International Institute of Information Technology, Bangalore

AI511-Machine Learning

Project Report

Prepared By: -

Name: - Syed Hashir Mahmood, Vipul Manohar Ahuja

Roll Number: - MT2022122, MT2022132

Degree: - Master of Technology

Branch: - Computer Science and Engineering

Year: - 1st Year

Semester: - 1st Semester

Guided by Prof. Dinesh Babu and Prof. Neelam Sinha

Assisted by Debmalya Sen

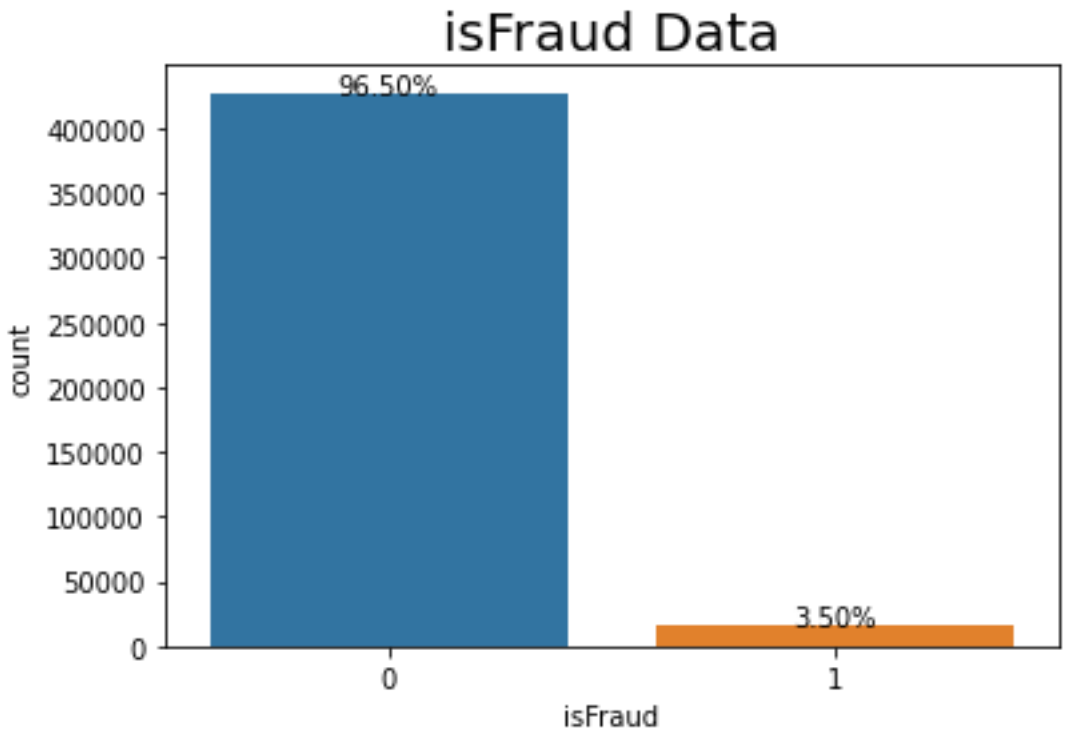
Abstract

1) The data consist of two excel file train.csv (on which we trained our model) and test.csv (on which we predict our result) and then we upload the predicted result on Kaggle. We created two ipynb files. In first file we uploaded train.csv into data frame.
2) In train.csv, initially there were 434 columns along with target, so we removed the columns which has more than 40% null values in it.
3) Now, we replaced the unique values of "P_emaildomain" (61 in count) with the mapped function created by us to reduce the count from 61 to 35. [Example yahoo.com and yahoo.co.uk is treated as yahoo]
4) Then, we replace the null values of numerical features with mean and null values of categorical values with mode.
5) Then, we apply corr () method to check the correlation among all the remaining columns and again we removed the columns which has absolute correlation value more than 0.90.
6) After, this we remaining with total 104 columns (including 5 categorical features)
7) Then, we downloaded the preprocessed train file in csv format.

