IME-672A
Data Mining & Knowledge Discovery

Project Report
CREDIT CARD FRAUD DETECTION

Group Number: 12

GROUP MEMBERS:

Neil Rajiv Shirude	- 170429
Mukesh Kumar	- 170405
Mayur Kumar	- 170384
Nimish Agarwal	- 170440
Mahajan Deepak Anil	- 170368
Kushagra Gupta	- 170358

1 Acknowledgements

We wish to express our sincere gratitude to our instructor **Dr. Faiz Hamid** for his valuable advice and guidance in completing this project. The way he presented each and every topic in the class made the topics very interesting and understandable, which helped a lot in making our project possible.

2 Introduction

Nowadays, there are a lot of credit card companies who are facing the problem of fraudulent credit card transactions. So, the task is to recognize these fraudulent transactions according to the dataset given to us.

The dataset is highly imbalanced and contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.

3 Our Approach in Brief

After visualizing and preprocessing the data, we split the data into training and testing data using stratified sampling, splitting it in a 4:1 ratio. We used the training dataset for training 5 models - **Isolation Forest, Gaussian Multivariate Anomaly Detection, Logistic Regression, K-Nearest Neighbours Classifier** and **Support Vector Classifier**. We tested these models on the testing dataset and compared their performance on 3 factors:

a. **Accuracy** of the model (in correctly classifying fraud and non-fraud transactions)
b. the number of **fraud transactions** that were **missed**.
c. the number of **non-fraud transactions** that were **incorrectly** classified as **fraud**.

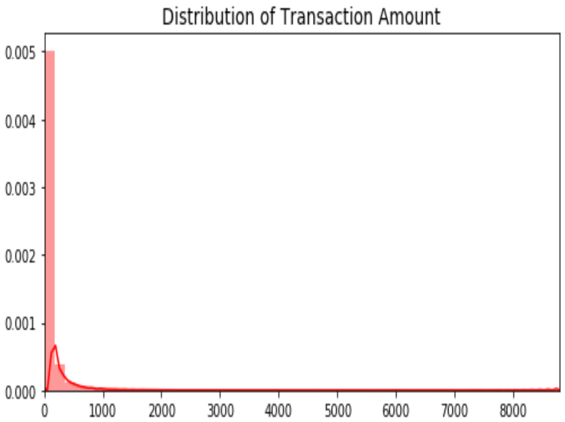
4 Salient Features Of Dataset

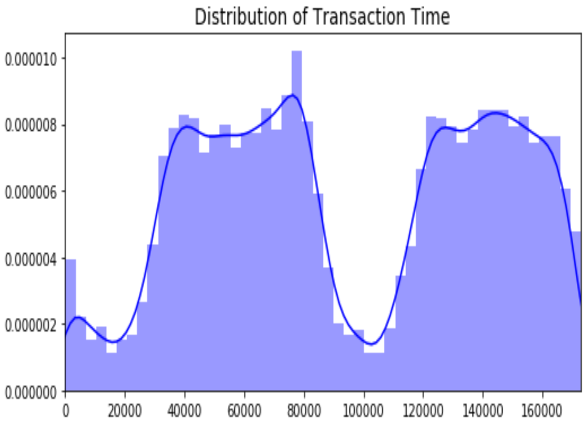
- Total **31** attributes (including **class**).
- **Time** is a Discrete-valued numeric attribute.
- **V1 to V28** are Principal Components of the original dataset not available to us.
- They are a result of Principal Component Analysis.
- They are continuous valued numeric attributes. We cannot say whether they are ratio-scaled or interval-scaled.
- **Amount** is a continuous-valued numeric attribute.
- **Class** is a discrete-valued Binary attribute that takes value **0** for non-fraudulent transaction and **1** for fraud transaction.
- **V1 to V28** are distributed around **0** and are scaled.
- From **V1 to V28**, the variance of attributes decreases from left to right, as expected from a PCA output.

5 DATA VISUALISATION

- Checked the data corresponding to '**Amount**' and '**Time**' attribute and concluded that these two attributes are not scaled.

