Illinois Institute of Technology

Michael Neely (A20393733), Hao-Yu(A20485217), Abhishek Shastry HM(A20468925)

Credit Risk Predictions using Classification Modeling

Contribution Percentage of Each Group Member: 33.33%

Prof. Lulu Kang

I. Introduction

The motivation behind this paper and project is centered around the topic of credit risk analysis in the context of banks approving loans. Since its inception, the institution of banking has had many problems with discriminatory practices when it comes to issuing loans and credit to ethnic minorities. Banks in America were notoriously known in the past for its practice of redlining which relegated people of color and immigrants to certain neighborhoods in cities. The practice made it as to where people who lived in sections of the city that were labelled red, which stood for 'hazardous' were very high risk and mortgage lenders would refuse to issue loans to this demographic. This one of the many forms of discrimination that banks were allowed to do. While today this and many other forms of historical discrimination were outlawed, discrimination when it comes to issuing loans is still a prevalent problem. A year long study, conducted by the Center For Investigative Reporting found that in 2019, analyzed more than 2 million mortgage loan applicants found that people of color were more likely to be denied home loans compared to their white counterparts [1]. This was true even when you control for level of income, debt ratio and other factors. The only difference was the race of the applicant. The goal of this project is to analysis and train models for prediction on a public credit risk dataset to determine what are important features that are grounds for determining whether or not someone is likely to default on the loan and can certain feature engineering techniques can aid in the accuracy of predictions.

This paper is organized as follows. In section 2, we will give an overview of the machine learning task of classification and the classification models used in our experiments. In section 3, we discuss the exploratory data analysis conducted on our training data, the methodology for running our experiments, and the challenges that were encountered. In section 4, we discuss the results from these experiments and the feature engineering techniques that were helped improved prediction accuracy. In section 5, the conclusion of our results we will be presented.

II. Data Sources

For the entirety of this project, we utilized the Credit Risk Dataset from Kaggle [2]. The dataset contains 12 features which describe relevant characteristics of an individual such as his age, income and history with loans.

III. Overview of Classification

Classification in machine learning is a supervised learning process where based on features of the given training dataset we make predictions on a label. Classification has applications in many different domains. For our project, we utilized 5 different classification models: 1. KNN, 2. Logistic Regression, 3. Decision Trees, 4. Random Forest, and 5. Naive Bayes to test the performance of our predictions.

A. KNN

K-Nearest Neighbors uses the category of near data point to find the category of target point. It's supervised learning so the target attribute should be categorical variable.
KNN doesn't need to use training data to build models. It predicts every data every time. Therefore, KNN is the simplest to do.

B. Logistic regression

Logistic Regression is a supervised classification algorithm that utilizes a logistic function to frame binary output. The output of the logistic regression will be a probability ($0 \leq x \leq 1$), and can be adopted to predict the binary 0 or 1 as the output (if $x<0.5$, output= 0, else output=1). Although logistic regression is a form of regression analysis, what makes logistic regression useful for binary classification tasks is introduction of this decision threshold variable.

C. Decision Trees

Decision Trees is a tree-based algorithm that classifies data from the root of the tree through inductive rules , finding the best segmentation points section by section to divide the data into small units. If the sample number is too low and the variable is too much, the ability of

classification will be worse. Otherwise, the path of the Decision Tree is fixed so it doesn't have fault tolerance.

D. Random Forest

Random Forest combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. To be brief, Random Forest uses random sampling to get many Decision Trees and sets of the trees is forest. It is more accurate.

E. Gaussian Naive Bayes

Naive Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.

