CAPSTONE PROJECT REPORT

Project Name: FindDefault (Prediction of Credit Card fraud)

Developer/Author: Saurav Bhagat

Email: sauravbhagat@gmail.com

Git Link: https://github.com/Saurav2108/Final_Capstone_Project_UpGrad

➢ Problem Statement :

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

The dataset 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. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

We have to build a classification model to predict whether a transaction is fraudulent or not.

➢ Key Areas to focus on

The following is recommendation of the steps that should be employed towards attempting to solve this problem statement:

- **Exploratory Data Analysis:** Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.
- **Data Cleaning:** This might include standardization, handling the missing values and outliers in the data.
- **Dealing with Imbalanced data:** This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building.
- **Feature Engineering:** Create new features or transform the existing features for better performance of the ML Models.
- **Model Selection:** Choose the most appropriate model that can be used for this project.

- **Model Training:** Split the data into train & test sets and use the train set to estimate the best model parameters.
- **Model Validation:** Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.
- **Model Deployment:** Model deployment is the process of making a trained machine learning model available for use in a production environment.

➢ Dataset Overview : Credit card data with following details

- Total Features: 31
- Total (Rows)Transactions: 284,807
- Fraudulent Transactions: 492

➢ Exploratory Data Analysis (EDA)

- **Count Plot:** A count plot showing the distribution of fraudulent vs non-fraudulent transactions. Which shows the skewness present in data,
- **Heat-Map**: A correlation heatmap to visualize the relationships between different features in the dataset.
- **Histogram**: plot of all features showing normally distributed trend.

➢ Data Preprocessing (Cleaning)

- **Standardization**: The Amount feature was standardized.
- **Missing Values Treatment:** No missing values were found in the dataset.
- **Feature selection:** Time has low correlation with the target variable hence dropping from the datasets

➢ Dealing with Imbalanced Data

- **SMOTE**
- We are implementing Over-sampling technique to deal with imbalanced data: Increasing the number of fraudulent transactions
- (Synthetic Minority Over-sampling Technique) imblearn.over_sampling ADASYN

➢ Feature Engineering

- **Interaction Terms:** Created interaction terms to capture relationships between features.
- **Feature Scaling:** Applied Min-Max scaling to ensure all features are on the same scale

➢ Model Building and Evaluation

Below Models are trained and evaluated to find best model that generalize the model well

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **XGBoost**

➢ Model Performance Metrics:

Base on the all model performance we select the best model

Recall (Sensitivity): The ratio of correctly predicted fraudulent transactions to all actual fraudulent transactions.

➢ Model Results: The models are evaluated using Imbalanced and smote

Smote technique to handled imbalanced data gave the best results, so we are proceeding with that and below scores are gave for it

Best Model –XGBoost using Smote

Model Name	AUC	Average Precision
Logistic Regression	0.9710	0.6968
Decision Tree	0.8970	0.3378
Random Forest	0.9644	0.8747
XGBoost	0.9831	0.8671

➢ Model Deployment:

To deploy the trained model, we used the Pickle module to save it in a serialized format. This allows the best-performing model to be easily loaded for future use.

```
import pickle

# Define filenames for saving the model and scaler
model_filename = 'best_fraud_detection_model.pkl'  # Model saved as .pkl
scaler_filename = 'scaler.pkl'  # Scaler saved as .pkl

# Save the model using pickle
with open(model_filename, 'wb') as model_file:
    pickle.dump(best_model_instance, model_file)

# Save the scaler using pickle
with open(scaler_filename, 'wb') as scaler_file:
    pickle.dump(scaler, scaler_file)

print(f"Model saved as {model_filename}")
print(f"Scaler saved as {scaler_filename}")

# Make predictions on the test data
predictions = loaded_model.predict(X_test)
```

➢ Project Result Summary (Conclusion)

- Leveraged advanced machine learning techniques to effectively detect fraudulent transactions within a highly imbalanced dataset.
- Tackled class imbalance using SMOTE (Synthetic Minority Over-sampling Technique), enhancing the dataset’s representativeness.
- Conducted comprehensive experimentation across multiple classifiers, optimizing model performance for fraud detection.
- Achieved impressive precision and recall metrics, underscoring the model's reliability in identifying fraudulent activity.
- High AUC score illustrates the model's robust ability to distinguish between fraudulent and legitimate transactions, validating its effectiveness.