Loans for Small Businesses

Team Members:

- Ryan, Jacobowitz; Email: rjacobow@seas.upenn.edu
- Saibernard, Yogendran; Email: bernie97@seas.upenn.edu
- Tanvi, Gupta; Email: guptanvi@seas.upenn.edu

home pod: daft-dinosaurs

Abstract

Small businesses are a vital part of the American economy, and the US Small Business Administration (SBA) plays a crucial role in supporting these businesses through its loan programs. For lenders and financial institutions, being able to predict the likelihood of a loan being repaid can help them manage their risk and make more informed lending decisions. This project aims to analyze which features affect the performance of loans and predict whether a loan is likely to be repaid or defaulted. A combination of financial and non-financial factors are used to predict the success of SBA-guaranteed loans and various models such as logistic regression, random forest and Tab Net are implemented. The performance of each model is evaluated using a data set of past loan applications. On comparing the accuracy of the models, it is observed that the best performing algorithm for the data set is random forest with an F1 score of 0.95.

Keywords: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, Tab Net

1 Motivation

Small business loans are a type of financing that is provided to small businesses to help them grow and expand. These loans are typically offered by banks and other financial institutions, and can be used for a variety of purposes, including purchasing equipment, hiring employees, and expanding into new markets. The ability to predict the likelihood of a small business obtaining a loan is an important skill for banks and other financial institutions. By using predictive modeling techniques, these institutions can identify which businesses are most likely to repay their loans on time, and can make more informed decisions about which businesses to lend to.

The SBA[1] was founded in 1953 to help lend money to small businesses. They help small businesses get funding by setting guidelines for loans and reducing lender risk. These SBA backed loans make it easier for small businesses to get the funding they need. Small business can help decrease unemployment and foster the creation of new jobs. The SBA has historically helped many successful businesses get loans such as Apple or FedEx. However, there have also been stories of small businesses and/or start-ups that have defaulted on their SBA-guaranteed loans. Financial institutions have been turning to AI and machine learning not only for credit risk management, but also to tackle other issues like financial fraud and evaluation of client behavior that could potentially lead to a financial loss. If the risk associated with a given loan can be predicted well ahead of time, financial institutions can attempt to prevent loan defaults and grant loans to businesses that could potentially thrive. Machine learning models that predict the success of a loan can be used by financial institutions to improve their loan underwriting process. Understanding and evaluating the risk involved in a particular loan will help them decide on optimal loan offers.

2 Related Work

Prior research related to the use of machine learning on finance, banking, and lending was expanded upon to produce accurate results. In their paper, Hamid and Ahmed [5] implemented three algorithms: j48, bayesNet, and Naive Bayes. These algorithms were used to build models to classify whether loans were likely or unlikely to default by investigating customer behavior and previous pay back credit. This classification method was then used to predict whether a given loan should be lent. They implemented the model using the Weka application and found that best algorithm for loan classification

was j48.

Shubham Gadekar[4], in his Kaggle project, predicted the performance of mortgage loans. He evaluated the effects of each feature, like age, loan term and credit scores on the success of a loan. He goes on to observe that longer term loans were more popular as compared to shorter term loans and that employed applicants have a better credit score than self-employed applicants. Gadekar uses a random forest classifier to predict the success of loans with a 76% accuracy.

Ereiz[3] used OptiML to identify the best suited prediction model. Based on the input data, BigML returned 33 candidate models out of which 3 were short-listed: a 64-model 465-node decision forest, a neural network based on 128 evaluated networks and a logistic regression model. Ereiz then compares the accuracy, precision and recall of the three algorithms and concludes that for most classification problems precision is usually less important than the recall because predicting a loan that will be good as bad (false positive) is not as costly as predicting a loan that will be bad as good (false negative).

3 Data Set

The data set being used for this project shows the success of SBA loan guarantees given to small businesses. It was released publicly by the U.S. Small Business Administration (SBA) and is available on the website, Kaggle [6].

