Capstone Project Proposal: Fraud Detection

1. Domain Background

Fraud detection is a critical area in the financial services industry. With the rise of digital banking and online transactions, fraudulent activities have become more sophisticated and frequent. Banks must proactively detect and prevent fraud to protect customers and maintain trust. As a software engineer in the Fraud Prevention CoE at a bank, I have firsthand exposure to the challenges and importance of real-time fraud detection systems.

2. Problem Statement

The goal of this project is to build a machine learning model that can accurately classify whether a financial transaction is fraudulent or legitimate. The model should be able to detect patterns in transaction data that indicate potential fraud, helping reduce false positives and improve detection rates.

3. Datasets and Inputs

I plan to use the Kaggle Credit Card Fraud Detection dataset, which contains anonymized credit card transactions made by European cardholders in September 2013. The dataset includes:

- 284,807 transactions
- 492 frauds (0.172%)
- Features: 30 numerical input variables (PCA-transformed), including Time, Amount, and Class (target)

The dataset is publicly available and suitable for this project.
Dataset link https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

4. Solution Statement

I will build a supervised classification model using algorithms such as Random Forest, XGBoost, and Neural Networks. The model will be trained to distinguish between fraudulent and non-fraudulent transactions. I will also explore techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

5. Benchmark Model

As a baseline, I will implement a simple Logistic Regression model. This will serve as a benchmark to compare the performance of more complex models. The benchmark will help evaluate whether advanced models provide significant improvements.

6. Evaluation Metrics

Given the class imbalance, accuracy is not a reliable metric. Instead, I will use:

- Precision
- Recall

- F1-Score
- AUC-ROC Curve

These metrics will help assess the model's ability to detect fraud while minimizing false positives.

7. Project Design

The project will follow this workflow:

1. Data Exploration & Preprocessing
   - Handle class imbalance
   - Normalize/scale features
   - Split into train/test sets
2. Model Training
   - Train baseline and advanced models
   - Use cross-validation
3. Model Evaluation
   - Compare models using evaluation metrics
   - Visualize ROC curves and confusion matrices
4. Model Optimization
   - Hyperparameter tuning
   - Feature selection
5. Deployment
   - Deploy the best model using AWS SageMaker
   - Create a simple API or web interface for inference