Problem overview

Churn, also known as customer attrition or turnover, refers to the process of customers ceasing to do business with a company or service. In the context of the banking industry, churn refers to customers closing their accounts, switching to a different bank, or taking their business to a non-bank competitor. Churn can have a significant impact on a bank's revenue and profitability, as it costs more to acquire new customers than to retain existing ones.

Banks use a variety of methods to measure churn, such as the number of accounts closed per month or the percentage of customers who leave the bank in a given period. Additionally, churn is not only important for the banking industry but also for other industries such as telecommunications, subscription-based services, and e-commerce.

Churn is a big problem for most companies. Losing customers requires investment in new customer acquisition to replace them. This could be several times more expensive than retaining existing customers, depending on the domain. Research on churn in the banking system can help banks identify the factors that contribute to customer churn, as well as develop strategies to reduce it. This can include identifying at-risk customers, understanding their reasons for leaving, and implementing targeted retention programs.

Additionally, research on churn can help banks improve their customer acquisition efforts by identifying the characteristics of customers who are most likely to stay with the bank. Overall, research on churn in the banking system can help banks improve their bottom line and better serve their customers.

Data source and detailed description

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.

Data Wrangling

Dataset was a .csv file that was loaded to Jupyter Notebook for further analysis.

Shape of the data is 10000 rows and 14 columns.
There was no missing data in the dataset.

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. And we will need to look for a solution to overcome this imbalance for further modeling.

Numerical features are: Age, Tenure, CreditScore, EstimatedSalary, Balance, NumOfProducts

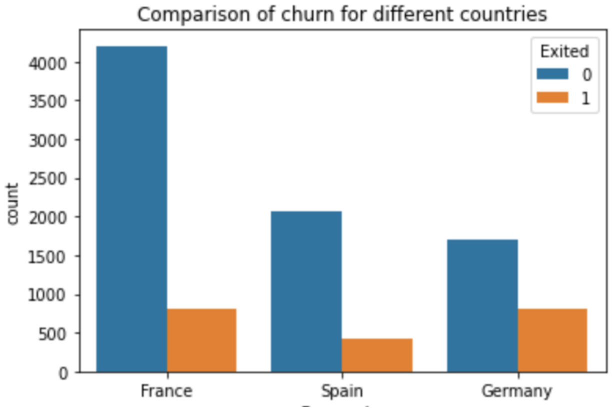
Categorical features are: Geography (Germany, Spain, France), Gender (Male, Female), HasCrCard(1,0), IsActiveMember(1,0)

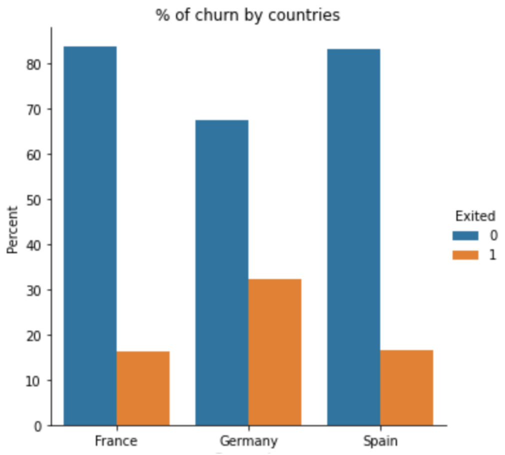
Target variable is a field Exited that is categorical and can take values 1, 0 where 1 - is the customer who exited the bank program.

We won’t use fields 'RowNumber', 'CustomerId', 'Surname' for further analysis and predictions so they can be deleted from the dataset.

Data visualization and exploration

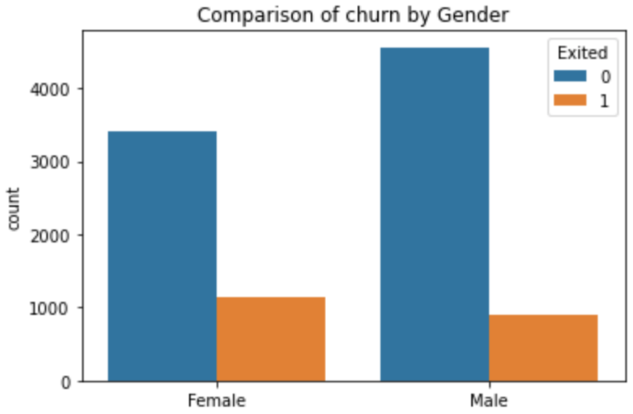
Let’s check how our users are distributed by geography and whether there is any relation to the churn:

