GROUP PROJECT PROPOSAL

Group 01

Team member: Le Ba Hung
Nguyen Trang Nhung
Pham Thi Ngoc Mai
Tran Tue Nhi

1. Business Problem

a. Introduction:

The financial sector is continually challenged by the increasing prevalence and sophistication of credit card fraud, which not only leads to significant financial losses but also diminishes consumer trust, thereby jeopardizing the stability of financial institutions. A study conducted by Vellore Institute of Technology in 2019 revealed alarming statistics: 1,579 data breaches and nearly 179 million compromised records, with credit card fraud being the most prevalent issue. The widespread adoption of credit cards has created ample opportunities for fraudsters to exploit security weaknesses (Dornadula & Geetha, 2019). Existing fraud detection methods struggle to keep up with evolving tactics. In fact, research from Tabuk University in Saudi Arabia indicates that despite efforts by financial institutions to enhance security, the incidence of fraud tends to rise alongside the increasing number of credit card users (Alenzi & Aljehane, 2020).

b. Objective:

The primary objective of this project is to develop a machine learning model using logistic regression to detect fraudulent activities in credit card transactions, thereby minimizing financial losses and enhancing transaction security.

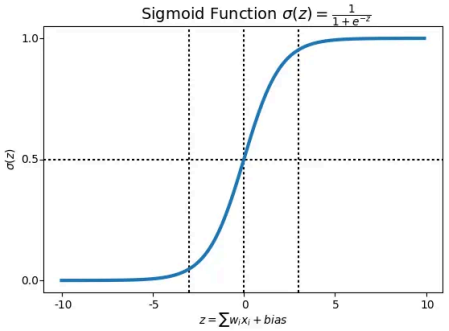
2. Model Background: Logistic Regression

a. Overview:

Logistic Regression is a supervised machine learning algorithm that predicts the probability of a categorical dependent variable based on one or more predictor variables. Although it is similar to Linear Regression and the name also suggests a “regression” model, Logistic Regression deals with classification problems (Müller & Guido, 2016), particularly binary classification ones where there are only two limited outcomes such as yes/no, 0/1, true/false or fraud/non-fraud (Kanade, 2022).

b. General Properties:

- **Input:** Feature vectors from training data.
- **Output:** A probability that a given transaction is fraudulent, which is mapped to a binary category based on a threshold.
- **Logistic function: Sigmoid** maps predictions and their probabilities (Kanade, 2022), outputs values between 0 and 1.
- **Parameter:** The **parameter C** governs the balance of regularization strength, where higher values of C indicate reduced regularization (Müller & Guido, 2016).



c. Advantages:

- **Ease of Use:** Logistic Regression is easy to train and implement, with minimal maintenance required (Alenzi & Aljehane, 2020).
- **Interpretability:** Logistic regression provides clear interpretations, which is essential in regulated sectors like banking. It allows for assessing the impact of each input variable on the predicted outcome.
- **Real-time Prediction:** Logistic regression supports quick retraining with each new example, enabling near real-time responses (The ultimate guide to logistic regression, 2020).
- **Scalability:** Thanks to its efficient algorithm and low computational demands, logistic regression can easily be scaled, even with growing data volumes and speeds.

3. Data Description

a. Data Overview and Sources:

The dataset encapsulates transactions made by European cardholders in September 2013, collected over a period of two days. It comprises 284,807 transactions, of which 492 are fraudulent, reflecting a highly unbalanced dataset with a positive class (frauds) making up only 0.172% of all transactions.

b. Types of Data: The dataset contains both numerical and categorical data:
- Numerical Data:
    + 'Amount': transaction value.
    + 28 'V' features: Principal Component Analysis (PCA) transformed variables to preserve transaction details while safeguarding sensitive information.
- Categorical data: Binary 'Class' label indicating fraud status (1 for fraud, 0 for non-fraud), making it apt for classification tasks.

c. Data Relevance:

