BANK CHURN PREDICTION

Table of contents

01 PROBLEM OVERVIEW

- Churn is the process of customers leaving a company or service.
- In the banking industry, it's the closure of accounts or switching to another bank.
- Churn negatively impacts a bank's revenue and profitability.
- Banks measure churn using various methods such as closed accounts or percentage of customers leaving.
- Research can help banks reduce churn by identifying factors and implementing retention programs.

02 OBJECTIVE

- Understand the key drivers of customer churn and their impact on business performance.
- Develop a predictive model to identify at-risk customers and proactively implement retention strategies.
- Enhance existing retention strategies, such as targeted incentives and personalized communication, to improve customer loyalty and reduce churn.
- Continuously monitor and evaluate the effectiveness of retention efforts to ensure maximum return on investment.

03 DATASET

Current study provides analysis of the dataset provided by bank (Kaggle website)

which contains next fields:

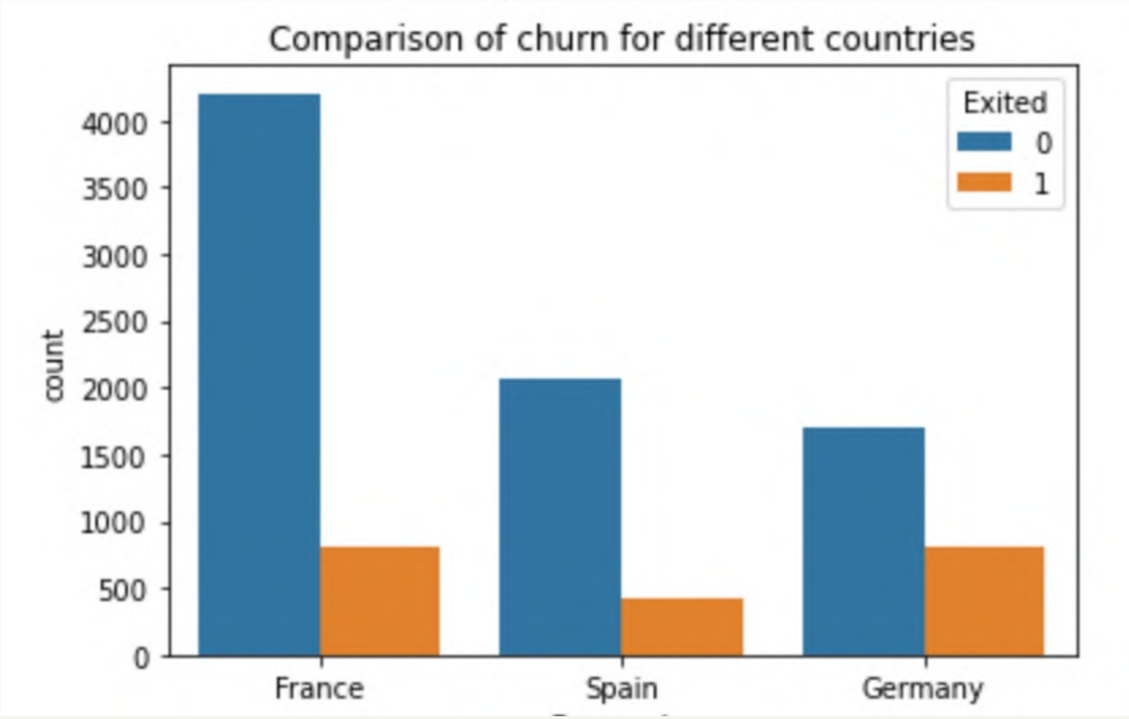
- RowNumber - Number of the row in the table
- CustomerId - Unique customer ID
- Surname - Surname of the customer
- CreditScore - Credit Score
- Geography - Country of the bank's branch for customer
- Gender - Gender of customer
- Age - Age
- Tenure - How many years customer is with this bank
- Balance - Current balance on the deposit account
- NumOfProducts - Number of banking products this customer uses
- HasCrCard - Does this customer have a credit card in this bank?
- IsActiveMember - Was this customer active during this report period
- EstimatedSalary - Salary range estimated by bank
- Exited - Target variable that shows whether a customer is churned from banking services.

