CREDIT CARD DEFAULT PREDICTION

Presented by Junaid Khan & Shoaib Akhter Syed

OVERVIEW

- Introduction
- Dataset Overview
- Data Preprocessing
- Dataset Visualization
- Model Selection
- Model Deplpyment
- Results
- Conclusion

INTRODUCTION

In today’s financial landscape, understanding credit card defaults is crucial for both consumers and financial institutions. With increasing credit card usage, especially among younger users, the risk of default has become a significant concern. Defaults can lead to financial losses for banks and damage individuals’ credit, affecting their future borrowing ability.

This project aimed to develop a predictive model that estimates the likelihood of credit card default using historical data. By applying machine learning, we sought to identify patterns and factors contributing to defaults, enabling more informed lending decisions. Through data exploration, preprocessing, and model building, our objective was to enhance the credit assessment process and reduce default rates effectively.

DATASET OVERVIEW

We utilized a dataset from the UCI Machine Learning Repository, containing essential demographic, financial, and behavioral information on credit card holders. Key features included:

Demographic Attributes: Age, gender, marital status, and education level, offering insights into user backgrounds.

Financial Attributes: Credit limits, monthly bill amounts (BILL_AMT1–6), and recent payments (PAY_AMT1–6), which track payment behaviors.

Behavioral Attributes: Payment status over the last six months (PAY_0–PAY_6), indicating whether payments were made on time or delayed.

Our target variable, “Default Payment Next Month”, denotes whether the client defaulted (1) or did not default (0) the following month.

DATASET OVERVIEW

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	\
0	1	20000	2	2	1	24	2	2	-1	-1	
1	2	120000	2	2	2	26	-1	2	0	0	
2	3	90000	2	2	2	34	0	0	0	0	
3	4	50000	2	2	1	37	0	0	0	0	
4	5	50000	1	2	1	57	-1	0	-1	0	

	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	\
0	...	0	0	0	0	689	0	
1	...	3272	3455	3261	0	1000	1000	
2	...	14331	14948	15549	1518	1500	1000	
3	...	28314	28959	29547	2000	2019	1200	
4	...	20940	19146	19131	2000	36681	10000	

	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
0	0	0	0	1
1	1000	0	2000	1
2	1000	1000	5000	0
3	1100	1069	1000	0

DATA PREPROCESSING

- **Handling Missing Values:** We filled missing demographic values using the mode and numerical values with mean imputation to ensure dataset consistency.

```
ID                          0
LIMIT_BAL                   0
SEX                         0
EDUCATION                   0
MARRIAGE                    0
AGE                         0
PAY_0                       0
PAY_2                       0
PAY_3                       0
PAY_4                       0
PAY_5                       0
PAY_6                       0
BILL_AMT1                   0
BILL_AMT2                   0
BILL_AMT3                   0
BILL_AMT4                   0
BILL_AMT5                   0
BILL_AMT6                   0
PAY_AMT1                    0
PAY_AMT2                    0
PAY_AMT3                    0
PAY_AMT4                    0
PAY_AMT5                    0
PAY_AMT6                    0
default payment next month  0
dtype: int64
```

