BU425 - Business Analytics

Final Project

Dr. Hamid Elahi

Credit Rating Prediction

April 10th, 2023

Stefan Bukarica 190563930

Krishna Khajuria 183206760

Rufael Musa 180571340

Jake Lambkin 180658600

Table of Contents

Abstract

Credit ratings predictions are extremely important in business, economics and finance. And with the drastic increase of data available to companies due to the advancement of technologies etc., companies are trying to leverage this abundance of data to make accurate informed decisions to maximize their profits and competitive advantage. Predicting the credit rating of a customer allows a financial institution to determine how to market various credit related products to current and future customers (should the data exist).

In this project we classify the credit ratings on 3 levels: Good, Standard, and Poor. We reduced the dimensions and features of our data to suit model training as well as modified our features to be numeric and scaled. We also made sure to not scale categorical (numeric) values where the scale is irrelevant.

Using this cleaned and pre-processed data, we created 3 models to properly classify the credit rating of individuals. These include a KNN Classifier, a Random Forest Classifier, and a Support Vector Classifier. Each with their respective benefits and drawbacks. Ultimately, we found that the Random Forest Classifier performed best when predicting the credit rating of unseen (test) data. Since accuracy is an important measure of how the financial institution would use these models we recommend using this model as it is most accurate.

Introduction

Credit rating prediction is a crucial area of research in finance and economics. Credit rating refers to a quantified assessment of a borrower's creditworthiness in general terms or with respect to a particular debt or financial obligation. A credit rating can be assigned to any entity that seeks to borrow money—an individual, a corporation, a state or provincial authority, or a sovereign government(Limited, *Predictive modeling mitigates credit risk*). Credit ratings play a vital role in financial markets as they are used by lenders, investors, and other stakeholders to make informed decisions about the creditworthiness of borrowers.

In recent times, with the increasing availability of data and advancements in machine learning and artificial intelligence, there has been a growing interest in using predictive models to automate the credit rating process. Such models can help improve the accuracy and efficiency of credit rating assessments, reduce the risk of defaults, and increase access to credit for borrowers.

This report aims to provide a comprehensive overview of credit rating prediction, including the key concepts, methodologies, and challenges in building and deploying credit rating models. Additionally, we will review the business question/problem that we are trying to answer/solve, the steps we took to pre-process the data, the models and techniques used to analyze the data. Finally, we will draw conclusions from our analysis and give recommendations for our business problem.

Business Problem

As briefly hinted in the previous section as more and more data becomes available to companies, companies have started leaning towards creating robust predictive models to automate their credit rating processes.

Some advantages of using predictive credit models are: 1. Increased access to credit and profits: One of the biggest challenges for individuals or businesses with limited credit history or poor credit scores is accessing credit. Credit rating prediction models can help lenders identify creditworthy borrowers who might otherwise be overlooked, allowing them to offer loans and other credit products to a wider range of customers. Which would directly help increase profits. 2. It can help reduce risk of fraud: Credit rating prediction models can also be used to detect fraudulent activities, such as identity theft or loan stacking, where borrowers take out multiple loans simultaneously. By identifying potential fraudsters early on, lenders can mitigate their risk exposure and prevent losses. 3. Improved customer experience and customer retention: Obtaining credit products usually tends to be a time consuming process especially if done manually, but with the help of predictive models lenders can expedite the application process and come to decisions quickly saving time which helps improve the overall customer experience and as a result increases customer retention. 4. Better pricing and risk management: By predicting creditworthiness

lenders are able to manage their risk exposure more effectively. And are able to adjust their interest rates and other credit risks which lead to more profitable leading portfolios.

Given these advantages which act as the main motivation for creating a model of this nature and given how important it is in the business context we want to answer the following Business questions: What attributes does a person with bad credit have? Which variables impact a person's credit score the most? As well as how banks can use this information to make better marketing decisions about clients in the future?

