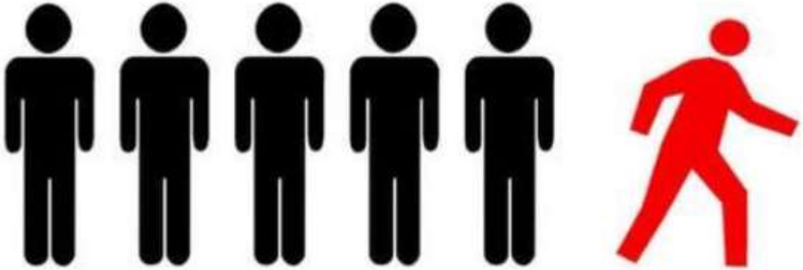
Bank Churn Predictions

Shivendu Kishore , Ankit Tiwari, Ayush Singh,
January 23, 2024

Churning in Bank Sector – Overview

CUSTOMER CHURN



Customer Churn in Banking

- Churn is defined as exiting of customer from the service and moving from one company to another.

- The reasons can for example be:
  - Availability of latest technology
  - Customer-friendly bank staff
  - Low interest rates
  - Location

- Churn rate usually lies in the range from 10% up to 30%.

Why is this important for the bank?

- The cost of attracting new customers can be five to six times more than holding on to an existing customers

- Long term customers become less costly to serve, they generate higher profits, and they may also provide new referrals

- Losing a customer usually leads to loss in profit for the bank.

How to define a churner in a bank

- Customer who closes his account or has decreasing number of transactions over a specific period in time
- Focus on customers who have three or less products with the bank
- Active customer is a customer with two or more active products
- A churner can be defined as a customer who has not been active over a specific time

Churning Prediction Model – Concept

Customer Churn Model

- Prediction models are used to identify customers who are likely to churn

- The model uses historical data on former churners and tries to find some similarity with existing customers

- If some similarity is found those customers are classified as potential churners.

What needs to be considered when setting up a churn model?

- How to define a churner in the bank
- Churn prediction variables to use in the model
- Methods/techniques used to build the model

Data Source

- Kaggle Churn Modelling Dataset
    - 165034 records
    - Demographic information
    - Banking information

Churn prediction variables

- Customer demographics variables
    - Age
    - Surname
    - Gender
    - Geographical data

Churn prediction variables

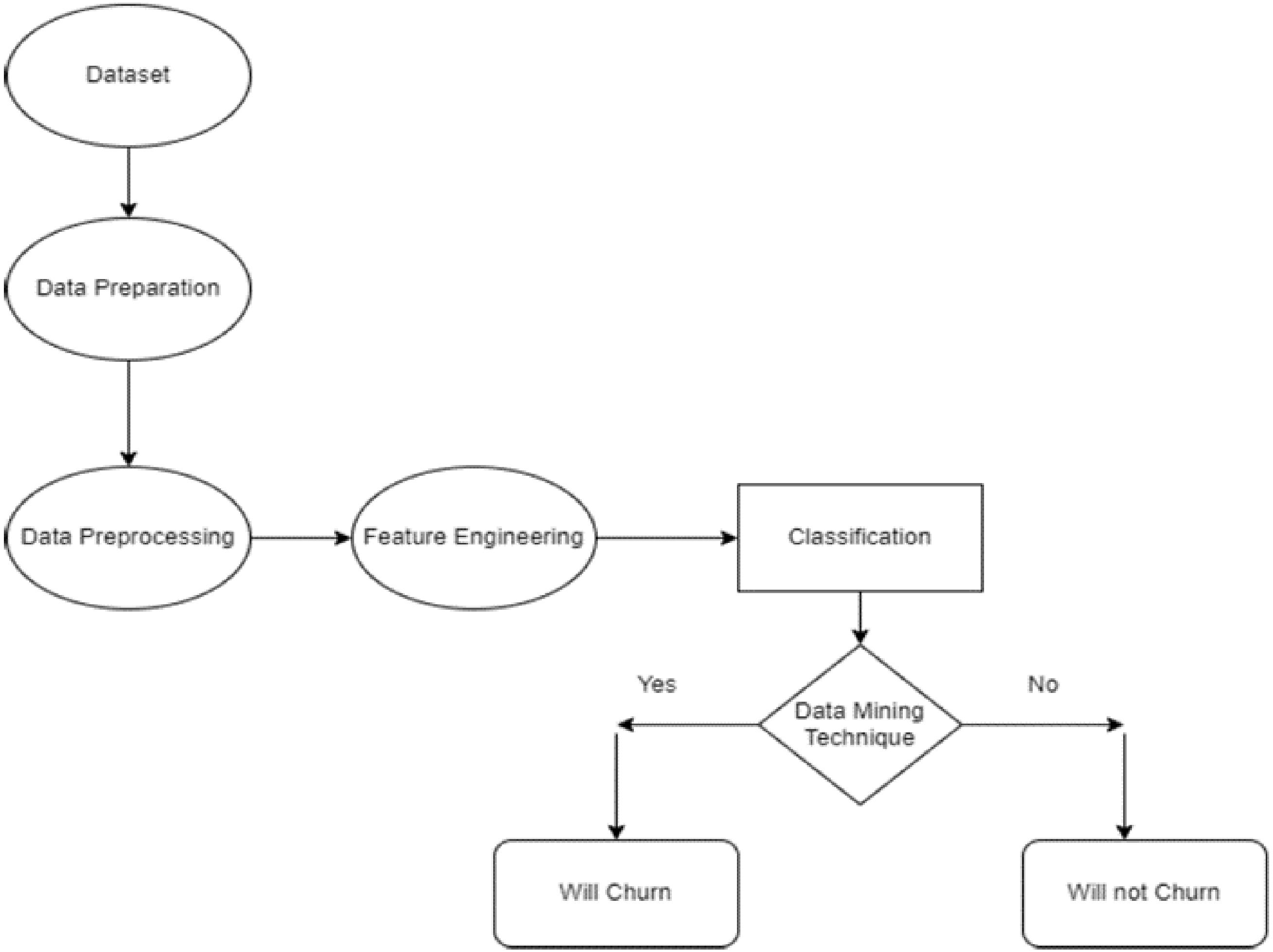
- ❖ Bank-Related Variables
  - Balance
  - Credit Score
  - Tenure with bank
  - Number Of Products of bank used by customer
  - Has Credit Card
  - Is Active Member
  - Estimated Salary