3.1 Feature Descriptions

The data set is a list of 899164 loan applications with 27 different features and a label signifying whether the loan was charged off or paid in full. The name, city, state, and zip code represent information about the borrower's location while the bank name and state refers to the bank's location. The NAICS code refers to a standardized way to look at the type of industry the business is located in. The approval data and year have a wide range, from 1961 to 2014. The term refers to the length of the loan in months. The total jobs created signifies the total number of jobs created by the business. The number of jobs retained refers to the number of jobs created subtracted by the number lost by the creation of the business. The franchise code refers to whether there is a franchise of the business or if it is not a franchise. Codes 00000 and 00001 designate a non-franchise business while any other code designates a franchise. The Urban/Rural feature specifies what type of location the business is located at, and whether it is in an urban or rural location. Revolving line of credit refers to a loan that can be borrowed again upon being paid off without needing to apply for a new loan. The low document feature details whether the loan application used very few documents during the loan approval process. The gross disbursement is the total amount that is disbursed to the companies. The gross balance is the amount of the loan that still needs to be paid back. The charge off amount refers to how much of the loan is cancelled during the lifetime of the loan. The gross approval amount is the total amount that the bank is willing to lend. The SBA approval amount is the pre-approved amount the SBA is willing to guarantee borrowers will receive. The features and their description are represented in Table 2.1.

3.2 Data Preprocessing

The noise in the data set had to be cleaned out before any models could be trained on it. Features like unique ID and borrower's organization were dropped because they were not particularly relevant to model creation. DataTime features like ChgOffDate, ChgOffPrinGr were dropped because they could directly tell us that the loan was charged-off. To make the model time-independent, we dropped the features 'ApprovalDate', 'ApprovalFY', 'DisbursementDate'. Some features that should have had binary values also had noisy data. Such data points had to be dropped and string values of Y and N were converted to binary values of 1 and 0. The data set also has many columns with Object dtype and hashing had to be applied to convert them to numerical data types. The target, 'Defaulted', consists of two classes , 0 and 1. 0 means that a business does not default on their loan, 1 means that they do.

One of the key challenges in predictive analysis for classification is dealing with imbalanced data, where there are significantly more data points in a class as compared to other classes. This can make it difficult for predictive models to accurately identify the data points belonging to the minority class. The data set used was highly imbalanced and after pre-processing, it had 112,080 defaulted loans and 511,303 fully paid loans, as shown in Figure 5.1. A class imbalance creates a bias where the machine learning model tends to predict the majority class. To account for this bias, we used different sampling techniques on the data set.

Synthetic Minority Oversampling Technique (SMOTE) was used to oversample the data. SMOTE works by utilizing a k-nearest neighbour algorithm to create synthetic data until the minority class has the same proportion as the majority class, as shown in After oversampling the data, there were 818,084 data

Feature	Description
LoanNr_ChkDgt	Identifier Primary key
Name	Borrower name
City	Borrower city
State	Borrower state
Zip	Borrower zip code
Bank	Bank name
BankState	Bank state
NAICS	North American industry classification system code
ApprovalDate	Date SBA commitment issued
ApprovalFY	Fiscal year of commitment
Term	Loan term in months
NoEmp	Number of business employees
NewExist	1 = Existing business, 2 = New business
CreateJob	Total number of jobs created by the business
RetainedJob	Net number of jobs retained
FranchiseCode	Franchise code, (00000 or 00001) = No franchise
UrbanRural	1 = urban, 2 = rural, 0 = undefined
RevLineCr	Revolving line of credit: Y = Yes, N = No
LowDoc	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	The date when a loan is declared to be in default
DisbursementDate	Disbursement date
DisbursementGross	Amount disbursed
BalanceGross	Gross amount outstanding
MIS_Status	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Charged-off amount
GrAppv	Gross amount of loan approved by bank
SBA_Appv	SBA's guaranteed amount of approved loan

Table 2.1: List of Features

points, with equal distribution for the two classes, 0 and 1.

RandomUnderSampler, with a sampling strategy (hyper-parameter) of 0.5 , was used to reduce the size of the majority class to twice the size of our minority class. After undersampling the data, there were 280,200 data points. Since our minority class had enough data points, we were still left with a good amount of data to work with. The class distribution is shown in Figure 4.

4 Problem Formulation

4.1 Project Approach

Loan prediction can be viewed as a machine learning problem because it involves the analysis of historical data to predict the likelihood of a loan being repaid. Various algorithms can be implemented which use given data to learn patterns and relationships that can be used to make predictions about the likelihood of a loan being approved. The objective of this project is to predict the success or failure of a prospective applicant to pay off a loan taken to support their small business. The target variable is the loan status which signifies whether a loan is a good or a bad loan for the bank. The data set used consists of information about past loan applications such as the amount of the loan approved, the loan term, the type of loan and the type of business being funded.

