Customer Churn

1202 – Data Analysis Tool Analytics

Project Report

Submitted by: Group 9

Submitted to: Prof. Mizanur Rahman

Contents

1. Introduction
2. Exploratory Data Analysis
3. Data Preprocessing and Feature Engineering
4. Model Training and Evaluation
5. ROC and AUC Analysis
6. Adding Changes
7. Conclusion

Introduction

The goal of this project is to predict whether a customer will churn (leave the bank) or remain with the organization. By leveraging machine learning models to understand customer behavior and identify churn risk.

Methodology

The project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring a systematic approach:

1. **Business Understanding**: Define churn as a significant problem requiring predictive solutions.
2. **Data Understanding**: Explore the dataset to uncover patterns and correlations.
3. **Data Preparation**: Clean, preprocess, and engineer features for machine learning.
4. **Modeling**: Train and evaluate multiple models to identify the most effective one.
5. **Evaluation**: Use classification metrics (accuracy, precision, recall, F1-score, and AUC) to compare model performance.
6. **Deployment**: Propose the best model for real-world application.

Exploratory Data Analysis

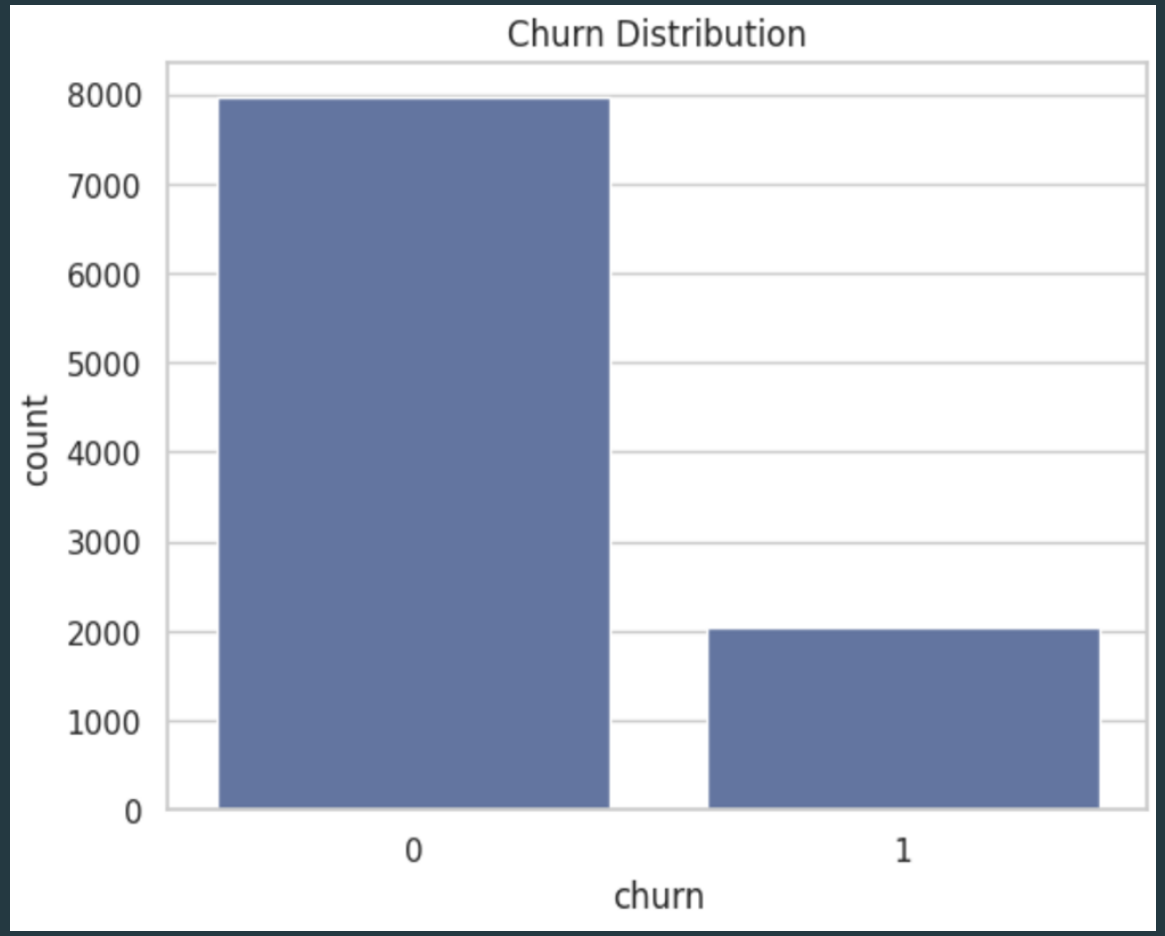
Objective

Understand the distribution and characteristics of the data, identify patterns and trends in customer behavior that correlate with churn, and detect and address potential data quality issues.

