

A background image showing a group of business professionals in a meeting. One person is holding a smartphone, and another is pointing at a laptop screen. The image is dimmed to allow text to be read.

# Customer Turnover at Banks

---

DATA MINING PRINCIPLES

GROUP 7

VINCENT CALDWELL, ROSELYN ROZARIO, NORMA FERREL, ANKIT GUBILIGARI, AND ALBERTO BERMEA TREVINO



# Agenda

## Introduction

- Executive Summary

## Summary of Analysis

- Analysis Goals & Analytical Plan
- Data Description
- Expected Results

## Unsupervised and Supervised Methods & Solutions

## Strategic Implications for Stakeholders & Business Value

# Executive Summary

---



## **Understanding Customer Churn Drivers:**

- Identified and analyzed key factors driving customer turnover through comprehensive data analysis.
- Uncovered insights into customer behavior, preferences, and satisfaction levels contributing to churn.



## **Segmentation for Targeted Market:**

- Clustered the customer base into distinct groups based on behavioral patterns, credit scores, and purchasing habits



## **Predictive Modeling for Churn Forecasting:**

- Leveraged historical data and algorithms to develop robust predictive models that predict future customer churn patterns



## **Recommendations for Retention/Acquisition:**

- Formulated tailored strategies to optimize customer retention and acquisition by using the forecasts from our models.

# Goals

---

1

UNDERSTAND THE KEY  
DRIVERS OF CUSTOMER  
TURNOVER (CHURN).

2

SEGMENT THE  
CUSTOMER BASE INTO  
DISTINCT GROUPS FOR  
TARGETED MARKETING  
STRATEGIES.

3

DEVELOP PREDICTIVE  
MODELS TO FORECAST  
CUSTOMER CHURN.

4

FORMULATE  
RECOMMENDATIONS  
BASED ON THE  
ANALYSIS TO SUPPORT  
RETENTION AND  
ACQUISITION EFFORTS.

# Analytical Plan

---

SEGMENTATION ANALYSIS

---

FEATURE IMPORTANCE

---

RELATION AND INTERACTIONS BETWEEN  
VARIABLES

---

SURVIVAL ANALYSIS

---

EXPLAINABLE AI

---









BEST MODEL

---

RECOMMENDATIONS

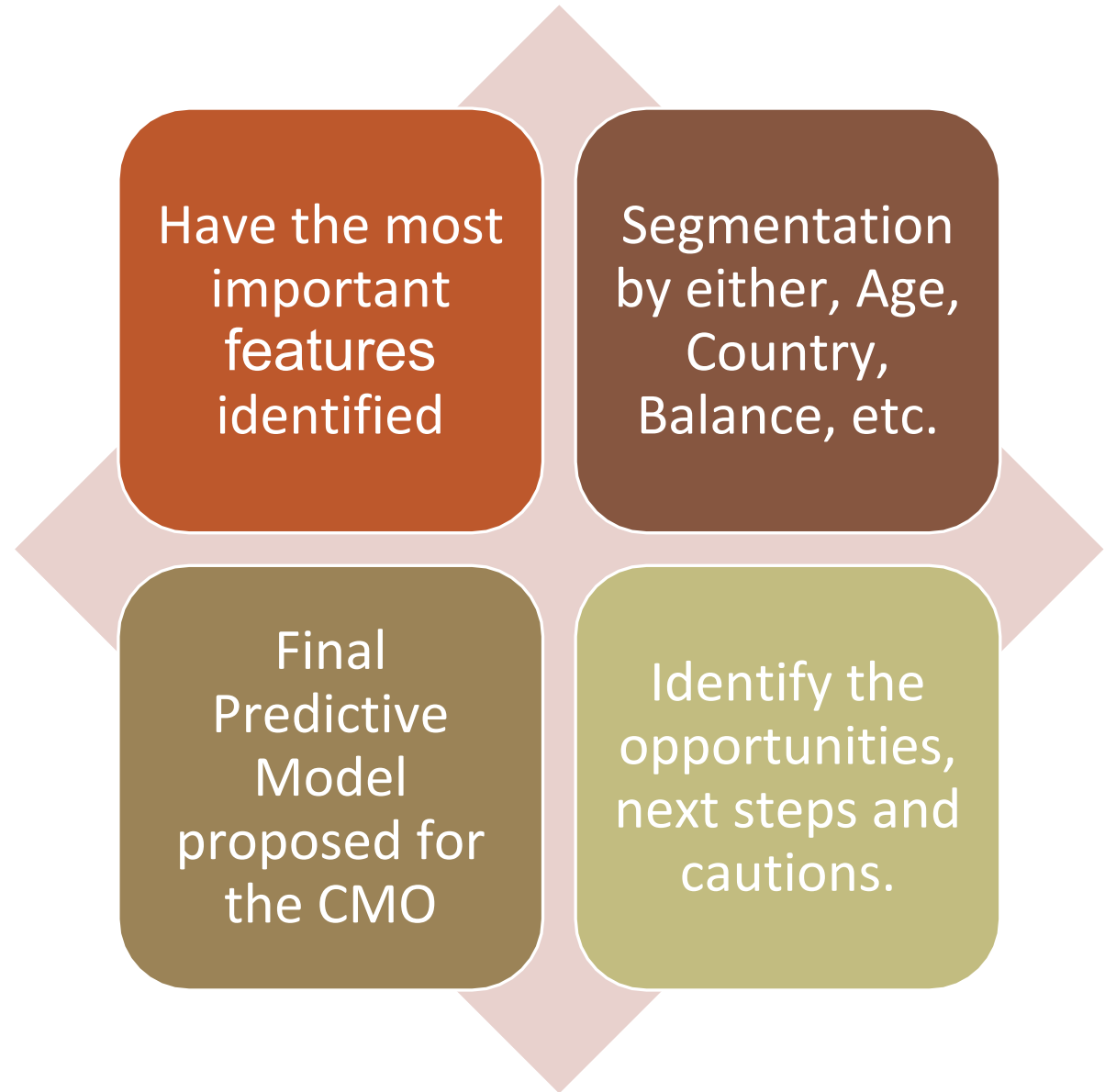
# Data Description

---

-  **CreditScore:** Ranges from 350 to 850 with a mean of approximately 650.52, indicating a wide range of credit scores among the customers.
-  **Age:** The customers' ages range from 18 to 92 years, with a mean age of around 38.92 years. This shows a broad age distribution among the bank's customers.
-  **Tenure:** Tenure with the bank ranges from 0 to 10 years, with an average tenure of about 5.01 years.
-  **Balance:** Bank balances vary significantly among customers, from 0 to approximately 250,898.09, with an average balance of around 76,485.89. The presence of balances at 0 could - - indicate customers without a current account balance or savings.
-  **NumOfProducts:** Customers use between 1 and 4 products, with an average slightly above 1.5 products per customer.
-  **HasCrCard** and **IsActiveMember:** These are binary variables indicating whether a customer has a credit card and whether they are considered an active member, respectively.
-  **EstimatedSalary:** Estimated salaries of customers vary widely, ranging from 11.58 to almost 200,000, with an average salary of approximately 100,090.24.
-  **Exited:** This binary variable indicates customer churn, with around 20.37% of customers having exited.

There are no duplicate rows.

