BANK CUSTOMER ATTRITION CLASSIFICATION

Maggie Huang, Jintao Li, Jean Nieto Cordova, Juan David Ovalle

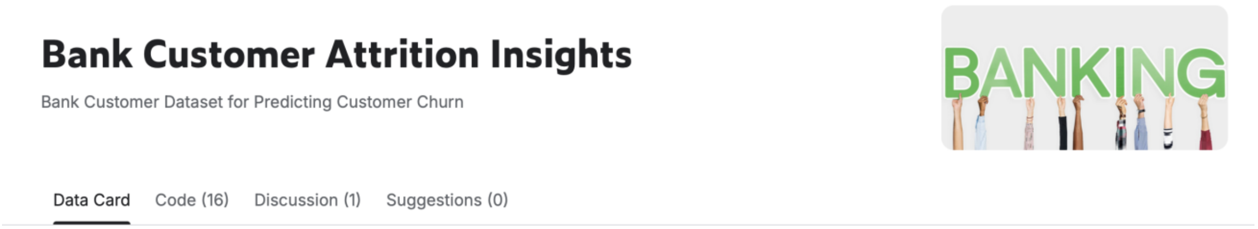
AGENDA

- Framing the Problem
- Dataset Insights
- Exploratory Data Analysis
- Feature engineering & Pipeline
- Predictive Modeling Approach & Evaluation
- Causal Inference
- Conclusion
- Threats to Validity
- Conclusions & Insights
- Lessons Learned and Next Steps

FRAMING THE PROBLEM

- In today's high competitive banking landscape, losing **even 5% of customers** can erode profitability by 25%.
- Acquiring new customers typically usually **costs 5-25 times** more than retaining them.
- Customer attrition isn't just a metric—it's a **warning signal** for brand loyalty gaps, unmet needs, and rising acquisition costs.
- Our goal is to detect the **core causes** of bank attrition, determine which features increase the likelihood of attrition through predictive modeling, and make tangible strategies to transform **short-term losses** into **long-term loyalty.**

DATASET INSIGHTS



Bank Customer Attrition Insights

Bank Customer Dataset for Predicting Customer Churn

Data Card Code (16) Discussion (1) Suggestions (0)

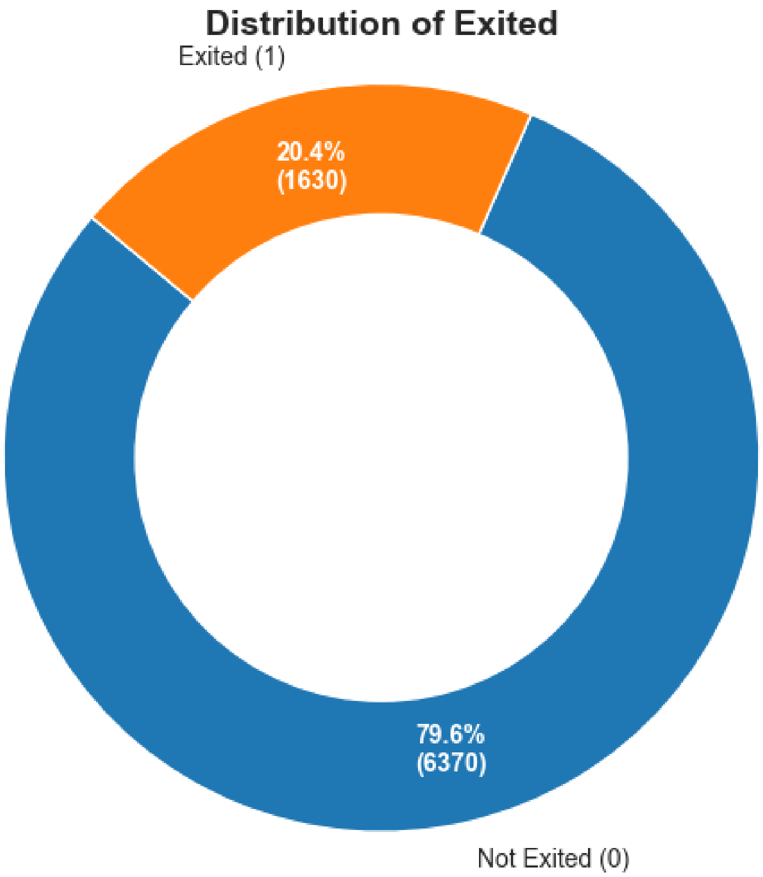
- Dataset is downloaded from Kaggle
- Dataset has **10000 rows, 18 variables**.
- Our Target variables is **“Exited”**. (Binary Variable includes 1 or 0)
- Drop 3 irrelevant columns: ‘RowNumber’, ‘CustomerId’, ‘Surname’.
- Dataset has **no missing values**

DATA EXPLORATION – CATEGORICAL VARIABLES

Target variable is highly imbalance with about 80% no attrition and 20% attrition, which means **1 in 5 customers choose to leave** the bank, causes long-term services problems.

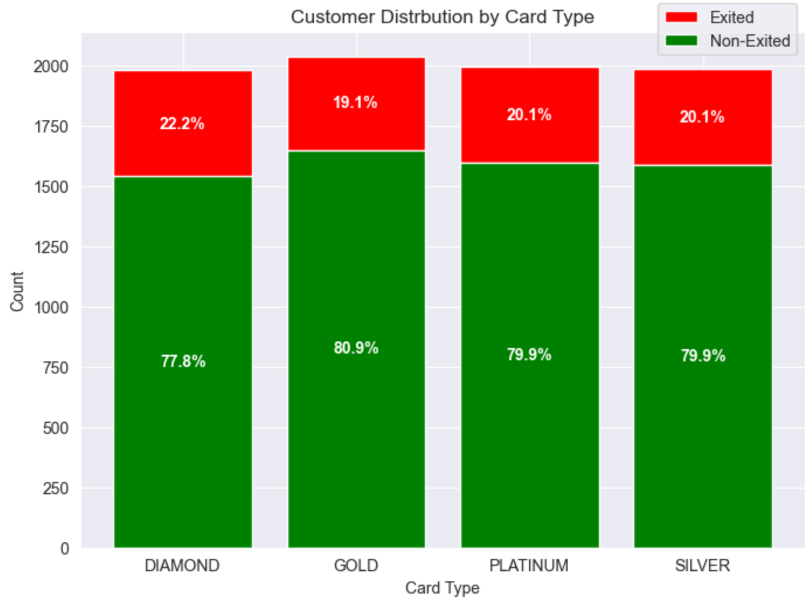
Imbalance problem will make the model to tend to predict the majority of classes, potentially ignoring lost customers

To deal with imbalance target variable problem, we will use a combination of **SMOTE and Under Sampling** techniques.



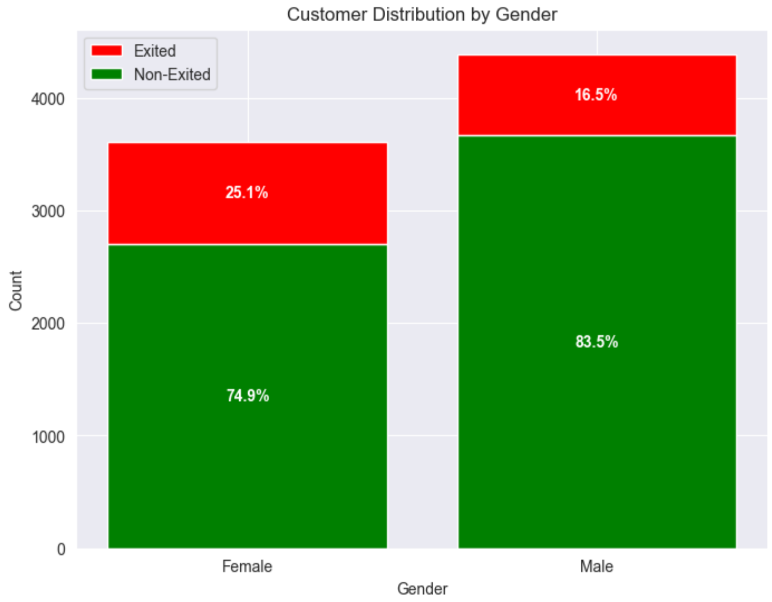
DATA EXPLORATION – CATEGORICAL VARIABLES