IV. Methodology and Experiment Setup

A) Data

The dataset contains 32581 records. The loan_status variable is the predictor column(dependent feature). It consists of 0 and 1. The 0 indicating the credit risk free and 1 indicating the existence of credit risk. The data set clearly indicates it to be a classification Machine Learning task. So we decided to go out with classification modeling techniques such as Logistic Regression, Naive Bayes, Decision tree, Random Forest, K Nearest neighbor. The data set contains a huge number of records. There are 21 features after using encoders.Therefore it is wise to split the train test into 75:25 and n>10,000. The reason 5 models we are choosing is because of the fact that we intend to bring out the best accuracy, but making sure that the bias and variance remains balanced.

B) Exploratory Data Analysis

The implementation part starts with the Exploratory Data Analysis(EDA), different Visualization, feature Engineering, Ensembling methods, modeling, hyperparameter tuning. The goal is to produce a low bias and low variance model in the long run. Therefore trail and testing with sampling, tuning, feature selection methods have been carried out consistently.
The Visualization techniques were used initially as the data set was large to find the patterns, understand the data, extract the hidden information.
As per EDA is concerned it started with data cleaning such as finding null values, Nan etc. Zero value was checked as we cannot have a loan amount , person age values zero. Pandas profiling was used to get the full fledged report.
Visualization consisted of a distribution plot to check the skewness(Appendix 1.), box plot was used to remove any outliers(Appendix 2.), pair plots were used to check the relationship of each independent variable with the dependent variable. Heatmap(Appendix 3) was used to check the correlation values. Histogram was used to see if there were any deviating values. Principal component analysis(Appendix 4.) was used to see if there was any possibility of dimensionality reduction.

C) Modeling

Modeling initially began with Logistic Regression then Decision tree was used. Later random forest, thereafter Knn and naive bayes. Hyperparameter tuning was to improve the accuracy thereby finding the best parameters.

The experiment setup: We divided the process of building the model into 3 parts. EDA, Feature engineering techniques and modeling. Version control was used to work in a collaborative environment. Jupyter notebook was used to build the model.

V. Analysis and Results

Through visualization we could see that only a couple of features displayed a linear relationship. Majority of them were non linear. By thinking of this it was evident that decision Trees would do a better job in classification. The heatmap and pairplot showed similar results. After using PCA we could see a decrease in the value of the accuracy. This shows that all features carry equal are important in predicting the laon status. The loan grades were graded A,B,C which indicates A>B>C. Therefore label encoding was used to maintain the weights. For the features loan-intent, loan-owndership and cb_person_default on file one hot encoding was used.

Variance inflation factor values not greater than 6. As 6-10 requires further investigation and above 10 has serious multicollinearity. Logistic Regression produced the output aorunf 85 percent with True positives 5387 and True Negative 822. Ensemble technique such as bagging(sampling with replacement) using logistic regression produced an accuracy more over same. One of the reason could be due to absence of mulitcollinearity and the model seemed to have absence of overfitting problem as training and test results were having minor difference in their accuracy score. Knn classifier produced an accuracy of 89 percent. Knn internally implements any of the methods such as' ball_tree', 'kd_tree', 'brute'. Hyperparamter tunning was used to get the best features. Using GridSearch Cross validation ball tree was prompted. This was expected as records were large brute would be very computationally expensive.

The Gausian naive bayes classifier produced least accuracy among all with 81 percent of accuracy. One of the strongest reason could be due to imbalance class ratio of loan_status. It has 22603 zero's and 6327 one's. Naive bayes is sensitive to imbalance class labels. Zero frequency problem may be the biggest reasons.

The random forest which is aggregated sampled ensemble groups of decision tress does best with accuracy of 93 percent. Random forest deals well with these imbalanced class ratios.