# Expected Results

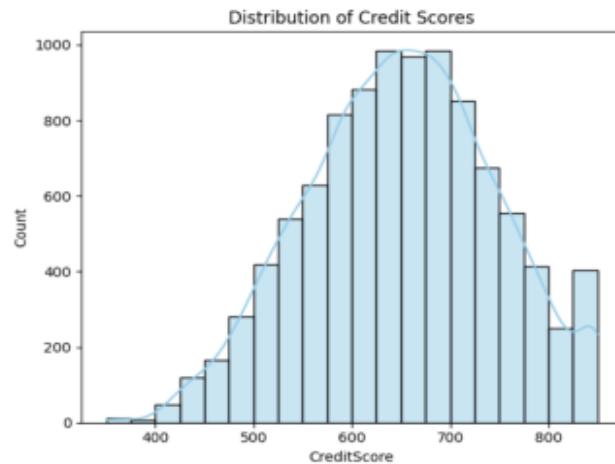


# Preliminary Analysis

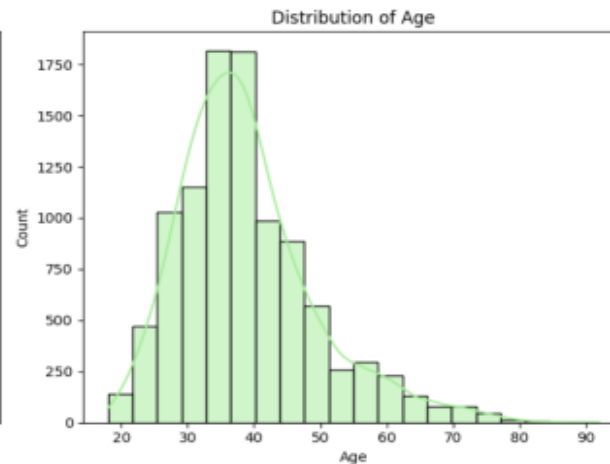
1. UNIVARIATE  
ANALYSIS
2. BIVARIATE/  
MULTIVARIATE  
ANALYSIS



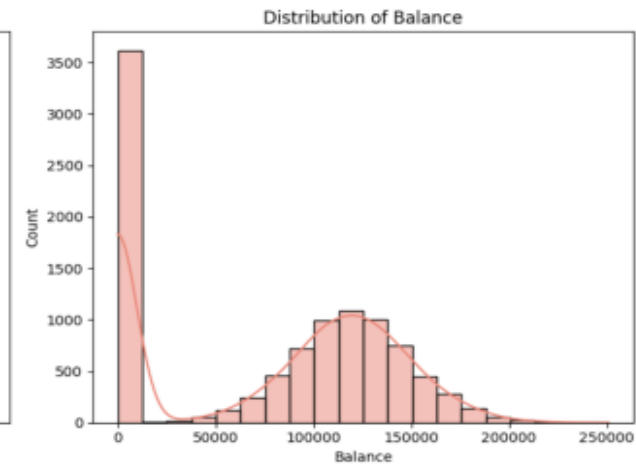
# Customer Profiles



- Most customers have credit scores around the 600-700 range
- Few have very low or very high credit scores.

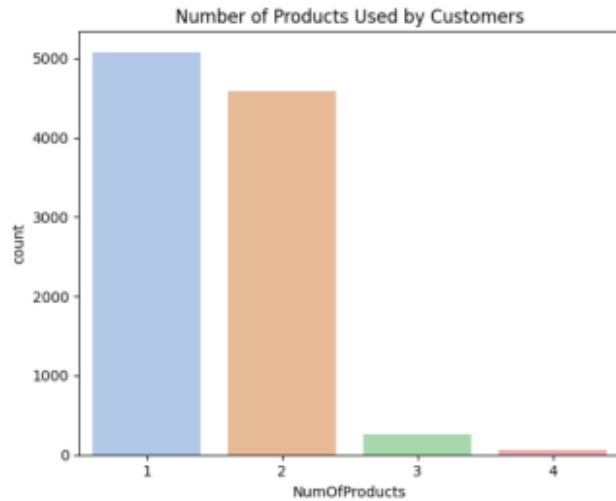


- There are more young customers than older ones in this dataset, with most people being around the 30-40 age range.
- There are fewer customers as age increases.

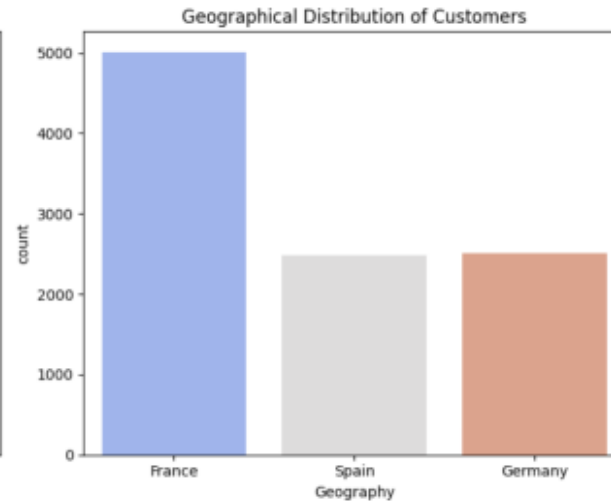


- There appears to be many customers with little to no balance in their accounts.
- Another common balance amount is between \$100,000 and \$150,000.
- The nature of this spread might suggest different segments in the customer base (i.e., accounts not used for holding balances versus accounts that maintain a moderate balance).

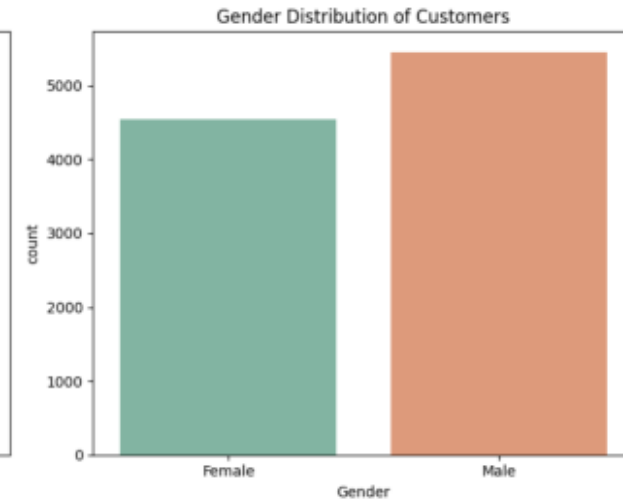
# Customer Profiles, Continued



- Most customers use 1 or 2 bank products, with a significant drop in the number of customers using 3 or 4 products.
- Customers likely prefer a smaller number of banking services.

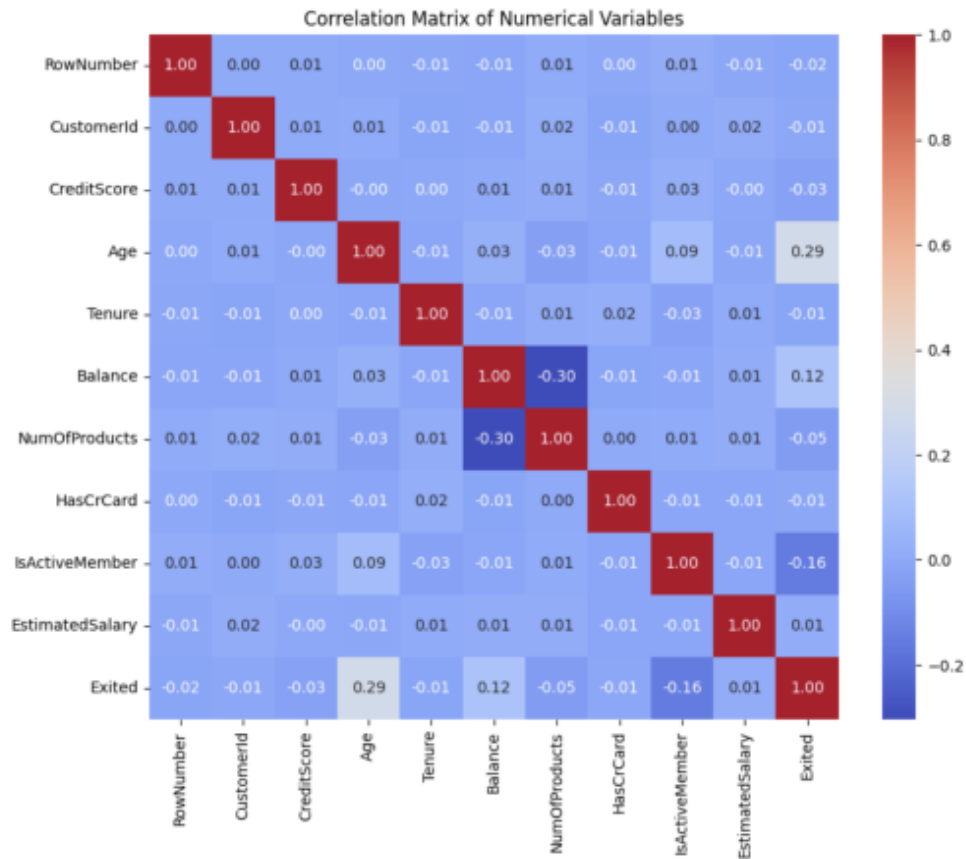


- The customers are primarily from France, with a notably lower number of customers from Spain and Germany.



