Predicting Loan Defaults

By: Jose Echevarria

Presented to:

Business Analytics Staff
Dolan School of Business
Fairfield University

For the Degree of Master of Science in Business Analytics

Under the supervision of:
Doctor Illya Mowerman

Fairfield, CT

Jul 16, 2023

Table of Contents

Research Question and Data

How can we effectively assess and mitigate the risk of loan default in our lending portfolio by leveraging borrower characteristics and loan attributes?

The dataset I will be using to answer the research question was acquired from Kaggle's Loan Default Prediction Competition.

Introduction

This study aims to investigate whether individuals who appear similar are more likely to default on a loan compared to others. By analyzing various factors and characteristics, the goal is to identify predictors that can differentiate between borrowers who are more likely to default and those who are not. By leveraging these predictors we want to be able to detect borrower applications that will default.

Lending is always accompanied by risk and it is important to mitigate this risk. Machine Learning is still a relatively new field that can prove useful in detecting patterns in lending data that humans can not. By understanding the factors that contribute to loan defaults, banks and financial institutions can develop more accurate and effective models for assessing creditworthiness. This, in turn, helps lenders make informed decisions, minimize default rates, and manage their lending portfolios more effectively.

Previous studies have explored the prediction of loan defaults using various modeling techniques. Commonly used models include logistic regression, decision trees, random forests, support vector machines, and neural networks. With such a wide spread of options it is safe to say that the most optimal model for a given project will depend on the data available and the preprocessing approach used. For example, a paper conducted at the university of Beijing in 2022 by Herui Chen found that the best models for that study were a logistic regression and a Neural Network. Another study by Wanjun Wu from ByteDance in the same year concluded that

the best models for their Loan Default prediction were Random Forest and XGBoost. Typically, these models leverage different combinations of features, such as credit history, income, debt-to-income ratio, employment status, and other relevant variables, to predict loan defaults. These studies have provided insights into the factors associated with default risk and have laid the foundation for further research in this area. Like previous studies I will have to dive deep into the features available in my dataset to perform analysis, to feature engineer and select the best performing model, therefore identifying the most important features for predicting loan defaults.

In order to paint a clear picture of my study, this paper will follow a well outlined process to ensure a professional Loan Default prediction project is delivered. This introduction serves as the first deliverable in the overall process. Here I have described the topic, its importance and discussed previous studies. The next phase is the literature review which is the foundational piece for my research. More will be discussed surrounding previous studies on the Loan Default prediction topic in hopes of gaining useful insights for my research and data set. Furthermore this will serve as a deep dive into the features that will be used (or not) for this research. After Literature Review, will be the modeling phase in which insights gathered from the data will be highlighted after analysis and key features will be discussed as potentially important predictors of loan defaults. Finally, I will wrap things up with a conclusion providing a summary of my findings, best performing models for this study, important features and potential routes for future research. Through the framework that is used for this project I will be able to share the steps I take to come to the conclusions of my study in a sound comprehendable format.

Literature Review

Loan Default Prediction has been a widely experimented area in Data Science with many different models used to predict outcomes of lending. In order to conduct an effective experiment I decided to narrow my research down to the algorithms shown to be the most successful at loan default prediction in order to use as a ‘North Star’ for building my own model.

According to the sources I have gathered, Decision Trees, Logistic Regression, Random Forest, XGBoost, Naive Bayes and LightGBM have been the most successful in recent years. Some considerations taken in the study are model complexity and performance. Typically we see that XGBoost is largely popular in that aspect since it is a relatively new advancement in

machine learning often outperforming other algorithms in classification problems. LightGBM, which is a similar approach to XGBoost that runs much faster and saves memory has been a top performing model in some studies as well. However, another important aspect often considered is model explainability. Machine Learning oftentimes is referred to as black box learning since many models are not able to be interpreted easily. When there is a lot of weight on explainability models to build faith in a model, algorithms such as Decision Trees, Logistic Regression and Naive Bayes are great options. A third consideration hinted at before when I explained LightGBM is the time it takes to train and test a model. Given time constraints it may be difficult to test every and all combinations to arrive at the best result and even if such a result is reached the Data Scientist must consider whether the performance of the better model outweighs that of another model that may be faster and/or more explainable.

