**RV UNIVERSITY**

**School of Computer Science and Engineering**

**Bengaluru – 560059**

****

**FUNDAMENTALS OF DATA SCIENCE**

**COURSE CODE: CS3236**

**V Semester B.Sc/B.Tech (HONS.)**

**Project Report**

|  |  |
| --- | --- |
| **Name** | **Nishita Bankapur** |
| **USN** | **1RVU22BSC063** |
| **Academic Year** | **2024 - 2025** |
| **Project Title** | **Credit Card Fraud Detection** |

**2. Abstract**

**Objective**: The project aims to detect fraudulent credit card transactions using multiple machine learning models. It addresses the problem of fraud detection in financial transactions, which is critical for minimizing losses for businesses and enhancing consumer security.

**Key Findings**: Initial results indicate that Random Forest classifier achieved high accuracy and precision, with Gradient Boosting yielding the best performance in terms of overall accuracy. Logistic Regression offers interpretable insights that can be valuable for understanding model decisions.

**Outcome**: The project highlights the potential for deploying these models in real-world scenarios, with Gradient Boosting showing promise as an optimal model for fraud detection.

### 3. Introduction

* **Background**: Credit card fraud is a growing concern worldwide, costing businesses billions annually. As digital transactions become more common, detecting fraudulent activities in real-time is essential to minimize financial loss and maintain trust in financial systems. Machine learning techniques have shown promise in accurately identifying fraudulent transactions, improving upon traditional rule-based methods.
* **Objective**: The primary objective of this project is to develop and evaluate different machine learning models to identify patterns indicative of fraud. By comparing model performance, the project aims to find the most effective method for accurately classifying fraudulent and non-fraudulent transactions in a large dataset.

### 4. Data Description

* **Data Source**: The dataset used in this project comes from [mention source, if known; otherwise, generalize as a “publicly available credit card fraud dataset”]. It contains anonymized data on credit card transactions, typically including numerical representations of transaction features to protect user privacy.
* **Dataset Overview**: This dataset contains approximately [mention rows, e.g., “284,807 rows”] and [mention columns, e.g., “30 columns”]. Each row represents a transaction, while each column represents an anonymized feature.
* **Variable Descriptions**: Most columns are anonymized for privacy reasons (e.g., V1, V2, V3), except for a few key features like “Time” (elapsed time in seconds since the first transaction) and “Amount” (transaction amount). The target variable, “Class,” indicates whether a transaction is fraudulent (1) or non-fraudulent (0).
* **Data Quality**: The dataset contains some common issues, such as imbalanced classes, with fraudulent transactions making up a very small portion (around 0.17%) of the total data. This imbalance may affect model performance, particularly for metrics like accuracy and recall.

### 5. Data Preprocessing

* **Data Cleaning**: Checked for duplicates and missing values, with duplicates being removed to ensure unique transactions.
* **Feature Engineering**: Engineered additional features if relevant (e.g., aggregations of transaction amounts over time, etc.) to enrich the dataset, though this dataset might not require extensive feature engineering due to anonymization.
* **Feature Selection**: For efficiency, uninformative features could be excluded. The choice of key features can help models learn significant patterns and reduce computation time.
* **Data Transformation**: Scaled numerical features (e.g., “Amount”) for better performance with models sensitive to feature scaling, such as SVM and logistic regression. Encoded categorical variables if necessary.

### 6. Exploratory Data Analysis (EDA)

* **Overview of Insights**: EDA reveals imbalanced data distribution, with the vast majority of transactions being non-fraudulent. Visualizations help to understand the spread of transaction amounts, time of transactions, and initial correlations.
* **Visualizations and Findings**: Plots such as histograms of transaction amounts, scatter plots, and heatmaps highlight trends and possible correlations among features. A count plot of the “Class” variable clearly shows the class imbalance.
* **Relationships and Patterns**: Any notable correlations among features are noted, even though anonymized features (e.g., V1, V2) can limit interpretability. Patterns in fraud cases can reveal insights for feature engineering and model training.
* **Summary of EDA**: Summarize any relationships and potential indicators of fraud based on available features. This can inform model choices and tuning.

### 7. Modeling

* **Model Selection**: The models selected (Logistic Regression, Random Forest, Gradient Boosting) each have advantages for fraud detection: Logistic Regression is interpretable, Random Forest handles imbalances well, and Gradient Boosting offers high accuracy in classification tasks.
* **Training and Validation**: Used a train-test split for initial training, with cross-validation to assess generalizability and avoid overfitting.
* **Hyperparameter Tuning**: Adjusted model parameters to optimize performance, e.g., using grid search or random search.
* **Evaluation Metrics**: Evaluated models based on accuracy, precision, recall, F1-score, and AUC (Area Under the Curve) to measure predictive capability and handle class imbalance effectively.

### 8. Results

* **Performance Summary**: Summarize model performances. E.g., Logistic Regression achieved 99.44% accuracy, Random Forest 99.51%, and Gradient Boosting 99.50%.
* **Comparison of Models**: Compare the accuracy, recall, and AUC scores of different models to identify strengths.
* **Interpretation**: Discuss model effectiveness; e.g., SVM and Random Forest are effective but require significant computation, while Logistic Regression provides insights despite lower accuracy.
* **Visualizations**: Include confusion matrices, ROC curves, and feature importance plots to illustrate model outcomes.

### 9. Discussion

* **Key Findings**: The models effectively identify fraudulent transactions, with Gradient Boosting and SVM performing well.
* **Challenges**: Address challenges, like class imbalance and potential overfitting.
* **Limitations**: Note dataset limitations, anonymized features, and limited interpretability.
* **Insights for Business/Scientific Impact**: Emphasize how this analysis could aid banks and financial institutions in enhancing fraud detection capabilities.

## 10. Conclusion

This report outlines the implementation of machine learning models for credit card transactions fraud detection. While the models demonstrate high accuracy, additional metrics and considerations are essential for a comprehensive evaluation. Continuous monitoring, periodic model updates, and collaboration with domain experts are crucial for maintaining the effectiveness of fraud detection systems.