To predict the success of a loan using machine learning, first the data was pre-processed to extract the relevant information and prepare it for modeling. Next, a logistic regression baseline model was trained on this data and then more sophisticated models were trained, such as random forest, XGBoost, Naive Bayes, gradient boosting, and Tab Net. By modelling the data, the most important features for predicting the target were determined. The trained model was then used on test data to evaluate the accuracy with which it could predict the success of loans. The performance of each model was evaluated and compared to find the best algorithm for our dataset.

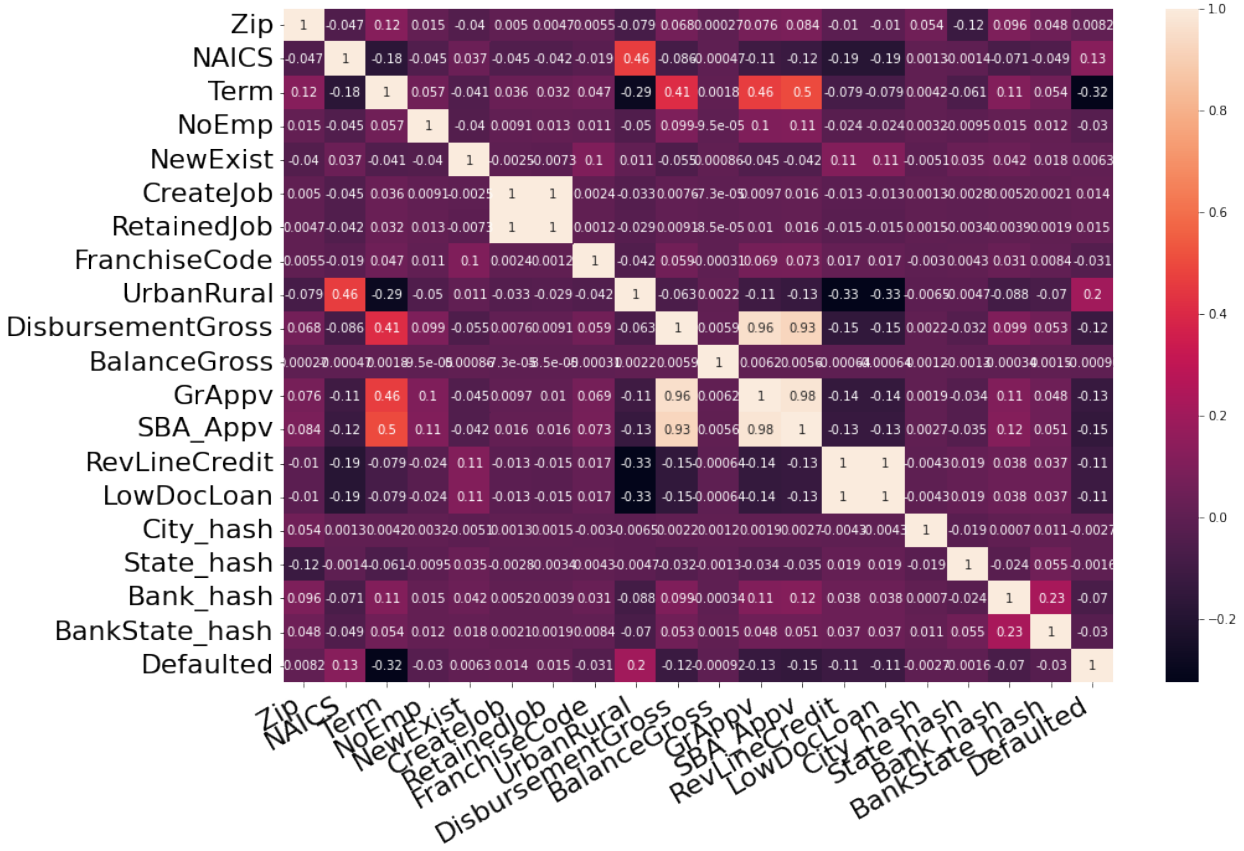

Figure 2.1: Feature Correlation Matrix

4.2 Correlation of Features

To get a better idea on how different features relate to one another, a correlation matrix was created, as can be seen in Figure 2.1. The number of created jobs and retained jobs are directly correlated which is likely because the total amount of jobs created by a business and the net amount of jobs retained because of that business are very closely related values. The other highly correlated values are the gross approval, SBA approval, and gross disbursement since the bank approved amount (grAppv), pre-approved loan amount (SBA Appv), and the total value of the loan are all directly correlated. A key observation to note from the matrix is that the loan term and the loan amount approved were correlated. This is intuitive as a small business would need a longer period of time to repay a loan with a high principal amount.

