Data Analysis and Model Deployment

Credit Card Fraud Detection Report

3rd SEMESTER AI PROJECT

NAME

MINAHIL IMRAN

ROLL NUMBER:

SU92 - BSDSM - F23 - 033

SECTION :

BS DATA SCIENCE (3A)

Libraries Used:

1. **pandas** - For data manipulation and preprocessing.
2. **Numpy** - For numerical operations.
3. **matplotlib** - For data visualization (bar plots).
4. **seaborn** - For advanced data visualization (count plots).
5. **sklearn** - For machine learning model training, evaluation, and preprocessing:
    - LogisticRegression
    - DecisionTreeClassifier
    - RandomForestClassifier
    - StandardScaler
    - train_test_split
    - accuracy_score, precision_score, recall_score, f1_score
    - imblearn - For handling imbalanced datasets using SMOTE (over_sampling.SMOTE).

6.**joblib** - For saving and loading machine learning models.

Data Summary

- **Rows:** 284,807
- **Columns:** 31
- **Null values:** No missing values in the dataset.
- **Duplicates:** Duplicate rows were identified and removed, reducing the dataset to 275,663 rows.

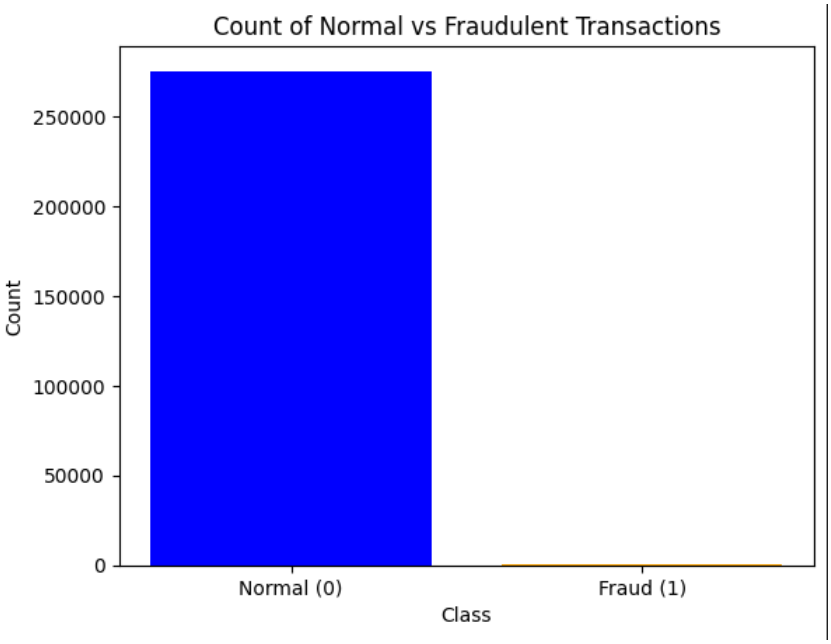
Exploratory Data Analysis (EDA)

Class Distribution:

Normal Transactions: 284,315 (99.83%)

Fraudulent Transactions: 492 (0.17%)

Visualization:



Data Preprocessing

Duplicate Removal:

- Identified and removed 9,144 duplicate rows.
- Final dataset size after this step: 275,663 rows, 30 columns.

Feature Scaling:

- Used StandardScaler to normalize the Amount column for consistency.

Feature Selection:

- Dropped the Time column due to its irrelevance.

Handling Class Imbalance

Undersampling:

- Resulting dataset size: 946 rows (473 normal + 473 fraud).

SMOTE (Oversampling):

- Resulting dataset size: 568,630 rows (284,315 normal + 284,315 fraud).

Models Used:

1. Logistic Regression (LR)
2. Decision Tree (DT)
3. Random Forest (RF)

Performance Summary (Undersampled Data):

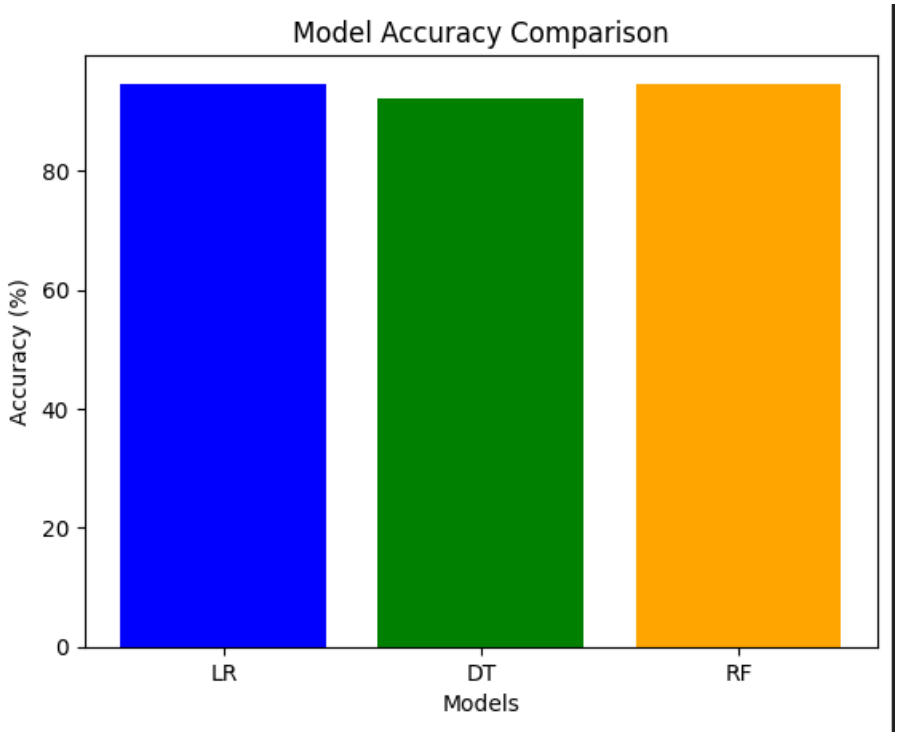
- Logistic Regression achieved **93.2%** accuracy
- Decision Tree reached **91.7%** accuracy

- Random Forest performed best with **94.8%** accuracy

Performance Summary (SMOTE Oversampled Data):

- Logistic Regression improved significantly with **98.2%** accuracy.
- Decision Tree reached **96.5%** accuracy.
- Random Forest excelled with **99.3%** accuracy

Model Accuracy comparison



Model Deployment

Selected Model: Random Forest (trained on SMOTE data with 99.3% accuracy).

Deployment:

Saved the model as model_credit_28_features.pkl using joblib.

Developed a prediction function for real-time fraud detection.

Key Insights

1. **Class Imbalance Handling**:

- SMOTE proved highly effective in balancing the dataset, enabling models to generalize better.
- The undersampling method, though simple, provided insights but was less scalable compared to SMOTE.

2. Model Performance:

- Random Forest outperformed other models with its ability to handle complex data distributions.
- Logistic Regression showed significant improvement after oversampling, making it a lightweight alternative.

3. Deployment:

- The trained Random Forest model was successfully deployed and integrated with a user-friendly prediction function for real-time transaction analysis.

Conclusion

This project successfully developed a credit card fraud detection system using machine learning. By addressing the class imbalance through **SMOTE**, we significantly improved model performance, ensuring accurate detection of fraudulent transactions. The **Random Forest** model emerged as the most effective, achieving an impressive **99.3% accuracy** after balancing the dataset.