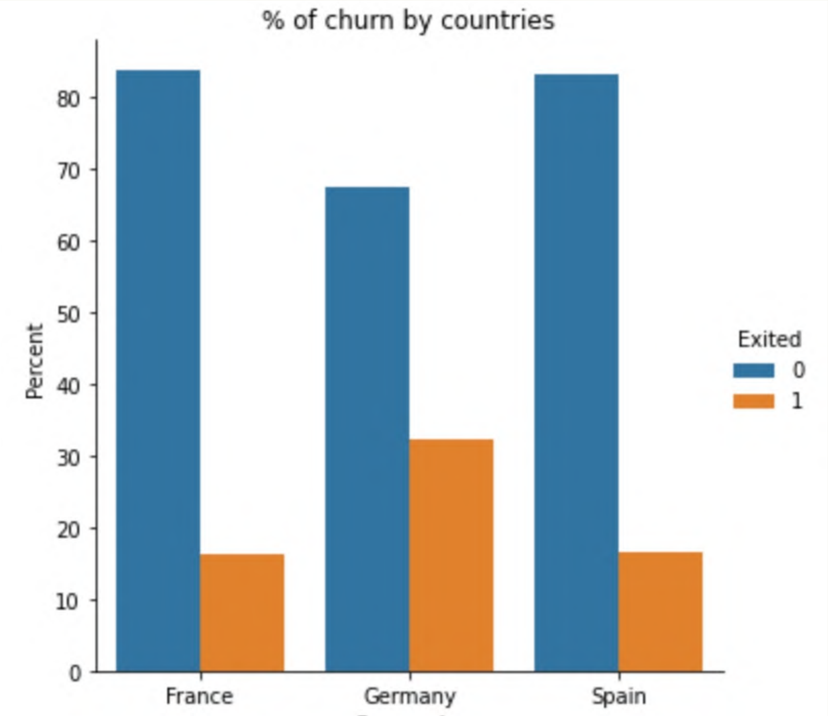
04 DATA OVERVIEW

- ☑ Shape of the data is **10000 rows and 14 columns.**
- ☑ Number of clients that have **exited** and **stayed** with the bank services are : 2037 (**20.4%**) and 7963 (**79.6%**) correspondingly. **Dataset is unbalanced** with a minor class of customers who exited the bank services.
- ☑ Numerical features are: Age, Tenure, CreditScore, EstimatedSalary, Balance, NumOfProducts
- ☑ Categorical features are: Geography (Germany, Spain, France), Gender (Male, Female), HasCrCard (1,0), IsActiveMember (1,0)
- ☑ Target variable is a field **Exited** that is categorical and can take values 1, 0; where **1 - is the customer who exited** the bank program.

05 CHURN BY GEOGRAPHY

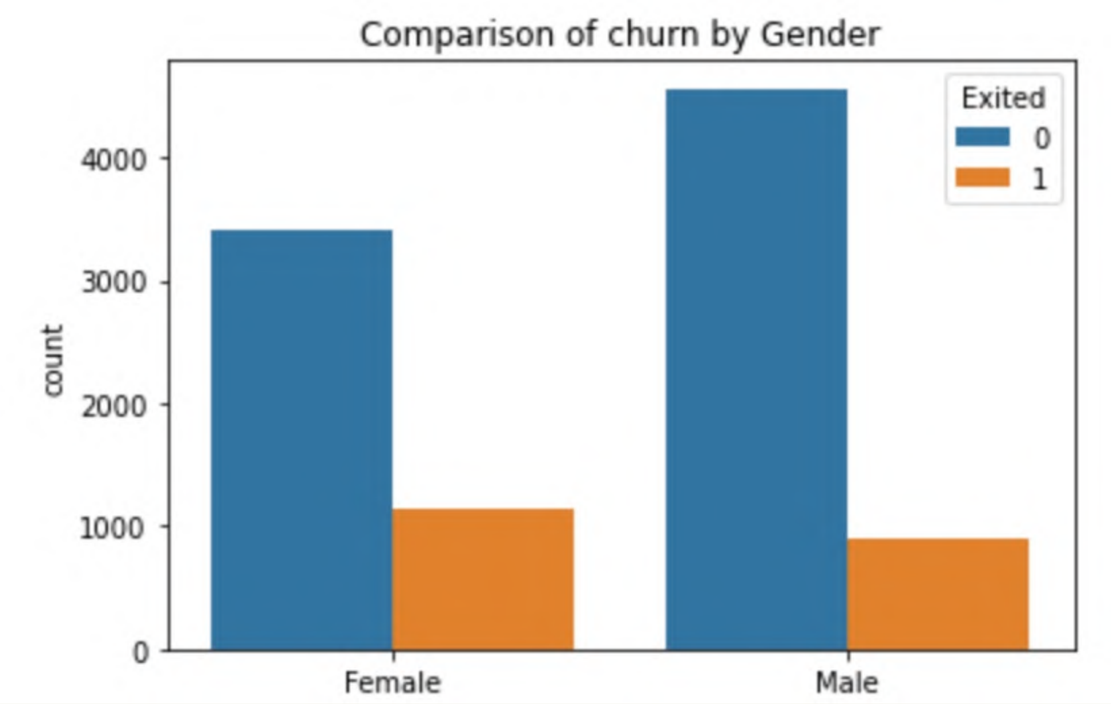
Germany has less users than other countries but there is highest percentage of churn - 32.44%