All algorithms used data related to customer loan applications and used some sort of statistical metric to measure performance. Some of the methods used were F1-Score, Accuracy, Precision, Recall and AUC/ROC. This is worth mentioning in relation to the algorithms themselves because some algorithms may perform better in one metric while performing worse in others. The method that will be used for my model will have to make sense to convey how the model's performance will affect their business, i.e how much money gained by dishing out more loans to qualified applicants vs how much money will be lost for mistakenly classified loans.

In order to build a loan default prediction model of my own I needed to be assured that I had features in my dataset that had historical backing for predicting loan default in past studies. By referencing past research to validate the features of my dataset I am able to gain confidence in the data and more explainability on a feature level. Being able to string together logic behind why some features are important can greatly enhance the story to my audience in case questions are raised. With that said, any features out of the 36 that I was not able to find past research on will not be used as they are not justified and the relationship to loan defaults not explainable.

Of the features available in my dataset, some of the most commonly used to predict loan default are type of the loan, term of the loan, interest rate on the loan, debt to income ratio and transaction cost of the loans(might be extracted from Loan Amount variables). Loan Amount is another feature often used which in this dataset has three related features in Loan Amount

(amount applied for), Funded Amount (Actual), and Loan Amount Investor (Investor approved). It is assumed that a better candidate will receive an amount of more or closer to the amount they applied for therefore all three features may prove useful here. We can subtract the actual loan amount from the investor approved amount to see if we can extract a variable that can act as the transaction cost of the loans. Other variables that have been used in loan default prediction models and available in this dataset are delinquencies (30+days) in past 2 years, credit grades (and sub grades), number of inquiries, number of open accounts, revolving balances, revolving credit utility, total interest received (payments), late fees, recoveries, collections in the last 12 months,account-type (joint or individual) collections amount, public derogatory records, number of delinquent accounts, total balance across accounts, total revolving credit limit. Lastly, is the target variable, loan status which indicates whether a loan has defaulted or not with a label of 1 or 0 assigned respectively. In total there are 29 usable features representing the aforementioned variables in this dataset for building a loan prediction model.

Although I excluded some features for not having any backing in previous studies, I will still have to consider dropping more of the features upon conducting exploratory data analysis since there may be issues with data such as single value columns, missing values or other factors making the selected features unusable. With this initial feature selection I should be able to produce a well structured model and be able to speak on the features backed by my references.

Overview The Data and Preparation

The dataset used for this analysis was gathered from Kaggle but was originally composed by MachineHack, a leading hub for Machine Learning and AI in partnership with Deloitte. The data consists of 67,463 rows and 35 columns. As discussed in literature review only variables backed by research will be used for analysis and modeling. Below are tables summarizing the meaning of each variable in the data, a univariate analysis of the numerical variables and a summary of the categorical variables.

Data Dictionary

Feature Name	Definition	Type
Loan Amount	Loan Amount applied for in dollars	int64
Funded Amount	Actual amount funded by investor after fees	int64
Funded Amount Investor	Amount to be funded before fees	float64
Term	Loan Term in months	int64
Interest Rate	Interest rate assigned to loan	float64
Grade	Credit Grade of borrower application	object
Sub Grade	Sub Grade of borrower application	object
Employment Duration	(Homeownership) Variable name is misaligned this is actually homeownership status	object
Home Ownership	(Income) After research on MachineHack this appears to be meant labeled as Income in dollars	float64
Verification Status	Verification status of reported income	object
Loan Title	The reason for the borrower requesting a loan as written by the borrower	object
Debit to Income	(Debt to Income Ratio) Monthly debt divided by monthly income	float64
Delinquency - two years	Number of delinquencies in last two years	int64
Inquires - six months	Number of credit inquiries in last 6 months	int64
Open Account	Number of active open credit line accounts	int64
Public Record	Number of derogatory public records	int64
Revolving Balance	The balance that carries over from one month to the next	int64

