Final Draft

Ribhay Singh[1], Preston Overberg[1], and Ramon Aguirre Gomez[1]

[1]Arizona State University, Tempe, AZ 85281, USA

January 2, 2026

1 Abstract

Small businesses play an important role in the economies of the world. They spark innovation, create jobs, and increase GDP. Most small businesses that are starting out need loans, which are often labeled as a high risk loan. Much analysis has been labored to try to predict approval or denial of loans, but much less has gone into the predictions of the loans after approval. Our team wanted to analyze these loans to predict whether the loan would be charged off or paid in full.

Our team came upon a data set from the SBA, which included 3 smaller data sets separated into 10 year periods, each containing hundreds of thousands of records for small business loans under the 7(a) loan program. This data set contains characteristics about the loans as well as characteristics of the business itself.

Analysis and machine learning application yielded positive results as our best model, the random forest classifier, achieved 0.934 Accuracy and a 0.969 AUC Score. The team also experimented with a categorical boost model as well as logistic regression, which both yielded slightly lower scores than the random forest.

During the analysis, limitations emerged that were ultimately addressed, but hindered our findings from being as convincing as possible. In general, our research provided valuable insights and takeaways that answered our questions about the components of small business loans and their implications.

2 Introduction

2.1 Background

The SBA began as a loan offering program in the 1950's to help start, build and grow small businesses [14]. The SBA like many financial institutions follow certain protocols to decide which lenders are eligible setting a standard across their loans. They are only given to applicable businesses which meet certain criteria, criteria that's increased over time[11]. According to the consumer financial protection

bureau, 99.9% of U.S. businesses are small businesses, and only 16.5% use loans from banks or other financial institutions[4]. In addition, there has been a decrease in loans in the past, from 2007 to 2012 over 100,000 businesses opened, but loans decreased by 344,000 and despite a recovery from the 2008 recession, there was still a decline of 5.3% from 2022 to 2023 [18, 11].

Financial institutions use credit scores of owners, collateral, along with many other techniques to predict the likelyhood of loan default from lenders[12]. In addition machine learning is being introduced in default prediction [3, 8],but mainly in order to predict approval. Once the loan is approved, less methods of prediction are used on the loan information. Banks and loan institutions use payment history, advances, business income and various measures of success to predict default[19]. Yet, there is little report on data on the loan itself to predict the default of a loan. Data like a bank's location, FDIC number or NCUA number, along with fund delivery method and jobs supported is rarely mentionded or considered [8, 10].

By exploring the dataset from the U.S. Small Business Administration on loans 7(a), the study aims to explore **whether it is possible to predict the survival or failure of a small business based on loan characteristics after approval and business profile.**

2.2 Project plan

The majority of the project will be implemented using Python in Jupyter notebook. Python is particularly useful because of its extensive ecosystem of Data Science and Machine Learning libraries. The plan is to utilize libraries such as Numpy, scikit-learn, pandas, seaborn, Matplotlib and many more. These libraries will enable us to inspect and clean the dataset, handle missing values or outliers, visualize patterns and apply machine learning models for predictive analysis.

EDA will play a crucial role in understanding the SBA 7(a) loan dataset and finding patterns as well as relationships within its variables. The very first step would be to gain a better overall understanding of the dataset by inspecting its structure, data types, checking for missing values as well as outliers and inconsistencies. The data will then be visualized thoroughly with the help of a multitude of plots such as box plots, histograms, heatmaps and boxplots. While visualizing the dataset, we will pay special emphasis towards detecting patterns and spotting outliers which might come into use while applying feature engineering and machine learning in the later steps.

For feature importance, we can count on variable importance in a similar manner as Raphael Couronné, in the study "Random forest versus logistic regression: a large-scale benchmark experiment". We will use built in variable importance measures (VIM) and perhaps permutation based VIM to obtain the most important features/ variables [5]. Since some instances and features will be removed, this could affect the classification ratio due to the removal of default or possible payed off loans, but with over 20 million lines the impact will be minimal. In addition we will not use the whole data set at once, but rather sections which means we can also focus on areas where the loans and their data are fully present. Considering the data travels back to the 1990's we can segment by the years and build a model to focus on times of distress such as it was the 2008 recession. The division by state can also be done to focus on states with lower loans to target more lending in those areas. On the other end, we can focus on states with the most capita flow, Giving us more data and loans to work with. The possibilities with the data set we have are endless, we can focus on specific areas, times, or even banks, giving us flexibility and leniency.

There will also be 2 data sets, one with normalized variables and one with raw variables. The data

frame with raw variables will also be less processed in an attempt to reduce data noise. The normalized dataset will also have more columns as the columns with dates will be processed and year, months and quarters will be extracted. The extraction will be done to attempt to obtain more meaningful data and columns that provide useful information for our models.

3 Literature Review

Although some of these businesses are successful, a large proportion of them end up failing, with only 33.8% of them surviving after ten years of operation [15]. The US Small Business Administration was created in 1953 in order to address the challenges faced by small businesses in the nation. It serves as the primary resource for these enterprises, providing counseling, capital, and contracting expertise. The SBA 7(a) is their flagship loan program that provides financial support to small businessess that meet their eligiblity criteria [13].

The primary aim of the study is to look at prior research that has leveraged such data in order to predict the survival or failure of small businesses. Furthermore, the study also aims to compare a range of analytical approaches that have been used for prediction, spanning all the way from simple statistical models such as logistic regression to more complex models such as Random Forest and CatBoost.

Many projects have covered this topic in an attempt to gather insights on the success of small businesses around the world. One such study is from the European Journal of Operational Research for small businesses in Chile. This project looked into the adaptation of credit scoring to micro-entrepreneurs. In this project, the researchers used methods such as logistic regression to ultimately decide a binary variable of loan approval. The model created cutoff points using the probability of default, thus creating the binary classification. The project then analyzes the costs of false negatives and false positives, which in this case, are rejecting a "good" applicant and accepting a "bad" applicant. One of the results that was documented in this study was that the "average income of applicants with access to a loan is 26.44% higher than the ones without this form of financial help"[1]. We can utilize many insights and methods from this study to inform our research, which is similar in nature.

For context, there are usually two types of entrepreneurs in the realm of small businesses. The first type are necessity entrepreneurs who venture into business because of lack of employment opportunities and for survival reasons. The second type are the opportunity entrepreneurs who venture out into unexploited or underexploited business opportunities [7]. This aligns with the SBA's mission to offer financial and general guidance to both categories of entrepreneurs. The loan programs offered by SBA such as 7(a) enable struggling entrepreneurs that often lack capital to seize market opportunities [15].

Another study that occurred in the UK aimed to provide loans to start-ups that were previously excluded from the traditional credit market. Through the Start-up Loan (SUL) Scheme, the government of the UK offered these loans to over 85000 start-ups from 2012 to 2021. Many of these loans were issued to previously unemployed people, who are categorized as necessity entrepreneurs. This study attempts to learn about these differences in entrepreneurs and whether the results of their businesses differ from those coming from waged employment. The results showed that the unemployed founders received smaller loans and faced a higher default risk, whereas older and more educated unemployed entrepreneurs performed better. The loans to unemployed start-ups were cost-effective, and when considering broader societal benefits, such as enabling transitions from unemployment to self-employment, the impact of the scheme was highly positive [6].

To add to the previous, financial institutions have been around for a long time; previously, they have tried dynamic monitoring and credit scoring models, many of which use Logistic Regression or Random forest along with other neural networks [5, 10, 9, 8]. The article by Instefjord, "Loan Monitoring and Bank Risk", dives deeper into the use of loan risk management. and delineates how banks sell risky loans to drop their risk levels, and they keep loans of lower risk to keep risk low and make a profit from those loans. In order to achieve this, there is a need for a predictive model, but previous models have had little to no success. As stated in the text "The reason is that the learning process that is generated by the dynamic monitoring systems is slow and which will increase the number of risky loans under observation, which in turn leads to extra net risk." [10]. This gives an opening for a model that is not constantly monitoring information about loans. Considering we don't have live data this will not be an issue, on the other hand, it will allow us to explore a different model, which won't have to learn through the slow passage of time and can predict the risk a new loan brings along based on previously learned patterns.

Previous prediction models are one of two, pre-approval models and dynamic monitoring models. Pre-approval prediction models help institutions reduce their risk by predicting bad loans. Dynamic models on the other hand, reduce risk by monitoring live data, such as payments and credit credibility [10]. In contrast, we are trying to predict the risk of default on approved loans without the need for continuous monitoring due to its high cost and inherited risk through such an investment. In addition, positive results could benefit banks and their shareholders, but also their customers through more lenient regulations on loans.

Previous models have had issues with data imbalance along with data validity [9, 5]. This is better discussed in the article by Yazhe Li where they use linear regression since its the most common and solve the issue by clustering the smaller class and labeling it that way. In a similar manner, we will handle the imperfections through regularization such as lasso and ridge regularization along with feature selection and various parameter regularization.

Developing new models and new ways to characterize the validity and risk of loans can be helpful in many ways. The first is due to the failure of previous models, as stated earlier. In the study by Instefjord, models required extensive training which means an investment not at use by the bank while it learns. In the case that it is used, it increases the risk burden, as it compels banks to keep the loans under surveillance even if the loans are of high risk. Therefore, a system that does not monitor but can assess risk can be helpful to financial institutions. As mentioned before, unemployed individuals or "struggling individuals" are less likely to obtain funding or obtain less. In addition, there are also untapped business markets which are also less likely to receive funding. Therefore, it can help more individuals if banks can lower their loan standards and have a better picture of individuals with credit problems, unemployment, and more, without experiencing a high risk.

Throughout the years, models for banks and other financial institutions have been made using various data from countries like Greece and Chile [1, 8]. The most popular features used come from credit history and other personal information. In a research study done in Greek banks during the period from 2010 - 2012, analyzed the use of various credit scoring models and used the five C's, "the applicant's character, the collateral of the loan, the capital, the capacity, and the economic conditions"[8], all factors which will not be implemented here. In a way, we will do a long tail feature analysis for our model since we won't use the previously mentioned features, and will be using the left over data after a loan is approved and resolved. Variable selection will be important due to the nature of our data set. Unlike previous research, we are not dealing with multiple data sets which means all of our data uses the same variables and in-turn features. Along with this we have not picked specific variables to extract like it was the case with Giannapoulous which means not all of our variables will be used. For this we can

count on variable importance in a similar manner as Raphael Couronné, in the study "Random forest versus logistic regression: a large-scale benchmark experiment".Will use built in variable importance measures (VIM) and perhaps permutation based VIM in order to obtain the most important features/ variables [5].