This dataset is suitable for the Logistic Regression model, which predicts a binary outcome using feature vectors (such 'Amount' and PCA components). Because this model can produce probabilities instead of depending on simplistic accuracy metrics, which can be deceptive in imbalanced situations, it is well-suited to address the class imbalance in the dataset. Furthermore, including various transaction attributes in an anonymous form allows the Logistic Regression model to detect complex patterns that indicate fraudulent activity.

d. Compatibility with Logistic Regression:

- Data Types & Size: The dataset contains both numerical and categorical data, making it compatible with logistic regression. Additionally, with over 550,000 records, the dataset size is sufficient for logistic regression.

- Feature Independence: The PCA-transformed features V1-V28 in the dataset support Logistic Regression's assumption of feature independence by minimizing multicollinearity. However, 'Time' and 'Amount' may require further analysis to ensure they align with the model's expectations of independent predictors.

- Linearity: Linearity between the log odds of the target variable and the predictors is assumed but should be verified for each feature to optimize model performance.

- Balance of Classes: Logistic regression performs best when the classes are balanced. However, the dataset is skewed since fraudulent transactions are much rarer.

- Interpretability: The dataset aligns with logistic regression in terms of interpretability, as it includes a clear input-output structure suitable for supervised learning.

4. Pipeline

Step 1: Collect Data. Acquire a dataset with both fraudulent and legitimate transactions. This dataset should include variables like the amount, timing, and location of transactions.
Step 2: Prepare the Data.
- Clean: Address any missing or corrupt data.
- Normalize: Scale the data to prevent bias towards higher magnitude features.
- Encode: Transform categorical data into numeric formats, such as through one-hot encoding.
- Feature Engineering: Create new features or modify existing ones to enhance model performance (e.g., transaction frequency, transaction amount per time interval, or transaction amount per merchant category).
- Exploratory Data Analysis: Visualize the distribution of dataset to gain further insights to refine existing features or create new ones (Back to Feature Engineering).
Step 3: Split the Data. Divide the data into 70% training, 15% validation, 15% testing.
Step 4: Train the Model. Configure and train a logistic regression model using the training data.
Step 5: Evaluate the Model. Assess the model with the testing data using metrics suited for imbalance, such as precision, recall and F1-score.
Step 6: Model Optimization and Tuning.
- Regularization Techniques: Experiment with different regularization techniques (L1, L2 or Elastic Net) to prevent overfitting and help generalize the model better.
- Adjust Decision Threshold: Instead of using the default 0.5 cutoff in logistic regression, adjust the threshold value to balance between precision and recall, depending on the business requirements (e.g., prioritizing reducing false negatives over reducing false positives).
Step 7: Re-evaluate the Optimized Model. Re-evaluate the model, assess the improvements, and compare these results with the baseline model (the model before optimization).

5. Constraints

- **Data Availability:** There might not be sufficient suitable data to train the model well.
- **Data Quality:** We need to ensure the completeness, consistency, accuracy of data by handling missing values, normalizing data formats and rigorous data validation.
- **Class Imbalance:** There's a risk that the model will overfit to the majority class and fail to generalize well to new and unseen data, especially those representing the minority class.

References

Alenzi, H. Z., & Aljehane, N. O. (2020). *Fraud detection in credit cards using logistic regression.* International Journal of Advanced Computer Science and Applications, 11(12).

Dornadula, V. N., & Geetha, S. (2019). *Credit Card Fraud Detection using Machine Learning Algorithms.* Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Kanade, V. A. (2022, April 18). *Logistic regression: Equation, assumptions, types, and best practices.* Spiceworks Inc. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Practices%20for%202022-,What%20Is%20Logistic%20Regression%3F,1%2C%20or%20true%2Ffalse

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists.* O'Reilly Media, Inc.

The ultimate guide to logistic regression for Machine Learning. Keboola. (2020, August 24). https://www.keboola.com/blog/logistic-regression-machine-learning