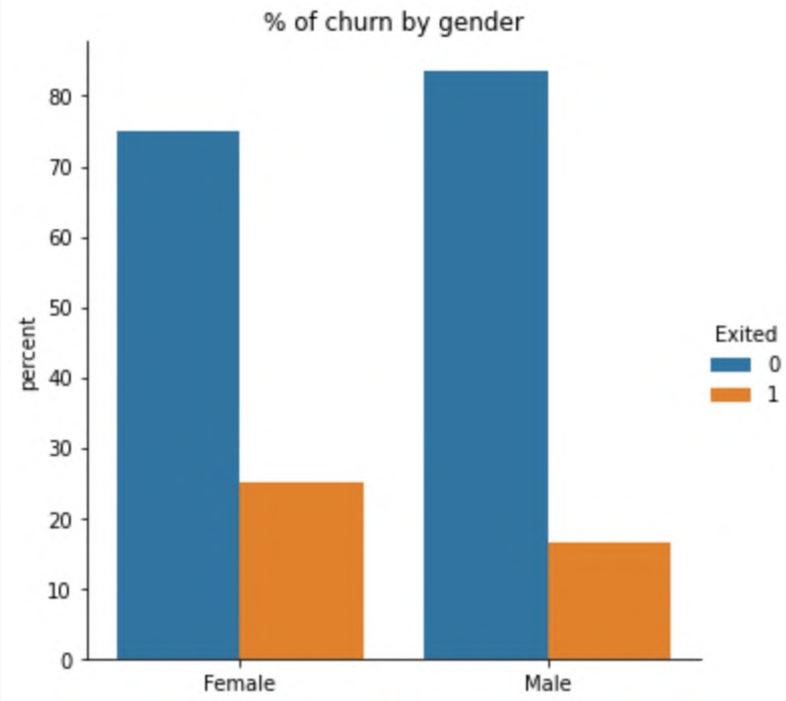




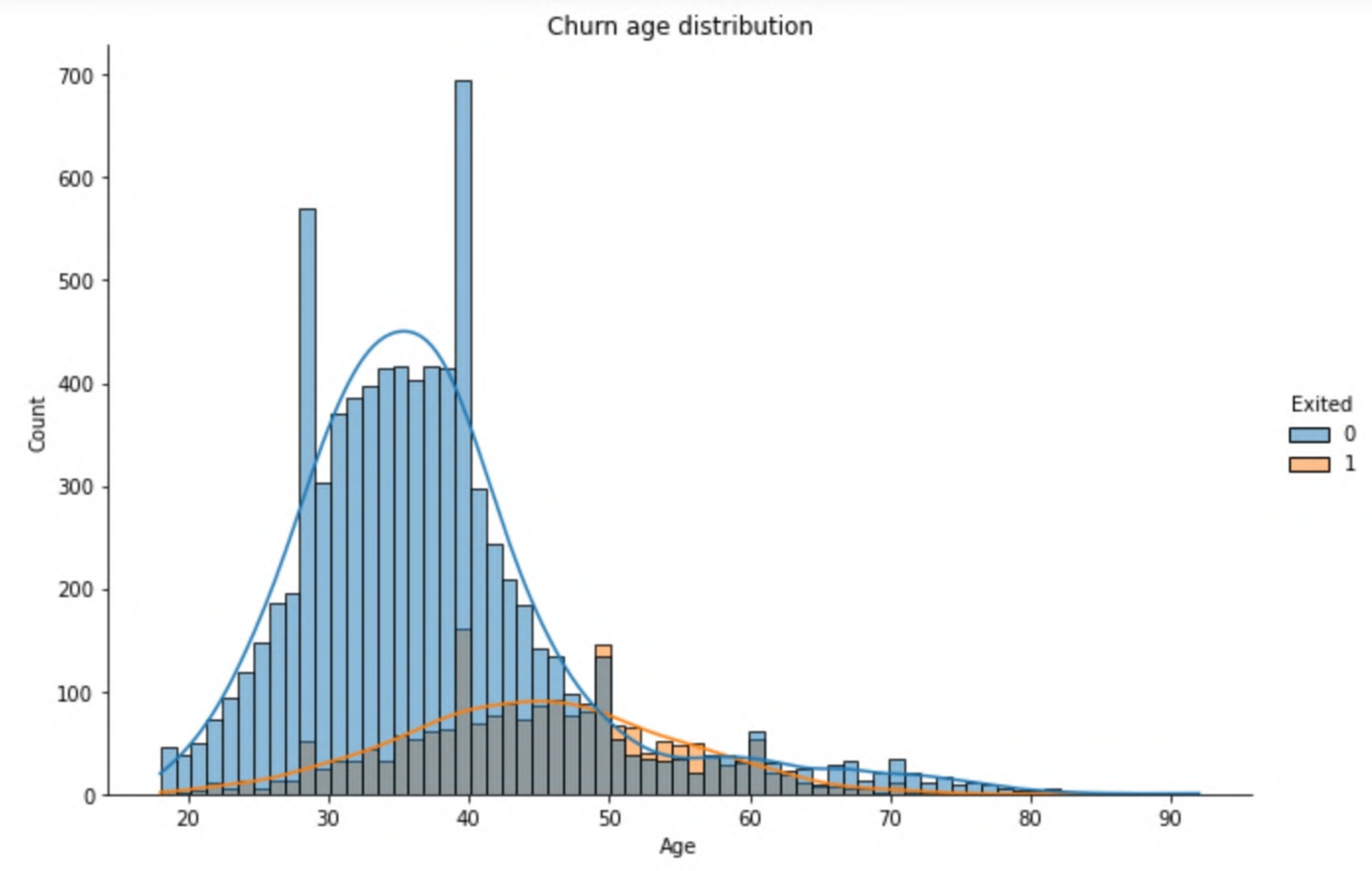
06 CHURN BY GENDER

The overall number of women customers is less than the number of male customers and the churn rate among females is 25% compared to 16% of males.