Revolving Utilities	Utilization refers to how much of your credit balance you're using at a given time (monthly)	float64
Total Accounts	Total accounts (open and closed).	int64
Total Received Interest	Total Interest payments received to date	float64
Total Received Late Fee	Total late payments received	float64
Recoveries	Recovery refers to the total amount of money recovered by a lender after a loan has been charged off. It means that the lender has declared the loan as uncollectible.	float64
Collection Recovery Fee	The Collection Recovery Fee is typically a percentage of the amount recovered by the collection agency	float64
Collection 12 months Medical	Number of collections excluding medical	int64
Total Collection Amount	Total amount ever recovered by collections	int64
Total Current Balance	Total credit line balance from all accounts	int64
Total Revolving Credit Limit	Total credit card limit across all accounts	int64
Loan Status	Whether the account defaulted or not. 1= Defaulter , 0 = Non-Defaulter	int64

Univariate Analysis (Continuous)

feature	min	max	mean	std	missing
Loan Amount	1014	35000	16848.90278	8367.865726	0

Funded Amount	1014	34999	15770.59911	8150.992662	0
Funded Amount Investor	1114.590204	34999.74643	14621.79932	6785.34517	0
Term	36	59	58.17381379	3.327440547	0
Interest Rate	5.320005799	27.18234758	11.8462579	3.718628715	0
Home Ownership	14573.53717	406561.5364	80541.50252	45029.12037	0
Debit to Income	0.675299086	39.62986189	23.29924062	8.451823721	0
Delinquency - two years	0	8	0.3271274625	0.8008883778	0
Inquires - six months	0	5	0.1457539688	0.4732912877	0
Open Account	2	37	14.26656093	6.225060448	0
Public Record	0	4	0.08143723226	0.3466057424	0
Revolving Balance	0	116933	7699.342425	7836.14819	0
Revolving Utilities	0.00517236	100.8800498	52.88944256	22.5394504	0
Total Accounts	4	72	18.62792938	8.319246431	0
Total Received Interest	4.736746327	14301.36831	2068.992542	2221.918745	0
Total Received Late Fee	3.06E-06	42.6188823	1.143968627	5.244365117	0
Recoveries	3.56E-05	4354.467419	59.69157773	357.0263463	0
Collection Recovery Fee	3.62E-05	166.833	1.125140937	3.489884545	0
Collection 12 months Medical	0	1	0.02130056475	0.1443854553	0
Total Collection Amount	1	16421	146.4679899	744.382233	0
Total Current Balance	617	1177412	159573.9336	139033.2456	0
Total Revolving Credit Limit	1000	201169	23123.00554	20916.7	0
Loan Status	0	1	0.09250996843	0.2897466455	0

Categorical Variables

Feature	Number of Unique Values	Missing
Grade	7	0
Sub Grade	35	0
Homeownership	3	0
Verification Status	3	0
Loan Title	109	0

Preparing Data for Modeling

For the modeling following this phase, I have decided that I will be using only three algorithms as candidates for my loan prediction model. The first of the three models I will be using is Logistic Regression as it is classically used for such problems as Loan Prediction and offers a lot of explainability and tuning options. The second algorithm will be an XGBoost classifier as extreme gradient boost is a recent breakthrough often outperforming most other models. The last model will be a Random Forest model as it is one of the top performing models and is able to handle imbalanced data while having the benefit of being able to extract Feature Importance just like the XGBoost classifier.

The variables in the dataset needed to be reworked so that they were more useful for my models. First, the column names for Homeownership and Employment Duration were replaced as the first was actually the Income and the latter was the Home Ownership status. Next, the numerical variable for Loan Term had categories of 36, 58 and 59 months. I made the assumption that loan terms were meant to be 3 and 5 years so I converted the type to category and reduced the categories to 3_years and 5_years. The numerical variable for whether a borrower had a collection on their record in the last 12 months was also converted to categorical with values of 0 and 1. Income verification status had 3 categories of Not Verified, Verified and Source Verified. I decided to consolidate them into just Verified and Not Verified. The Loan Title variable has 109 unique entries which overlapped and were only differentiated by spelling and alternate descriptions. I was able to consolidate the values to just 9 unique categories based on

the loan types and the words contained in each category. Upon inspection of the distribution of values and Loan Amounts, I concluded that the bottom four categories did not have enough instances to draw conclusions from as over 95% of the default loan instances belonged to the top 5 classes. For this reason, rows containing those values were dropped. The remaining 5 categories were Refinance, Consolidation, Other Loan, Home Improvement and Major Purchase. Lastly, the variable for Accounts Delinquent was dropped since it only contained 1 unique value which would not be helpful for my models to make decisions.

