CAPSTONE PROJECT 2

Analysis of Bank Customer Churn of Bailey's Bank

Thomas Sheridan

INTRODUCTION

In an era where customer loyalty is paramount, Baileys Bank faces a pivotal challenge with a 21% customer turnover rate. This high churn rate not only threatens potential revenue loss but also undermines the long-term sustainability and reputation of the bank. Understanding and effectively managing customer churn is crucial for Baileys Bank to maintain financial stability and reinforce its market position.

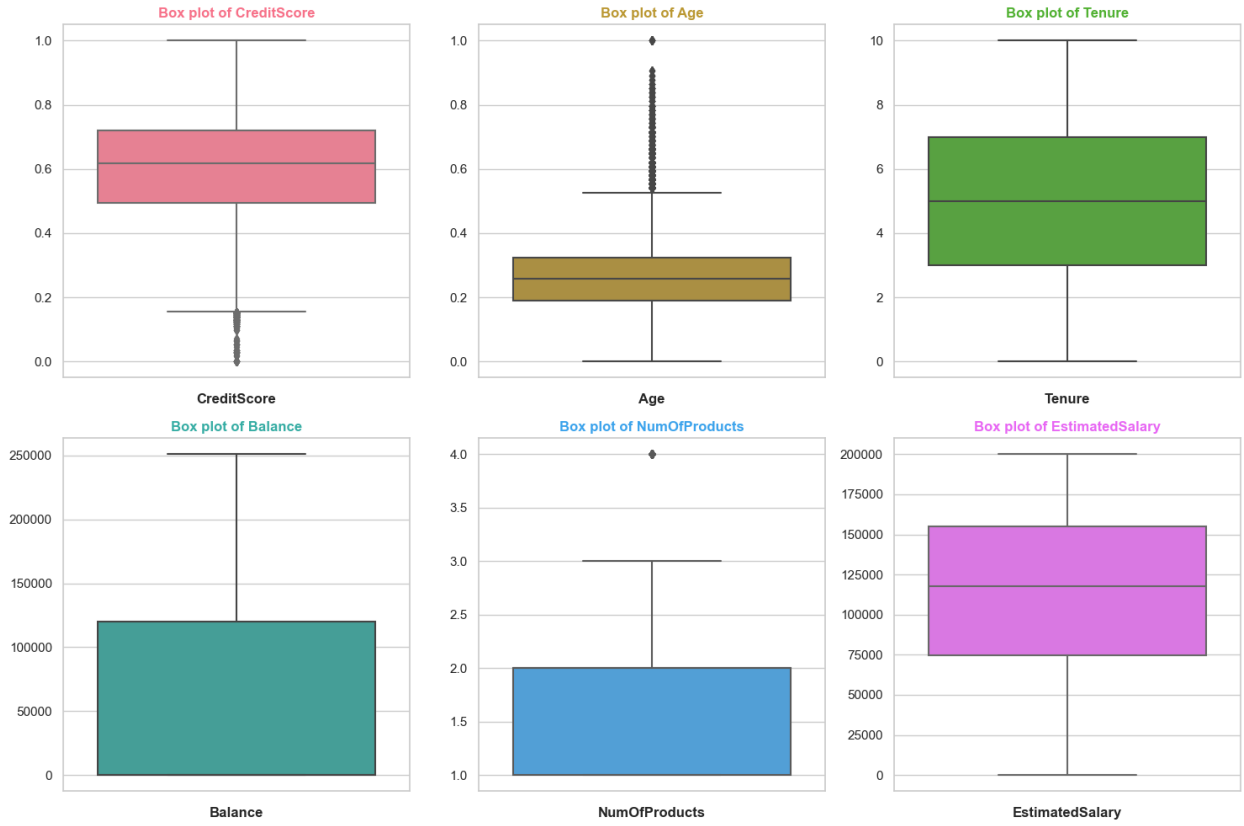
This project report outlines the comprehensive analysis conducted on customer data to understand the underlying reasons for customer departures and the factors influencing their decisions to leave. This report aims to provide practical insights that will help Baileys Bank improve customer retention and reduce the churn rate to less than 10% by the start of the next financial quarter by using advanced data analysis techniques.