DATA PREPROCESSING

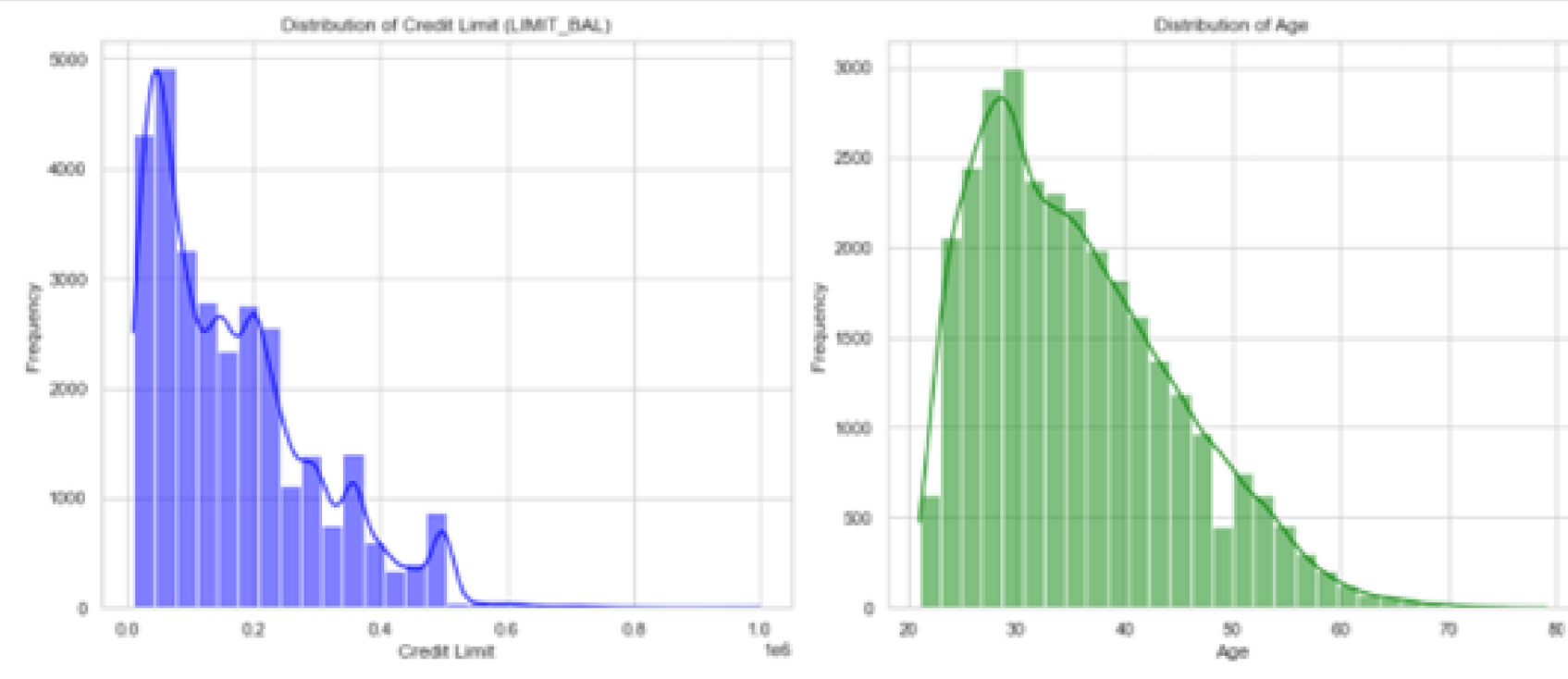
- **Outlier Detection:** Outliers were identified using box plots, especially in high-variance features like credit limit (LIMIT_BAL). We applied quantile-based capping to limit the impact of extreme values.

- **Feature Scaling:** Standardization was applied to numeric features, setting a mean of 0 and standard deviation of 1. This scaling improves model performance by aligning feature magnitudes.

- **Encoding Categorical Variables:** We used one-hot encoding for Logistic Regression and integer encoding for Random Forest to ensure compatibility with each model.

