**Credit Card Fraud Detection System**

**Introduction:**

Credit card fraud is most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. To detect the fraudulent activities the credit card fraud detection system was introduced. Our project aims to focus on building a machine learning based predictive model to detect whether a transaction is fraudulent or not.

We used algorithms such as Logistic Regression, Decision tree, Random Forest, KNN, XGBoost. The result of these algorithms is based on accuracy, precision, recall, and F1-score. The Random Forest algorithm has the greatest accuracy, precision, recall, and F1-score and is considered as the best algorithm that is used to detect the fraudulent transaction

**Problem Statement:**

In this project we want to identify fraudulent transactions with Credit Cards. Our objective is to build a fraud detection system using Machine learning techniques.

**About dataset:**

This project uses a dataset of 18,52,394 transactions. Each transaction is labelled either fraudulent or not fraudulent. The prevalence of fraudulent transactions is very low in the dataset. Less than 0.1% of the card transactions are fraudulent. This means that a system predicting each transaction to be normal can reach an accuracy of over 99.89% despite not detecting any fraudulent transaction. This will necessitate adjustment techniques.

**Insights:**

* Target variable is highly imbalanced as there are more legit transactions compared to fraudulent transactions
* Fraudulent transactions are happening more during mid nights
* Fraudulent transactions are happening more during weekends
* The average amount of fraudulent transaction is $390
* People between the age groups of 20-60 are more victim of fraud or prone to fraud

(Working age groups are more prone to fraud)

* More number of fraudulent transactions occurred in New York, Texas, Pennsylvania states compared to other states
* More no. of frauds occurred in less dense cities
* Frauds are happening more when customer is far from the merchant location
* Fraudulent transactions are more for purchases in grocery and shopping stores.
* Fraudulent transaction may occur when the second transaction occurred within one hour of first transaction

**Techniques used in the project:**

The project compares the results of different Machine learning techniques:

* Logistic Regression
* Decision tree
* Random Forest
* KNN
* Naive Bayes
* XGBoost

**Results before SMOTE:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.no | Model name | Accuracy | Recall score | Precision score | F1 score |
| 1 | Logistic regression | 99.48% | 0% | 0% | 0% |
| 2 | Decision tree | 99.79% | 81.58% | 79.17% | 80.36% |
| 3 | Random forest | 99.86% | 75.66% | 97.62% | 82.25% |
| 4 | XGBoost | 99.89% | 82.41% | 94.91% | 88.22% |
| 5 | KNN | 99.52% | 15.33% | 69.01% | 25.09% |
| 6 | Naive Bayes | 99.48% | 0% | 0% | 0% |

**Results after SMOTE:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.no | Model name | Accuracy | Recall score | Precision score | F1 score |
| 1 | Logistic regression | 99.48% | 0% | 0% | 0% |
| 2 | Decision tree | 99.55% | 76.7% | 54.94% | 64.02% |
| 3 | Random forest | 99.81% | 76.83% | 85.93% | 81.12% |
| 4 | XGBoost | 99.87% | 85.11% | 89.73% | 87.36% |
| 5 | KNN | 94.74% | 71.4% | 6.77% | 12.36% |
| 6 | Naive Bayes | 9.61% | 90.83% | 0.05% | 0.01% |

**Results after PCA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.no | Model name | Accuracy | Recall score | Precision score | F1 score |
| 1 | Logistic regression | 71.26% | 67.68% | 72.96% | 70.22% |
| 2 | Decision tree | 96.50% | 97.59% | 95.53% | 96.55% |
| 3 | Random forest | 97.03% | 71.56% | 94.41% | 81.76% |
| 4 | XGBoost | 85.48% | 83.59% | 86.92% | 85.22% |
| 5 | KNN | 97.87% | 99.86%% | 96.04% | 97.92% |
| 6 | Naive Bayes | 76.45% | 69.72% | 77.83% | 75.45% |

**Conclusion:**

The best results are achieved by Random Forest performed before SMOTE. With this approach, the model is able to detect 99.89% of all fraudulent transactions in the test set. This satisfies the primary objective to detect the vast majority of abnormal transactions. And this technique and model used are simple to implement, easy to use and can be updated in real-time.

In addition, the number of false positives remains acceptable. This means a lot less verification work (on legitimate transactions) for the fraud detection compare to some other approaches which failed on this aspect. Key results are shown below:

**Confusion matrix achieved using Random Forest:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted values | | |
| Actual values |  | Negative | Positive |
| Negative | 7,37,038(TN) | 71(FP) |
| Positive | 937(FN) | 2,912(TP) |

**Comparison of key performance indicators between the tested approaches:**