Card type Insights



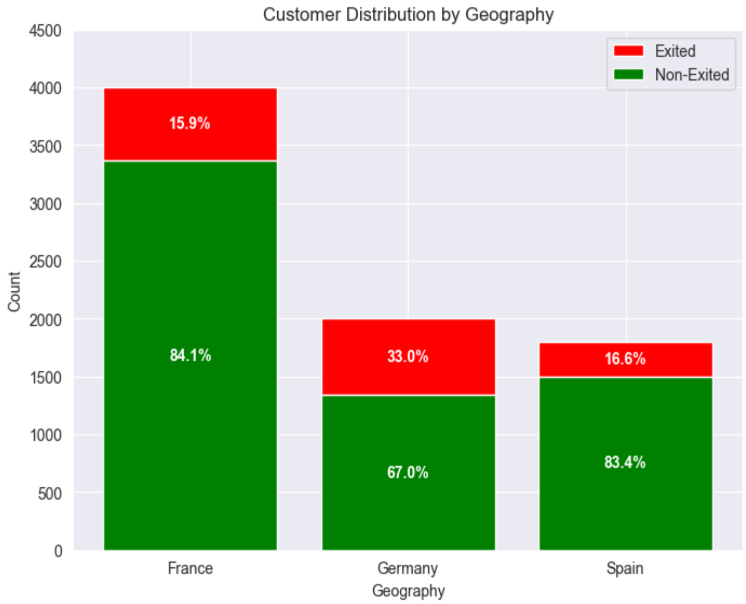
- Four types of Cards: **Gold, Platinum, Silver, Diamond**
- **Diamond** has the highest rate of attrition **with 22.2%,** followed by Gold 19.1%, this is **anti-common sense**
- With customers who exited, proportion of card types are quite similar

Gender Insights



- Compared with Male, Female has higher attrition rate, with about **25.1%** of them already exited the bank
- Female customers are **more than twice** as likely to churn as male customers
- Retention strategies for female customers should be a priority

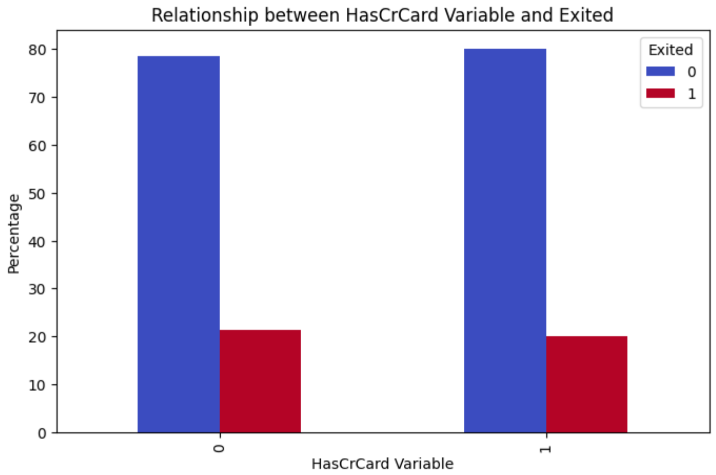
Geographical Insights



- High customer attrition in **Germany** with about **33%** of customer exited, almost the sum of exited customers from both France and Spain

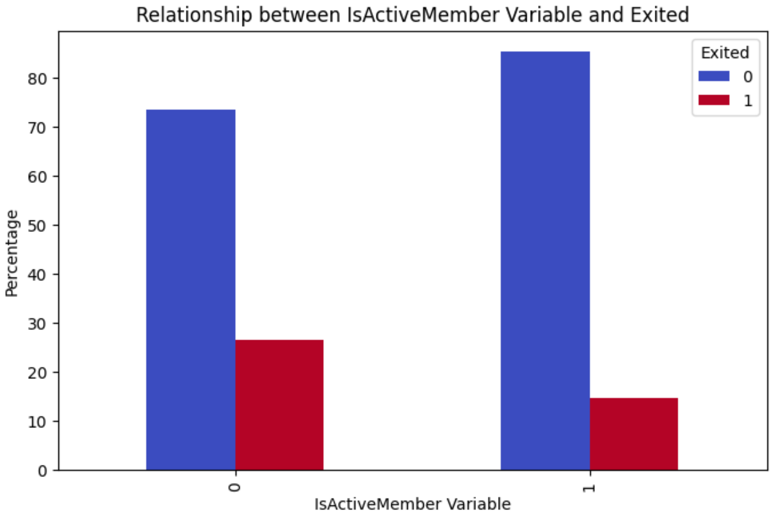
DATA EXPLORATION – BINARY VARIABLES

Credit Card Holding



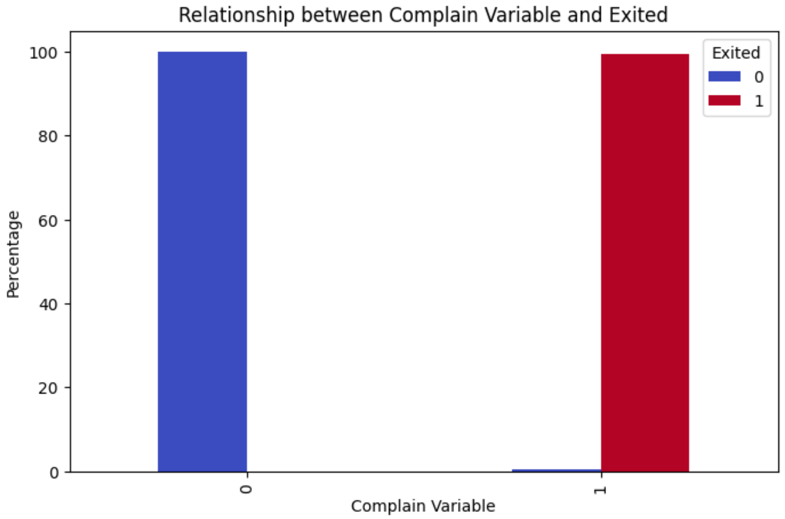
- Regardless of whether the customer has a credit card, the **attrition rate remains around 20%**
- The difference between the two groups was minimal, indicating **bank credit card may not a valid indicator** of customer loyalty

Card activation



- Obvious correlation between customer activity and loyalty
- Compared with active member 15%, **inactive member has higher attrition rate** with about 27%
- The retention rate for active customers (blue bar) is as high as 85%, compared to 73% for non-active customers

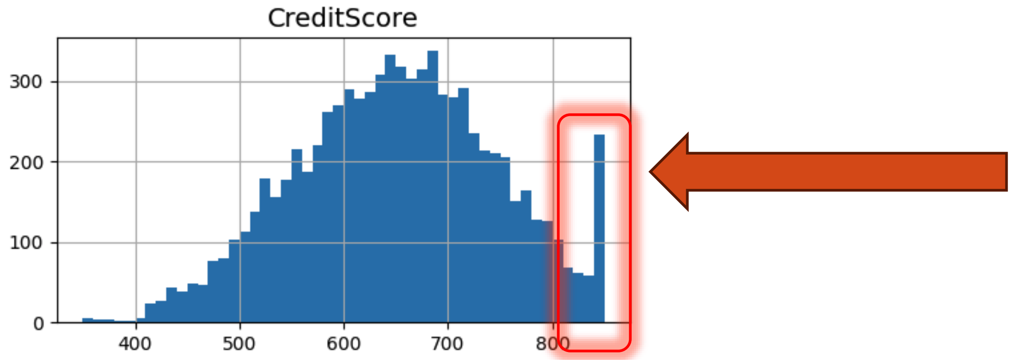
Customer Complain



- Complain is the strongest predictor of customer churn
- Almost 99% of customers with complaints left the bank. However, only a very tiny number of people who complained and stayed.

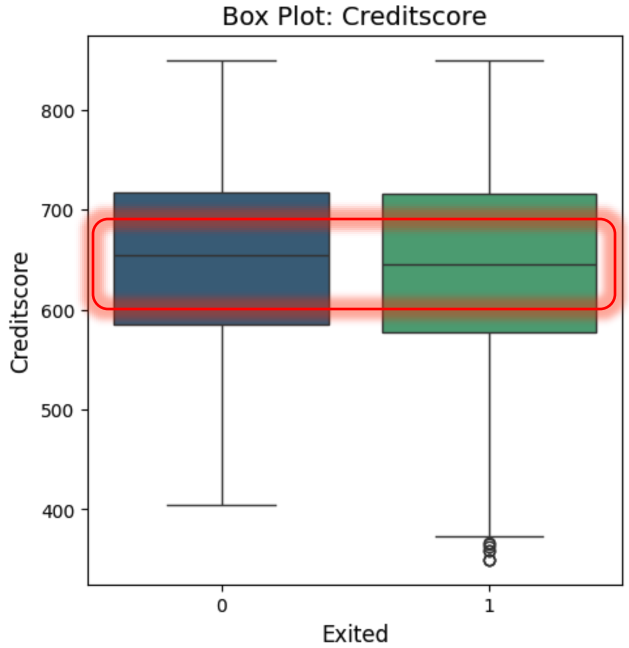
DATA EXPLORATION – NUMERICAL VARIABLE

- Distribution of Credit Score



Credit Score

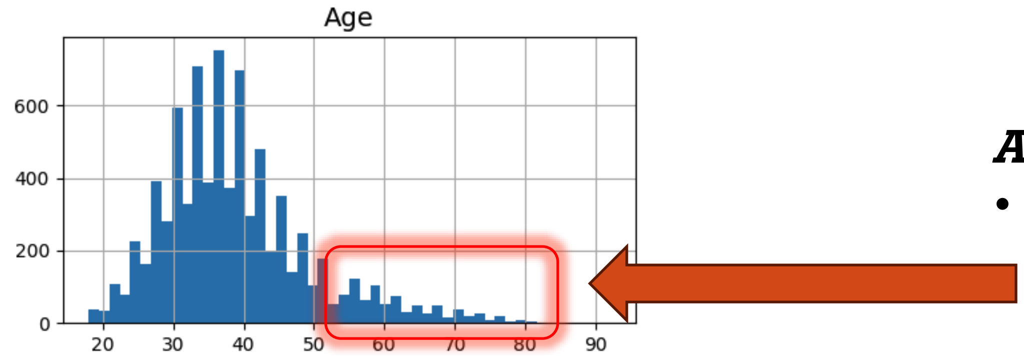
- **concentrated in between 600-700** scores; a small number of extreme low score (<500) and **big number of high score (>800)** customers



