Bank Loan Classification

ECO421 Final Group Paper

Instructor: Professor Marlène KOFFI

Anqi Chen, chenan26, aaq.chen@mail.utoronto.ca
Arshvir Singh Bhangoo, bhango10, arshvir.bhangoo@mail.utoronto.ca
Anthony Kutsman, kutsmana, anthony.kutsman@mail.utoronto.ca

Introduction

Motivation

The majority of small businesses in the United States use loans in the form of credit from the Small Business Administration (SBA) to help their businesses grow. These SBA loans are used by small businesses across the United States to help pay labor costs, purchase inventory, equipment and technology, as well as general working capital. In 2019 alone, over 20.8 billion dollars has been lended out to small businesses across the United States in the form of an SBA loan[1], which shows how prevalent these loans are in fostering small business growth. In 2019, around 43% of small businesses applied for an SBA loan, and the overall average SBA loan amount is $107,000 (Shepherd, 2021), suggesting that a relatively large percentage of small businesses take on a significant amount of debt through SBA loans. Moreover, given that the failure rate for small businesses is around 50% in the 5th year of operation (Lewis, Daniel Lewis, 2021), and that around 17% of all SBA loans went into default from 2006 to 2015 (Voigt, 2020), it is important to predict what contributes to SBA loan default and to understand the top characteristics that contribute to default on SBA loans.

Research Questions and their answers

We examine which classifiers, out of logistic regression, naive bayes, decision tree and random forest, allow us to best predict SBA loan default by comparing their accuracy rate and AUC scores. Moreover, we then use these optimal classifiers to examine the most important features that contributed to SBA loan default from 1987 to 2014. As an extension, we split our data set based on loan size into small, medium, and large loan sizes and examine the top three characteristics that contribute to SBA loan default within each of these three groups.

We find that the random forest classifier is the optimal classifier to predict SBA loan default. Moreover, we see that the top features that contribute to the SBA loan default are the loan amount that was charged off, the loan term in months, and the fiscal year that the loan was issued. We also find that the most important features contributing to SBA loan default among the three loan size groups are the same: the loan amount charged off, the loan term in months, and the fiscal year that the loan was issued.

Contribution and Literature Review

Given the importance of SBA loans to the success of small businesses, many papers employing machine learning methods have been written about the implementation of machine learning

[1] https://www.sba.gov/sites/default/files/aboutsbaarticle/WebsiteReport_asof_20190830.pdf

classifiers for prediction of SBA loan default rates, and to find the features that contribute to SBA loan default. Dennis Glennon and Peter Nigro, in 2005, analyzed how the maturity of SBA loans is related to SBA loan default using a logistic regression model (Glennon & Nigro, 2005). Glennon and Nigro use a sample from the SBA 7(a) loan-guarantee program, and split up the sample into three groups based on the loan's maturity: 3 years, 5 years, and 15 years to maturity. They found that the top factors contributing to loan default rates vary across loan maturity, and moreover that loan default rates overall vary across the three groups of loans with different maturities. Specifically, they found that corporate structure was an important feature for short term SBA loan default, but not necessarily for longer term loans.

Moreover, Juhi Bhargava and Prashanth Musuku used the decision tree, random forest, and the logistic regression classifiers, as well as artificial neural networks to study the top factors that contribute to SBA Loan default based on SBA loan data from 2015 (Foley, 2012). Bhargava and Musuku found that Total Principal Received, and Outstanding Principal were some of the most important features contributing to SBA loan default. In their analysis, they split up the data into two groups that included good loans and bad loans, and found that the features mentioned above contributed to SBA loan default overall, and within both groups. These groups (good loans and bad loans) were determined by the repayment amount of the loan, the regularity of payments, as well as various other features.

We build off of the analysis presented by Glennon and Nigro, as well as Juhi Bhargava and Prashanth Musuku, by using a different time period from 1987 to 2014, and naive bayes classifier in addition to the logistic regression, decision tree, and random forest classifiers in order to find the optimal classifiers to predict SBA loan default rate. Moreover, building off of the work of Glennon and Nigro, instead of splitting up our data into three groups based on Loan Maturity, we look at the the top features that contribute to SBA loan default based on 3 groups of SBA loan size, as well as the top three features contributing to SBA loan default for loans of all sizes.