The objectives of this analysis are twofold: to identify key patterns and triggers of customer churn at Baileys Bank and to develop targeted interventions to keep customers at risk of departure. This initiative involves collaboration among various stakeholders within the bank, including the CEO, CFO, CTO, and Customer Service Managers, ensuring a united front in enhancing service quality and customer satisfaction.

This report will help Baileys Bank improve customer retention and enhance its reputation as a customer-focused and innovative financial institution by analyzing data and engaging with customers.

Data Wrangling

- Removing Irrelevant Columns: You removed several columns that were not relevant to the analysis.
- I also dropped the 'id' column because it is a unique identifier that does not contribute to the analysis of churn.
- I removed the 'left_digit' because of rounding issues.
- We dropped the 'Surname' column to de-identify the data because it was not useful for the churn analysis.
- Even when just wrangling, the age factor of churn shows itself.



1. Credit Score (Pink Box Plot):

- Shows the distribution of customer credit scores.
- Most scores are centrally clustered, with a few outliers on the lower end.

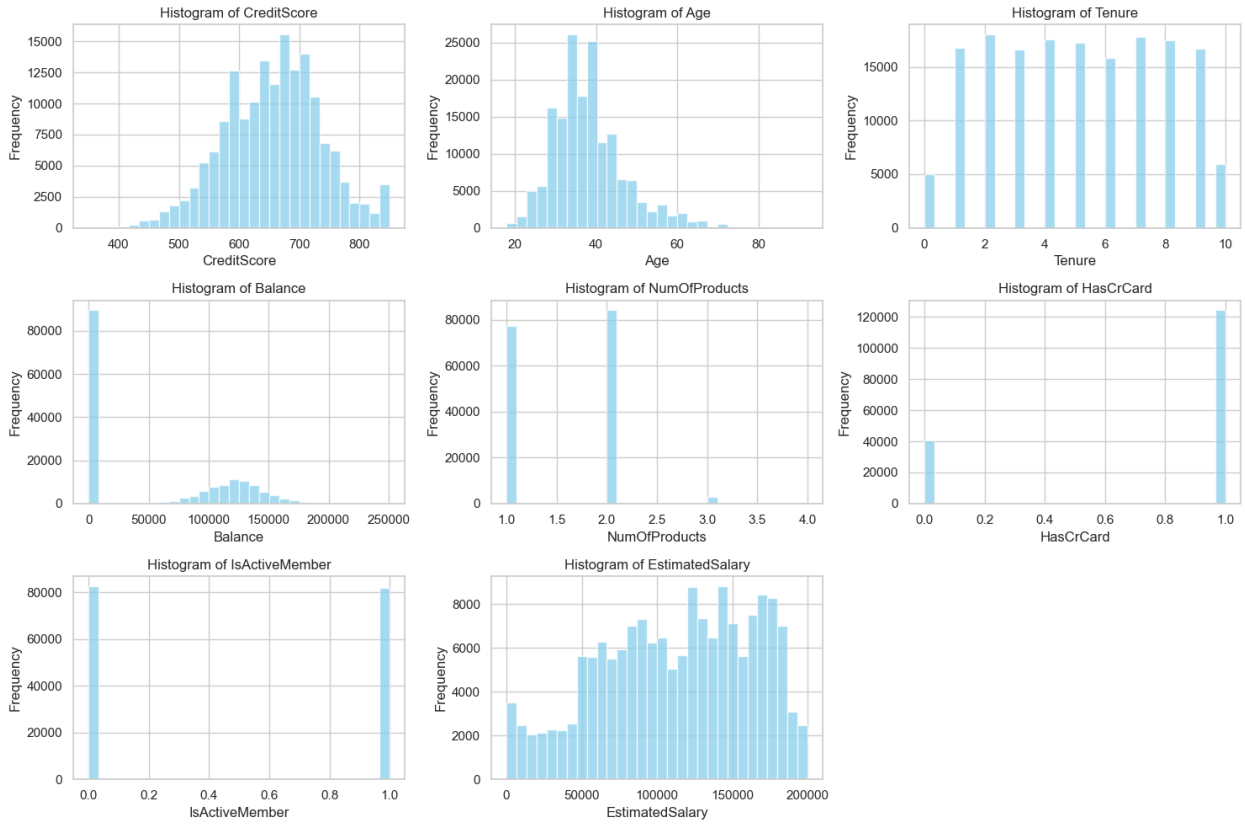
2. Age (Yellow Box Plot):
   - Displays the age distribution of customers.
   - The data has a right-skewed distribution, showing a younger customer base with some older outliers.
3. Tenure (Green Box Plot):
   - Represents the distribution of the number of years customers have been with the bank.
   - Data is symmetric around the median, showing a uniform distribution of customer tenure.
4. Balance (Cyan Box Plot):
   - Shows the balance amount in customers' accounts.
   - Most customers have a significant amount of money in their accounts, with a few having much higher balances, as shown by the upper whisker.
5. Number Of Products (Blue Box Plot):