Based on research none of the models I have chosen strictly require specific preprocessing techniques however they can still benefit from using normalized and unskewed data and such. For this reason I created a preprocessing pipeline where I have applied skewness removal with a Power Transformer which is similar to a log transformation and can handle negative values. I have also applied a scalar function to make the data more normally distributed as well as performed outlier removal. For each step in the pipeline I have experimented with the available alternatives. Lastly, I included a normalizer for good measure so that all values are between 0 and 1. Each model will react differently to the pipeline so the best combinations for each will have to be documented. In my pipeline I have also included an Imputer which will apply the most common value for any missing values detected; however, I did not detect any missing values in the data. It is simply a formality. With these preparations set I am ready to start experimenting with models and evaluating the results.

Modeling

As explained in the previous section, I have chosen 3 models to use for Loan Default prediction. The 3 models are Logistic Regression, XGBoost and Random Forest. For testing purposes the data was split 80/20 with 20% of the data reserved for the testing set. For each algorithm, I conducted a 10 fold cross validation for training to ensure consistent performance. Since our client will care about not just detecting loan defaults but also missed opportunities, I have chosen F1 as the main metric of success. F1 scores represent the harmonic mean between Precision and Recall. Precision measures the correctly classified positive predictions against all labels predicted to be

positive. Recall measures the correctly predicted classes out of all the actual positive labels.

Logistic Regression

Logistic regression is the most widely used model in loan default prediction and often works well even when features are not highly correlated. In this case however, it was the worst performing model with an F1-Score averaging around .55. Even after hyper tuning I was not able to increase the F1- Score. Interestingly, this model performed best at detecting Loan Default with and F1-Score for the Loan Default class of 0.16. This is still however very low of a score to call this anything close to reliable. The accuracy score for this model is .52.

	precision	recall	f1-score	support
0	0.91	0.52	0.66	12063
1	0.10	0.50	0.16	1231
accuracy			0.52	13294
macro avg	0.50	0.51	0.41	13294
weighted avg	0.84	0.52	0.62	13294

The coefficients for each variable are weights used to determine how our target variable y responds to the movements of each independent variable X. After pulling the coefficients and ordering them by greatest absolute value to least I was able to determine that the top five most predictive variables for this model are Total Collection Amount, Collection Recovery Fee, “Refinance” as the Loan Title, a credit Subgrade of B1and if the Loan Title is “Consolidate”. While the model’s performance is poor on test data it does appear that these variables would make sense as top indicators of Loan Default. According to the results if a value is positive or the higher the value is among Total Collection Amounts, Refinance Loans, Sub-Grade of B1and or a Consolidation loan, the more likely the result of the target class will be positive. However, if the

Collection recovery fee is lower, it has a negative weight affecting the target variable in the opposite direction.

