CSE 343: Machine Learning Project Report

Credit Card Default Prediction

Devansh Arora	Devraj Sharma	Aayush Kapoor	Ishita Sindhwani
2020053	2020054	2020007	2020305

Abstract

Credit cards are crucial financial instruments in today's world economy. Banks give customers money on credit with the hope that they will pay it back. But if customers default on the credit amount it leads to huge losses for the issuing bank. Thus, predicting the default rate of customers for a bank is instrumental for avoiding risk associated with issuing credit cards. This project aims to assess various machine learning algorithms like Logistic Regression, Support Vector Machines(SVMs), Random Forests and Artificial Neural Networks in predicting whether a customer would default or not.

1. Motivation

In today's life online payments are a boon to people's life, but with all the digitalization arises problems that have to be dealt with. One such problem is how do card issuers know whether the customer will pay the credit? That's a complex problem with many existing solutions—and even more potential improvements. Credit default prediction is central to managing risk in a consumer lending business. This allows lenders to optimize lending decisions, which leads to sound business economics experience

2. Problem Statement

The use of machine learning in fraud detection has been an interesting topic in current times. There are many types of frauds that are present in the world, but we decided to look at a fraud that has been overlooked by many i.e. credit card default. Billions of dollars have been lost in this default, but with the use of machine learning algorithms, if we can increase the accuracy of predicting defaults by even 2-4% from the existing model's accuracy it can be instrumental in increasing profits manifold. Thus our goal is to create the best machine learning model we can for the given dataset and try to reduce the credit default rate.Machine learning is divided into 3 categories and we will use supervised machine learning to deal with this issue. In supervised learning we will apply different algorithms and pick the one which is best suited in terms of accuracy.

3.Literature Survey

>>*Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, Kuiyi Liu, "A study on predicting loan default based on the random forest algorithm"* (*https://www.sciencedirect.com/science/article/pii/S1877050919320277*)

Lin Zhu et al. uses a random forest classifier to predict loan default rate on data provided by Lending Club, an online lending platform. The dataset consisted of 115,000 rows of customer data and 102 attributes. After feature selection, 15 attributes were selected for training. SMOTE(Synthetic Minority Oversampling Technique) was used to deal with the problem of class imbalance. On comparing with various other ML algorithms, random forest turned out to be the best.

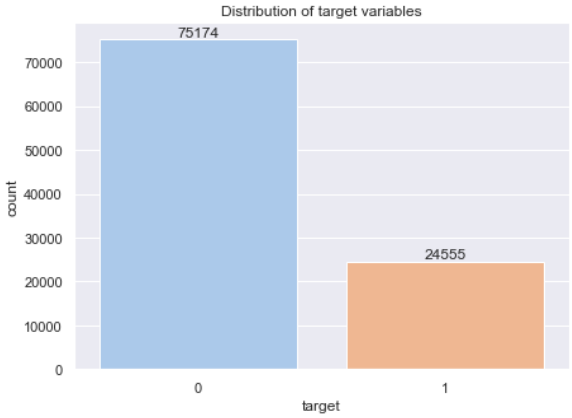
>>*Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran "Credit Card Default Prediction using Machine Learning Techniques"* (*https://ieeexplore.ieee.org/document/8776802*)

Y. Sayjadah et al. tries to predict default rates for credit card users. It uses bank data of customers with 30,000 instances and 24 attributes. Correlation based Feature Selection techniques were used to reduce the dimensionality. Finally, logistic regression, rpart decision tree and random forest were trained on the data. Random forest emerged as the best method.

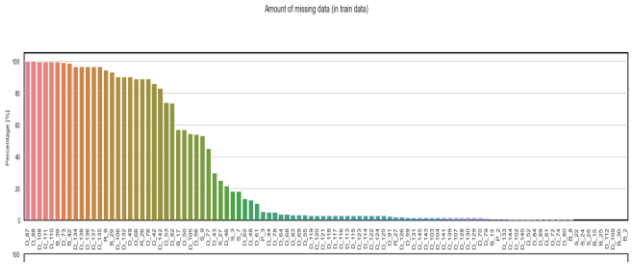
4.Dataset Description and visualization

The Dataset consists of 191 columns and 99730 rows.
The dataset contains aggregated profile features for various customers for each statement date. The features fall into 5 categories namely
D: Delinquency variables
B: Balance variables
S: Spend variables
P: Payment variables
R: Risk variables
Most of the columns contain continuous data and only 11 features are categorical.

For each customer a target value corresponding to whether they defaulted or not is given. The value is 0 for non defaulters and 1 for defaulters and we need to predict the probability of a future payment default.
We observe that the dataset is imbalanced as 75 percent of it is for good customers.



All variable types contain missing values and outliers. The percentage of missing values for the columns in descending order has been attached.



