I Cloud 7 Technical Assessment

Data Scientist Jr.
Presented by: Jorge Forero

Executive Summary

Objective:

The primary objective of this project was to predict customer churn for a European bank and to identify key drivers behind customer attrition across three distinct markets: France, Spain, and Germany. By leveraging predictive models, we aimed to provide actionable insights and recommendations that could guide the bank's efforts to improve customer retention and reduce churn, ultimately enhancing profitability and customer loyalty.

Key Findings:

- Age is the most consistent driver of churn across France, Spain, and Germany, with older customers (particularly those aged 41 and above) being at significantly higher risk of leaving. This highlights the need for targeted retention efforts focused on older demographic segments.
- Balance and financial standing show varying effects on churn. While higher balances correlate with higher churn in France and Germany, balance has less impact in Spain, where customer engagement and product variety play a more significant role in retention. This suggests that retention strategies should be tailored based on regional customer behavior.
- Product engagement is a critical factor in reducing churn. Customers with more bank products are less likely to churn across all regions, emphasizing the importance of cross-selling and maintaining strong customer relationships.

Recommendations:

- Customers aged 41 and older consistently show a higher likelihood of churn across all markets. Implementing tailored retention strategies such as exclusive financial products, personalized advisory services, or loyalty programs could help retain these high-value customers.
- Utilize the insights from the Predictive Churn-Risk System developed in this project to offer relevant products to customers who currently only hold one bank product. By accurately predicting customer behavior, the bank can increase product adoption and reduce the likelihood of churn through enhanced customer engagement.

- Build an Advanced Regional Metrics Dashboard to monitor real-time data for key churn drivers (e.g., Age, Balance, Product Usage) across different regions. The dashboard should be based on the predictive models developed during this project and provide dynamic insights into customer behavior, retention metrics, and the impact of new initiatives by region.

1.Introduction

This report presents a comprehensive analysis of customer churn using advanced data science techniques, aimed at empowering the bank to make data-driven decisions to improve customer retention. By performing an in-depth Exploratory Data Analysis (EDA) and applying machine learning models, this report seeks to identify the key factors influencing customer attrition and develop predictive models that can accurately forecast churn risk.

1.1.Scope of Analysis

This analysis examines a dataset of 10,000 customer records from France, Germany, and Spain, indicating whether each customer is active or has churned. The focus is on exploring the data to identify churn patterns and building predictive models to classify churn. The goal is to uncover key churn drivers and provide actionable insights to help the bank reduce customer attrition.

1.2.Ethical Considerations

Several ethical considerations were taken into account during the analysis:

- Data Privacy: Since the dataset used include personally identifiable information (PII), all customer-related data was handled responsibly to ensure privacy and confidentiality. Measures were taken to anonymize any sensitive data.
- Transparency and Reproducibility: All steps of the analysis, from data preprocessing to model development, were documented clearly to ensure transparency and reproducibility. This enables other analysts to replicate the analysis and verify the results.
- Bias in Predictions: Predictive models can sometimes introduce or reinforce bias, particularly if the data contains underlying biases (e.g., Gender disparities). We made efforts to ensure the fairness of the models by carefully analyzing feature importance and ensuring balanced evaluation metrics.

2.Methodology

2.1.Data Collection and Processing

Data for this analysis was provided by the client, containing key information about 10,000 customers. The dataset included the following fields:

- CustomerId: A unique identifier for each customer
- Surname: The customer's last name
- CreditScore: A numerical value representing the customer's credit score
- Geography: The country where the customer resides (France, Spain or Germany)
- Gender: The customer's gender (Male or Female)
- Age: The customer's age
- Tenure: The number of years the customer has been with the bank
- Balance: The customer's account balance
- NumOfProducts: The number of bank products the customer uses (e.g., savings account, credit card)
- HasCrCard: Whether the customer has a credit card (1 = yes, 0 = no)
- IsActiveMember: Whether the customer is an active member (1 = yes, 0 = no)
- EstimatedSalary: The estimated salary of the customer
- Exited: Whether the customer has churned (1 = yes, 0 = no)

In line with the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, the following phases were followed to structure this project:

- Business Understanding: The primary objective was to analyze customer churn and create predictive models that could help identify customers likely to leave the bank.
- Data Understanding: After initial exploration, the 'Exited' field was identified as the target variable, and a detailed exploratory data analysis (EDA) was conducted to understand customer behavior in each country (France, Spain, Germany).

To address the class imbalance in the target variable ('Exited'), we employed two strategies:

- Undersampling: Reducing the number of majority class instances to match the minority class size.
- Oversampling: Increasing the number of minority class instances by generating synthetic samples or replicating existing ones to balance with the majority class.

We evaluated the performance of the models using a combination of metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), to assess the models' ability to predict customer churn accurately.

After comparing the models on both training and validation data, Random Forest consistently outperformed other models, showing the best balance between precision and recall. This model was selected as the best-performing model for further tuning.

To further enhance the performance of the Random Forest model, we performed hyperparameter tuning using grid search. The parameters tuned included:

- Number of trees (n_estimators).
- Maximum depth of the trees (max_depth).
- Minimum samples required to split a node (min_samples_split).
- Minumim Samples per leaf (min_samples_leaf).
- Max_features
- Bootstrap

Finally, we conducted a feature importance analysis to identify the key drivers of customer churn. The Random Forest model provided insights into which features (e.g., Age, Balance, Active Membership) had the greatest impact on churn. This analysis played a crucial role in developing strategic recommendations for customer retention.

3.Key Findings

3.1. France Branch

Economic Observations:

- France's GDP growth is projected to be subdued at 0.7% in 2024, following a significant slowdown in 2023. This slow growth may affect consumer spending and overall banking activity, influencing churn rates as customers may seek better financial options elsewhere due to limited economic opportunities.
- Inflation is expected to decrease to 2.5% in 2024 from 5.7% in 2023. A decline in inflation can improve purchasing power, potentially reducing churn as customers feel more financially secure and less inclined to switch banks.

Banking Market Observations:

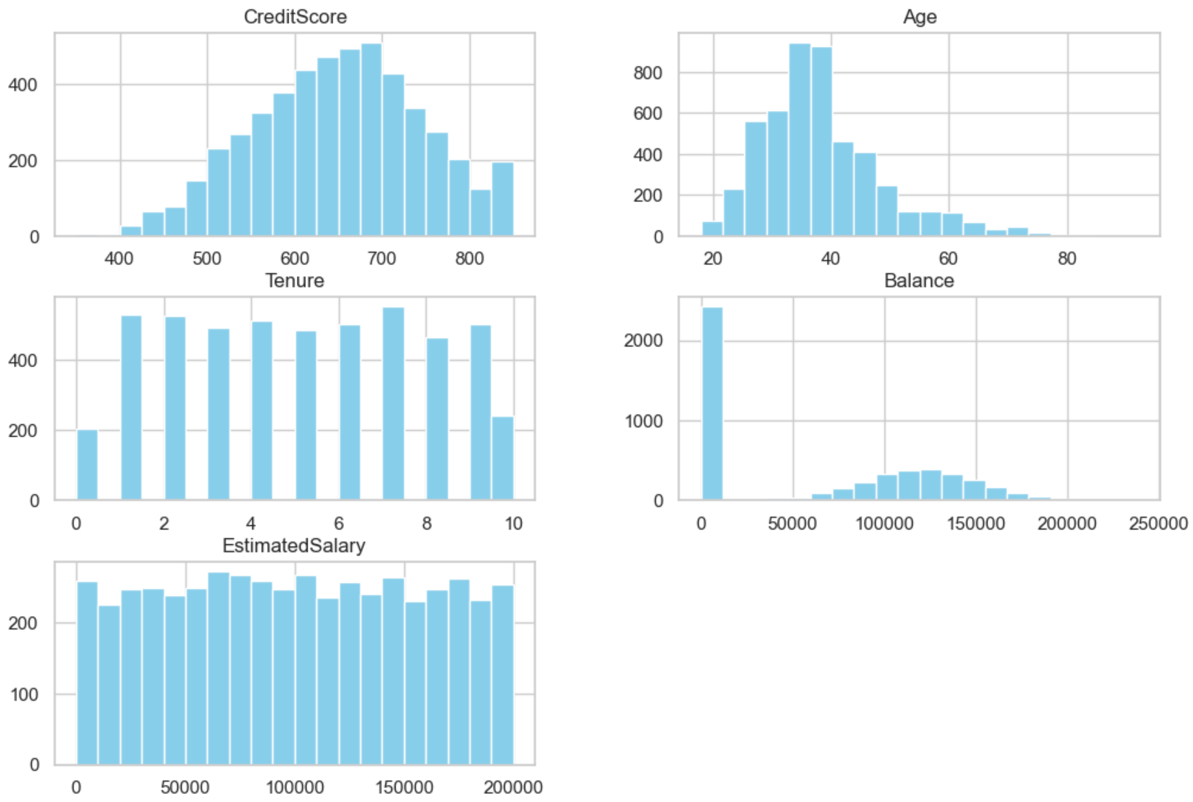
- French banks are experiencing pressure on revenues due to rising interest rates not translating into higher profits. This scenario may lead banks to increase fees or reduce services, which could contribute to higher churn rates if customers feel they are not receiving adequate value.
- Despite challenges, asset quality is expected to remain stable, which could influence customer confidence in their bank's stability and reliability, impacting their likelihood of staying or leaving.

