

Project Title: Credit Card Fraud Detection using Logistic Regression

Abdullah Mazharuddin Khaja

Objective:

The focus of this particular project is to identify medicated credit card frauds using machine learning and particularly by using logistic regression. The objective is to design a system that provides a reliable means of detecting fraud when such transactions are being undertaken while at the same time avoiding false positives and false negatives that make real time protections for the financial institutions and customers.

Motivation:

Although not very frequent, fraudulent credit cards are a major threat for both customers and companies due to the financial and reputational losses that they produce. Therefore, it is important for schemes to be developed towards the effective detection of fraud with a view of ensuring that confidence in financial processes is not eroded. Because fraud data sets normally have a significant number of legitimate transactions and only a few fraud cases the challenge is to develop a model that will be able to identify them while not flooding users with a high number of false positives.

Modules:

Module 1 Data Acquisition and Preprocessing

- Input: Data containing the credit card transactions where each transaction is described by the amount and the time of the transaction.
- Functionality:
 - Previous step: Load the data and preprocess the data (e. g. , impute missing values and delete duplicates).
 - Implement feature selection and feature engineering (for example, it is necessitating to create more features such as time period between two transactions).
 - Synthetic Minority Over-sampling Technique (SMOTE) – Use it to employ more instances in the minority class which in this case is the malicious/ fraudulent class.
- Output: A clean dataset adjusted for missing values in order to feed into a model.

Module 2: Model Training

- Input: Preprocessed dataset.
- Functionality:
 - Perform binary classification on fraud samples and non-fraud samples with logistic regression.
 - Apply a form of regularization specifically, L2 to avoid over fitting.
 - Take training on training set and for validation cross-validation techniques must be used.
- Output: Implemented and trained logistic regression with best set hyperparameters.

Module 3: Model Evaluation

- Input: Test dataset.
- Functionality:

- Measure the validity of the created model's performance on the basis of Precision, Recall, F1-Score, and ROC AUC.
- True positives and false positives can also be measured along with true negatives, false negatives by using confusion matrices.
- Output: A text detailing the results of the models' performance; the scores used in the assessment as well as the confusion matrices.

Module 4: Real-Time Processing and Applications and the Future Improvements

- Input: Data that pertains to the real time credit card transactions (plan for future improvement).
- Functionality:
 - Implement the model in real-time applications in areas such as finance transaction to facilitate fraud identification.
 - Notify administrators and customers of possible fraudulent activities.
- Output: Real time fraud detection model with simultaneous retraining of the model to identify new patterns of fraud.

Timelines:

S.No	Activity	Expected Due Date
1	Data Acquisition and Preprocessing	9/30/2024
2	Model Training	10/10/2024
3	Model Evaluation	10/20/2024
4	Real-Time Processing and Future Enhancements (optional)	10/30/2024
5	Presentation and Documentation Preparation	11/10/2024
6	Final Report	11/15/2024