Task 2
Critical visual analysis

VISUALISATION PRINCIPLES AND VISUAL ANALYTICS

FABIO PALLIPARAMBIL

SUBMITTED ON: 15/01/2021

Table of Contents

1.0 Introduction

The paper was published by Wiseley Wu who is a student at the University of California, Berkeley on 11th MAY 2018 as his Final Project(2). The report topic is **Prediction of Loan Default with Machine Learning Method**. The dataset used is taken from Kaggle and the data is from a financial company based in America known as **LendingClub**.

2.0 Part1 – Evaluate an existing visual analytic approach for the problem

The issue here is that the prediction of the loan is not done correctly, because the model has been rejecting most of the customer which means this potentially can lead the finance company to lose its customers. The approach taken to find the problem was to implement multiple machine learning algorithms such as Logistic regression, SVM(Linear Kernel), SVM(RBF Kernel), Gradient Boosted Trees and a basic Neural network to get the best predicting model. But it all starts with a base machine learning model known as the Naïve Bayes Algorithm and this was the model used by the **LendingClub**.

- What kind of data set it used?
    - The dataset used is from Kaggle as mentioned in Introduction. It is in a CSV format and the data is from a period 2007 to 2018. The data set is split into Accepted.csv and rejected.csv. It has 2260701 instances and 151 features(acceptance.csv) and 27648741 instance and 9 features(Rejection.csv).
- What type of visualisation does it use?
    - It uses visualisation like pie chart for a breakdown of a Loan status, Correlation matrix between features, Line chart for evaluating the model, a bar graph for top feature selection, heat map for hyperparameters and multiple histograms for feature evaluation.
- What data processing and/or transformation it utilises?
    - It deals with missing values by deleting the feature if it has 100000 rows of NaN in a feature. It had removes feature with negative values of debt-to-income ratio. It has removed all unrelated and unusual data to reduce the dimensionality. It uses extracts feature by plotting the coefficient so it can have the best features possible as it a big dataset.
- What data analysis method it utilises?
    - It looks at the data and carries out all the data pre-processing to make sure the data is ready to be trained. The Feature has been picked carefully, it helps to increase the performance of the models. It analyses different model accuracy though line graphs and bars charts.
- What is the analytical reasoning outcome of this approach on this specific data?
    - According to the reasoning the neural network and Gradient boosted Tree perform the best in the bunch. And It shows the pre-processing has a big role in the success of the models.

3.0 Part 2: Suggest your solution

Conduct an Experiment with the same dataset as above paper and visualize the data in depth and get some insights on the dataset. As each dataset has 2Million instances, it is very computationally intensive. Therefore, I cut it down to 10000 for usability and to understand the data better. Figure 1 is used to find whether the loan amount has been funded, also finding which term got the highest funds.

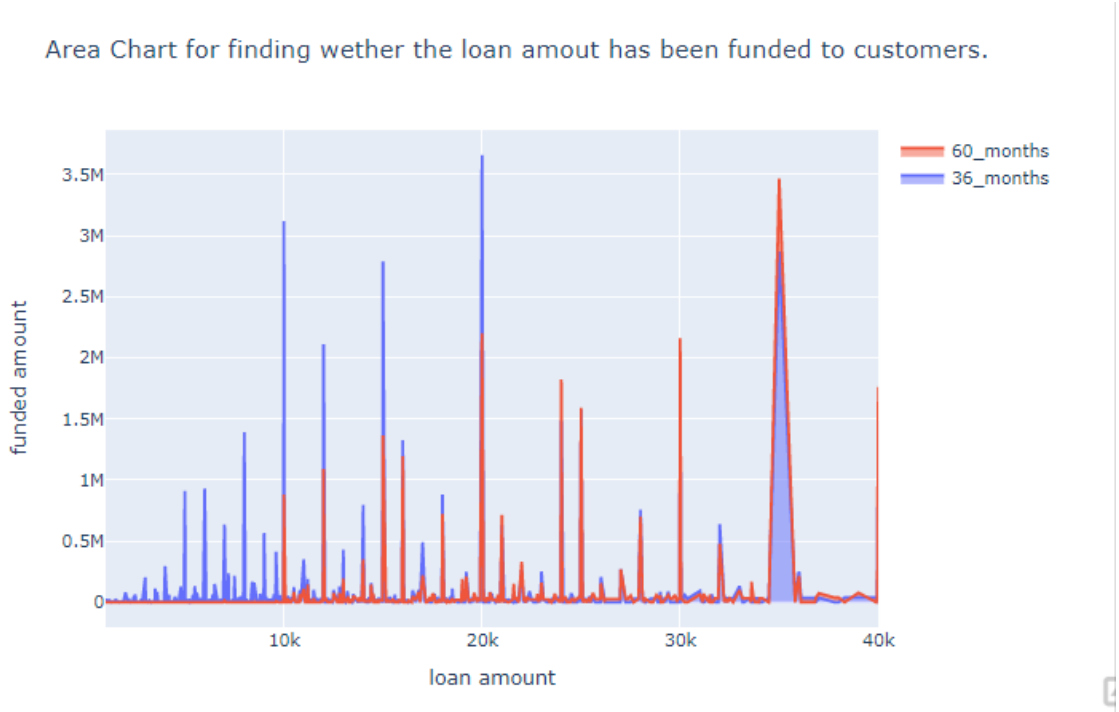

Figure 1 Area Chart (Loan Accepted dataset)

Scatter plot is used to find the interest rate charged for the different loan amounts, relating their grades. And this will plot will give an idea of how much would a customer pay interest according to their grade for the specific loan amount.