Data Analysis and Findings

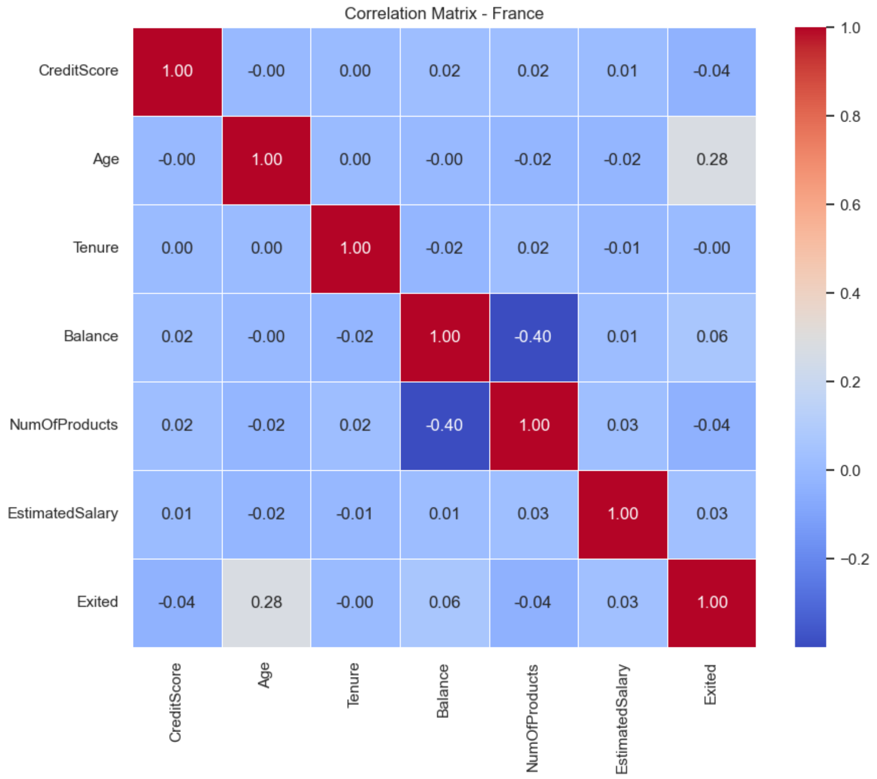
- The credit score distribution is approximately normal, with most customers having scores between 600 and 700. The age distribution is skewed towards younger customers, primarily between 30 and 50, indicating a solid base of financially responsible young clients. *(See Chart 1)*
- The balance and estimated salary data suggest that many customers either don't save significantly with the bank, or there may be issues with how balance is measured. It might be worth considering using an average balance over a year for better accuracy. *(See Chart 1)*
- There is a positive correlation between age, balance, and churn, suggesting that older customers or those with higher balances are more likely to churn. *(See Chart 2)*
- Customers with more products are slightly less likely to churn, which is a key correlation that will be further explored with our ML models. *(See Chart 2)*
- Active membership appears to play a critical role, with active customers showing lower churn risk. Similarly, customers holding only one product are more prone to churn, which follows logical expectations. *(See Chart 3)*

Visualizations and Data

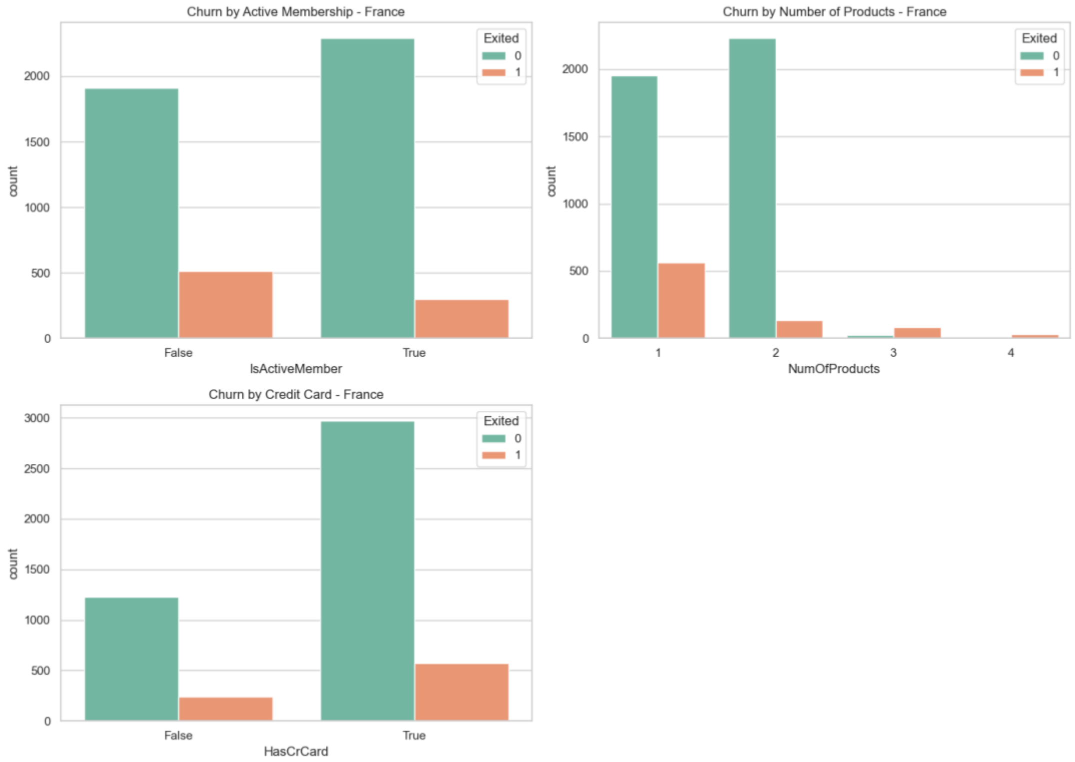
- Chart 1: Distribution of Numerical Features



