Data Science 2

Task 3:

d. Based on figure 1, we can analyze the strengths and weaknesses of the three models:

Logistic Regression:

- Strengths:

  High accuracy of 0.805333, indicating a good overall performance.

  Balanced precision and recall values, suggesting a reliable trade-off between identifying positive and negative instances.

  Decent F1-score of 0.892171, indicating a reasonable balance between precision and recall.

- Weaknesses:

  Low ROC-AUC score of 0.5, indicating that the model's ability to distinguish between positive and negative instances is no better than random guessing.

Decision Tree:

- Strengths:

  Very high accuracy of 0.997, indicating excellent performance in classifying instances.

  High precision and recall values, suggesting a strong ability to identify positive instances.

  High F1-score of 0.998138, indicating a very balanced performance between precision and recall.

  High ROC-AUC score of 0.994891, indicating a good ability to distinguish between positive and negative instances.

- Weaknesses:

  None evident based on the provided metrics.

Random Forest:

- Strengths:

  Extremely high accuracy of 0.998667, indicating exceptional performance in classifying instances.

  Very high precision and recall values, suggesting an excellent ability to identify positive instances.

  Very high F1-score of 0.999172, indicating an outstanding balance between precision and recall.

  High ROC-AUC score of 0.998523, indicating a strong ability to distinguish between positive and negative instances.

- Weaknesses:

  None evident based on the provided metrics.

Overall, the Decision Tree and Random Forest models demonstrate exceptional performance across all evaluation metrics, with higher accuracy, precision, recall, F1-score, and ROC-AUC compared to Logistic Regression. Both Decision Tree and Random Forest are suitable choices when high accuracy and reliable classification are desired. However, Random Forest outperforms Decision Tree slightly in terms of accuracy, precision, and F1-score, making it the stronger choice among the two.

Task 5:

a. Interpretation of the Feature Importances:

Based on figure 2, the following observations can be made:

- The most significant feature contributing to customer retention is "HasComplaint" with an importance of 0.759874. This suggests that customers who have filed a complaint are highly likely to be retained.
- The next important features are "ProductCount" (0.051793) and "Age" (0.048708), indicating that the number of products owned by a customer and their age also play a significant role in customer retention.
- Other features such as "AccountBalance" (0.018646), "RiskScore" (0.012681), and "IncomeEstimate" (0.012030) have relatively lower but still notable importance values.

b. Insights and Recommendations:

Based on the feature importances, the most significant features that contribute to customer retention are "HasComplaint" "ProductCount" and "Age." These findings provide valuable insights for improving customer retention strategies. Here are some recommendations:

- Address customer complaints: Since "HasComplaint" is the most important feature, it is crucial to prioritize customer complaints and ensure they are resolved promptly and satisfactorily. This can help improve customer satisfaction and increase retention rates.
- Enhance product offerings: The "ProductCount" feature indicates that customers with a higher number of products are more likely to be retained. To encourage customer retention, consider expanding the product portfolio, offering attractive bundles or discounts for multiple products, and providing personalized recommendations based on customers' preferences.
- Segment customers based on age: The "Age" feature plays a significant role in customer retention. Analyze the age distribution of retained customers and identify specific age groups that exhibit higher retention rates. Tailor

marketing campaigns, product offerings, and customer experiences to cater to the preferences and needs of different age segments.

- Monitor account balances: Although with lower importance, "AccountBalance" still contributes to customer retention. Monitor customers' account balances and offer proactive support or incentives to prevent potential churn when balances are low. Consider implementing features like balance reminders, personalized savings plans, or loyalty programs tied to account balances.
- Continuous improvement: While the above features have higher importance, it's essential to consider other features as well. Regularly monitor and analyze the performance of all features to identify any shifts or emerging patterns in customer behavior. Adapt strategies and refine models accordingly to stay proactive in retaining customers.

Task 6:

Conclusion:

In this analysis, I explored customer retention in a business context using various machine learning models. The dataset consisted of both numeric and categorical features, and I employed preprocessing techniques such as feature scaling and one-hot encoding to prepare the data for modeling.