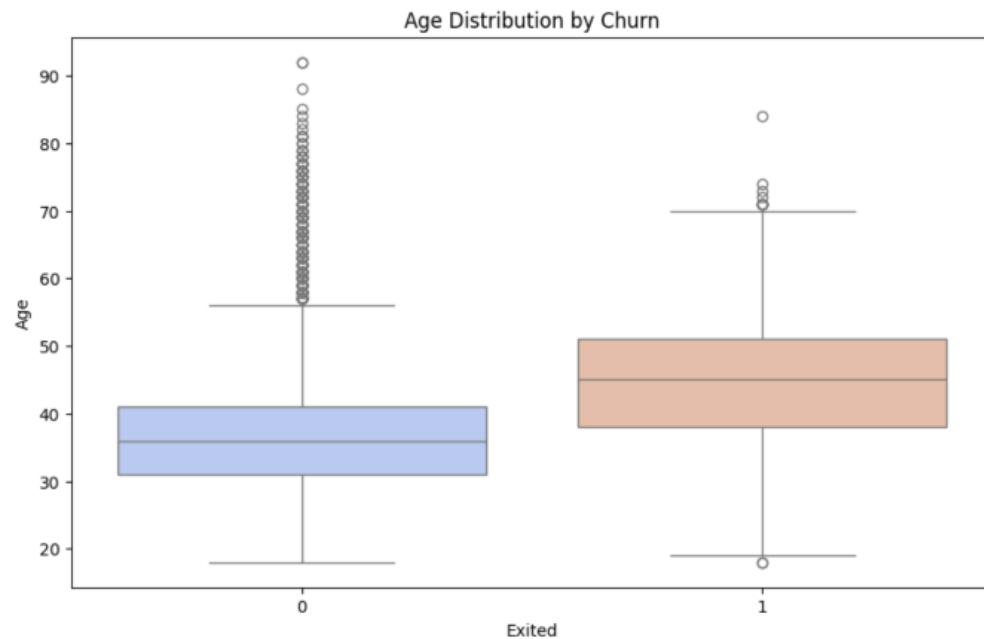
- There are more male customers than female customers in the dataset.

# Factors Influencing Churn

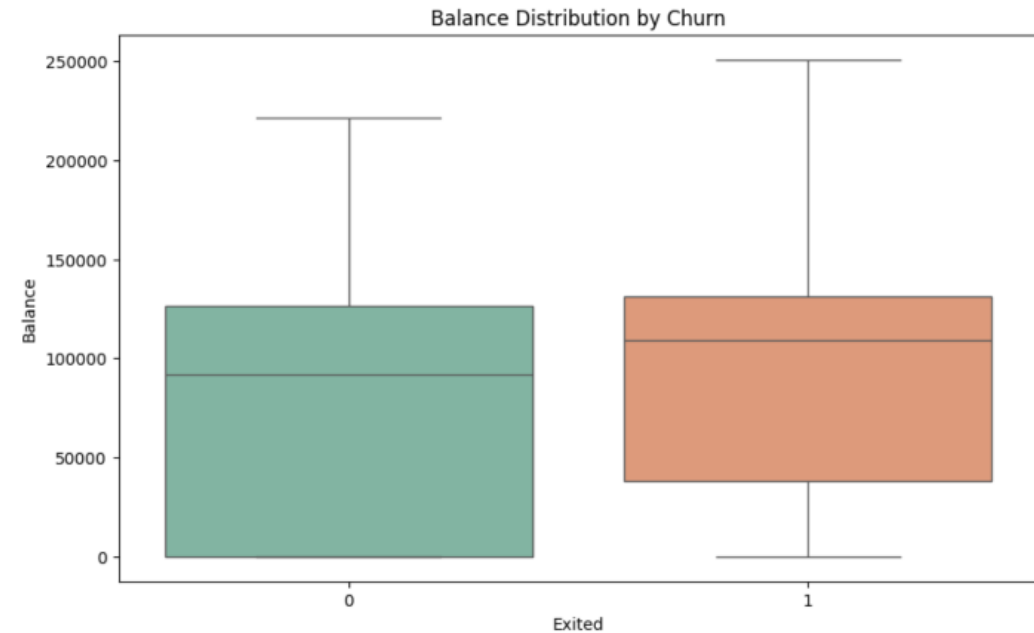


- Certain variables like Age, Balance, and IsActiveMember have more of a relationship with customer churn (Exited).
- NumOfProducts seems to play a role in both balance and churn rate.
- Variables like CreditScore and EstimatedSalary appear to have little linear relationship with churn.
- The features and relationships were further investigated through models.

# Factors Influencing Churn, Continued

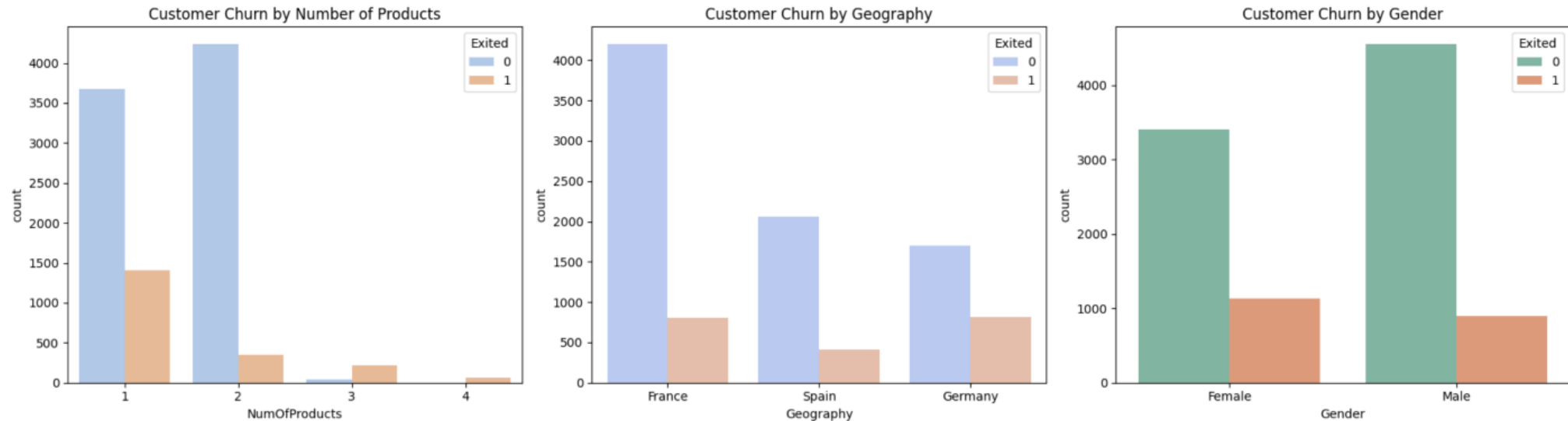


- Customers who exited (churned) tend to be older than those who stayed, with the median age of churned customers being notably higher.



- Customers who have exited generally have a higher median balance compared to those who stayed.
- The distribution for customers who stayed (not exited) shows a significant number of customers with a balance of zero, which is not as prevalent among customers who exited.

# Factors Influencing Churn, Continued Again



- Customers using multiple products have a varied churn rate, notably higher among those using 3 products.
- Churn rates vary by geography, with some regions showing higher churn.
- Gender also influences churn, with some differences observed between male and female customers.