- Chart 2: Correlation Matrix France



- Chart 3: Churn by Categorical Variables

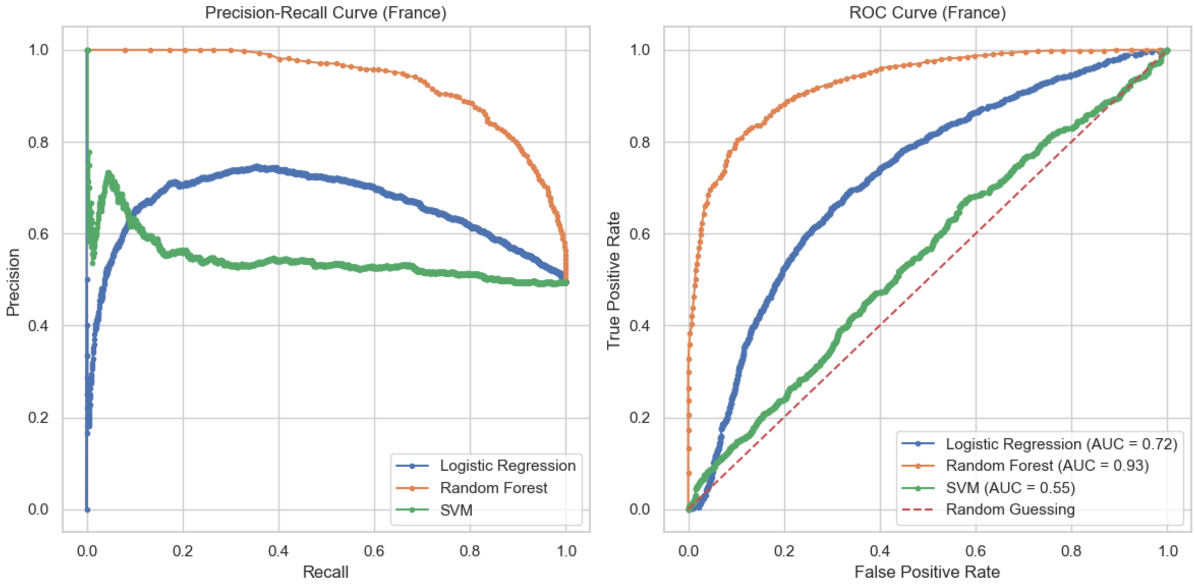


ML Processing and Model Insights:

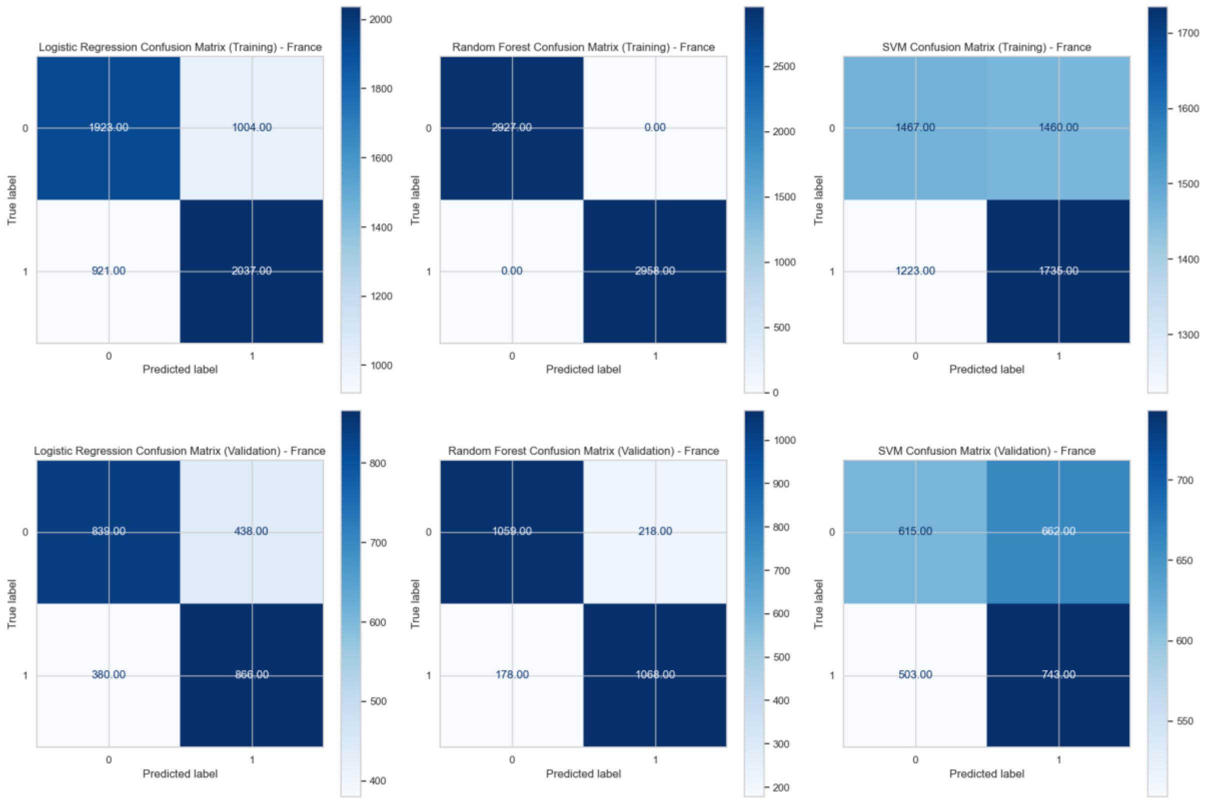
- Random Forest model performed exceptionally well, providing the best balance between precision and recall across all models tested. This means it effectively identified customers likely to churn without sacrificing accuracy, making it a reliable tool for churn prediction. *(See Chart 4)*
- Key Drivers of Churn Identified *(See Chart 6)*:
    - Age: Older customers are at significantly higher risk of churn, highlighting the need for targeted retention efforts focused on this demographic.
    - Balance: Customers with higher balances show a slightly higher likelihood of churning. Retention efforts could focus on enhancing the value proposition for this group, such as offering exclusive products or personalized financial services.
    - Active Membership: Active customers are less likely to churn. This emphasizes the importance of maintaining customer engagement through loyalty programs or proactive customer service initiatives.

Visualizations and Data

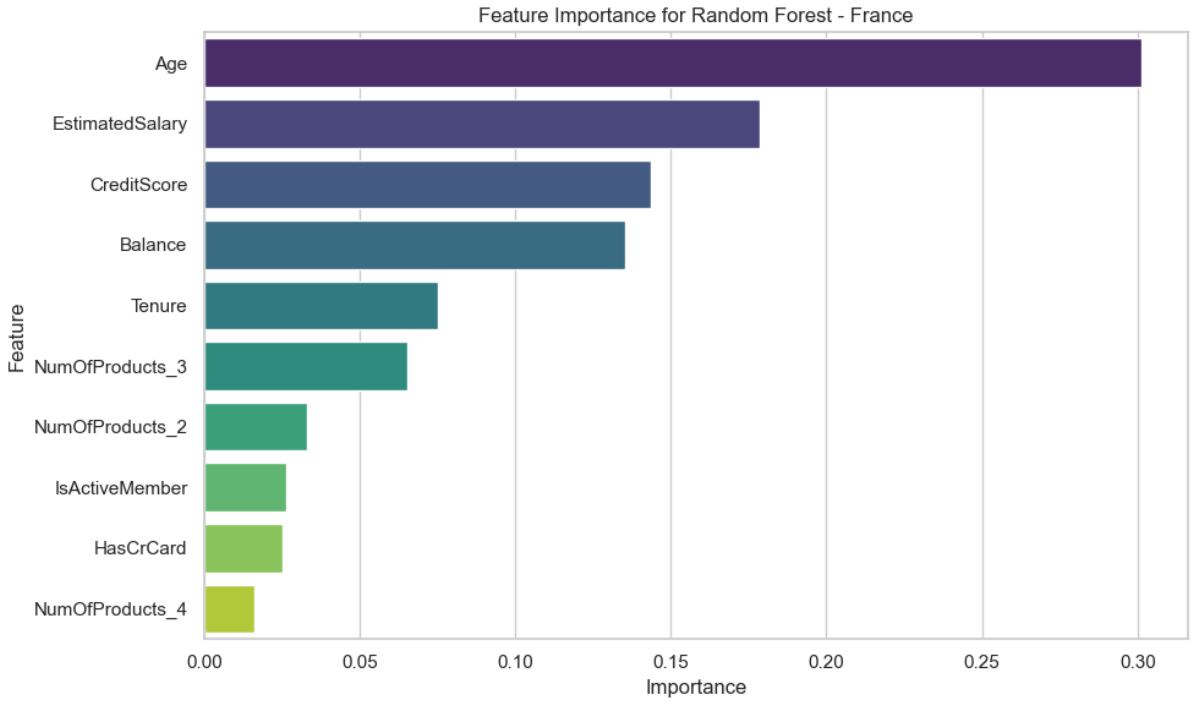
- Chart 4: Precision-Recall & ROC Curve



- Chart 5: Confusion Matrix – All Models



- Chart 6: Feature Importance Chart



3.2. Spain Branch

Economic Observations:

- Spain's GDP is forecasted to grow by 2.5% in 2024, driven by external demand and recovery plans. A robust economic outlook can enhance consumer confidence, potentially reducing churn as customers feel more secure in their financial decisions.
- The European Central Bank (ECB) has begun easing interest rates, improving credit performance and boosting deposit growth. This environment may lead to more favorable lending conditions for customers, thereby decreasing the likelihood of churn as clients find better financing options available.