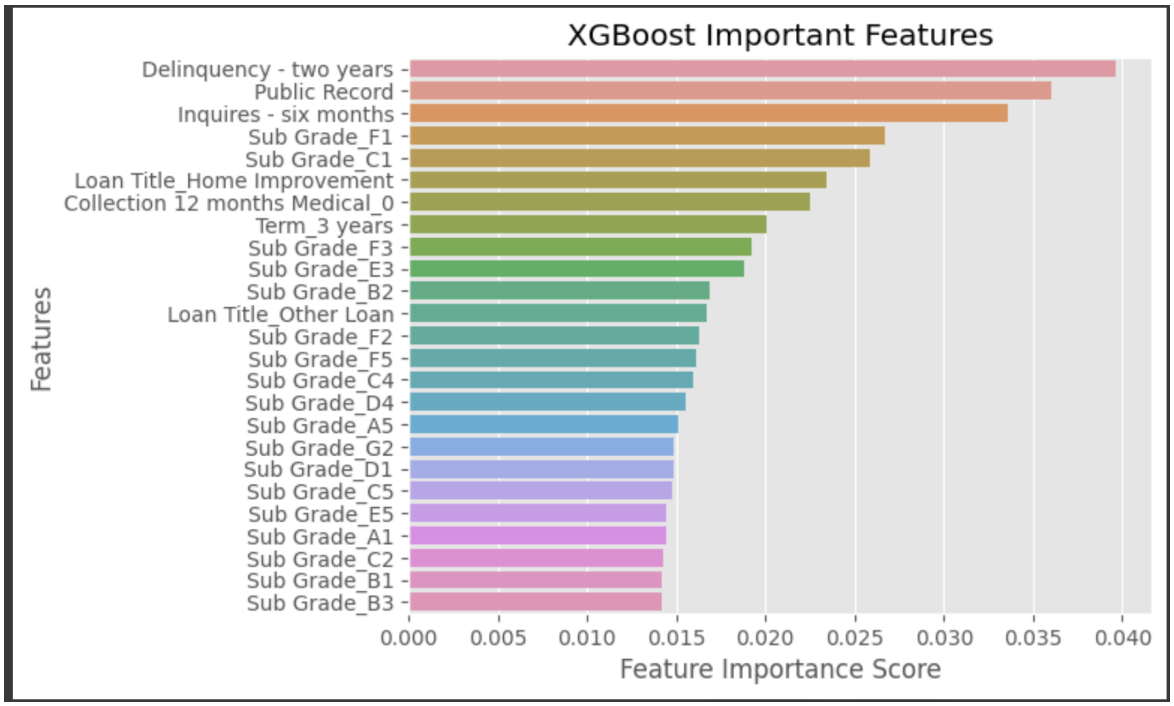
	feature	coef
17	Total Collection Amount	0.511459
16	Collection Recovery Fee	-0.255351
40	Loan Title_Refinance	0.222934
27	Sub Grade_B1	0.215073
39	Loan Title_Consolidation	0.200778
25	Grade_G	-0.162447
30	Sub Grade_B4	0.154479
8	Open Account	-0.142061
26	Sub Grade_A5	0.129500
15	Recoveries	0.125072

XGBClassifier

The XGBoost is a powerful model based on Decision trees and Gradient Boosting, that I expected to be the greatest performer. Similar to a multi-layer perception, XGBoost will try to minimize the error calculated from each tree to improve upon it in the following trees. The model had about a .95 average F1 Score after 10 cross validations during training and closer to .96 after hypertuning. However, this model performed poorly on the test data as the F1-score was .02 of the actual loan default class. The accuracy score for this model is .90. Overall, this model is not an effective predictor of loan default.

	precision	recall	f1-score	support
0	0.91	0.99	0.95	12063
1	0.08	0.01	0.02	1231
accuracy			0.90	13294
macro avg	0.49	0.50	0.48	13294
weighted avg	0.83	0.90	0.86	13294

According to the extracted feature importance graph, the top 5 features in order from most significant to least are Delinquencies in two years, a credit Subgrade of F1, Public Record, Subgrade of G5 and and number of inquiries in the last 6 months. These scores ranged from .025 to .04. The top 3 features of Delinquencies, Sub-grade F1 and Public Record were significantly more important than the rest of the features. This group of top 5 features seems logical in deducting that they would be good predictors of Loan Default as they all have to do with poor credit and a history of not being able to pay back previous loans. Compared to the coefficients from the logistic regression this seems to make more sense.

