Team Member's Details:

• **Group Name: "I am the TEAM"**

• **Name: Krishnakant Sonji**

• **Email: cronjobs alo.edu, krishnasonji@gmail.com**

• **Country: United States**

• **College/Company: University of New York at Buffalo**

• **Specialization: Data Science**

Problem Description: XYZ Bank wants to send personalized Christmas offers to different customer groups instead of sending the same o er to all. The goal is to create no more than 5 customer segments using machine learning to improve the efficiency of the campaign.

GitHub Repo link: https://github.com/Thir13een/DG-week-13

Report: Customer Segmentation Model Selection

Introduction

Objective: The primary goal of this project was to analyze and build a predictive model for customer segmentation using multiple machine learning algorithms and select the most suitable solution. We aimed to achieve high predictive accuracy while also meeting business requirements for model interpretability.

Dataset Overview:

- **Total Records**: 976,320
- **Total Features**: 44
- **Key Features**: Included customer_age, household_income, customer_seniority_months, and start_of_month_customer_relation.
- **Goal**: To balance predictive accuracy with interpretability to make the model both accurate and understandable for business use.

Solution Overview

1. **Logistic Regression (LR)**
    - **Approach**: Logistic Regression served as our baseline model due to its simplicity and high interpretability. It calculates the probability of each class based on feature weights.
    - **Performance**: Achieved an accuracy of 99.98% with near-perfect precision and recall, indicating minimal misclassification.
    - **Strengths and Limitations**:
        - **Strengths**: High interpretability, straightforward feature analysis, and efficient computational performance.
        - **Limitations**: May struggle with capturing complex, non-linear relationships in the data.
    - **Visual Aid**: The confusion matrix illustrated strong classification results with minimal errors across classes.

2. **Random Forest (RF)**
    - **Approach**: A powerful ensemble model that combines multiple decision trees to handle complex data interactions.
    - **Performance**: Achieved 99.91% accuracy with high feature importance, identifying start_of_month_customer_relation as a key predictor.
    - **Strengths and Limitations**:
        - **Strengths**: High accuracy, ability to handle complex data structures, and moderate interpretability through feature importance scores.
        - **Limitations**: More computationally intensive and slightly less interpretable than simpler models.
    - **Visual Aid**: Feature importance plot displayed the influence of top features on predictions.

Comparison of Models: Logistic Regression vs. Random Forest

Metric	Logistic Regression (LR)	Random Forest (RF)
Accuracy	99.98%	99.91%
Interpretability	High	Moderate
Computation	Fast	Slower, resource-intensive
Complexity Handling	Limited	Excellent

Key Insights:

- **Logistic Regression** offers high accuracy and interpretability, making it suitable for simpler relationships. However, it may struggle with complex, non-linear patterns.
- **Random Forest** delivers high accuracy and effectively handles complex patterns, although it is computationally more demanding.

Conclusion:

- If interpretability is the top priority, **Logistic Regression** is preferred.
- If the focus is on handling complexity and achieving robustness, **Random Forest** is the better option.

Selected Solution

Chosen Model: **Random Forest Classifier**

Rationale:

- **Balanced Performance**: Provides high accuracy (99.91%) with excellent recall and precision across all classes.
- **Complexity Handling**: Suitable for datasets with complex relationships.
- **Interpretability**: Offers feature importance scores, providing insight into feature influence, which aligns with business requirements for interpretability.

Visual Aids:

- **Confusion Matrix**: Demonstrates classification accuracy.
- **Feature Importance Plot**: Highlights key features like start_of_month_customer_relation.

Conclusion

Summary of Findings:

- The Random Forest model met performance goals with high accuracy and interpretability. It provides a good balance for handling complex data relationships and understanding feature influence.

Final Model Justification:

- The Random Forest model was selected based on its robust performance, capacity to handle complex data, and alignment with the project's requirements for both accuracy and interpretability.

Future Improvements

Model Enhancements:

- **Hyperparameter Tuning**: Further tuning to optimize performance and efficiency.
- **Experimentation with Boosting Models**: Testing models like Gradient Boosting or XGBoost for potentially improved accuracy.

Data Improvements:

- **Feature Engineering**: Creating new features to enhance predictive power.
- **Data Collection**: Collecting more balanced data to address any class imbalances.

Other Considerations:

- **Explainability Tools**: Tools like SHAP or LIME can be used for enhanced interpretability if required by business stakeholders.