Data and Methodology

Data

The dataset (SBAnational.csv) used in our analysis is from the U.S. Small Business Administration (SBA). The raw SBA data is very large and contains 899,164 observations and 27 variables. We removed the variables that included a significant number of missing values, such as the variable “ChgOffDate” which has 736,465 missing values. Then we dropped all the observations that contained missing values. This resulted in a cleaned dataset of 886,240 observations and 26 variables (Table 1).

	count	unique	top	freq
LoanNr_ChkDgt	886240	886240	9264993008	1
Name	886240	769838	SUBWAY	1259
City	886240	32298	LOS ANGELES	11467
State	886240	51	CA	129398
Zip	886240	33501	10001	919
Bank	886240	5788	BANK OF AMERICA NATL ASSOC	86075
BankState	886240	56	CA	116737
NAICS	886240	1311	0	198267
ApprovalDate	886240	9786	7-Jul-93	1120
ApprovalFY	886240	48	2005	76905
Term	886240	411	84	225820
NoEmp	886240	597	1	151454
NewExist	886240	3	1	636139
CreateJob	886240	246	0	619802
RetainedJob	886240	356	0	433397
FranchiseCode	886240	2754	1	631412
UrbanRural	886240	3	1	465149
RevLineCr	886240	18	N	415439
LowDoc	886240	8	N	775189
DisbursementDate	886240	8435	31-Jul-95	9747
DisbursementGross	886240	117753	$50,000.00	42928
BalanceGross	886240	15	$0.00	886226
MIS_Status	886240	2	P I F	730199
ChgOffPrinGr	886240	82645	$0.00	726032
GrAppv	886240	21922	$50,000.00	68424
SBA_Appv	886240	37935	$25,000.00	49277

Table 1: Summary Statistics of All Variables

The outcome variable of this classification is the “MIS_Status" variable, which represents the loan default status of the business. The two levels of this qualitative variable are being charged off or paid in full. There are 156,041 observations that were charged off and 730,199 observations that were paid in full. Categorizing by states, California had the largest number of businesses that paid in full, as well as businesses that defaulted (Figure 1).

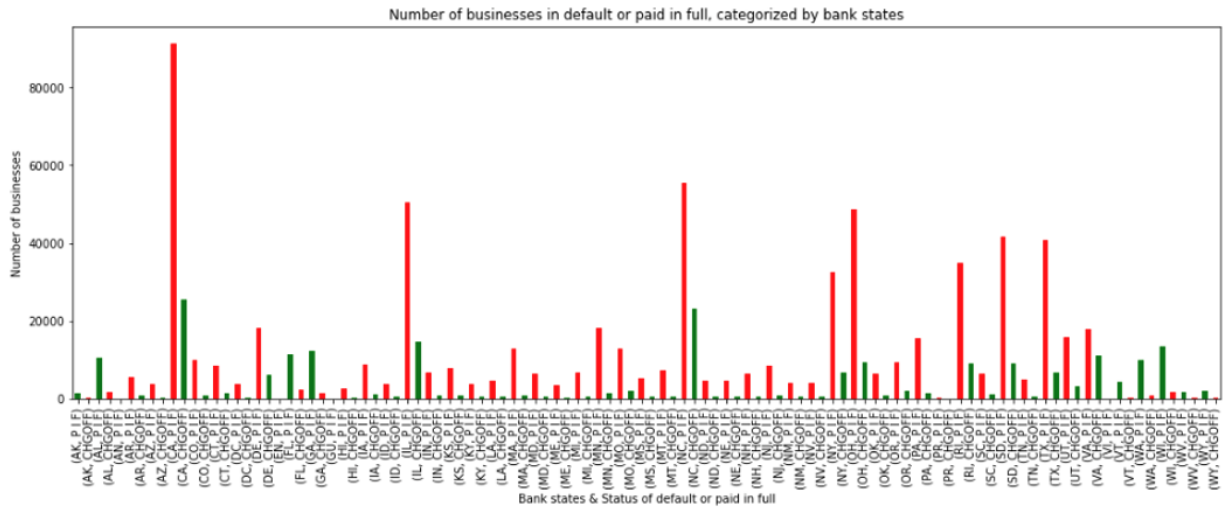

Figure 1: Number of businesses in default or paid in full, categorized by bank states

Moreover, the observations are also categorized by the status of whether the business was in an urban or rural area. Urban areas had the highest number of businesses that paid in full as well as businesses that defaulted (Figure 2).