Churn Prediction Modelling

- Predict if a bank’s customer will stay or leave the bank by taking insights from historical dataset

  - Exited – 0 – Not leave
  - Exited – 1 – Leave

Process Theory – Working


Dataset
Data Preparation
Data Preprocessing
Feature Engineering
Classification
Data Mining Technique
Yes
No
Will Churn
Will not Churn

Steps for Prediction Model

- Loading Data
- Data Exploration
- Data Visualization
- Feature Preprocessing & Engineering
- Model Selection
- Validation
- Submission

ETL

ETL

Extract, Transform, Load

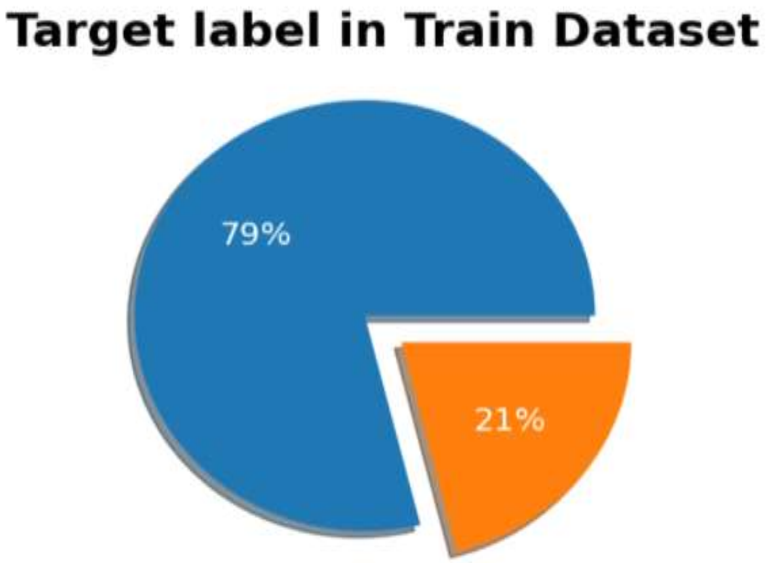
- Kaggle Dataset → Cleaning
- Removing features not needed for ML:
  - RowNumber
  - CustomerId

- Save it as the analytical base table (train_mod, test_mod)
- (train_mod, test_mod) → input to all modelling notebooks