Churn Distribution

Approximately 80% of customers are retained (class 0), while only 20% have churned (class 1).

This imbalance is a common challenge in classification problems, as it can lead to models biased toward the majority class. Techniques like SMOTE were later applied to address this issue.



Exploratory Data Analysis (cont.)

Feature Exploration

Age, balance, tenure, and products number are key features related to churn.

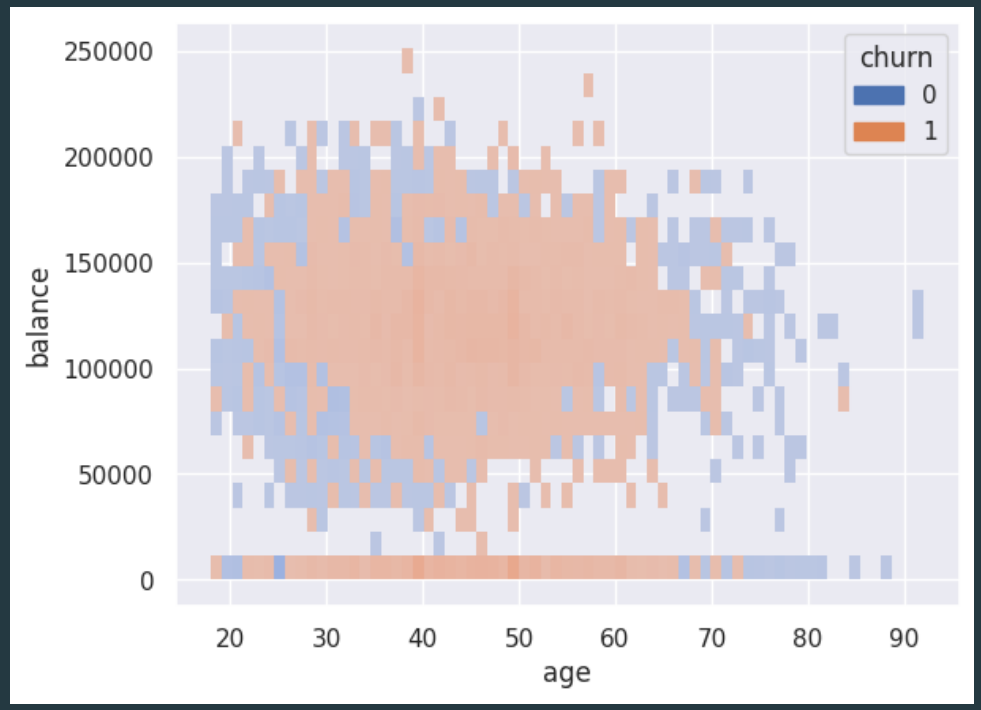
Older customers show a higher likelihood of churn.

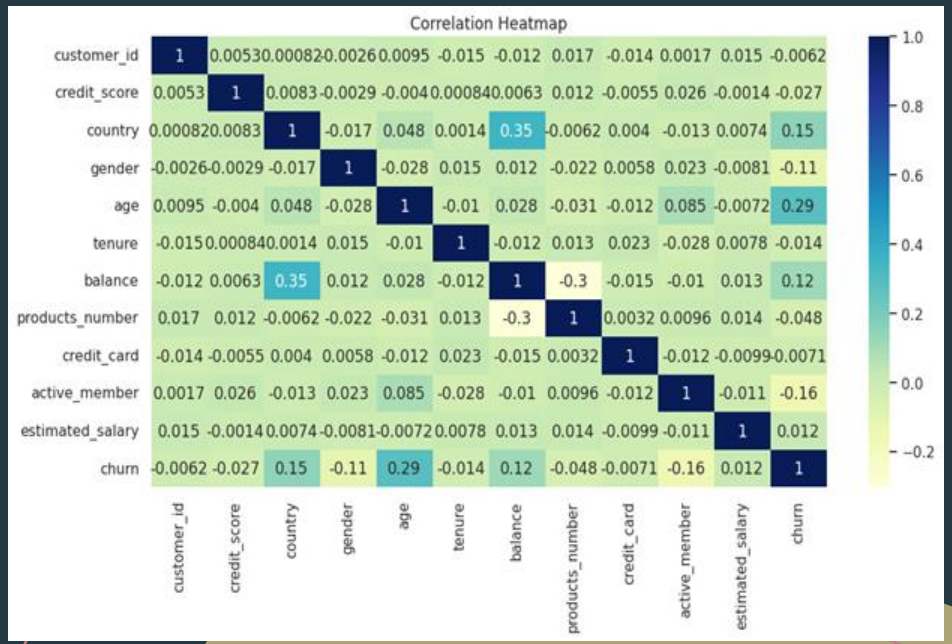
A notable portion of customers has a zero balance. These customers are more likely to churn

Customers with moderate to high balances tend to stay with the bank

Correlation Analysis

Positive correlation between age and churn, negative correlation between balance and churn, and other features showing weak to moderate correlations.