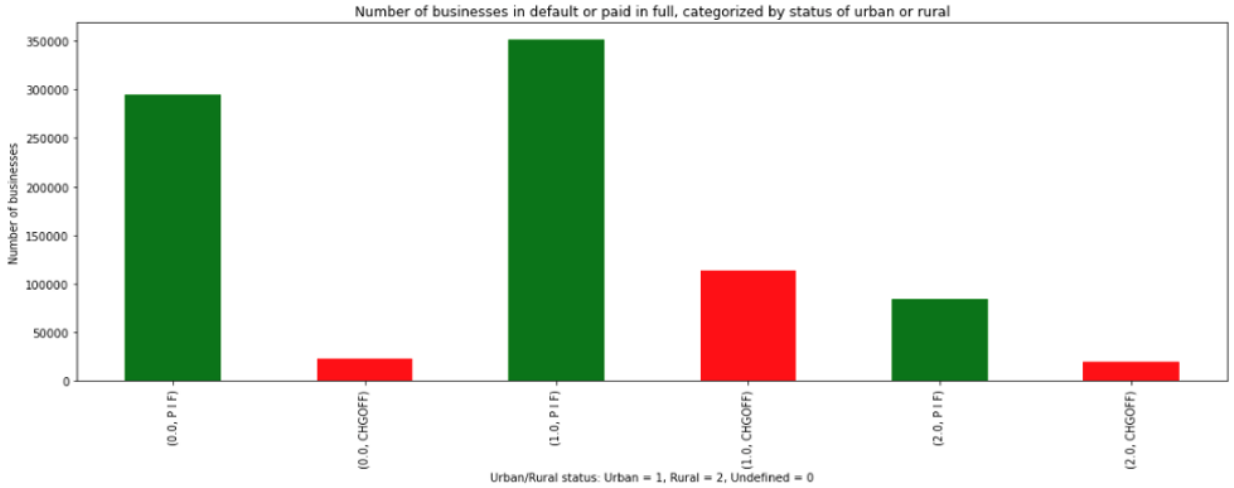

Figure 2: Number of Businesses in default or paid in full, categorized by status of urban or rural

The identifier variables, such as primary key and borrower name, were only used for identifying purposes, therefore they were not included in the list of predictors. Initially, some data cleaning was conducted to these predictors. Since all the continuous variables are of type object, their dollar sign symbols were removed and they were changed into floats. Then, integers are used to represent each level of the categorical variables. Since a part of our research question is focused on predicting the top features that contributed to SBA loan default among the three loan

subgroups (based on the SBA loan amount), a new variable "SBA_amount_level" was created to hold three levels. The SBA loan amount was divided evenly into tertiles (0% - 33%, between 33% - 66% including 66%, between 66% - 100% including 100%), and was assigned to number 1, 2, and 3, corresponding to the small, medium, and large loan size groups. A bar plot is also created to count the number of businesses that default, categorized by "SBA_amount_level" (Figure 3).

