CS229 - INNOVATION LAB REPORT

CREDIT CARD FRAUD DETECTION REPORT

Prepared by

GAURAV KUMAR
1701CS21
KANAV GHAI
1701CS24

Project Aim

The aim of this project was to use some specific set of machine learning algorithms and selecting the best according to the dataset to classify the transactions as fraud or non fraud

What was the inspiration

The health of the credit card industry is best measured not by the number of people with cards, but rather by the number of credit card holders who pay their bills.

Hence, it is very necessary for credit card company to have such a algorithm so that it can detect if its card holders are going to do pay back or not by going through their past transactions and using algorithm to find whether or not the probability of a transaction being fraud is there or not.

Project Scope

The scope of the project is somewhat restricted by the missing sensitive data. The PCA data might be useful for prediction but the high dimensional credit history data is really necessary in order to produce quality visualizations which are missing in the selected dataset. The customer name, a date with an exact time of transfer, location, device used for money transfer are some of the crucial information necessary for clustering/segmenting the data and the number of defaults is very low and the unbalanced data might have some impact in the prediction and insights extracted.

Dataset description

Dataset was taken from kaggle website. It contains only numerical input variables which are the result of a PCA transformation. Due to confidentiality issues, they cannot provide the original features and more background information about the data. Features V1, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise

Data limitations

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

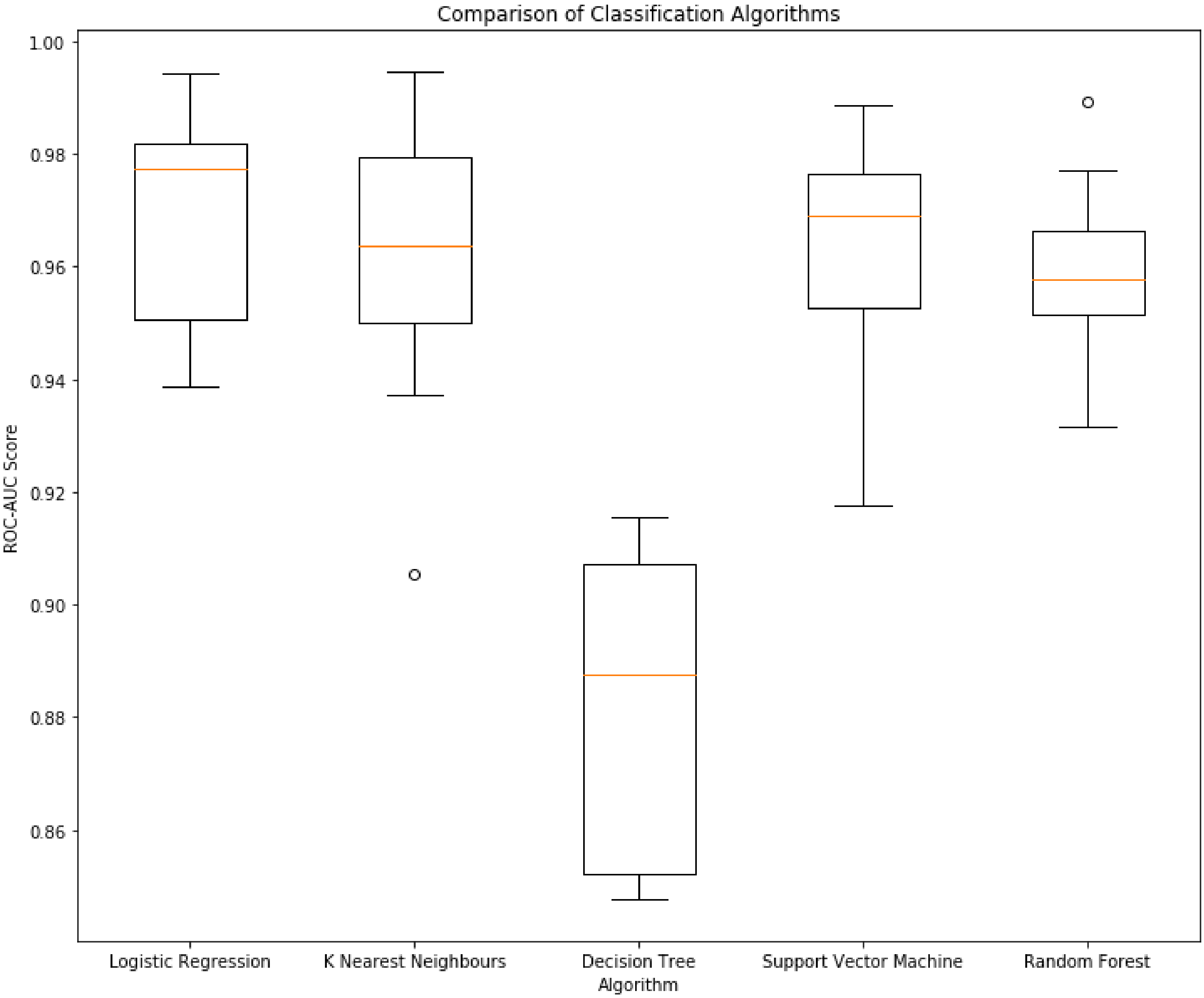
Novelty of the Project

Integration of features like solving the problem of imbalance dataset, Using Outlier removal technique to improve the results, plotting relationship between different variables to see which variables are highly correlated for a transaction to be classified as a fraud, testing the dataset with different algorithms to select which one would work best depending on the dataset given.