5 Methods

5.1 Feature Importance

Some of the key factors that are used in small business loan prediction include the type of loan, amount disbursed, the loan term, the size and type of business being funded, etc. By analyzing these factors, predictive models can estimate the likelihood that a small business will be able to repay a loan.

To evaluate the relative importance of features towards predicting the target, i.e. 'Defaulted', we plotted a bar graph, as shown in Figure 5.2. The plot is an indicator of the impact each feature has on the model. We can observe that 'loan term' has the most effect on the success of a loan, followed by whether the small business is a franchise or an independent venture.

We then used SHAP values to measure the contribution of each feature towards the target. SHAP values, or SHapley Additive exPlanations, is a method based on cooperative game theory and used to increase transparency and interpretability of machine learning models. The summary of the contribution of each feature is shown in Figure 5.1. As shown by figure 5.1 and 5.2, loan term has the most effect on the target, followed by franchise code. The SHAP scores for each feature are outlined in table 5.1.

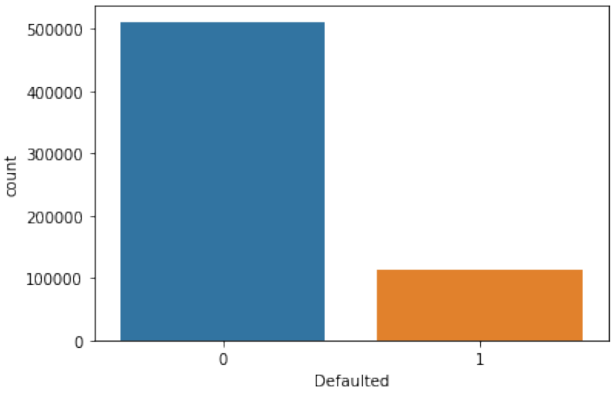

Figure 3.1: Distribution with Class Imbalance

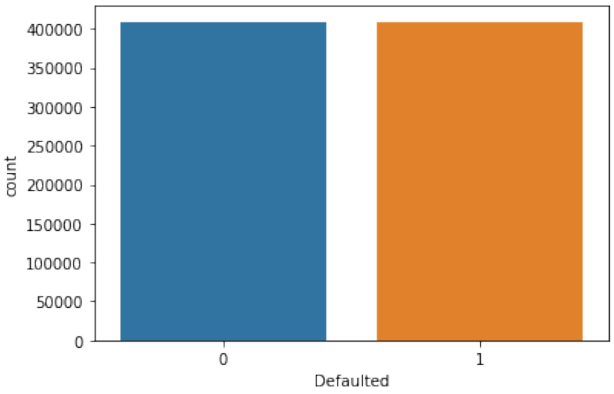

Figure 3.2: Distribution of Oversampled Data

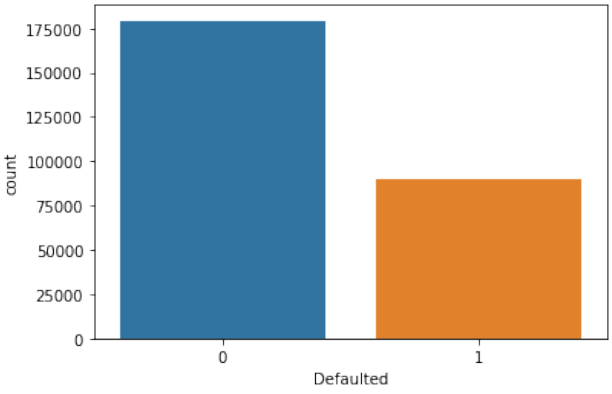

Figure 3.3: Distribution of Undersampled Data

When studying the correlation between the target and the features, it was found that higher term loans tend to perform worse than shorter term loans. As the term increases, the chances of the business defaulting on a loan also increase. Also, it was observed that loans given to small businesses that were franchises tend to perform better than loans given to small business that were independent.