- Illustrates the number of products customers use.
- Most customers use 1 or 2 products, with very few using 3 products, and outliers using 4 products.

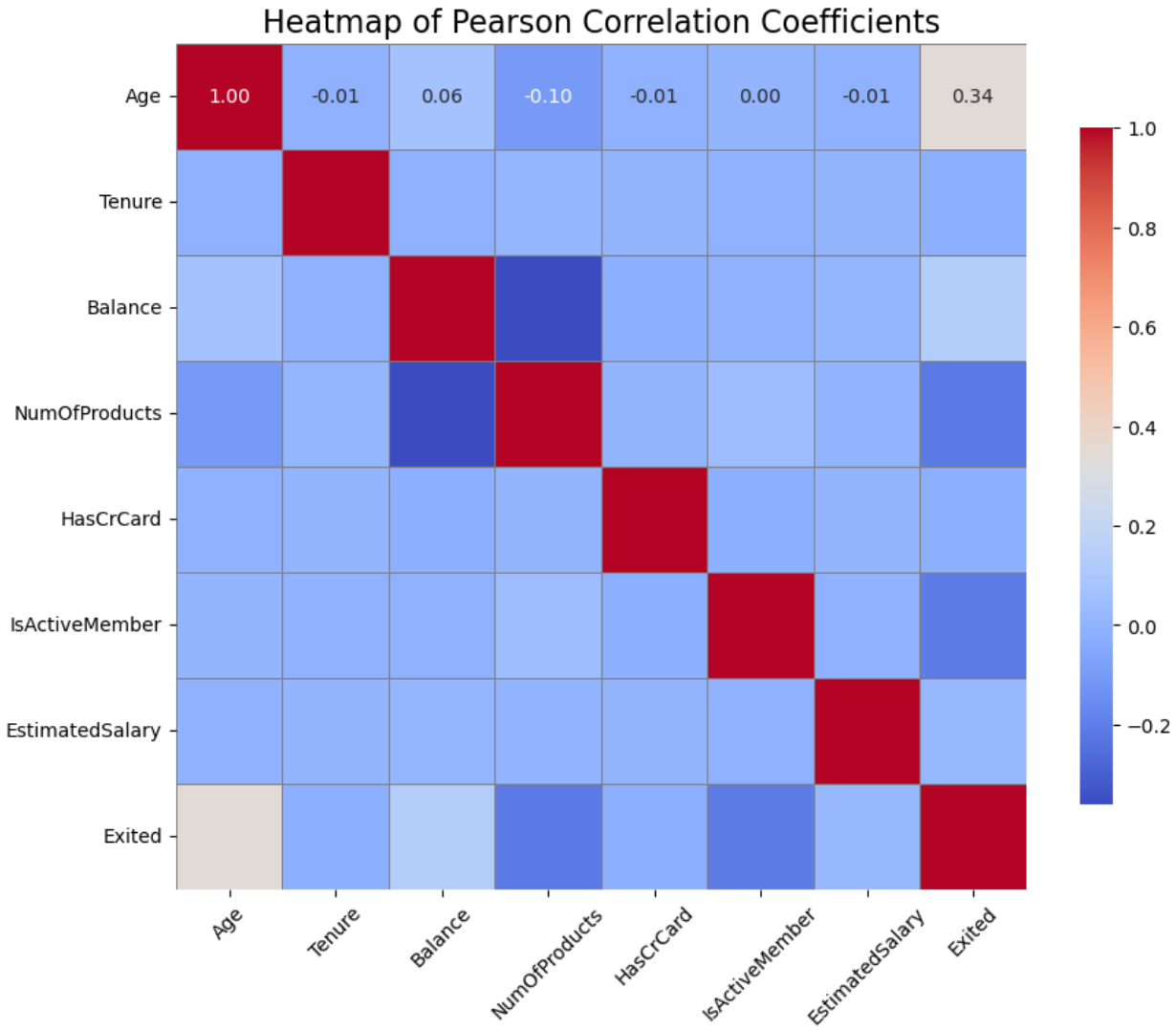
6. Estimated Salary (Purple Box Plot):
    - Displays the estimated salary of customers.
    - Salary distribution is uniform across the range, showing no significant skew.