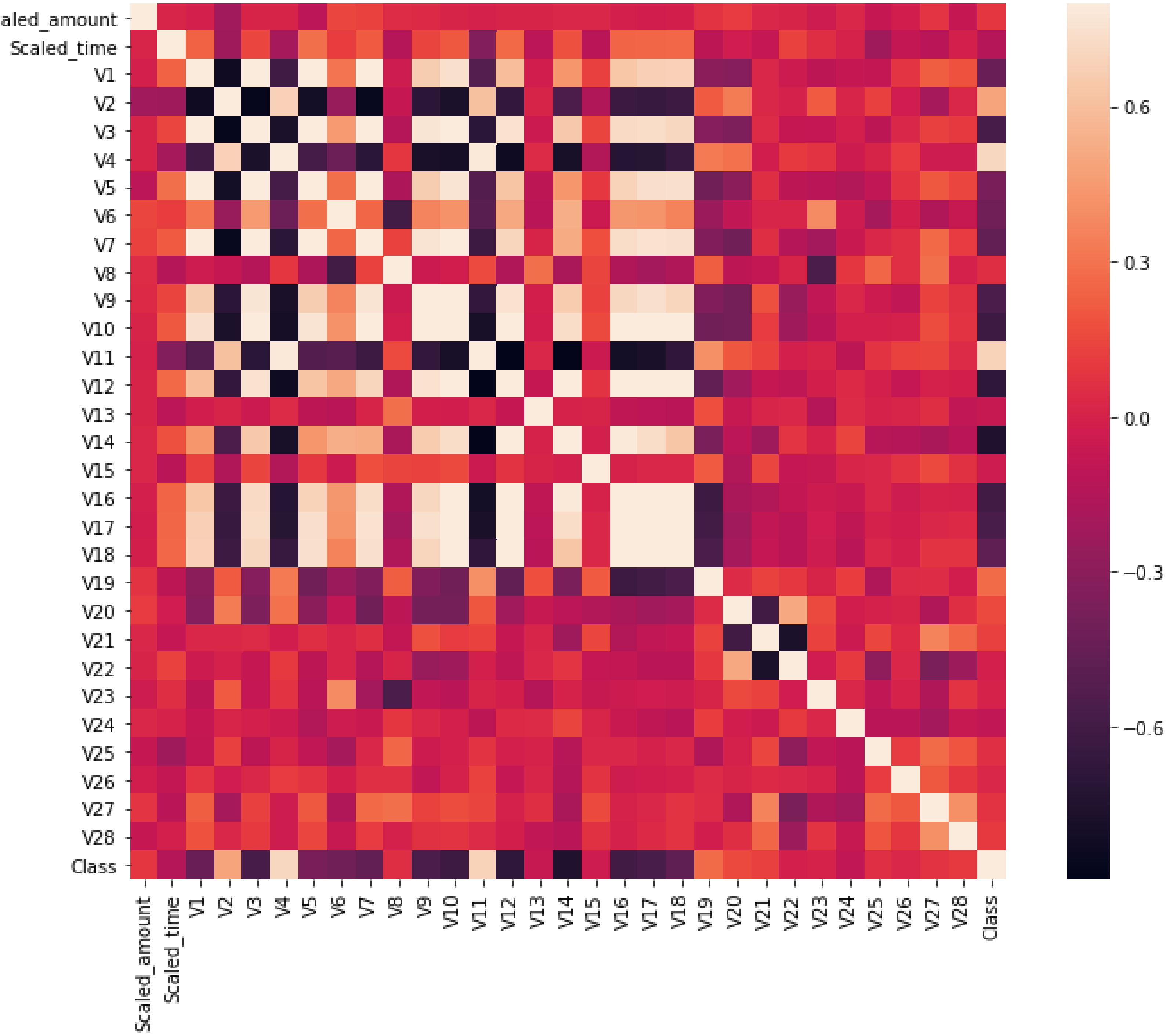
Project Description

- Exploratory Data analysis
- Fixing Data Imbalance problem
- Exploring the Subsample and Outlier Removal
- Dimensionality reduction using t-SNE
- Choosing the best algorithm to work on our dataset
- Results on test data using the best chosen algorithm

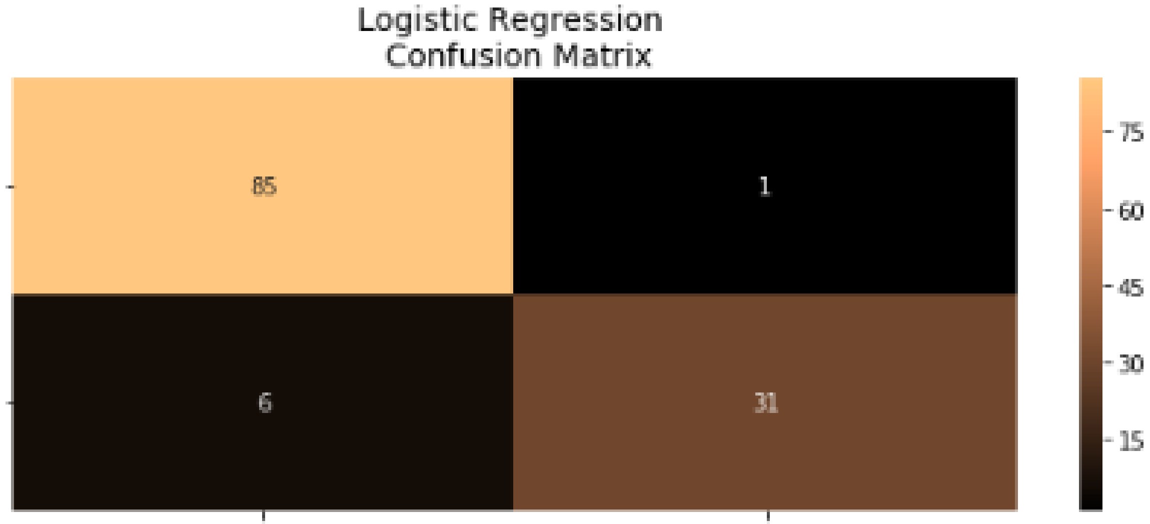
Results obtained on selecting the best model