The very first approaches to predicting small business loan outcome mainly involved straightforward statistical methods such as logistic regression. Logistic regression was favored due to its well-understood features and wide availability, making it simple to implement [17]. Even though logistic regression is easy to implement, it underperforms on highly imbalanced datasets which is a common characteristic in credit card scoring [2]. In order to fix this, researchers came up with a new approach using Lasso Logistic Regression. This approach extends the existing logistic regression model and applies the L1 constraint, allowing both shrinkage and variable selection at the same time. Lasso Logistic Regression outperforms standard logistic regression, random forest and CART models in terms of AUC and F-measure [17].

Over the past few years, there has been a shift towards ensemble machine learning models such as random forest and gradient boosting because of their strong predictive power. In fact, these newer approaches have proven to be more optimal performers than the earlier used statistical methods in terms of credit scoring and loan default prediction. In the case of problems involving imbalanced datasets such as credit card scoring or business survival, models such as random forest had far better predictive performance than the conventional logistic regression approach [2].

Gradient boosting algorithms such as CatBoost have also become a popular option for credit card scoring and loan prediction. CatBoost is noted to perform quite well with datasets containing categorical variables while also being optimal for capturing complex and non linear relationships. This model was found to outperform other models such as XGBoost and Random Forest when directly applied to the SBA dataset. CatBoost also has a feature importance metric which can be particularly useful in highlighting which borrower and loan characteristics had the greatest influence in the prediction [16].

4 Exploratory Data Analysis

4.1 Data visualization

4.1.1 Uni-variate analysis

Initial exploration was done by examining each feature individually, below we can observe the multitude of outliers. In addition, we can observe that there are more paid off loans with higher approvals. The defaulted loans are usually smaller than those paid off, which can represent a trend that, the bigger loans/investments are more likely to be paid off.

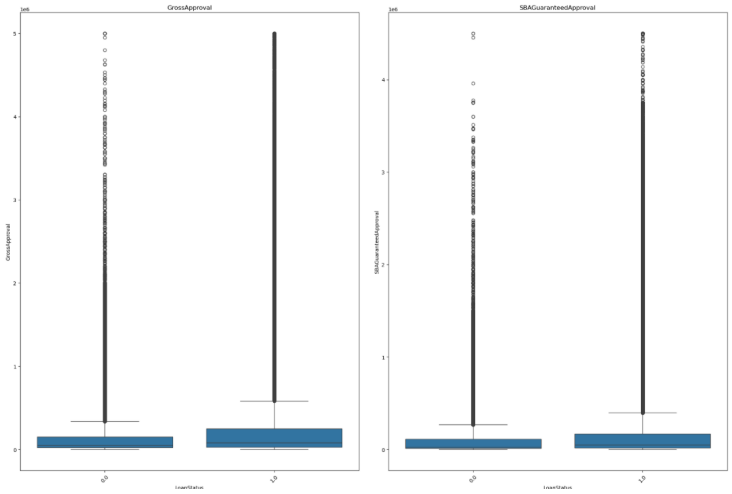

Figure 1: The right hand side of the graphs, 1, shows the greater gross approval and SBA guaranteed approval for loans that were paid off meaning loans with smaller approvals could be at a higher risk of default

After Examining the trends in approved loans we look for patters in loan length and any trends that may arise.

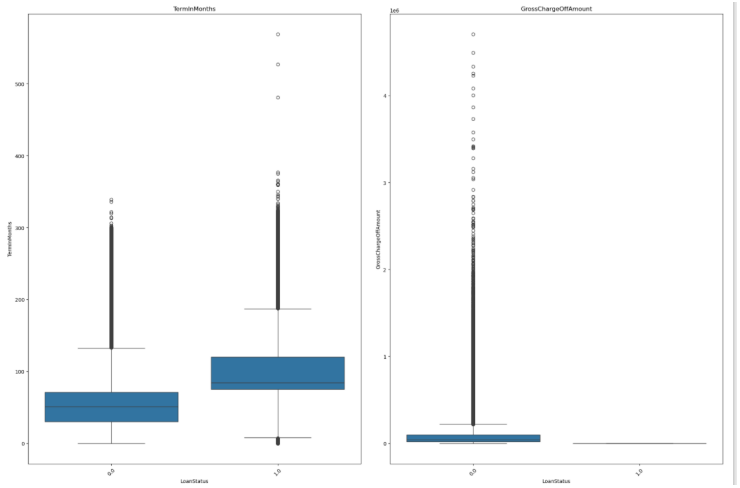

Figure 2: The Loans that were paid off, rather than defaulted or charged off show they had longer terms. This means the loans which were given longer periods to be paid off seem to be more common in the paid off section, while shorter loans were more likely to default.

From the above graphs we can observe that there is a possibility of data leakege if the column GrossChargedOffAmount is used therefore it will also be removed.

When comparing features against the target variable, boxplots are some of the most useful visualizations to examine the distributions of features for each value of the target variable and the outliers. However, the box plots from this data set don't provide very many useful insights. It appears there are

a lot of outliers, but this data set is extremely large, so there will naturally be a lot of outliers for each feature. There are some insights that these box plots show us, like the summary statistics and how the variable can behave based on the value of the target variable, which in this case is LoanStatus. Many of these box plots are compressed, however, making them hard to read.

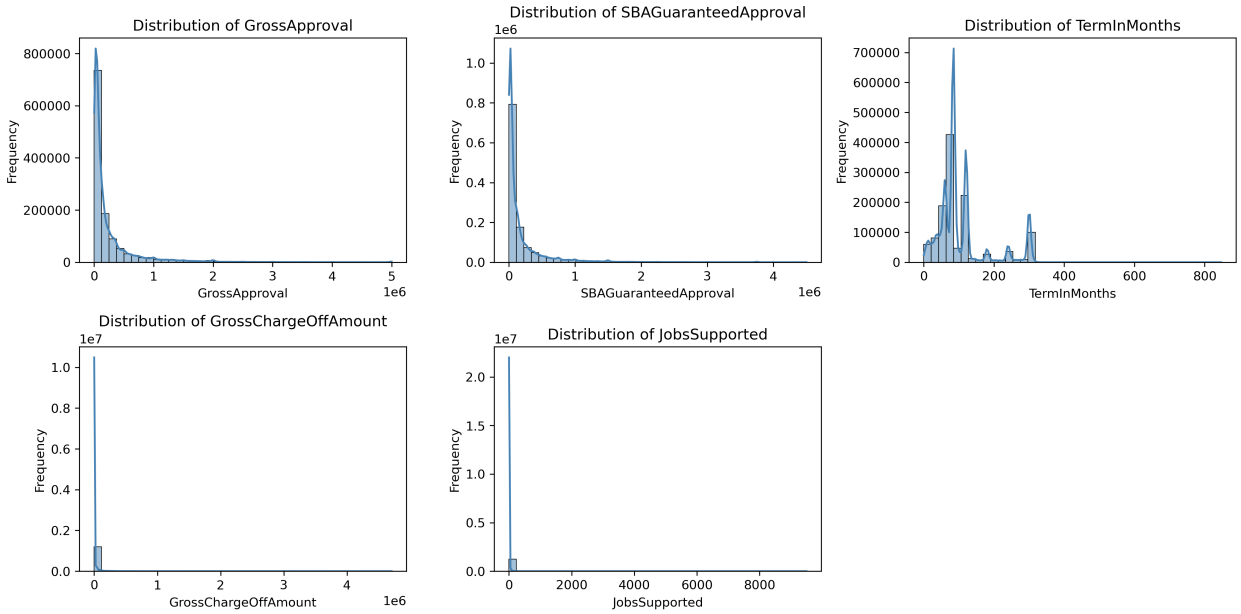

Figure 3: Distribution of key numeric variables including Gross Approval, SBA Guaranteed Approval, Term in Months, Gross Charge-Off Amount, and Jobs Supported. The heavy right skew indicates that most loans and charge-offs are small, while a few high-value outliers dominate the upper range.

To gain a deep initial understanding of the dataset, our team conducted univariate analysis focusing on the distribution of key numerical and categorical variables. For the numerical variables, we observed factors such as Gross Approval, SBA Guaranteed Approval, Term in months, Gross Charge off amount and jobs supported.

The histograms revealed that most variables show significant right skewness which is a common pattern seen in financial data such as this one. For instance, we uncovered that factors such as Gross Approval and SBA Guaranteed Approval show that the majority of the loans were approved for smaller amounts with just a few abnormally large loans, creating long tails in the overall distribution. On a similar note, Gross Charge off Amount and Jobs Supported are concentrated along lower values as well indicating that the majority of the loans involved limited charge off and supported a small number of jobs even though there are rare cases of significant financial exposure or loans with high impact.

4.1.2 Multivariate Analysis

Year, months and quarterly values have been extracted from date columns like ... and ... for better handling for the models.

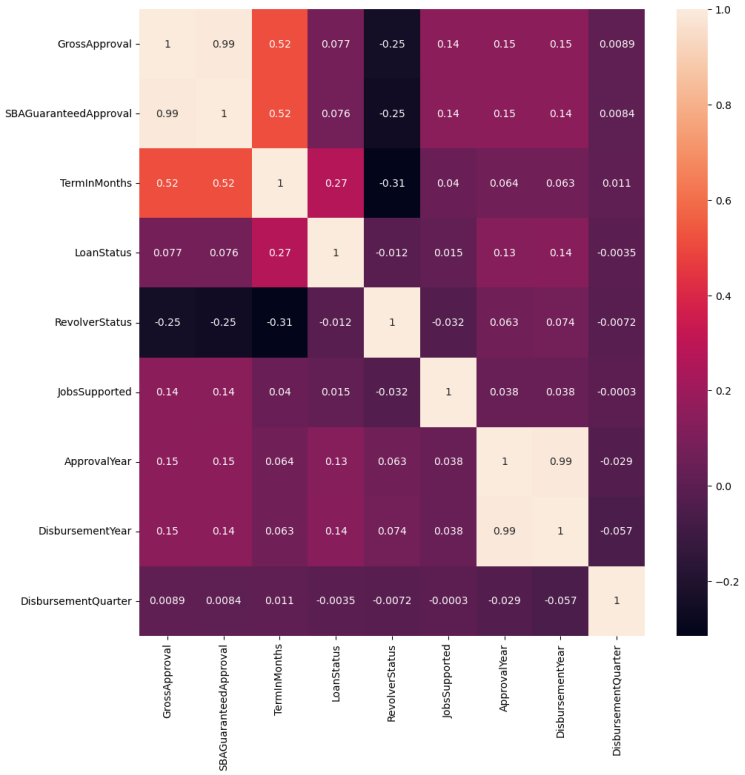

Figure 4: Correlation heatmap showing relationships between key loan-related numeric variables such as Gross Approval, SBA Guaranteed Approval, and Term in Months. Strong correlations indicate potential feature importance for predictive modeling

