Model Development Report for Fraud detection

### Detailed Analysis and Insights

### Jayanth Vodnala

### October 29, 2024

**Abstract**

This report provides an in-depth analysis of the development and implementation of machine learning models to identify fraud credit uses cases.

**Contents**

1. [Introduction](#_bookmark0) 3
   1. [Stakeholder Identification](#_bookmark1) 3
   2. [Problem Statement](#_bookmark2) 3
2. [Dataset Description](#_bookmark3) 3
   1. [Source](#_bookmark4) 3
   2. [Relevance](#_bookmark5) 3
3. [Methodology](#_bookmark6) 3
   1. [Feature Engineering](#_bookmark7) 3
   2. [Model Selection and Hyperparameters](#_bookmark8) 4
4. [Model Evaluation](#_bookmark9) 4
5. [Recommendations and Future Work](#_bookmark10) 4
   1. [Model Recommendation](#_bookmark11) 4
   2. [Improvements and Future Work](#_bookmark12) 4
6. [Conclusion](#_bookmark13) 4
7. [Appendices](#_bookmark14) 5
   1. [GitHub Repository](#_bookmark15) 5

# Introduction

## Stakeholder Identification

Financial institutions and fraud protection teams housed inside banks or credit card companies constitute the main project stakeholders. These players are concentrated in using data-driven insights to increase customer confidence, lower financial losses, and improve fraud detection accuracy.

## Problem Statement

The objective is to develop a machine learning model capable of accurately classifying transactions as either fraudulent or non-fraudulent, based on user demographic and transaction data. This solution aims to enhance fraud detection rates, reduce false positives, and enable proactive measures against fraud.

# Dataset Description

## Source

The dataset has multiple attributes, including user income, occupation, and credit card information. This data serves as the basis for constructing a model that differentiates between fraudulent and lawful transactions. It is sourced from Kaggle. [Fraud Detection Dataset](https://www.kaggle.com/datasets/sameerk2004/fraud-detection-dataset)

## Relevance

The dataset contains essential information regarding user attributes and transaction details, allowing the model to learn patterns associated with fraudulent activities. This information is crucial for developing a predictive model tailored to real-world fraud detection.

# Methodology

## Feature Engineering

#### Selected/Engineered Features:

* + - Income Binning: The Income variable was categorized into discrete bins (Low, Medium, High) to identify trends within defined income ranges.
    - Categorical Encoding: Categorical variables, including Profession and Income Bin, were converted using one-hot encoding to ensure interoperability with machine learning techniques.
    - Standardization: Numerical features were normalized to enable interoperability with algorithms such as SVM, which are sensitive to feature scales.

## Model Selection and Hyperparameters

#### Models and Hyperparameters:

**Models and Hyperparameters**

The following models were implemented, each with a set of hyperparameters tuned to achieve optimal performance:

Decision Tree Classifier:

Parameters tuned: max\_depth, min\_samples\_split, and min\_samples\_leaf.

Random Forest Classifier:

Parameters tuned: n\_estimators, max\_depth, and min\_samples\_leaf.

XGBoost Classifier:

Parameters tuned: n\_estimators, max\_depth, and learning\_rate.

Support Vector Machine (SVM):

**Parameters tuned: C, kernel, and gamma.Reasoning:** The selection of these models aligns with the need for both robust performance across varied data distributions and actionable insights into feature impacts, facilitating strate- gic business decisions.

**Rationale**

Each model was chosen for its distinct benefits: Decision Trees and Random Forests provide interpretability, XGBoost provide strong performance on structured data, and Support Vector Machines are proficient in high-dimensional feature spaces. Hyperparameter tuning was performed with GridSearchCV, employing F1-score or AUC-ROC as the evaluation metric to address the uneven characteristics of fraud detection.

# Model Evaluation

**Evaluation Metrics:** Accuracy,

The models were evaluated using F1-score and AUC-ROC, given their suitability for imbalanced datasets. F1-score balances precision and recall, while AUC-ROC assesses the model’s capability to differentiate between classes.

**Performance Analysis:** Both models d.

* **Decision Tree**: Demonstrated reasonable classification ability, though prone to overfitting on training data.
* **Random Forest:** Showed balanced performance, especially with well-tuned n\_estimators and max\_depth, yielding a more stable model compared to Decision Tree.
* **XGBoost:** Provided the highest AUC-ROC score, highlighting its effectiveness in handling complex patterns within the dataset.
* **SVM:** Performed well on balanced data, but required extensive tuning due to sensitivity to C and gamma parameters.

# Recommendations and Future Work

## Model Recommendation

XGBoost is recommended due to its robustness, flexibility to non-linear relationships in the data, and superior AUC-ROC score, according to evaluation metrics. It is ideally suited for prompt implementation in fraud detection owing to its equilibrium of precision and recall.

## Improvements and Future Work

Anomaly Detection Integration: Implementing Isolation Forest or One-Class SVM to detect rare fraud cases as anomalies.

Advanced Feature Engineering: Developing additional features based on transaction frequency and timestamp information.

Ensemble Techniques: Combining predictions from multiple models for a more comprehensive approach to fraud detection.

# Conclusion

This research illustrates the utility of machine learning in fraud detection, providing financial institutions with a data-driven methodology for controlling and decreasing fraud risk. The developed models offer actionable insights that enhance the accuracy of fraudulent transaction identification.

# Appendices

## GitHub Repository

Access all project-related code and documentation through the GitHub repository at [https:](https://github.com/msfdev1234/Project_Wine_Dataset)

[//github.com/msfdev1234/Project\_Wine\_Dataset](https://github.com/msfdev1234/Project_Wine_Dataset), ensuring full transparency and repro- ducibility of the results.