Data Preprocessing and Feature Engineering

Categorical Features

Country and gender were converted into numerical representations.

This allowed the model to differentiate between male and female.

Addressing Class Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the dataset.

After applying SMOTE, the dataset was balanced, ensuring that both classes were equally represented during model training.

Feature Scaling

Numerical features such as (credit score, balance, age, etc.) had different ranges and were standardized using “StandardScaler” to ensure uniformity.

Feature Engineering

New variables were created to enhance the dataset's predictive power, such as a binary activity indicator for customers with a balance of 0.

These transformations added depth to the dataset, allowing models to capture more nuanced patterns in the data.

Data Preprocessing and Feature Engineering (cont.)

Challenges in Preprocessing

- **Class Imbalance**: The application of SMOTE resolved the imbalance but introduced synthetic data, which might not perfectly represent real-world patterns. This trade-off was accepted to improve model performance.
- **Scaling and Interpretability**: While standardization improved model performance, it slightly reduced interpretability for certain models like Logistic Regression, where coefficients directly correspond to feature importance.

Impact of Preprocessing

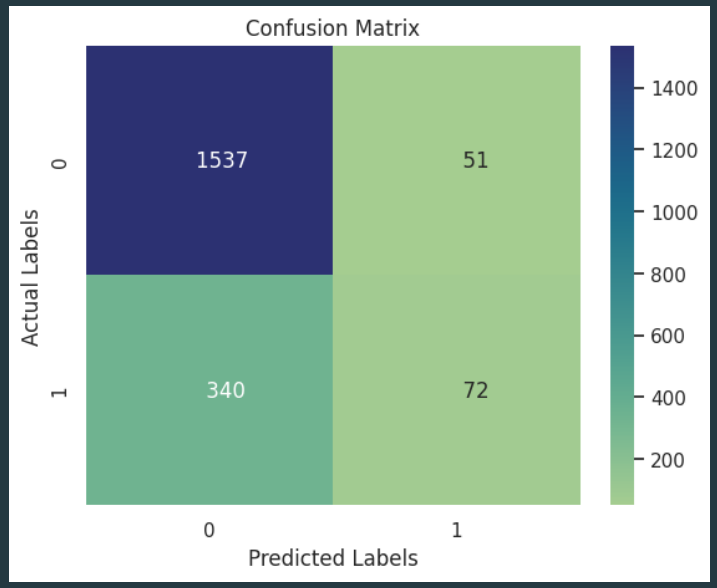
The preprocessing steps significantly enhanced the quality of the dataset, ensuring that it was ready for machine learning:

- The balanced dataset provided fair learning opportunities for both classes.
- Scaled features ensured consistent performance across different algorithms.
- Engineered features added complexity and depth, improving the model's ability to identify patterns.

Model Training & Evaluation

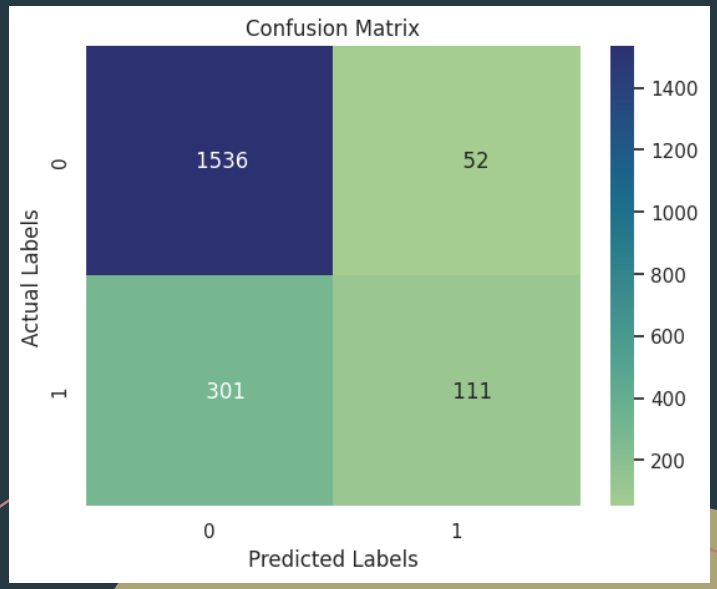
Model 1: Logistic Regression

- Accuracy: 80%
- Precision: 0.59
- Recall: 0.17
- F1-Score: 0.27
- ROC-AUC: 0.75 Logistic Regression served as a baseline model. It performed well for the majority class (non-churn) but struggled with the minority class due to the linear separability assumption.