Only 11 features are categorical namely 'D_120','B_30', 'B_38', 'D_114', 'D_116', 'D_117','D_126', 'D_63', 'D_64', 'D_66', 'D_68'

4.1Analysis for categories of features

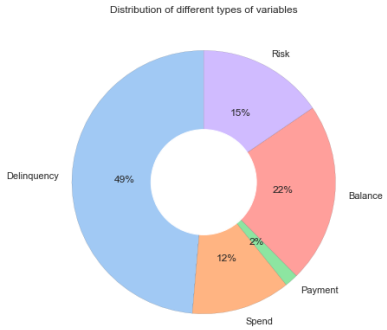
Delinquency Variables:
Total:92
Number of variables with dtype float64: 90
Number of variables with dtype object: 2
Number of variables with categorical data:9
Number of variables with missing values:81

Spend Variables:
Total:22
Number of variables with dtype float64: 21
Number of variables with dtype object: 1
Number of variables with categorical data:0
Number of variables with missing values:13

Risk Variables:
Total:28
Number of variables with dtype float64: 28
Number of variables with dtype object: 0
Number of variables with categorical data:0
Number of variables with missing values:4

Payment Variables:
Total:3
Number of variables with dtype float64: 3
Number of variables with dtype object: 0
Number of variables with categorical data:0
Number of variables with missing values:2

Balance Variables:
Total:40
Number of variables with dtype float64: 40
Number of variables with dtype object: 40
Number of variables with categorical data:2
Number of variables with missing values:19



This figure shows the distribution of columns(features) into those of various categories

4.2 Dataset analysis

The values for data in most columns lie in the range of 0 to 1. Missing values are present in most columns as shown previously. The data is skewed and contains outliers. The following plots are done for columns with highest information value from each category.

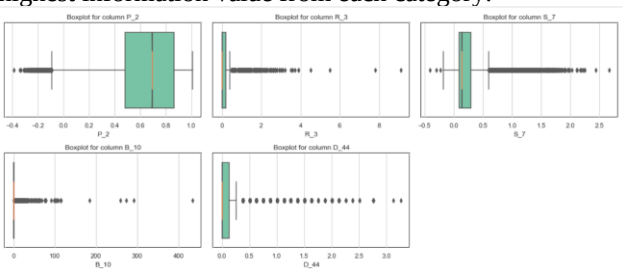


Figure 4

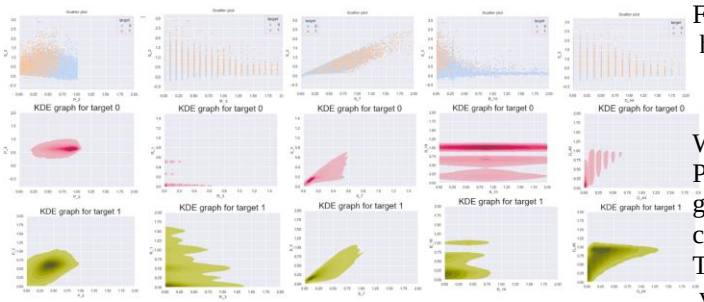


Figure 5

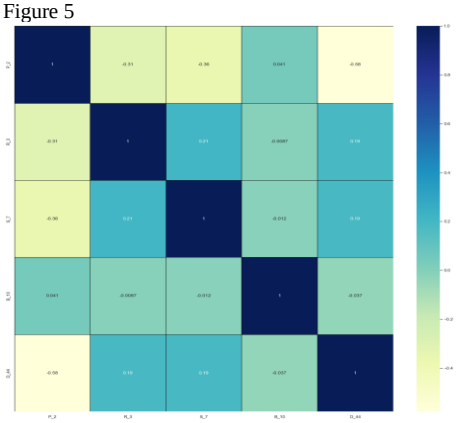


Figure 6

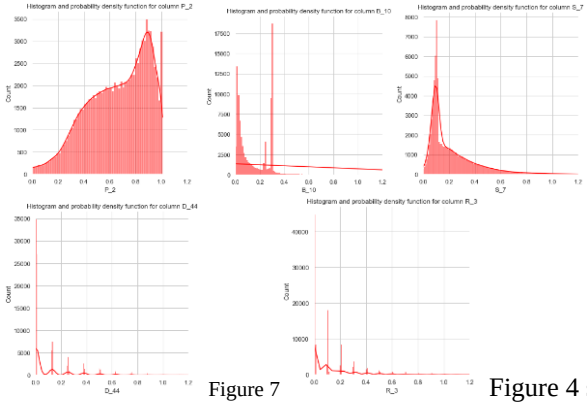


Figure 7

Figure 4 shows the boxplots for them and we can observe that our data mostly lies somewhere between 0 and 1 and many outliers are present. The scatter and kde plots in Figure 5 show the correlation between top 2 Information value columns from each category Except category B, the other category graphs are not correlated. We can also observe in the heatmap that the columns don't have much correlation. The pdf plots show that the data has a skewed distribution and confirms the presence of outliers.