Key insights gleaned from the bank's customer data analysis provide a comprehensive understanding of the customer base. A significant majority of customers maintain good credit scores, although there remains a smaller fraction with lower scores, indicating a diversity in financial health. The age distribution shows a predominance of younger customers, with an appreciable representation of older individuals, suggesting a wide demographic reach. In terms of tenure, customers exhibit an even spread across various durations of association with the bank, pointing to both new and long-term customer relationships. Financially, there is a notable range in account balances, predominantly skewed towards higher values, which may reflect a wealthier clientele. Regarding banking products, most customers tend to hold between one and two products, indicating a preference for simplicity or specific product loyalty. Lastly, the variation in estimated salaries is considerable, spanning a broad spectrum that does not favor any particular salary bracket, highlighting the bank's appeal across different income levels. These insights are pivotal for tailoring customer retention and acquisition strategies.

Data Exploration



- This bank customer dataset reveals several interesting patterns in the distribution of key features. The histogram of Credit Score shows a roughly normal distribution centered around 650 to 700, showing that most customers have a mid-range credit score.
- The Age histogram is right-skewed, with most customers falling between 30 and 50 years old. Tenure appears evenly distributed, suggesting that customers have a wide range of account longevity.
- The Balance histogram displays a notable number of customers with zero balance and a concentration around 100,000 for those with balances.
- The Number of products histograms show most customers have one or two products, with very few having more. A significant number of customers in the.
- The Is Active Member histogram reveals an almost even split between active and inactive members. Finally, the Estimated Salary histogram suggests a uniform distribution across various salary ranges, showing diverse income levels among the customers.
- The estimated salary histogram suggests a uniform distribution across various salary ranges, showing diverse income levels among the customers.



1. **Age**: The most notable correlation is between Age and Exited (0.34). This positive correlation suggests that older customers are more likely to churn, making age a significant factor in predicting customer churn.
2. **Tenure**: Tenure shows a near-zero correlation with Exited (-0.01), showing that the time a customer has been with the bank does not significantly influence their likelihood of churning.
3. **Balance**: Balance has a small positive correlation with Exited (0.06), suggesting that customers with higher balances might have a slightly higher tendency to leave the bank, though this relationship is weak.
4. **Number Of Products**: This feature shows a slight negative correlation with Exited (-0.10), implying that customers with more products are less likely to churn, albeit the relationship is weak.
5. **Has Credit Card**: This feature has a negligible correlation with Exited (0.00), showing that whether a customer has a credit card does not affect their churn probability.

6. **Is Active Member**: The correlation between IsActiveMember and Exited is negative (-0.01), but very close to zero, suggesting that being an active member has a minimal impact on churn likelihood.
7. **Estimated Salary**: There is a slight positive correlation between EstimatedSalary and Exited (0.01), showing that higher salaries do not significantly affect churn probability.

In summary, age stands out as the most significant predictor of customer churn in this dataset. The other features have either weak or negligible correlations with churn, highlighting the importance of focusing on age when developing strategies to mitigate churn in the bank's customer base.