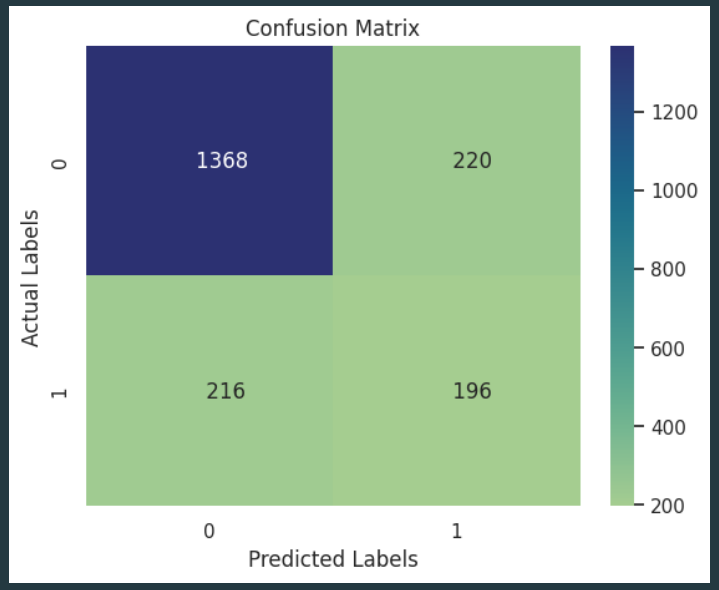


Model 2: Naïve Bayes

- Accuracy: 82%
- Precision: 0.68
- Recall: 0.27
- F1-Score: 0.39
- ROC-AUC: 0.79 Naive Bayes improved recall for the minority class but still lacked sufficient discriminatory power.

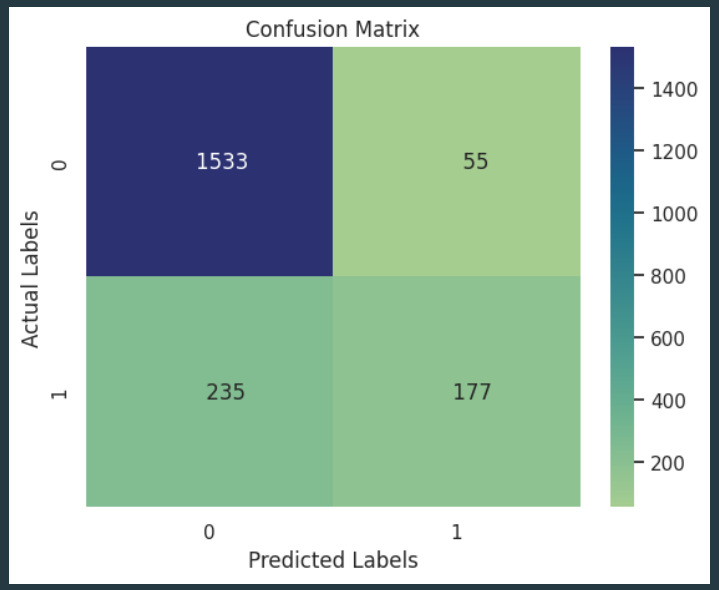


Model Training & Evaluation (cont.)



Model 3: Decision Tree

- Accuracy: 78%
- Precision: 0.47
- Recall: 0.48
- F1-Score: 0.47
- ROC-AUC: 0.68 The Decision Tree model achieved a balanced precision and recall, but its performance suffered from overfitting due to its tendency to capture noise in the training data.