- The median credit score for both groups **were similar, around 650-680**
- Retained customers (blue on the left) have a **slightly wider spread** of credit scores
- From the green boxplot, it seems like **people who has lower credit score tend to leave the bank**
- The relationship between credit score and customer churn is not very clear

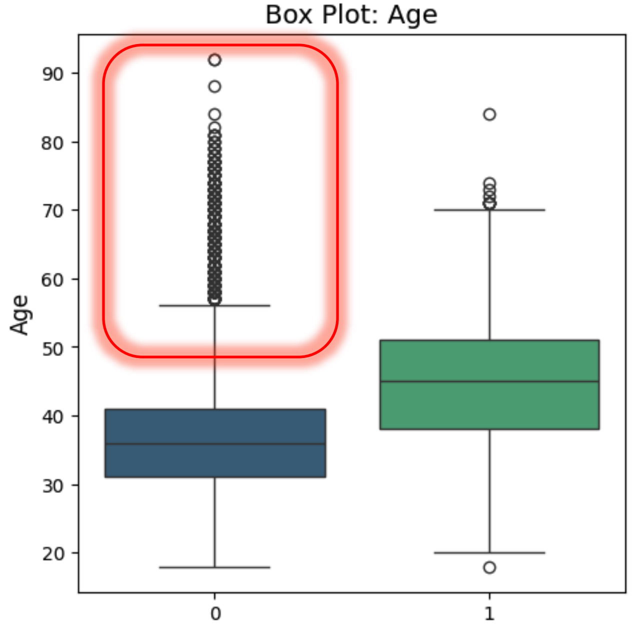
DATA EXPLORATION – NUMERICAL VARIABLE

- Distribution of Age



Age

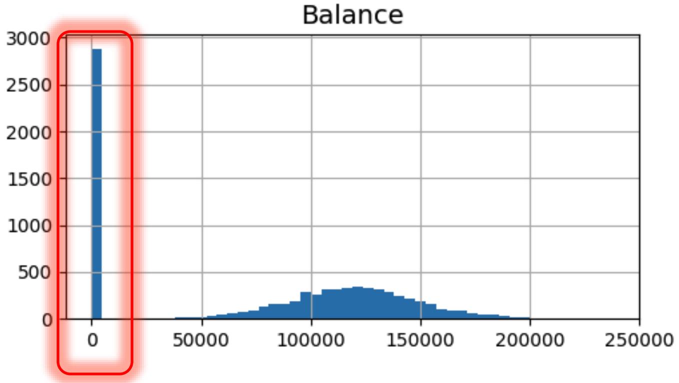
- distribution is **right skewed**, and most customers are between 25 and 45 years old

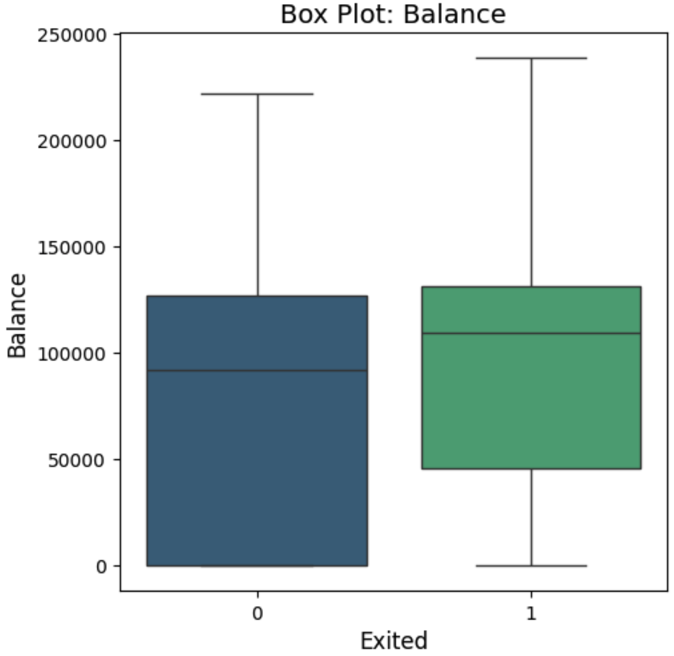


- Retained customers (blue on the left) are **generally younger**, with a median age of about 35
- The churn customers (green on the right) are **significantly older**, with a median age of about 45
- The age distribution of the two groups was clearly separated.
- This indicates that Age is a strong predictor of customer churn. The risk of leaving **increases significantly** with age

DATA EXPLORATION – NUMERICAL VARIABLE

- Distribution of Balance





Balance

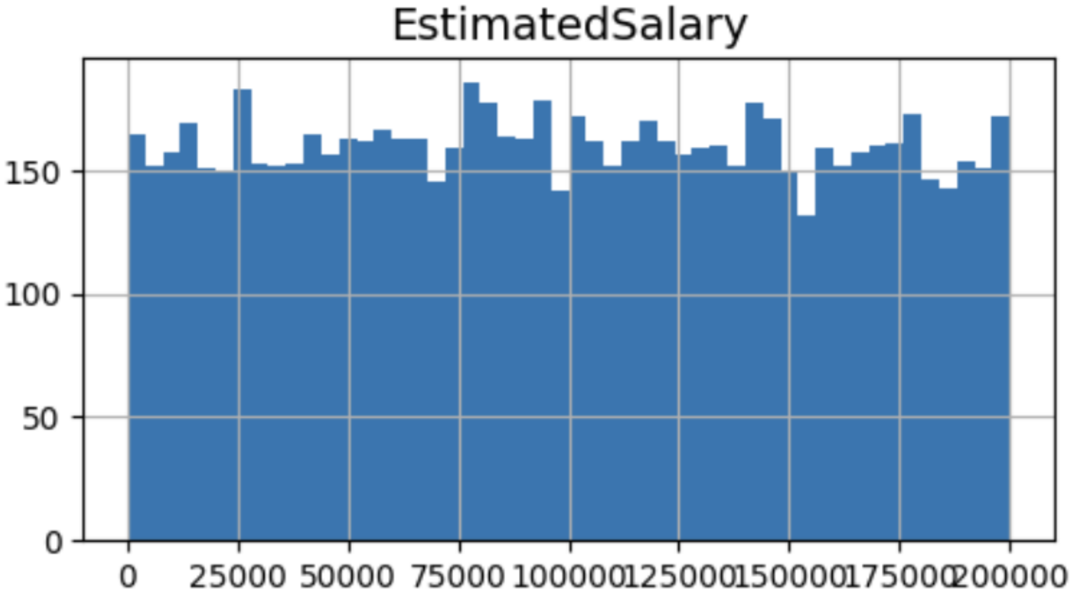
- Non-zero balance customers are **concentrated in the 100,000-150,000** range
- We may have to **treat "zero balance" as a separate feature**

- **Almost 3,000 customers** (nearly 30 per cent of the total) have completely **zero balances**

DATA EXPLORATION – NUMERICAL VARIABLE

Estimated Salary

Salary Group	Customer Not Exited	Customer Exited	Total Count
Above 1000	79.7%	20.3%	7957
Below 1000	74.4%	25.6%	43

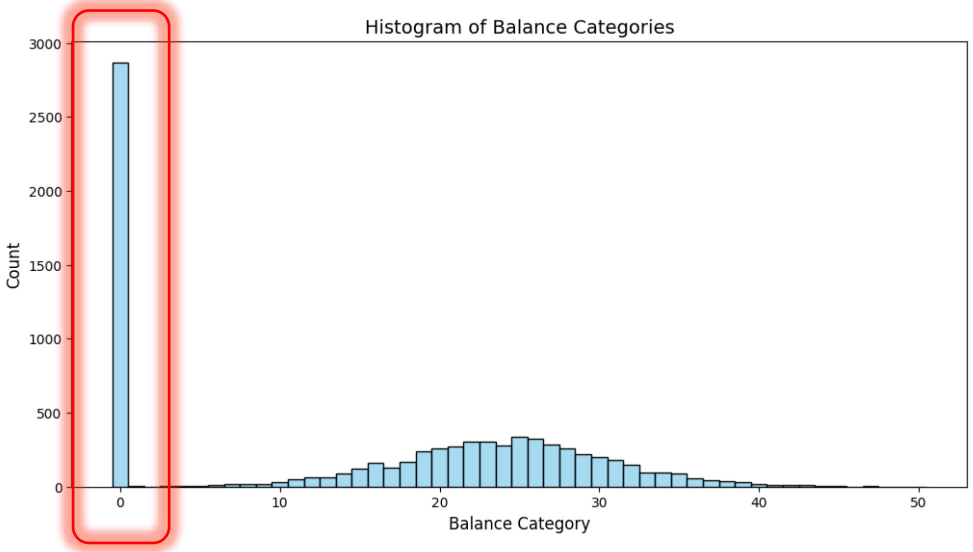


From the distribution, it is obvious that the estimated salary variable is **uniformly distributed**. However, it is abnormal and unrealistic that it has big number of salary that below 1000 and high deviation existed in our sample.