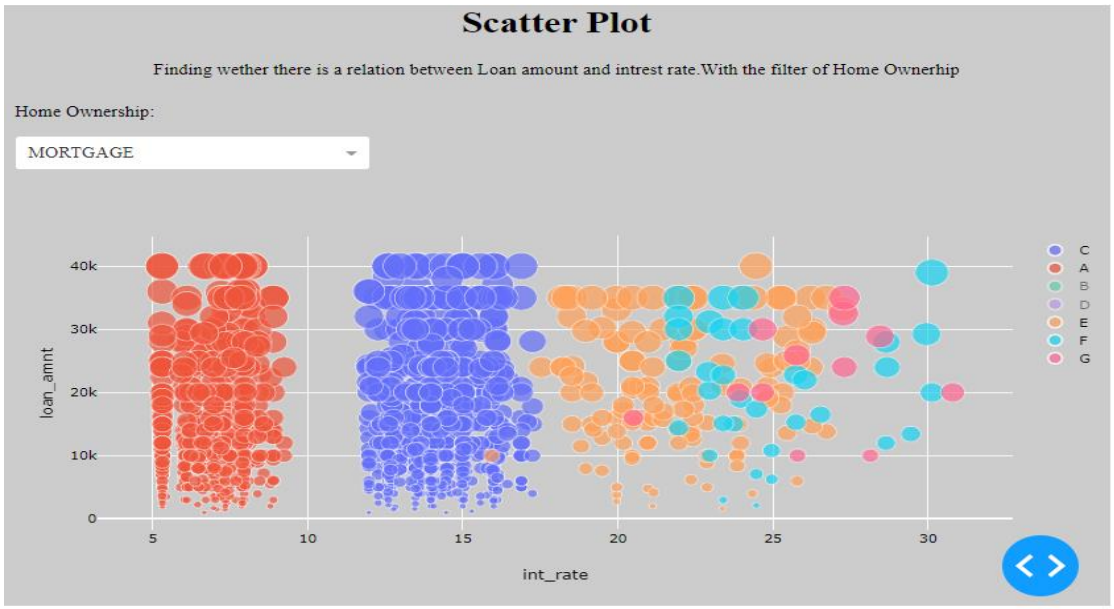

Figure 2 Scatter Plot ((Loan Accepted dataset)

The histogram is plotted to understand different application dates and the amount requested by the applicants. Also, you can understand whoever had less than 1 year of employment length has been rejected for the loan(figure 3).