Data Visualization-Insights from Data

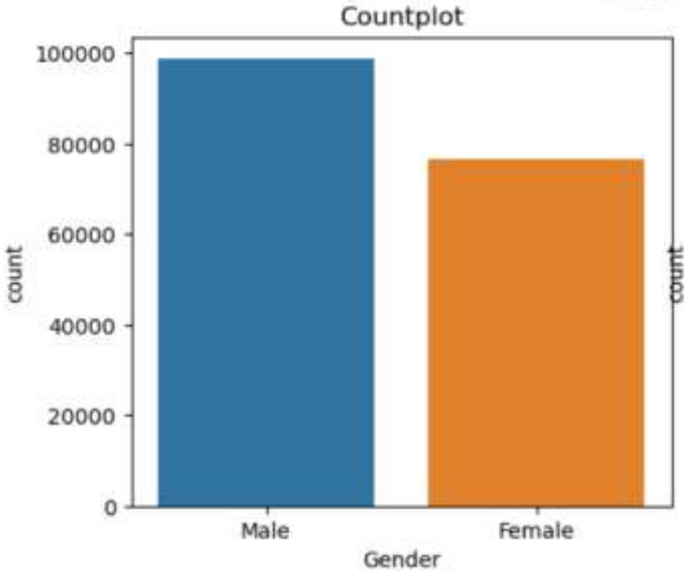
Target Feature Distribution

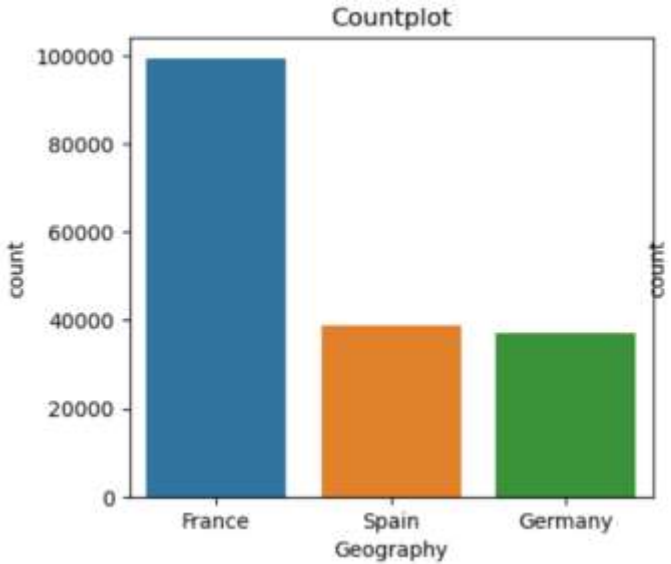
```
0    138076
1     36958
Name: Exited, dtype: int64
```



- ❖ Imbalanced dataset:
  - ➢ Stays - 79%
  - ➢ Exits - 21%
- ❖ Handling imbalanced classes:
  - ➢ StratifiedKFold

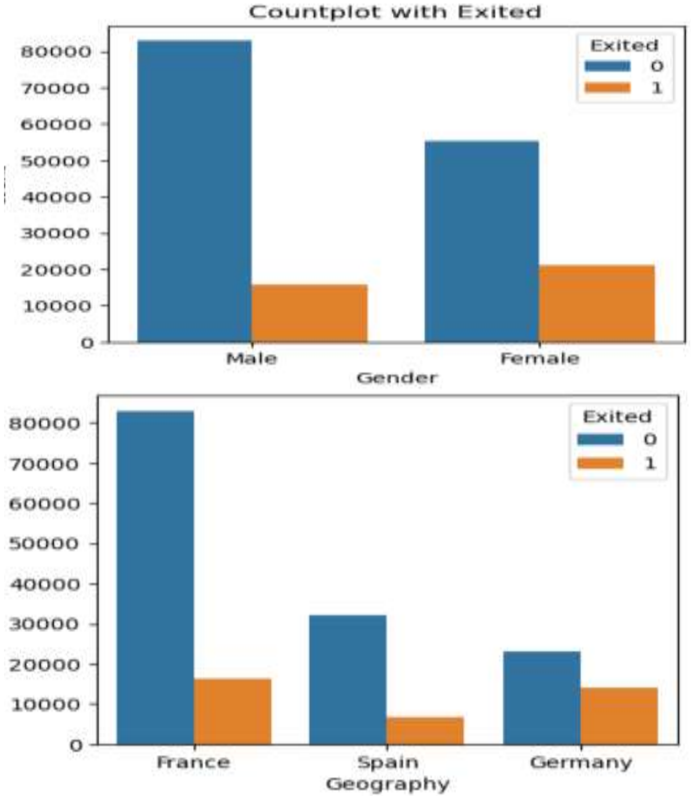
Distributions of Categorical Features





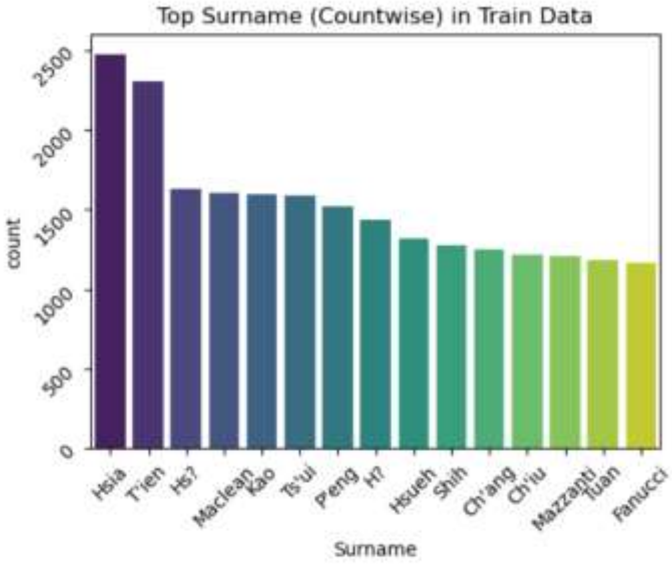
- ❖ More males than females
- ❖ France 50%; Spain, Germany 25% each