Banking Market Observations:

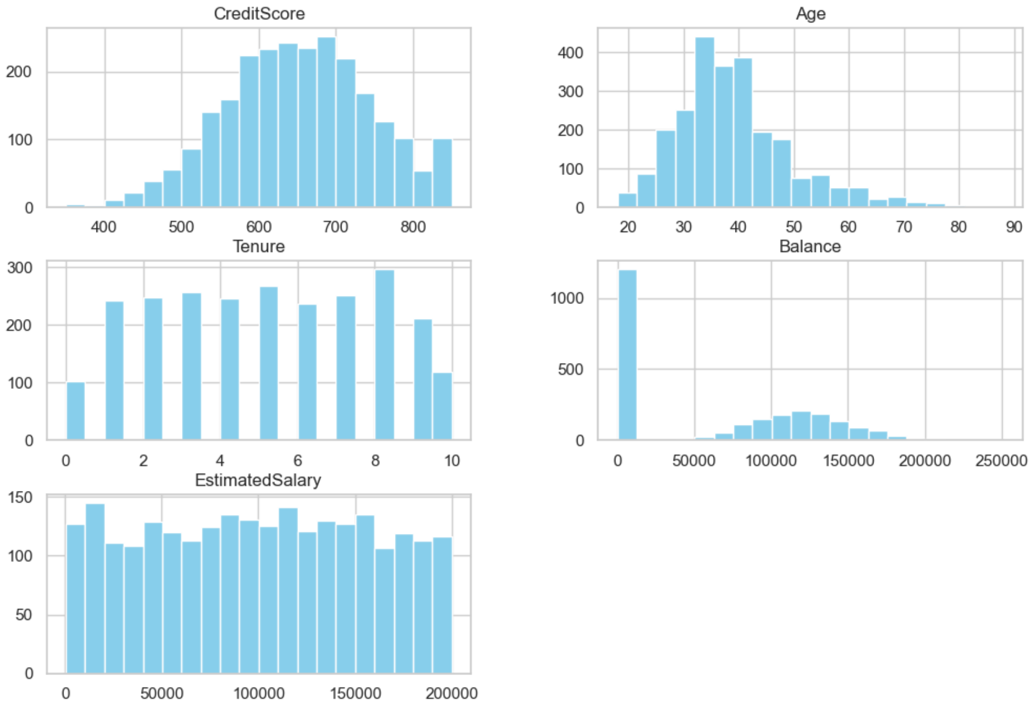
- The Spanish banking sector has seen profitability (ROE) increase to 12.6% at the end of 2023, indicating a healthy banking environment that could foster customer loyalty as banks perform well financially.
- The NPL ratio stands at 3.15%, higher than the European average but stable relative to previous years. This stability may reassure customers about the bank's risk management practices, potentially reducing churn rates if they perceive their bank as financially sound.

Data Analysis and Findings:

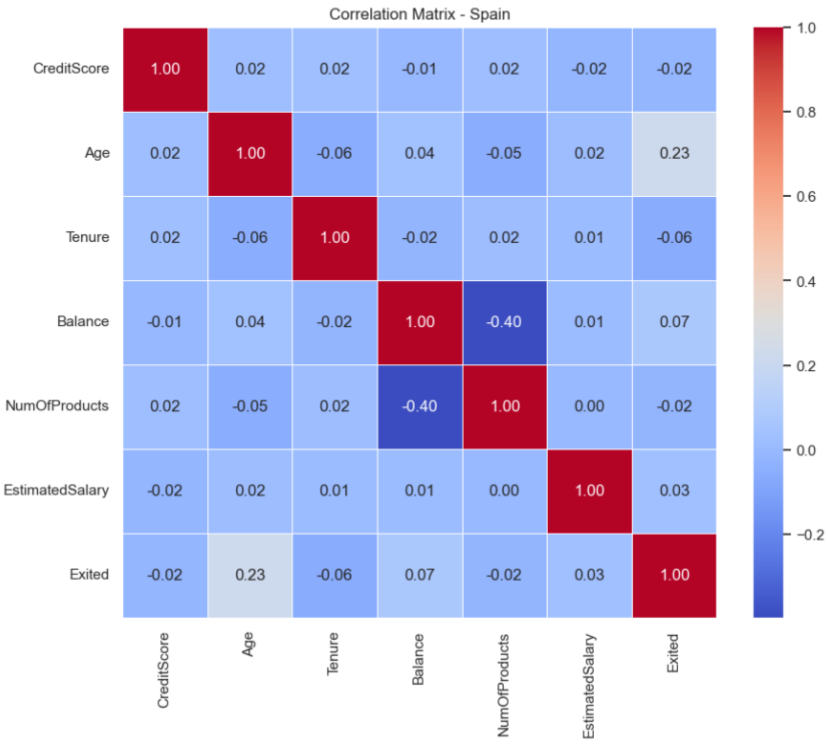
- Similar to France, the credit score distribution is normal, with mos customers socring between 600 and 700. The age distribution is also skwed towards younger customers, with most between 30 and 50 years old. *(See Chart 7)*
- A significant proportion of customers have zero balance, wich mirrors the French dataset. The remaining balances show a wide spread between 50,000 and 150,000. The estimated salary distribution remains uniform, as observed in France, with no major differences across different customer segments. *(See chart 7)*
- The 31–40 age group represents the largest share of churners and non-churners, making it a key segment to target for retention strategies. Customers aged 41–50 and 51–60 also show moderate churn levels, further indicating that older clients are more prone to churn. *(See Chart 10)*