Analysis of Variance (ANOVA) on Age by Geography

To explore whether geography has a significant effect on the age of customers, we performed an ANOVA test. The results show a very low P-value (4.06e-320), showing that the geographical category (Geography) has a statistically significant effect on age. This implies that the differences in customer ages across various geographical regions are unlikely because of random chance. This finding suggests that age demographics differ significantly between regions, which could influence customer behavior and churn rates. Understanding these geographical differences can help the bank tailor its strategies to different regions more effectively.

Logistic Regression for Predicting Churn

We used logistic regression to model the probability of customer churn (Exited) based on Age and Credit Score. The coefficients for Age (0.091361) and Credit Score (-0.000847) show that as age increases, the likelihood of churn also increases, whereas a higher credit score slightly decreases the likelihood of churn. The accuracy of the logistic regression model is approximately 78.46%, with a confusion matrix showing:

Random Forest Classifier for Improved Prediction

To improve model performance, we applied a Random Forest Classifier, which generally performs better with complex datasets. After handling categorical variables, scaling the features, and splitting the dataset into training and test sets, the Random Forest model achieved an accuracy of 86%. The confusion matrix and classification report for the Random Forest model are:

- **Confusion Matrix:**
  - True Negatives: 24,714
  - False Positives: 1,346
  - False Negatives: 3,217
  - True Positives: 3,724
- **Classification Report:**

    - Precision, Recall, and F1-Score for each class show the model performs well in identifying non-churners but is less effective in identifying churners.
    - Overall accuracy: 86%
    - Macro average F1-Score: 0.77.

The Random Forest model shows improved performance compared to the logistic regression model, particularly in balancing the precision and recall for both classes. This suggests that Random Forest is more effective for this dataset, likely because of its ability to handle non-linear relationships and interactions between features.

In summary, the EDA and model evaluations reveal significant insights into the factors influencing customer churn:

- Age is a significant predictor of churn, with older customers being more likely to leave.
- Geography significantly affects age distribution, which may contribute to regional differences in churn rates.
- While logistic regression provides a baseline model, Random Forest offers improved predictive performance, making it a more suitable choice for this dataset.

These insights can guide targeted strategies to reduce churn, such as focusing on retaining older customers and tailoring approaches based on geographical differences.

True Negatives (TN): 24,714

• These are the customers who exited (true label 0) and were correctly predicted by the model as exited (predicted label 0).