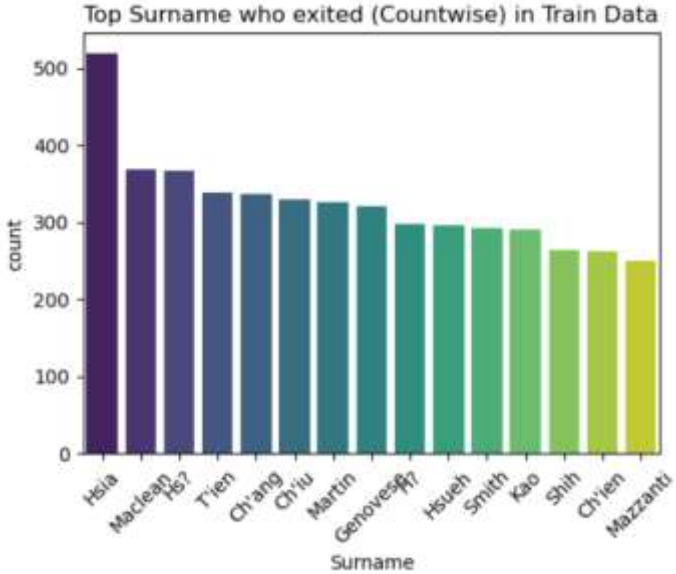
Churn Segmentation by Gender/Geography



- Leaving bank:
  - Females 28%
  - Males 16%
  - German customers 38%
- Germany: least num of customers → most churn