# Customer Segmentatio n

K-MEANS  
ANALYSIS

# Customer Segments

---

Cluster	Segment Description
Young Financial Novices	<ul style="list-style-type: none"><li>• Younger customers with average credit scores, low balances, and a high number of products.</li></ul>
Young and Wealthy Individuals	<ul style="list-style-type: none"><li>• Young customers with average credit scores, high balances, and typically only one product.</li></ul>
Middle Aged and Financially Sound	<ul style="list-style-type: none"><li>• Slightly older customers with slightly higher credit scores, high balances, and a significant number of products.</li></ul>
Older and Less Financially Active	<ul style="list-style-type: none"><li>• Much older customers with average credit scores, lower balances, and a low number of products.</li></ul>

# Statistical and Linear/Non- Linear Analysis

1. HYPOTHESIS TESTING
2. SURVIVAL ANALYSIS



# Customer Balances Comparison

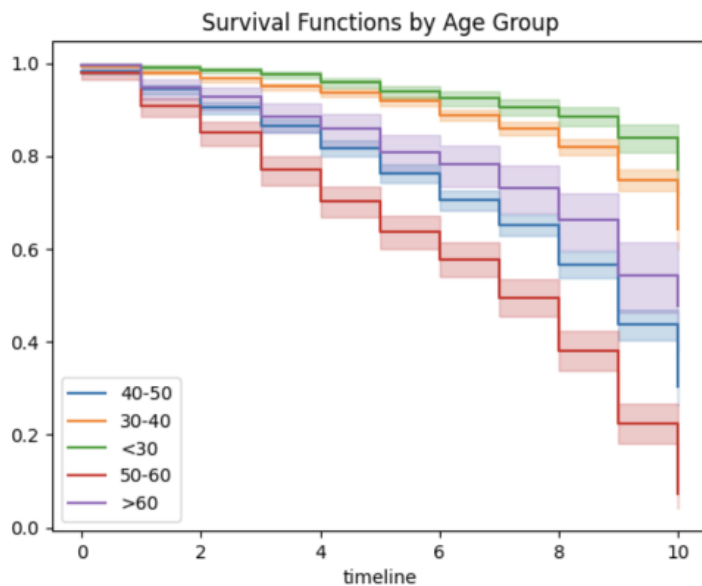
Multiple Comparison of Means – Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
France	Germany	57637.4796	0.0	54360.7642	60914.195	True
France	Spain	-274.4888	0.9791	-3565.2806	3016.3031	False
Germany	Spain	-57911.9684	0.0	-61707.2892	-54116.6476	True

	group1	group2	meandiff	p-adj	lower	upper	reject
0	France	Germany	57637.4796	0.0000	54360.7642	60914.1950	True
1	France	Spain	-274.4888	0.9791	-3565.2806	3016.3031	False
2	Germany	Spain	-57911.9684	0.0000	-61707.2892	-54116.6476	True

- Balances for customers in Germany are significantly different from those in France and Spain.
- The balances between France and Spain are not significantly different from each other.

# Tenure by Age Group

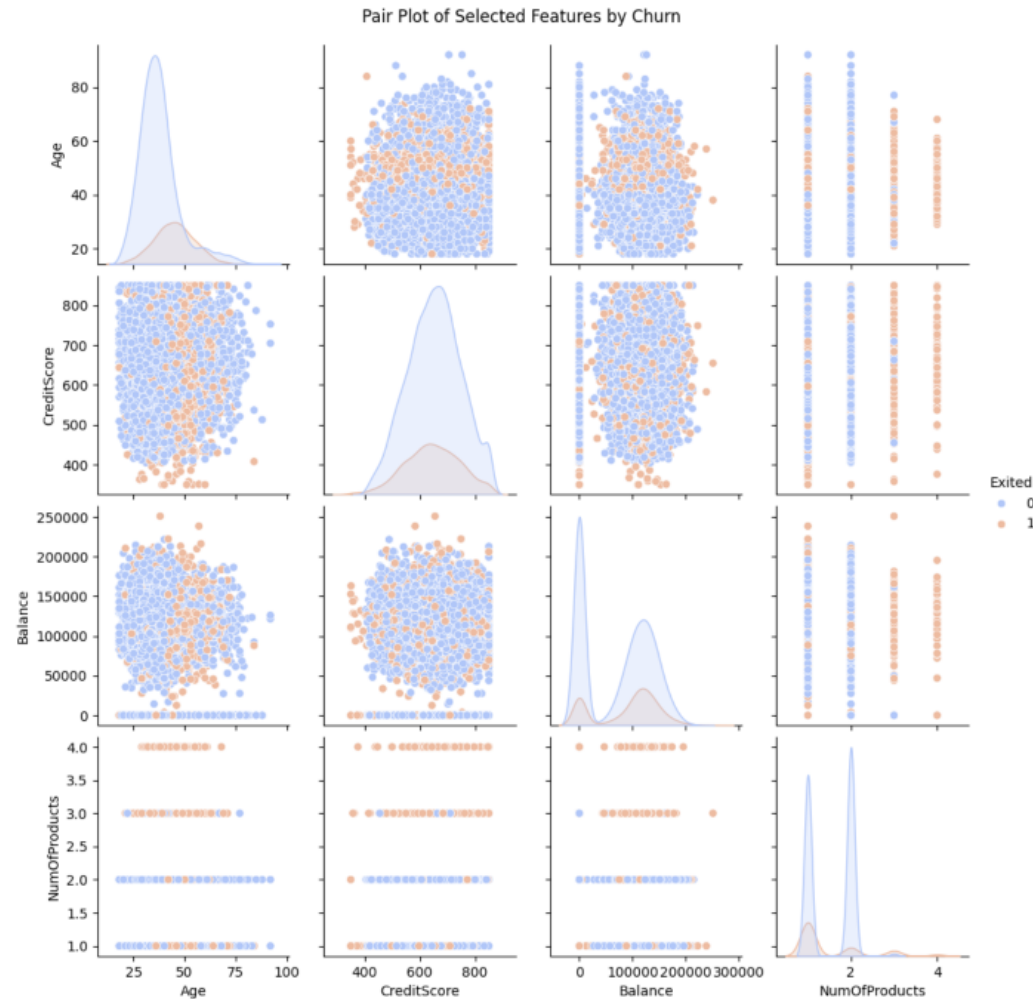


- The age groups "<30" and "40-50" appear to have the best retention over time, with the "<30" group being the most retained.
- The "50-60" age group exhibits the highest churn rate, particularly past the 5-year mark, indicating potential issues with product or service satisfaction among this cohort.
- The "30-40" age group, while starting strong, begins to churn at a faster rate after 2 years, possibly coinciding with life events that lead to a reassessment of banking needs.
- The ">60" group shows an increased risk of churn as time progresses, which might reflect changing financial needs or service expectations.
- Overall, the bank may need to investigate and address specific needs or pain points for the "50-60" and ">60" age groups to improve retention.
- Efforts to maintain the strong retention seen in the "<30" cohort could be beneficial.
- The bank might also consider targeted retention strategies for the "30-40" group after the initial few years of their relationship to reduce the observed increase in churn.