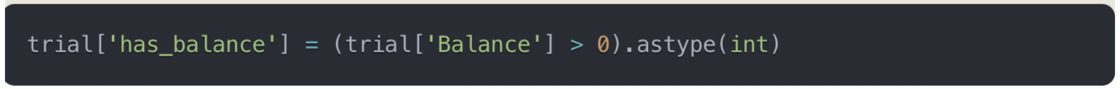
- Min value: 11.58 (Lower than 1,000 salary is low density and affect the deviation of the variable)
- These low values affect the target variable, therefore those should be removed

FEATURE ENGINEERING

- We found that there are plenty number of people whose account balance is **zero.**
- This indicates that Customers with a balance of 0 **constitute a distinct group from other customers**
- Zero balance customers may have completely different churn patterns
  - 0: balance – 13% attrition on average
  - 1: balance – 24% attrition on average

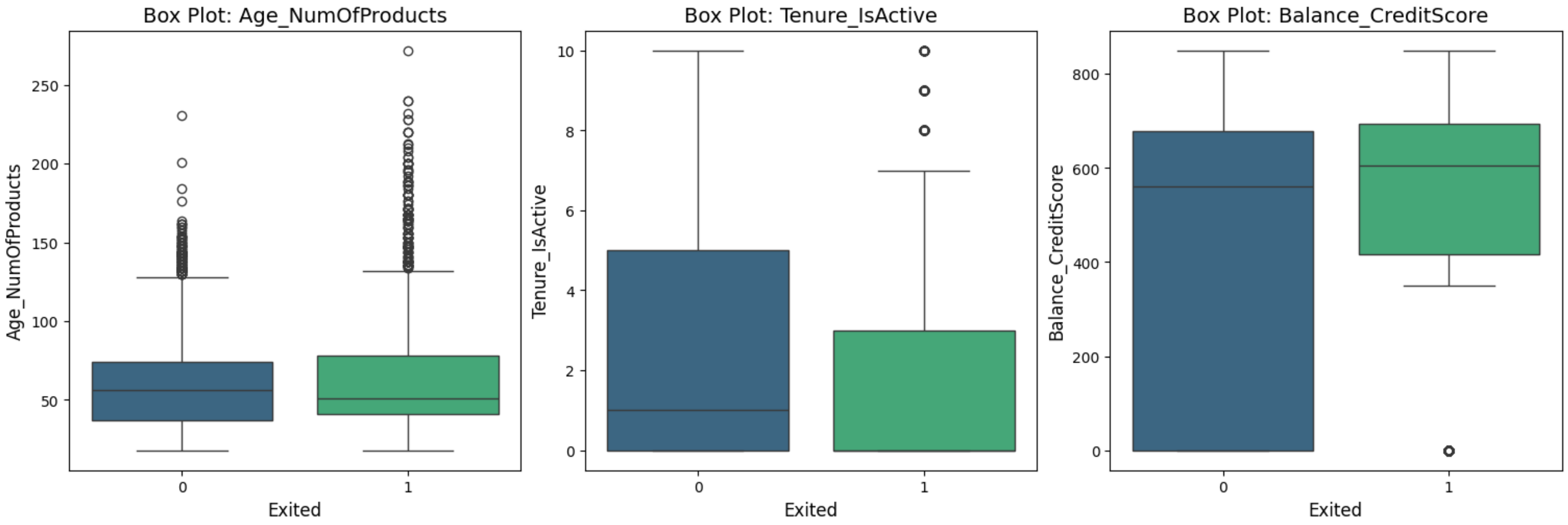


Therefore, we create a separate feature **‘has_balance’** to distinguish between **customers with a balance (>0) and those with no balance (=0)**



```
trial['has_balance'] = (trial['Balance'] > 0).astype(int)
```

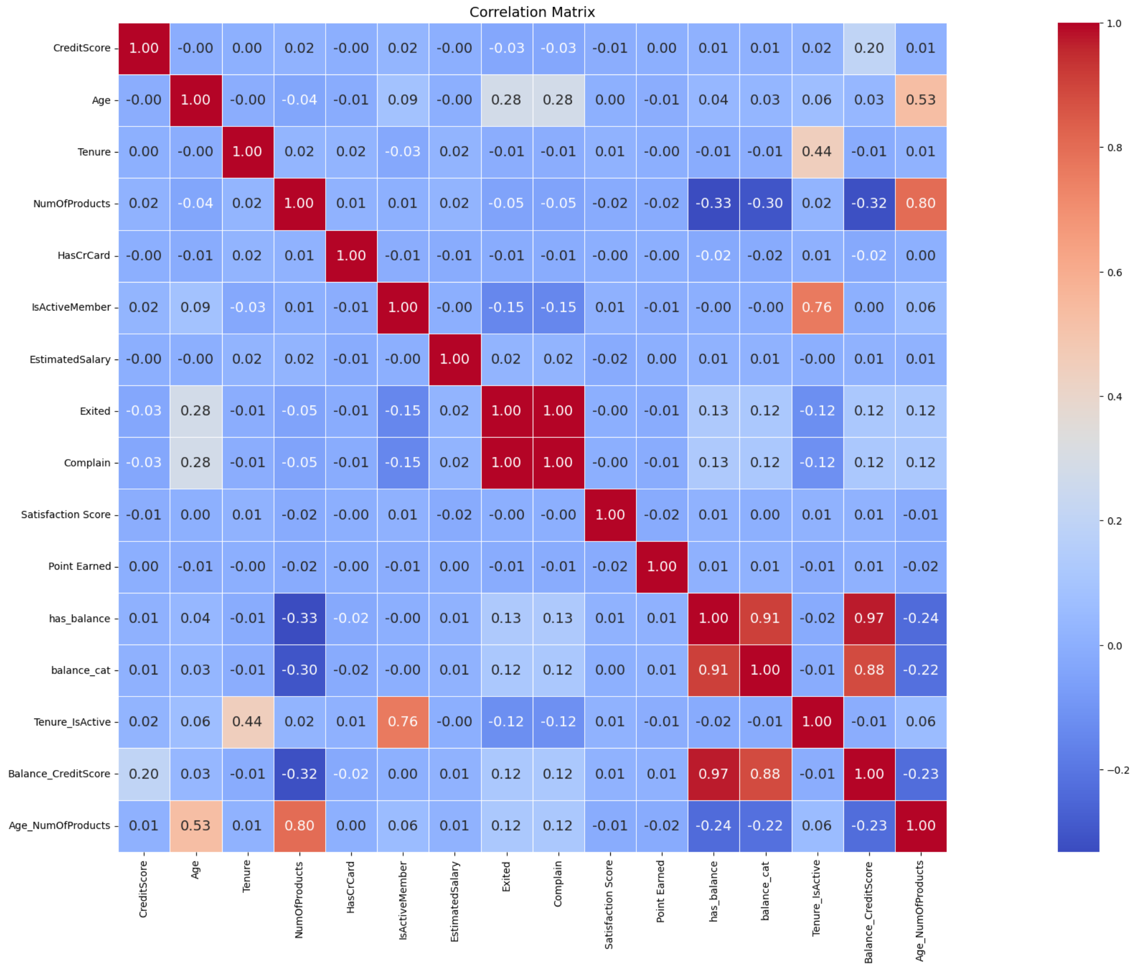
FEATURE ENGINEERING



- **Age_NumOfProduct:** A 25-year-old with 1 product is more likely to churn than a 50-year-old with 3 products
- **Tenure_IsActive:** A 2-year active customer is more likely to leave than a 10-year active customer. Higher volatility in customers with few years with the bank
- **Balance_CreditScore:** A customer with a positive balance and good credit might leave because more options available