07 CHURN BY AGE

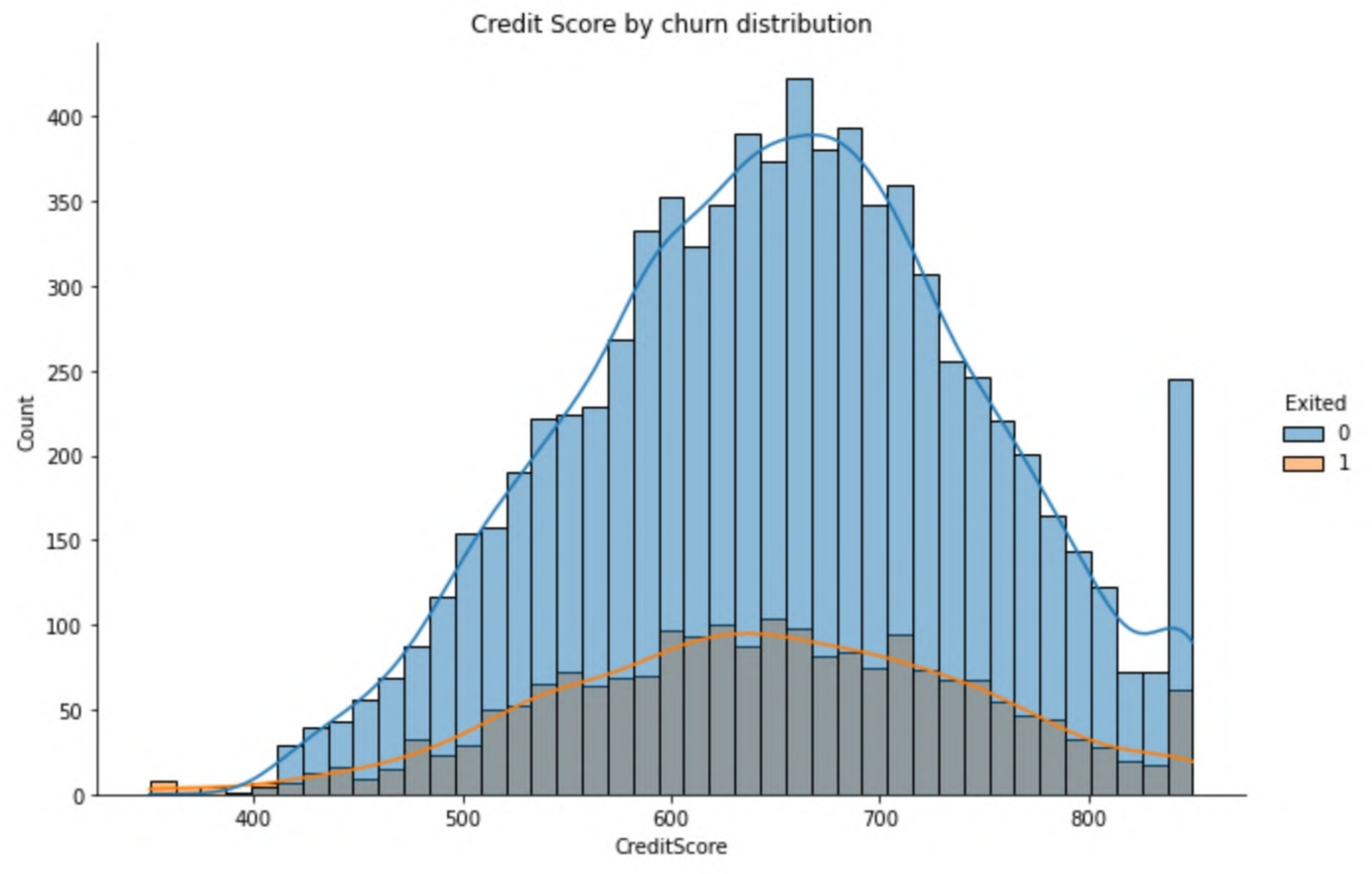


Age distribution for all customers and for those who stayed has a **long right tail**, however age distribution of customers who exited looks more normally distributed.

Mean age of churn customers is **44.84** years with **std of 9.76**

Non-parametric test using permutations had p-value=0 so we could **reject** H0: there is no difference in the mean age between customers who exited and those customers who stayed (H0: $\mu 1 = \mu 2$)