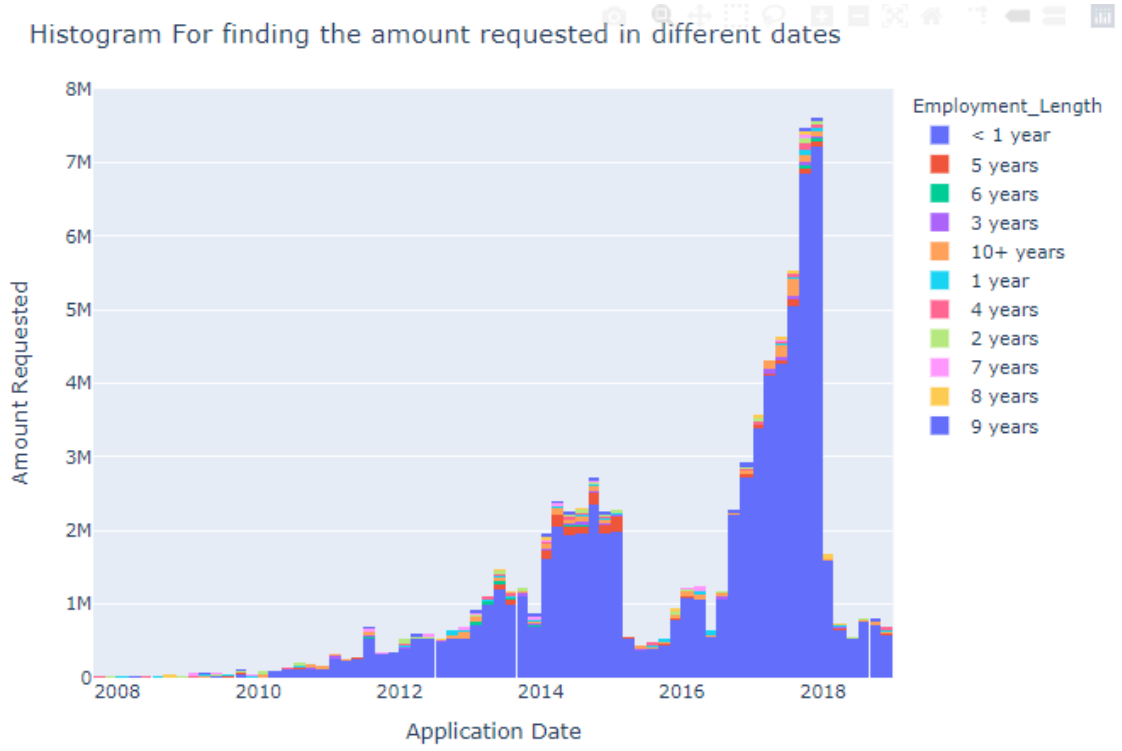

Figure 3 Histogram (Loan Rejected dataset)

The main reason for the pie chart is that feature is flexible, so you can find out which value in a feature has been rejected for the loan approval.

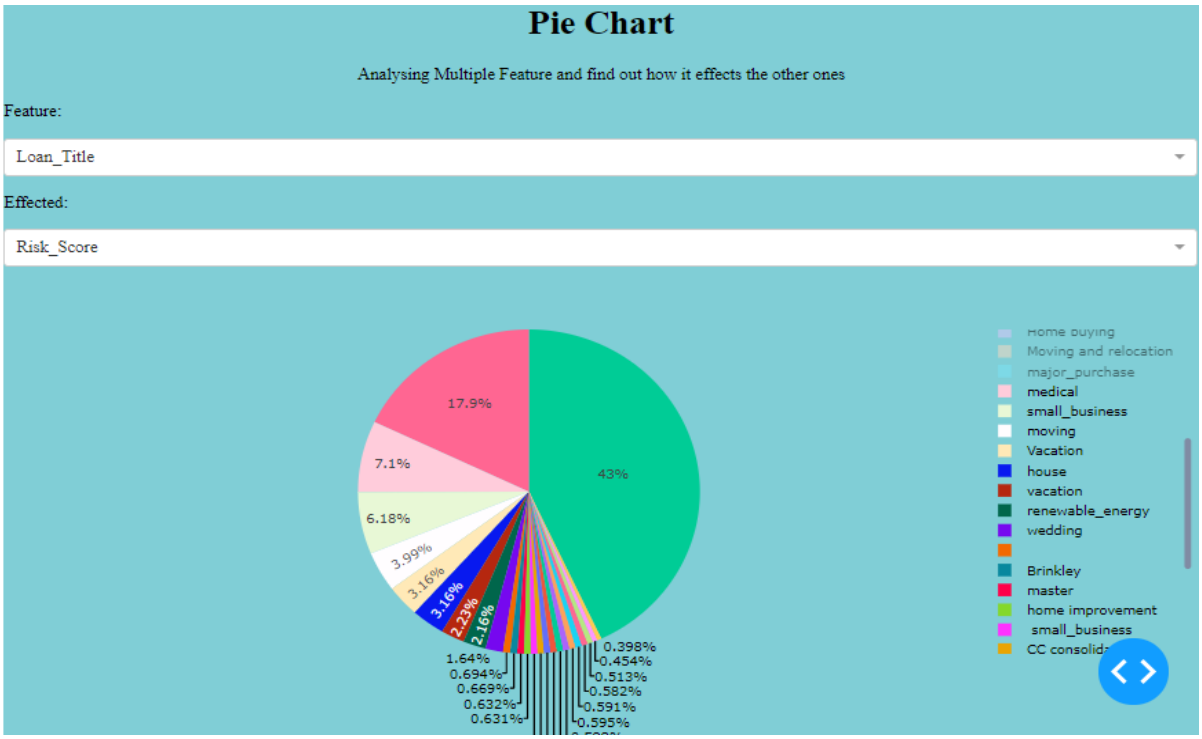

Figure 4 Pie Chart (Loan Rejected Dataset

4.0 Conclusion

The above visualisation has been done to visualisation the acceptance and rejection of loan requests. I have divided the **figures 1,2 for acceptance of loan request and figure 3,4 for rejection of loan request.** According to the visualisation you can see if you have recent employment(<1 YRS) you will be mostly to be rejected for the loan request. Also, your grade plays a big role in the interested amount offered from financial companies for the loan amount.

5.0 Reference

1. Kaggle.com, All lending Club Loan Data(2007-2018),Available at [Online] : https://www.kaggle.com/wordsforthewise/lending-club
2. Wiseleywu.com, Loan Report, Available [online at] : https://www.wiseleywu.com/loan_default/final_report.pdf