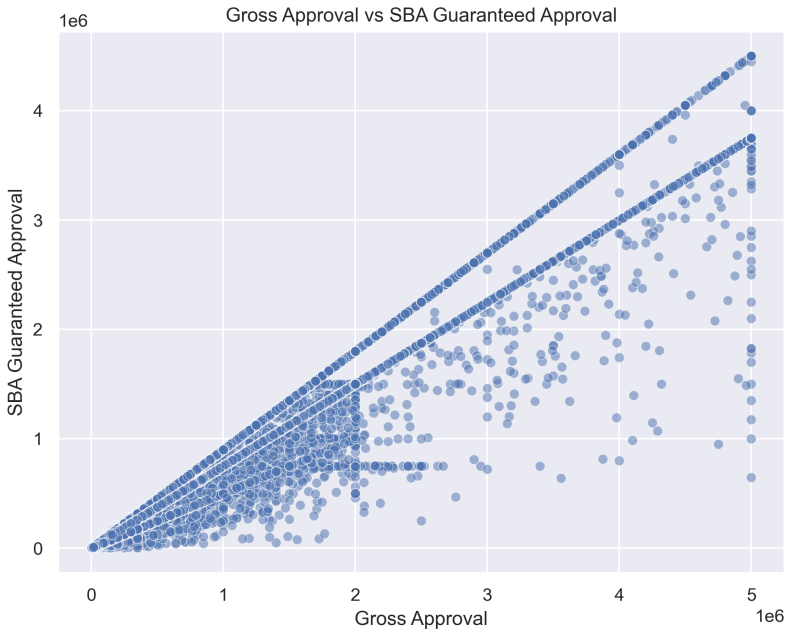

Figure 5: Scatterplot demonstrating the strong linear relationship between Gross Approval and SBA Guaranteed Approval, confirming the correlation observed in the heatmap.

Getting a good understanding of the relationships between key features is an essential part of the EDA process. This notion is further backed by Tukey's book on EDA which states that feature relationships through correlation analysis helps detect redundant or multicollinear variables prior to modeling [3] We first plotted a correlation heatmap to examine loan features such as Gross Approval, SBA Guaranteed Approval, Term in Months, Jobs Supported, Revolver Status, Gross Charge Off Amount ane Loan Status. The heatmap gave our team insights about strong positive correlation between the features such as Gross Approval and SBAGuaranteedApproval which indicated that large loan approvals has higher SBA backing.

4.1.3 Time-Based Analysis

Another important area of the EDA to examine is the changing of the data over time. Analyzing how the data moves, what happens, and maybe why it happens is key to understanding the past in order to predict the future. Analyzing multiple features over time gives us a much larger picture into the causation of data movement. These next few visuals capture important features of the data over time.

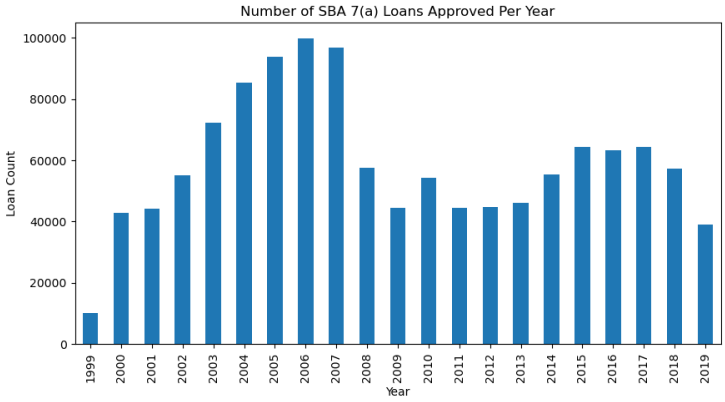

Figure 6: The number of SBA 7(a) loan approvals each year. Approvals steadily increased during the early 2000s, peaking around 2006–2007. A decline appears during the 2008–2009 financial crisis, followed by stabilization and recovery through the mid-2010s, reflecting trends in small business financing demand.

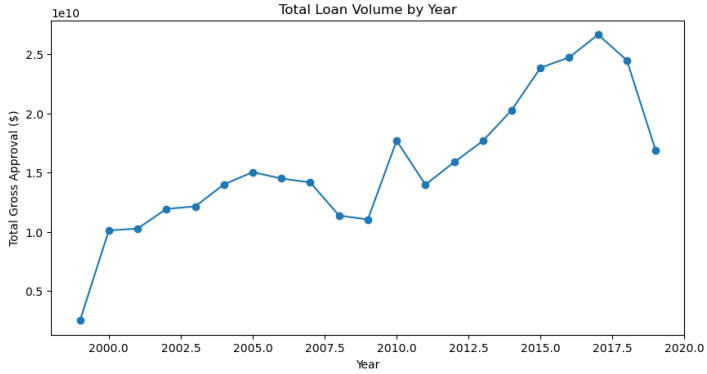

Figure 7: Total SBA 7(a) loan volume by year. Despite fluctuations in loan counts, total lending volume shows an upward trend, especially after 2010. This indicates larger loan amounts being approved over time, with a peak around 2015–2017, followed by a decline toward 2019.

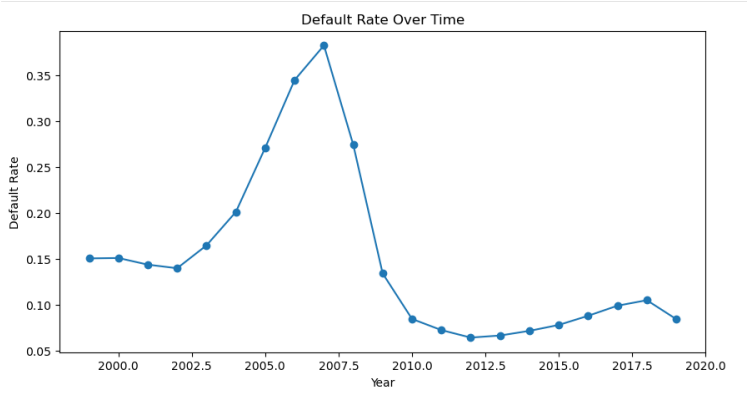

Figure 8: Default rate of SBA 7(a) loans over time. The default rate remains relatively stable in the early 2000s before sharply peaking around 2007–2008 coinciding with the financial crisis. After 2009, defaults decline dramatically and stabilize at lower levels, reflecting improved lending standards and economic recovery. Minor fluctuations after 2016 suggest modest increases in risk exposure in recent years.

When comparing the Number of Loans approved to the Gross Approval Amount, the trends don't seem to match up as much as one might expect. They are roughly correlated, but total Gross Approval for the most part is continuously increasing, while Number of Approvals drops and stays relatively low.

It is evident that something happens around 2008, because the amount of loans approved drops by a lot, and the default rate goes down. From prior knowledge we know there was a recession in 2008, which could certianly be the case for the change in these metrics. It is important to inspect visualizations like these to check for any unexpected changes that might require more inspection.

4.1.4 Geo-positional Analysis

The following visualizations provide our team with a comprehensive understanding of the geographic distribution patterns within the SBA loan dataset. Specifically, by expertly utilizing Tableau's advanced features, we were able to effectively leverage its Geo-spatial visualization capabilities. This allowed us to clearly and accurately identify distinct regional patterns, pinpoint areas of significant loan concentration, and gain detailed insights into regions exhibiting higher rates of either loan defaults or approval activities. These crucial geographic insights are instrumental in providing a more nuanced context for understanding broader trends within the dataset and in highlighting the considerable variations that exist in lending outcomes when analyzed across different states and geographical regions.

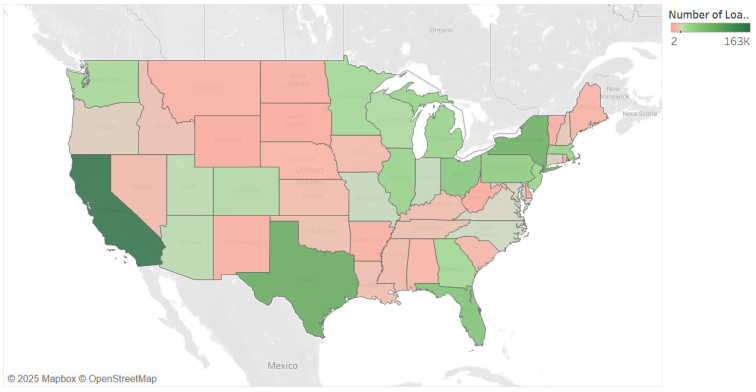

Figure 9: Number of loans per state. Highest being in California, followed by Texas, New York and Florida.

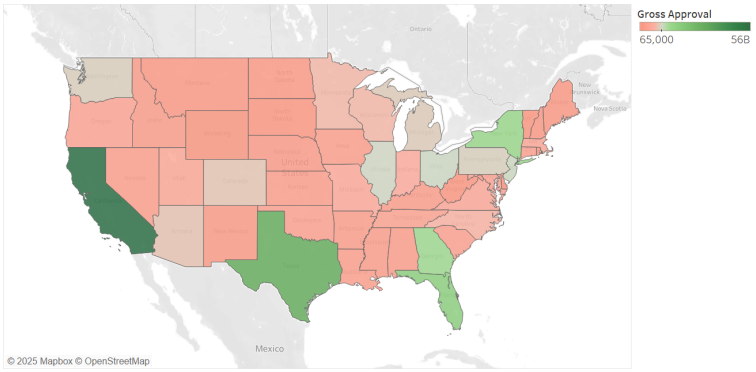

Figure 10: SBA loan gross approval per state. The pink states like Wyoming and South Dakota have lower gross approval than California and Texas. The darker the green, the more loans and money that have been approved for small businesses in those states

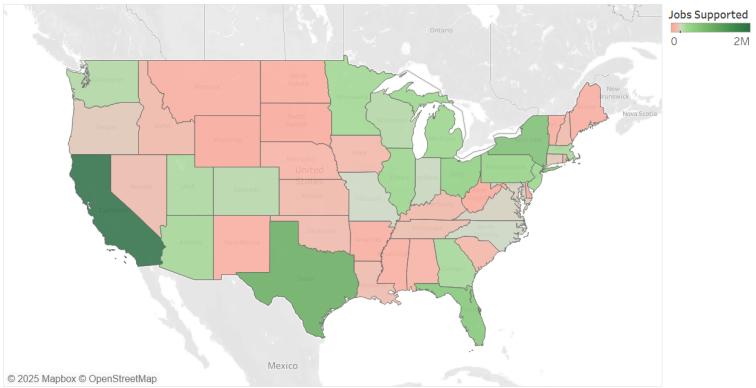

Figure 11: Total number of jobs supported by these loans in each state. Dark pink include the states with the least jobs supported by the SBA loans. Darker green states like California, New York and Florida, show higher job support by small business loans.