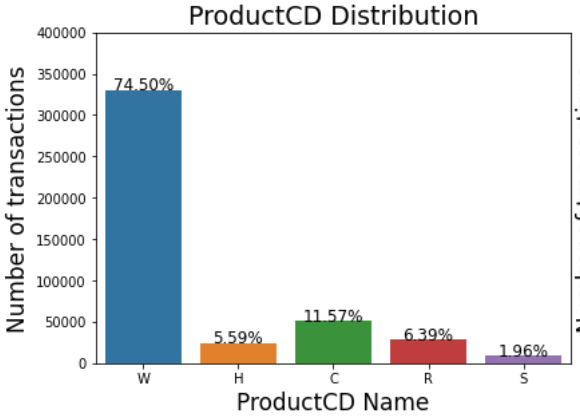
In the second file:

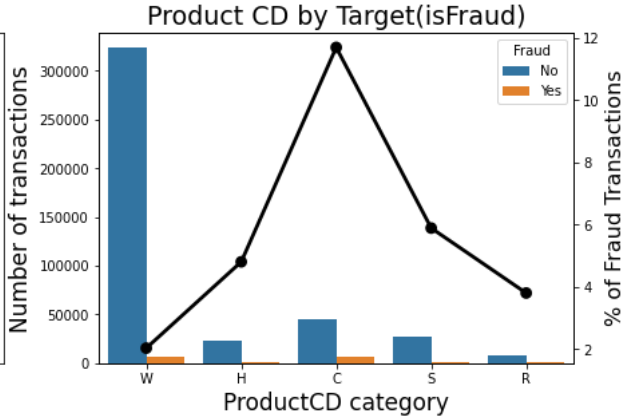
1) Firstly, we imported the test file and the preprocessed train file.
2) Then, we extract the columns name from preprocessed train file and then use the same column for test.csv
3) Now, we applied One Hot Encoding on categorical features of preprocessed train file and the categorical features of the new test file which we used the columns of preprocessed train file(without OHE).
4) Now, the column increased from 104 to 148.
5) Then we applied different models on it.