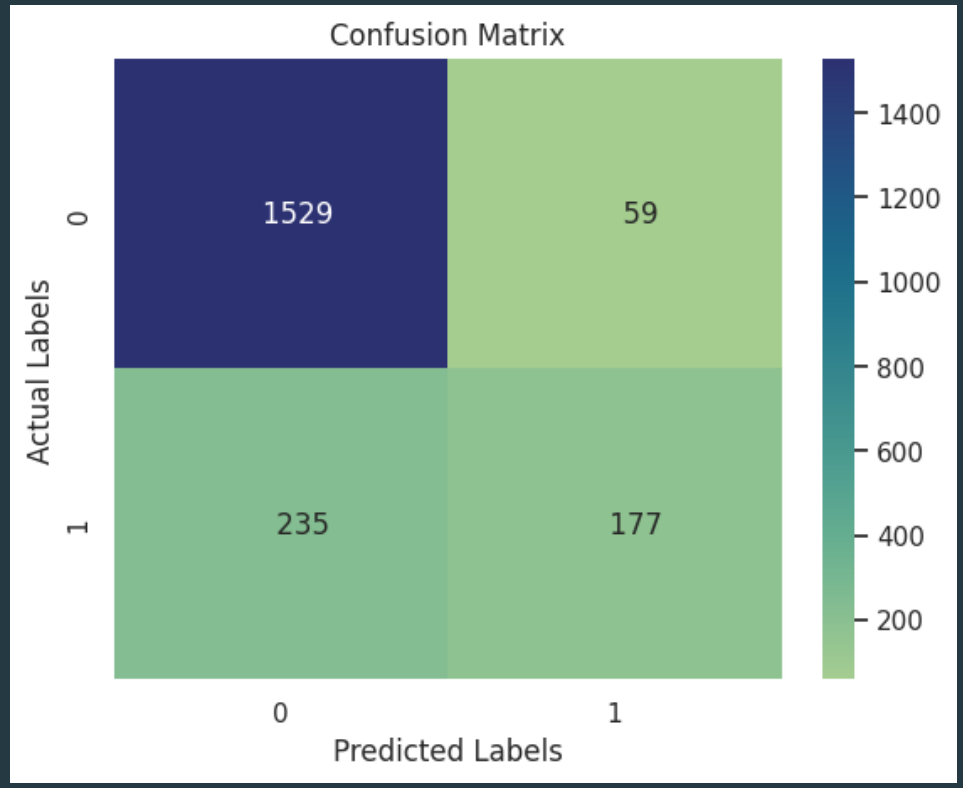
Model 4: Random Forest

- Accuracy: 85%
- Precision: 0.76
- Recall: 0.43
- F1-Score: 0.55
- ROC-AUC: 0.84 Random Forest emerged as one of the top performers, offering a good balance between accuracy, precision, and recall.

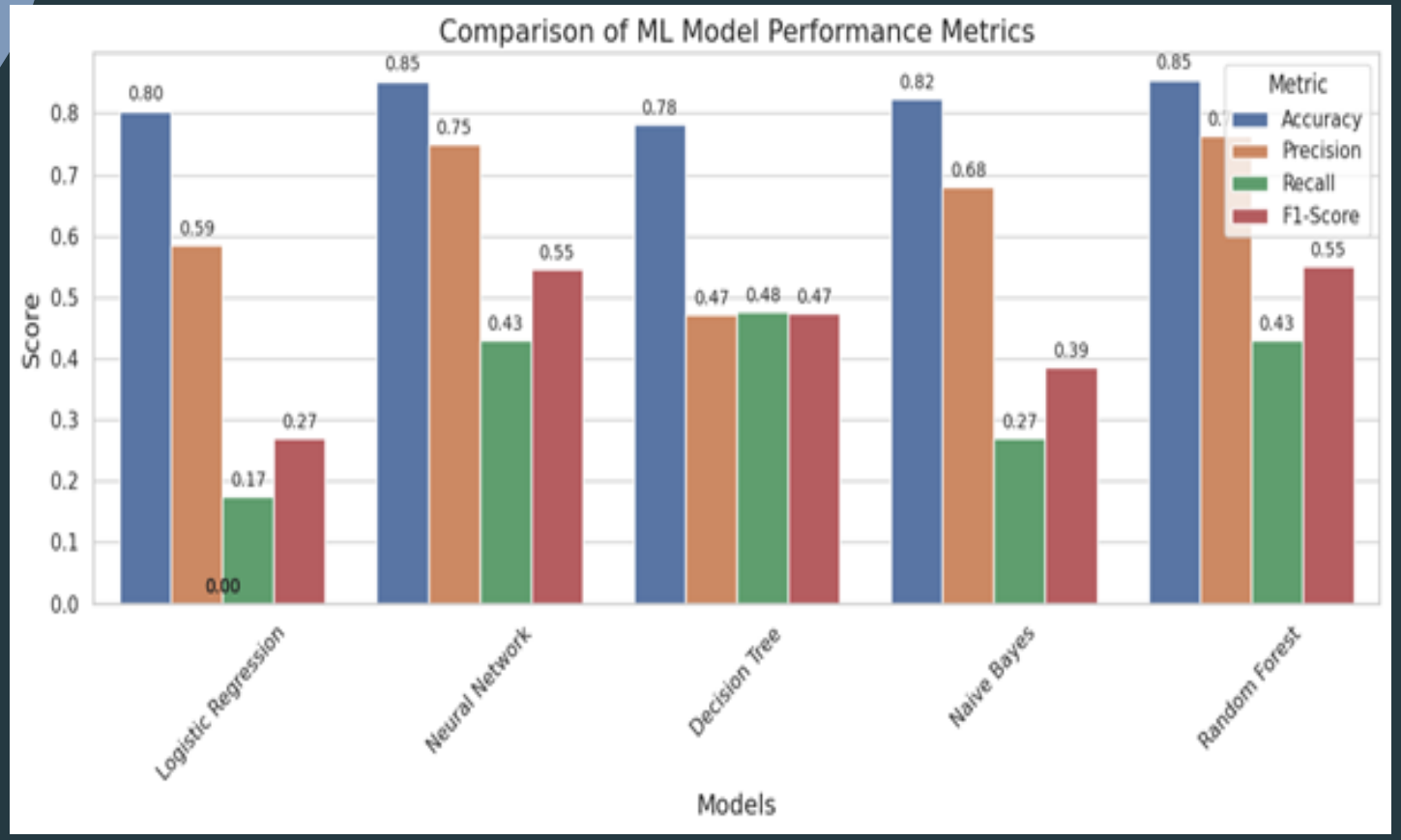
Model Training & Evaluation (cont.)