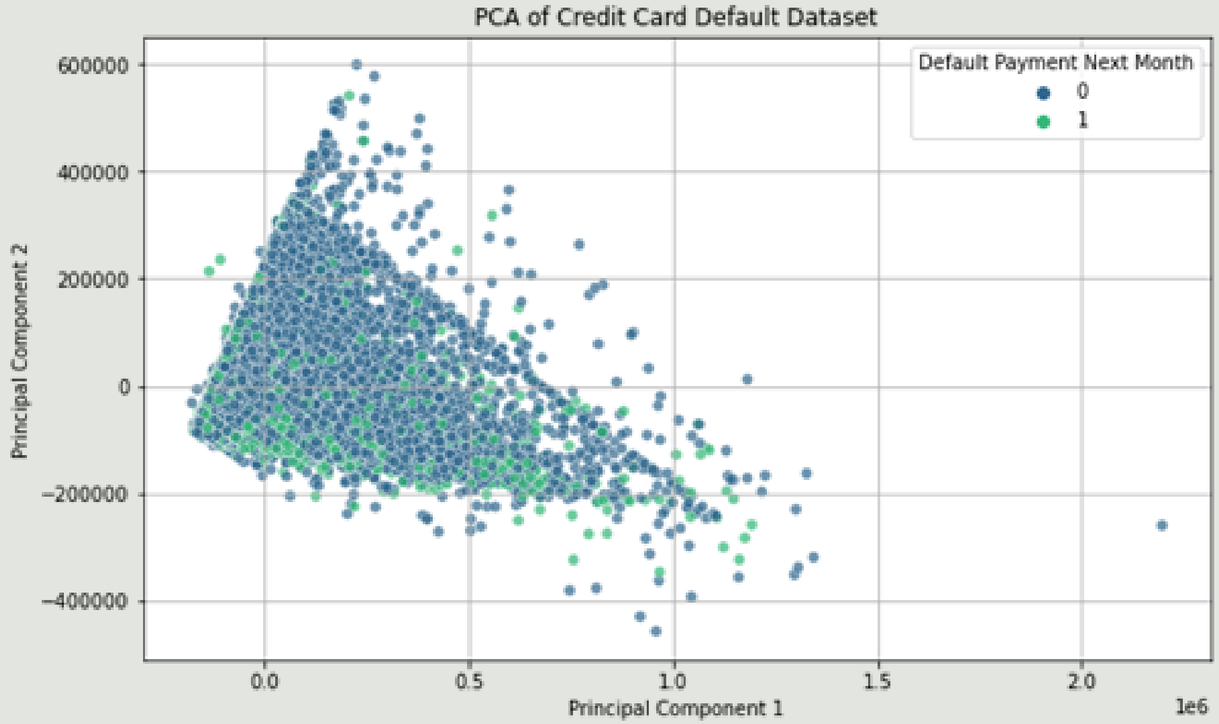
DATA VISUALIZATION

- **Histograms:** Histograms for LIMIT_BAL revealed a skewed distribution, with many clients holding lower credit limits, while AGE distribution peaked in the 30-40 range, hinting at a younger demographic.



DATA VISUALIZATION

- **Feature Transformation -** PCA helped reduce dimensionality, focusing on features with the most variance to simplify visualization and interpretation. By compressing data into fewer dimensions, it captured core patterns and made visual analysis more accessible. PCA components were selected based on variance contribution, allowing us to retain essential information while reducing noise, providing a clearer view of the data structure and aiding feature selection



MODEL SELECTION

1. Logistic Regression

Chosen for its interpretability and suitability for binary classification. After preprocessing, we used cross-validation and hyperparameter tuning to enhance its accuracy.

2.Random Forest

This ensemble model was selected for its ability to capture non-linear relationships. Hyperparameter tuning, including adjustments to tree depth and the number of estimators, improved its performance.

MODEL SELECTION

N-fold Cross-validation

Both models were evaluated using 10-fold cross-validation to obtain the mean and standard deviation of each performance metric, ensuring that our models generalize well to unseen data. Random Forest outperformed Logistic Regression in almost all metrics, with lower variance in its predictions, indicating a more stable model.