As you can see in each one of these graphs, the saturation is in largely populated states, those with more urban areas and large populations. This is expected as urban areas tend to have a higher opportunity for business to spring up and succeed compared to rural areas. In each one of these graphs, California is a dark green, signifying the highest state in each one of these metrics. Texas, New York, Florida, Georgia, and a few others are consistently green as well. States like Nebraska, Iowa, Arkansas, Wyoming have low numbers in each of these, which is also as expected given their population and lack of urban areas.

4.2 Dataset Overview

The dataset used in the study is sourced from the Small Business Administration 7(a) Loan program found in Kaggle issued between 1990 to 2019 [14]. The data was originally provided in separate files and categorized according to their respective time periods. Each time period consists of 10 years each, beginning with our first data set from 1990- 1999, then we have a period of loans from 2000- 2009, and lastly our data containing the years 2010-2019.

The first data set (1990-1999) has the most missing entries, out of 337,043 there are rows which have every value missing meaning we will have to get rid of rows with no information. Below we can see 'BankFDICNumber', 'BankNCUANumber', 'FranchiseCode' and 'BusinessAge' are among the columns with the most missing entries. This set has a default ratio of 35,362 Charged off to 255,851 Paid in full, giving us a 0.1382 ratio.

Then we have our second data set from 2000- 2009 with 690,347 lines missing data in various columns, yet like in the previous, it does not mean all rows are empty, but there are columns which have no data for all 600 thousand instances. In a similar manner below we see 'BankFDICNumber', 'BankNCUANumber', 'FranchiseCode' and 'BusinessAge' are columns with some of the most missing values. This set has a default ratio of 146,540 Charged off to 453,654 Paid in full, giving us a 0.3230 ratio.

Lastly, in our final data set we have 545,751 lines in which various columns are missing vast amounts of data. In the same fashion as the previous sets, the most missing values appear in columns like 'BankFDICNumber', 'BankNCUANumber', 'FranchiseCode' and 'BusinessAge'. This set has a default ratio of 27,916 Charged off to 319,229 Paid in full, giving us a 0.0874 ratio.

4.3 Data Cleaning and Preparation

Data cleaning was one of the most crucial portions of our project because the final results would be heavily dependent on it. Once the two datasets from the two time periods were concatenated, they were thoroughly inspected using a variety of operations such as df.info(), df.describe(), df.shape(), df.isnull().sum() among many others. These operations gave our team strong insights about the structure of the data, missing values and overall distribution of data. For example, we uncovered that certain variables such as LoanStatus and GrossApproval were almost entirely complete while others such as FranchiseCode and BusinessAge had very high numbers of missing values. We also uncovered that there were a large number of features that held little to no value in our analysis. For example, the feature AsOfDate has one singular value "2022/12/31" which is the date our data was retreived. This singular value throughout has no negative or positive effect on data; it is simply a feature of little to no importance.

AsOfDate	0
Program	0
BorrName	39
BorrStreet	30
BorrCity	11
BorrState	11
BorrZip	0
BankName	0
BankFDICNumber	109199
BankNCUANumber	1208533
BankStreet	45
BankCity	44
BankState	51
BankZip	44
GrossApproval	11
SBAGuaranteedApproval	11
ApprovalDate	0
ApprovalFiscalYear	0
FirstDisbursementDate	147754
DeliveryMethod	11
subpgmdesc	0
InitialInterestRate	653713
TermInMonths	0
NaicsCode	19742
NaicsDescription	20390
FranchiseCode	1148834
FranchiseName	1148947
ProjectCounty	188
ProjectState	11
SBADistrictOffice	11
CongressionalDistrict	2397
BusinessType	49
BusinessAge	1124425
LoanStatus	0
PaidInFullDate	463215
ChargeOffDate	1061647
GrossChargeOffAmount	0
RevolverStatus	0
JobsSupported	0

Figure 12: Null values per column in the main data frame merged from the 2000's and 2010's time periods before data cleaning

Based on these observations, a very specific data cleaning plan was implemented. Features that had a high amount of null values or those that were redundant, such as Franchise Code, Franchise Name, Bank NCUA number among others were dropped. Categorical fields such as NaicsDescription and Business Type were filled with "Unknown" and mode values to retain the overall structure but also prevent any bias. The target variable, which is LoanStatus also had significant revisions made to it during the cleaning phase. In specific, the LoanStatus column was converted into a binary variable to define the prediction target. Loans that were marked as "PIF" (paid in full) were encoded as 1 to signify successful repayment while those marked as "CHGOFF" (charged off) were encoded as 0 to signify failure. In addition, the loans marked as 'CANCLD', 'EXEMPT' and 'COMMIT', were dropped as they were neither paid off nor charged off, they were special cases which we were not interested in. From this point on, any columns with high amounts of data missing have been dropped along with any duplicates found.

Previously, our histograms show our data as right skewed, mainly in the GuranteedApproval and

GrossApproval columns which are full of financial information. While there are other columns which also show a right skewed-ness, they do not need to be normalized as their data is not linear or numerical as in numerically independent, they are more categorical than numerical. Given this, we have the following graphs after normalization of GuranteedApproval and GrossApproval columns.

After the data cleaning process was completed, the resulting dataset became much more usable and geared towards our objective. The cleaned dataset still has around a million high quality rows and 26 features, making it perfect for our analysis. Furthermore, it is now well aligned towards machine learning which will be implemented in the future. All in all, the cleaned dataset achieved the right balance between completeness and noise reduction.

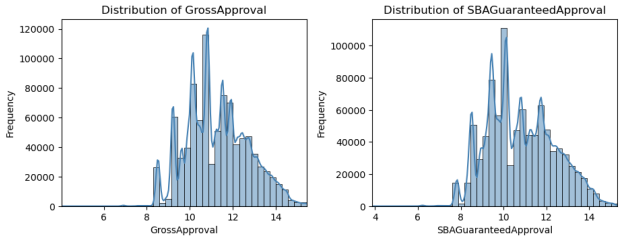

Figure 13: Normalized Gross Approval and Guaranteed Approval columns for better prediction.

5 Methods Used for Prediction

5.1 Logistic Regression

The very first model that was tested by the team was logistic regression. We decided to include logistic regression in our analysis because of its well understood feature and wide availability, making the implementation easy and straightforward [17]. Well before the implementation of this model, we were well aware that logistic regression would not be the best performing model as was revealed from previous research. Similar to all of the models implemented in this project, the logistic regression model was set up using Python's scikit learn library.

```
Accuracy : 0.7478
F1-score : 0.8267
ROC AUC  : 0.7949

Classification Report:
              precision    recall  f1-score   support

           0      0.406     0.791     0.537     34248
           1      0.940     0.738     0.827    151340

    accuracy                          0.748    185588
   macro avg      0.673     0.765     0.682    185588
weighted avg      0.841     0.748     0.773    185588
```

Figure 14: The Evaluation Metrics used on The Logistic Regression Base Model.

5.2 Random Forest

Random Forest was an obvious choice as a model candidate because of its ability to detect trends in complex data without overfitting. Our random forest model performed quite well on the testing data, achieving an accuracy of 0.93 and an AUC score of 0.969. One limitation that we can certainly see which has been outlined in other sections has been the imbalanced data set. With so many more Paid in Full observations than Charged Off observations, the random forest model was not as efficient in predicting the Charged Off loans. It achieved a precision score of 0.84, recall of 0.77 and f1-score of .80 for Charged Off loans. These numbers are all much lower than the scores of Paid in Full loans. Although these scores are not terrible for Charged Off loans, the imbalanced data set can skew the model into predicting PIF more than it should, considering that more often than not if it guesses PIF, then it is correct.

	precision	recall	f1-score	support
0	0.84	0.77	0.80	34248
1	0.95	0.97	0.96	151340
accuracy			0.93	185588
macro avg	0.89	0.87	0.88	185588
weighted avg	0.93	0.93	0.93	185588

Figure 15: This figure shows the scoring summary for the random forest model

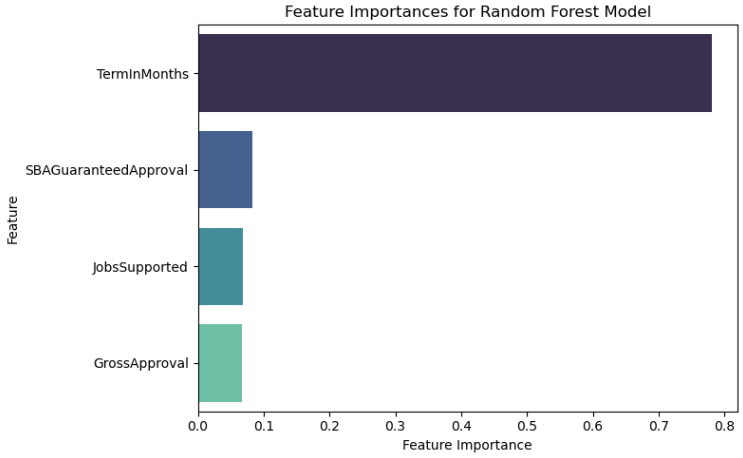

Figure 16: The figure shows the random forest coefficients for the predictors. The results are very similar to the logistic regression graph since most of the relationships hold in this case too.