Feature	**SHAP Score**
ZIP	0.0189
NAICS	0.0232
Term	0.2201
Num of Employees	0.0126
New/Exist	0.0091
Jobs Created	0.0073
Jobs Retained	0.0154
Franchise Code	0.0635
Urban/Rural	0.0450
Gross Disburse	0.0210
Gross Balance	0.0000
Gross Approved	0.0192
SBA Approved	0.0282
Revolving Credit	0.0042
Low Document	0.0036
City Hash	0.0125
State Hash	0.0115
Bank Name Hash	0.0283
Bank State Hash	0.0229

Table 5.1: SHAP Scores for Each Feature

5.2 Models

To solve for the most predictive feature, classification methods were used and the f1 and area under the curve (AUC) scores were calculated. Logistic regression was chosen as the baseline, as it is one of the most simple classification models and gave an initial result to compare more sophisticated models to.

To get a more accurate model, a random forest was implemented as it is commonly used for classification problems. Random forests are well-suited to this task because they can handle large datasets with a multitude of features, and they are able to make accurate predictions even when the data is highly unbalanced or when there are missing values. Additionally, random forests are relatively easy to interpret, which can be important in a domain like loan prediction where it is important to understand why a particular loan was approved or denied. The random forest classifier was implemented with a balanced class weight, random state = 42 and the

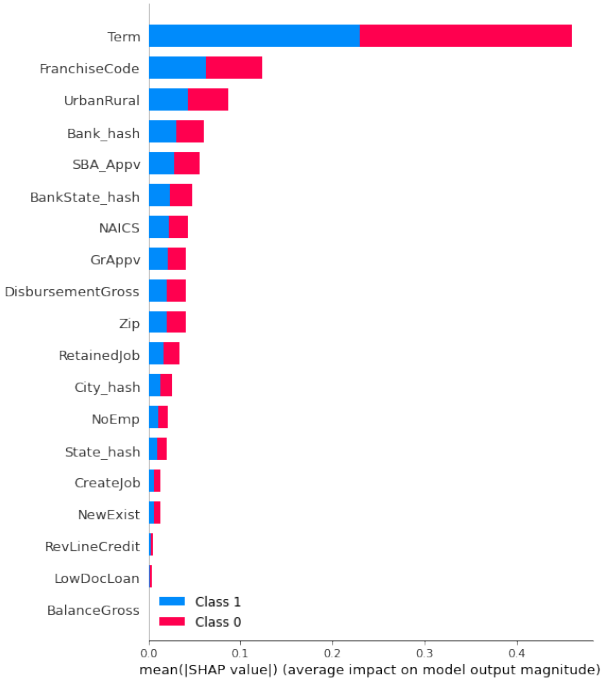

Figure 5.1: Summary of SHAP Scores

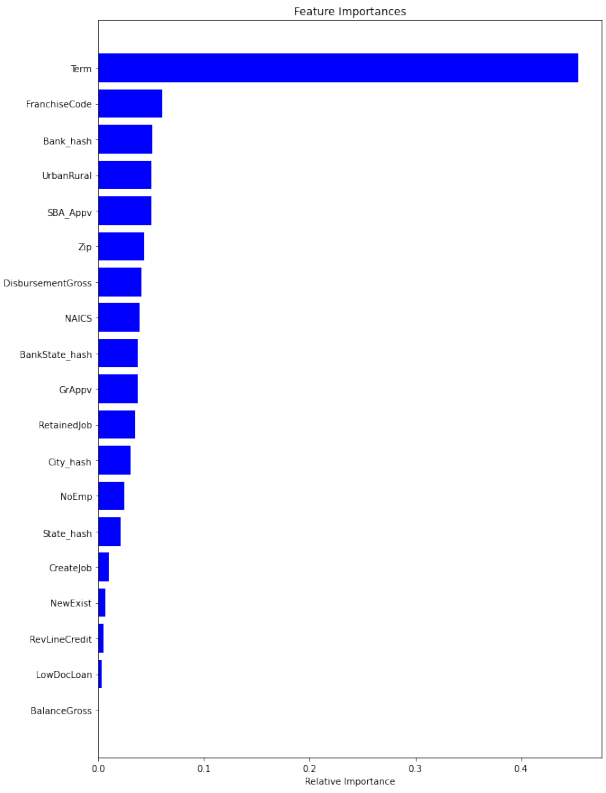

Figure 5.2: Feature Importance

maximum depth was specified as None.