Visualizations and Data

- Chart 7: Distribution of Numerical Features

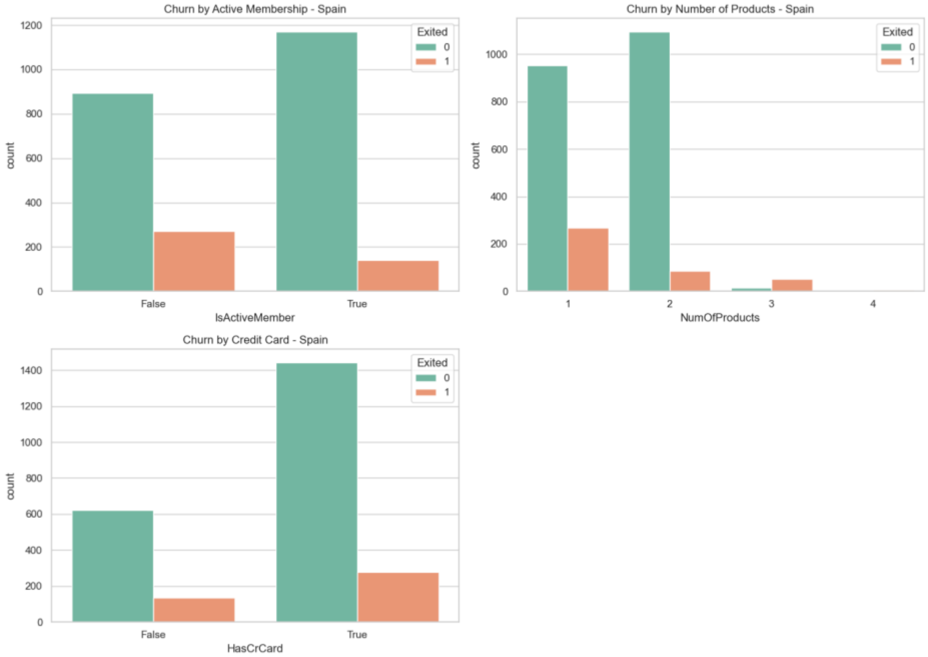


- Chart 8: Correlation Matrix France

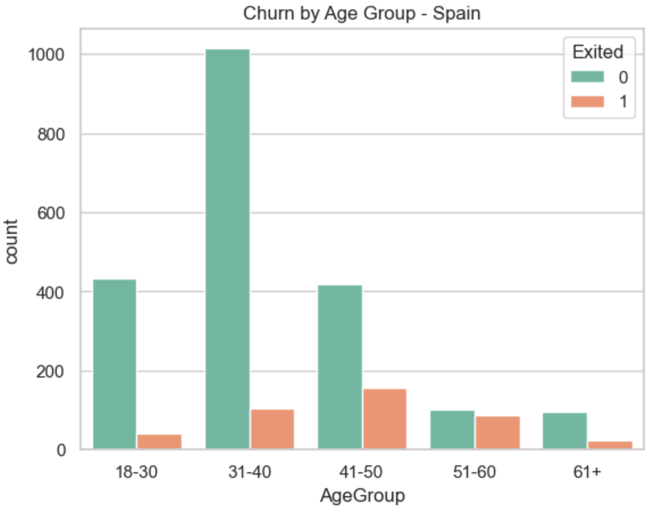


	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	Exited
CreditScore	1.00	0.02	0.02	-0.01	0.02	-0.02	-0.02
Age	0.02	1.00	-0.06	0.04	-0.05	0.02	0.23
Tenure	0.02	-0.06	1.00	-0.02	0.02	0.01	-0.06
Balance	-0.01	0.04	-0.02	1.00	-0.40	0.01	0.07
NumOfProducts	0.02	-0.05	0.02	-0.40	1.00	0.00	-0.02
EstimatedSalary	-0.02	0.02	0.01	0.01	0.00	1.00	0.03
Exited	-0.02	0.23	-0.06	0.07	-0.02	0.03	1.00

- Chart 9: Churn by Categorical Variables



- Chart 10: Churn by Age Group



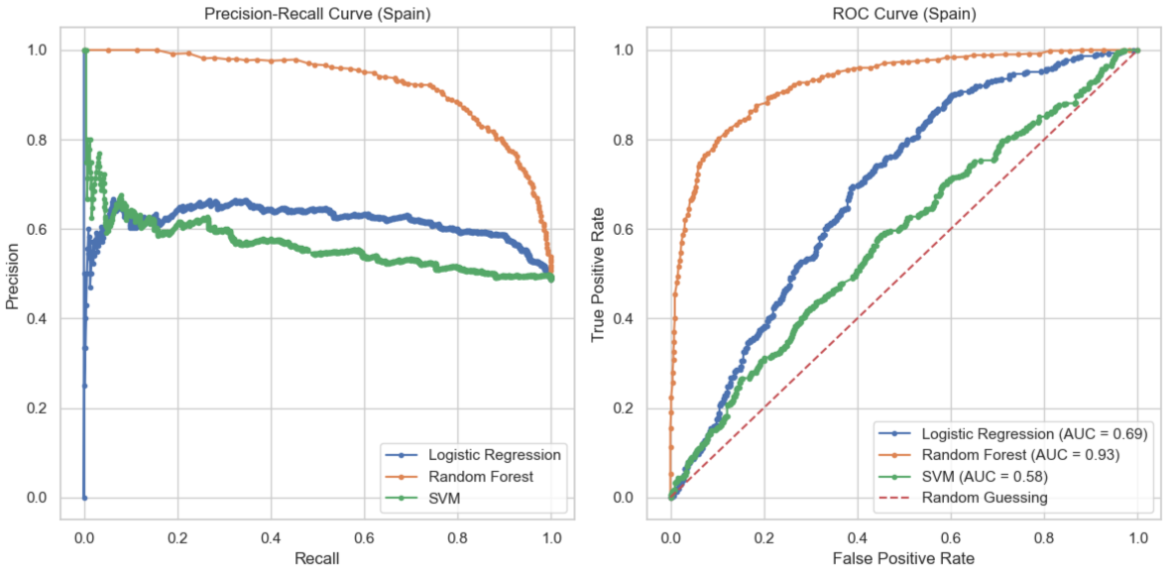
ML Processing and Model Insights:

- Random Forest outperforms Logistic Regression and SVM in all metrics, demonstrating the best balance between precision and recall. It achieves an AUC of 0.93, showing its ability to handle the complexity of the data and class imbalances, particularly compared to SVM's underperformance, which struggled with the complexity of the dataset *(See Chart 11)*
- Key Features influencing churn for Spain include age, balance, and estimated salary, as shown in the feature importance analysis. These insights can guide the bank in focusing retention strategies on specific customer segments, such as older clients with higher balances, who are more likely to churn. *(See Chart 13)*

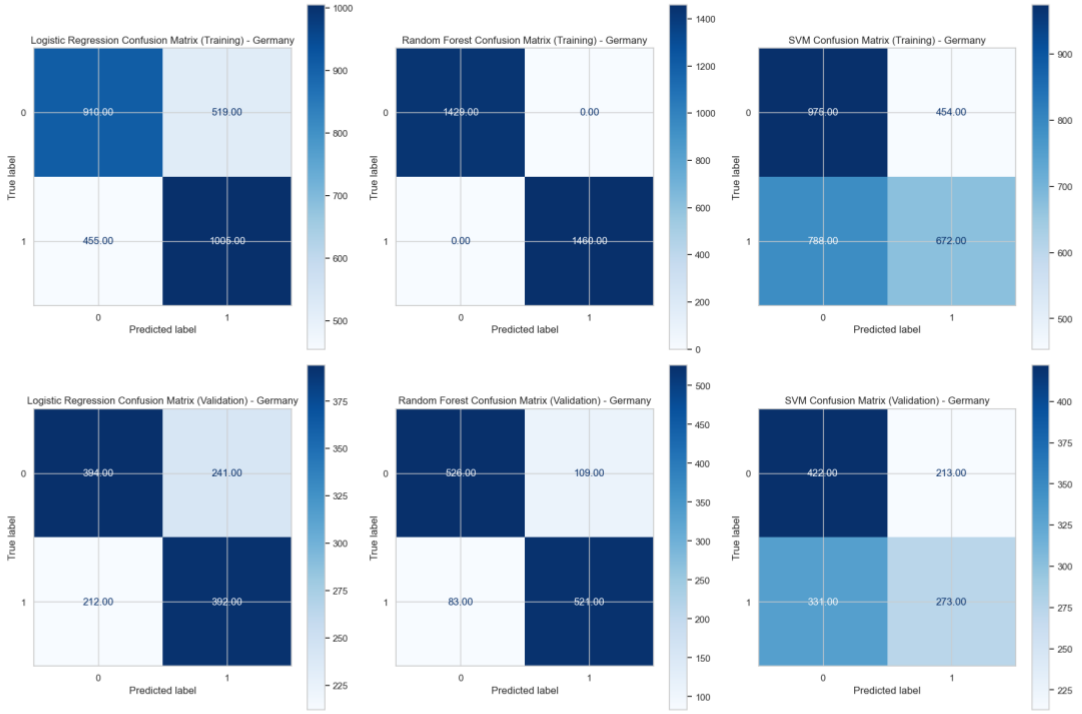
- The Confusion Matrix shows that the model is especially strong at predicting customers who will not churn (non-churners), with relatively few false positives. This suggests the model is reliable when identifying customers who are likely to remain, helping the bank prioritize high-risk clients for retention efforts.