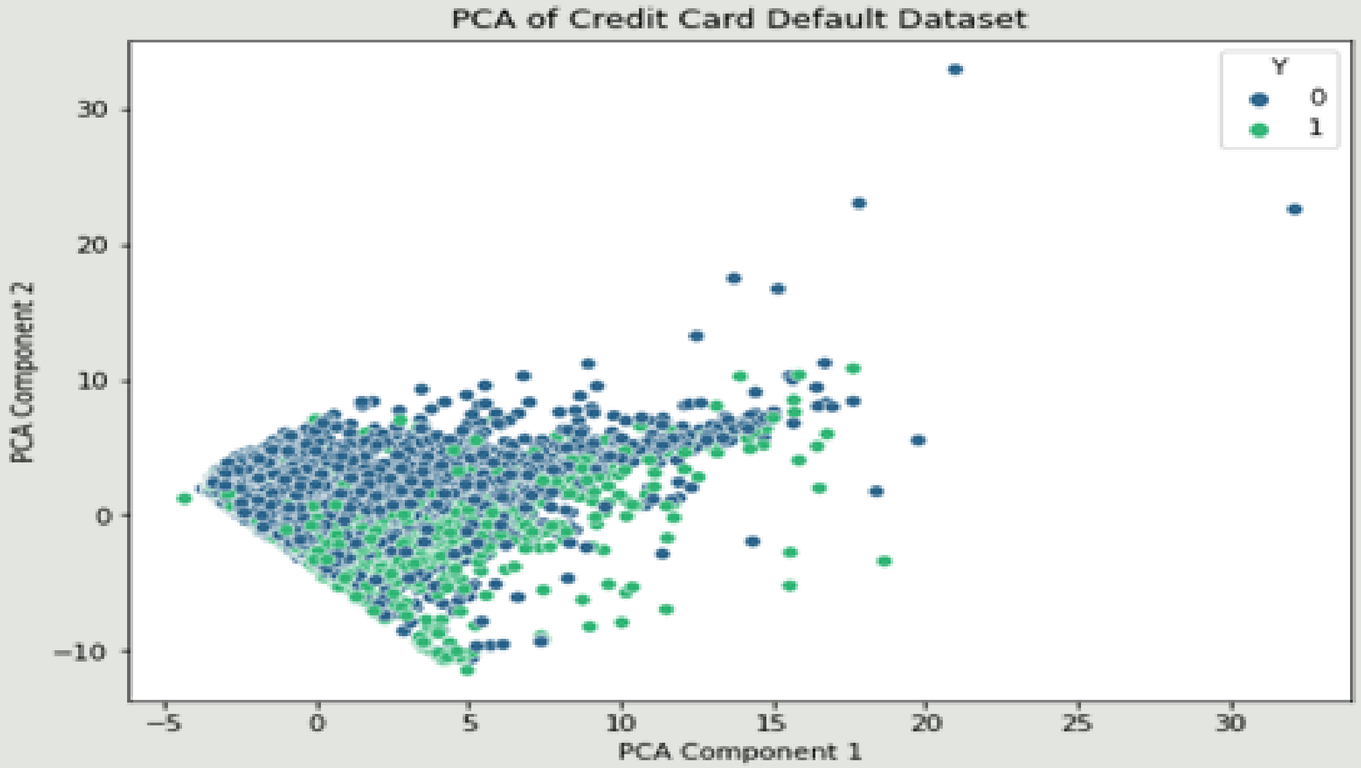
```
Cross-validation scores for each fold: [0.81452381 0.8202381  0.81547619 0.81142857 0.80857143]
Average cross-validation score: 0.8140476190476191
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      7040
           1       0.62      0.36      0.46      1960

    accuracy                           0.81      9000
   macro avg       0.73      0.65      0.67      9000
weighted avg       0.79      0.81      0.79      9000

Accuracy Score: 0.8131111111111111
```

MODEL SELECTION

PCA Scatter Plot and Correlation Heatmap

We visualized feature relationships and redundancies with PCA scatter plots and correlation heatmaps. PCA helped us explore the dataset's two-dimensional structure, while the correlation heatmap revealed interdependencies between features, guiding our feature selection.