We then implemented a gradient boosting classifier with the number of estimators as 100, learning rate of 1 and max depth of 1. Gradient boosting is an ensemble method that combines the predictions of multiple weaker models to produce a single, more accurate prediction. This can be especially useful for loan prediction, where the relationship between the features and the target variable (i.e. whether a loan will be repaid or defaulted) may be non-linear or otherwise difficult to model. An XGBoost (eXtreme Gradient Boosting) classifier was implemented on the data set as well.

A Gaussian Naive Bayes was implemented next because it is a simple and efficient algorithm. The hyperparameters chosen are priors = None and var_smoothing = 1e-09. It is able to make predictions based on the relative frequency of different classes in the training data, which can be useful when dealing with unbalanced datasets.

Logistic regression, random forest, gradient boosting, and Gaussian Naive Bayes were imported from the library "sklearn" and XGBoost was imported from its own library.

A Tab Net model was used as a final method. Neural networks are commonly used for image recognition and audio but are much more rarely used for tabular data. Tab Net was developed to fill in this underdeveloped area of neural networks[2]. Using Tab Net required installing the Tab Net library from pytorch and training a Tab Net classifier using the data set. It is a deep learning algorithm that was developed by Google AI, designed to be a more efficient and interpretable alternative to other deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). TabNet is a type of attention-based neural network, which means that it uses a mechanism called "attention" to focus on the most important parts of the input data while making predictions. This can help the algorithm to make more accurate predictions while using fewer computational resources.

6 Experiments and Results

Each model was compared to every other model using a 5-fold cross validation method for AUC and F1 scores. The F1 score is a metric used to evaluate the performance of a classification model, such as a model that predicts the likelihood of a loan being defaulted. It is calculated as the harmonic mean of the precision and recall of the model. Precision is the number of true positives divided by the total number of predicted positives, and recall is the number of

true positives divided by the total number of actual positives. The F1 score is a useful metric because it takes into account both the precision and recall of the model, and therefore provides a more complete picture of its performance. A high F1 score indicates that the model has both high precision and high recall, while a low F1 score indicates that the model has either low precision or low recall (or both). The AUC (Area Under the Curve) is a metric used to evaluate the performance of a binary classification model. AUC evaluates the same model's performance across different thresholds. It is the area under the receiver operating characteristic (ROC) curve, which is a plot of the true positive rate (TPR) against the false positive rate (FPR) of the model. The AUC ranges from 0 to 1, with a higher AUC indicating a better performing model. For example, an AUC of 0.8 would indicate that the model has a high true positive rate and a low false positive rate, while an AUC of 0.5 would indicate that the model has a random performance (i.e., it is no better than flipping a coin). The AUC is a useful metric because it provides a single value that summarizes the performance of a binary classification model across all possible thresholds, and is therefore insensitive to changes in the classification threshold.

To determine which preprocessing method produced the most accurate models, the models were run for the unbalanced, undersampled, and oversampled cases. As can be seen in tables 6.1 and 6.2, models trained on oversampled data provided the most accurate results. In some cases, models trained on undersampled data provided less accurate results than those trained on imbalanced case. This is likely because undersampling removes too many data points from the majority class and the accuracy might reduce. Models tend to run better on oversampled data because oversampling can help to balance the class distribution in the dataset, which can improve the model's performance. In an imbalanced dataset, the majority class dominates, and the model may have difficulty learning the patterns in the minority class because there are fewer examples to learn from. Oversampling can help to mitigate this problem by generating additional synthetic examples of the minority class, which can help the model to learn the patterns in the minority class more effectively. By oversampling the minority class, we can improve the performance of these algorithms by giving them more examples of the minority class to learn from. Additionally, oversampling can help to reduce the effects of bias and variance, which can further improve the performance of the model. On the other hand, undersampling can cause the model to lose important information from the majority class, which can negatively impact its performance.

The five scores calculated for each method using oversampling were then averaged and the results are in table 6.1 and 6.2. Logistic regression was the first viable model, but resulted in the second lowest f1 score of 0.693 due to low complexity of the model. Gaussian Naive Bayes performed the worst of all the methods, especially with respect to its f1 scores. Tab Net was the next best method, with an f1 score of 0.809. Gradient boosting received an f1 score of 0.918 and XGBoost was the second best method with a score of 0.931. The best method for classification was the random forest method, with an f1 score of 0.965 and an AUC score of 0.993. Random forest is a very effective method for classification, and so it was expected that it would perform well. One unexpected result was Tab Net being one of the lower scoring models. This might be because Tab Net doesn't perform as effectively for preprocessed data, as the model is built for raw data, which would explain why it performed worse for the preprocessed cases versus the unbalanced case.