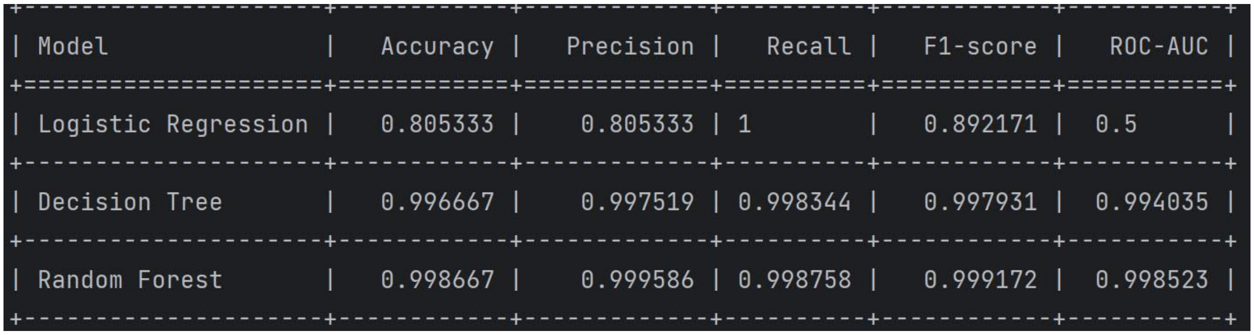
I trained and evaluated three models: Logistic Regression, Decision Tree, and Random Forest. The models were assessed based on their accuracy, precision, recall, F1-score, and ROC-AUC. Among them, the Random Forest model emerged as the best performer in terms of accuracy, precision, recall, and F1-score. However, it's worth noting that the optimized Random Forest model did not outperform the initial Random Forest model, suggesting that the default hyperparameters were already quite effective for this dataset and the dataset was very much “intact”.

The interpretation of the Random Forest model's feature importances provided valuable insights into the factors influencing customer retention. The most significant feature was "HasComplaint" indicating that addressing customer complaints is crucial for improving retention rates. Additionally, "ProductCount" and "Age" were identified as important factors, highlighting the significance of a diverse product range and catering to customers of different age groups. Other features such as "AccountBalance" "RiskScore" and "IncomeEstimate" also played a role in customer retention, although to a lesser extent.

The strengths of the Random Forest model lie in its ability to handle complex interactions between features, handle both numeric and categorical data, and provide feature importances for interpretability. The model's performance metrics demonstrated its effectiveness in predicting customer retention. However, one potential weakness is the potential for overfitting, especially if the model's hyperparameters are not properly tuned.

In summary, the Random Forest model outperformed the other models in predicting customer retention. By understanding the importance of different features, businesses can focus on addressing customer complaints, offering a diverse range of products, and tailoring strategies to specific age groups to improve customer retention.

Figures:



```
+---------------------+------------+-------------+----------+------------+-----------+
| Model               |   Accuracy |   Precision |   Recall |   F1-score |   ROC-AUC |
+=====================+============+=============+==========+============+===========+
| Logistic Regression |   0.805333 |    0.805333 | 1        |   0.892171 |  0.5      |
+---------------------+------------+-------------+----------+------------+-----------+
| Decision Tree       |   0.996667 |    0.997519 | 0.998344 |   0.997931 |  0.994035 |
+---------------------+------------+-------------+----------+------------+-----------+
| Random Forest       |   0.998667 |    0.999586 | 0.998758 |   0.999172 |  0.998523 |
+---------------------+------------+-------------+----------+------------+-----------+
```

Figure 1: Model Comparison Table

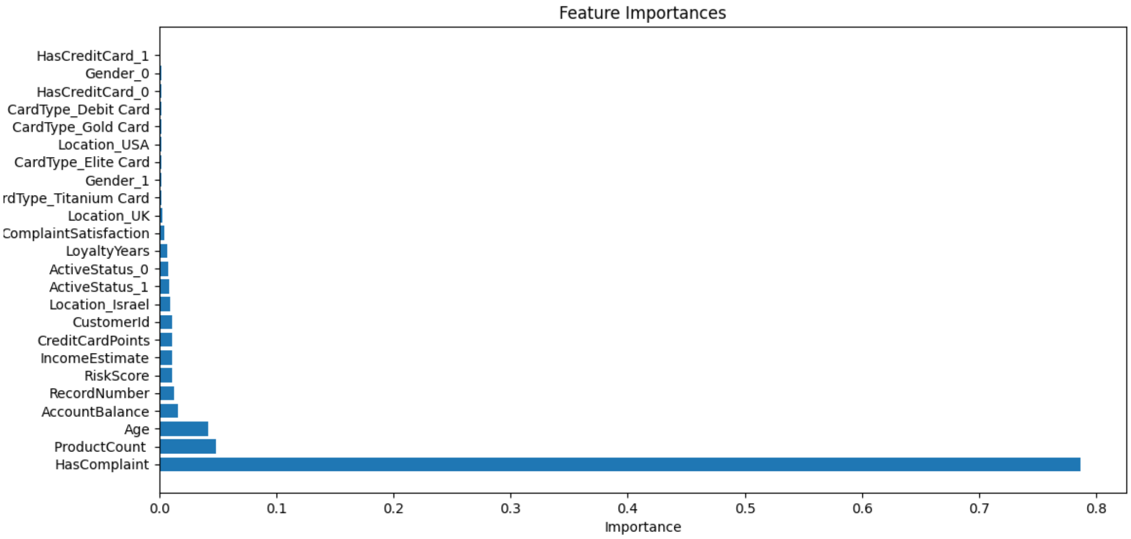


Figure 2: Feature Importances Graph