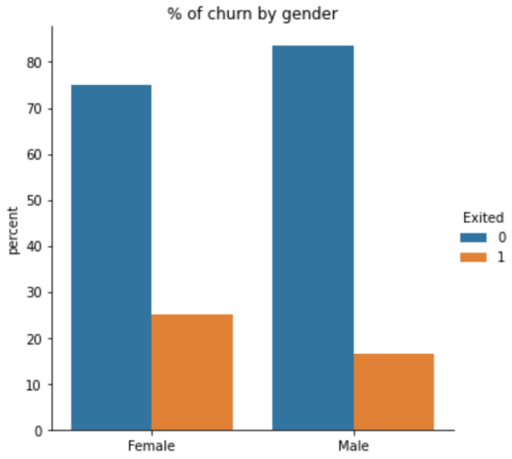




Germany has less users than other countries but there is highest percentage of churn - 32.44%

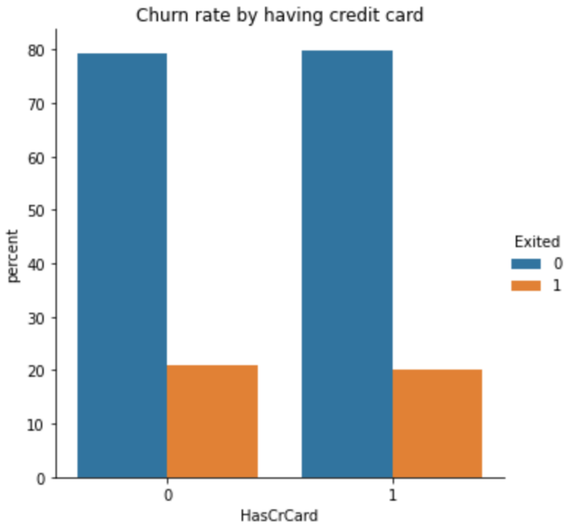
Gender distribution of customers and their churn:

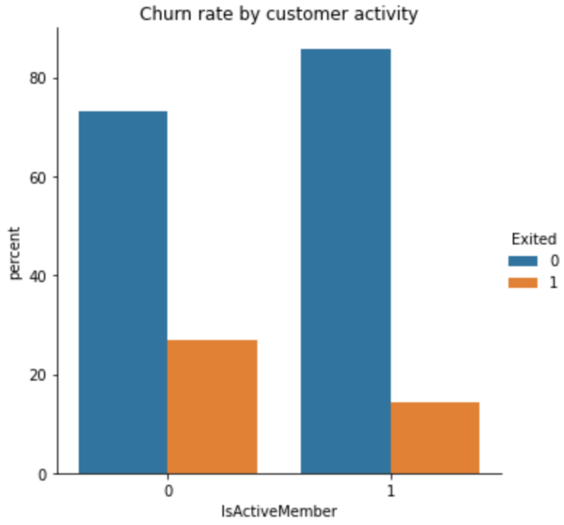




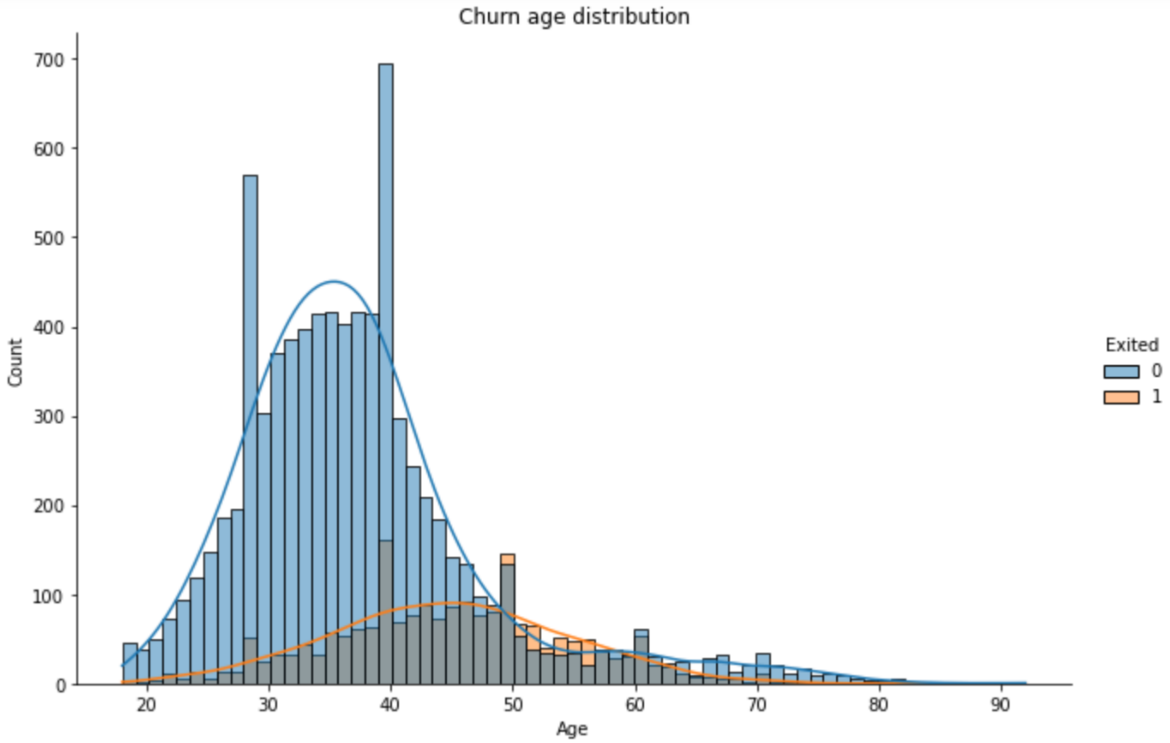
In numerical value all bank branches have almost the same number of churn among female and male customers. However, the overall number of women customers is less than the number of male customers and the churn rate among females is 25% compared to 16% of males.