EDA and Pre-Processing

In this section of the report we go over the data that we used and what we did in order for it to be appropriate to train our models on. Initially the dataset consists of 150,000 records, 100,000 in the train.csv and 50,000 in the test.csv. We combine these two csv's and redo the train_test_split since we want an 80/20 split. The dataset has a mix of categorical and continuous features. We are trying to predict the feature "Credit_Score" which indicates the credit rating of a customer of this bank with 3 levels; "Good", "Standard", "Poor". It is important to note that each record in the dataset represents an event regarding a customer in the bank where credit is considered, this could be anything from a car payment to a business loan application. The raw dataset consists of the following 27 features:

- ID - Represents a unique identification of an entry
- Customer_ID - Represents a unique identification of a person
- Month - Represents the month of the year
- Name - Represents the name of a person
- Age - Represents the age of the person
- SSN - Represents the social security number of a person
- Occupation - Represents the occupation of the person
- Annual_Income - Represents the annual income of the person
- Monthly_Inhand_Salary - Represents the monthly base salary of a person
- Num_Bank_Accounts - Represents the number of bank accounts a person holds

- Num_Credit_Card - Represents the number of other credit cards held by a person
- Interest_Rate - Represents the interest rate on credit card
- Num_of_Loan - Represents the number of loans taken from the bank
- Type_of_Loan - Represents the types of loan taken by a person
- Delay_from_due_date - Represents the average number of days delayed from the payment date
- Num_of_Delayed_Payments - Represents the average number of payments delayed by a person
- Changed_Credit_Limit - Represents the percentage change in credit card limit
- Num_Credit_Inquiries - Represents the number of credit card inquiries
- Credit_Mix - Represents the classification of the mix of credit scores (in the test set)
- Outstanding_Debt - Represents the remaining debt to be paid (in USD)
- Credit_Utilization_Ratio - Represents the utilization ratio of credit card
- Credit_History_Age - Represents the age of credit history of the person
- Payment_of_Min_Amount - Represents whether only the minimum amount was paid by the person
- Total_EMI_per_month - Represents the monthly EMI payments (in USD)
- Amount_invested_monthly - Represents the monthly amount invested by the customer (in USD)
- Payment_Behaviour - Represents the payment behavior of the customer (in USD)
- Monthly_Balance - Represents the monthly balance amount of the customer (in USD)
- Credit_Score - Represents the bracket of credit score (Poor, Standard, Good) (in the train set, this is the same as credit_mix so when we combine the train and test set we drop credit_mix and take rows with non-null ‘Credit_Score’)

Due to the sheer size of the dataset we opted to drop all records containing NA’s as this left us with ~58,000 records, which is still a large segment to use for training and testing for our data. Next, we dropped all ID columns these included the following: 'ID', 'Customer_ID', 'Month', 'Name', 'SSN', and ‘Credit_Mix’ as it is the same as Credit_Score, and ‘Type_of_Loan’ as there were too many types of

loans within that categorical feature. Next we removed all inconsistent data from the features, this could be things such as negative ages of customers, values set to '-3333' or '____'. As well as removing extra symbols within a column. Then we converted all categorical variables to be numeric. This is important because we want to use KNN which is a distance based algorithm that needs numeric input in order to work. After these adjustments we are left with around ~45,000 records for training and testing. The last step before finalizing our dataset was to make dummy variables for the 'Occupation' feature and scale all of the non-categorical data using a StandardScalar. We needed to scale the data since many of the continuous features are on drastically different scales such as Age and Monthly_Balance. Now that we have done all of these things, our final dataset is ready to be used for our models.

Techniques

In order to solve the business problem at hand we need to create models capable of predicting a multi-class variable both efficiently and effectively. With that in mind, this section of the report will cover the process we took to pick, train, test, and improve upon the models chosen in order to produce the best results possible. Prior to training we were able to choose a few models which meet the requirement of predicting multi class variables. Decision trees were the most obvious choice as they are capable of classifying multi-class variables in an easily interpretable way. Also to reduce some of the downsides to decision trees, we planned to use the random forest model of decision trees in order to reduce overfitting and increase robustness while maintaining interpretability. K Nearest Neighbors (KNN) is another obvious choice as it is able to predict multiple classes and is also easily interpretable. Although base Support Vector Machines (SVM) are not capable of multi-class prediction, SVM classifiers are capable of it with the right kernel and other parameters. With that being said we decided to go forward with training these models to determine which one works best with the chosen dataset.