# Driving Features of Churn Per Analysis

1. PAIR PLOT
2. FACET GRID
3. COHORT  
ANALYSIS
4. FEATURE  
IMPORTANCE  
ANALYSIS

# Relationship Between Variables

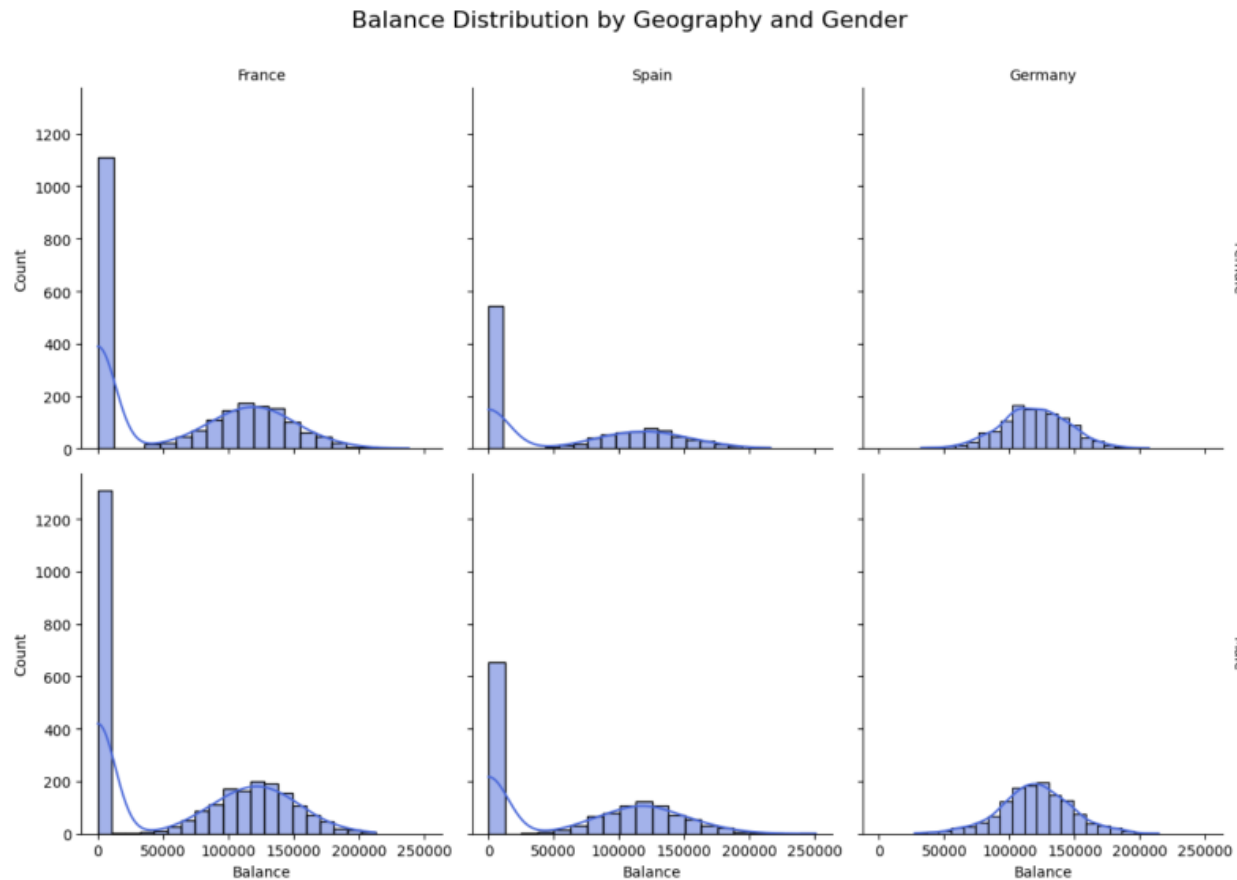


Blue: Not Churned (Exited = 0)

Orange: Churned (Exited = 1)

- Churned customers tend to be older and have higher balances across the dataset.
- There is no clear relationship between CreditScore and churn.
- Customers with a higher number of products show some increased churn, particularly for those with 3 products.
- The plot suggests that while Age and Balance may have some association with churn, CreditScore and NumOfProducts do not have a clear linear relationship with churn when considered individually.
- The data might benefit from models and further analysis to capture complex interactions between features.

# Geographic Variation and Gender Differences



- Customers in Germany tend to have higher balances on average compared to those in France and Spain.
- France has a significant number of customers with a balance close to zero, more so than Spain and Germany.
- The balance distribution between genders within each country appears relatively similar
  - Gender does not significantly influence the balance distribution within this dataset.
- For the overall level of balances, especially the higher balance accounts in Germany, regional economic factors or banking habits may influence account balances more than gender.

# Deeper Dive into Ages

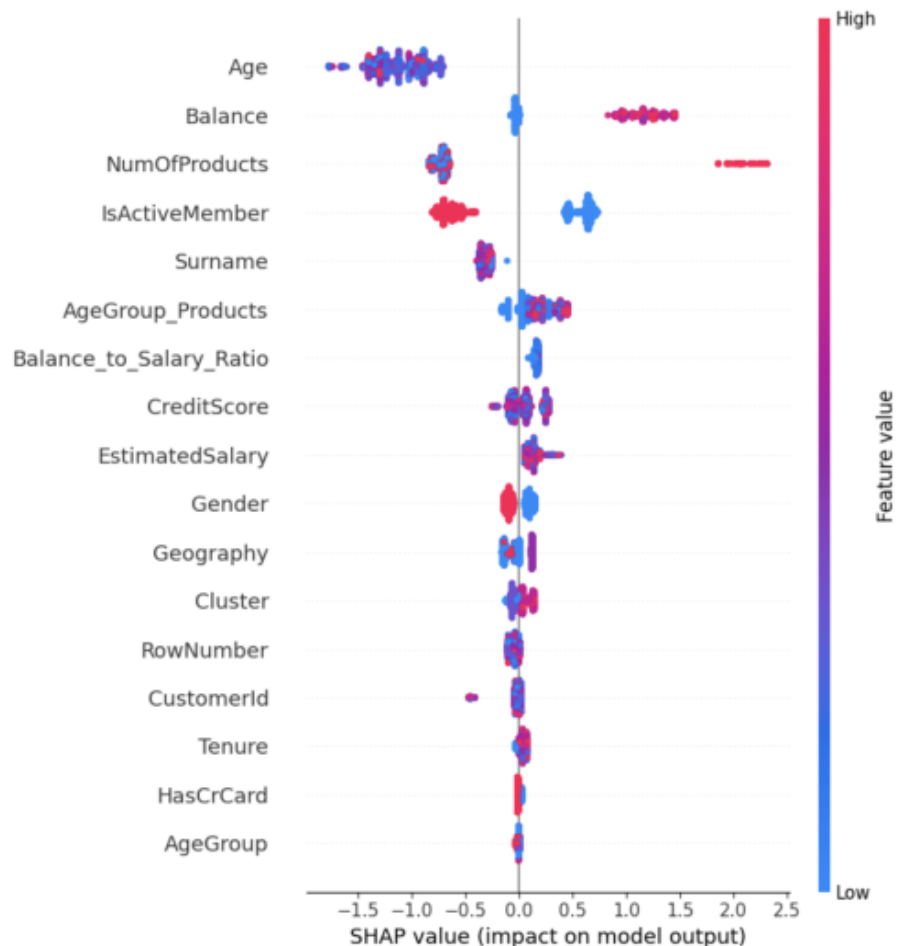
---

- Average Balance increases with age, peaking in the 50-60 age group.
  - This could indicate that older customers tend to accumulate more wealth in their accounts, possibly due to longer banking relationships or more significant financial stability.
- Churn Rate also increases with age but shows a notable spike in the 50-60 age group, with over 56% churn rate.
  - This significant churn in the 50-60 age group could indicate dissatisfaction or specific needs not being met for customers in this life stage.
- Rate slightly decreases for the >60 age group.
- These insights suggest that while older customers might hold higher balances, they also represent a higher churn risk, particularly those in the 50-60 age group.
- Tailoring services or products to the needs of this cohort could be crucial for retention strategies.

	AgeGroup	Average_Balance	Churn_Rate
0	<30	73198.764050	0.075203
1	30-40	75583.362420	0.120872
2	40-50	79122.193461	0.339655
3	50-60	82401.663162	0.562108
4	>60	75742.596401	0.247845

# Importance Ranking of Features

	Importance
<b>Age</b>	0.248300
<b>EstimatedSalary</b>	0.167262
<b>CreditScore</b>	0.161797
<b>Balance</b>	0.148759
<b>NumOfProducts</b>	0.131229
<b>Tenure</b>	0.083798
<b>IsActiveMember</b>	0.042137
<b>HasCrCard</b>	0.016718



- This SHAP summary plot represents the impact of each feature on the model's output
- Values present the strength and direction, of the relationship between features and the target variable
- Age has the most substantial overall impact on model output.
  - We see a mix of positive and negative impacts, suggesting that for some instances, increasing age may lead to a higher likelihood of the predicted outcome (likely churn), while for others, it may decrease it.
- Surname, RowNumber, and CustomerID show little to no impact, suggesting they don't have meaningful predictive power. This is expected since these features are likely identifiers with no real connection to the target variable.
- Tenure and HasCrCard appear at the bottom, indicating their overall impact on model output is minimal.

# Insights and Business Application

STRATEGIC  
IMPLICATIONS &  
BUSINESS VALUE



# Strategic Model Value

---

**Targeted Marketing:** Leveraging age and product usage insights for personalized campaigns.

**Engagement Programs:** Highlighting the significance of active membership in reducing churn, suggesting the development of loyalty and engagement initiatives.

**Financial Well-being Focus:** Utilizing insights on balance and financial ratios to offer customized financial advice or products, enhancing customer satisfaction and retention.

**Precision in Retention Efforts:** Early identification of churn risks enables proactive intervention, optimizing resource allocation and increasing ROI on retention programs.

## Outcome Benefits:

**Reduced Churn Rate:** By addressing key factors identified by the model, implement strategies that directly impact customer retention.