- Visualised the data distribution of **Amount** and **Time** attribute





- Box plot of **Amount** vs **Class** :
  Through this boxplot, we observed that all the fraud transactions are very low amount so we removed **7** transactions belonging to the non-fraudulent class whose transaction amount > 10,000, i.e, these 7 data points correspond to **extreme outliers**.

- Plotted the distribution of '**Amount**' vs '**Time_min**' showing the data points as the red dot representing the fraudulent transaction and blue dot representing the non-fraudulent transaction.

- Plotted the histogram of '**Amount**' vs '**Time_hour**' showing the fraudulent transaction in red color and non-fraudulent transaction in green color.

- Plotted the distribution of each class for all the attributes **V1-V28**. In this plot, we observed that the distribution of three features **'V10', 'V12**', and '**V14**' seemed like Gaussian distribution. So, to ensure that we are plotting the distribution of Fraud transactions of the three attributes mentioned above.

- Further, we plotted the pair plot of all the possible pairs to visualize the dependency of one attribute over another so that we can decrease the no. pf attributes.

- We plotted the boxplot of each attribute just to observe the distribution and outliers.

- Then we plotted the heatmap of the correlation matrix:
  Most of the pixels are dark pink in color, which means most of the attributes are independent of each other.

- Some cases are of positive correlation and some are negatively correlated.
- But Pearson's product coefficient of all lies between (-0.5 to +0.5). Hence, we are not removing any attribute in this step.

- Negative correlation with class: **V10, V12, V14, V17.**

- We have to make sure we use the subsample in our correlation matrix or else our correlation matrix will be affected by the high imbalance between our classes. This occurs due to the high-class imbalance in the original data frame.

6 Data Preprocessing

- There is neither any missing value nor any noisy data in our dataset. In our problem statement, outliers are the indication of fraudulent cases. So, explicit data smoothing is not performed. Extreme outliers will be removed while training the model. That's why, we first plotted boxplot of amount attribute and observed 7 out of 284,807 datasets having amount greater than 10000 lies to Class 0. We excluded these rows.

- As we have only one file of dataset so there is no need of data integration.

- We used robust scaling to scale the features 'Time' and 'Amount' as this method of scaling uses 'median' as the central tendency, so robust scaling is immune to outliers.

- We plotted the pair plot of all the possible pairs to visualize the dependency of one attribute over another so that we can decrease the no. of attributes. Also, we plotted the heatmap of the correlation matrix. Most of the pixels are dark pink in colour, which means most of the attributes are independent of each other. Some are of positive correlation and some are negatively correlated. But Pearson's product coefficient of all lies between (-0.5 to +0.5). Hence, we are not removing any attribute in this step.

- Since our classes are highly skewed, we should make them equivalent in order to have a normal distribution of the classes. We shuffled the data before creating the subsamples. Cosine Similarity Analysis not performed as data has very few zeros. Parametric Methods for numerosity reduction- NOT Applicable as we need to detect outlier. So, we did numerosity reduction via Random under-sampling to get the balanced consisting of 492 fraudulent cases and 492 non-fraudulent cases.

- And again, we plotted Correlation matrix heat-map on new Balanced Data. (V16, V12), (V17, V1), (V18, V16) & (V18, V17) are correlated attributes of this balanced dataset. But We are not dropping any of the correlated columns because after performing classification, we found that the accuracy of our model decreases in case of dropping the columns.

7 Gaussian Multivariate Anomaly Detection

- Gaussian Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called **outliers**.

- The **intuition** behind applying this technique in our case:-
  - It is preferred over Supervised Learning Algorithms when there are very small positive examples (y=1) compared to large negative examples (y=0) which makes it **ideal** for our case due to **high-class imbalance** nature of our Dataset.
  - Since credit card fraud in the future may be completely different, it gives us the liberty of classifying different types of anomalies. In contrast to other classification algorithms which demand future examples similar to training ones.
  - The features provided to us are already gone through PCA meaning independent of each other thus satisfying the prerequisite of this technique.