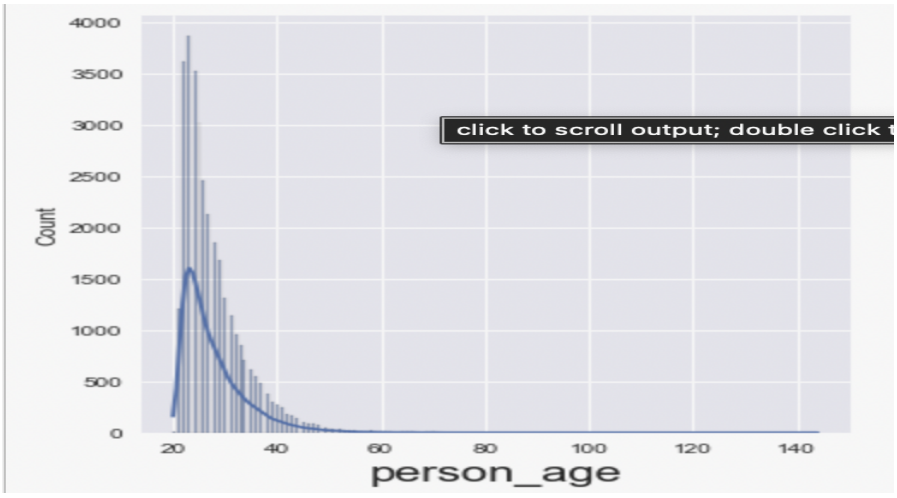
K-Fold cross validation was used on gaussian naive bayes and thus the accuracy was increased by five percent(new accuracy: 86 percent).

VI. Conclusion

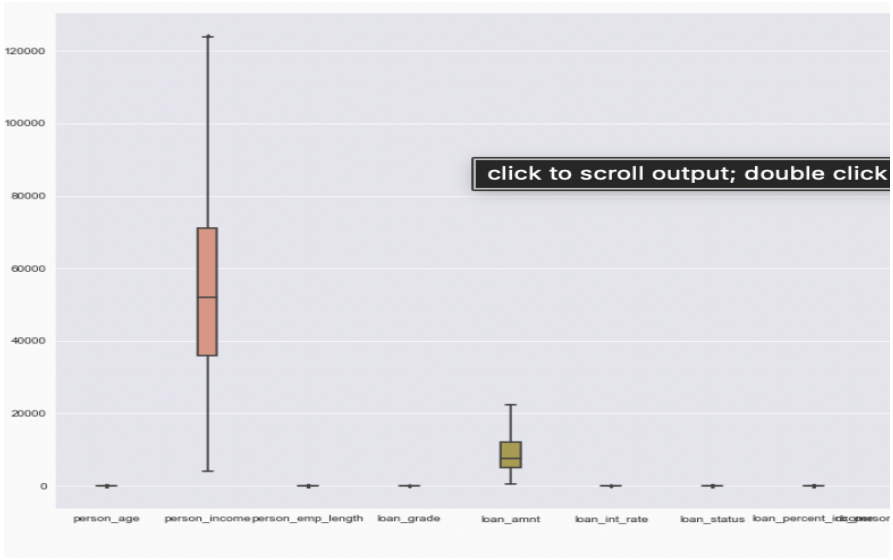
Random forest is a better classification model when the dataset is large and has imbalance class ratio in dependent column. The low performing models can be improved by cross validation, hyperparameter tunning, ensemble techniques and so on. Only after understanding the data, patterns, domain knowledge it is better to implement Machine Learning algorithms.