Visualizations and Data

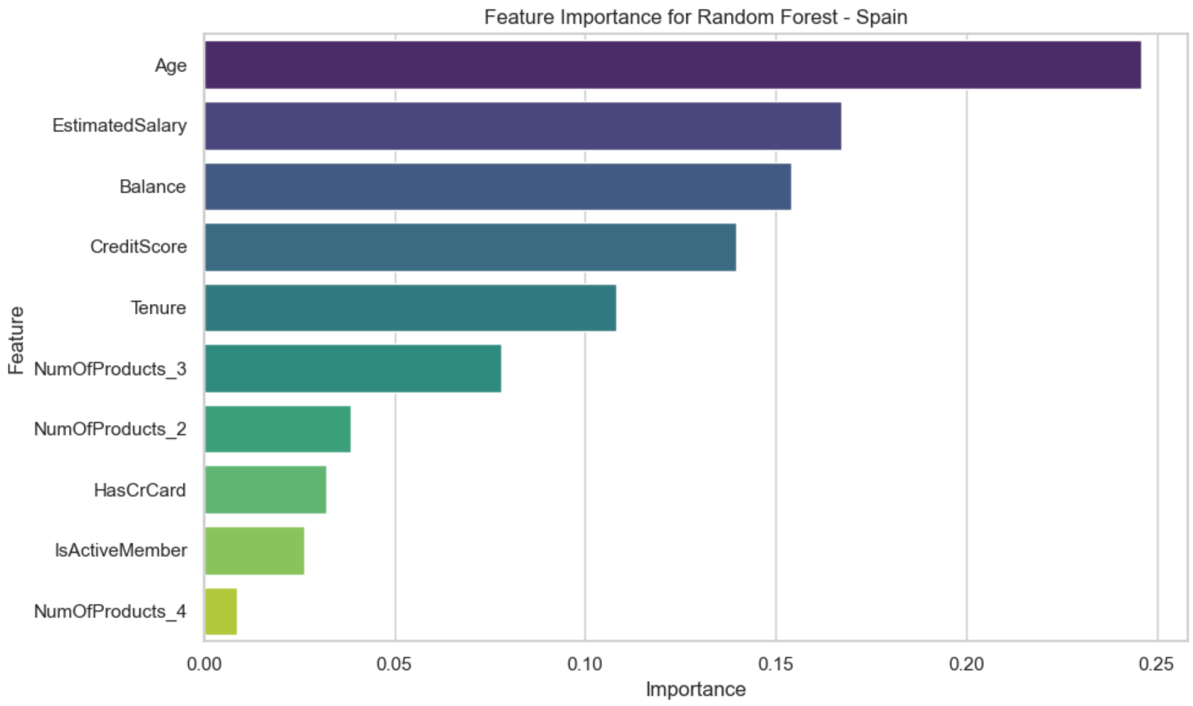
- Chart 11: Precision-Recall & ROC Curve



- Chart 12: Confusion Matrix – All Models



- Chart 13: Feature Importance Chart



3.3. Germany Branch

Economic Observations:

- The German economy is expected to grow by only 0.2% in 2024, indicating a slow recovery from recession. This sluggish growth could lead customers to reconsider their banking relationships if they perceive limited financial opportunities.
- Inflation Rates: Inflation is projected at 2.2% in 2024, down from higher levels previously experienced. Lower inflation can enhance consumer purchasing power and stability, possibly reducing churn as customers feel more secure financially.

Banking Market Observations:

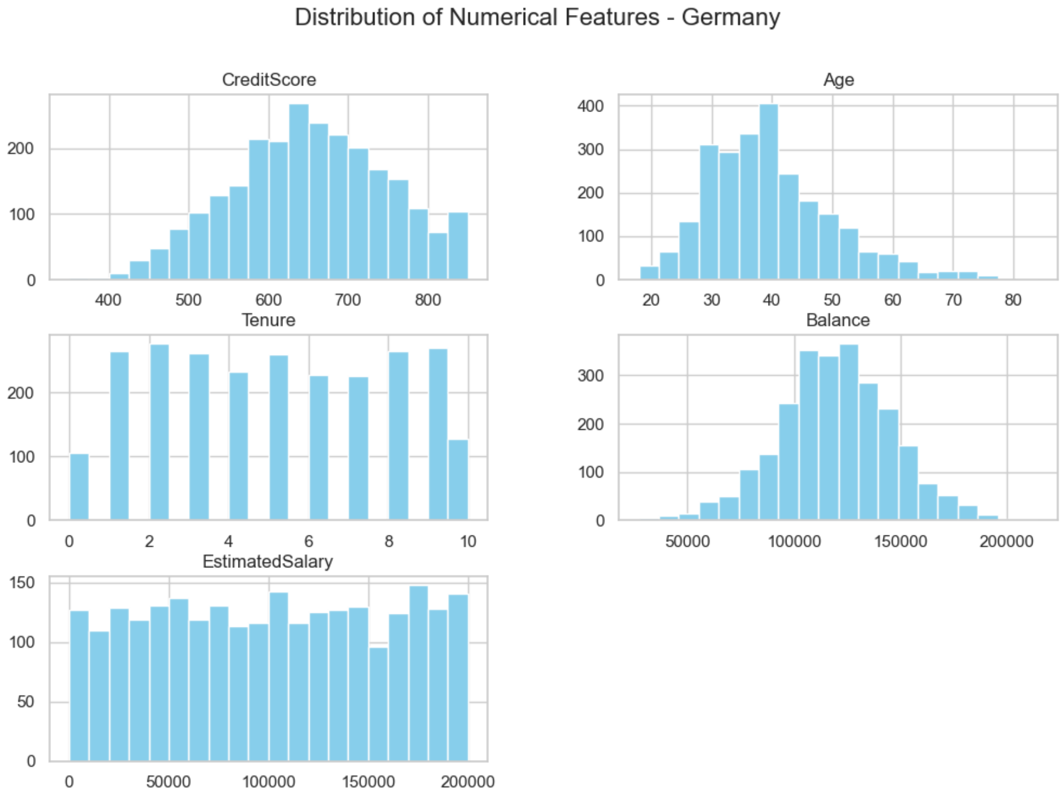
- Business and consumer confidence remain low despite a gradual recovery, which may lead to increased churn if customers seek banks with better service or offerings during uncertain times.
- Germany's government financing deficit is expected to decrease significantly by 2025, indicating improved fiscal health that could positively influence customer perceptions of bank stability and reliability.

Analysis and Findings

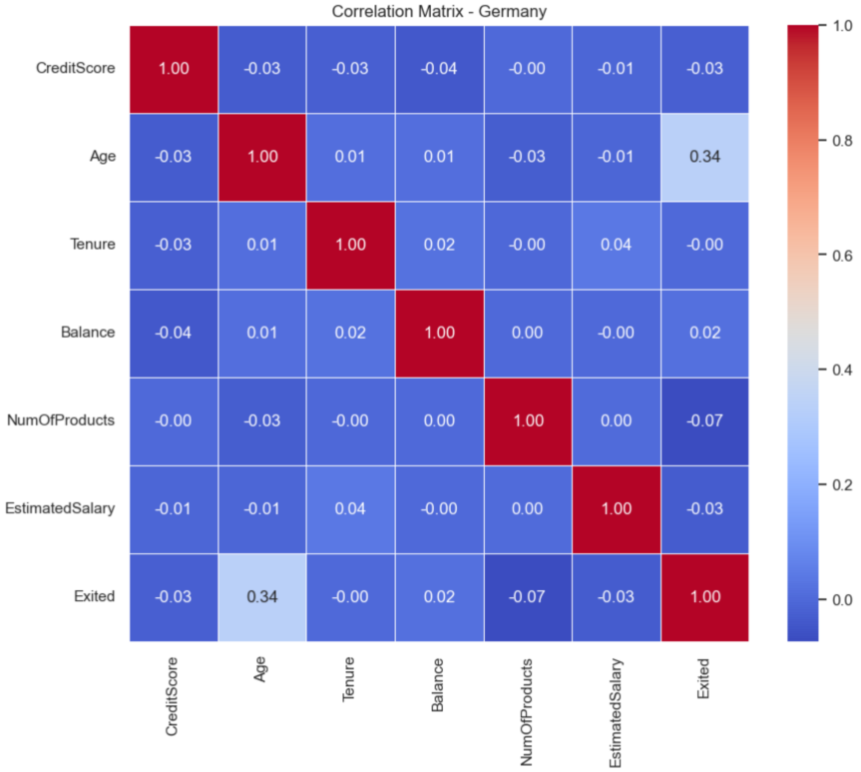
- The credit score distribution is fairly normal, with most customers scoring between 600 and 750. The majority of customers are aged 30 to 50, similar to France and Spain, with fewer customers over 60. *(See Chart 14)*
- Age has a stronger correlation with churn (0.34) compared to France and Spain, making it a key churn predictor for Germany. The number of products is negatively correlated with churn (-0.07), reaffirming that customers with more products are less likely to churn. Balance shows a weak positive correlation (0.02) with churn, while credit score, tenure, and salary show minimal impact. *(See Chart 15)*
- The 31–40 age group has the highest churn rate, similar to France and Spain. The 41–60 age groups also show moderate churn, indicating that older customers are more likely to leave. The 18–30 and 61+ groups have lower churn rates. *(See Chart 17)*