- Results:
  - **Accuracy Obtained:** 98.81%
  - **Confusion Matrix-**

Non-fraud Transactions identified correctly: **56806**	Non-fraud Transactions identified incorrectly: **55**
Non-fraud Transactions identified incorrectly: **33**	Fraud Transactions identified correctly: **65**

8 Isolation Forest

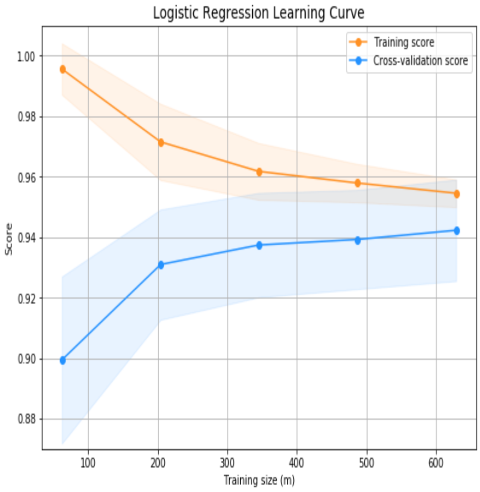
- Isolation Forest is an **anomaly detection algorithm** that uses- outliers are less frequent and spaced further away as compared to inliers.
- The advantages of this algorithm are linear time complexity and small memory requirement.
- The algorithm isolates every data tuple from its surroundings, and notes the number of steps required. Tuples which need a smaller number of steps for isolation are regarded as outliers.
- Results:
  - **Accuracy Obtained:** 97.21%
  - **Confusion Matrix-**

Non-fraud Transactions identified correctly: **55301**	Non-fraud Transactions identified incorrectly: **1560**
Non-fraud Transactions identified incorrectly: **29**	Fraud Transactions identified correctly: **69**

9 Logistic Regression

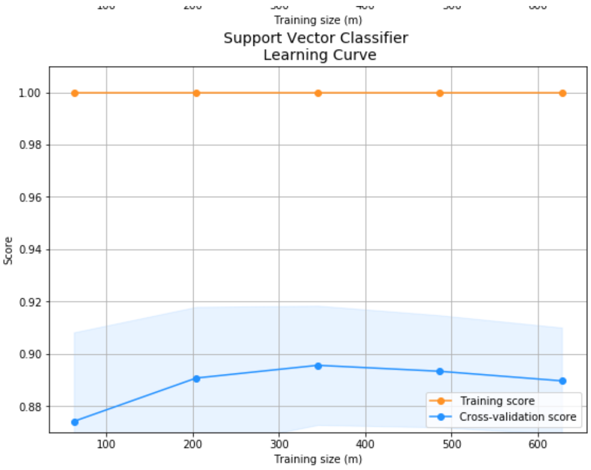
- Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. It is an example of supervised learning. Here we use a function which gives us continuous values between 0 to 1. We use the output of the sigmoid function to predict the class of new transactions.
- As we had to predict the class of the transactions between fraudulent and non-fraudulent we used various algorithms and logistic regression is one of them.
- The confusion matrix for the model tells us that out of the 87 fraudulent transactions the model was able to identify 83 transactions as fraud. The confusion matrix and the learning curve can be found below.

class	precision	recall	f1-score	support
Fraudulent	1.00	0.98	0.99	56861
Normal	0.08	0.87	0.14	98



10 Support Vector Classifier:

- Support vector machine (SVM) is a supervised machine learning algorithm in which we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the optimal hyperplane that maximizes the margin between the two classes. The vectors that define the hyperplane are support vectors.
- It maps data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problems so that data is mapped implicitly to this space.



Class	precision	recall	f1-score	support
Fraudulent	1.00	1.00	1.00	56861
Normal	0.14	0.17	0.15	110

- SVC Has a training score of 93.0 % accuracy score.
- When we were training the data with undersampled data after splitting it into train and test data set consisting of nearly 200 data points then out of the 87 fraudulent transactions, the model was able to predict 77 correctly. But as the total number of data points was very much more than that we thought it would be better to test it for 1/5th stratified data after which the performance of the model was very bad.