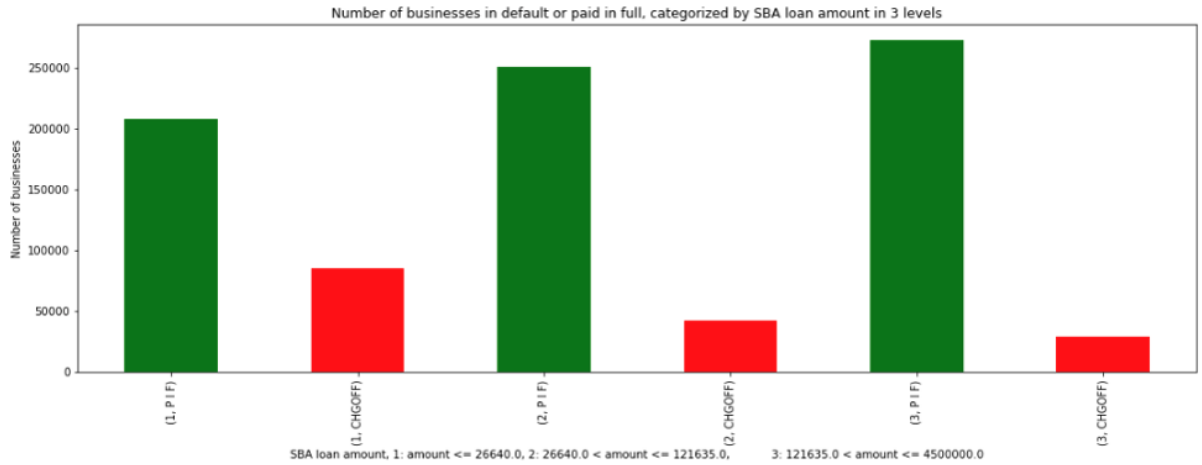

Figure 3: Number of Businesses in default or paid in full, categorized by SBA loan amount in 3 levels

After the data cleaning process, all the predictors were float types. Therefore, the analysis will use a list of 14 predictors, which are fiscal year of commitment, loan terms, number of business employees, new or existing business, number of jobs created, number of retained jobs, urban or rural, revolving line of credit, LowDoc loan program, amount disbursed, charged-off amount, amount of loan approved, SBA's guaranteed amount, and SBA's amount categorized into three quantiles (Table 2).

	count	mean	std	min	25%	50%	75%	max
ApprovalFY	886240.0	2001.149622	5.880586	1968.0	1997.0	2002.0	2006.0	2014.0
Term	886240.0	110.954647	78.990583	0.0	60.0	84.0	120.0	569.0
NoEmp	886240.0	11.420650	74.187995	0.0	2.0	4.0	10.0	9999.0
NewExist	886240.0	1.279900	0.451509	0.0	1.0	1.0	2.0	2.0
CreateJob	886240.0	8.463092	237.301746	0.0	0.0	0.0	1.0	8800.0
RetainedJob	886240.0	10.842406	237.739546	0.0	0.0	1.0	4.0	9500.0
UrbanRural	886240.0	0.759725	0.646074	0.0	0.0	1.0	1.0	2.0
RevLineCr	886240.0	10.332192	4.947374	0.0	3.0	12.0	12.0	17.0
LowDoc	886240.0	4.359118	0.997451	0.0	4.0	4.0	4.0	7.0
DisbursementGross	886240.0	202141.902469	287938.000625	4000.0	42837.0	100000.0	240000.0	11446325.0
ChgOffPrinGr	886240.0	13593.781748	65466.157464	0.0	0.0	0.0	0.0	3512596.0
GrAppv	886240.0	193499.681127	283505.287898	1000.0	35000.0	90000.0	227000.0	5000000.0
SBA_Appv	886240.0	149983.856349	228169.354635	500.0	21250.0	62125.5	175000.0	4500000.0
SBA_amount_level	886240.0	2.009993	0.818474	1.0	1.0	2.0	3.0	3.0

Table 2: Summary Statistics of 14 Predictors

Before constructing classifiers, a preliminary analysis of the default rate was conducted, and a map of the default rate was built (Figure 4). The default rate was calculated by using the proportion of businesses that default in each state. Then the default rate for contiguous United States from 1987 to 2014 was mapped. States that appear darker have a higher default rate and reversely states with lighter color have lower default rates. There was no data for states that appear white. Florida state appears to be the darkest, meaning it had the highest default rate over the period. While Mantona and Wyoming states appear to be the lightest, meaning they had the lowest default rate.

