01436  0.23574  1.15342

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.326e+01  1.209e+02   0.606 0.545382
Ranking                  2.589e-03  6.869e-04   3.769 0.000234 ***
Region2                 -1.074e-02  1.217e-01  -0.088 0.929854
Region3                  1.103e-02  1.454e-01   0.076 0.939612
Region4                 -6.341e-02  1.606e-01  -0.395 0.693544
Region5                  8.320e-02  1.284e-01   0.648 0.517954
Region6                  1.140e-01  1.570e-01   0.726 0.468766
Region7                 -2.278e-02  1.626e-01  -0.140 0.888751
Region8                 -1.210e-01  1.375e-01  -0.880 0.380422
Total.females.Stud      -1.483e-06  8.044e-06  -0.184 0.853963
NPT4_PUB                 1.282e-05  5.049e-06   2.540 0.012092 *
GRAD_DEBT_MDN            2.379e-05  1.183e-05   2.012 0.046018 *
PLUS_DEBT_ALL_NOMALE_N   1.860e-04  6.687e-05   2.782 0.006092 **
PLUS_DEBT_ALL_NOPELL_MD -1.214e-05  4.341e-06  -2.797 0.005829 **
YEAR                    -3.590e-02  5.992e-02  -0.599 0.550065
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3753 on 151 degrees of freedom
Multiple R-squared:  0.5879,    Adjusted R-squared:  0.5497
F-statistic: 15.39 on 14 and 151 DF,  p-value: < 2.2e-16
```

Figure 2.1(c) - Residual Plot for Model at 75-25 Split

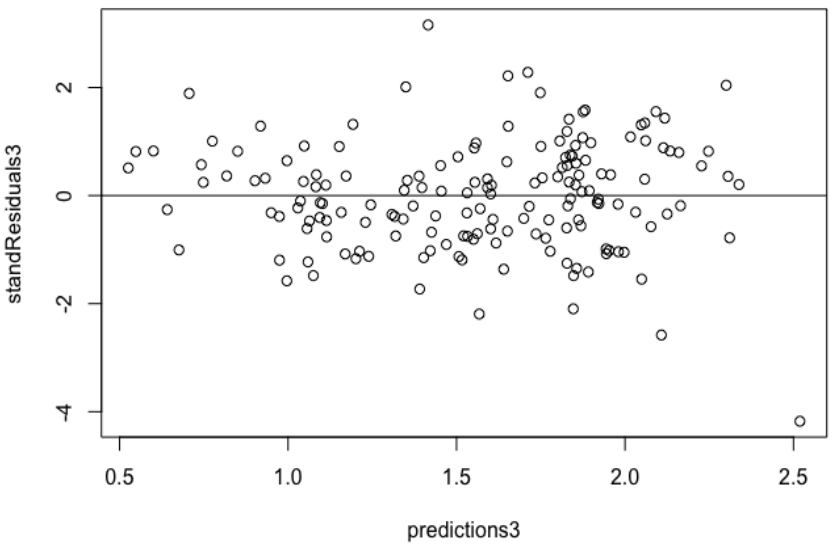


Table 6.6

Model Summary, Quality and Performance Check

Results	**Model 1**	**Model 2**	**Model3**
Partition	70/30(%)	80/20(%)	75/25(%)
R^2	0.5983	0.5846	0.5948
Adjusted R^2	0.5613	0.5515	0.5601
p-value:	4.83e-22	6.16e-22	8.93e-22
RSE	0.3728	0.3817	0.371
RMSE	0.3556101	0.3663031	0.3549788
AIC	149	177	157
BIC	195	224	204
Accuracy	0.7618139	0.7571624	0.7557983

The validation model built we built is sufficient and proves to be consistent with our theories that default rates are correlated with the variables included in the model. The results of the model were evaluated with different partitions, and several measurements were done to check consistency with our theories.

The next steps are to deploy the model by making the model available for prediction. First, we will create a deployment plan where we select the algorithm, then specify the training dataset that will be used as a baseline to detect model drifts. We will build a production workflow that allows us to make predictions as well as process new data. This can be done periodically as data is only made available monthly and yearly. Thereafter, we can measure the accuracy of the prediction, as well as how effective the decisions that were taken were based on the prediction. Finally in the Iteration stage we will address model drift and gain new insight and other behaviors observed in the process.

Having satisfied all assumptions, producing a good prediction accuracy, our model needs no revision. Although there were limitations to accessing expected data due to privacy suppression and unavailability of such data, our model was still able to produce significant variables that accounts for about 60% of the dependent variable at a predictive accuracy of 76.2%.