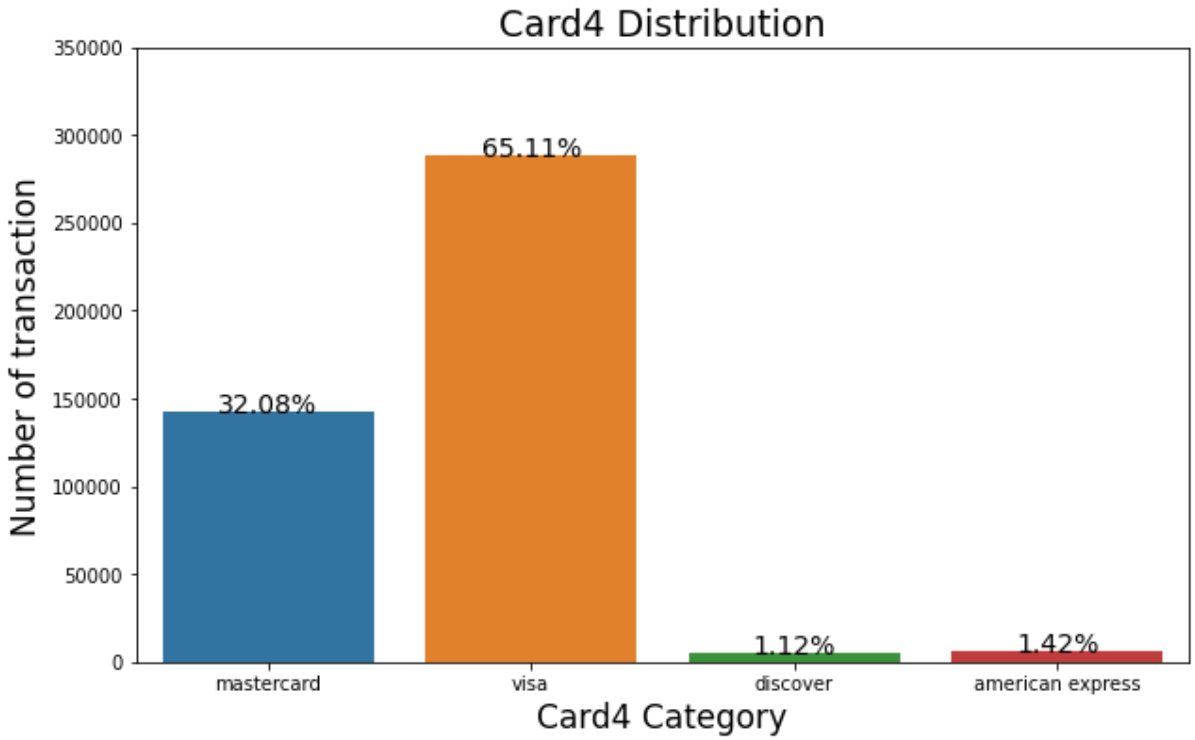
Graph Plots

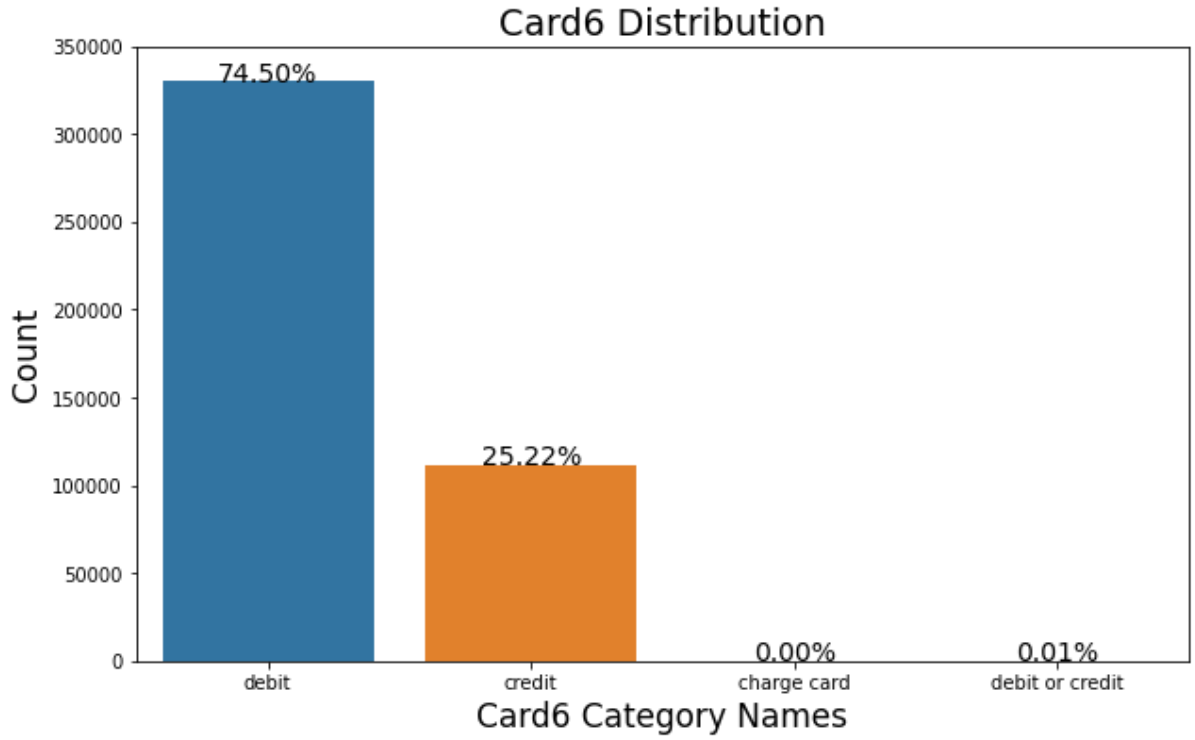


ProductCD Distributions






Card4 Distribution
Number of transaction
350000
300000
250000
200000
150000
100000
50000
0
32.08%
65.11%
1.12%
1.42%
mastercard
visa
discover
american express
Card4 Category


Card6 Distribution
Count
350000
300000
250000
200000
150000
100000
50000
0
74.50%
25.22%
0.00%
0.01%
debit
credit
charge card
debit or credit
Card6 Category Names

1) First Image is the Bar Plot of Target Variable “IsFraud”
2) Second Image is the Bar plot of the “Product_CD” and comparison of “Product_CD” and “Target”.
3) Third Image is the Bar Plot distribution of Card4.
4) Fourth Image is the Bar Plot distribution of Card6.

Results of Different Models

Sr No	Different Models	Result as per Kaggle before Competition End
1	Logistic Regression without Hyper Parameter	0.56
2	Logistic Regression with Hyper Parameter	0.52
3	KNN Classifier without Hyper Parameter	0.63
4	KNN Classifier with Hyper Parameter	0.59
5	XGBoost without Hyper Parameter	0.81
6	XGBoost with Hyper Parameter	0.84
7	SVM without Hyper Parameter	0.83
8	SVM with Hyper Parameter	0.84
9	Random Forest without Hyper Parameter	0.864
10	Random Forest with Hyper Parameter	0.869
11	Light Gradient Boost without Hyper Parameter	0.85
12	Light Gradient Boost with Hyper Parameter	0.89

Observations and Conclusions

1) Our Best prediction was 0.89 before contest end.
2) Generally, XGBoost gives better result than Random Forest, but in our case, Random Forest has given better result.
3) Surprisingly, the result after applying Neural Network was extremely bad(50.5%), Hence we did not include it.