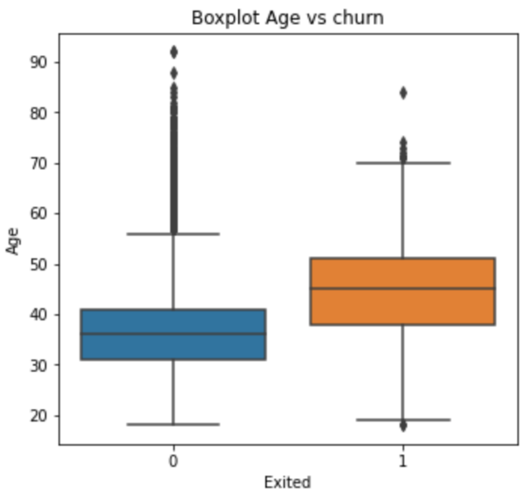
Almost 70.5% of customers have a credit card, and it could be interesting how this can influence customers to stay or leave the bank services, nevertheless both groups of customers have the same churn rate. The same time, customers who were active during the last reporting period tend to churn less than inactive users.





Age distribution for customers:

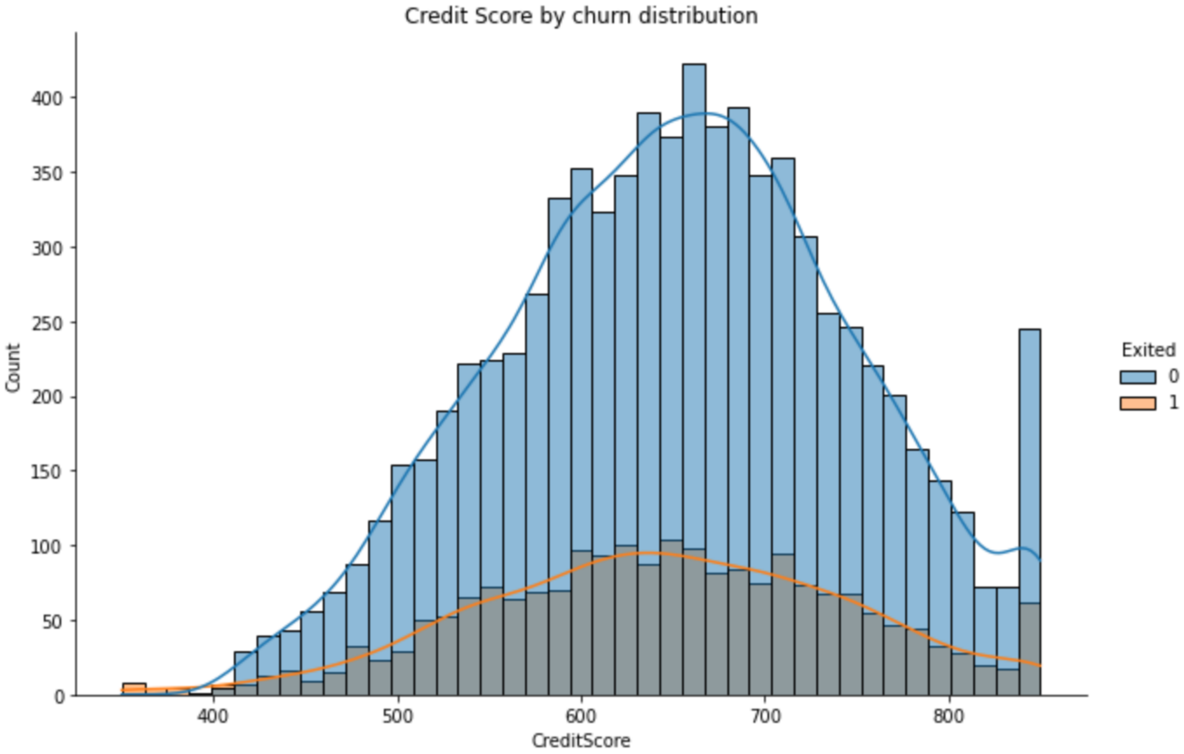