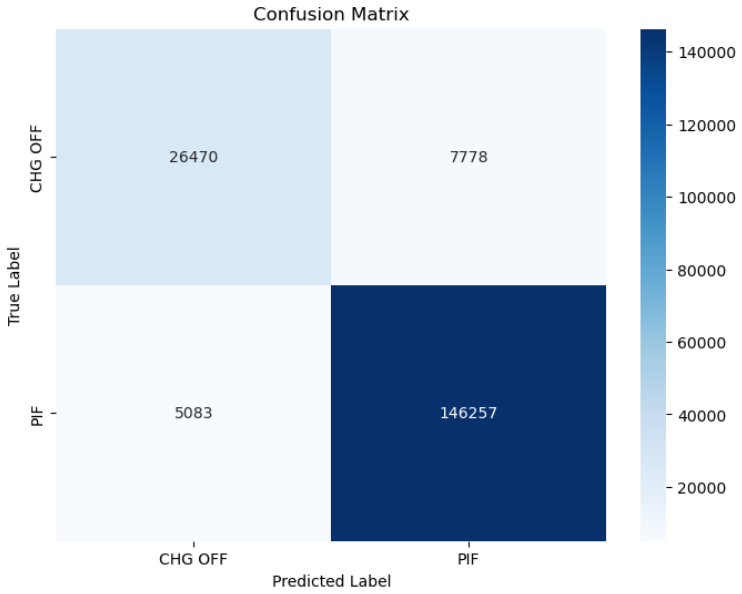

Figure 17: Confusion Matrix for Random Forest Model

From the confusion matrix above, we can see the imbalance of the target variable with Paid In Full being much larger for true predictions than charged off loans. One item to note is that even though there were so many more Paid In Full loans, there were more false positives than false negatives, meaning the model classified 7778 Charged Off loans as Paid in Full compared to 5083 Paid in Full loans classified as Charged Off by the model. This accounts to 23% of Charged off loans classified incorrectly and 3% of Paid in Full loans classified incorrectly. These numbers are demonstrated by the recall in the outputs as well. This difference was expected by our team as we were aware of what the imbalance in the target variable would influence.

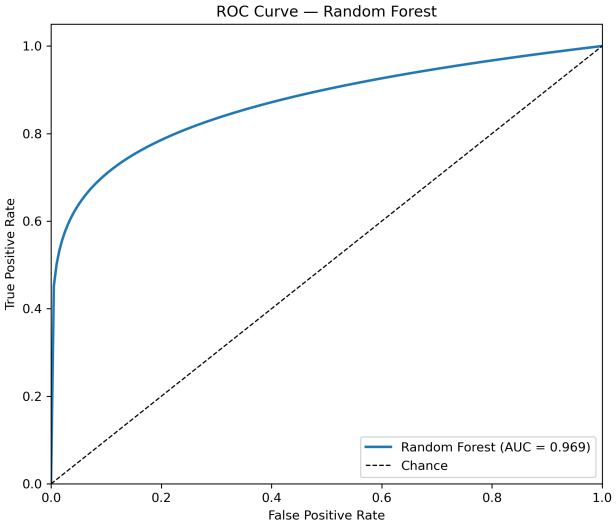

Figure 18: Random Forest ROC Curve (AUC = 0.969)

5.3 Categorical Boosting

Catboost was chosen because it is noted to perform quite well with datasets containing categorical variables while also being optimal for capturing complex and non linear relationships. This model was found to outperform other models such as XGBoost and Random Forest when directly applied to the SBA dataset. CatBoost also has a feature importance metric which can be particularly useful in highlighting which borrower and loan characteristics had the greatest influence in the prediction [16]. It also does internal cross verification also simplifies the creation of the model but increases the internal complexity of the model. This results in longer interpretation time which means the model takes longer to render, load and compile.

```
Accuracy:  0.9071545574067289
ROC AUC: 0.8188657654471414
              precision    recall  f1-score   support

           0       0.79      0.68      0.73     34248
           1       0.93      0.96      0.94    151340

    accuracy                           0.91    185588
   macro avg       0.86      0.82      0.84    185588
weighted avg       0.90      0.91      0.90    185588
```

Figure 19: Evaluation Metrics for base CatBoost Model

Here we bagan with a model that uses accuracy as is evaluation metric and uses only one text feature 'NAISCDescription'. This results in a model with a high accuracy of 0.907, yet it's far from perfect. The resulting accuracy comes with a AUC Score of 0.819 which is better than .5 as it tells us we are not guessing in our classifications. But while our AUC score is high, so is our false positive and false negative rates as we can see in our confusion matrix below. This means we are not getting all the zeros correct or we are labeling some zeros as ones when it should be the opposite.

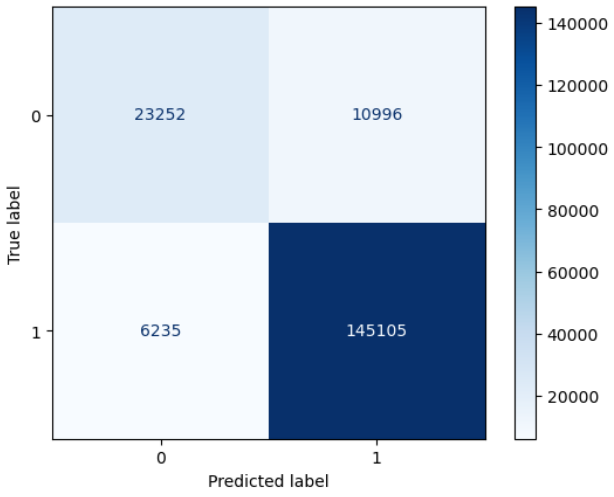

Figure 20: Confusion Matrix for base CatBoost Model

Using the Confusion Matrix along with the ROC AUC curve below we can observe that the True positive rate is not in the best condition. The curvature of the line indicates that the True Positive rate starts to decrease as we begin increasing the FPR and its incremental rate slows down. In our confusion matrix this is visible as we see a high number of false positives on the top right corner.

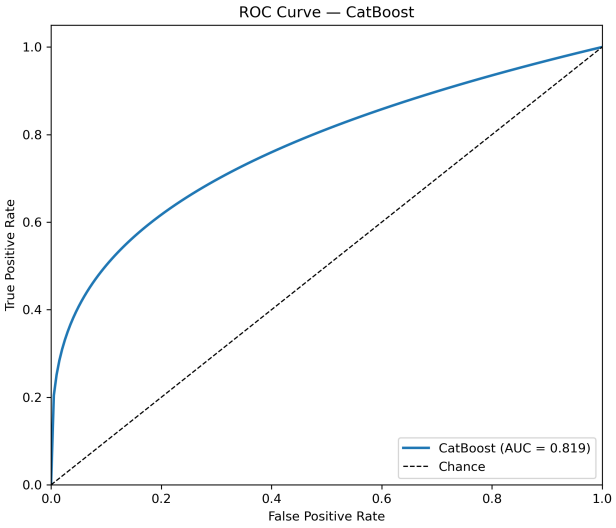

Figure 21: CatBoost ROC Curve (AUC = 0.819)

The base model uses features such as 'TermInMonths', 'subpgdesc', 'ApprovalFiscalYear' the most. On the other hand, features such as 'GrosApproval' and 'SBADisstrictOffice' have little to no impact on the model as a whole. Overall, this gives us a good insight into which features are truly important and which ones are of negligible relevance.

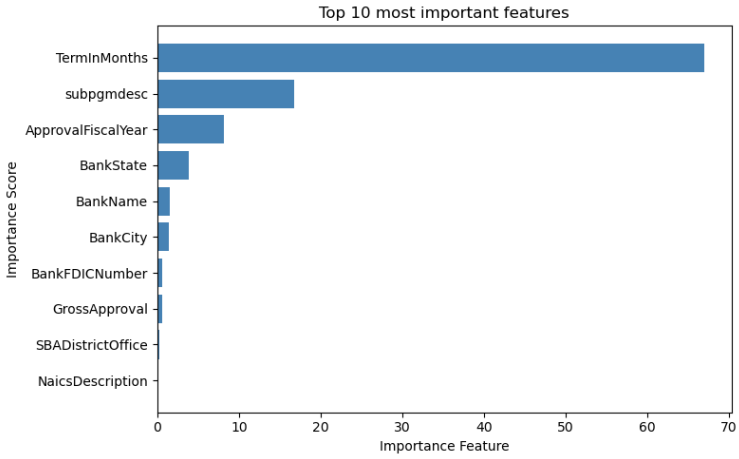

Figure 22: The Graph above demonstrates the variable importance in Ascending order. These are the most useful features for our base model.

6 Final Results

6.1 Overview

The results from our base models and hypertuned models both result in Logistic regression being subpar when compared to Random Forest and Categorical Boost. Although base logistic regression did have an Accuracy score of 0.746 and an AUC score of 0.794, it is not as high as Random Forest with 0.934 Accuracy and 0.969 AUC Score. When compared with CatBoost it also is lower as Catboost achieved accuracy of 0.907 and 0.819 AUC Score. This all means we are not guessing with any of our models nor are they randomly assigning classes. The Models are using data to classify our instances but some are using more data than others.

Logistic regression and Random Forest use 4 out of 29 Features available which means outliers and any noise in those 4 columns will greatly impact the model. Categorical Boost on the other hand uses 8 columns which can increase our predictive power by capturing more patters within the data. The Representation of complex relationships is also better handled and can decrease the bias within features. This does come with an increased risk of overfitting though as well as well as higher computational cost and time.

On hypertuned models, the final results of logistic regression fall significantly short when compared to both the random forest and Catboost models. Furthermore, from our tuned Categorical model we obtain an accuracy of 0.909 and AUC of 0.936. This Shows an increase from our base CatBoost model as our AUC is increase by 0.1 or 10%. The content that follows will go in depth into the results, performance, and the nuances of each model and how they stack up against each other. (shorten)

6.2 Final Logistic Regression Model Performance

The team utilized grid search on the logistic regression model due to the low amount of numerical columns available which limits the number of possible combinations. This limit in complexity negatively affected our ability to predict since we were not able explore as many patterns or correlations that may be present. Although we did see an improvement in Accuracy to 0.7478 and ROC AUC to 0.7949, the F1-Score remained the same at 0.8267. This was performed using the same features as above.

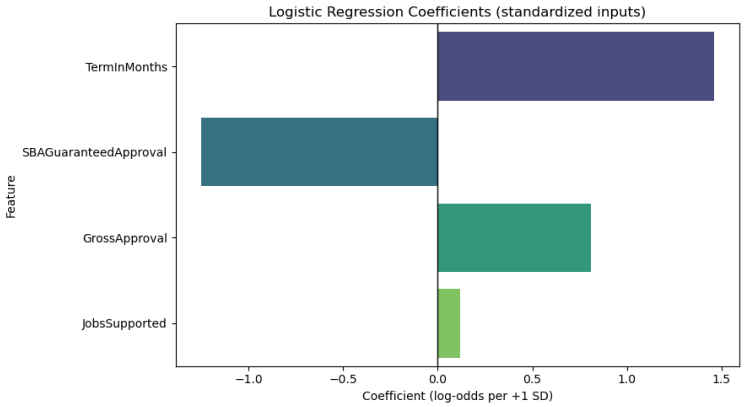

Figure 23: Feature Importance for the model with the best parameters found through Grid Search.