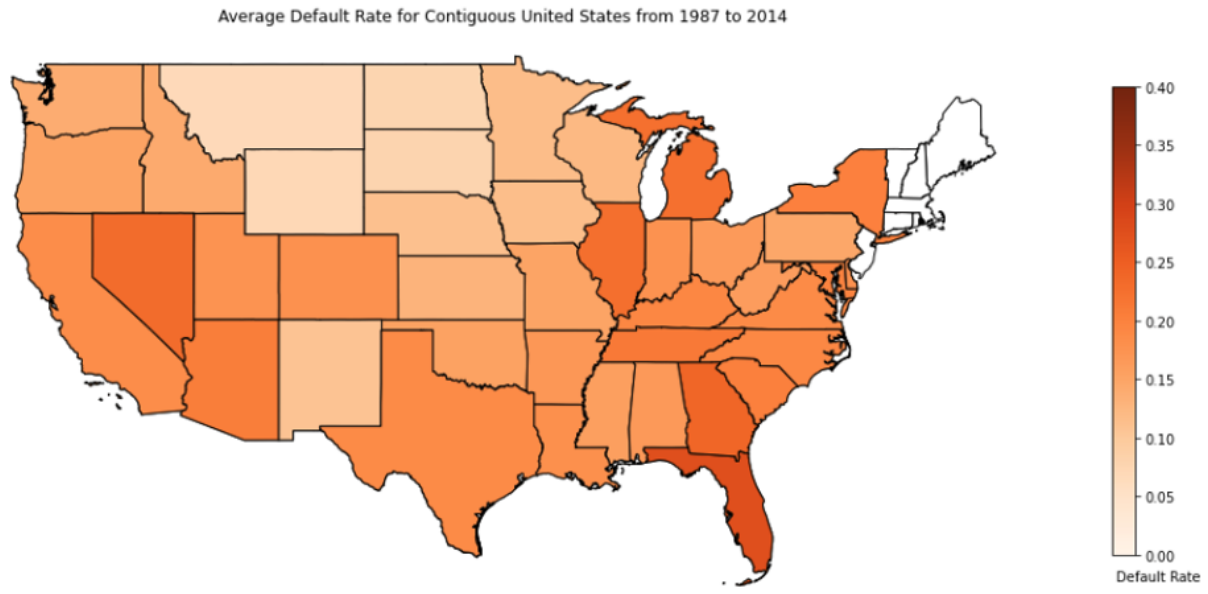

Figure 4: Average Default Rate by States for Contiguous U.S. from 1987 to 2014

Methodology

In this analysis, four machine learning classifiers were constructed and the classifier with the optimal test accuracy and AUC score was used to predict the top characteristics of businesses that default on SBA loans. The four classifiers are logistic regression, naive bayes, decision tree, and random forest classifier. The outcome variable is "MIS_Status" and the predictors were the list of 14 variables listed in the previous section. The whole dataset was split into two parts: 70% of the data were used as training dataset, and the remaining 30% were used as testing dataset. After building all the classification models, the optimal classifier was used to select top ten characteristics by comparing classifiers' test accuracy and AUC score.

Furthermore, in order to examine the impact of different levels of SBA's guaranteed loan amount, a further analysis was done by separating the dataset according to the three tertiles of SBA's guaranteed loan amount. The three tertiles are: 0% - 33%, 33% - 66%, 66% - 100%. Loans below or equal to $26640 are considered as small size loans, loans between $26,640 and $121,635 are considered medium size loans, and loans above $4,500,000 are considered large size loans. Then using the three separate datasets, a random forest classifier was used to classify each dataset. Lastly, the top ten characteristics for each dataset were shown. Detailed results will be discussed in the following section.

Results

As mentioned in the earlier sections, we constructed different machine learning classifiers namely, logistic regression, Naïve Bayes, Decision Tree and Random Forest and fitted them to the entire dataset. From the table below we can see that the optimal classifier based on the test accuracy rate and AUC score is "Random Forest" as it has an accuracy rate of 99.37% and AUC of 0.9960 (Table 3).

	Accuracy	AUC score
Logistic	0.981382	0.993536
Naive Bayes	0.961737	0.990129
Decision Tree	0.988186	0.978061
Random Forest	0.993783	0.996043

Table 3: Comparison between Accuracy Rate and AUC scores

We then used the random forest classifier to extract the important features that contributed to SBA loan default overall (not within the three subgroups). The most important features, from the first most important to the third, that we discovered are loan default amount, loan term in months, and the fiscal year of loan issue respectively. We found that the correlation between unpaid loan amount and loan default status is 0.44457, suggesting a relatively low but positive correlation between these variables. This shows, according to our data, that higher unpaid loan amount is correlated with a higher risk of default on SBA loans. In addition, our data predicts that long term loans are associated with lower risk of default, with a correlation score of -0.31579. Lastly, the fiscal year of loan issue is the third important feature in this prediction. The top features, beyond the three mentioned above are shown in Figure 5.