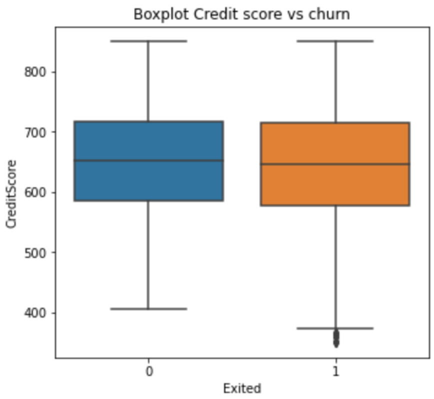


Age distribution for all customers and for those who stayed has a long right tail, however age distribution of customers who exited looks more normally distributed.

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

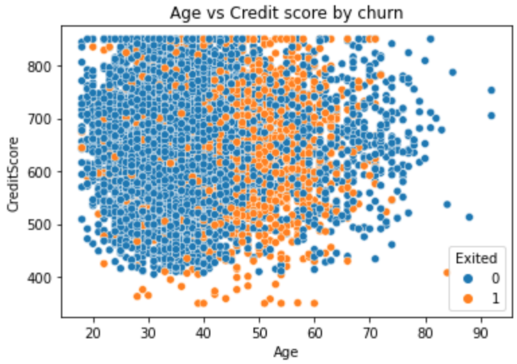
Credit Score distribution:





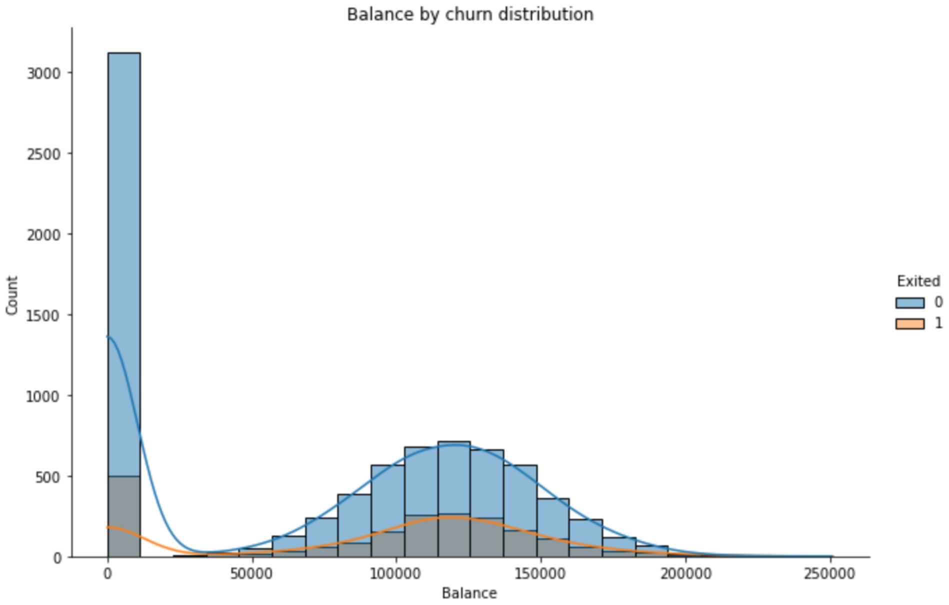
Most values of credit score for all customers is higher than 600. With mean of 650.52 and std of 96.65