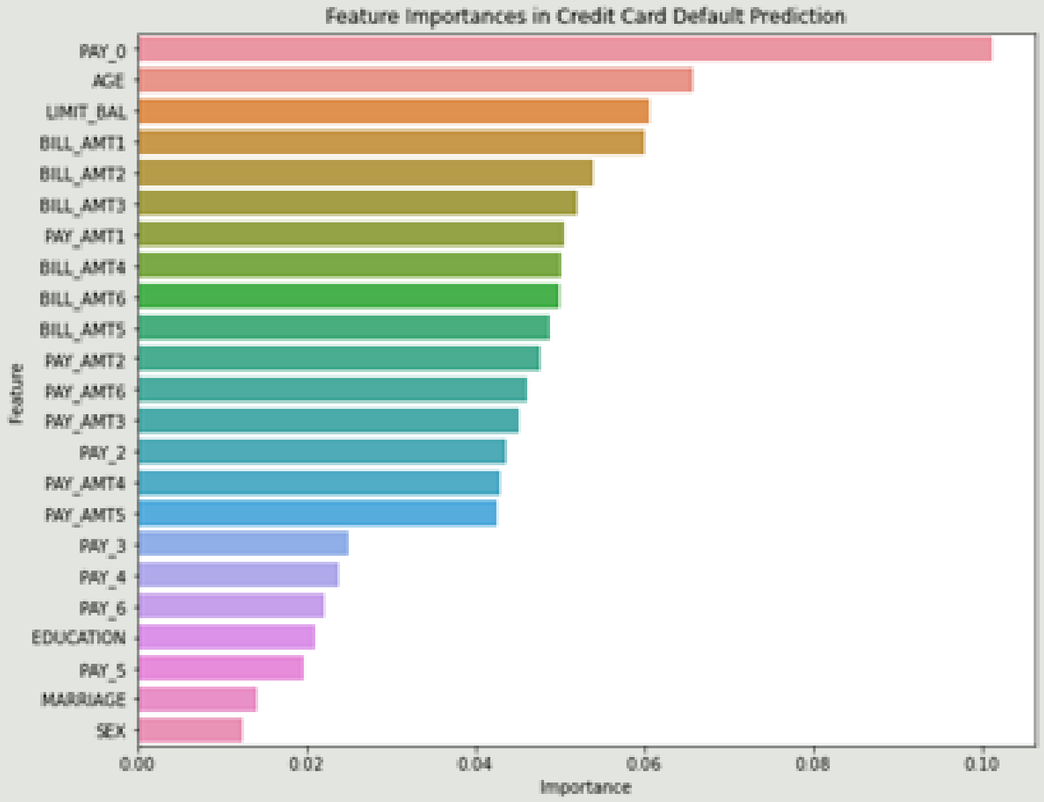
.



MODEL SELECTION

Feature Importances

Understanding feature importance provides valuable insights for credit risk analysis and can help financial institutions in developing more tailored credit policies.



MODEL DEPLOYMENT

Gradio enabled us to deploy the final model as an interactive app. Users could input details and immediately receive a prediction, making the model accessible and practical for real-world use, with potential for integration into broader risk.

MODEL DEPLOYMENT

- **Challenges:** The dataset's class imbalance, with fewer defaults, made capturing true default cases challenging. Selecting the right features and avoiding noise was also critical.
- **Mistakes and Improvements:** Early feature selection could have been more selective, as some irrelevant features added noise. Future iterations may involve more systematic validation and multiple model comparisons.
- **Future Work:**Future projects might explore neural networks, employ more current data, and enhance the app with real-time updating capabilities and user feedback mechanisms."