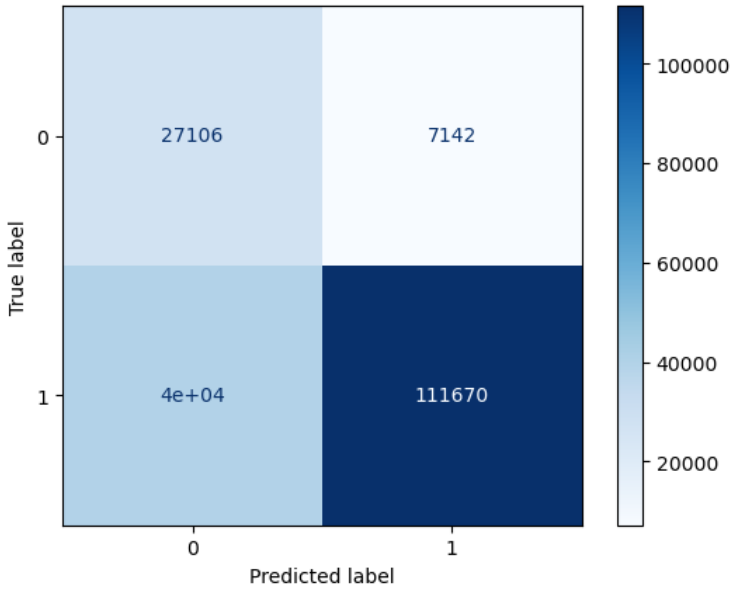

Figure 24: Confusion Matrix for the model with the best parameters found through Grid Search.

The final results as shown by the ROC curve from the model with the best parameters found demonstrates the low amount of changes in the models.

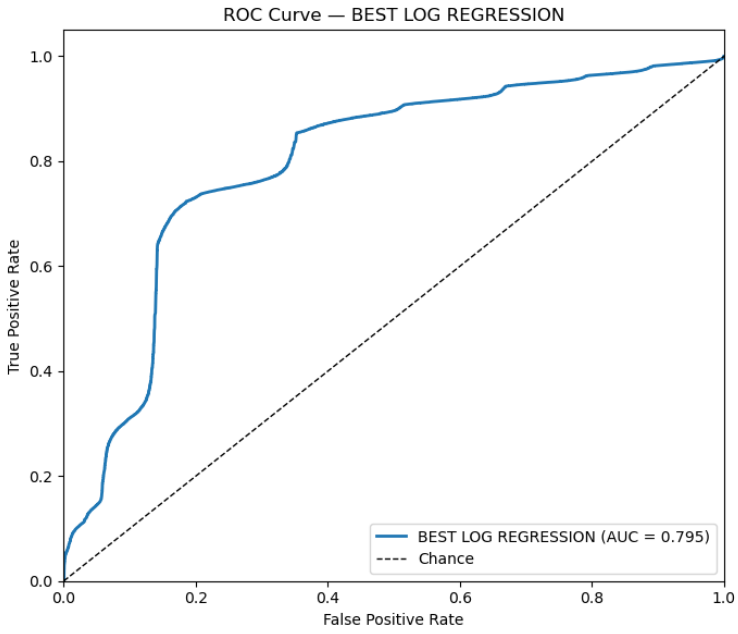

Figure 25: ROC curve for the model with the best parameters found through Grid Search.

Therefore, despite additional attempts to improve the performance of the logistic regression model, it still fared the poorest when directly compared to the other models.

6.3 Normalized Logistic Regression Model Performance

Logistic regression saw a slight change in performance scores from the normalization of the data. However, these increases were quite small and even negative, with no significant impacts of the model in the big picture. Thus, the normalized data did not improve our logistic regression model.

```
Accuracy : 0.748
F1-score : 0.8268
ROC AUC  : 0.7965

Classification Report:
              precision    recall  f1-score   support

           0      0.407     0.792     0.537     34217
           1      0.940     0.738     0.827    150866

    accuracy                          0.748    185083
   macro avg      0.673     0.765     0.682    185083
weighted avg      0.841     0.748     0.773    185083
```

Figure 26: Normalized Logistic Regression Classification Report

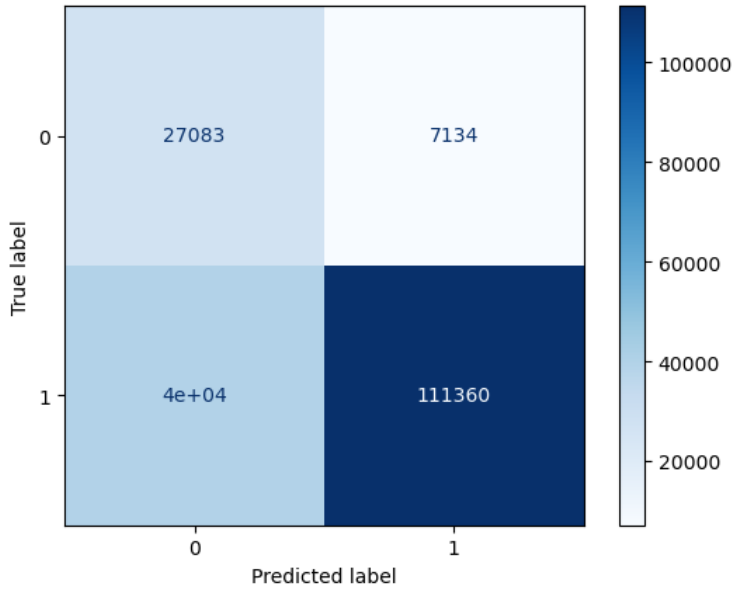

Figure 27: Normalized Logistic Regression Confussion Matrix

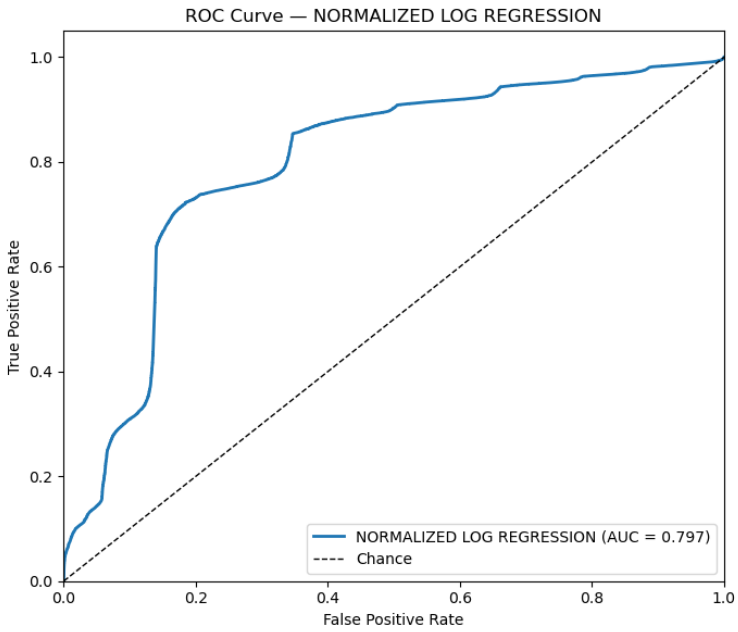

Figure 28: Normalized Logistic Regression ROC Curve

6.4 Final Random Forest Model Performance

Best estimators for random forest through random search cross verification gave us lower values in both Accuracy and AUC score.

```
Accuracy for best model: 0.9073269823479967
ROC AUC for best model: 0.9505898021511606

Classification Report for best model:
              precision    recall  f1-score   support

           0       0.70      0.88      0.78     34248
           1       0.97      0.91      0.94    151340

    accuracy                           0.91    185588
   macro avg       0.83      0.90      0.86    185588
weighted avg       0.92      0.91      0.91    185588
```

Figure 29: Random Forest with best estimators from the search yield

The confusion matrix shows lower false positives and higher false negatives. This trade off means there are less predicted Paid in full loans that are truthfully going to be Charged Off. The increase in false negatives means the predicted Charged off but actually paid off loans are more abundant. what this means for a financial institution is there are less loans which they will have to default but will also have to reject more loans which would have been paid off. The trade off mitigates risk but loses the bank clients who would have other wise paid off their loans. Considering there is no way of knowing which loans those would be, adopting a model like this would lower risk burden on financial institutions but would in turn lower the amount of loans given out.

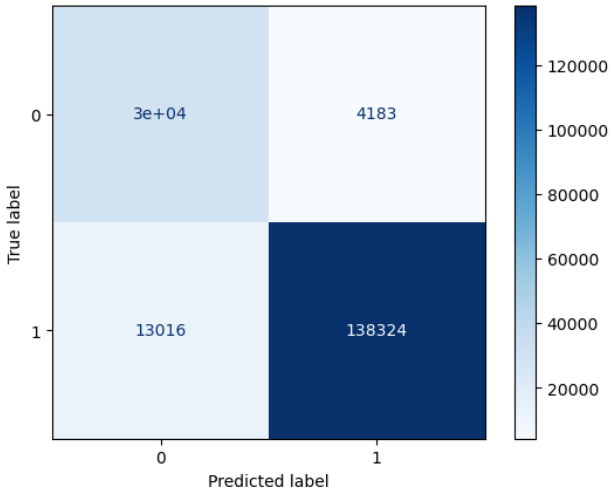

Figure 30: Best CV Random Forest Confusion matrix

The Forest with the best estimators also use the same four features with one exception. The features Gross Approval and Jobs Supported change ranking and Gross approval has more influence than Jobs

supported.

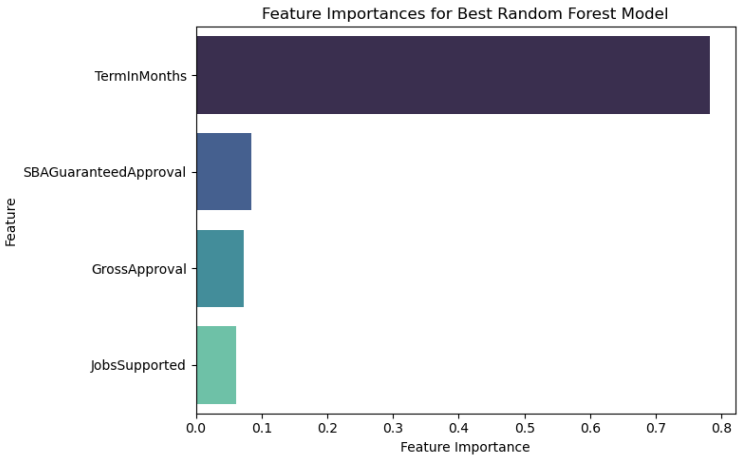

Figure 31: Best CV Random Forest Feature Importance

These changes are also visible in the ROC curve which shows a better inclination towards the left upper corner or in other words to a perfect model.

