Individual Reflection on Payment Propensity Prediction Model

Introduction

The Payment Propensity Prediction Model project provided an insightful experience in applying machine learning techniques to financial risk assessment. The primary goal was to predict whether customers would default on payments based on historical data, leveraging classification models such as **Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost**. Throughout this project, I gained valuable experience in data preprocessing, feature engineering, model selection, hyperparameter tuning, and evaluation.

Key Learnings

1. Data Preprocessing & Feature Engineering

- We identified **no missing values** in the dataset, simplifying the preprocessing steps.
- **Outlier detection using boxplots** helped in understanding anomalies in features like credit limits and payment history.
- **Feature normalization (StandardScaler)** ensured consistent scaling of numerical variables, improving model performance.
- **Splitting data into 80% training and 20% testing** provided a balanced evaluation strategy.

2. Exploratory Data Analysis (EDA)

- **Correlation analysis** revealed that **payment history (PAY_0 - PAY_6)** was the strongest predictor of default.
- Customers with **higher outstanding balances** had a higher risk of default.
- Younger customers (**under 30 years old**) showed an increased likelihood of delayed payments.
- **Higher credit limits correlated with lower default risk**, suggesting that financially stable customers are less likely to default.

Model Performance & Comparison

We experimented with multiple machine learning models, each offering unique advantages:

- **Logistic Regression**: Achieved **80.78% accuracy**, but struggled to capture complex non-linear relationships.
- **Decision Tree**: Improved accuracy to **81.72%**, but had high variance and was prone to overfitting.
- **Random Forest**: Provided **81.5% accuracy** with better generalization due to ensemble learning.

- **XGBoost, LightGBM, and CatBoost**: Boosting models delivered superior ROC-AUC scores, with **CatBoost achieving the highest at 77.53%**.

Model	Accuracy	ROC-AUC Score
Logistic Regression	80.78%	70.76%
Decision Tree	81.72%	74.23%
Random Forest	81.5%	77.0%
XGBoost	81.4%	76.3%
LightGBM	81.67%	77.28%
CatBoost	**81.69%**	**77.53%**

Key Takeaways:

- **CatBoost was the best-performing model**, handling class imbalance and categorical features effectively.
- **Boosting models (XGBoost, LightGBM, and CatBoost) outperformed traditional models**, showcasing their advantage in financial applications.
- **False negatives (missed defaulters) remained a challenge**, highlighting the need for further tuning or alternative modeling approaches.

Challenges Faced

1. Class Imbalance

- Since there were significantly more non-defaulters than defaulters, the models tended to predict non-default more frequently.
- **Addressed using oversampling techniques like SMOTE** to balance the dataset.

2. Model Interpretability

- While boosting models provided the best accuracy, their black-box nature made interpretability challenging.
- **SHAP (SHapley Additive Explanations) was explored** to provide insight into feature importance and improve transparency.

Business Impact & Future Enhancements

Business Benefits:

- **Early identification of high-risk customers** allows financial institutions to take proactive measures.
- **Optimized debt collection strategies** based on risk segmentation, leading to improved cash flow management.
- **Reduction in operational costs** by automating risk assessment using machine learning.

Future Enhancements:

- **Exploring Deep Learning Models (LSTMs, Transformers)** for sequential payment pattern prediction.
- **Incorporating Behavioral Analysis** using transaction history and external financial indicators.
- **Deploying a Real-Time API** to allow businesses to integrate the model into financial decision systems.

Conclusion

This project demonstrated the power of **machine learning in financial risk assessment**. Through rigorous data analysis, model experimentation, and performance tuning, we successfully built a robust **payment propensity prediction model**. The project provided valuable insights into credit risk modeling and reinforced the importance of **explainable AI, data preprocessing, and hyperparameter tuning**. Moving forward, refining this model with real-time deployment and alternative learning techniques will further enhance its effectiveness in financial applications.

Prepared by:
SivaRamaKrishna Yarra
Date: March 2025