VII. Appendix

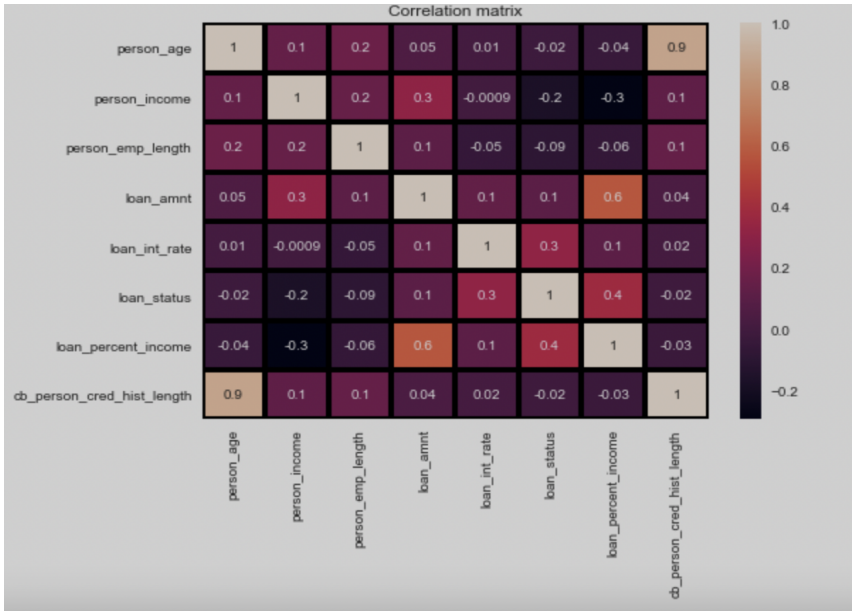
Appendix 1:



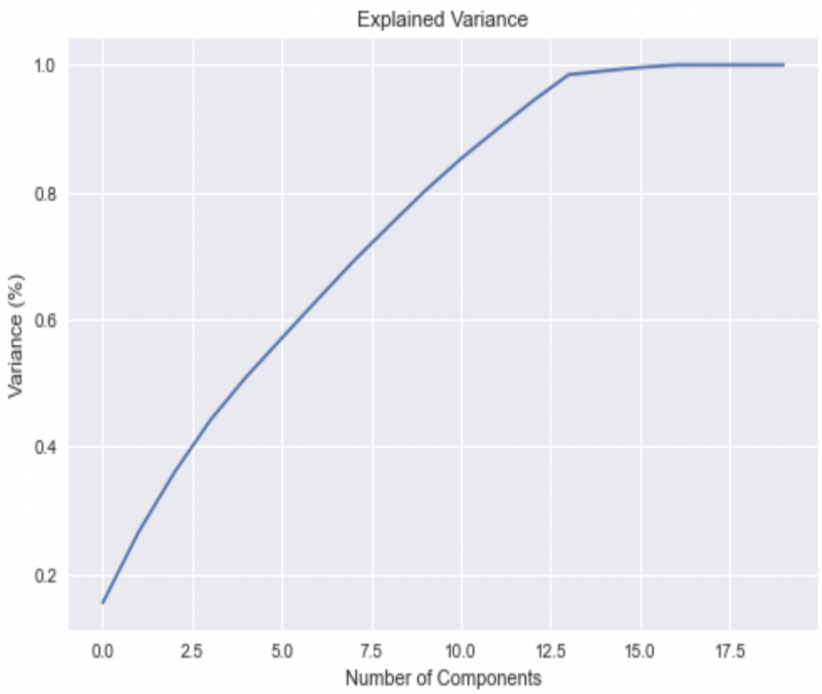
Appendix 2



Appendix 3



Appendix 4


Explained Variance
1.0
0.8
0.6
0.4
0.2
Variance (%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Components

VIII. References

[1] Aaron Glantz and Emmanuel Martinez. (2021, June 30). *Modern-day redlining: Banks discriminate in lending*. Reveal. Retrieved May 1, 2022, from https://revealnews.org/article/for-people-of-color-banks-are-shutting-the-door-to-homeownership/

[2] Tse, L. (2020, June 2). *Credit risk dataset*. Kaggle. Retrieved May 1, 2022, from https://www.kaggle.com/datasets/laotse/credit-risk-dataset

[3] Soner Yildirim (2020,Feb 11). *Decision Trees and Random Forests — Explained*. Retrieved May 1, 2022 from https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd

[4] Rohan Vats (2021,Feb 22) *Gaussian Naive Bayes: What You Need to Know?* Retrieved May 1, 2022 from https://www.upgrad.com/blog/gaussian-naive-bayes/