**Increased Customer Lifetime Value (CLV):** Improved targeting and retention strategies lead to higher customer loyalty and spending over time.

**Data-Driven Decision Making:** Empowers the marketing team with actionable insights, guiding strategy with precision and evidence.

# Key Insights

---

**Age Impact:** Customers aged 50-60 show a higher churn rate, with a notable spike over 56% churn in this age group.

**Active Membership Effect:** Active members are 54% less likely to churn, emphasizing the value of engagement programs.

**Balance and Churn:** High balance customers (>100,000) are more prone to churn, suggesting targeted financial advisory services.

**Geographical Variations:** Significant balance differences between regions, with Germany showing higher average balances and potential churn rates.

**Product Usage Influence:** Using multiple products reduces churn risk, with a negative correlation of -0.16 between NumOfProducts and Exited.

## Data Points:

Credit Score distribution peaks around 600-700, affecting churn minimally.

20.37% overall churn rate, providing a baseline for improvement targets.

Geographic analysis shows notable differences, with an ANOVA F-statistic of 958.43 indicating significant regional balance differences affecting churn.

Survival analysis reveals tenure-related churn patterns, with steeper drops around the 4-year and 7-year marks.

# Business Application

---

## Revolutionizing Customer Retention Strategies

- Leverage age-specific data to personalize marketing and enhance customer experience, significantly reducing churn among key demographics.
- Harness the power of active engagement to lower churn by 54%, transforming how loyalty programs and customer interactions are managed.
- Deploy targeted financial advisory and product offerings to high balance customers, turning potential churn into an opportunity for deeper engagement.
- Utilize regional churn insights to customize strategies, addressing unique needs and maximizing customer satisfaction across geographies.
- Promote cross-selling strategies with evidence-backed approaches, reducing churn risk by encouraging diversified product usage.

## Economic Impact

- With a 20.37% churn baseline, strategic interventions based on these insights can substantially boost retention, directly impacting the bottom line.
- Regional balance differences highlight untapped opportunities for market-specific offerings, potentially unlocking new revenue streams.
- Addressing churn at critical tenure milestones (4 and 7 years) with targeted initiatives can enhance customer lifetime value and loyalty.

# Recommendatio ns

BUSINESS VALUE

# Recommendations Per Analysis

## Targeted Engagement & Retention Initiatives

- Implement personalized engagement strategies for customers aged 50-60, potentially incorporating health, retirement planning, and lifestyle services.
- Enhance loyalty programs to significantly increase active membership benefits, focusing on digital engagement and rewards for financial product usage.

## Financial Health-Focused Offerings

- Introduce personalized financial advisory services for high balance customers, leveraging AI and machine learning for predictive financial health insights.
- Develop regional-specific product offerings based on unique churn insights, especially in high-churn regions like Germany, to cater to local financial needs.

## Product Cross-Selling and Bundle Strategies

- Launch targeted cross-selling initiatives to encourage the use of multiple products, reducing churn by showcasing the comprehensive value of your services.

## Geographical Customization

- Tailor marketing and service strategies to regional needs, based on churn and balance data, ensuring cultural and economic factors are considered.

Thank You!

# Appendix

FURTHER  
INFORMATION

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Age	0.05	1.05	0.00	0.05	0.05	1.05	1.05	0.00	27.69	<0.005	558.36
Balance	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	9.37	<0.005	66.97
NumOfProducts	-0.02	0.98	0.04	-0.10	0.06	0.91	1.06	0.00	-0.53	0.60	0.75
CreditScore	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-2.30	0.02	5.55
EstimatedSalary	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	0.28	0.78	0.36
IsActiveMember	-0.78	0.46	0.05	-0.87	-0.68	0.42	0.50	0.00	-16.35	<0.005	197.14

Concordance

0.70

Partial AIC

33147.58

log-likelihood ratio test

984.65 on 6 df

-log2(p) of ll-ratio test

693.38

# Cox Model



# Cox Model, Explained

The results from the Cox Proportional Hazards Model provide insight into the effects of various customer features on the hazard, or risk, of churn (the event) occurring.

Model and Data:

The model used is a Cox Proportional Hazards model, fitted with 10,000 observations and 2,037 observed churn events.

Model Fit:

The partial log-likelihood of -16567.79 is a measure of the fit of the model; by itself, it's not interpretable, but it's useful for comparison with other models. The log-likelihood ratio test statistic of 984.65 with 6 degrees of freedom and a highly significant  $-\log_2(p)$  indicates that the model as a whole is significantly better than an empty model (i.e., a model without any covariates). Concordance:

The concordance index of 0.70 indicates a good predictive ability of the model. A value of 0.5 would suggest no predictive discrimination, and 1.0 would indicate perfect separation of risk.

Coefficients (coef):

Age: The positive coefficient of 0.05 suggests that with each additional year, the hazard of churn increases by 5%. The  $\exp(\text{coef})$  of 1.05 supports this, indicating a 5% increase in the hazard ratio per year of age.

Balance: The coefficient is close to 0, implying a minimal direct effect on churn risk. Since balance values are likely large, even small coefficients can be significant.

NumOfProducts: The negative coefficient of -0.02 suggests that with each additional product used, the hazard decreases by 2%, although this is not statistically significant ( $p = 0.60$ ).

CreditScore: The small negative coefficient implies a very slight decrease in hazard with higher credit scores, and this effect is significant ( $p = 0.02$ ).

EstimatedSalary: The coefficient is close to 0 with a high p-value ( $p = 0.78$ ), suggesting no significant effect on churn risk.

IsActiveMember:

The negative coefficient of -0.78 is the most substantial in magnitude and is highly significant. It suggests that active members have a 54% lower risk of churn compared to inactive members ( $\exp(\text{coef}) = 0.46$ ).

Statistical Significance (p-values and confidence intervals):

Age, Balance, and IsActiveMember have p-values  $< 0.005$ , indicating their effects on churn risk are statistically significant. The confidence intervals (coef lower 95% and coef upper 95%) do not cross zero, which supports this significance.

NumOfProducts and EstimatedSalary are not statistically significant in this model, as indicated by p-values above 0.05 and confidence intervals that include zero. Partial AIC: The Akaike Information Criterion (AIC) is used for model comparison; lower values indicate a better model fit.

In summary, the model suggests that Age and IsActiveMember status are the most significant predictors of customer churn, with higher age increasing churn risk and active membership significantly reducing it. Balance shows a minimal effect, and CreditScore shows a slight but significant effect. NumOfProducts and EstimatedSalary do not appear to have a significant impact on churn risk in this model. The model's concordance index suggests it has a good ability to discriminate between those who will churn and those who will not.