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

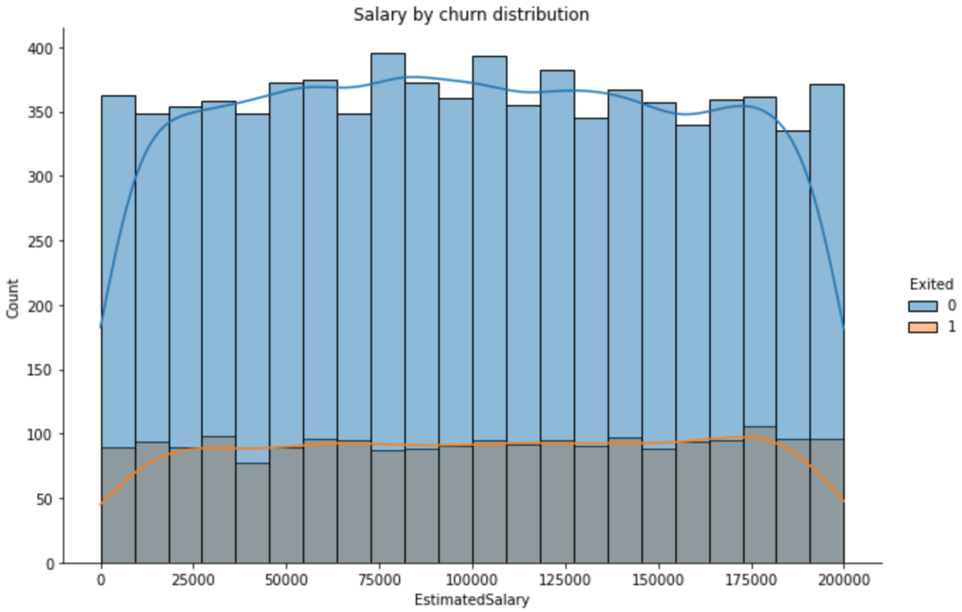


There is no visible relation between Age and Credit score and how they influence customer decision to leave.

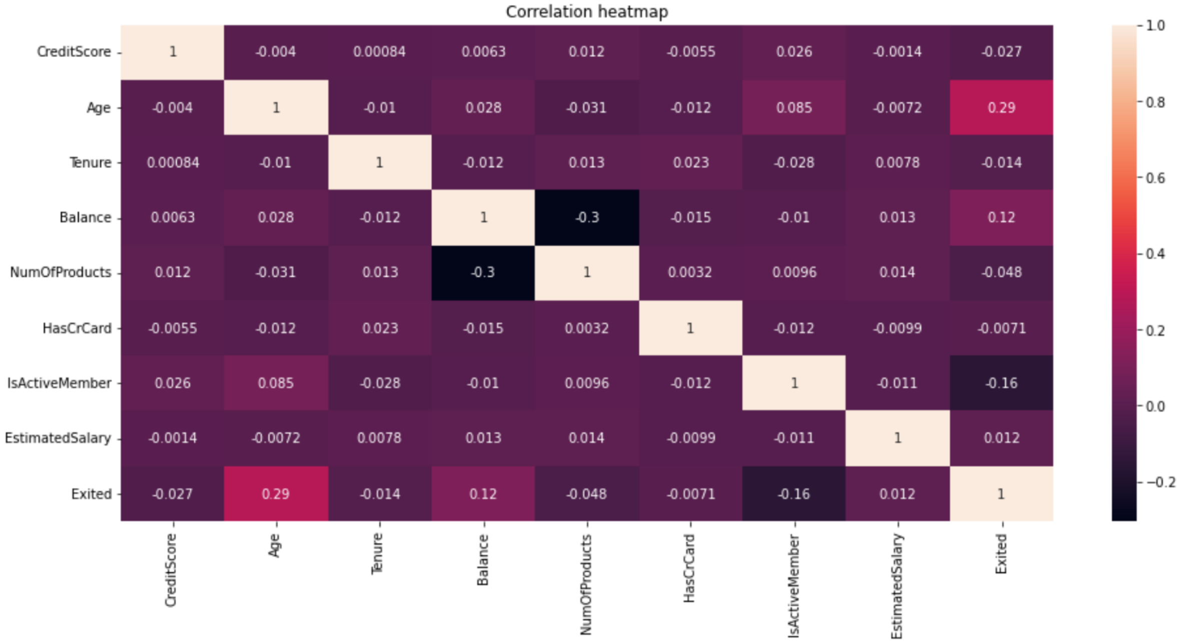
Surprisingly 35% of customers have zero balance on their accounts for the reporting period. Also for non-zero balance values are distributed normally:



Estimated salary is uniformly distributed for all customers:



Correlation between features:



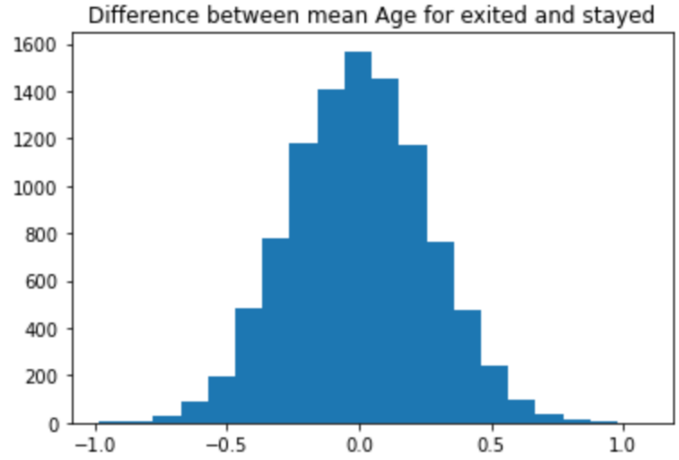
Pearson correlation values are quite low for all features in the dataset. The highest value for correlation coefficient to churn is 0.29 for Age field.

Statistical tests

So let's run Hypothesis testing to check whether there is a difference for age median between those customers who churned and those who stayed.
Hypothesis formulation H_{null}: there is no difference in the mean age between customers who exited and those customers who stayed (H0: μ1 = μ2)
$H_{alternative}$: there is a difference in the mean age between customers who exited and those customers who stayed (HA: μ1 ≠ μ2). We're also going to pick a significance level of 0.05.

First of all we need to determine distribution of the data and then decide how hypothesis testing can be run. Two tests were run to determine whether Age distribution is normal but in both cases p-value was 0. Due to the logic of the tests checking whether data is normally distributed the p-value less than 0.05 reject H_{null} that stated as our data was normally distributed. In this case we can't use T-test and will have to use a non-parametric test here using permutations. In our case we used 10000 permutations to calculate the difference between mean of Age for exited and stayed customers. This is simply a label for statistical tests used when the data aren't normally distributed. These tests are extraordinarily powerful due to how few assumptions we need to make.

As we are checking whether means are equal or not, then we have two-tailed test and need to count values on both sides of the histogram that are greater than observed difference between means and less than -1*observed difference between means. In this case we can compare abs values. P-value for run test is 0 so we can reject Hnull that means are equal. So feature Age has influence on whether a customer exits or stays with the bank.



Similar test with the same logic was done for the Estimated salary feature but the p-value was 0.225. This means that EstimatedSalary feature doesn't have significant influence on user exiting outcome. The corresponding correlation value is 0.012. Estimated salary feature can be removed from the dataset as we couldn’t reject the H_{null} hypothesis.

For categorical features Chi-square tests were run.
An assumption of few variables showing positive impact are true or not.

Do these variables have a significant impact on churn?

The Chi-square test of independence determines whether there is a statistically significant relationship between categorical variables. It is a hypothesis test that answers the question — do the values of one categorical variable depend on the value of other categorical variables?

The Chi-square test of association evaluates relationships between categorical variables. Like any statistical hypothesis test, the Chi-square test has both a null hypothesis and an alternative hypothesis.

H_{null} hypothesis: There are no relationships between the categorical variables. If you know the value of one variable, it does not help you predict the value of another variable.

$H_{alternative}$ hypothesis: There are relationships between the categorical variables. Knowing the value of one variable does help you predict the value of another variable.

Alpha was set to 0.05

Name of Feature	Chi-sq statistics	P-value	Degree of freedom
Geography	301.26	3.83e-66	2
Gender	112.92	2.25e-26	1
NumOfProduct	1503.63	0	3
IsActiveMember	242.99	8.79e-55	1
HasCrCard	0.47	0.49	1
Tenure	13.9	0.18	10

From these results we can see that we couldn’t reject H_{null} hypothesis for HasCrCard and Tenure fields. So we can remove those from our dataset.

Feature preparation

Categorical features transformation

Categorical features ‘Geography’ and ‘Gender’ were transformed using dummies and the first column for each of them was dropped.

Train/Test split

Dataset was splitted into Train and Test datasets with 80/20 proportion. All further manipulations with data were fitted to the Train set and applied to the Test set in order to avoid data leakage from the Train set.

Unbalance class approach

Initial dataset was unbalanced with a minority class of Exited users of 20%. In order to improve further classification we can apply an oversampling technique - SMOTE (Synthetic Minority Oversampling TEchnique) that performs data augmentation by creating synthetic data points based on the original data points. The advantage of SMOTE is that you are not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

A general downside of the approach is that synthetic examples are created without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes.

Data scaling

Standard scaler technique was chosen to scale data for ‘Age’, ‘Balance’ and ‘CreditScore’. Standardization of datasets is a common requirement for many machine learning estimators implemented in scikit-learn; they suppose that features do more or less look like standard normally distributed data: Gaussian with zero mean and unit variance.