For ease of use in testing, we created a class to represent a classifier model, where to create a classifier object of the class we simply pass in a model and the data as parameters. We created a variety of class methods which can be used to train, evaluate and improve a model. With this setup we were easily able to train each of our models and adjust parameters as needed. The first method we created called train,

simply splits the data into the train test split, fits the model as is. It then predicts using the test data and provides the object with accuracy, recall, precision, and F1 scores through .SCORE for the object and a confusion matrix through .cm. We then create a method called evaluate which performs k-fold cross validation on the given model and outputs the cross validation score in a variety of metrics with .CV_SCORE. By splitting the dataset into a number of folds we can evaluate how effective each model is when given new data. This is important to check as it can help us to recognize if the model is over or underfitting so that we can make needed adjustments to parameters. In order to improve upon our models in the most effective manner we created a grid search function as a part of the class. This allows us to test various hyper parameter configurations for each model as they are placed within a grid which is searched and tested so that the most optimal configurations are chosen. The downside to this method is it can increase computation times quite a bit, especially on large datasets. However, with time it is a much better solution compared to manually testing different parameter configurations through trial and error.

With our classifier model class complete we began testing it with the three models we were going to try. The KNN classifier performed fairly poorly at first with scores around 60% across all metrics. Grid search greatly improved the results to around 65% after quite a bit of testing (figure 1). KNN may have performed the best with continued hyper parameter tuning and changes to initial data processing, but with the dataset being quite large KNN was quite time consuming and was not deemed worth the continued effort. The Random Forest Classifier although also suffering from lengthy computation resulted in the most accurate results by far after hyper parameter tuning (figure 2). Of course compared to the base decision tree model as well as KNN some interpretability is lost. However with results well over 75% it was the best model to use in our analysis of the business problem at hand. With our process being outlined we can now move onto discussion of our analysis, its outcomes, and our recommendations in order to answer the questions we would like to answer with our models.

Analysis, Outcomes, and Recommendations

With the random forest classifier producing the best results after training and evaluation, we decided to use the information we gain from the model to answer the questions outlined earlier on in the report. In this portion of the report we aim to provide answers to these questions and give recommendations to banks or other businesses based upon the results of our data analysis. First we would like to determine the factors which indicate a person has good or bad credit. In order to achieve this we can determine the variables with the greatest impact on a person's credit score and build profiles for them from there. The random forest model has an attribute called feature_importances_ which computes scores relating to how important each feature was in predicting the target variable during the training and evaluation process. Figure 3 shows the feature importance scores for each of the independent variables used to train the models. Using these scores we can gain insight on the most important attributes contributing to one's credit.

The most important factor by far according to the random forest is a person's outstanding debt. This is an expected result as a person with poor credit is likely to have a lot of outstanding debt and a person with good credit is likely to have less debt as they are more likely to be making payments in time. However, it is not a tell-tale sign of someone's credit, for example, a person with very little credit history is likely to have a poor credit score but less debt. This leads us to another of the most important factors affecting a credit which is a person's credit history age. Although a person's actual age often correlates with their credit history age it is a far more important metric in predicting a person's credit. As a person goes throughout life and uses credit their score is likely to increase overtime. The second most important factor according to the random forest model is the person's interest rate. This is another more obvious one, as people with higher interest rates are more likely to be in debt as banks set higher rates for people less likely to pay. Another of the more interesting factors with a higher importance score is a change in credit limit. An increase in credit limit generally brings a person's credit score up, as if they were accepted for a higher limit they are likely to be someone who is good at making their payments. However, this is not always true as some may take on an amount of debt they shouldn’t have with an increased limit. The

remaining factor more important than the majority is the delay in someone's payment from the due date. This is another rather obvious factor as the longer someone delays making credit payments the lower their credit score is likely to be or will become.