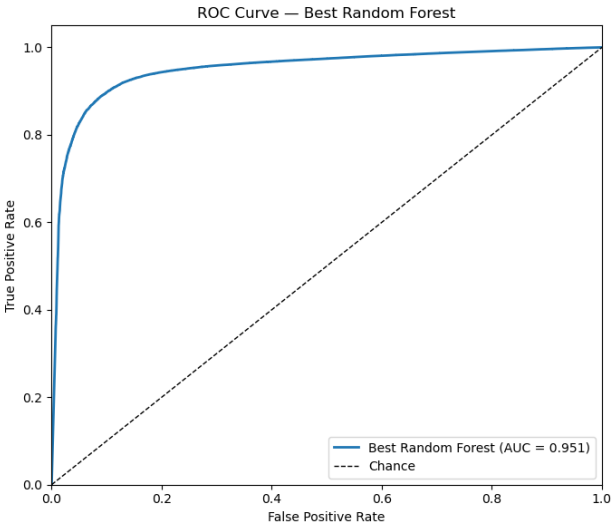

Figure 32: Best CV Random Forest ROC Curve

6.5 Normalized Random Forest Model Performance

Similar to the logistic regression model, with normalization, the random forest model did not exhibit much of an improvement from the original model with accuracy dropping slightly but ROC-AUC score increasing slightly.

```
Accuracy for best model: 0.9060259451165152
ROC AUC for best model: 0.9512597561841604

Classification Report for best model:
              precision    recall  f1-score   support

           0       0.69      0.88      0.78     34217
           1       0.97      0.91      0.94    150866

    accuracy                           0.91    185083
   macro avg       0.83      0.90      0.86    185083
weighted avg       0.92      0.91      0.91    185083
```

Figure 33: Normalized Random Forest Classification Report

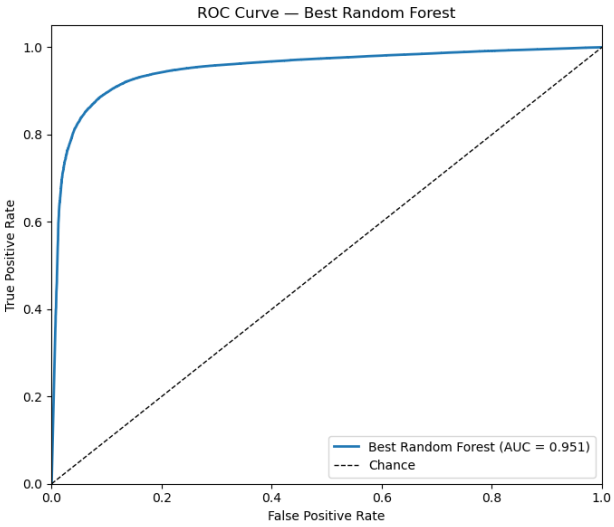

Figure 34: Normalized Random Forest ROC

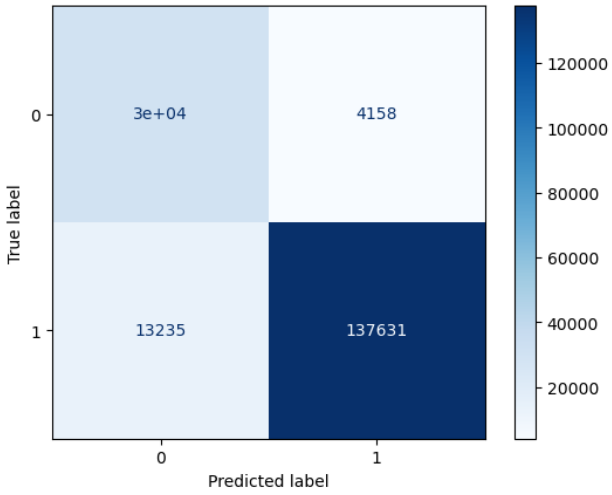

Figure 35: Normalized Random Forest ROC

6.6 Final Categorical Boost Model Performance

The Improved Catboost model stems from the base model but changes in features and various hyperparameters were introduced. Though the features 'BankName' and 'RevolverStatus' were re-introduced as text columns rather than categorical columns its most important features were still 'TermInMonths' and 'subpgmdesc'. In our Feature importance Graph, we can observe that that the most important features clearly dominate the model's decsion making.

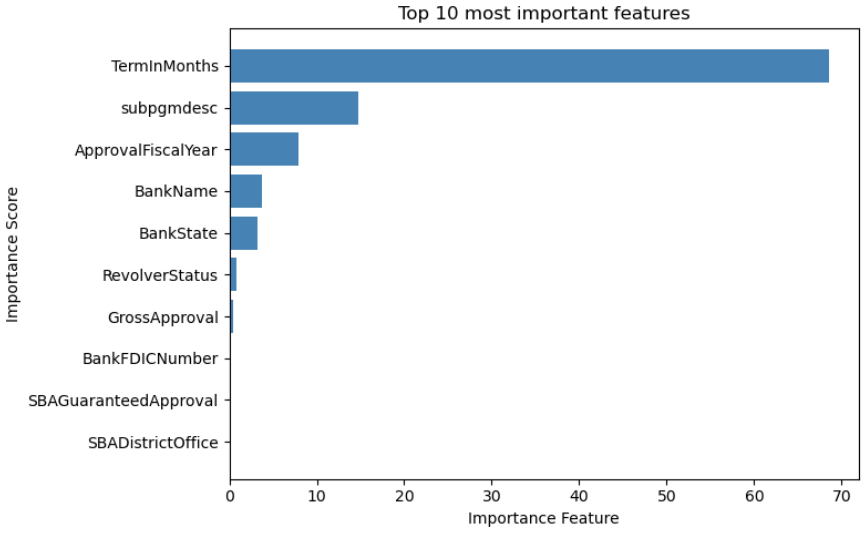

Figure 36: Features Used for Improved CatBoost Model

Hyper-parameters like 'eval_metric' were also changed from 'Accuracy' to 'PRAUC'. This was done

in order to better handle the class imbalance present on the data set. We also tuned the model further by adjusting factors such as learning rate and regularization parameters so as to reduce overfitting. This then led to a lower Accuracy score but a higher AUC score.

```
Accuracy:  0.9086363342457486
ROC AUC: 0.935763784929071
              precision    recall  f1-score   support

           0       0.80      0.68      0.73     34248
           1       0.93      0.96      0.94    151340

    accuracy                           0.91    185588
   macro avg       0.86      0.82      0.84    185588
weighted avg       0.90      0.91      0.91    185588
```

Figure 37: Evaluation Metrics Results for Improved CatBoost Model

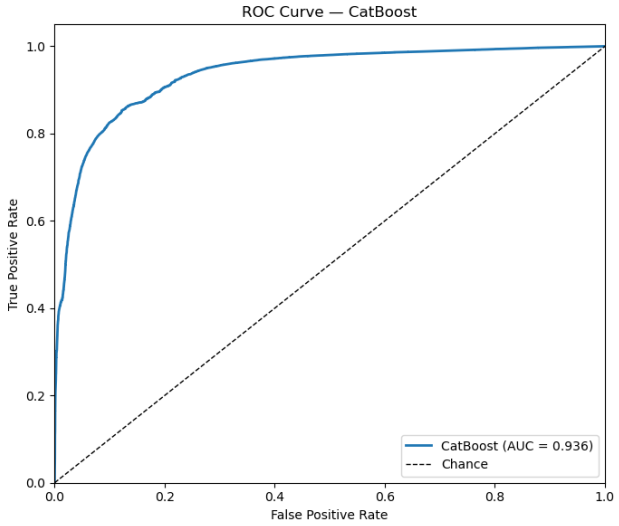

Figure 38: ROC AUC results from Improved CatBoost Model

The curve is better fit towards the left side of the graph, closer to a 1 meaning a better model.

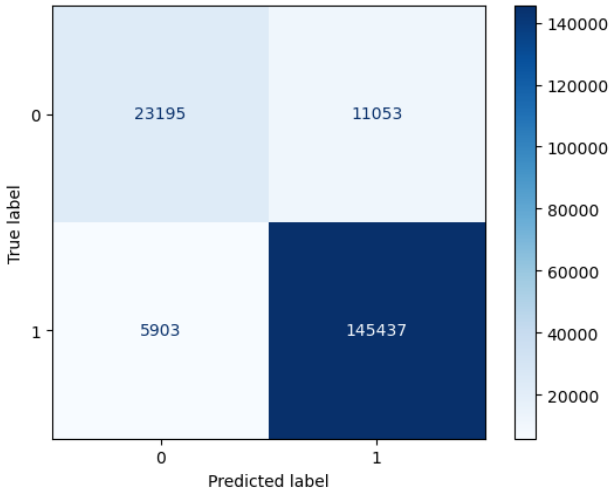

Figure 39: Confusion Matrix for Improved CatBoost Model

6.7 Normalized Categorical Boost Model Performance

The CatBoost model appears to be another model that did not improve with normalization. Precision, recall, and f1-score all went down for charged off loans, while Paid in Full stayed the same.

The confusion matrix shows that false predictions actually went up as a result of the normalization, and true predictions went down.

```
Accuracy:  0.9072794367932225
ROC AUC: 0.9357810536302135
              precision    recall  f1-score   support

           0       0.79      0.67      0.73     34217
           1       0.93      0.96      0.94    150866

    accuracy                           0.91    185083
   macro avg       0.86      0.82      0.84    185083
weighted avg       0.90      0.91      0.90    185083
```

Figure 40: Classification Report for Normalized CatBoost Model

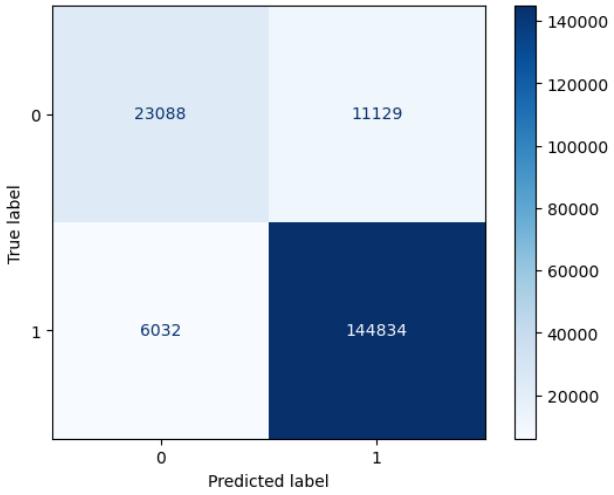

Figure 41: Confusion Matrix for Normalized CatBoost Model

The confusion matrix shows that false predictions actually went up as a result of the normalization, and true predictions went down.