>>Requirement for Pre-processing:

The dataset available is filled with a large number of inconsistencies, missing values, noise which leads to erroneous prediction. Before the model is trained on the actual dataset the a dataset needs to be pre-processed. This includes removing noise, null values and non-important features.

5.Methodology:

For the data-pre-processing WOE transformation technique has been used. The formula is as follows:

$$WOE = ln\left(\frac{percentage\ of\ events}{percentage\ of\ non\ events}\right).$$

Percentage of good events for a category is the number of good events that happened in that value of the attribute compared to the total number of good events.

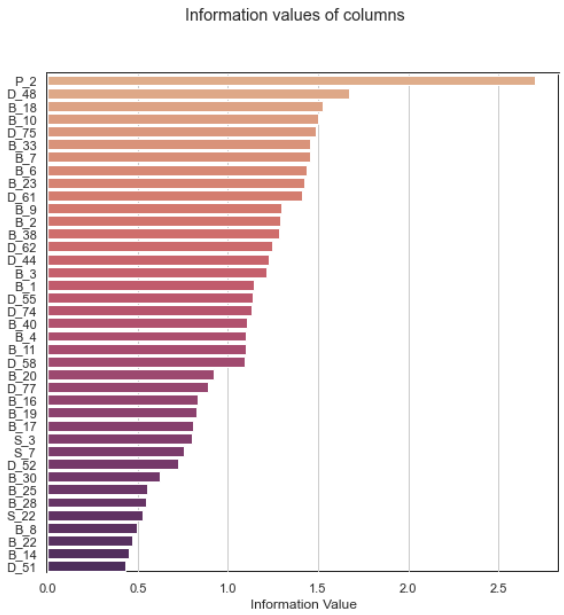
The attribute is then selected based on the information value

that is present in the feature which is calculated using the following formula:

IV=WOE*(Percentage of events-Percentage of Non-Events).

The feature having an IV>0.1 is considered as a good predictor rest all are discarded.

With the help of this transformation all the values are replaced with their WOE values and null values are also removed. Outliers in the values are also separated.

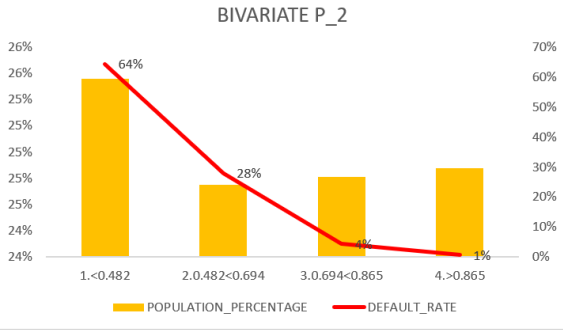


6.Result and Analysis

The columns that were finally chosen after this process are:

[P_2, P_3, P_4, S_3, S_5, S_7, S_8, S_15, S_22, S_24, S_25, S_27, R_1, R_3, R_27, B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8, B_9, B_10, B_11, B_12, B_13, B_14, B_16, B_17, B_18, B_19, B_20, B_22, B_23, B_25, B_28, B_30, B_33, B_37, B_38, B_40, D_106, D_112, D_113 , D_114, D_115, D_117, D_118, D_119, D_120, D_121, D_122, D_128, D_130, D_132, D_134, D_135, D_136, D_137, D_138, D_39, D_41, D_43, D_44, D_46, D_48, D_50, D_51,

D_52, D_55, D_56, D_58, D_59, D_60, D_61, D_62, D_70, D_71, D_74, D_75, D_77, D_64]



After the application of WOE transformation 84 features were finally chosen for final model fitting.
Each feature was analysed using a bivariate graph. From the graph inferences could be drawn basis the default rate in each population segment of the feature.
We ran 3 basic models to get an idea of how our dataset was responding. The results obtained have been attached below

```
Decision Tree 0.8474213710351282
Random Forest 0.864300277415689
Logistic Regression  0.8626625221431198
```

We can observe that random forest performed best with an accuracy of 86.4%. We have various other algorithms and techniques to be performed which we will showcase during final report.

7.Conclusions:
Learnings:
We learnt about various Explanatory data analysis techniques including various types of plots for data visualization and data analysis. We learnt how to read those graphs and draw conclusions about distribution, range and correlation of various features. We also learnt how to identify and deal with missing values.
Our dataset was huge and had 190 columns. We learnt how to identify their importance and do feature selection based on that.
We also learnt various pre processing techniques and used WOE transformation because it best suited our dataset.

Work Left:
Data training, testing validation and analysis is left. We need to explore many algorithms on our dataset and make changes to increase the accuracy as much as possible.

Individual contribution:
Ishita: Data description and visualization
Devansh: Data preprocessing and feature selection
Devraj:Data pre processing and feature selection
Aayush: Data pre processing and feature selection