Credit Card Fraud Detection
Kavya Nagaraju(hc3344)

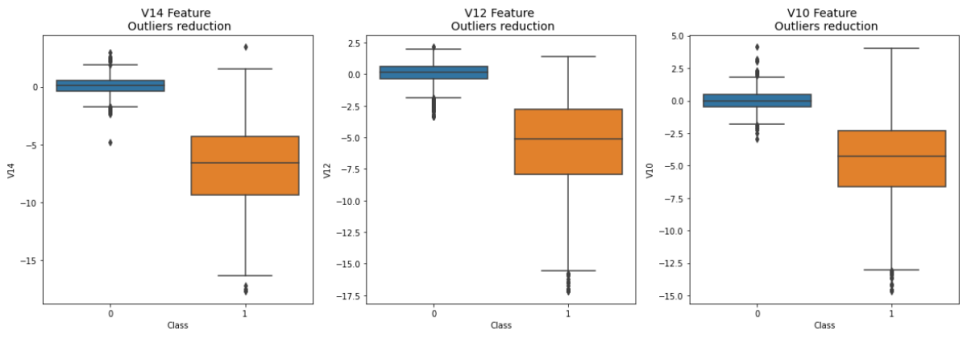
Background/ Introduction

The purpose of the present study involves analyzing a dataset to detect if a transaction is a fraudulent or normal payment. The dataset is highly imbalanced as most of the transactions are non-fraudulent.

Data Pre-Processing, Visualization Methods

Insights from data pre-processing are

- There are no null values in the data frame.
- Most of the transactions are Non-Fraud,99.83% of the time while Fraudulent transactions occur 0.17% of the time in the dataset.
- The Features 'Amount' and 'Time' are also scaled by 'RobustScaler'. Robust Scaler algorithm is being used as it is robust to outliers.
- As Dataset is highly imbalanced, to avoid issues like 'Overfitting' and 'Incorrect correlations', a subset of the data frame is created by performing Random Under Sampling.
- The Subset data frame contains fraudulent cases of 492 and equal cases of randomly chosen nonfraudulent transactions to have an equal distribution.
- Plotting correlation matrix on the subset data frame, it was noted that
  - V17, V14, V12, V10 are Negatively Correlated.
  - Positive Correlation is found in V19, V11, V4, V2 likely to be fraud transactions.
- Anomaly Detection is performed by plotting Box plot and by eliminating extreme outliers, Elimination is done by first determining thresholds subtracting and adding the interquartile range IQR75 and IQR25, and then multiplying IQR. Lastly, if the threshold exceeds both sides conditional dropping is done.



- Dimensionality reduction is performed using T-SNE and could be observed that t-SNE can accurately cluster fraud and nonfraud transactions.

The Data frame is split to Train (70), Validate (10), Test (20) sets. Supervised Models and Unsupervised models such as 'Logistic Regression', 'K nearest Neighbor', 'Support Vector Classifier', 'Decision Tree' and 'K means' is applied.
Under sampling is performed on Train and Validation set by Near Miss technique to balance class distribution by random elimination majority of the class distribution.
Oversampling is performed on Train and Validation set by SMOTE technique has also been implemented to balance class distribution by randomly increasing the minority class.

The Results can be viewed below:

```
Classifiers:  LogisticRegression Has a accuracy of 94.74 % accuracy score
Classifiers:  LogisticRegression Has a f1-score of 94.38 % accuracy score

Classifiers:  KNeighborsClassifier Has a accuracy of 93.68 % accuracy score
Classifiers:  KNeighborsClassifier Has a f1-score of 93.18 % accuracy score

Classifiers:  SVC Has a accuracy of 92.63 % accuracy score
Classifiers:  SVC Has a f1-score of 91.76 % accuracy score

Classifiers:  DecisionTreeClassifier Has a accuracy of 89.47 % accuracy score
Classifiers:  DecisionTreeClassifier Has a f1-score of 89.13 % accuracy score

Classifiers:  KMeans Has a accuracy of 63.16 % accuracy score
Classifiers:  KMeans Has a f1-score of 36.36 % accuracy score
```

Oversampling Logistic Regression, the average Precision recall score was 0.92.
On the test set, Oversampling was performed on Logistic regression and Under sampling was performed on all other models. The results observed were:

Logistic Regression:

	precision	recall	f1-score	support
0	0.95	0.90	0.92	99
1	0.90	0.95	0.92	91
accuracy			0.92	190
macro avg	0.92	0.92	0.92	190
weighted avg	0.92	0.92	0.92	190

KNears Neighbors:

	precision	recall	f1-score	support
0	0.99	0.86	0.92	99
1	0.87	0.99	0.92	91
accuracy			0.92	190
macro avg	0.93	0.92	0.92	190
weighted avg	0.93	0.92	0.92	190

Support Vector Classifier:

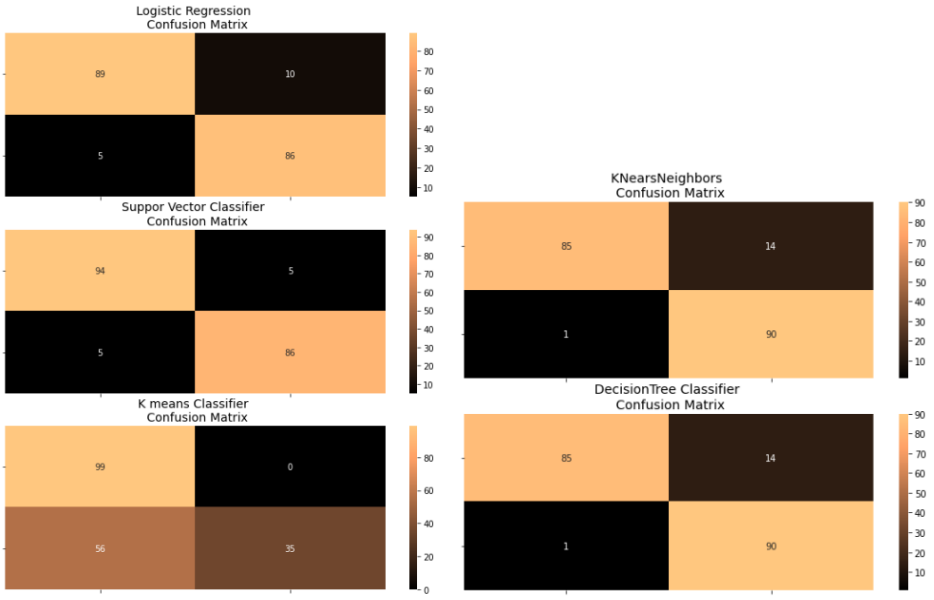
	precision	recall	f1-score	support
0	0.95	0.95	0.95	99
1	0.95	0.95	0.95	91
accuracy			0.95	190
macro avg	0.95	0.95	0.95	190
weighted avg	0.95	0.95	0.95	190

Decision Tree Classifier:

	precision	recall	f1-score	support
0	0.99	0.86	0.92	99
1	0.87	0.99	0.92	91
accuracy			0.92	190
macro avg	0.93	0.92	0.92	190
weighted avg	0.93	0.92	0.92	190

K means Classifier:

	precision	recall	f1-score	support
0	0.64	1.00	0.78	99
1	1.00	0.38	0.56	91
accuracy			0.71	190
macro avg	0.82	0.69	0.67	190
weighted avg	0.81	0.71	0.67	190



Conclusion: Observing the results, it can be noted that Logistic Regression and Support Vector classifier provides more accurate results compared to other models meaning it can distinguish between fraudulent and nonfraudulent transacti