11 K-Nearest Neighbours Classifier

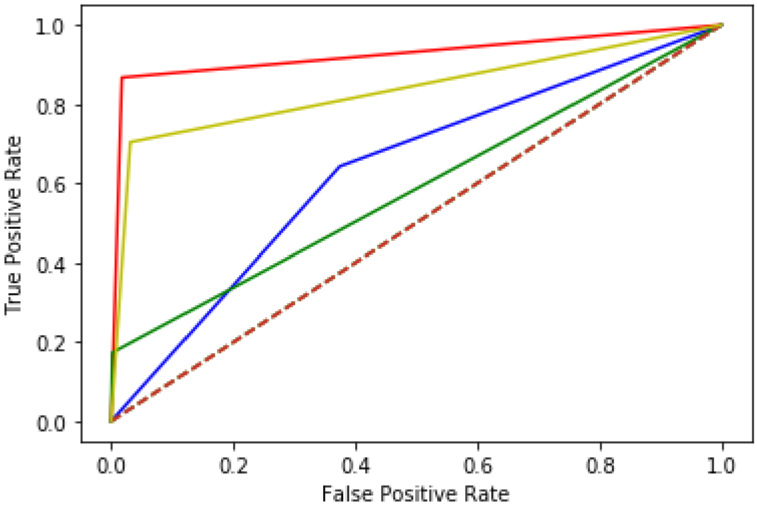
- K-Nearest Neighbours Classifier is a model in which for each data point we find the k nearest neighbours using similarity analysis. Then we assign the majority class of the k nearest neighbours to the given data point . The only variable in the model remains the value of k. We use Cross-Validation data set to find the accuracy of model for some values of k and then use the value of k that gives best performance.
- We applied this method because if we look at the data we have some points that belong to a particular group considering the similarity between them and the pairplots gave us an idea that the fraudulent transactions are closer to each other. That gave us the idea that we should look at the neighbours of the data point which in turn would give us the idea of the class of the given data point .

Class	precision	recall	f1-score	support
Fraudulent	1.00	0.63	0.77	56861
Normal	0.00	0.64	0.01	98

Classifier	True Positives	False Positives	False Negatives	True Negatives
Isolation Forest	55301	1560	29	69
Gaussian Anomaly	56806	55	33	65
Logistic Regression	56000	861	13	85
KNN	55000	1861	81	17
SVC	36000	20861	35	63

12 Conclusion

The ROC curve for the Logistic Regression(Shown in Red), Isolation Forest(Shown in yellow), KNN(Shown in blue) and SVC(Shown in green) give us the idea about the relative performance of those models and we can compare them easily using area under curve. We can clearly see from the figure that the ROC curve of Logistic Regression has the highest area under curve amongst the classifiers.



- Among all algorithms, **Logistic Regression** has the **highest percentage of correctly identified transactions**, while **Gaussian Anomaly Detection** has the **highest accuracy**.
- **Logistic Regression-**
    - **Logistic Regression** classified 85 out of 98 fraudulent transactions correctly, with **92%** accuracy overall.
    - Only about **10%** of the **fraudulent transactions** were **missed** by Logistic Regression.
    - Number of non-fraudulent transactions classified as fraudulent: approx 1000.
    - Total Number of Transactions needed to be verified: approx 1100 in 10 hours
    - Assuming that 1 employee would need 10 minutes for manually verifying whether a transaction is actually fraud or not, he/she can verify 60 transactions in 10 hours.
    - Practically, the bank needs to have a team of 40 dedicated individuals, who would work in 2 shifts, if Logistic Regression is used.
- **Gaussian Anomaly Detection-**
    - If **Gaussian Anomaly Detection** is used, only 120 transactions need to be manually verified in 10 hours, with **99.84% accuracy** which would need a team of 4 dedicated individuals only.
    - But **33%** of the **fraudulent transactions** were missed by this model.
- Hence, the decision regarding which model to use lies with the bank, depending on whether it is fine with a lot of fraud transactions going undetected or spending resources on manually verifying detected transactions.