Model 5: Neutral Network

- Accuracy: 85.3%
- Precision: 0.75
- Recall: 0.43
- F1-Score: 0.55
- ROC-AUC: 0.84 The Neural Network achieved similar results to Random Forest, demonstrating strong generalizability and the ability to capture complex patterns. However, its interpretability was limited compared to simpler models.

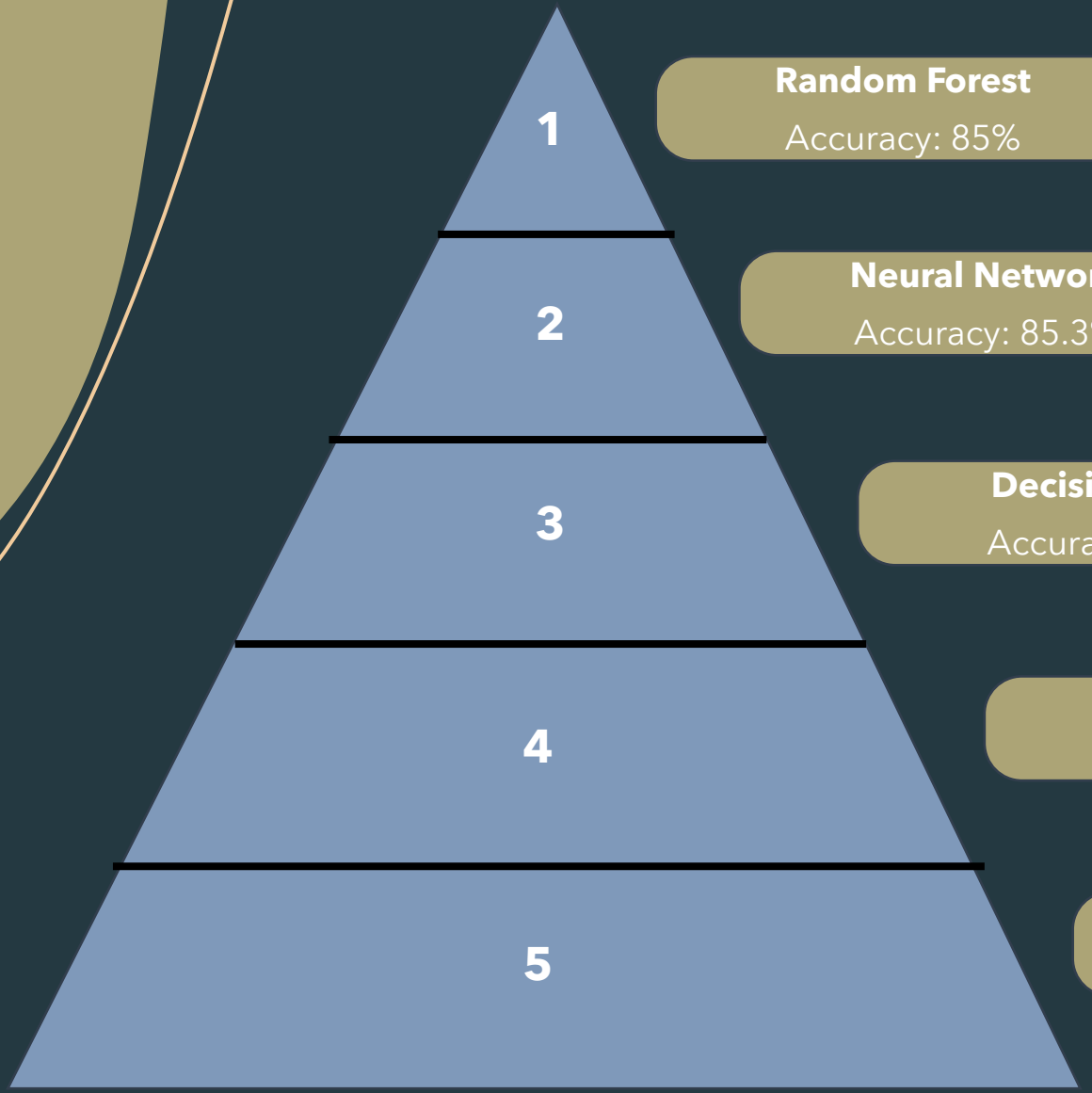


Insights and Comparative Analysis



- **Random Forest** and **Neural Network** emerged as the best-performing models, achieving high accuracy and balanced precision and recall. These models effectively handled class imbalance and non-linear relationships.
- Simpler models like Logistic Regression and Naive Bayes struggled with the minority class, highlighting the need for more sophisticated techniques.
- Decision Tree, while interpretable, was prone to overfitting, limiting its reliability for generalization.

Model Ranking



Random Forest
Accuracy: 85%