Distributions of Categorical Features

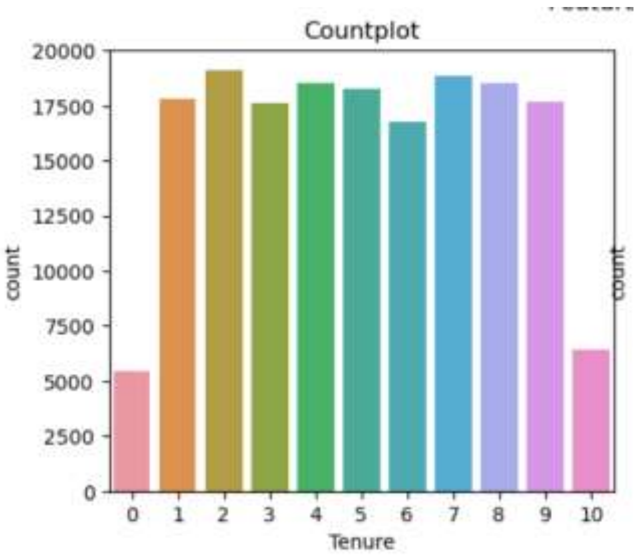


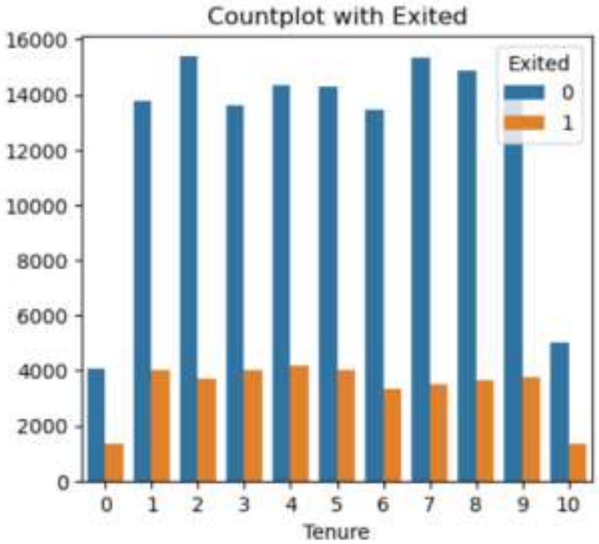


❖ Top Surnames in the dataset

❖ Top Surnames Exited in the dataset

Distributions of Categorical Features

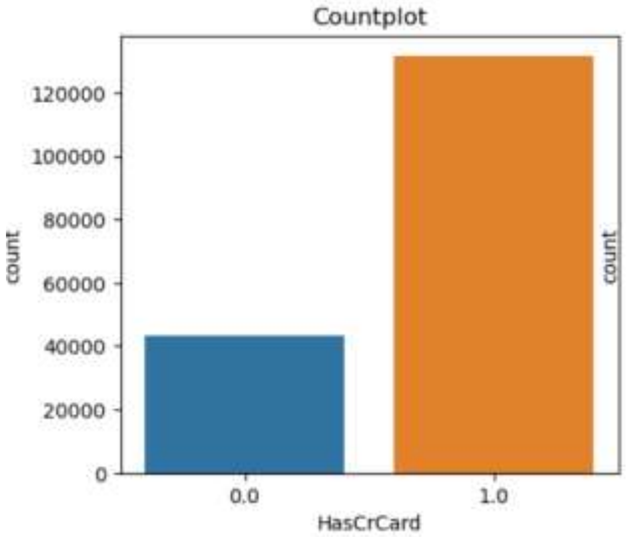


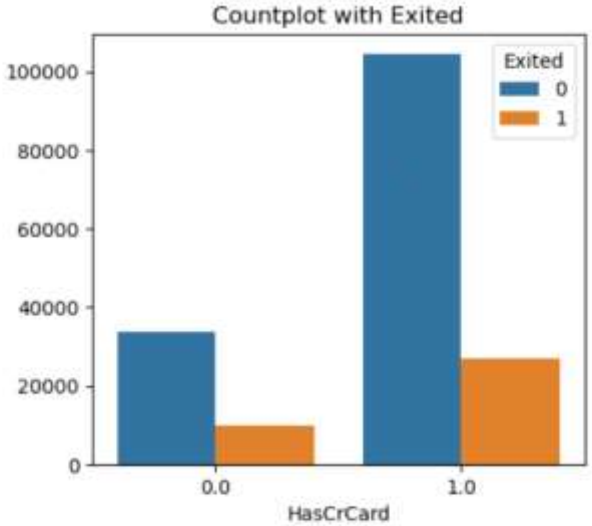


- ❖ Customers with bank in years(Tenure) in the dataset