08 CHURN BY CREDIT SCORE

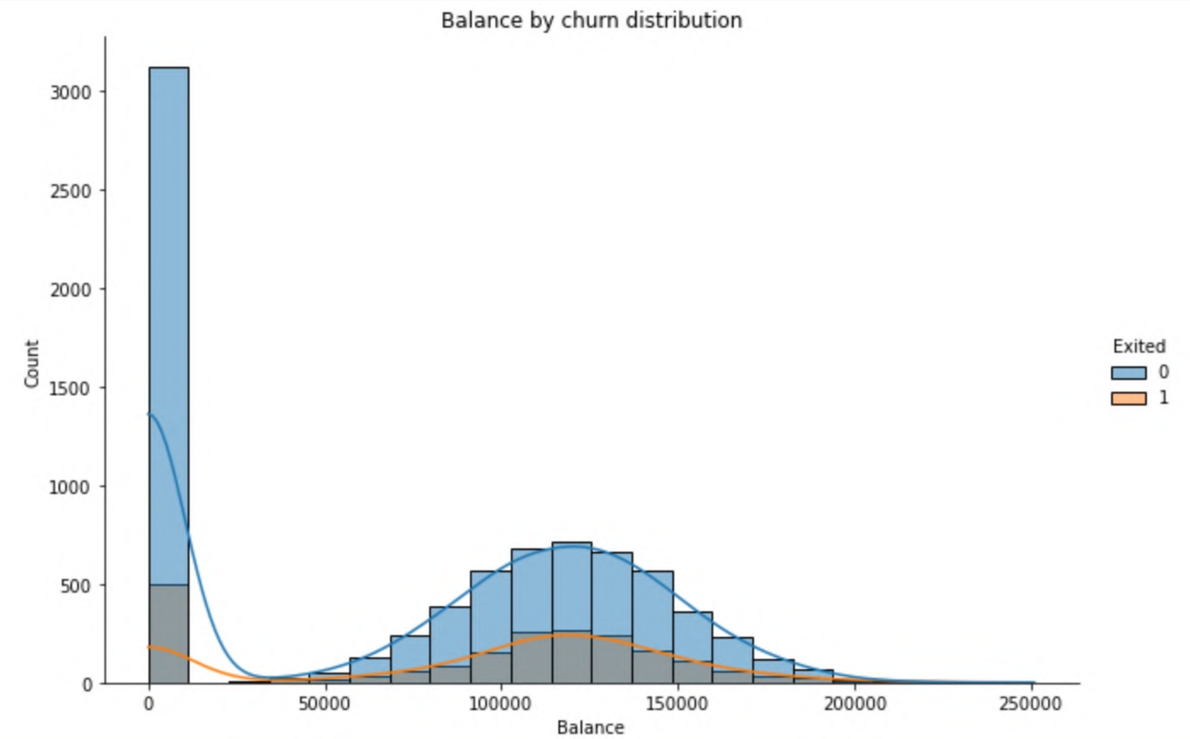


Most values of credit score for all customers is higher than **600**. With **mean of 650.52** and **std of 96.65**

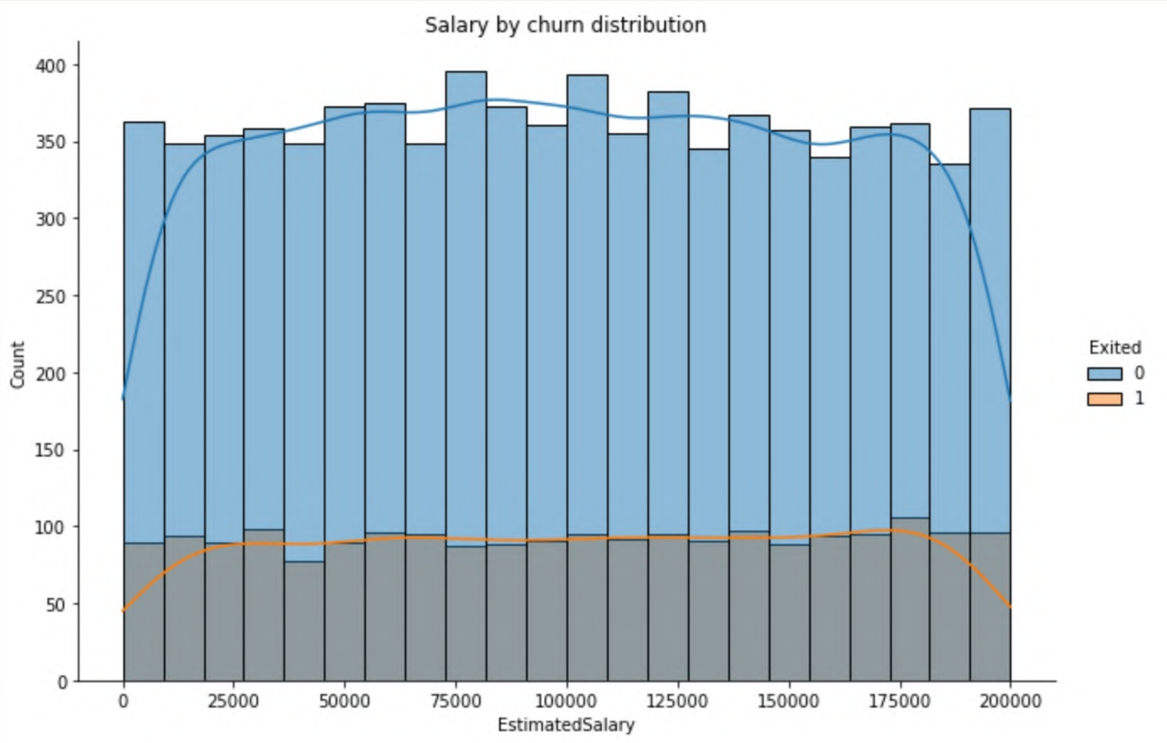
There is **no difference** between distributions of Credit score values among exited and stayed customers.