Visualizations and Data

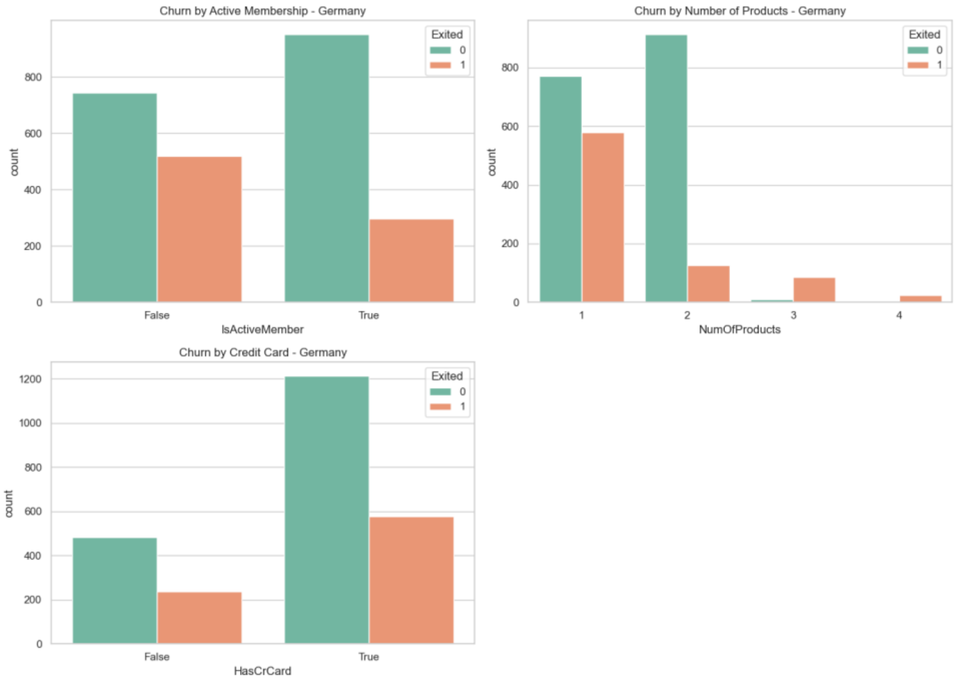
- Chart 14: Distribution of Numerical Features