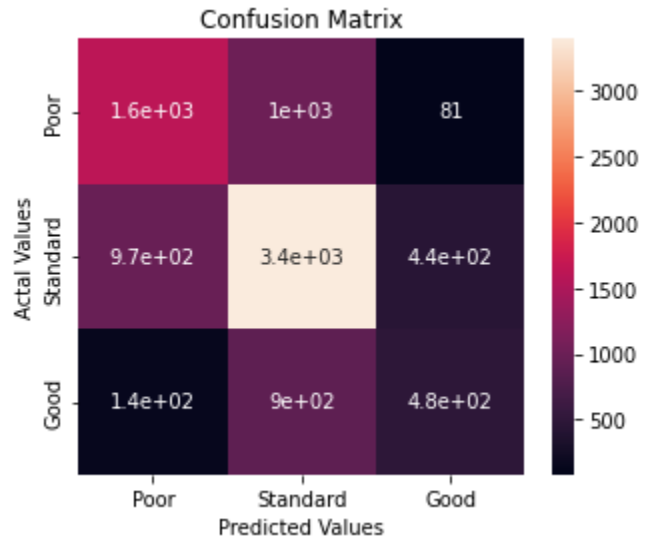
With the above factors in mind, we can determine that a person with poor credit is likely to have more debt, less credit history, higher interest rates, and a lesser ability to make regular payments. This profile perfectly fits that of students and young people new to the workforce as all or most of these factors are likely true to them. Some of these factors are also likely to affect immigrants who are new to a country and have less credit history there. We can also determine a person with good credit to be someone with the opposite factors, generally people with more established careers and a longer history of credit as they are more likely to have the ability to make more payments resulting in less debt.

All of this information and analysis allows us to answer our final question and give recommendations to interested parties. Of course the information gained from the model can help in its most obvious use case of predicting the credit new clients are likely to have. The results of this project can also be used to guide new marketing and program strategies which could lead to improved customer retention and increased profits. Marketing programs towards different demographics such as students that are tailored towards them with factors of poor credit scores, a lot of debt and being new to using credit in mind. This would bring a lot of customers in the younger generation which are in school and likely to become valuable customers in years to come. Another recommendation we would make is to use the analysis done here and in other similar areas to aid in determining interest rates as they are such a large factor in the credit programs customers choose and good programs lead to better and lasting customers.

Conclusion

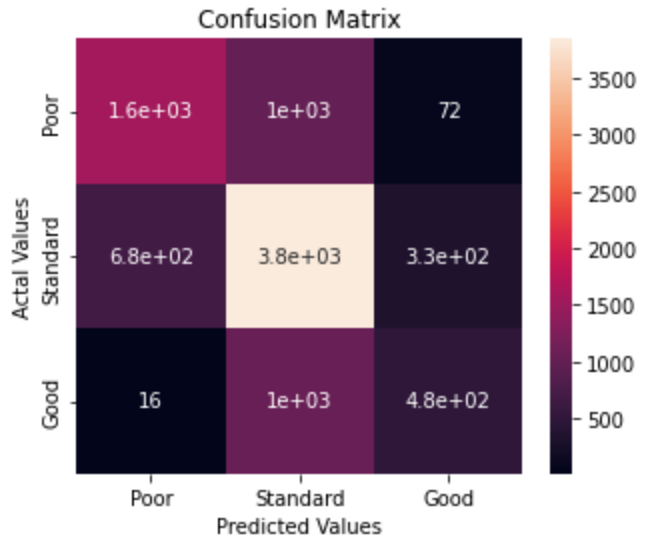
As machine learning becomes a more and more integrated part of many industries it is more important than ever that businesses use the tools accessible to them to keep up with the rest of the industry. This doesn't exclude the finance industry, with so much rich financial data available there is much we can learn from it. Throughout this report, we have described our dataset, the business problem we are trying to solve, our process in creating models to solve the problem, and the analysis of our results. For our analysis, we used the random forest classifier and provided banks with useful information regarding the factors determining a person's credit and how they can use this information to make better decisions in the future.