09 CHURN BY BALANCE & SALARY

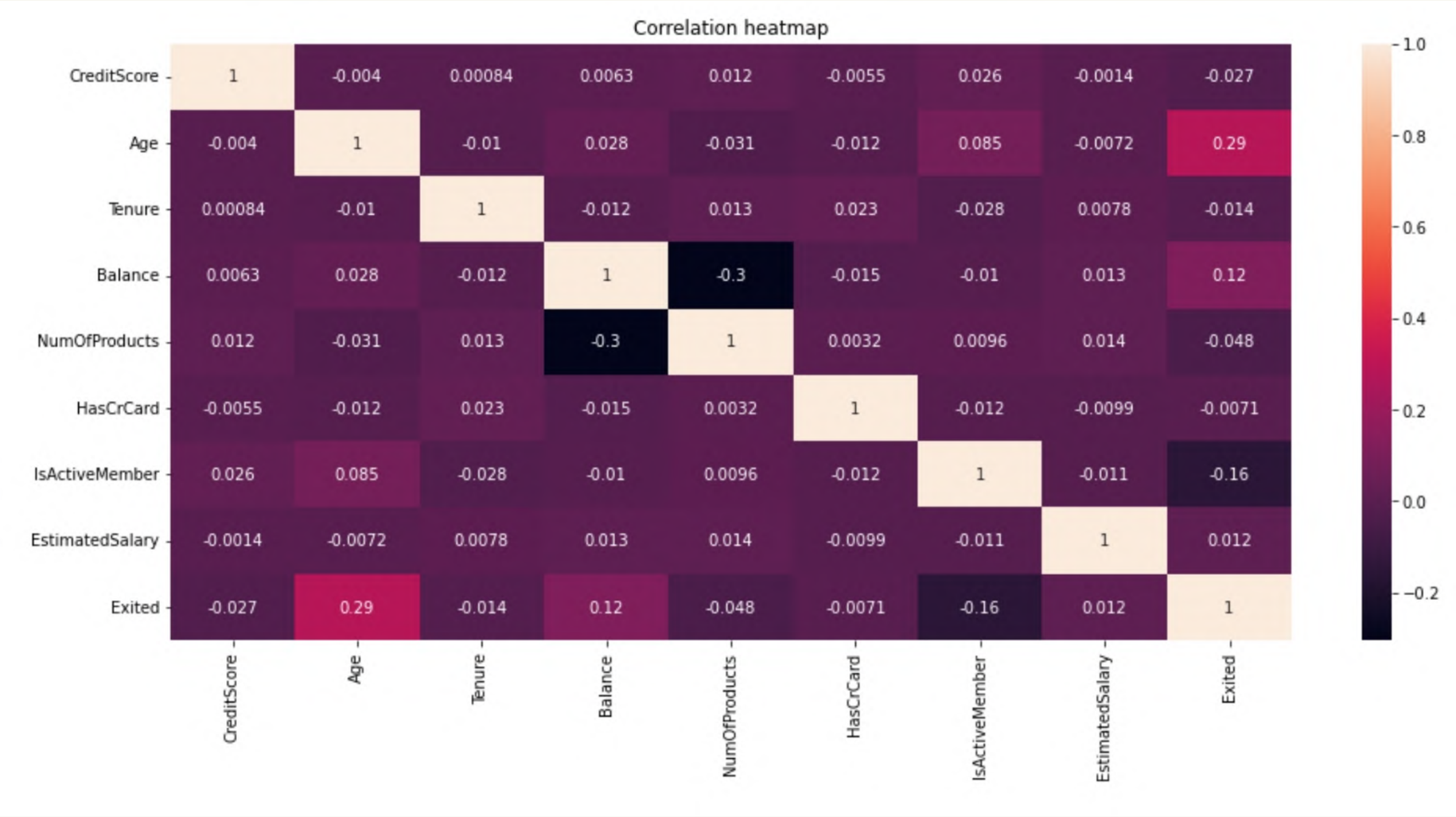
35% of customers have zero balance on their accounts for the reporting period



Estimated salary is uniformly distributed for all customers



10 CORRELATION MATRIX



Pearson correlation values are quite low for all features in the dataset.

The highest value for correlation coefficient to churn is **0.29** for **Age** field

11 FEATURE PREPARATION

- ☑ 'Tenure' and 'HasCrCard' were removed by checking Chi-square test. p-values are 0.18 and 0.49 correspondingly.
- ☑ 'EstimatedSalary' was removed from the dataset as we couldn’t reject the Hnull hypothesis with p-value of 0.225
- ☑ 'RowNumber', 'CustomerId', 'Surname' were removed as they don't have meaningful information for the modeling process.
- ☑ Categorical features ‘Geography’ and ‘Gender’ were transformed using dummies and the first column for each of them was dropped.