• False Positives: 1,346

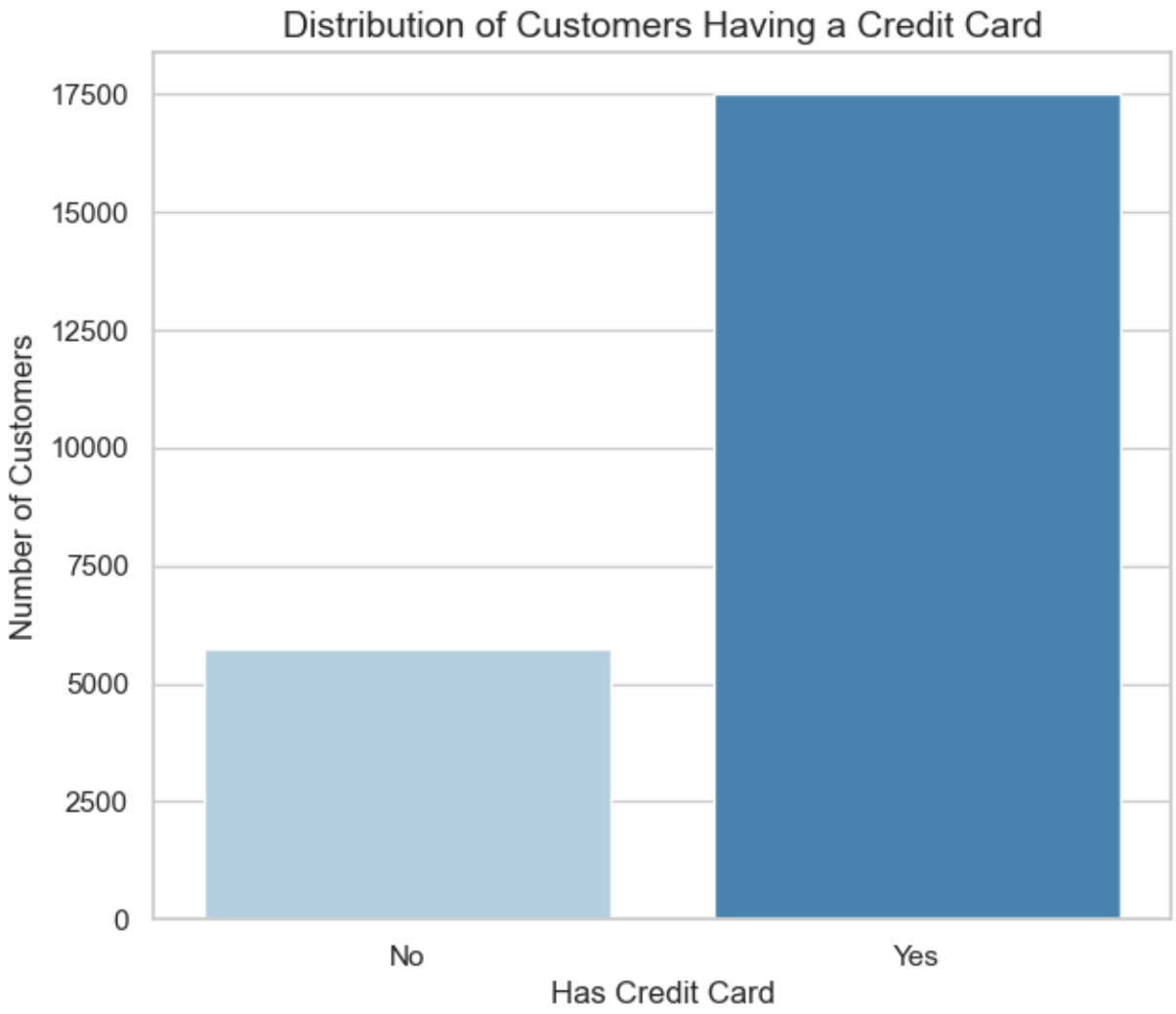
• These are the customers who exited (true label 0) but were incorrectly predicted by the model as stayed (predicted label 1).

• False Negatives (FN): 3,217

• These are the customers who stayed (true label 1) but were incorrectly predicted by the model as exited (predicted label 0).

• True Positives (TP): True Positives: 3,724

The model correctly identifies non-churners more accurately than churners.



Product Penetration: The high number of credit card owners shows strong product penetration. It indicates that the credit card offering is popular among the bank's customers, which could be due to various factors such as rewards, interest rates, or marketing strategies.

Customer Engagement: Credit card owners might be more engaged with the bank's ecosystem, using more of its services. This can increase the 'stickiness' factor, meaning customers are less likely to churn because they are accustomed to and rely on the bank's services.

Revenue Implications: Credit cards are a significant source of revenue for banks through interest, fees, and transaction charges. Customers with credit cards are valuable to the bank, and losing them can have a notable impact on revenue. Thus, understanding credit card holders' behavior is critical to mitigating churn.