# Logistic Regression

---

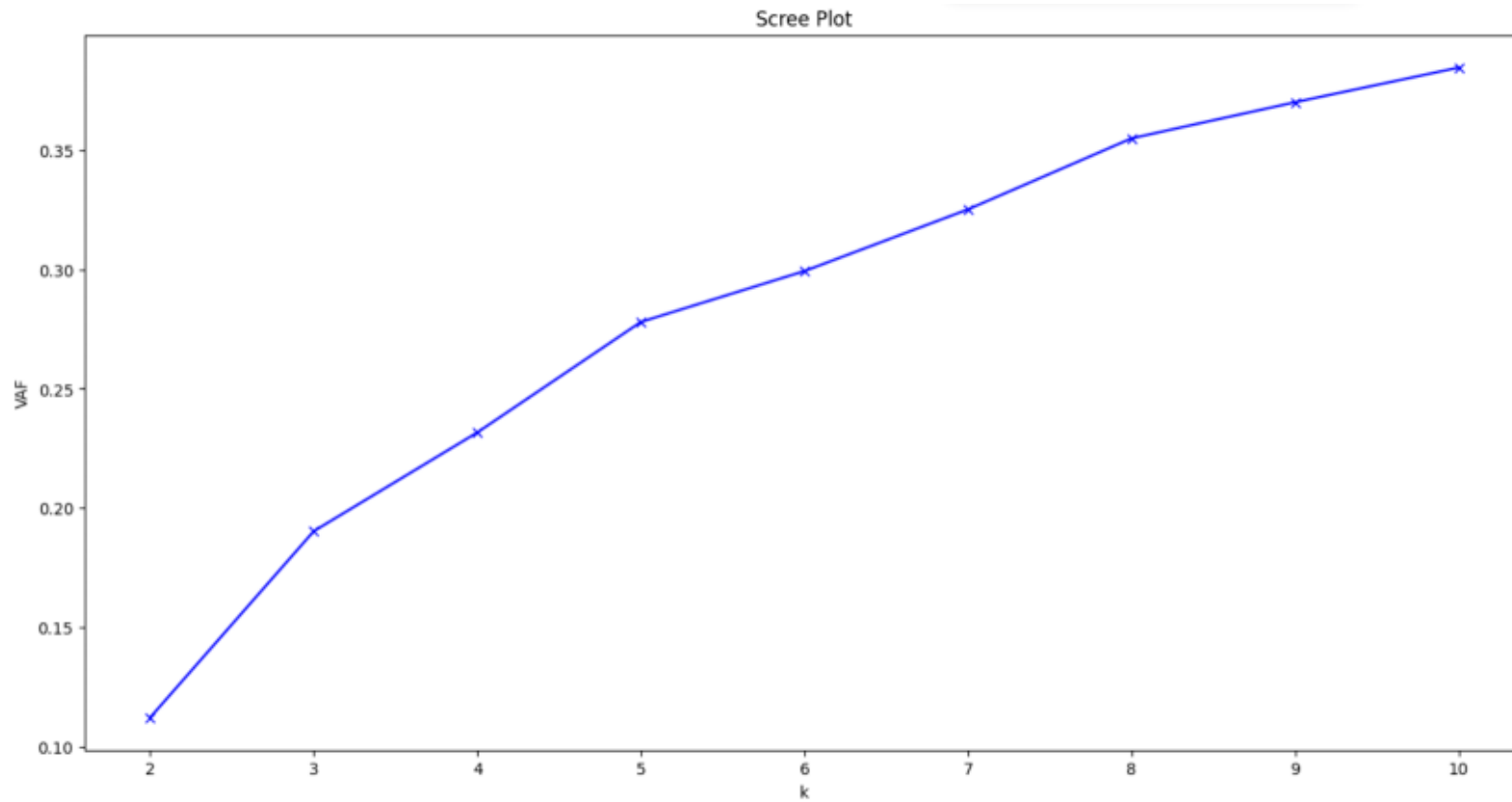
- The logistic regression model achieved an overall accuracy of 80% on the testing set, which indicates that it correctly classified 80% of the instances.
  - Accuracy alone might not provide a complete picture of the model's performance, especially in the case of imbalanced datasets, such as the one we have here.
- Looking at the classification report, we observe that the precision for predicting class 0 (customers who did not churn) is relatively high at 81%, indicating that when the model predicts a customer will not churn, it is correct around 81% of the time. However, the precision for predicting class 1 (customers who churned) is much lower at 44%, suggesting that the model's ability to correctly identify customers who churned is not as strong.
- The recall, or true positive rate, for class 1 is only 8%, indicating that the model identified only 8% of the actual churn cases correctly. This low recall suggests that the model has difficulty capturing instances of churn accurately.
- The F1-score, which considers both precision and recall, is considerably lower for class 1 (14%) compared to class 0 (89%), indicating an imbalance in the model's performance across the two classes.
- Looking at the confusion matrix, we can see that out of 584 actual churn cases, the model correctly identified only 48, while incorrectly labeling 536 of them as non-churn cases.

Accuracy: 0.80

Classification Report:					
	precision	recall	f1-score	support	
0	0.81	0.97	0.89	2416	
1	0.44	0.08	0.14	584	
accuracy			0.80	3000	
macro avg	0.63	0.53	0.51	3000	
weighted avg	0.74	0.80	0.74	3000	

# Gaussian Mixture

---



# Gaussian Mixture

FOR K = 3

Centers for the test data clusters are:

```
[[ -0.00573215  0.06680375  0.00536444]
 [ -0.06963967 -0.09993935 -0.01508731]
 [ -0.04295787 -0.02561991  0.09007848]
 [  0.10215378 -0.03404859 -0.02609881]
 [ -0.01194246  0.04983272  0.05525789]
 [  0.02405968 -0.01348148 -0.03903383]
 [  0.64920267 -1.54035103  0.64920267]
 [ -1.02583358  0.04257667  0.97481699]
 [ -0.0122847  -0.0330812  -0.04298647]]
```

Cluster Test Indicators

[1 2 0 ... 0 0 2]

```
[[ -0.00573215 -0.06963967 -0.04295787  0.10215378 -0.01194246  0.0240596
  0.64920267 -1.02583358 -0.0122847 ]
 [  0.06680375 -0.09993935 -0.02561991 -0.03404859  0.04983272 -0.01348148
 -1.54035103  0.04257667 -0.0330812 ]
 [  0.00536444 -0.01508731  0.09007848 -0.02609881  0.05525789 -0.03903383
  0.64920267  0.97481699 -0.04298647]]
```

Cluster Sizes

[684 573 743]

Centers for the original clusters are:

```
[[ -0.02047018  0.0156081  0.00682862]
 [ -0.03908934  0.02024722  0.02095549]
 [ -0.09608305  0.03148521  0.06665175]
 [  0.02451889 -0.03148259  0.00241148]
 [  0.0179642  0.02411442 -0.03730882]
 [ -0.02541531 -0.00539756  0.02899822]
 [  0.64920267 -1.54035103  0.64920267]
 [ -1.02583358  0.01835082  0.97481699]
 [  0.00997872  0.0197153  -0.02595875]]
```

The test cluster proportions are [0.342 0.2865 0.3715]

The original cluster proportions are [0.3455 0.2965 0.358 ]

VAF for the test data is:

0.18907700771335617

VAF for the original data is:

0.19024460071391294

FOR K = 4

Centers for the test data clusters are:

```
[[ 2.72455981e-02 -5.21472013e-02  6.21965474e-02  5.95389231e-02]
 [-2.01637617e-02 -9.02358301e-03 -1.05875026e-01 -8.68538899e-02]
 [ 1.89512390e+00 -2.96306611e-01 -1.68092892e-01 -2.46839567e-01]
 [ 1.58375632e-03  4.07976719e-02 -4.60622685e-03  1.03458183e-02]
 [ 1.35455691e-01  7.77899969e-01  4.65340072e-02 -9.11837332e-01]
 [-1.81508321e-01 -5.46833540e-01 -4.11549042e-02  7.25850438e-01]
 [ 3.35089482e-01  6.49202671e-01 -1.54035103e+00  6.22027359e-01]
 [ 5.52738810e-01  2.09835850e-02 -6.70518790e-03 -1.67398406e-01]
 [-1.58132884e-01  2.76509656e-02 -1.82632953e-02 -5.41610748e-02]]
```

Cluster Test Indicators

[2 3 3 ... 3 3 1]

```
[[ 2.72455981e-02 -2.01637617e-02  1.89512390e+00  1.58375632e-03
  1.35455691e-01 -1.81508321e-01  3.35089482e-01  5.52738810e-01
 -1.58132884e-01]
 [-5.21472013e-02 -9.02358301e-03 -2.96306611e-01  4.07976719e-02
  7.77899969e-01 -5.46833540e-01  6.49202671e-01  2.09835850e-02
  2.76509656e-02]
 [ 6.21965474e-02 -1.05875026e-01 -1.68092892e-01 -4.60622685e-03
  4.65340072e-02 -4.11549042e-02 -1.54035103e+00 -6.70518790e-03
 -1.82632953e-02]
 [ 5.95389231e-02 -8.68538899e-02 -2.46839567e-01  1.03458183e-02
 -9.11837332e-01  7.25850438e-01  6.22027359e-01 -1.67398406e-01
 -5.41610748e-02]]
```

Cluster Sizes

[230 670 533 567]

Centers for the original clusters are:

```
[[ 0.17501357 -0.20639674  0.01011199  0.15574961]
 [-0.04570089 -0.03044084  0.0281275  0.03137606]
 [ 1.74394266 -0.31777474 -0.15831588 -0.31512448]
 [-0.13989306  0.01850793 -0.02271479  0.06984393]
 [ 0.04827195  0.67806113  0.02172737 -0.87683171]
 [-0.3270267  -0.51534308 -0.00338682  0.7971503 ]
 [ 0.28221533  0.64920267 -1.54035103  0.63078586]
 [ 0.54341494 -0.19413674 -0.04665488  0.01562658]
 [-0.08278236  0.02684235  0.03398648 -0.02636466]]
```