Neural Network
Accuracy: 85.3%

Decision Tree
Accuracy: 78%

Naïve Bayes
Accuracy: 82%

Logistic Regression
Accuracy: 80%

ROC and AUC Analysis

Logistic Regression:

- **AUC**: 0.75
- **Insights**: Logistic Regression achieved moderate discriminatory power, indicating that it performed better than random guessing but struggled with accurately predicting churners due to the linearity of the model.

Naive Bayes:

- **AUC**: 0.79
- **Insights**: Naive Bayes showed a slight improvement over Logistic Regression. Its probabilistic nature helped capture some patterns in the data, but it was less effective for the minority class (churn).

Decision Tree:

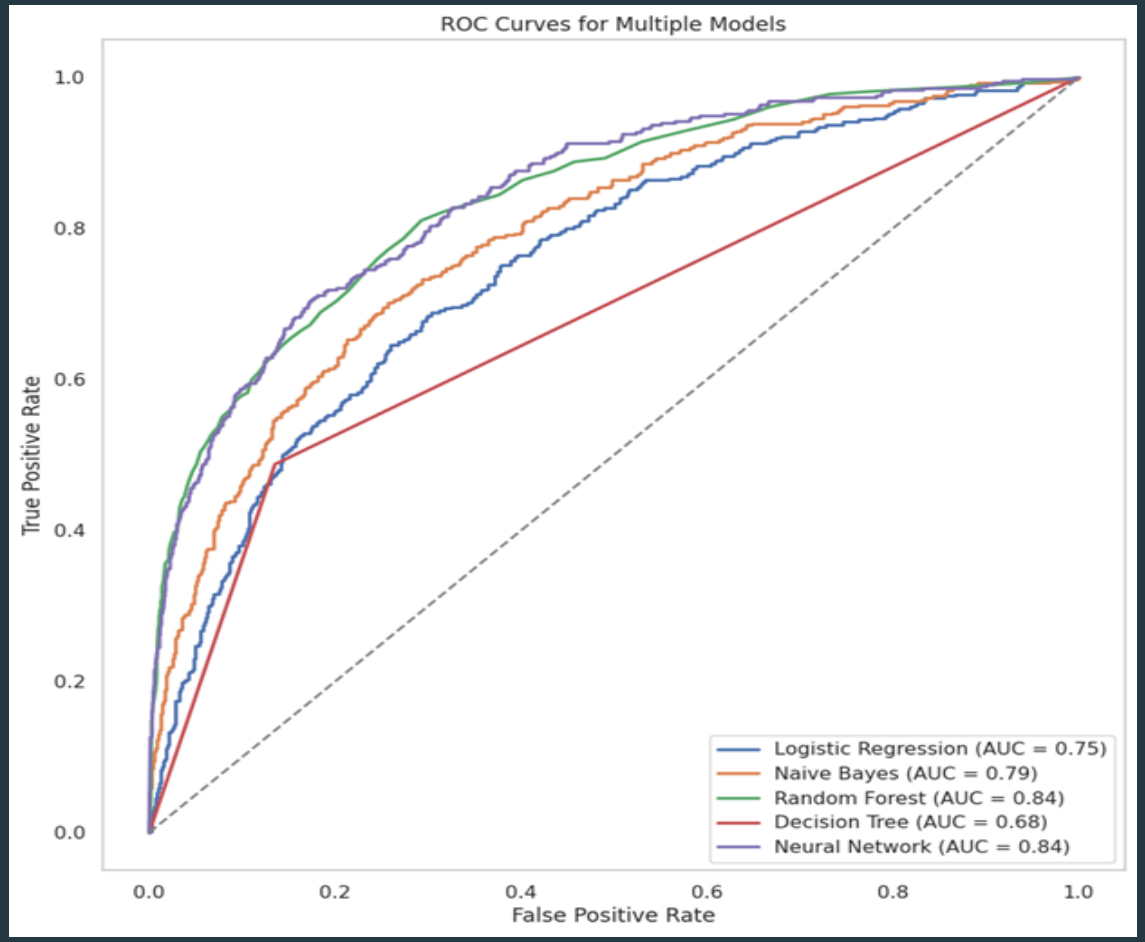
- **AUC**: 0.68
- **Insights**: The Decision Tree model had the lowest AUC score. While it performed well on the training data, overfitting reduced its ability to generalize, leading to a weaker discriminatory ability on the test data.