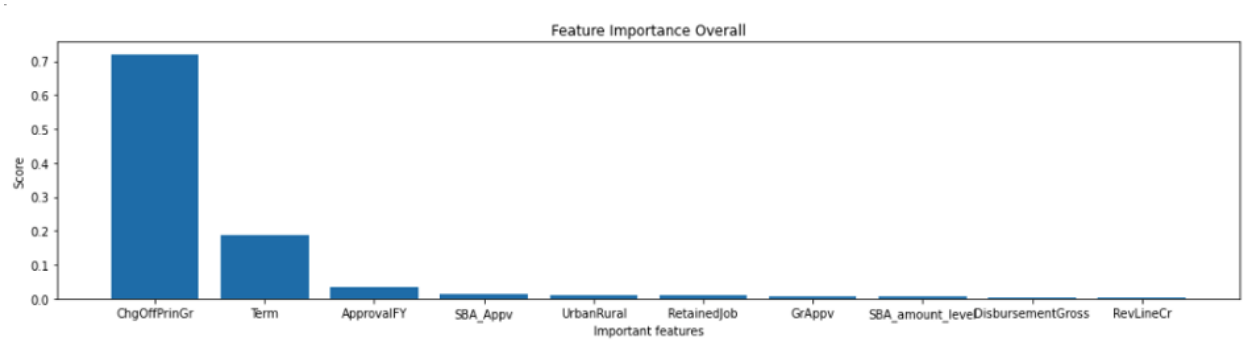

Figure 5: Feature Importance Overall

In order to examine which features contributed the most to the loan default among all three loan size groups, we used the random forest classifier (which we found to be the optimal classifier overall) to extract the top features among the three subgroups. The data suggests that the top three features for the small SBA loan size group are the same as the top three features in the overall dataset, which are loan amount that was charged off (unpaid), loan term (in months), and fiscal year of loan issue. Figure 6 shows the top features that contributed to small size SBA loan default, beyond the top three that were mentioned above.

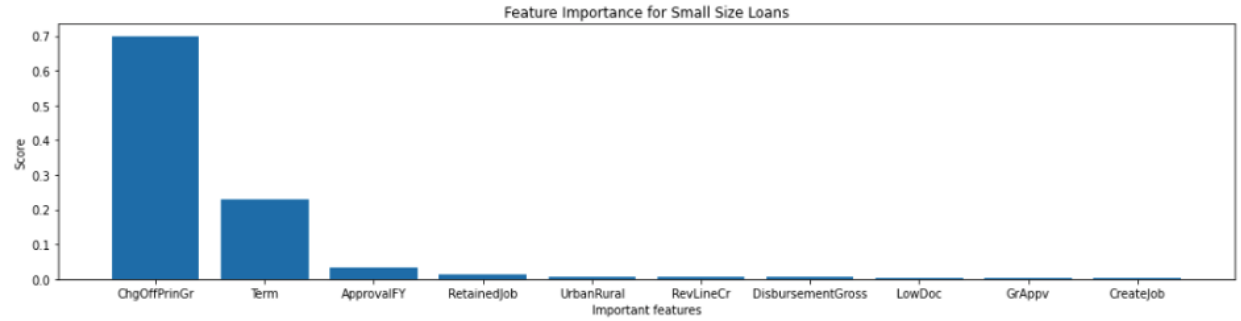

Figure 6: Feature Importance for Small Size Loans

Similarly, the top three features for medium and large size loans are the same as those for small size loans, and they are the loan default amount, loan term in months, and fiscal year of loan issue. Figure 7 and 8 show the top features that contributed to the medium and large size SBA loan default, beyond the top three that were mentioned above. Overall, our analysis suggests that there is not a major difference in the top features that contribute to the SBA loan defaults between small, medium, and large loan size loans. Furthermore, the top three features are consistent in the overall dataset and among the three subgroups, though with varying feature importance scores.