Modeling

In this section we used different classification algorithms for churn classification modeling:

1. The Decision Tree was used as a baseline model to understand how it can fit our train data and to see feature importance.
2. Grid Search CV algorithm was applied to find out hyperparameters for next models:
    - Logistic Regression
    - Random Forest
    - Support Vector Machine classification
    - KNN

3. All best estimators for these models were fit to the whole training set. Altogether with performing learning curve algorithms to understand how models perform with different data volumes.
4. Applied fitted models to test sets and analyzed performance.

Taking into account that that initial data set was unbalanced and we are trying to predict churn that is minority class, our main metrics for model performance will be F1, Recall, Accuracy.

Accuracy is the ratio of the number of correct predictions to the total number of input samples.

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.

The **F1**-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean.

Models overview

Decision Tree: A decision tree is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The benefits of decision trees include that they are easy to understand and interpret, can handle both categorical and numerical data, and can handle multi-output problems. Drawbacks include that they can be prone to overfitting, especially with large trees.

Logistic Regression: Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The benefits of logistic regression include that it is a simple and efficient algorithm, it is easy to interpret the results, it can handle non-linearly separable data, and it can handle multi-class problems. Drawbacks include that it may fail when the relationship between the independent and dependent variables is non-linear.

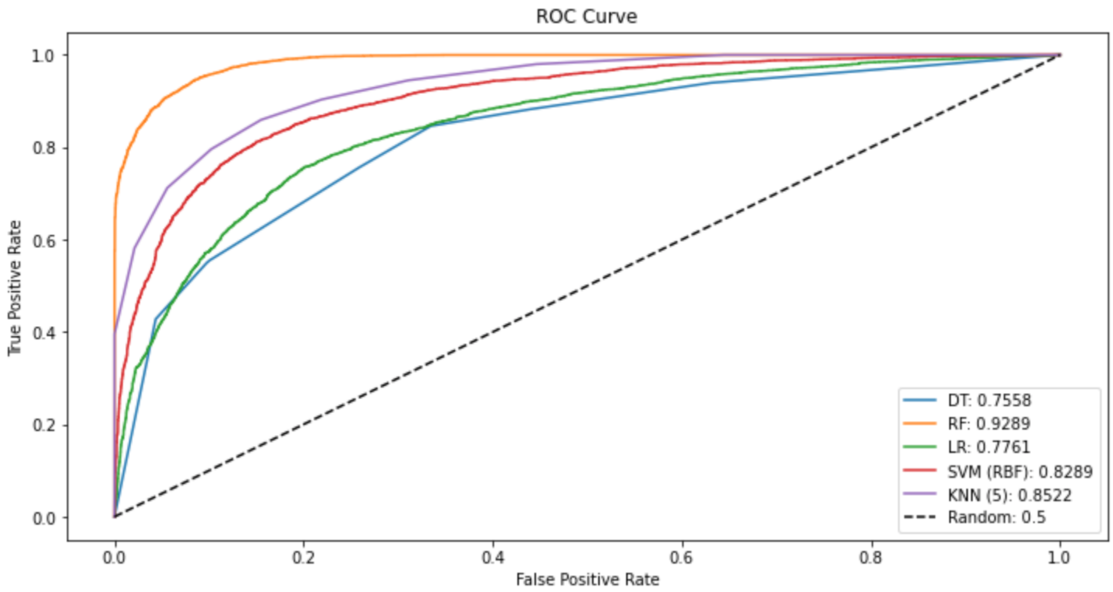
Random Forest: Random Forest is an ensemble learning method for classification and regression. It constructs a multitude of decision trees at training time and output the class that is the mode of the classes (classification) of the individual trees. The benefits of Random Forest include that it is less prone to overfitting than a single decision tree, it can handle large amounts of data, it can handle missing data, and it can handle multi-output problems. Drawbacks include that it can be computationally expensive and it can be difficult to interpret the results.

SVM: Support Vector Machine (SVM) is a supervised learning algorithm that can be used for classification and regression problems. The benefits of SVM include that it can handle non-linearly separable data and it can handle high-dimensional data. Drawbacks include that it may be sensitive to the choice of kernel and it can be computationally expensive.

KNN: k-Nearest Neighbors (KNN) is a simple and effective algorithm for classification and regression problems. The benefits of KNN include that it is easy to understand and implement, it can handle multi-class problems, and it can handle large amounts of data. Drawbacks include that it can be computationally expensive and it may be sensitive to the choice of distance metric.