Random Forest:

- **AUC**: 0.84
- **Insights**: Random Forest was one of the top-performing models, with excellent discriminatory power. Its ensemble nature, which combines multiple decision trees, enabled it to effectively capture complex relationships and avoid overfitting.

Neural Network (Multi-Layer Perceptron):

- **AUC: 0.84**
- **Insights:** The Neural Network matched Random Forest in terms of AUC, demonstrating its strength in capturing intricate patterns. However, its complexity and lack of interpretability limited its usability for business stakeholders.



Interpretation of ROC Curves

The ROC curves highlighted the trade-offs between TPR and FPR:

- Logistic Regression and Naive Bayes showed relatively steep curves at the initial thresholds, indicating better performance for the majority class (non-churn).
- Random Forest and Neural Network had more balanced curves, achieving high TPR with a low FPR across thresholds. This indicates strong performance for both classes.
- The Decision Tree curve demonstrated overfitting, as the model struggled to generalize beyond the training data.

ROC and AUC Analysis

Interpretation of ROC Curves

The ROC curves highlighted the trade-offs between TPR and FPR:

- Logistic Regression and Naive Bayes showed relatively steep curves at the initial thresholds, indicating better performance for the majority class (non-churn).
- Random Forest and Neural Network had more balanced curves, achieving high TPR with a low FPR across thresholds.
- The Decision Tree curve demonstrated overfitting, as the model struggled to generalize beyond the training data.

Challenges in ROC Analysis

- Imbalance in the Dataset: Despite using SMOTE, the original imbalance in the dataset influenced the ROC curves. Models like Logistic Regression and Naive Bayes showed weaker performance due to their inherent limitations in handling imbalance.
- Threshold Selection: While ROC provides a visual tool for threshold selection, it requires careful consideration of business priorities (e.g., prioritizing recall vs. precision).

Insights from ROC and AUC

- Random Forest and Neural Network demonstrated the best ability to differentiate between churners and non-churners, as reflected by their high AUC scores (0.84).
- While Logistic Regression and Naive Bayes provided moderate discriminatory power, their performance was not sufficient for practical deployment in a highly imbalanced setting.
- Decision Tree was the least effective, with an AUC of 0.68, primarily due to overfitting.

Changes/Tuning and Feature Engineering

Addressing Class Imbalance with SMOTE

•**Challenge**: Imbalanced target variable (20% churners).

•**Change**: Applied SMOTE to generate synthetic data for the minority class.

•**Impact**: Improved recall and F1-scores for the minority class, especially in Random Forest and Neural Network models.

•**Rationale**: SMOTE prevents overfitting and maintains full majority class data

Feature Engineering

•**Change:** Created binary activity indicator, added interaction terms (e.g., products_number * active_member), and normalized features.

•**Impact:** Improved model performance by capturing complex patterns.

•**Rationale:** Provides deeper insights into relationships between variables for better predictions.

Hyperparameter Tuning

•**Change**:

- **Random Forest**: Increased trees (n_estimators), limited depth (max_depth), adjusted leaf samples (min_samples_leaf).
- **Neural Network**: Modified architecture, fine-tuned learning rate, increased iterations (max_iter).

•**Impact**: Enhanced precision, recall, and F1-scores; AUC value of 0.84.

•**Rationale**: Optimizing hyperparameters prevents overfitting and underperformance.

Conclusion

- The Bank Customer Churn Prediction project successfully leveraged machine learning to address the critical issue of customer churn. Through structured data analysis, preprocessing, and model training, the project identified Random Forest and Neural Network as the best-performing models, achieving high accuracy (85%+) and balanced metrics.
- Key insights revealed age, balance, and active membership as strong predictors of churn, providing actionable strategies for targeted retention efforts, such as loyalty programs and engagement campaigns. Addressing challenges like class imbalance with SMOTE and refining models through hyperparameter tuning ensured robust and reliable predictions.
- Future improvements include incorporating additional data sources, exploring advanced ensemble methods, and deploying the model for real-time churn prediction. This project demonstrates the potential of data-driven solutions to improve customer retention and enhance profitability, positioning the bank for sustainable success.