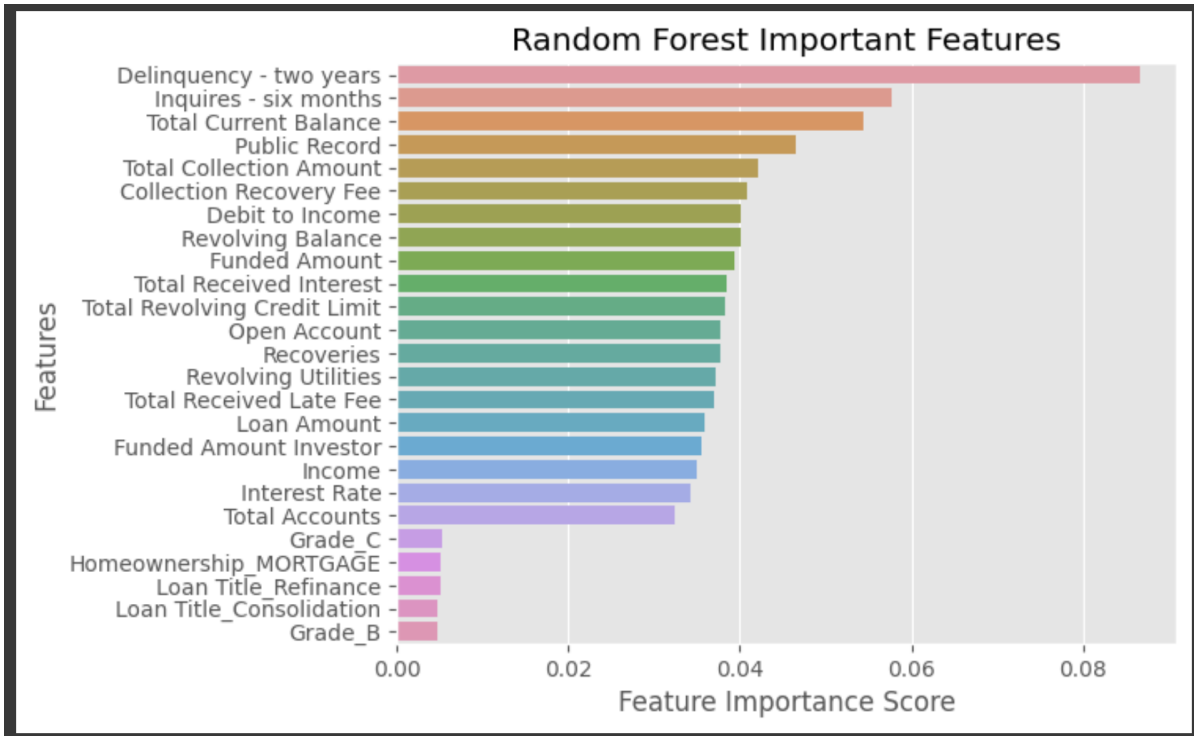


Random Forest

The last model used for predicting loan is the Random Forest which is another Decision Tree based model that takes the average of its trees' classifications as the final result for each instance. Interestingly, this model performed with an average F1-Score of .97 during the same training process as the last two models. However, hypertuning only worsened the model for training as the average score dropped to .95. This process was done with a RandomSearchCV in python. It may have been more complex and required a GridSearch but I did not have the time or computing power to execute this type of Hypertuning efficiently. Regardless, the model performed poorly on testing data. The model has an F1 score of .04 for the Loan Default class and a .94 for the non-Loan Default class.

	precision	recall	f1-score	support
0	0.91	0.98	0.94	12063
1	0.10	0.02	0.04	1231
accuracy			0.89	13294
macro avg	0.51	0.50	0.49	13294
weighted avg	0.83	0.89	0.86	13294

The feature importance for this model showed that the top 5 features had more significance than the ones for the XGBoost model. Feature Importance Scores ranged from .041 to .091. The most significant features are Delinquencies in the last two years, Inquiries in the past 6 months, Total Current Balance, Public Record and Total collection amount. This model seems to suggest that Delinquency is the most important feature of all as the score is .091 with the next feature being Inquiries which has a score of .056. In conclusion, this model also performed poorly in terms of predicting Loan Default and separating out enough features to determine the most important.



Summary

In this project I explored answering the question: How can we effectively assess and mitigate the risk of loan default in our lending portfolio by leveraging borrower characteristics and loan attributes? With the resources gathered and results of my predictive analysis I have come to inclusive results. A model able to predict loan default was not achievable in the time allotted however, I do believe it is possible, with good data, to predict Loan Defaults using the same or similar variables. For example, in all three model features signaling poor credit or a record of inability to pay back loans were given higher levels of importance or coefficient values. This tells us that the models were able to identify such a correlation and naturally as humans we can deduct that these should in fact be variables that are heavily considered when giving out a loan as they indicate that the borrower is high risk.