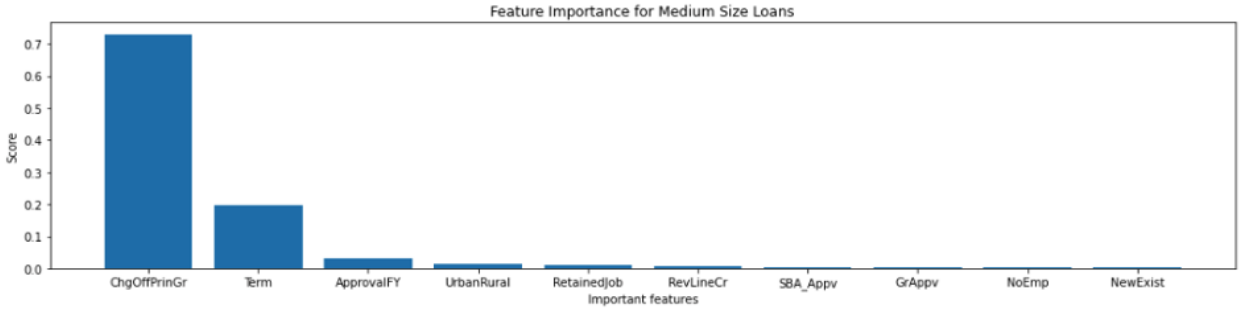

Figure 7: Feature Importance for Medium Size Loans

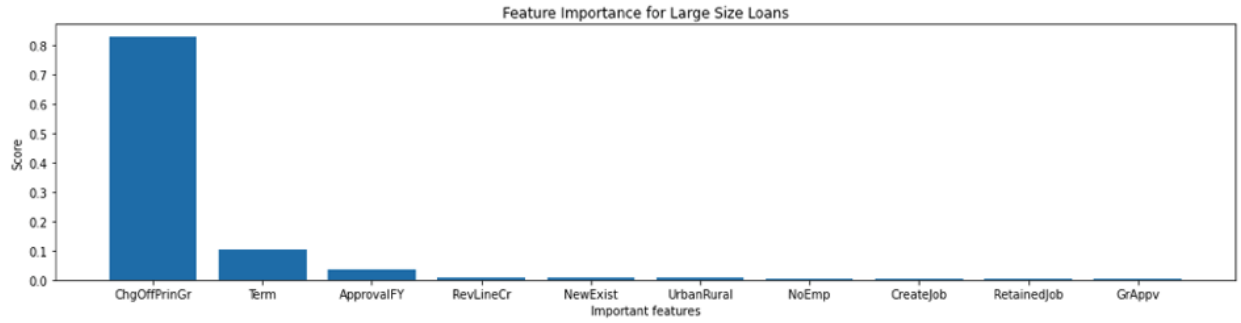

Figure 8: Feature Importance for Large Size Loans

Conclusion

We implemented four machine learning classifiers and compared their accuracy rate and AUC score in order to find the optimal classifier that best allowed us to predict SBA loan default. Out of the four classifiers, which are logistic regression, naive bayes, decision tree, and random forest, we found that random forest was the optimal classifier in predicting SBA loan default. Using the random forest classifier we examined the most important features that contributed to SBA loan default from 1987 to 2014, and found that the top three important features are loan default amount, loan term in months, and fiscal year of loan issue. As an extension to this we aimed to find the top characteristics for loan default among small, medium and large size loans by splitting the dataset into three groups based on loan amounts. Unlike Peter Nigro and Dennis Glennon's analysis that found a difference in the top features that contributed to SBA loan default among loans with different maturities, when splitting the data based on groups of loan size we found no differences in the top features that contribute to SBA loan default. Although we found that the Random Forest Classifier was an optimal way to predict SBA loan default among the classifiers we examined, a further analysis could implement the use of neural networks to predict SBA loan default.

References

Foley, R. (2012). Predicting micro LENDING loan Defaults USING SAS® Text Miner. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications,* 417-455. doi:10.1016/b978-0-12-386979-1.00019-0

Glennon, D., & Nigro, P. (2005). An analysis of sba loan defaults by maturity structure. *Journal of Financial Services Research, 28*(1-3), 77-111. doi:10.1007/s10693-005-4357-3

Lewis, D., Daniel Lewis. (2021, January 08). Small business Loan Statistics 2021: How your industry affects your Loan CHANCES. Retrieved April 11, 2021, from https://www.finimpact.com/small-business-loan-statistics/#:~:text=rates%20per%20industry.-,Latest%20SBA%20Loan%20Failure%20Rates%20by%20Industry%20Code,even%20accounting%20for%20economic%20upsets

Shepherd, M. (2021, January 27). The average small business loan amounts in 2021. Retrieved April 11, 2021, from https://www.fundera.com/business-loans/guides/average-small-business-loan-amount

Voigt, K. (2020, October 09). 1 in 6 small Business Administration Loans FAIL, study finds. Retrieved April 11, 2021, from https://www.nerdwallet.com/article/small-business/study-1-in-6-sba-small-business-administration-loans-fail