Future Recommendations

We would recommend that the model be fed with a complete and more relevant dataset. According to Shaun Burke (2001), in his research on the Valid Analytical Method, I would recommend that analysts should not rely on the regression statistics alone to indicate a linear relationship but instead data should be plotted for visual analysis as statistics alone is not enough to determine the linear relationship between the variables.

We would also recommend that when assumptions are not fulfilled, data need to be checked before transformation, there were a lot of inconsistencies in our data at the initial stage

and this was making the analysis difficult, an assumed continuous variable had more than 5% of zeros and it made transformation difficult, resulting in very low R-squared for our model, and heteroscedasticity, after we discovered that zeros were due to unavailable data for those observations, they were removed, and transformation showed better results and our model produced better results.

Model Deployment and Model Life Cycle

Deployment Cost

Some particularly important considerations to make before deploying a model includes knowing what type of deployment approach would best answer the business problem. The benefits of the deployment type should be weighed against the complexity or cost implication that comes with it when compared to other approaches like real-time predictions. Based on the nature of our data collection, we resolved to use batch predictions as data collected is only being updated over a period from monthly to yearly. Deployment cost was an important consideration for the deployment approach chosen, comparing the deployment cost of a batch prediction to real-time reveals that our model does not need the computing power needed to spread the load emanating from live data being fed into the system throughout the day, that could force our model into unprecedented costs such as SLAs (Service-Level Agreement) to meet up with the real-time loading.

The model that we built is a multiple linear regression model that utilizes a batch inference to make predictions, thus it falls into the third category of machine learning model types. The cost of deploying a batch inference model is typically free of such costs associated with endpoint requirements. According to Susilo (2021), other models like real-time inference

models will require the endpoint to be available 24/7. Furthermore, computing resources may need to be scaled up during peak load periods and scaled down during low load periods. This management of scaling means that a team of professionals may be required to manage the real-time model.

We believe the model we built will not be subjected to these costs because as information comes in from the U.S Department of Education on federal student aid and student loan repayment rates, the model will only be fed with statistics by the user periodically. In addition, this includes other significant variables from the various Universities.

Predictive Models Cost Benchmark

To better illustrate the cost of deploying a predictive model, we researched an article that talks about a default prediction model for Italian SMEs (Small and medium enterprises). The model in question adopts a logistic regression with data provided by an Italian credit rating agency. The analysis focuses on the role that financial and economic factors, drawn from balance sheets and income statements, can play in affecting the probability of default of firms (Modina & Pietrovito, 2014).

We can use this example to estimate deploying a predictive machine learning model. In understanding the process of developing a machine learning model, MLOps comes as one of the best approaches and in deploying the model as a production system. This approach is important as it streamlines operations and reduces the long-term burden on engineering to produce a new model from scratch each time (Coop, 2021).

MLOps Framework

Model infrastructure- This involves the implementation of load balancers to regulate data flow.

Data support- An independent data pipeline manager would have to be hired for continuous updates of analytic data.

Engineering and deployment- This involves the implementation of a continuous integration and continuous deployment system to pull information from the registry.

As stated by Coop (2021), a company utilizing the MLOps approach can expect to spend up to $90,000 over a total investment period of five years. This would also hold true for the Italian company use case example.

Bare-Bones Framework

Equally as argued by Coop (2021), if the MLOps approach is expensive, a company can opt for a bare bones approach which is a down scaled version of the former. The total investment over a five-year period can amount to $60,000.

Model infrastructure- A single machine in the cloud with no load management.

Data support- Timed script executed on infrastructure to pull data.

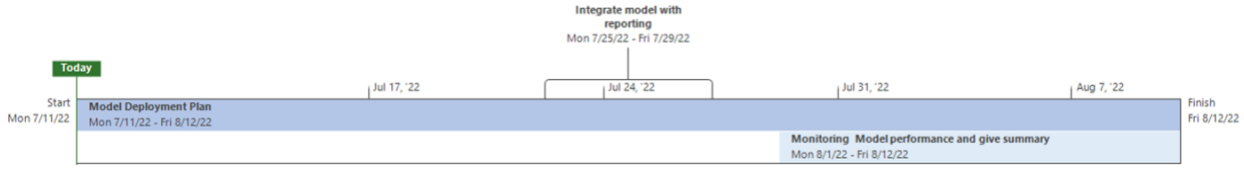
Engineering and deployment- This involves implementing copying data from a data scientists' machine then uploading to a cloud-based machine.