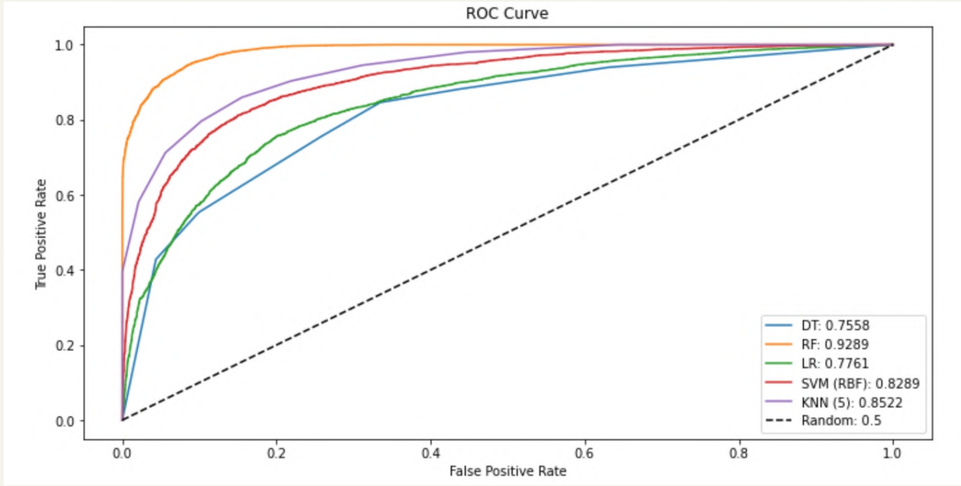
11 FEATURE PREPARATION

- ☑ Dataset was split into Train and Test datasets with 80/20 proportion.
- ☑ Imbalance of the dataset was solved with oversampling technique - SMOTE (Synthetic Minority Oversampling TEchnique) that performs data augmentation by creating synthetic data points based on the original data points.
- ☑ Standard scaler technique was chosen to scale data for ‘Age’, ‘Balance’ and ‘CreditScore’.

12 MODELING

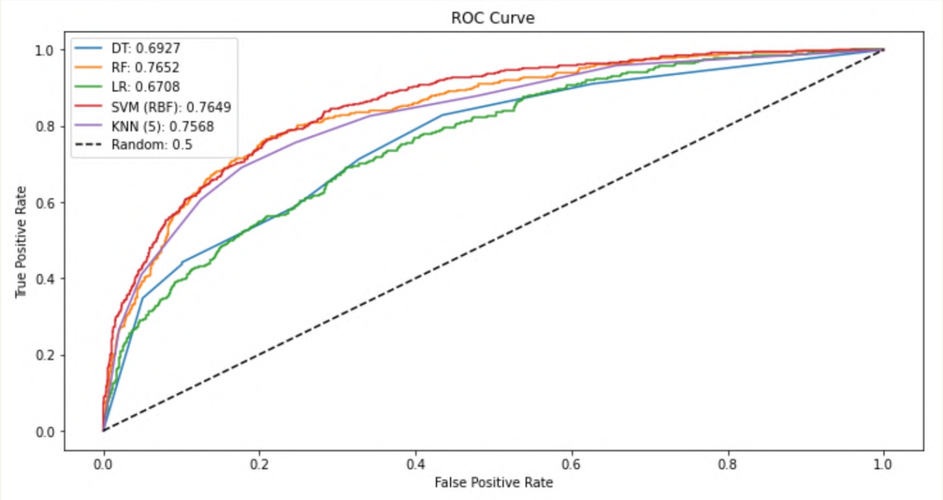
- Main metrics for evaluation of the models were: F1, Recall and Accuracy.
- Grid Search CV algorithm was applied to find out hyperparameters for models.
- Results for Train dataset is below:

Model name	Best parameters	Time to fit (s)	Accuracy	Recall	F1
Decision Tree	criterion='gini', max_depth=3	0.42	0.76	0.85	0.78
Logistic Regression	C= 0.1, max_iter=100, penalty= 'l1', solver= 'liblinear'	0.07	0.78	0.78	0.78
Random Forest	max_depth= 10, max_features= 'auto', min_samples_leaf=2, min_samples_split=5, n_estimators= 100	0.94	0.87	0.87	0.87
SVM	C=2, gamma=0.1, kernel='rbf'	61.52	0.83	0.82	0.83
KNN	n_neighbors=9	0.019	0.85	0.86	0.85