- ❖ Customers with bank in years (Tenure) Exited in the dataset

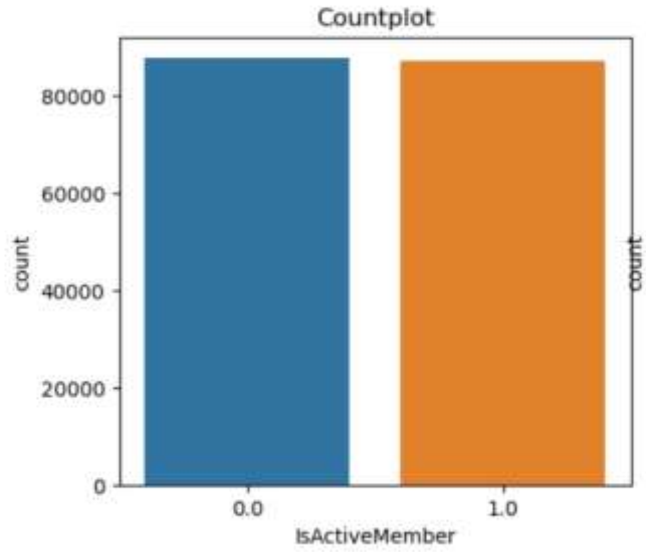
Distributions of Categorical Features

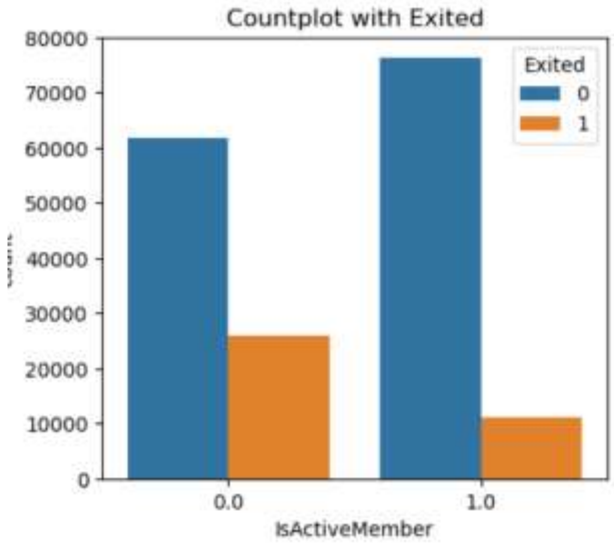




- Customers having credit card
- Customers having credit card exited

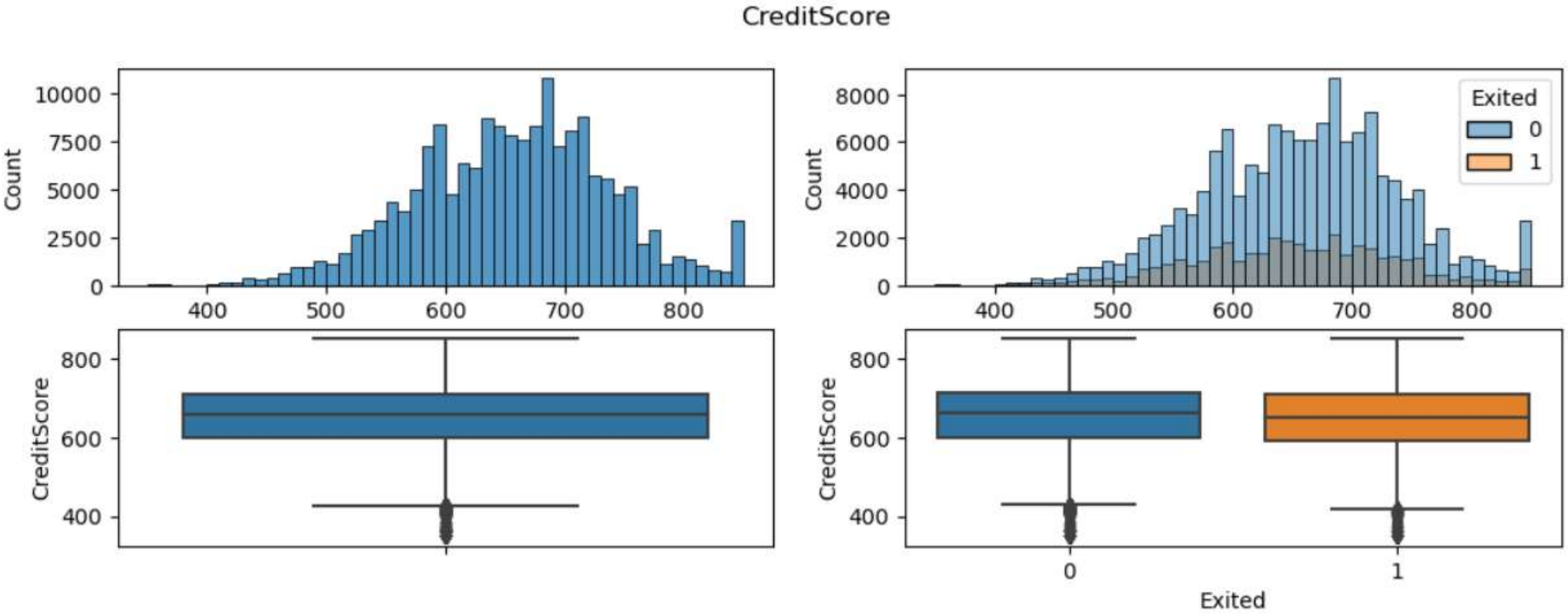
Distributions of Categorical Features





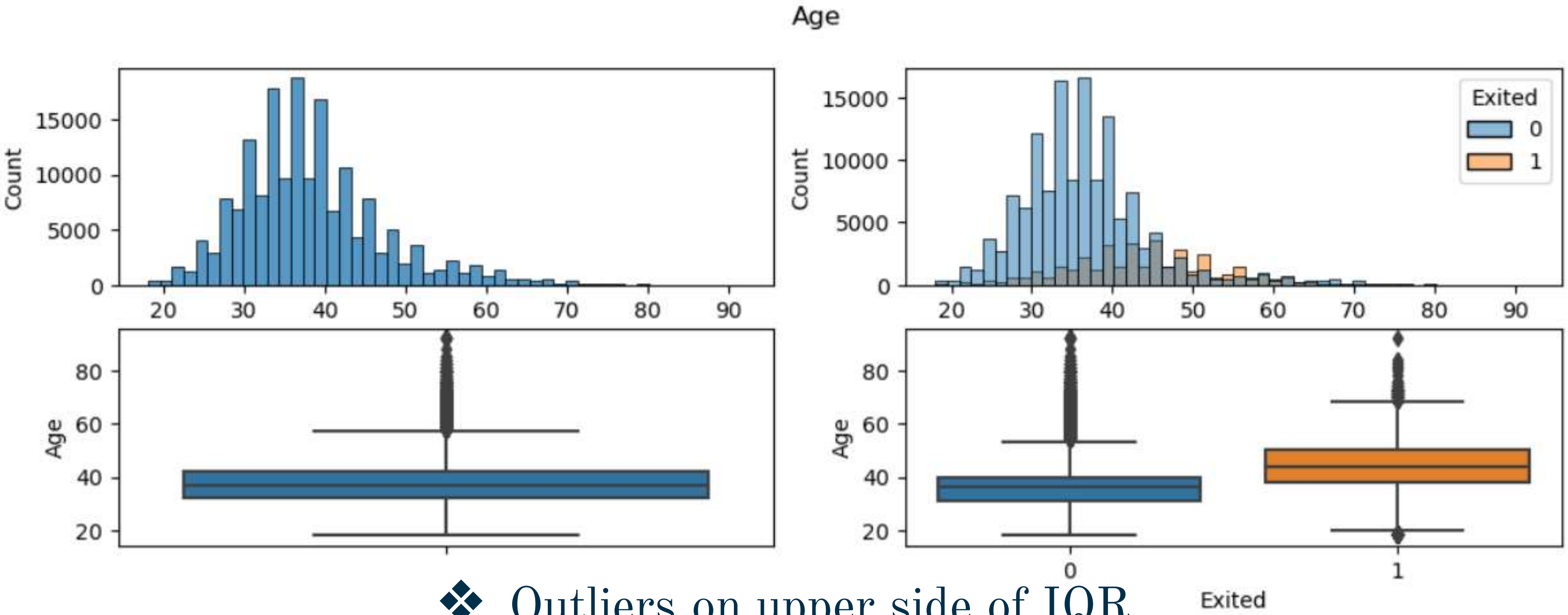
- Active Member of bank
- Active Member of bank Exited