Schedule, Training, and Risk

As shown in figure 3.1, the proposed task for deploying our model involves steps that would incorporate analytical results into everyday decision-making process. The model is first moved to production by doing a scoring system of the model with a new dataset without

dependent variables. Our model deployment is expected to be done in 25 days. The steps for the model deployment are stated in the timeline below with allotted period for each step:

Figure 3.1



ID	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		**Model Deployment Plan**	**25 days**	**Mon 7/11/22**	**Fri 8/12/22**	
2		Validate the model	5 days	Mon 7/11/22	Fri 7/15/22	
3		Score the model	5 days	Mon 7/18/22	Fri 7/22/22	
4		Integrate model with reporting	5 days	Mon 7/25/22	Fri 7/29/22	3
5		Integrate model with application	5 days	Mon 8/1/22	Fri 8/5/22	
6		Monitoring Model performance and give summary	10 days	Mon 8/1/22	Fri 8/12/22	4

After model validation is done and indicates that the model is good enough to use, the next step is to start the deployment process with scoring the model. The first five days of our model deployment is to validate the model.

Score the model: The obtained value from scoring the model is used in decision making. The scoring is done by applying a new dataset to the model to find practical insights good for solving business problems. The duration for this stage is five days.

Integrate with reporting: Our approach is to use business intelligence tools in integrating the model and making reports. This will serve as a reference point for future consultation. Five days were allotted to this task.

Integrate with application: Our mode of integration is to use other applications to be parts of the model integration process for operational and business efficiency. We allotted the next five days to carry out this task.

Monitoring Model Performance: This is necessary to see how the model is performing after deployment, feedback of this performance for the next allotted ten days would help identify further problems and help give information that would improve the model performance.

To train the users who will be utilizing the model on a regular basis. An instructional system development model will be implemented. There is need to define training objectives based on job respective descriptions, job responsibilities and defined objectives (Chand, 2014). The first step consists of defining need assessments, job analysis and the target audience. This involves an in-depth analysis of job roles and the specific training required for each job role. For instance, on the business intelligence front employees will need to be trained in the use of a data visualization tool like tableau or power BI.

The second step, otherwise known as the planning phase, consists of setting the goal of the learning outcome as well as the training methodology that will be deployed. Continuing with the business intelligence example, this type of training can be provided through an eLearning platform which consists of video tutorials and brief assessments for the user to complete to move onto the next stage of training. These assessments can be used to get an idea of how well the employee is understanding the material and thus lead to changes in content selection and sequencing if necessary. The third step involves development of the training program with components related to handouts, visual aids, and demonstrations. The fourth step, otherwise known as the execution phase, focuses on logistical arrangements associated with training.

The most important step is to make sure the training program has achieved high work performance. According to Chand(2014), this phase consists of identifying strengths and weaknesses of the training program while taking note of the necessary amendments to be made.

As mentioned prior there are three broad categories of models namely, a one-off model, batch models and real-time models. The model we built is a multiple regression model which broadly falls under the batch inference category as such it can be used on a repetitive basis because the predictive model can feed information periodically to get a result. Furthermore, as stated by Kervizic (2019), the complexity of the model itself plays a role in its ability to be used frequently. Implementation of linear regression models are easier to implement, much space is not usually required for storage and deployment.

Benefits

A standard practice among most organizations when addressing a business problem is to do a cost benefit analysis. This refers to a systematic and analytical process of comparing benefits and costs in evaluating the desirability of a project (Mishan & Quah, 2020). This is exceedingly important as it determines whether an organization will go ahead with a project or program.

Firstly, it is important to note that there is no specific way to carry out a cost-benefit analysis but there are core considerations to make in the analytical process. According to Weller (2016), the first step is to establish a framework to outline the parameters of the analysis, the second and third steps involve identifying costs and benefits whilst assigning monetary value to

them then lastly make a comparison between the costs and benefits. Thereafter, analyze results and provide recommendations.

We conducted a cost benefit analysis with these steps in mind to ascertain the specific benefits to the organization that will be obtained from using our model. The main consideration in our analysis comes in the form of transactional costs that are incurred from transferring a good or service across a technological interface. Data mining can be an expensive process even for the most established organizations. Therefore, we will save the organization money by carrying out data mining processes for them thus saving them labor related costs from hiring personnel to carry out these procedures for them. All the organization needs to do is provide access to the datasets while making sure to grant us full authorization to avoid a situation where we come across privacy suppressed information thus impacting the analytical process. Furthermore, another benefit to the organization is the maintenance of the system will be done by our team on our servers thus eliminating the need for the organization to perform this task. Additionally, this is compounded by the fact that the predictive model will enable the organization to gradually lower their default rate by acting on the statistical variables that they have control over. Identifying and acting upon the results of the predictive model will be cardinal to the organization to avoid losing access to title IX federal student aid.