Appendix



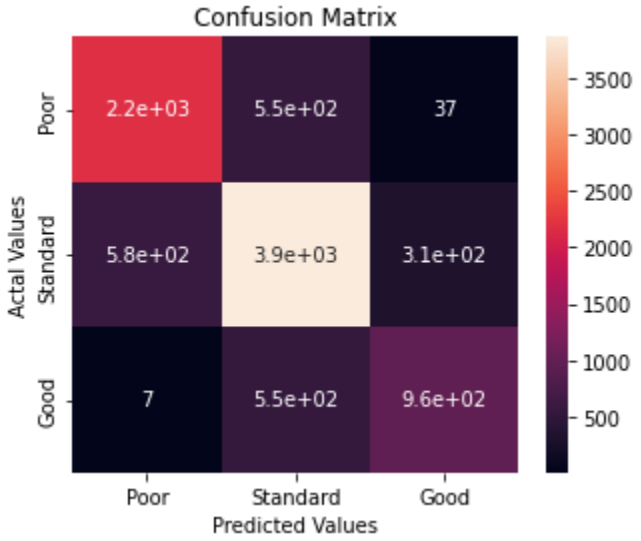

```
Non-tuned: {'Accuracy ': 0.6076369673491976, 'Recall': 0.6076369673491976, 'Precision': 0.5985703395453034, 'F1': 0.5994591137410256}
```

Figure 1 - Non-Tuned KNN Confusion Matrix and Scores



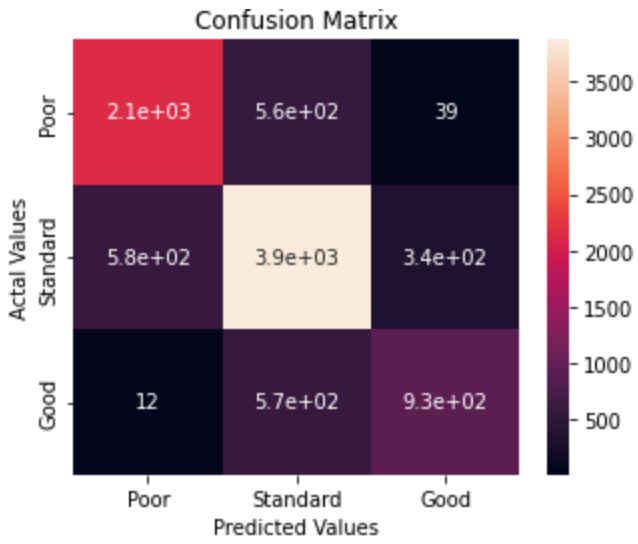
Tuned: {'Accuracy ': 0.6515771997786386, 'Recall': 0.6515771997786386, 'Precision': 0.64580842365112, 'F1': 0.6382880706848123}

Figure 2. - Tuned KNN Confusion matrix and Scores



Non-tuned {'Accuracy ': 0.7748754842280022, 'Recall': 0.7748754842280022, 'Precision': 0.773835607267528, 'F1': 0.7735788304400069}

Figure 3 - Non-Tuned Random Forest Confusion Matrix and Scores




```
Tuned {'Accuracy ': 0.7683453237410072, 'Recall': 0.7683453237410072, 'Precision': 0.7669091350769917, 'F1': 0.7668385982102705}
```

Figure 4 - Tuned Random Forest Confusion Matrix and Scores

```
[0.03780962 0.04262955 0.04301738 0.03779695 0.04176953 0.08331468
 0.02610976 0.06521    0.0473212  0.06233572 0.05009197 0.11259655
 0.04603662 0.06224595 0.02892959 0.04370715 0.04519724 0.02069067
 0.04718281 0.00377649 0.00373208 0.00367462 0.00382521 0.00350111
 0.0040528  0.00364423 0.00377374 0.0036883  0.00378912 0.00340211
 0.00366246 0.00411253 0.0038874  0.00348487]
```

Figure 3 - Feature Importances from the Random Forest Classifier

- The values in the above array represent each feature's importance score (the higher the more important). The numbers appear in the same order of the list of features shown in the preprocessing section of the report.

Resources:

Github Repository: https://github.com/1-stefan/BU425-Project

Paris, R. (2022, June 22). Credit Score Classification. Kaggle. Retrieved April 10, 2023, from https://www.kaggle.com/datasets/parisrohan/credit-score-classification

Limited, I. (n.d.). Predictive modeling mitigates credit risk. Retrieved April 10, 2023, from https://www.infosys.com/industries/mining/case-studies/predictive-modeling.html