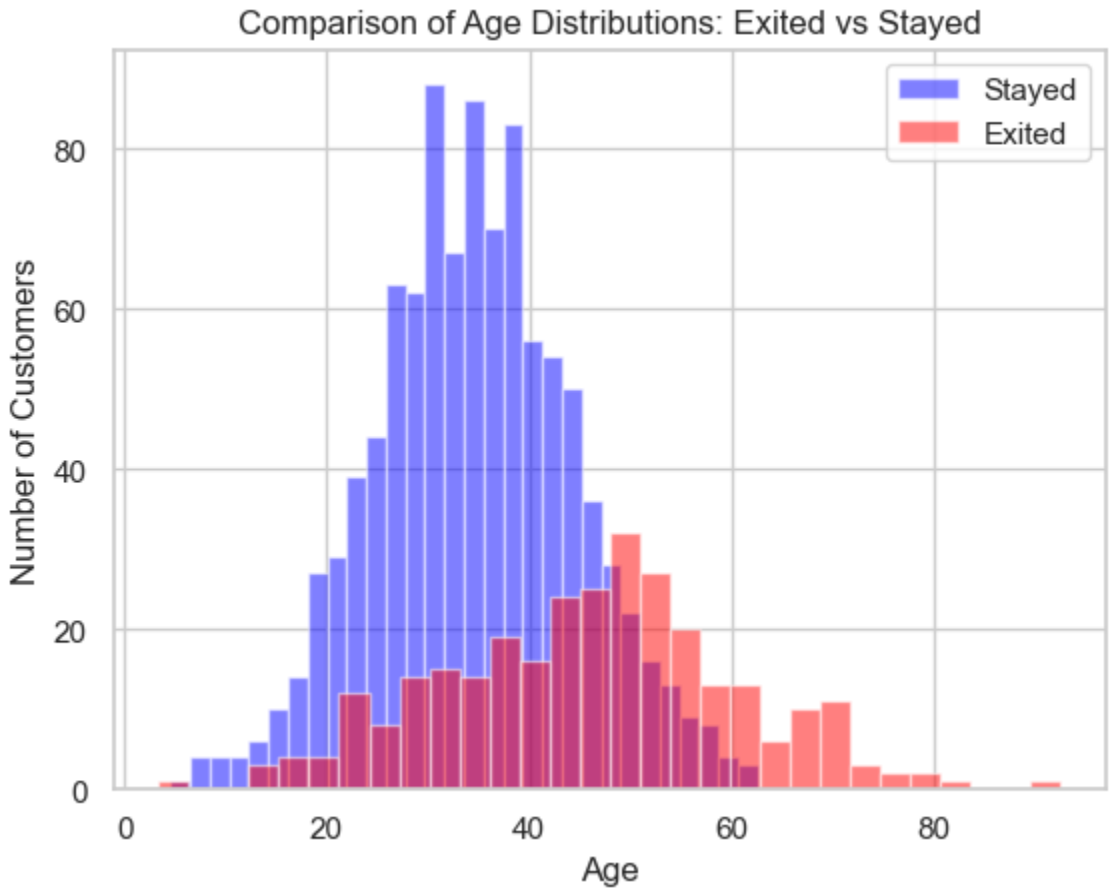
Customer Loyalty: The availability of a credit card might foster a sense of loyalty, especially if the card is linked with a rewards program. Loyal customers are less likely to churn, and banks often have specific retention strategies for them.

Risk Assessment: There could be a risk factor associated with credit cards. If a large portion of the bank's revenue depends on them, losing credit card customers could pose a financial risk. Churn models can help predict which credit card customers are at risk of leaving.

Churn Prevention Strategies: By analyzing the churn rate specifically within the group of customers who have credit cards versus those who do not, the bank can tailor its retention strategies. For instance, if credit card holders are churning at a higher rate, the bank might need to review its credit card policies, fees, or rewards program.

Cross-Selling Opportunities: There's also a cross-selling angle. Customers without a credit card could be potential targets for this product, but the strategy should be different from retention strategies for existing cardholders.

Customer Service and Satisfaction: Since credit card holders are a major part of the customer base, ensuring high satisfaction in this segment can be pivotal in reducing churn. Feedback and support services for these customers can be a focal point for improvement.



1. **Overall Age Distribution:**
    - The majority of customers fall between the ages of 20 and 60, with a noticeable peak around the 30-40 age range.
2. **Exited Customers:**

- The red bars show the distribution of customers who have churned. There is a significant number of churned customers in the 40-60 age range. This indicates that older customers are more likely to exit.
- The peak of churned customers is around the age of 50, suggesting that middle-aged to older customers have a higher propensity to leave the bank.

3. **Stayed Customers:**
    - The blue bars represent the distribution of customers who have stayed. There is a higher concentration of staying customers in the 20-40 age range.
    - The peak of staying customers is around the age of 30, indicating that younger customers are more likely to remain with the bank.
4. **Comparison:**
    - There is a clear shift in the distribution between exited and stayed customers. While younger customers (20-40) predominantly stay, older customers (40-60) show a higher tendency to churn.
    - The overlap in the middle (around ages 35-45) suggests a transitional age range where the likelihood of churning increases.