Recommendations

Recommendations for practice.

The responses to the research question revealed an institution can use variables at their disposal to lower default rates or at the very least keep them below the at-risk threshold. One of

the most significant variables is median debt for students who have not completed school. This would indicate that students who receive federal aid, and then drop out of school contribute to the student loan default rate for that institution. Therefore, we recommend that the various institutions work towards improving student retention rates across the board.

Recommendations for future research.

Based on our findings from this research and existing literatures on the topic, we would first recommend that for future research, there should be availability of access to privacy suppressed information from the U.S. Department of Education's database.

This aforementioned information is exceedingly cardinal to the creation of a stronger predictive model. Furthermore, through this study we ascertained that the U.S. The Department of Education and the institution's themselves can potentially produce a more efficient way to distribute federal aid to prospective and existing college students. We believe this could come in the form of federal aid caps limiting the amount a student could borrow in the beginning, then a release of more funding pending performance obligations being met.

Conclusions

This quantitative study addressed the problem of student loan default rates among Universities in the United States. The objective of the study was to predict student loan default rates using university statistical data and identify which variables an institution can use or investigate to lower default rate. The results of the quantitative study we conducted found that the following variables; Institutional ranking, region locale, open admissions policy, average net price for Title IV institutions (private for-profit and nonprofit institutions), median debt for students who have not completed school and student recipient count for median plus loan debt

disbursed to Pell recipients at all institutions were all significant in the final factors that lead to high or low student loan default rate. Therefore, we can conclude that the capacity to lower default rate is in the institution's hands to a substantial extent.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). x cc*CRISP-DM 1.0: Step-by-step data mining guide.* Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Chand, S. (2014). *Models of training employees: Steps, transitional and instructional system development model*. Your Article Library. Retrieved from https://www.yourarticlelibrary.com/training-employees/models-of-training-employees-steps-transitional-and-instructional-system-development-model/29548

Coop, R. (2021). *What is the Cost to Deploy and Maintain a Machine Learning Model*? Retrieved from https://www.phdata.io/blog/what-is-the-cost-to-deploy-and-maintain-a-machine-learning-model/

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now? *Big Data, 6*(3), pp. 176-190. doi:10.1089/big.2018.0083

Irv Lustig, *PHD* (n,d.). *Bringing QA to Data Science.* retrieved from https://www.softwaretestingnews.co.uk/bringing-qa-to-data-science-2/

Kervizic, J. (2019). *Overview of the Different Approaches to Putting Machine Learning Models in Production.* Retrieved from https://www.topbots.com/putting-ml-models-in-production/#real-time-prediction-integration

Kim H. Y. (2019). Statistical notes for clinical researchers: simple linear regression 3 - residual analysis. *Restorative dentistry & endodontics*, *44*(1), e11. https://doi.org/10.5395/rde.2019.44.e11

Khan, M. (2021). *A Basic Guide to Testing the Assumptions of Linear Regression in R*. Retrieved from https://www.godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r

Martin, J., de Adana, D. D. R., & Asuero, A. G. (2017). Fitting Models to Data: Residual Analysis, a Primer. In (Ed.), Uncertainty Quantification and Model Calibration. Intech Open. https://doi.org/10.5772/68049

Mishan, E.J., & Quah, E. (2020). Cost-Benefit Analysis (6th ed.). Routledge. https://doi.org/10.4324/9781351029780

M. Modina & F. Pietrovito. (2014) A default prediction model for Italian SMEs: the relevance of the capital structure, Applied Financial Economics, 24:23, 1537-1554, DOI: 10.1080/09603107.2014.927566

Susilo, M. (2021). *Machine Learning Model Deployment Options*. Retrieved from https://towardsdatascience.com/machine-learning-model-deployment-options-47c1f3d77626

Using Plots to Check Model Assumptions. (n.d.). Retrieved from https://web.ma.utexas.edu/users/mks/statmistakes/modelcheckingplots.html

Wu, S. (2020). *Multicollinearity in Regression.* Retrieved from https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

Zach. (2020). *How to identify Skewness in Box Plots*. Retrieved from

https://www.statology.org/box-plot-skewness/

Appendix A: Data Set

Student Loan Default Rate Data Set