FEATURE ENGINEERING

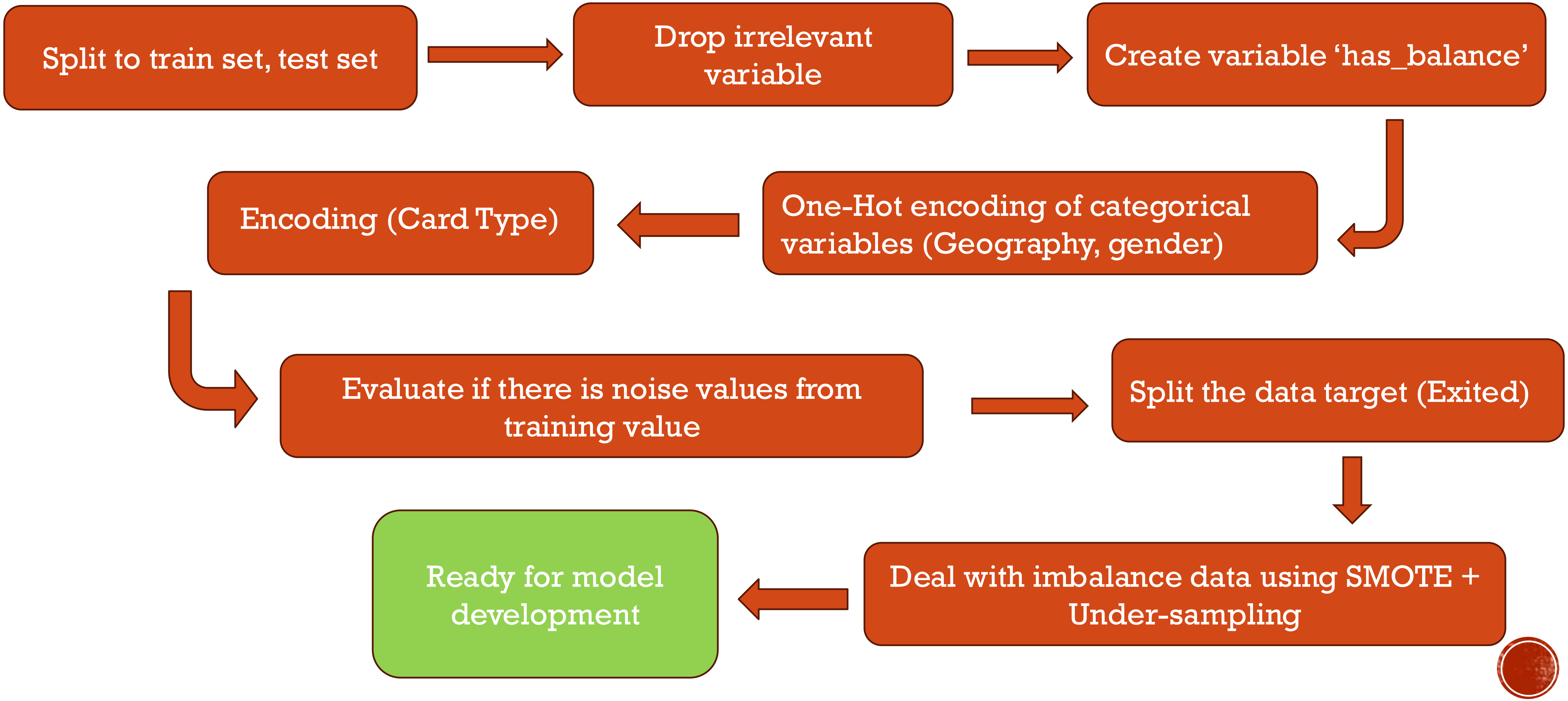
- **Complain:** Is perfectly correlated with the attrition, will be dropped
- **Balance_CreditScore:** High correlation, drop
- **Age_NumOfProduct:** High correlation, drop
- **Has_balance or Balance_cat:** use only one of this variables



Correlation Matrix

	CreditScore	Age	Tenure	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Point Earned	has_balance	balance_cat	Tenure_IsActive	Balance_CreditScore	Age_NumOfProducts
CreditScore	1.00	-0.00	0.00	0.02	-0.00	0.02	-0.00	-0.03	-0.03	-0.01	0.00	0.01	0.01	0.02	0.20	0.01
Age	-0.00	1.00	-0.00	-0.04	-0.01	0.09	-0.00	0.28	0.28	0.00	-0.01	0.04	0.03	0.06	0.03	0.53
Tenure	0.00	-0.00	1.00	0.02	0.02	-0.03	0.02	-0.01	-0.01	0.01	-0.00	-0.01	-0.01	0.44	-0.01	0.01
NumOfProducts	0.02	-0.04	0.02	1.00	0.01	0.01	0.02	-0.05	-0.05	-0.02	-0.02	-0.33	-0.30	0.02	-0.32	0.80
HasCrCard	-0.00	-0.01	0.02	0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00	-0.02	-0.02	0.01	-0.02	0.00
IsActiveMember	0.02	0.09	-0.03	0.01	-0.01	1.00	-0.00	-0.15	-0.15	0.01	-0.01	-0.00	-0.00	0.76	0.00	0.06
EstimatedSalary	-0.00	-0.00	0.02	0.02	-0.01	-0.00	1.00	0.02	0.02	-0.02	0.00	0.01	0.01	-0.00	0.01	0.01
Exited	-0.03	0.28	-0.01	-0.05	-0.01	-0.15	0.02	1.00	1.00	-0.00	-0.01	0.13	0.12	-0.12	0.12	0.12
Complain	-0.03	0.28	-0.01	-0.05	-0.01	-0.15	0.02	1.00	1.00	-0.00	-0.01	0.13	0.12	-0.12	0.12	0.12
Satisfaction Score	-0.01	0.00	0.01	-0.02	-0.00	0.01	-0.02	-0.00	-0.00	1.00	-0.02	0.01	0.00	0.01	0.01	-0.01
Point Earned	0.00	-0.01	-0.00	-0.02	-0.00	-0.01	0.00	-0.01	-0.01	-0.02	1.00	0.01	0.01	-0.01	0.01	-0.02
has_balance	0.01	0.04	-0.01	-0.33	-0.02	-0.00	0.01	0.13	0.13	0.01	0.01	1.00	0.91	-0.02	0.97	-0.24
balance_cat	0.01	0.03	-0.01	-0.30	-0.02	-0.00	0.01	0.12	0.12	0.00	0.01	0.91	1.00	-0.01	0.88	-0.22
Tenure_IsActive	0.02	0.06	0.44	0.02	0.01	0.76	-0.00	-0.12	-0.12	0.01	-0.01	-0.02	-0.01	1.00	-0.01	0.06
Balance_CreditScore	0.20	0.03	-0.01	-0.32	-0.02	0.00	0.01	0.12	0.12	0.01	0.01	0.97	0.88	-0.01	1.00	-0.23
Age_NumOfProducts	0.01	0.53	0.01	0.80	0.00	0.06	0.01	0.12	0.12	-0.01	-0.02	-0.24	-0.22	0.06	-0.23	1.00

FEATURE ENGINEERING – DATA PREPARATION



Split to train set, test set

Drop irrelevant variable

Create variable ‘has_balance’

One-Hot encoding of categorical variables (Geography, gender)

Encoding (Card Type)

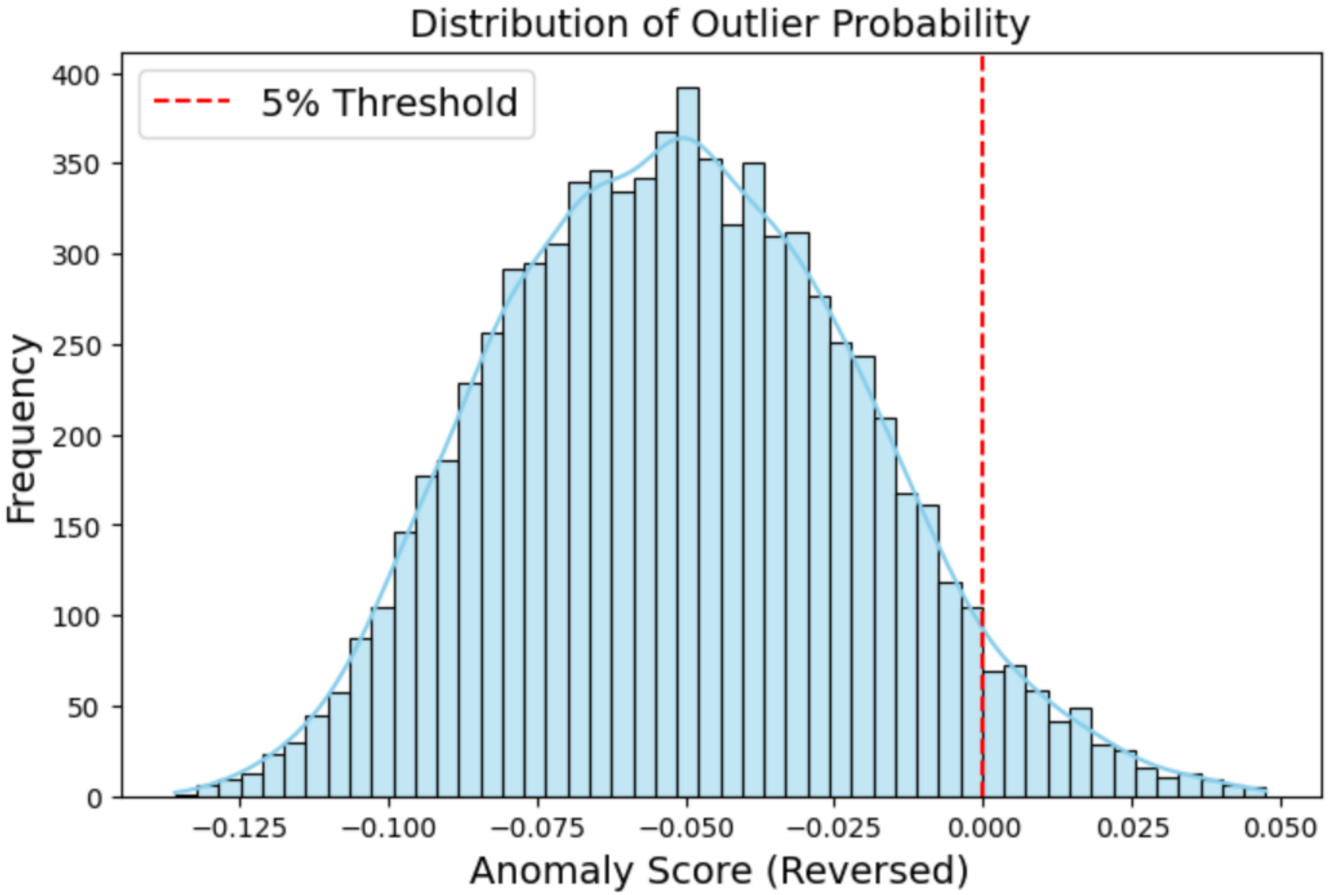
Evaluate if there is noise values from training value

Split the data target (Exited)

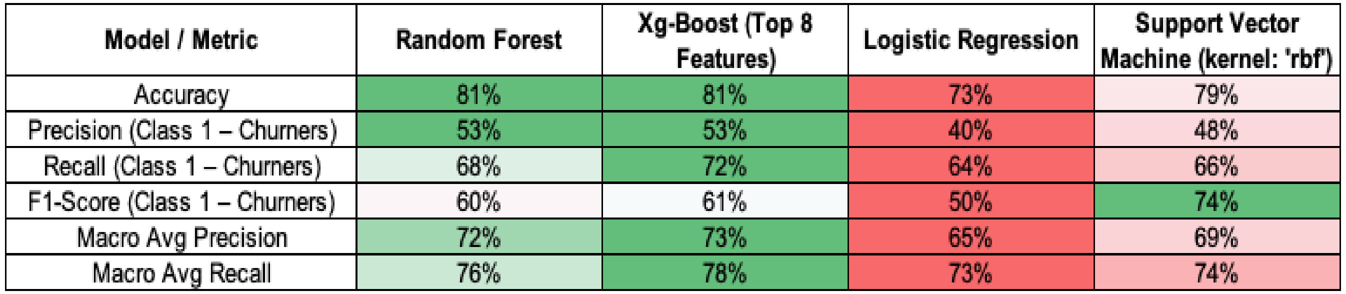
Deal with imbalance data using SMOTE + Under-sampling

Ready for model development

ISOLATION FOREST: OUTLIERS



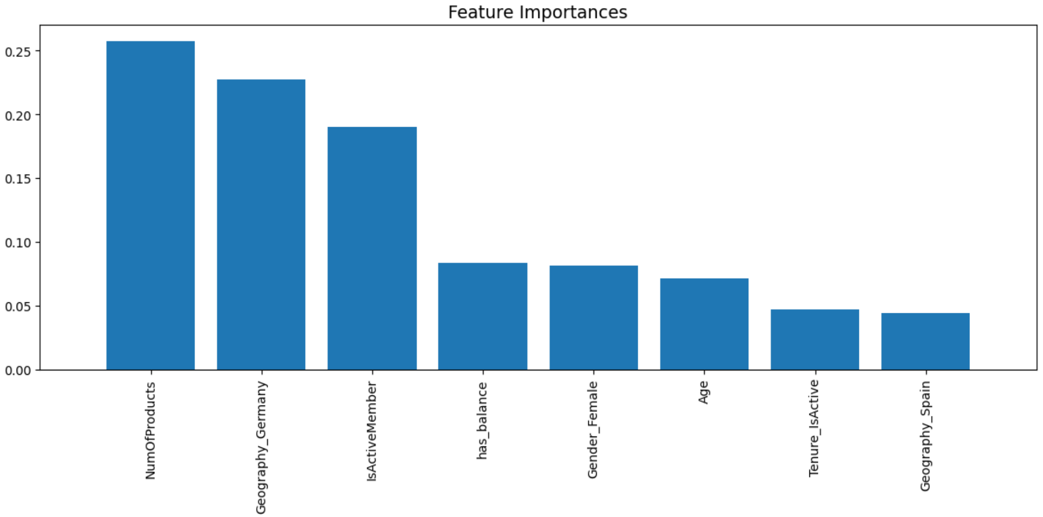
MODEL PERFORMANCE & COMPARISON



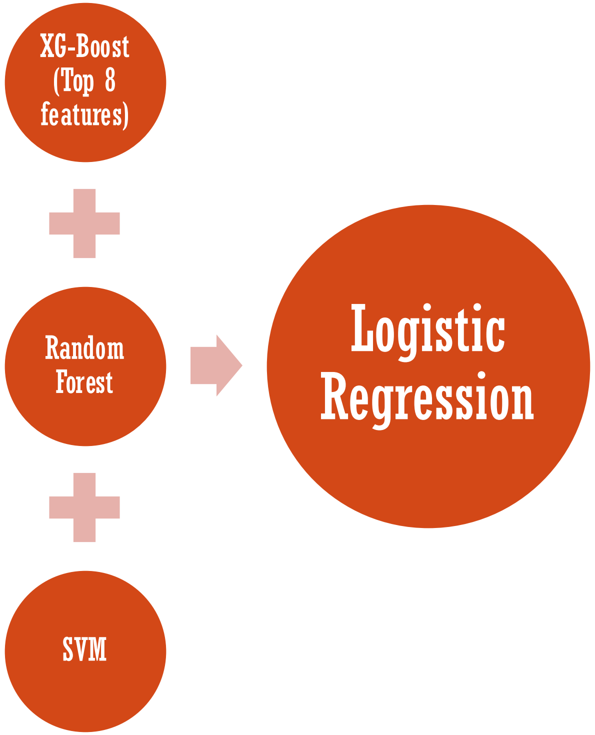
Model / Metric	Random Forest	Xg-Boost (Top 8 Features)	Logistic Regression	Support Vector Machine (kernel: 'rbf')
Accuracy	81%	81%	73%	79%
Precision (Class 1 – Churners)	53%	53%	40%	48%
Recall (Class 1 – Churners)	68%	72%	64%	66%
F1-Score (Class 1 – Churners)	60%	61%	50%	74%
Macro Avg Precision	72%	73%	65%	69%
Macro Avg Recall	76%	78%	73%	74%

- The number of products a customer holds has the highest impact on churn. Maybe too may products might overwhelm customers (Causal Inference).
- Geography significantly influences customer churn, highlighting the need for tailored services in different regions.

- Focus Metric (Recall): Prioritizing recall ensures that more churners are correctly identified, allowing the company to take proactive measures.
- To fine-tune the Xg-Boost model, a randomized search of the hyper-parameters was conducted; based on the best parameters, slightly modifications were tested through a Grid-Search Approach.



MODEL PERFORMANCE & COMPARISON



XG-Boost (Top 8 features)

Random Forest

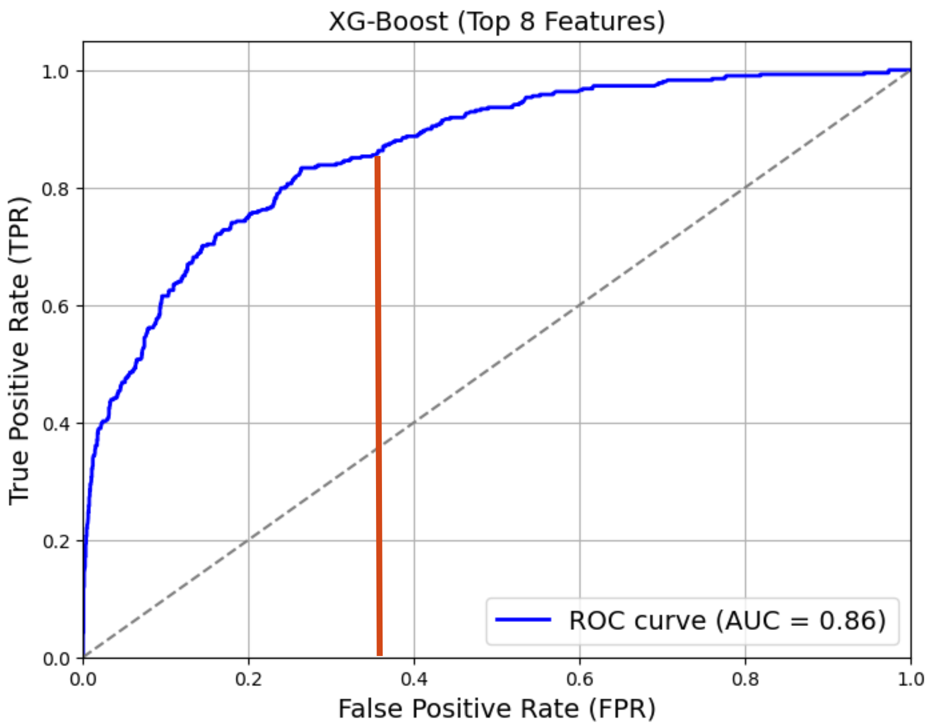
SVM

Logistic Regression

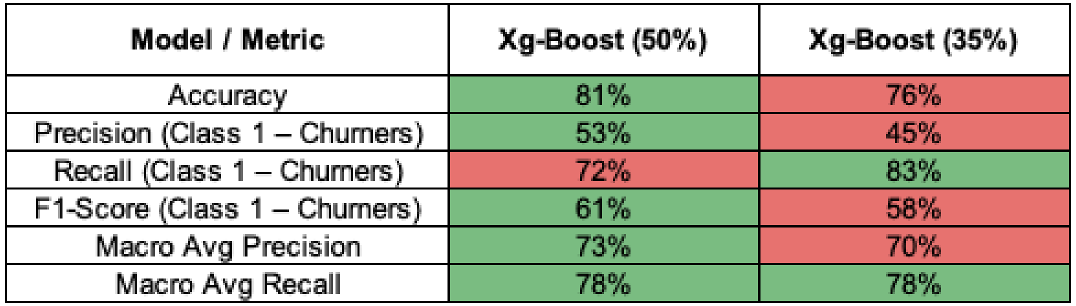
- Stacking is the most useful when base models have diverse strengths – if models are similar, the gain will be small (as seen here).
- Stacking adds complexity but does not bring significant recall improvement.

Model / Metric	Xg-Boost	Staking
Accuracy	81%	82%
Precision (Class 1 – Churners)	53%	54%
Recall (Class 1 – Churners)	72%	72%
F1-Score (Class 1 – Churners)	61%	62%
Macro Avg Precision	73%	73%
Macro Avg Recall	78%	78%

MODEL PERFORMANCE & COMPARISON



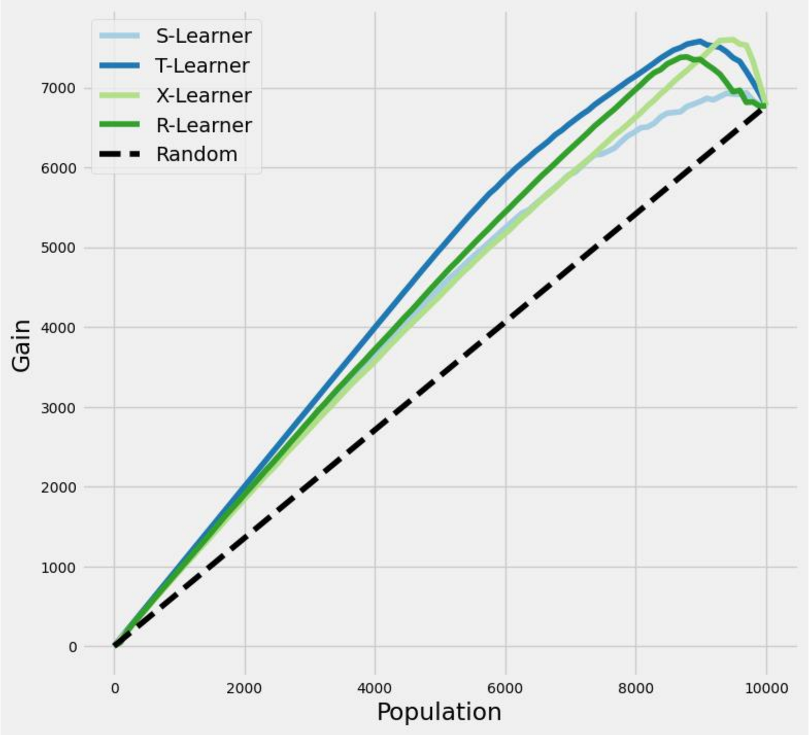
- Since our main metric is recall, adjusting the classification threshold is an effective way to improve the model's performance.
- We looked for the best threshold by evaluating different probability cutoffs and selecting the one that maximizes recall without sacrificing too much precision.
- Optimized Threshold = 0.35
- If false-positives are costly, a 0.5 threshold would be preferable



Model / Metric	Xg-Boost (50%)	Xg-Boost (35%)
Accuracy	81%	76%
Precision (Class 1 – Churners)	53%	45%
Recall (Class 1 – Churners)	72%	83%
F1-Score (Class 1 – Churners)	61%	58%
Macro Avg Precision	73%	70%
Macro Avg Recall	78%	78%

CAUSAL INFERENCE

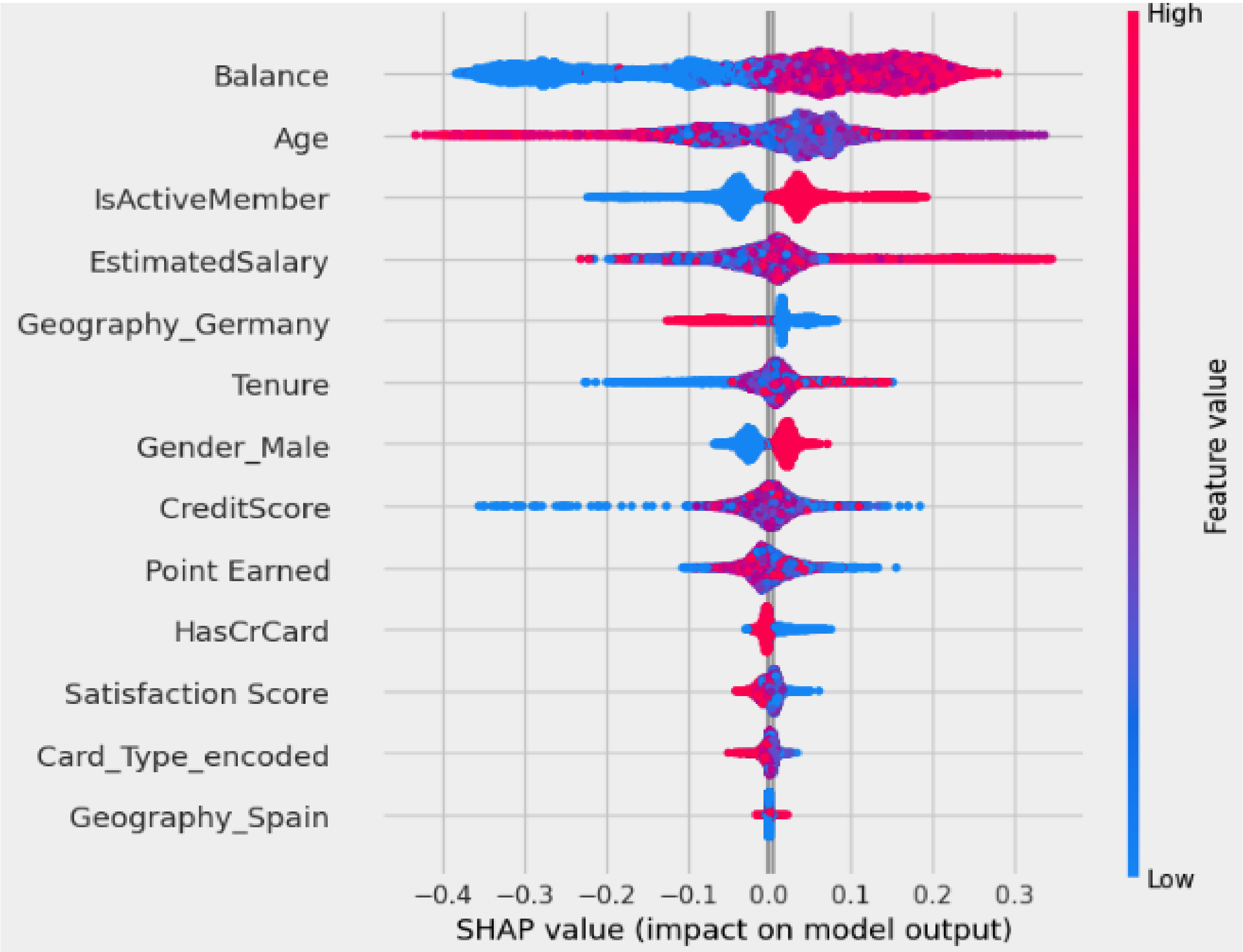
- Objective: Estimating the effect of having more than 2 products on the probability of leaving the bank.



- The model that best differentiates between those with more than 2 products and those who not is the T learner.
- Possible Reasons:
  - Product Overload & Complexity
  - High Fees & Hidden Costs
  - Financially Savvy Customers

Meta-learner	ATE	AUUC
S	0.566	0.612
T	0.598	0.675
X	0.580	0.625
R	0.575	0.642

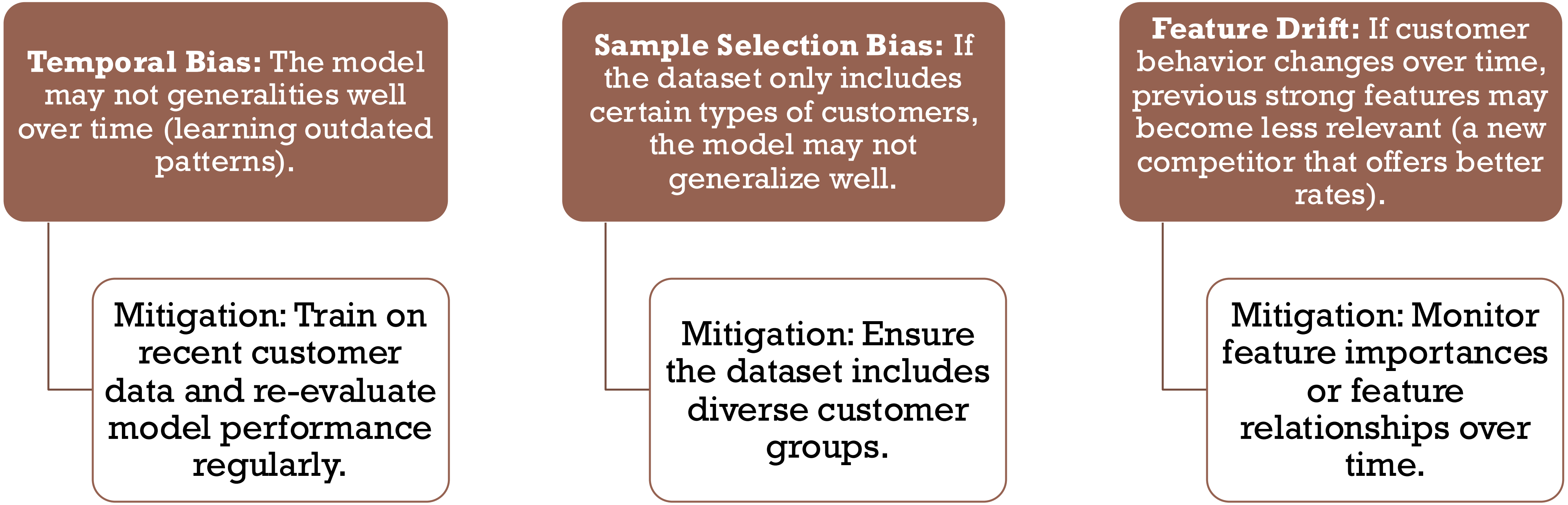
CAUSAL INFERENCE



- Customers with higher balance have a higher likelihood of leaving the bank when they have more than 2 products.
- Older customers are less likely to churn given that they have more than 2 products.

Feature	Feature importance
Age	0.295
Balance	0.230
Estimated Salary	0.128

THREATS TO VALIDITY



Temporal Bias: The model may not generalities well over time (learning outdated patterns).

Mitigation: Train on recent customer data and re-evaluate model performance regularly.

Sample Selection Bias: If the dataset only includes certain types of customers, the model may not generalize well.

Mitigation: Ensure the dataset includes diverse customer groups.

Feature Drift: If customer behavior changes over time, previous strong features may become less relevant (a new competitor that offers better rates).

Mitigation: Monitor feature importances or feature relationships over time.

CONCLUSIONS & INSIGHTS

High-balance customers are more financially aware, expect premium service, and won’t tolerate unnecessary fees or complexity—so they leave if they find better options

Older customers stay because they are loyal to the bank or find switching too time-consuming due to paperwork.

Regional differences play a significant role in customer churn, suggesting that a one-size-fits-all approach may be ineffective. To improve retention, businesses should implement region-specific strategies tailored to the unique behaviors and needs of customers in different locations.

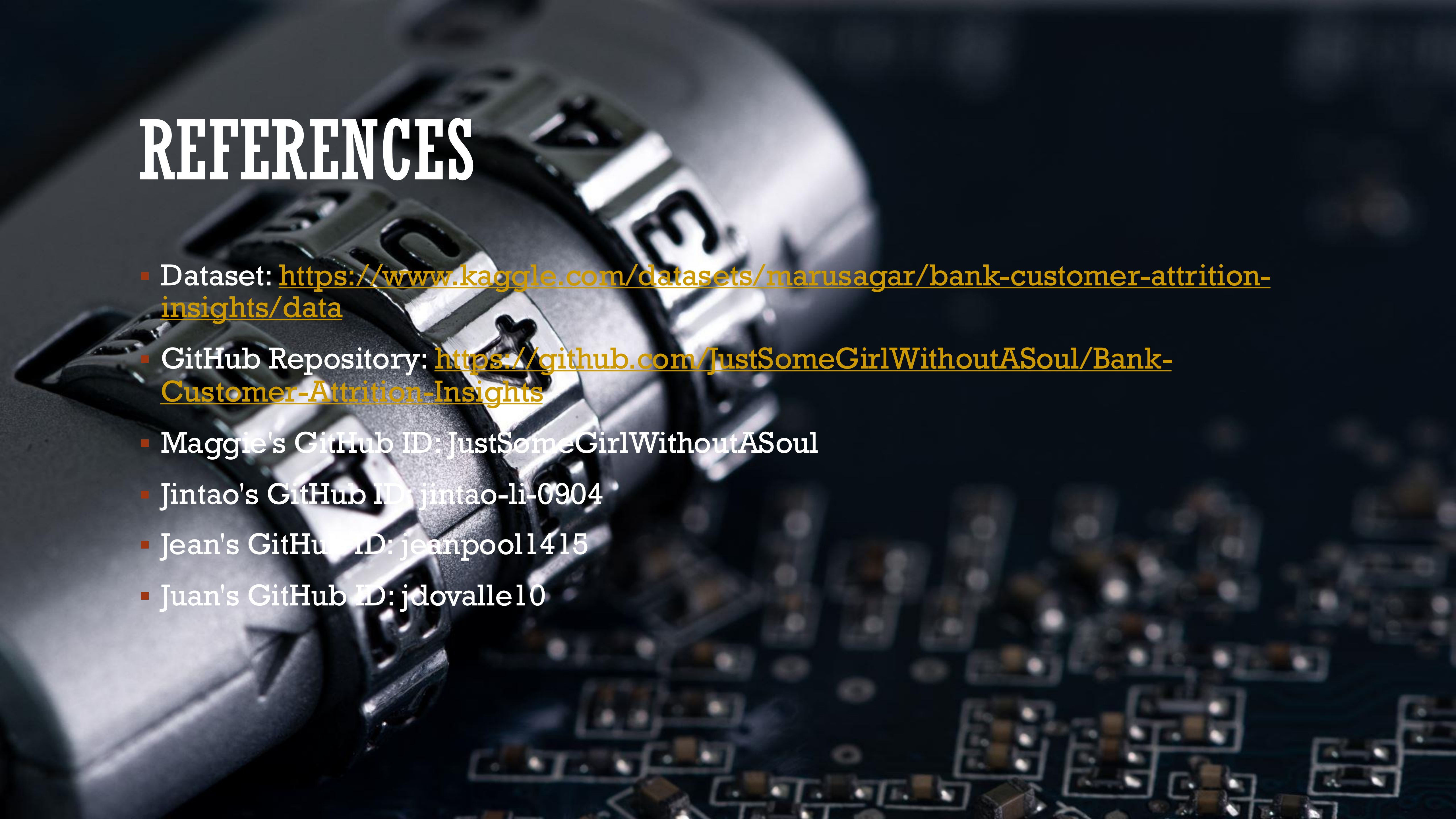
Modifying the classification threshold was the most effective way to improve recall, ensuring more churners are identified. However, care should be taken with this approach as false positive could end up representing more cost in the long-term.

LESSONS LEARNED & NEXT STEPS

- To maximize stacking effectiveness, models with **different learning approaches** (deep learning, tree-based models, probability-based models, etc) should be used.
- Review the complains info through text analytics, to determine the main reasons driving churning.
- Apply unsupervised learning techniques to identify customer segments, enabling the development of targeted strategies to manage churn more effectively in the bank.
- Repeat the causal inference analysis considering more possible interactions between features.

Thank you!

Any questions?



REFERENCES

- Dataset: https://www.kaggle.com/datasets/marusagar/bank-customer-attrition-insights/data
- GitHub Repository: https://github.com/JustSomeGirlWithoutASoul/Bank-Customer-Attrition-Insights
- Maggie's GitHub ID: JustSomeGirlWithoutASoul
- Jintao's GitHub ID: jintao-li-0904
- Jean's GitHub ID: jeanpool1415
- Juan's GitHub ID: jdovalle10