Numerical Features – Histplot and Boxplot



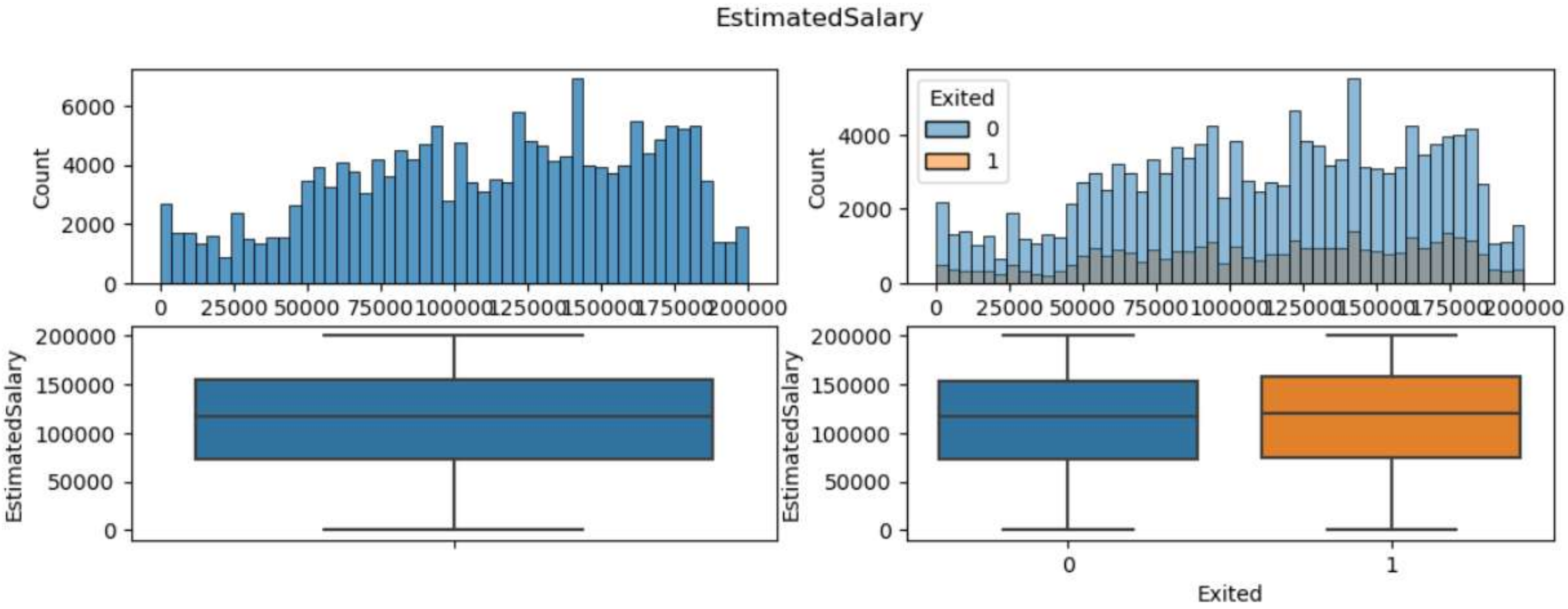
- Outliers on lower side of IQR

Numerical Features – Histplot and Boxplot

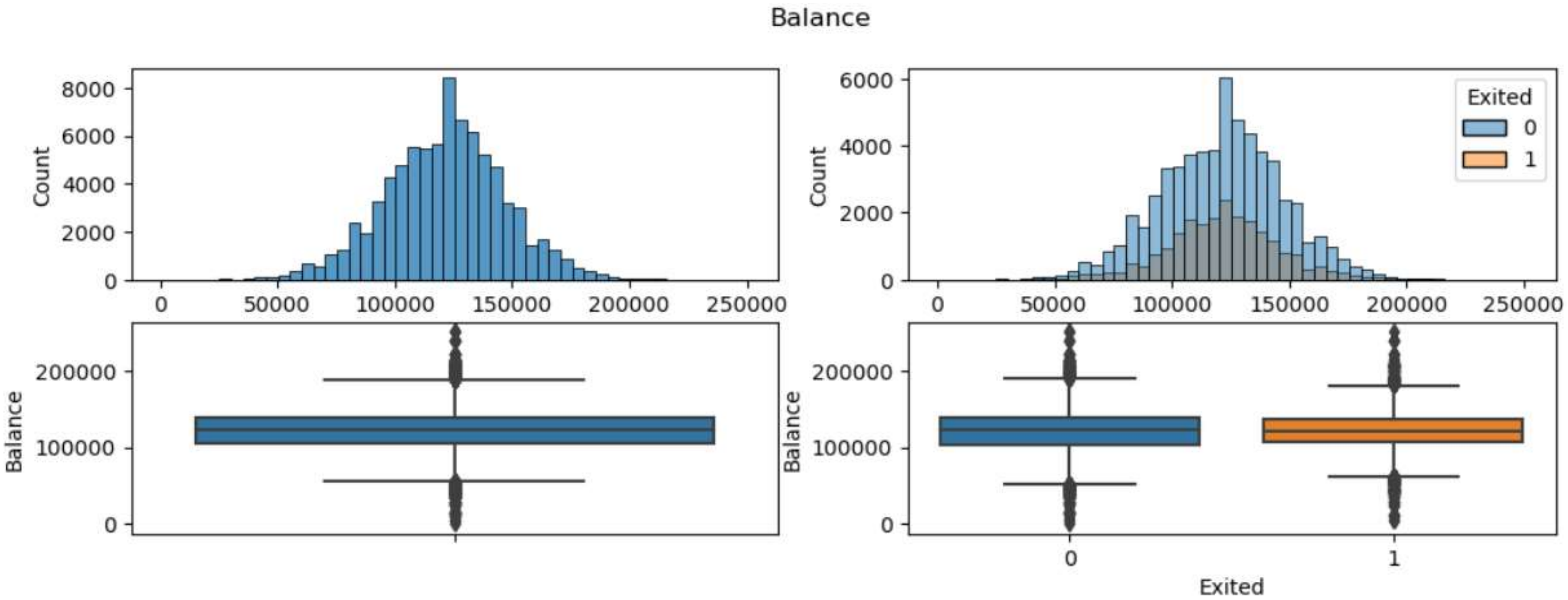


- ❖ Outliers on upper side of IQR

Numerical Features – Histplot and Boxplot



Numerical Features – Histplot and Boxplot



❖ In Feature – Balance, there are lot of customers who were having bank balance – 0 so avoided visualization from this and getting bigger picture of balance column. Also there are much outliers present in this

Selecting Model for Predicting Churn in dataset

Methods / techniques used to build a model

- Classification techniques
  - Logistic Regression Model
  - Tree Model – Random Forest Model
  - Boosting Method – Xtreme Gradient Boosting Model

Methods / techniques used to build a model

- <u>Logistic Regression Model</u>
  - Used for binary classification problems (predicting two outcomes)
  - Advantages:
    - Simple and easy to understand
    - Fast training and prediction times for smaller datasets
  - Disadvantages:
    - Assumes a linear relationship between features
    - Limited flexibility for capturing complex patterns in data

Methods / techniques used to build a model

❖ <u>Random Forest Model</u>

- Versatile for both classification and regression tasks
- Advantages:
  - Handles non-linear relationships well
  - Robust to overfitting, as it combines multiple decision trees
  - Can handle large datasets with many features
- Disadvantages:
  - Less interpretable compared to a single decision tree