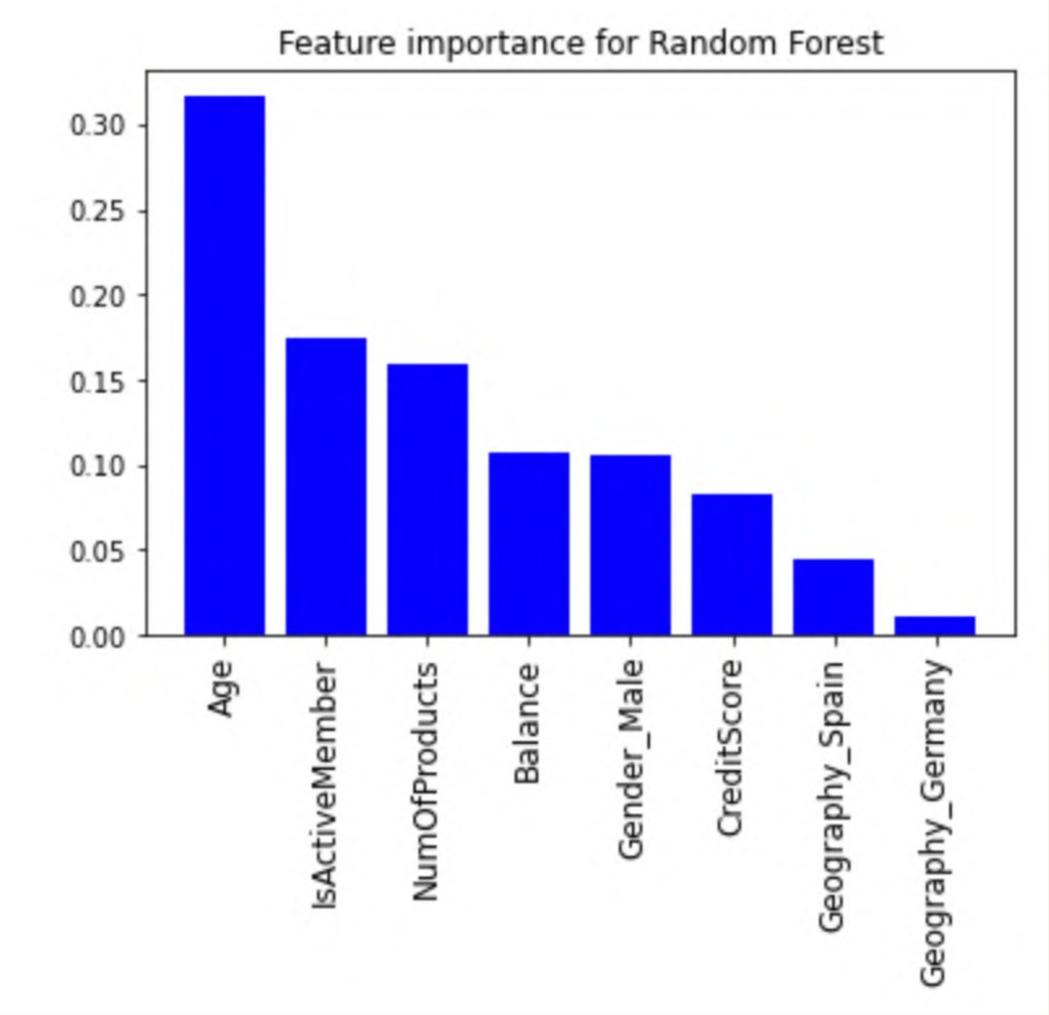
- Results of prediction on Test dataset:

Model name	Best parameters	Accuracy	Recall	F1
Decision Tree	criterion='gini', max_depth=3	0.68	0.71	0.48
Logistic Regression	C= 0.1, max_iter=100, penalty= 'l1', solver= 'liblinear'	0.72	0.59	0.46
Random Forest	max_depth= 10, max_features= 'auto', min_samples_leaf=2, min_samples_split=5, n_estimators= 100	**0.81**	**0.7**	**0.61**
SVM	C=2, gamma=0.1, kernel='rbf'	**0.8**	**0.7**	**0.6**
KNN	n_neighbors=9	0.8	0.69	0.58



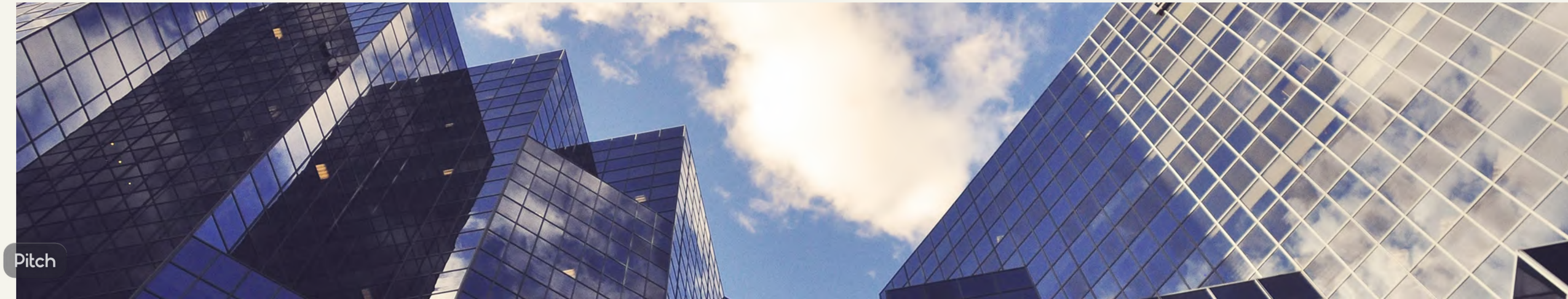
14 RESULTS

- ☑ Random Forest performed better among all models with an F1 score of 0.87 on Train set and 0.61 on Test set.
- ☑ Feature importance for Decision Tree and Random Forest showed common features - Age, IsActiveMember, NumOfProducts, Gender_Male.



15 IMPLEMENTATION

- ☑ Integrate classification model to rate customers based on probability of churn
- ☑ Set threshold for when to start customer success team interventions
- ☑ Continuously monitor and evaluate the effectiveness of retention efforts



16 FURTHER STEPS

Gather additional data for further analysis such as:

- ☑ Demographic data (marital status, education level, number of dependents, employment status, residence)
- ☑ Account information (types of accounts and products used)
- ☑ Transaction history (number of transactions, frequency, and amount)
- ☑ Customer interactions (phone, message, or visit contacts with bank representatives)
- ☑ Consider external factors (changes in economy or personal circumstances) that may influence churn.

THANK YOU