Machine Learning Model Report

1. Introduction

In this project, we deal with a highly imbalanced dataset with abnormal scattering of data points.

The primary objective was to develop a model to predict whether a given account is likely to be fraudulent (bad_flag).

The dataset had more than 98,000 rows, and the target class (bad_flag) was severely imbalanced, leading to challenges in model training, evaluation, and hyperparameter tuning.

2. Data Preprocessing

Data preprocessing was crucial to deal with missing values, data imbalance, and scalability of the model.

Initially, missing values were filled with zeros using the fillna method. Then, scaling was applied using StandardScaler to ensure that all features had equal weight.

To handle the imbalanced dataset, SMOTE was used for oversampling the minority class. PCA (Principal Component Analysis) was used to reduce the dimensions of the dataset, retaining 80% of the variance. This helped reduce overfitting and computational cost in further steps.

3. Model Selection

The logistic regression model was chosen because it is simple, interpretable, and effective in binary classification problems.

Logistic regression is a linear model that estimates the probability of an instance belonging to the positive class (fraudulent).

It was an appropriate choice given the imbalance in the dataset.

Additionally, grid search was used to fine-tune the hyperparameter "C" to achieve the best possible model performance. However, due to the large size of the dataset (over 98,000 data points), the grid search method introduced high computational costs, making it difficult to tune efficiently.

4. Challenges Faced

There were multiple challenges throughout the project:

1. **Overfitting to the Majority Class**: The dataset was highly imbalanced, and the model tended to overfit to the majority class, resulting in high accuracy but poor performance in predicting the minority class.

2. **SMOTE and Its Impact**: The synthetic data generated by SMOTE might have influenced the model to focus too much on the minority class, leading to overfitting on the minority data points.

3. **High Computational Cost**: Using GridSearchCV for hyperparameter tuning was computationally expensive, especially with oversampling, and the model training took significantly longer than anticipated.

4. **PCA and Data Loss**: Applying PCA reduced the dataset's dimensionality, but it might have led to the loss of some critical data, affecting the final model performance.

5. ** Outliers were not able to be separated due to the scattered nature of the data

5. Evaluation Metrics

Choosing the correct evaluation metrics was critical in assessing the model performance.

While accuracy can be misleading in the case of imbalanced datasets, metrics like precision, recall, and F1-score are more informative.

SMOTE helped balance the classes but may have skewed results towards the minority class. In some cases, the high accuracy was misleading, as the model was very good at predicting the majority class but struggled with the minority class.

6. Conclusion

The project highlighted the importance of handling class imbalance, proper scaling, and dimensionality reduction when dealing with large datasets.

While the Logistic Regression model performed well on the majority class, its performance on the minority class was poor due to the imbalance.

Future improvements could include using more advanced techniques for handling class imbalance, such **XGBoost**, which can better deal with imbalanced classes.

Additionally, using better computationally expensive hyperparameter tuning strategies and dimensionality reduction techniques may help in improving performance.