Methods / techniques used to build a model

- Extreme Gradient Boosting
  - Extremely powerful for classification and regression tasks.
  - Advantages:
    - High predictive accuracy due to the boosting technique
    - Handles complex relationships and interactions in the data
    - Regularization techniques to prevent overfitting
    - Can handle missing data effectively
  - Disadvantages:
    - Can be computationally intensive and require more resources.

Methods/techniques used to build a model – what to use for this dataset

- Conclusion:

  XGBoost is considered the best choice among these models for several reasons:

  - It combines the strengths of both random forests and gradient boosting.
  - It usually outperforms other models in terms of predictive accuracy.
  - It handles complex relationships and large datasets well.
  - Regularization techniques help prevent overfitting.

Model Training Steps

- Model selection and training on train data
- Hyperparameters tuning:
  - parameters grid with Optuna
  - GridSearchCV
  - StratifiedKfold
- Check ROC-AUC score
- Re-train model for best ROC-AUC score
- Save best model
- Feature Importance
- Predict Exited value in Test data

Conclusions

- Huge difference with StratifiedKfold split
- Optuna resulted speed and better hyperparameter tuning
- Ways for improvement
  - More data points for target variable imbalance situation
  - Feature engineering
  - Additional ML algorithms and imbalance handling techniques

Questions?

For detailed understanding now will move to Python Notebook