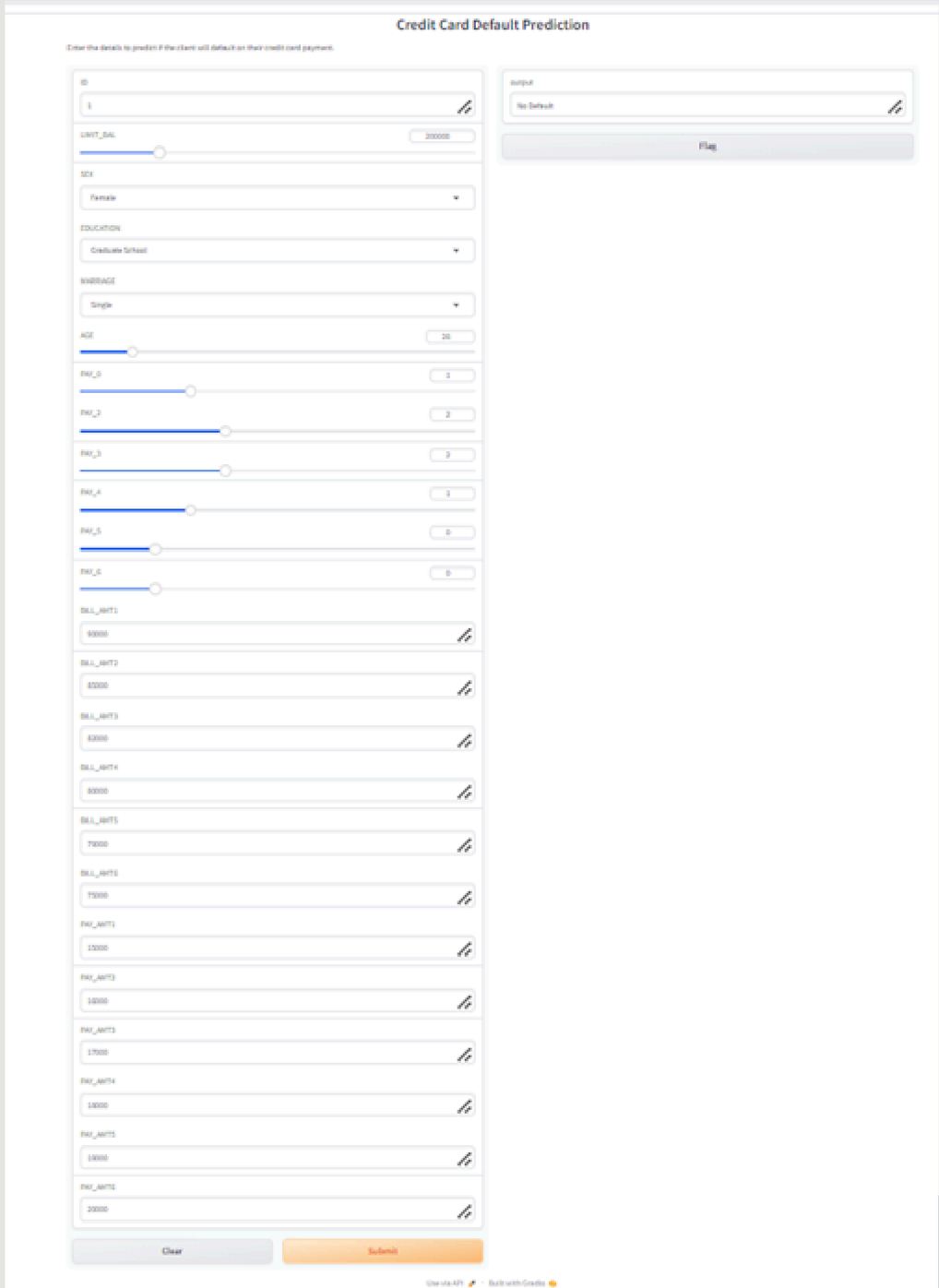
RESULTS

Both models were evaluated using 10-fold cross-validation, with Random Forest outperforming Logistic Regression in nearly all metrics:

- **Logistic Regression:** Achieved an accuracy of around 78%, with balanced precision and recall across classes after class-weight adjustments.
- **Random Forest:** Reached an accuracy of approximately 82%, showing improved precision and recall due to its capability to model complex patterns. SMOTE further boosted performance by addressing class imbalance.

Overall, Random Forest demonstrated higher accuracy and consistency, making it the more reliable model for predicting credit card defaults in our analysis.

FINAL IMPLEMENTATION



CONCLUSION

This project highlighted the importance of data preprocessing, feature scaling, and handling class imbalance for effective credit default prediction. Random Forest emerged as the stronger model, providing better accuracy and stability compared to Logistic Regression.

Key takeaways included the need for careful feature engineering and balanced data to improve predictive accuracy. Future improvements could explore advanced models or automated hyperparameter tuning to further enhance performance.

Thank You