The test cluster proportions are [0.115 0.335 0.2665 0.2835]

The original cluster proportions are [0.1275 0.3315 0.271125 0.269875]

VAF for the test data is:

0.23452446095566692

VAF for the original data is:

0.2316423418595197

AIC: 19101.945013657118

BIC: 20247.84529224568

# Decisions Trees, Random Forest, and XGBo

Decision Trees Best Score: 0.8518, Validation Accuracy: 0.8505

Random Forest Best Score: 0.8585, Validation Accuracy: 0.863

XGBoost Best Score: 0.8633, Validation Accuracy: 0.86

### Best Model Parameters:

Decision Trees: Max Depth: 5, Max Features: None, Min Samples Leaf: 1, Min Samples Split: 2

Random Forest: Max Depth: 10, Max Features: 'sqrt', Min Samples Leaf: 1, Min Samples Split: 2, N Estimators: 200

XGBoost: Col Sample Bytree: 0.8, Learning Rate: 0.1, Max Depth: 3, N Estimators: 100, Subsample: 1.0

### Strategic Implications:

XGBoost emerges as the best model with the highest validation accuracy and balanced precision-recall metrics, ideal for nuanced churn prediction strategies.

Fine-tuning model parameters has shown to significantly improve model performance, emphasizing the need for iterative modeling and optimization.

Integrating predictive modeling into customer relationship management can enhance decision-making, allowing for targeted intervention strategies to reduce churn.

Decision Trees		
	Class 0	Class 1
Precision	0.86	0.78
Recall	0.97	0.38
F1-Score	0.91	0.51
Random Forest		
	Class 0	Class 1
Precision	0.87	0.80
Recall	0.97	0.45
F1-Score	0.92	0.57
XGBoost		
	Class 0	Class 1
Precision	0.87	0.77
Recall	0.97	0.46
F1-Score	0.92	0.57

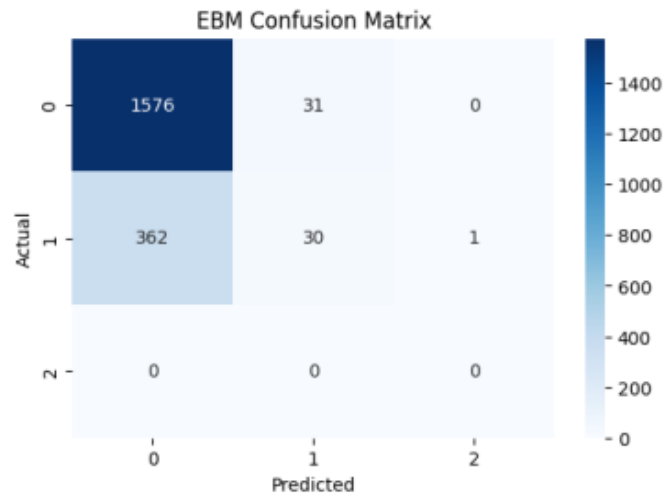
```

Classification Report:
              precision    recall  f1-score   support

     0.0         0.81      0.98      0.89       1607
     1.0         0.49      0.08      0.13        393
     3.0         0.00      0.00      0.00         0

 accuracy          0.80       2000
 macro avg         0.44      0.35      0.34       2000
 weighted avg      0.75      0.80      0.74       2000

```



# Explainable AI Method

## – EBM

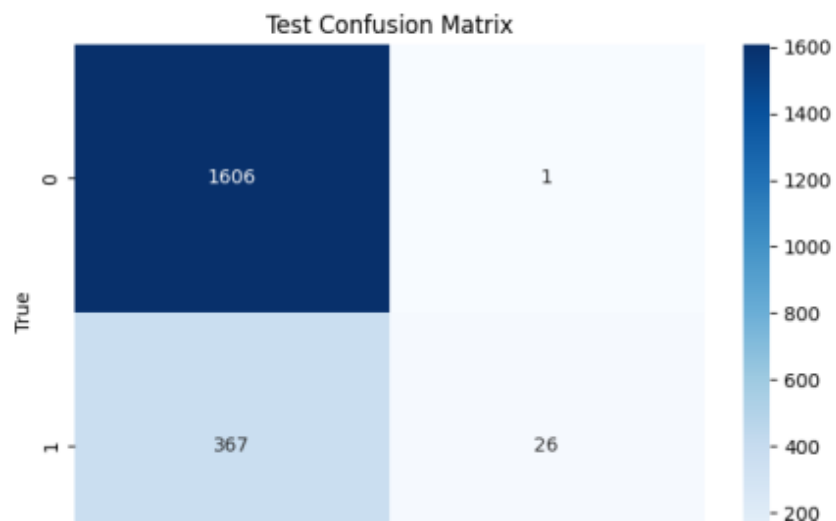
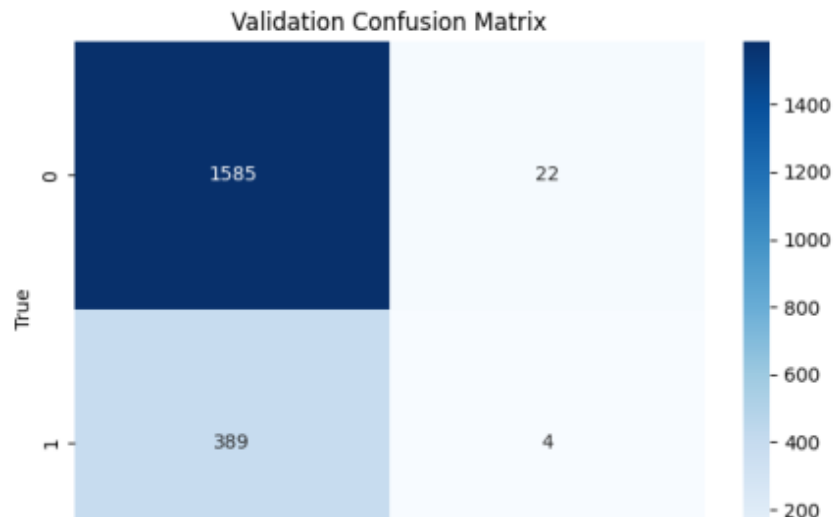
Precision for class 0 (customers who stayed with the bank) is 0.81, indicating that out of all the customers predicted to stay, 81% actually stayed. Precision for class 1 (customers who left the bank) is 0.51, suggesting that out of all the customers predicted to leave, only 51% actually left.

Recall for class 0 is 0.99, meaning that 99% of the customers who actually stayed were correctly classified. Recall for class 1 is 0.06, indicating that only 6% of the customers who actually left were correctly classified.

F1-score for class 0 is 0.89, indicating a good balance between precision and recall for customers who stayed. F1-score for class 1 is 0.11, which is quite low, suggesting that the model struggles to correctly classify customers who left.

Overall accuracy of the model is 0.80.

The model performs relatively well in predicting customers who stayed with the bank (class 0), but struggles significantly in identifying customers who left (class 1), as shown by the low recall and F1-score for class 1.



# Error Analysis

True Positives (TP): Increased from 4 to 26, indicating better positive class identification.

True Negatives (TN): Slight increase from 1585 to 1606, showing improved correct negative predictions.

False Positives (FP): Significant reduction from 22 to 1, demonstrating reduced incorrect positive predictions.

False Negatives (FN): Minor reduction from 389 to 367, indicating a slight improvement in identifying actual positives.

## Implications for Churn Prediction:

TP Improvement: Enhances the ability to engage proactively with potential churners.

TN Consistency: Essential for efficiently allocating resources to retention strategies.

FP Reduction: Crucial for avoiding unnecessary retention actions and customer dissatisfaction.

FN Reduction: Important for minimizing missed opportunities for retention and revenue preservation.

## Strategic Recommendations:

Focus on increasing TP and TN through targeted interventions based on predictive insights.

Minimize FP and FN by refining the prediction model and incorporating additional data sources for accuracy.

Regularly review model performance against new data to ensure consistency and adaptability over time.

Use error analysis insights to tailor customer communication and retention strategies, enhancing overall customer satisfaction and loyalty.