Some reasons I discovered as to why the models were not great at predicting loan default have to do with the data quality. When observing the 3 different variables for borrowed amounts there seemed to be very small correlations among them where it would be assumed there should be larger correlations. After conducting a formal analysis with a heatmap it was

found that none of the variables were highly or even moderately correlated. While high correlations can cause there to be bias, low correlations can be completely useless. In an attempt to improve correlations I applied outlier removal techniques as well as filtered the data by creating a new variable calculating the ratio between Funded Amount and Requested Amount and removing the lower and upper extremes. While this did slightly boost correlation the model ended up not being able to predict any of the Loan Default Class values. Another reason why there may have been poor results in predicting the loan default instances is due to the imbalance of the data. I don't believe there were enough instances of fraud in the data for the model to be applied on new data. I tried to improve results by applying SMOTE to increase the number of Fraud instances however it could be since there were so few, the model was not able to make effective 'replicas' of fraud instances. The last major limitation is the interpretation of the results. Rather than trying to predict default, the goal is to mitigate money loss due to high risk accounts defaulting. In this respect, I found that binning the prediction probabilities and comparing them to the actual instances of default is more useful for a bank than trying to predict loan default itself.

It is possible to create a great model for predicting loan default and there is proof because it has been done before. However, it is important to consider data quality and the source of the data before conducting such a study in a serious manner. In the future, I will more carefully evaluate a data set and run a test model (perhaps logistic regression) to ensure the dataset is a good candidate for a model to learn from. It is important for data to be correlated so that relationships can be identified and calculated. Catching this early on will prevent wasting time. In a true Data Science scenario I would utilize subject matter experts to better understand the data I am working with and how it is used on the Data Analyst side for decision making. As it stands the variables and their values were simply not effective despite attempts to preprocess the data for the algorithms used. Lastly, how the results are interpreted and presented to clients in a way that the results will be useful should not rely on a metric such as f1-Score because it is assuming banks will want to avoid giving out loans to defaulters in general. A better approach may be to raise interest rates on higher risk groups by finding a model that is optimal at isolating the most actual Loan Defaults. Below is the best example that I was able to come up with where the Logistic Regression model was able to isolate the largest group of actual defaults.

Logistic Regression Probability Bins and Actual Defaults:

actual	0	1
bins		
(-1.0, 0.0]	0	0
(0.0, 0.2]	0	0
(0.2, 0.4]	161	16
(0.4, 0.6]	11803	1207
(0.6, 0.8]	99	8
(0.8, 1.0]	0	0

In the table above you can see that the bin with a probability score between .4 and .6 was where 98% instances of loan default occurred. The same was done with XGBoost which captured 84% of fraud cases between .2 and .4 probability scores. The Random Forest model performed the worst with results scattered across different bins. With more time this strategy could be explored in depth and seems to be more useful for explaining the effectiveness of the chosen model which in this case would be the Logistic Regression as it is better at grouping the actual loan default cases within a given group of probabilities thus allowing me to identify which accounts should be given higher interest rates to mitigate losses from default. If I were to take the results as they are, I would interpret them to mean that most borrowers who default are of moderate risk and the interest rate should be adjusted accordingly. Ofcourse, this is still not an efficient model as more effort needs to be put in to separate the largest group of defaulters from non-defaulters but it is a start.

References

1. Cal State ScholarWorks - Download
2. Scaler - "Loan Default Prediction" - Link
3. Medium - Selena Zhao - "Predicting Loan Defaults Using Logistic Regression" - Link
4. EFMA - "Loan Default and Returns on Investment Analyses" - Link
5. Towards Data Science - "Machine Learning: Predicting Bank Loan Defaults" - Link
6. AWS Industry Lab - "Predict Loan Defaults" - Link
7. Journal of Financial Economics - "Loan Characteristics and Credit Risk" - Link
8. Scirp - "Predicting Loan Default Using Machine Learning Techniques" - Link
9. ScienceDirect - "A Deep Learning Approach to Loan Default Prediction" - Link
10. Towards Data Science - "Machine Learning: Predicting Bank Loan Defaults" - Link