Targeted Retention Efforts:

- Given the higher churn rates among older customers, the bank should consider targeted retention strategies for customers in the 40-60 age range. This could include personalized offers, enhanced customer service, or loyalty programs tailored to their needs.

5. **Youth Engagement Programs**:
    - To sustain the loyalty of younger customers, the bank could develop programs and products that cater to the interests and needs of the 20-40 age group. This group of people seems more inclined to remain, but it would be helpful to actively engage with them to ensure their satisfaction.
6. **Understanding Churn Drivers:**
    - We need to conduct further analysis to understand the specific reasons older customers are more likely to leave. Factors such as changes in financial needs, dissatisfaction with services, or better offers from competitors should be investigated.
7. **Product Development:**
    - The bank might consider developing or enhancing products that cater specifically to the financial goals and concerns of the middle-aged demographic, potentially reducing churn rates in this group.

Model Implementation and Evaluation

In the bank customer churn prediction project, various machine learning models were evaluated to determine their efficacy in identifying potential churn. Logistic Regression, used as the baseline, achieved modest accuracy, laying the groundwork for comparison with more complex models. The Random Forest model, enhanced through GridSearchCV for optimal hyperparameters like tree number and depth, demonstrated superior accuracy by effectively managing feature interactions. K-Nearest Neighbors (KNN) was explored with different ‘k’ values, providing deep insights into the dataset’s structure through instance similarity analysis. Naïve Bayes, offering a probabilistic viewpoint because of its assumption of feature

independence, was faster but less accurate than ensemble methods, serving as a useful baseline for probabilistic modeling. Ultimately, each model contributed uniquely to the understanding of customer churn; however, Random Forest stood out for its robustness and capability to handle unbalanced data. Meanwhile, Gradient Boosting distinguished itself by sequentially correcting errors and handling varied data features, ultimately being selected as the last model because of its comprehensive performance across multiple metrics.

Conclusion

Throughout this capstone project, we have undertaken a rigorous analysis of Bailey's Bank customer churn data to identify the underlying factors contributing to customer attrition. Using advanced data wrangling techniques, we prepared the dataset by removing irrelevant columns and correcting discrepancies, setting a firm foundation for in-depth analysis. Subsequent explorations revealed several critical insights into customer behaviors and preferences.

Our exploratory data analysis highlighted key variables influencing churn, such as age, balance, number of products used, and tenure with the bank. Notably, age emerged as a significant predictor of churn, with older customers showing a higher propensity to leave. This demographic trend was reinforced by our statistical tests, including ANOVA, which confirmed significant variations in customer age across different geographical regions, possibly affecting churn rates.

We applied multiple machine learning models to predict churn, beginning with logistic regression to establish a baseline and progressing to more complex algorithms like Random Forest and Gradient Boosting. Each model provided unique insights, but ultimately, the Random Forest and Gradient Boosting models showed superior performance in handling the dataset's complexities, including feature interactions and non-linear relationships.

The project's findings are critical for Bailey's Bank in several ways:

1. **Strategic Customer Retention**: Understanding that age is a key churn predictor allows the bank to tailor specific retention strategies to older customers, potentially through personalized service offerings or loyalty programs.
2. **Product and Service Improvement**: Insights into the number of products used and balance levels guide the bank in optimizing their product portfolio to better meet customer needs.
3. **Geographical Tailoring**: Recognition of regional differences in customer demographics can help in customizing marketing and service strategies to fit local customer profiles, enhancing satisfaction and loyalty.

The chosen models enable Bailey's Bank to proactively identify at-risk customers with high accuracy and predictive power and implement effective interventions before these customers churn. This capability is not just a technical achievement but a strategic asset that can lead to improved customer satisfaction, reduced attrition rates, and stronger financial performance.

In conclusion, this project not only underscores the importance of data-driven decision making in managing customer relationships but also sets a roadmap for Bailey's Bank to enhance its

competitive edge by transforming insights into actionable retention strategies. From now on, continual refinement of the models and integration of real-time data analytics should further enhance the bank’s ability to adapt to changing customer needs and market dynamics.