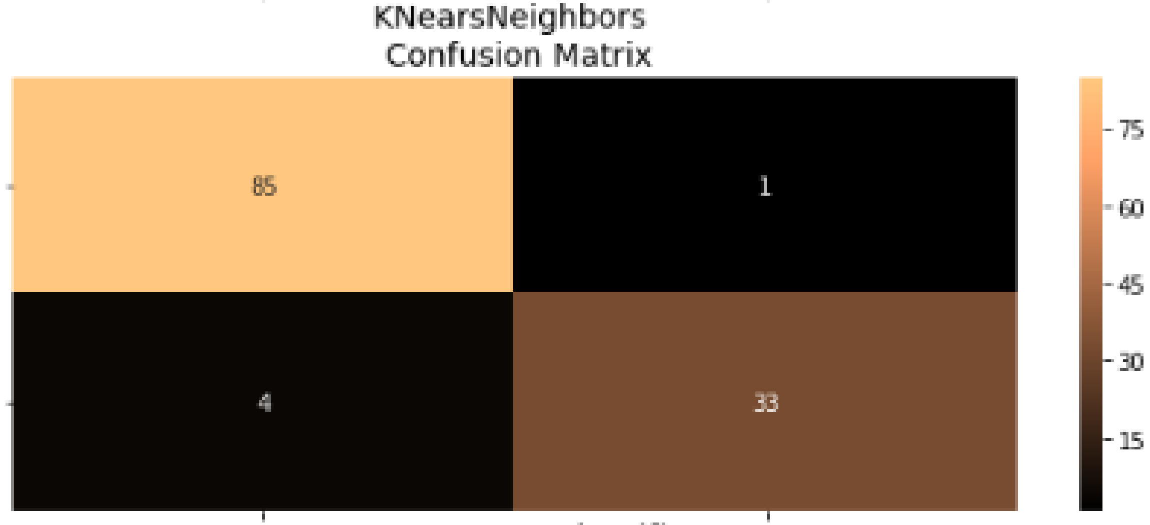


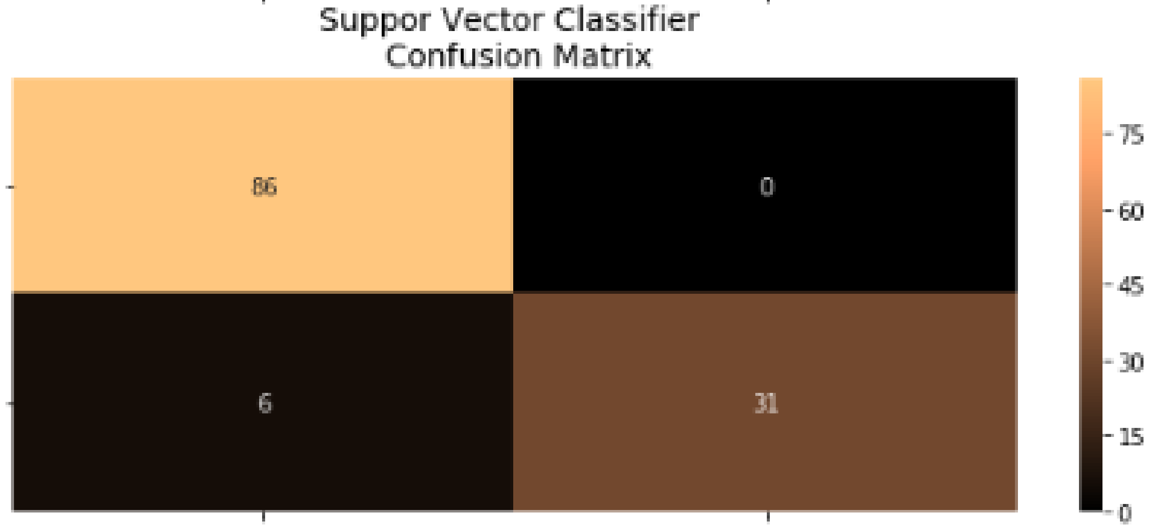
Results obtained on Correlation between different variables with a class being fraud/non fraud



Results obtained on test dataset using the best selected model


Logistic Regression
Confusion Matrix


KNearsNeighbors
Confusion Matrix


Suppor Vector Classifier
Confusion Matrix