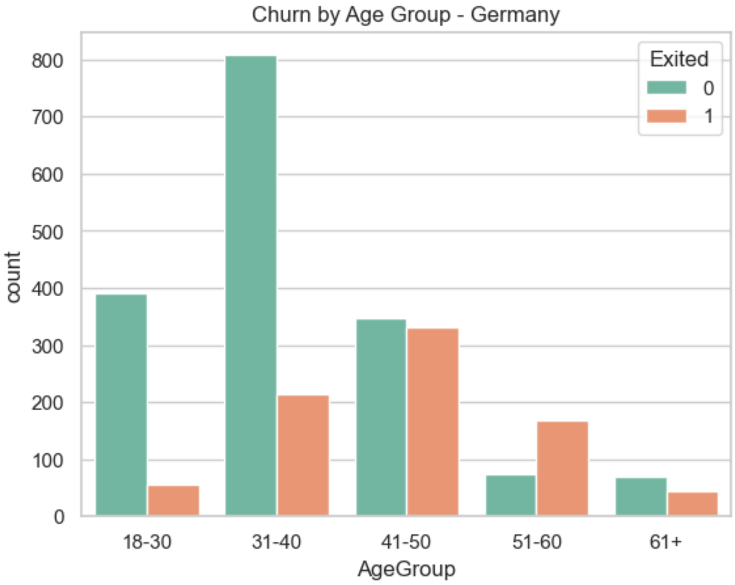
- Chart 15: Correlation Matrix France



- Chart 16: Churn by Categorical Variables



- Chart 17: Churn by Age Group

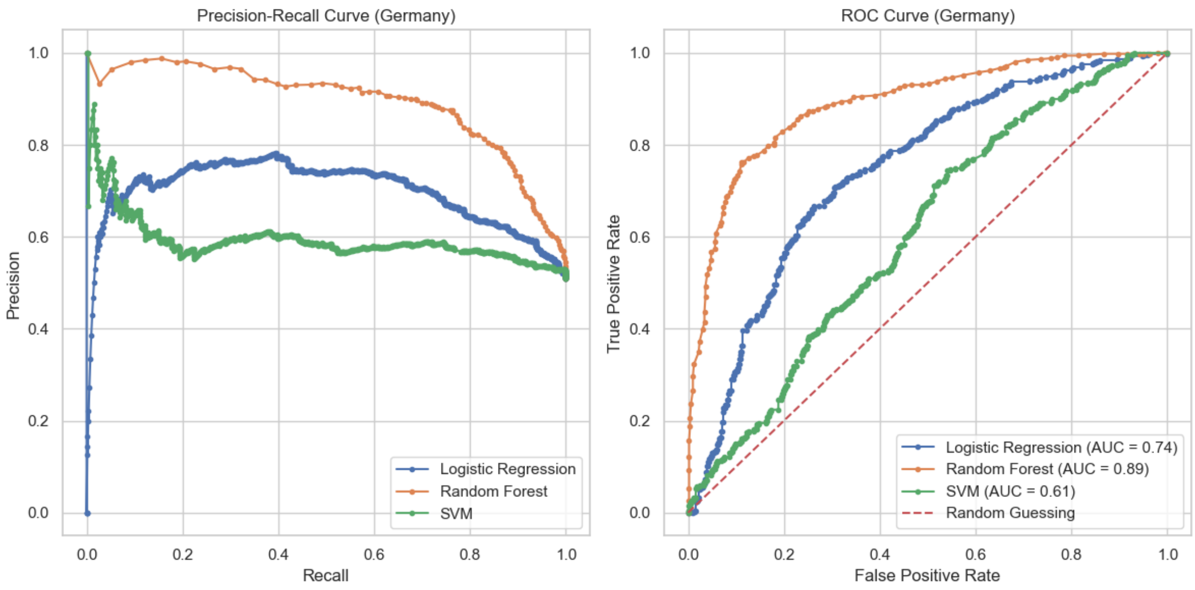


ML Processing and Model Insights

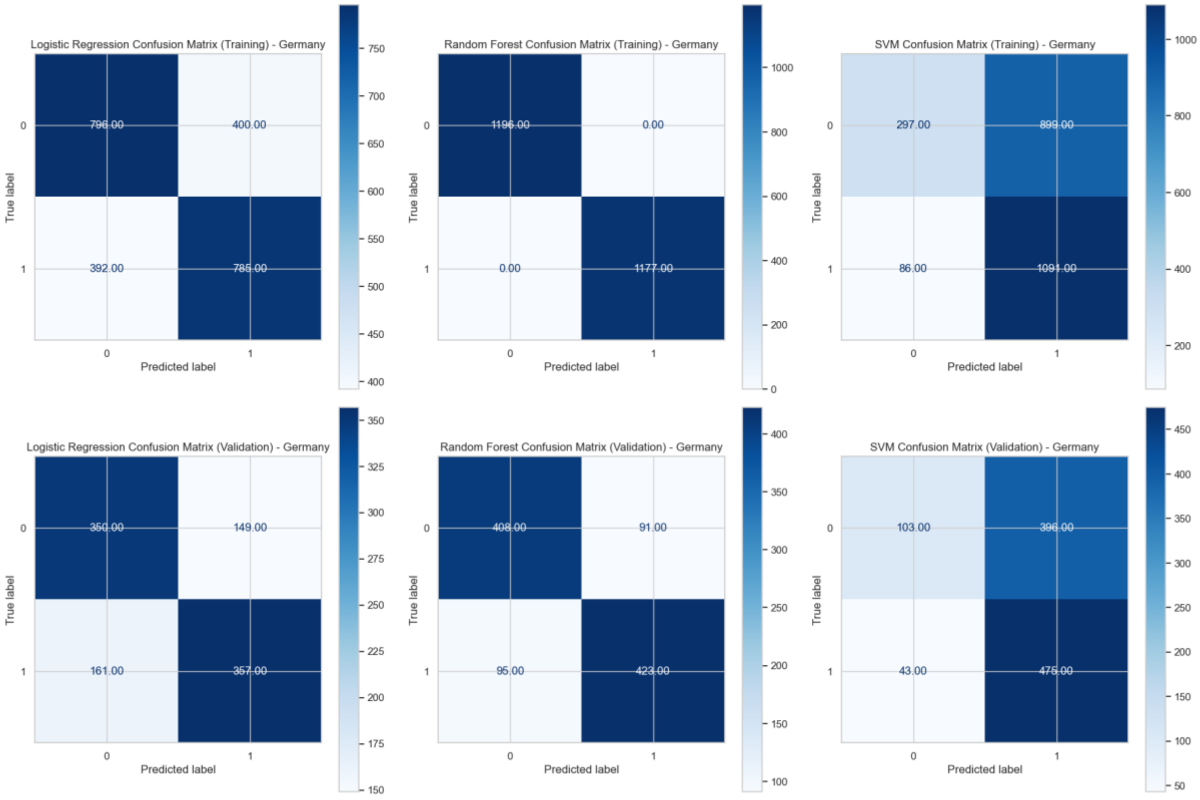
- The Random Forest model consistently performed the best, exhibiting the highest precision and recall across all models. The model also achieved an AUC score of 0.89, showing its strength in accurately predicting customer churn in the German market. *(See Chart 18)*
- The Random Forest model shows the most balanced confusion matrix. It achieved strong results in predicting both churners and non-churners, with 424 true positives and 402 true negatives on the test data, reflecting its capacity to handle class imbalance effectively. *(See Chart 19)*
- The most important feature in the Random Forest model for Germany was Age, indicating that older customers are more likely to churn. This aligns with our exploratory analysis, suggesting age should be a critical focus area for retention efforts. *(See Chart 20)*
- Balance and Estimated Salary were also significant predictors, with higher balances leading to a higher likelihood of churn, contrasting with some other countries like Spain, where churn was more prevalent among customers with lower balances. *(See Chart 20)*

Visualizations and Data

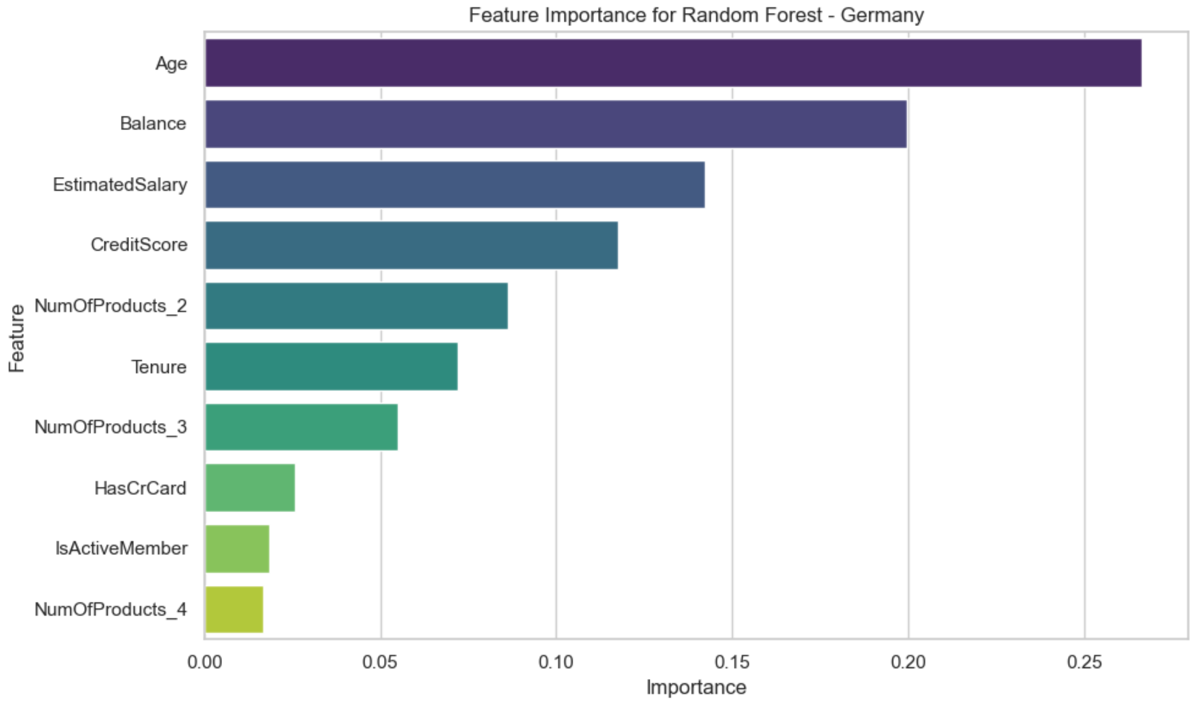
- Chart 18: Precision-Recall & ROC Curve



- Chart 19: Confusion Matrix – All Models



- Chart 20: Feature Importance Chart



4. Recommendations

4.1. Actionable Insights

- Across all countries, age emerged as a strong predictor of churn, particularly in the 31-50 age group. In Germany, the correlation between age and churn is even stronger, with older customers at higher risk of leaving. Retention strategies should focus on this demographic with tailored campaigns such as personalized financial advice, exclusive offers, or loyalty rewards to keep them engaged.
- Customers with only one product are more likely to churn across all countries. Cross-selling additional products, such as credit cards, savings accounts, or investment services, can help solidify the customer relationship. Focus should be on cross-selling to younger and inactive customers, who are at higher risk of leaving.
- Given that churn is influenced by multiple factors such as age, balance, and membership status, monitoring customer satisfaction through regular surveys and feedback mechanisms can provide early warning signs. Implementing proactive measures, such as offering financial health checks or surprise benefits for high-risk segments, can improve customer experience and reduce churn likelihood.

4.2.Strategic Initiatives

- Use specific data to create loyalty programs offering exclusive financial products, dedicated relationship managers or personalized financial planning for high-net-worth individuals and older clients.
- Enhance the customer experience via experimentation by investing in upgrading the digital banking platform with a focus on user experience, more intuitive features, and advanced self service tools.
- Develop an Advanced Regional Metrics Dashboard that continuously monitors key churn drivers and retention metrics, segmented by region (France, Spain, Germany). The dashboard will be built on the predictive models from this project (Random Forest) and will track customer behavior in real time, allowing the bank to evaluate the success of new retention initiatives.
- Develop a Predictive Cross-Selling System based on customer behavior analysis and the predictive models used in this project. The system will help the bank identify the right customers for product cross-sell opportunities, and also forecast potential demand for new product creation.

5.Next Steps

- Action Items:
  Immediate: Implement the Random Forest model as the core tool for predicting churn in each region. Use the identified key features (age, balance, active membership) to segment and target customers at higher risk of churn.
  Within 3 Months: Integrate the models into the bank's CRM system, enabling automated alerts for high-risk customers and real-time churn prediction.

- Future Analysis
  Collaborate with data engineers and data visualization experts to design a dashboard that offers real-time, actionable insights based on the models developed.
  Continuously track regional churn trends and customer behaviors, using the dashboard to measure the effectiveness of retention initiatives and adjust strategies as needed.