7 Conclusion and Discussion

The US Small Business Administration (SBA) helps support start-ups and small businesses through loan guarantees. In order to find which features best predict whether a business will default on their debt, multiple machine learning models were created. Logistic regression was used as a baseline and models using random forest, Gaussian Naive Bayes, gradient boosting, XGBoost, and Tab Net were done to determine the most accurate method for predicting debt default. It was found that the random forest model was the most accurate using 5-fold cross-validation, with the term limit of the loan, type of location of the business and franchise code being the most predictive factors.

Of all the models implemented, Random forest performed the best because it is able to handle high-dimensional data and complex relationships between the variables. This is because each decision tree in the random forest is trained on a different subset of the data, which allows the model to capture the interactions between different variables and make more accurate predictions. Additionally, random forest is able to handle missing values and outliers in the data, which can be common in datasets for loans, resulting in a model that's more robust and not influenced by noisy or anomalous data points. Furthermore, random forest is a type of ensemble method, which means that it combines the

Method	Unbalanced	Undersampled	Oversampled
Logistic	0.740	0.740	0.693
RF	0.946	0.946	0.965
Gaussian NB	0.424	0.457	0.554
Grad Boost	0.920	0.896	0.918
XGBoost	0.929	0.904	0.931
Tab Net	0.920	0.889	0.809

Table 6.1: Average F1 Scores

Method	Unbalanced	Undersampled	Oversampled
Logistic	0.730	0.730	0.768
RF	0.975	0.975	0.993
Gaussian NB	0.770	0.767	0.783
Grad Boost	0.946	0.948	0.968
XGBoost	0.962	0.963	0.979
Tab Net	0.945	0.947	0.883

Table 6.2: Average AUC Scores

predictions of multiple individual decision trees to reduce the variance of the predictions and improve the overall accuracy of the model. This can be especially useful in loan prediction problems, where the goal is to accurately identify which businesses are most likely to repay their loans on time. Overall, the combination of high-dimensional data handling, robustness to missing values and outliers, and ensemble learning make random forest a strong candidate for loan prediction problems, and can help the algorithm to achieve high accuracy and robust performance.

When studying the effect of features on the success of loans, we found that higher term loans tend to perform better than shorter term loans. For small businesses, a longer loan term can make it easier to manage their cash flow and to afford the monthly loan payments. This can be especially beneficial for businesses that are experiencing slow or unpredictable growth, or those that are facing temporary financial challenges. Overall, it was observed that loan term has a significant impact on the repayment of small business loans. Additionally, it was found that small businesses that were franchises tend to be more reliable in terms of loan repayment as they are associated with a bigger and already established business. The chances of such a business to repay the loan would be higher than that of a small independent business, like a start-up.

There are several possible directions for future work on this project. For example, the model could be further refined by incorporating additional data and features, such as the business owner's credit history, the rate of interest, and the financial performance of the business. In addition, the model could be tested on a more diverse dataset to better evaluate its performance and robustness. This could help ensure that the model is effective for a wider range of small businesses and lending situations. The model could also be used to develop a recommendation system for small businesses on how to improve their chances of receiving a loan guaranteed by the SBA.

Acknowledgments

We would like to thank Professor Lyle Ungar and our mentor, Keshav Ramji, for their valuable advice and support throughout the research process. We are grateful to them and the TAs for their helpful comments and suggestions, which have greatly improved the quality of the project. We would also like to acknowledge the support of our families and friends, who have encouraged and inspired us throughout this project.

References

[1] About US Small Business Administration. https://www.sba.gov/about-sba/organization.

[2] Pfister Arik. TabNet: Attentive Interpretable Tabular Learning, 2020. https://arxiv.org/abs/1908.07442.

[3] Zoran Ereiz. Predicting default loans using machine learning (optiml), 2019. Predicting Default Loans Using Machine Learning (OptiML),1-4.10.1109/TELFOR48224.2019.8971110.

[4] Shubham Gadekar. https://www.kaggle.com/code/shubhamgadekar/loan-prediction-with-random-forest.

[5] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed. Developing prediction model of loan risk in banks using data mining, 2016. Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016.

[6] Kaggle US SBA. Dataset. https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied.