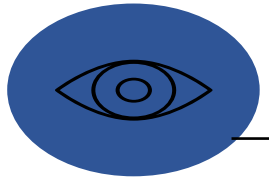
Predicting Bank Customer Churn

This project aims to predict bank customer churn using a dataset derived from the Bank Customer Churn Prediction dataset available on Kaggle. The dataset for this competition has been generated from a deep learning model trained on the original dataset, with feature distributions being similar but not identical to the original data.

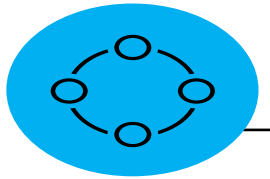
Raza Mehar

Created on: March 24, 2024
Last updated on: April 04, 2024
Version No: 1.0



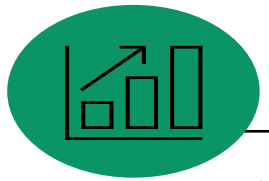
EXPLORE

- Reviewing metadata
- Inspecting heads and data types
- Understanding data dimensions



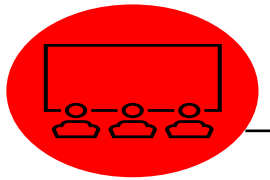
ANALYZE

- Descriptive statistics
- Univariate analysis
- Bivariate analysis
- Correlation analysis
- Predictive modeling
- Features Importance



SHARE

- Findings & Observations



ACT

- Recommendations

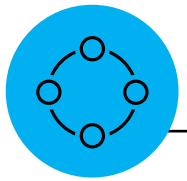
EXPLORE

Exploring Data Types and Dimensions and review descriptions

Feature	Data Type	Description
id	int	An identifier for each record to distinguish between different entries in the dataset.
CustomerId	int	Unique identifier for each customer, used to distinguish between individual customers.
Surname	str	The surname of the customer.
CreditScore	int	A numerical value representing the credit score of the customer, indicating creditworthiness.
Geography	str	The geographical location of the customer. (France, Spain, or Germany).
Gender	str	The customer's gender (Male or Female).
Age	float	The age of the customer.
Tenure	int	The number of years the customer has been with the bank.
Balance	float	The account balance of the customer.
NumOfProducts	int	The number of bank products (e.g., accounts, loans) held by the customer.
HasCrCard	float	Binary variable indicating whether the customer has a credit card (1) or not (0).
IsActiveMember	float	Binary variable indicating whether the customer is an active member of the bank (1) or not (0).
EstimatedSalary	float	The estimated salary of the customer.
Exited	int	Binary variable indicating whether the customer has churned (1) or not (0).

Dimension Rows: 65034; Columns: 14

Note: The dataset has not been processed yet. There are no missing or duplicate values.



ANALYZE

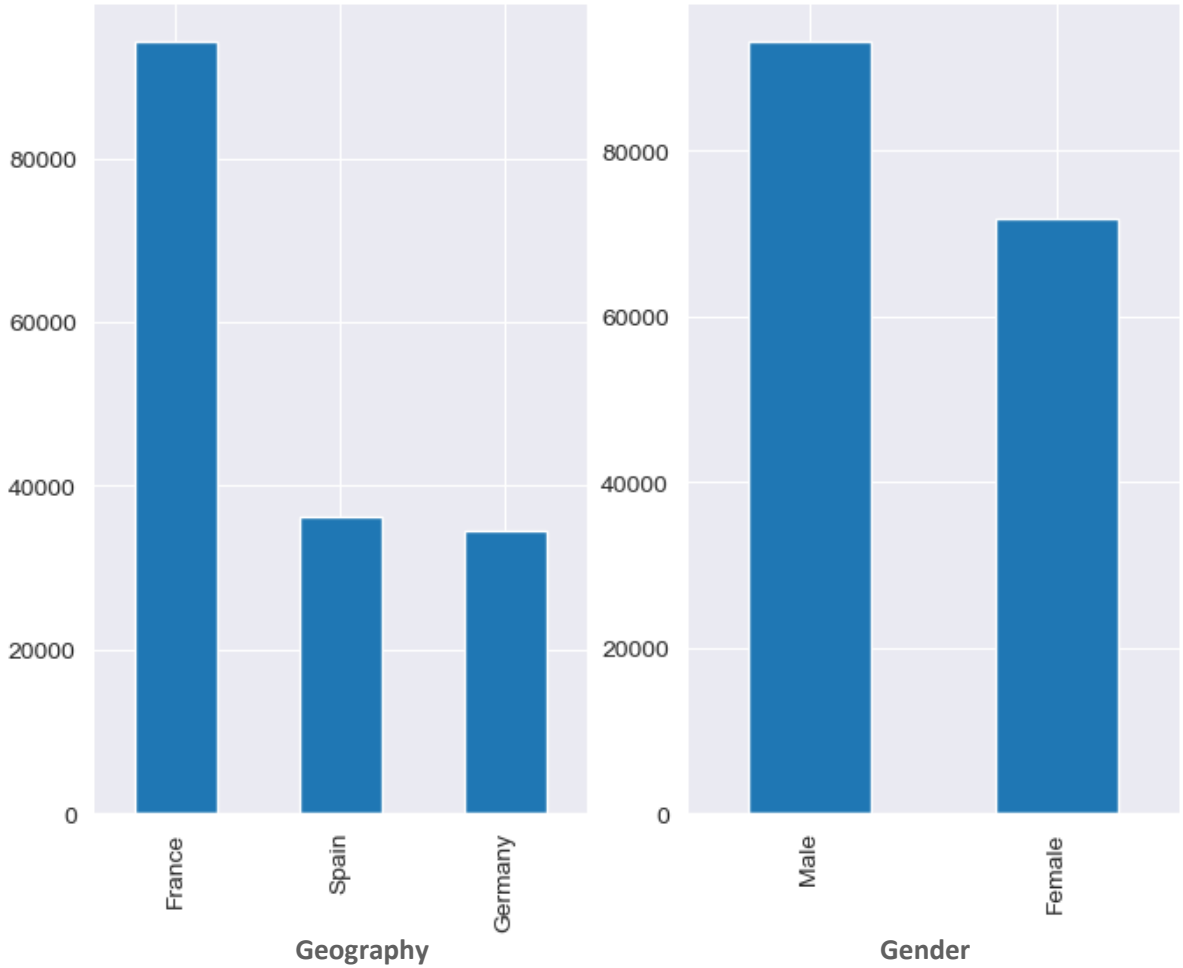
Exploring Descriptive Statistics

Numerical Feature	Min	Q1	Median	Mean	Q3	Max	SD
CreditScore	350	597	659	656	710	850	80
Age	18	32	37	38	42	92	9
Tenure	0	3	5	5	7	10	3
Balance	0	0	0	55478	119940	250898	62818
EstimatedSalary	12	74638	117948	112575	155152	199992	50293

Observation: As can be seen, there is a difference between the mean and standard deviation of the features. This will require data standardization.

ANALYZE

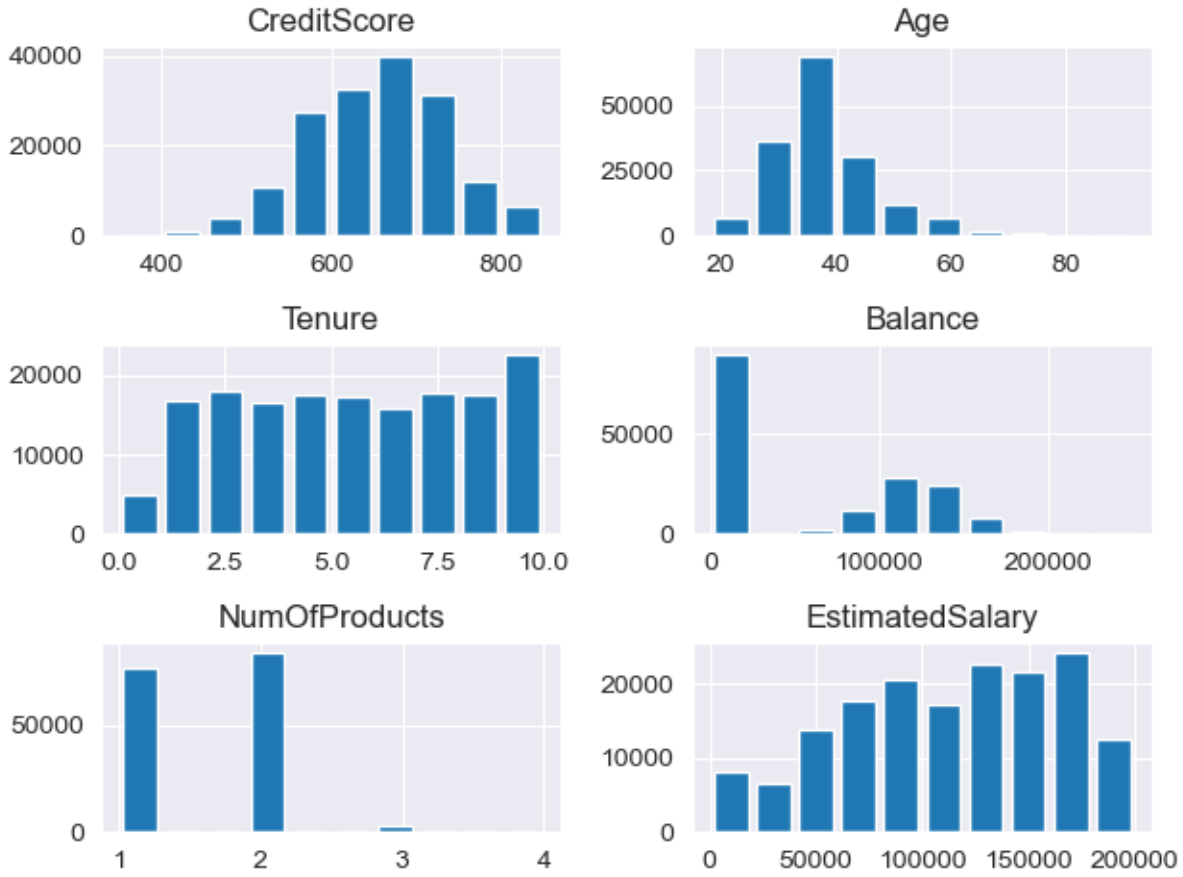
Univariate Analysis: Visualizing Categorical Features



Observation: The bar charts indicate that France has the highest number of customers, and most of the customers are male.

ANALYZE

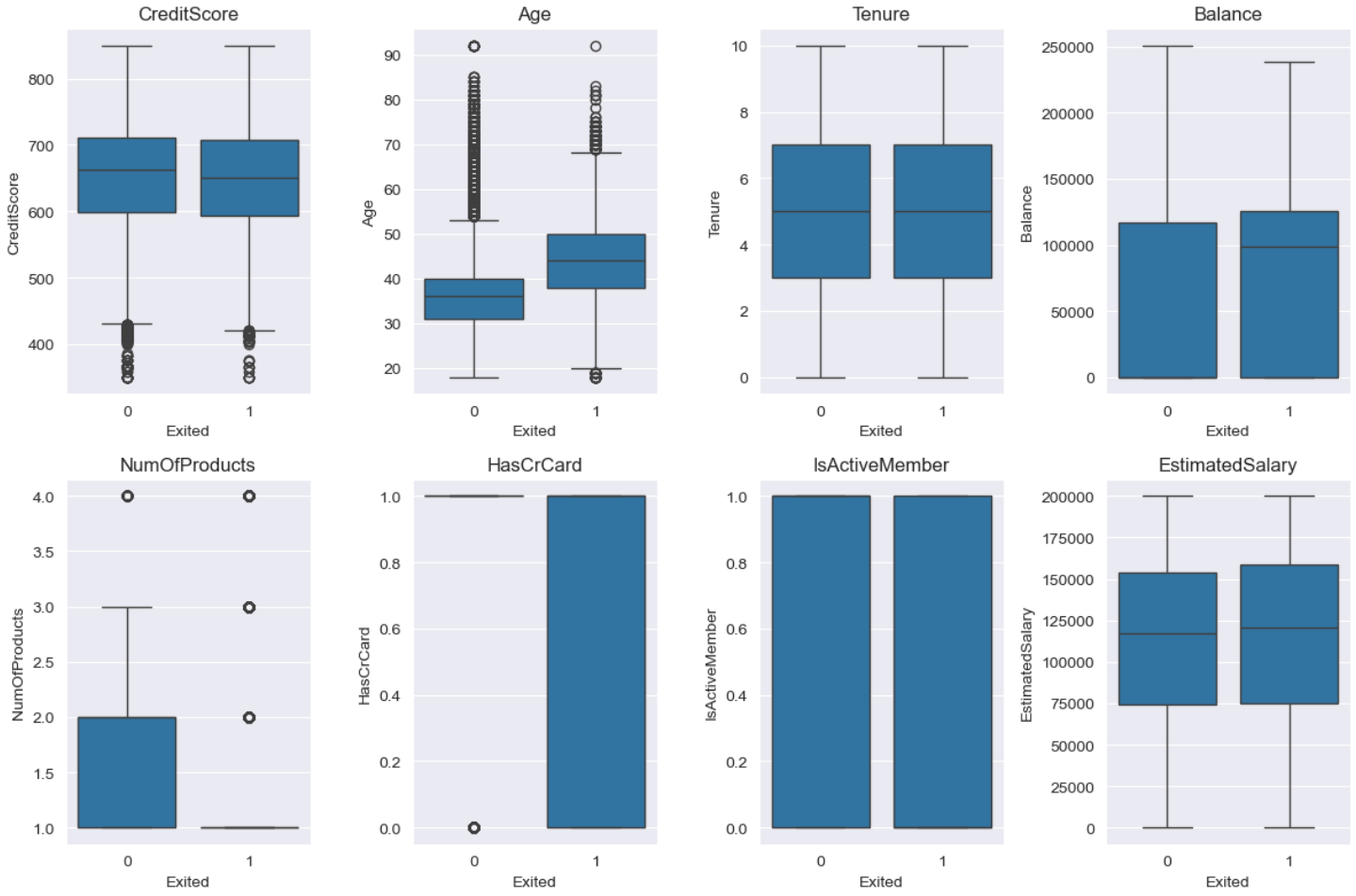
Univariate Analysis: Visualizing Numerical Features



Observation: CreditScore and Age follow a normal distribution, whereas other features do not. The majority of customers are in their mid-30s, with most maintaining a credit score in the range of 650 to 700. Additionally, the majority of customers have subscribed to 1 to 2 products only.

ANALYZE

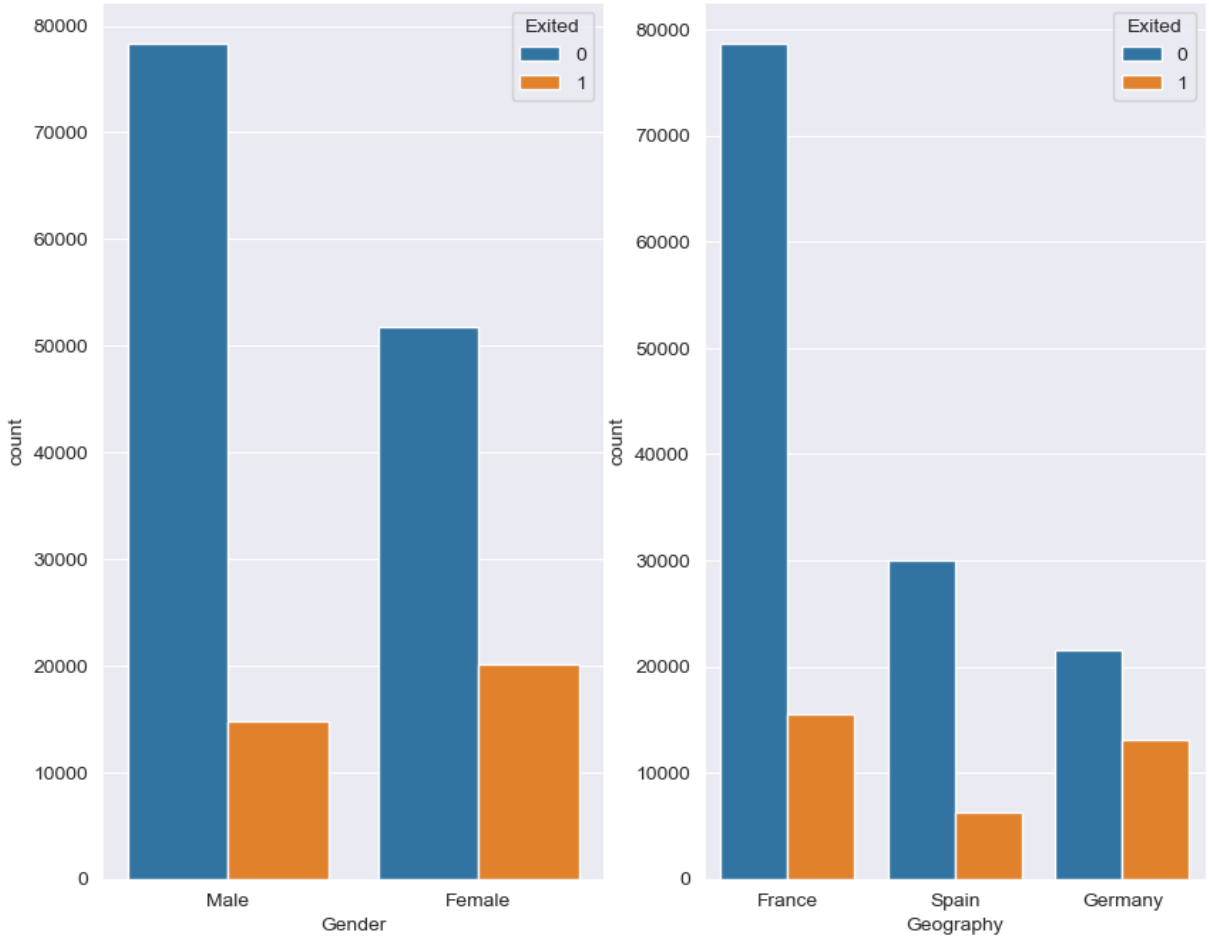
Bivariate Analysis: Exploring Relationships Between Numerical Predictors and Target Variable



Observation: It is evident that individuals in the age group of 40 to 50 have the highest churn rate. Additionally, customers with credit cards have also exhibited a high churn rate. Furthermore, those who have subscribed to 1- 2 products experience the highest churn.

ANALYZE

Bivariate Analysis: Exploring Relationships Between Categorical Predictors and Target Variable



Observation: Female customers have exhibited the highest churn rate, with France showing the highest churn rate followed by Germany.

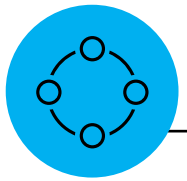
ANALYZE

Correlation Analysis: Examining Relationships Between Numerical Predictors and Target Variable

Feature	Data Type
Exited	1.00
Age	0.34
NumOfProducts	-0.21
IsActiveMember	-0.21
Balance	0.13
CreditScore	-0.03
HasCard	-0.02
Tenure	-0.02
EstimatedSalary	0.02

Using Point-Biserial Correlation as the target variable is binary.

Observations: It is apparent that there is no strong positive or negative correlation of the target variable with the numerical predictors.



ANALYZE

Feature Engineering and Oversampling

To enhance the model's performance, we introduced a new feature that calculates the ratio between the Balance and EstimatedSalary variables.

Additionally, it's important to note that there's a class imbalance issue within the dataset. Specifically, Class 1 comprises only 21% of the total data. To address this, we'll employ an oversampling technique like SMOTE to generate synthetic data.

ANALYZE

Predictive Modeling: Model Selection and Hyperparameter Tuning with Cross Validation

The model selection involved exploring various configurations for the K-Neighbors, Decision Tree, Random Forest, AdaBoost models, and XGBooost classifiers using 80% of the data for training while reserving 20% as a hold-out set—data that the model has not seen.

These model parameters were subjected to k-fold cross-validation (k = 10) to identify the best-performing configurations. The choice of k = 10 aligns with established best practices, supported by empirical evidence. Ron Kohavi's experiments on diverse real-world datasets indicate that a 10-fold cross-validation strategy strikes an optimal balance between bias and variance in model assessment.

Model	ROC AUC Score
K-Neighbors	0.9307874126090576
Decision Tree	0.8577441168823294
Random Forest	0.9636679016847967
Extra Tree	0.9705613648817577
Adaptive Boosting	0.9008086096227306
XG Boosting	0.9558999132345075

Random Forest, Extra Tree, and XG Boosting classifiers were selected for hyperparameter tuning based on their higher ROC AUC scores (more than 95%). After an exhaustive grid search the best hyperparameters for each model were identified, providing a foundation for subsequent model evaluation.

Model	n_estimators	max_depth	learning_rate	Optimal Parameters
Random Forest	50, 100	None, 5	Not Applicable	max_depth: None; n_estimators: 100
Extra Tree	50, 100	None, 5	Not Applicable	max_depth: None; n_estimators: 100
XG Boosting	50, 100	3, 5	0.01, 0.1	max_depth: 5; n_estimators: 100, learning_rate: 0.1

Note: A new feature: Balance to Salary ratio has been engineered. Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". International Joint Conference on Artificial Intelligence (IJCAI), 14 (2): 1137-43, 1995. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

ANALYZE

Predictive Modeling: Model Selection based on Optimal Hyperparameters

Following the hyperparameter tuning, the models underwent rigorous evaluation using a 10-fold cross-validation approach. The ROC AUC score was employed as a key metric to assess the performance of each model. ROC AUC score indicates how well the model differentiates between the two classes

Model	ROC AUC Error
Random Forest	0.963289027493175
Extra Tree	0.9691352976666237
XG Boosting	0.9464063993977739

Observation: The Extra Tree classifier exhibits the highest ROC AUC score compared to the other two models, suggesting superior performance in our evaluation.

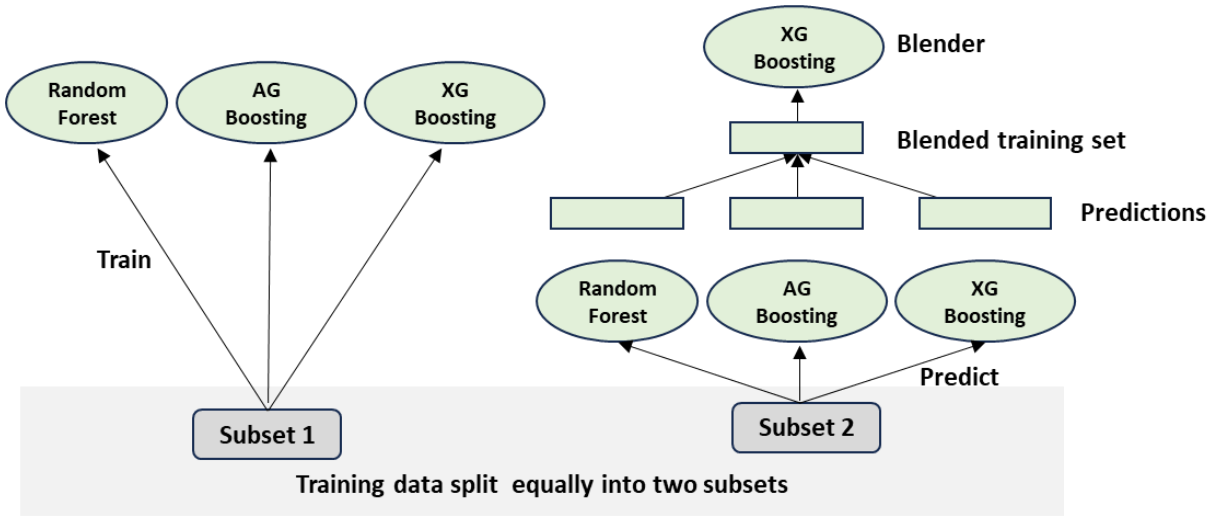
ANALYZE

Predictive Modeling: Performance Evaluation of the Model

The optimal model (XGB Classifier) was then trained on 100% of the training data. Subsequently, the model underwent evaluation on the reserved holdout set to assess its performance.

Performance Metrices	Training Data	Holdout Data
ROC AUC Score	0.9999999796487736	0.9661956764016835
Precision	1.0	0.8940567513219632

Additionally, model **staking and blending** were employed, and the performance was evaluated.

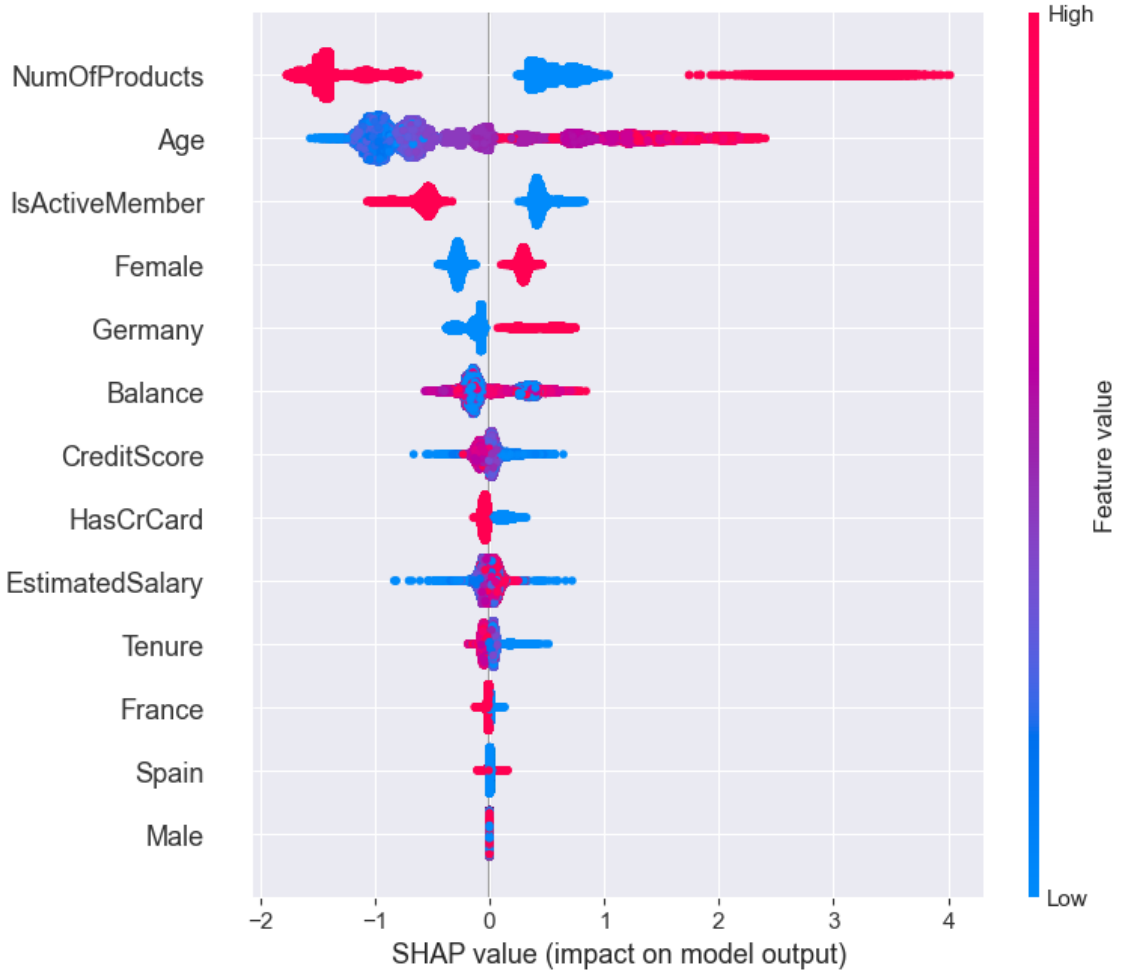


Performance Metrices	Holdout Data
ROC AUC Score	0.906778797186111
Precision	0.8865253402415442

Observation: The optimal model seems to perform well on both the training and the hold-out data sets. However, stacking and blending have reduced the performance significantly, hence this method will not be utilized.

ANALYZE

Feature Importance: Model interpretability using SHAP (Shapley Additive Explanations) Summary Plot



Observation: The higher values of NumOfProducts and Age tend to have higher predicted churn probability

SHARE

Findings & Observations

Customer Demographics

- France has the largest customer base, with a significant portion being male.
- Majority of customers are in their mid-30s, with credit scores typically falling in the range of 650 to 700.
- Most customers have subscribed to 1 to 2 products only.

Churn Patterns

- Customers with 1 -2 products experience the highest churn.
- The age group between 40 to 50 exhibits the highest churn rate.
- Customers holding credit cards show a notable churn tendency.
- Female customers, particularly in France and Germany, have the highest churn rate.

Note: Findings and observations are based on univariate, multivariate analysis, and correlation analysis.

ACT

Recommendations

Primary

Diversification of Product Portfolio: Explore opportunities to diversify the product portfolio to increase customer engagement and loyalty. Encourage customers to subscribe to additional products or services by offering bundled packages, personalized recommendations, or exclusive benefits for multiple product holders

Targeted Retention Strategies: Implement targeted retention strategies, especially focusing on customers aged between 40 to 50, as they exhibit the highest churn rate. Offer personalized incentives, such as loyalty rewards or discounts, to encourage their continued engagement with the bank.

Secondary

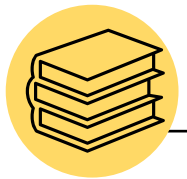
Enhance Credit Card Services: Enhance credit card experiences with better rewards, security, and tailored promotions to address high churn rates among cardholders.

Gender-Specific Retention Efforts: Develop tailored retention efforts for female customers to reduce their higher churn rates, including personalized financial advice and exclusive offers.

Geographically Tailored Strategies: Tailor retention strategies for regions like France and Germany with higher churn rates, using localized marketing and product offerings to resonate better with customers.

Continuous Monitoring and Analysis: Regularly monitor churn patterns and analyze trends to identify risks early. Utilize the model developed in this project to proactively identify customers at risk of churn and implement preemptive retention measures.

Note: The strategies will contribute not only to the retention of existing customers but also to the acquisition of new customers.



BIBLIOGRAPHY & REFERENCES

- Alan Agresti, Maria Kateri. “Foundations of Statistics for Data Scientists with R and Python”. CRC Press. 2022.
- Aurélien Géron. “Hands-On Machine Learning with Scikit-Learn & TensorFlow”. O’Reille Media. 2019.
- Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili. “Machine Learning with PyTorch and Scikit-Learn”. Packt Publishing Ltd. 2022.

LIBRARIES & MODULES

- pandas
- numpy
- seaborn
- matplotlib
- scipy
- sklearn