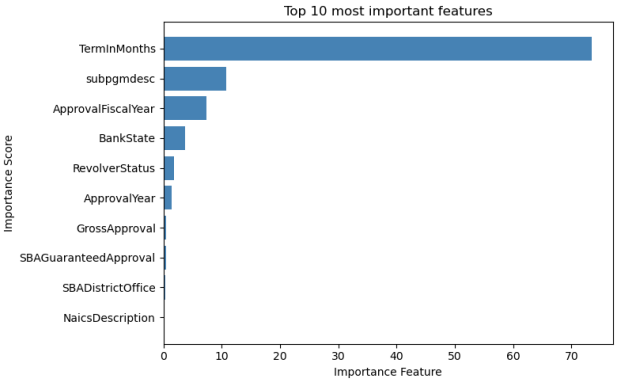

Figure 42: Feature importance for Normalized CatBoost Model

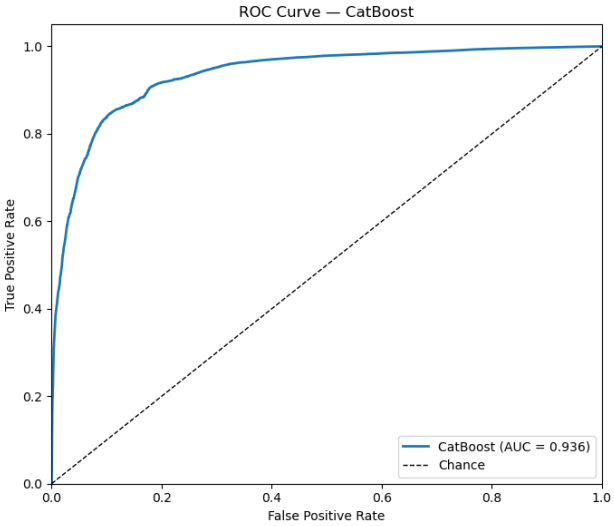

Figure 43: ROC for Normalized CatBoost Model

7 Challenges During Implementation

There were a multitude of challenges that were faced by our team while working on the project. To begin with, one of of the largest challenges we faced was related with the dataset itself. In particular, the initial raw dataset was extremely large in size (about 1.2M rows in total) and cluttered. It was a severe challenge to clean up the dataset while also ensuring that the dataset is still authentic and useful variables were kept. Another challenge related with the dataset was the fact the fact that it was highly imbalanced in nature which further complicated the work.

Another challenge was choosing the appropriate machine learning models depending on the nature of the dataset and project goals. For instance, logistic regression is easy to implement but less accurate for imbalanced dataset such as ours. On the other hand, models such as Random Forest boasted higher accuracy but required more tuning. At the end, our group decided to try out all the possible models which would fit in our context and compare them.

The size of the data set also caused other challenges within the machine learning process. With so many observations, this made running the machine learning algorithms time consuming. Specifically running the RandomSearchCV command, which allows us to view how different parameters affects the results of the random forest model. These commands had to be shortened however, to allow the computer to execute this command, which limited our potential combinations of hyperparameters to analyze. With the right equipment however, this concern could be easily mitigated with more powerful computers.

8 Discussion

The study conducted on the SBA 7(a) dataset gave a well versed view about the very potential of machine learning and statistical tools in predicting the success or failure of businesses. Each model that was implemented had its own caveats as well as strengths and they reflected the importance of model selection in terms of attaining good predictive performance. After several trials and tribulations, we were able to figure out the best possible modeling plan for our particular project.

Data cleaning occupied a huge chunk of the project, both in terms of length and complexity. The initial merged dataset was tremoundusly large in size, having 1.2 Million rows and 39 distinct features as mentioned earlier. While data cleaning was being performed, the team ensured that features that were irrelevant to our analysis were removed while still making sure that the dataset isn't stripped out of its usability. The largest issue of the dataset, by far was the fact that it was highly imbalanced in nature. It was essential to deal with class imbalances in order to ensure the fairness and readability of our final models.

The team discovered a clear pattern in the success of tree based algorithms, particularly Random Forest in terms of handling the complexity as well as the imbalanced nature of our dataset. This aligns perfectly with Couronne's claim about tree based ensemble methods faring better than logistic regression while dealing with high dimensional and complex datasets [5]. For this reason, random forest outperformed the logistic regression model in every possible metric. Categorical Boosting also performed quite well, especially after further tuning was performed at a later stage. However, it still underperformed when compared to the Random Forest model since it wasn't as effective in dealing with the complexity of the dataset.

Despite the decent performance, one major issue that posed as a large challenge for our team was handling the class imbalance in the SBA 7(a) dataset. Loans that were charged off made up a very small fragment of the overall data which had a negative influence on recall for our minority class. Wang et al discussed similar limitations in their paper, highlighting how imbalanced credit models are often biased towards majority classes unless balancing methods are applied [17].

Going beyond just performance metrics, the team made significant leap towards discovering the most notable factors influencing the success or failure of small businesses. Our models revealed features that were directly related to term length, loan size, and guarantee amount among others were strongest predictors of success. These findings tied in perfectly with Cheraghali's paper that discussed the factors influencing SME credit risk modeling such as size of loan and repayment structure [3].

The final results obtained prove that machine learning is an excellent tool in predicting loan default of small businesses in the SBA 7(a) dataset. In order to truly make the most out of machine learning, it is very crucial to choose the right model, employ the right split and apply the right balancing techniques when required. Ultimately, it is important to be cautious and responsible when conducting projects of this size and scale.

9 Conclusion

This study set out to determine whether the characteristics of SBA 7(a) loans can be used to predict whether or not the loans will be paid in full or charged off. This was done after the loans were ap-

proved, which is a key aspect of our study as many other research has gone into loans before they were approved or denied. Our analysis of two decades worth of loans from the SBA 7(a) program shows that machine learning techniques, especially those from tree-based ensemble methods, contribute meaningful predictive power in post-approval loan outcomes.

Across the three models that were evaluated, the Random Forest model demonstrated the strongest performance, yielding an accuracy of .934 and AUC score of .969. As stated before, this aligns with the work done by Couronne that tree-based ensemble methods outperform logistic regression in complex financial data[5]. Our CatBoost model improved with proper tuning, but ultimately stayed under the Random Forest in strength, while the logistic regression model consistently stayed below the other two models, reflecting its limitations in capturing nonlinear risk patterns as outlined in previous studies [8, 3]. This answers a subquestion of our research, which set out to understand which machine learning method would score the highest in our data. The Random Forest model outperformed the others, leading us to conclude the it is the most effective modeling technique for this data.

Another question that our team had set out to answer was which characteristics of the loans would impact our models' classification the most. Our models revealed that term length, loan size, guaranteed amount, and jobs supported were strong predictors of loan success. An time series analysis also revealed that charge off rated showed meaningful shifts in periods of economic stress, such as a recession, which aligns with macro lending trends in literature. This is shown by self employed salaries dropping 19% from 2007-2010 [18].

One of the most significant challenges that our team faced was the imbalance of Paid in Full and Charged off Loans. Imbalanced datasets tend to bias models toward the majority class and make minority-class recall more difficult to optimize[17]. This certainly showed up in our analysis of the models. The recall for charged off loans (the minority) was significantly less than those that were Paid in Full. Future work could include a more balanced dataset, or balancing techniques on the current dataset to increase prediction power of both targets.

Overall, our finding confirm that machine learning does in fact provide significant prediction power in the outcomes of small business loans. These models can be used and optimized to help lenders assess risk and monitoring strategies, as well as informing small business owners what they can do to create a more successful business and attain highly sought loans. With improved balancing strategies, richer macroeconomic features, and deeper model architectures, future research can further enhance predictive accuracy and support more responsible and effective small-business lending practices.

References

[1] Cristián Bravo, Sebastián Maldonado, and Richard Weber. Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research*, 227(2):358–366, 2013.

[2] Ian Brown and Clemens Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

[3] Hamid Cheraghali. Sme default prediction: A systematic methodology-focused review. *Journal of Small Business Management*, 2023.

[4] Shining a light on small business lending, 2024. Available at: https://www.consumerfinance.gov/about-us/small-business-lending/ (Accessed: 2025-08-29).

[5] Raphael Couronné. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 2018.

[6] Marc Cowling and Ondřej Dvouletỳ. Uk government-backed start-up loans: Tackling disadvantage and credit rationing of new entrepreneurs. *International Small Business Journal*, 41(7):714–733, 2023.

[7] Robert W. Fairlie and Frank Fossen. Opportunity versus necessity entrepreneurship: Two components of business creation. IZA Institute of Labor Economics Discussion Paper No. 11258, 2018. Accessed: 2025-09-11.

[8] Vasilios Giannopoulos. Predicting sme loan delinquencies during recession using accounting data and sme characteristics: The case of greece. *Intelligent systems in accounting,Finance and mangement*, 2019.

[9] Yazhe Li, Tony Bellotti, and Niall Adams. Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, 1(4):389–417, 2019.

[10] Nakata Hiroyuki Norvald Instefjord. Loan monitoring and bank risk. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2426149, 2014. Accessed: 2025-08-29.

[11] Small business loans decreased in 2023 as lending standards tightened, 2024. Available at: https://www.pymnts.com/smbs/2024/small-business-loans-decreased-in-2023-as-lending-standards-tightened/ (Accessed: 2025-08-29).

[12] Ray Tsaih and Yu-Ling Lien Yu-Jane Liu, Wenching Liu. Credit scoring system for small business loans. *Decision Support Systems*, 2004.

[13] U.S. Small Business Administration. About sba: Works to ignite change and spark action so small businesses can confidently start, grow, expand, or recover. Accessed: 2025-09-14.

[14] About page, 2025. Available at: https://data.sba.gov/about (Accessed: 2025-08-29).

[15] U.S. Small Business Administration, Office of Advocacy. Frequently asked questions about small business, July 2024. Accessed: 2025-09-11.

[16] Hong Wang and Li Cheng. Catboost model with synthetic features in application to loan risk assessment of small businesses, 2021. Preprint.

[17] Hong Wang, Qingsong Xu, and Lifeng Zhou. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLOS ONE*, 10(2):e0117844, 2015.

[18] Ann Marie Wiersch. Small business lending isn't what it used to be, 2013. Available at: https://www.clevelandfed.org/publications/economic-commentary/2013/ec-201310-why-small-business-lending-isnt-what-it-used-to-be (Accessed: 2025-08-29).

[19] Aikaterini Cheimarioti Yiannis Dendramis, Elias Tzavalis. Measuring the default risk of small business loans: Improved credit risk prediction using deep learning. *The Journal of Forecasting*, 2025.

A Appendix

The entire code snippet for the SBA project can be referred to here: https://github.com/Ribhay27/SBA/tree/main