After running grid search with cross-validation for described models we had next results for Train data:

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

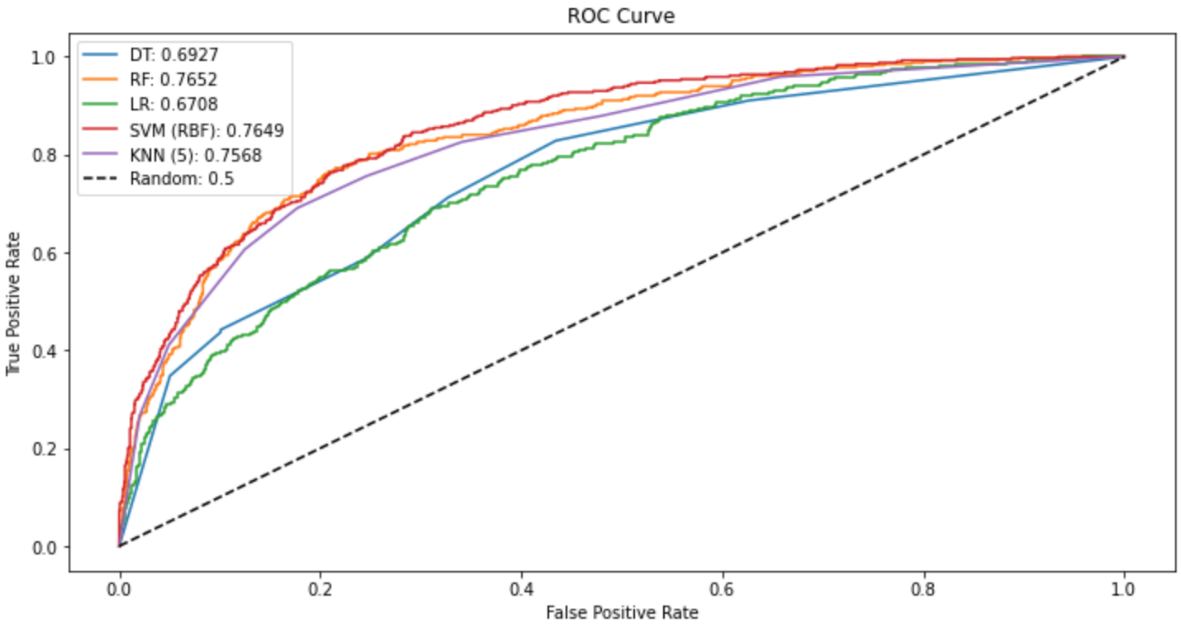


ROC-AUC curve is one of the most widely used metrics for evaluation. It is used for binary classification problems. The higher the AUC, the better the model's performance at distinguishing between the positive and negative classes.

Among all models Random Forest provided better results by all significant metrics.

We had to check these models' results on predicting unseen data from the Test set:

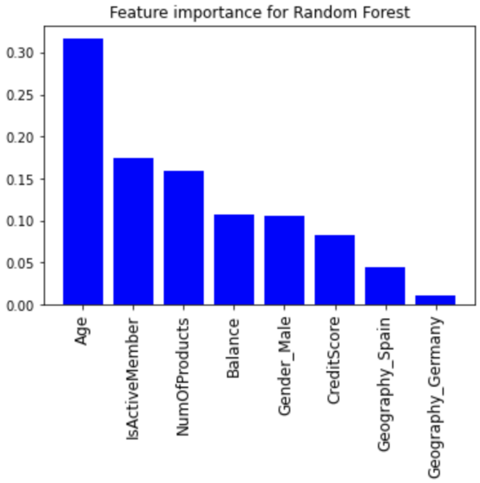
Model name	Parameters	Accuracy	Recall	F1
Decision tree	criterion=’gini’, max_depth=3	0.68	0.71	0.48
Logistic Regression	C= 0.1, max_iter=100, penalty= 'l1', solver= 'liblinear'	0.72	0.59	0.46
Random Forest	max_depth= 10, max_features= 'auto', min_samples_leaf=2, min_samples_split=5, n_estimators= 100	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



Random Forest classification can predict churn with 80% accuracy and 70% recall. ROC-AUC score for Random Forest is higher than for other algorithms.

Summary

Random Forest performed better among all models with an F1 score of 0.87 on Train set and 0.61 on Test set. This significant drop in performance can be related to overfitting on training data.



Feature importance for Decision Tree and Random Forest showed common features - Age, IsActiveMember, NumOfProducts, Gender_Male.

Second choice among models is SVM. It has an F1 score of 0.83 on the Train set and 0.6 on the Test. However, it took longer to fit the model to the dataset.

Further steps.

Bank can integrate classification model in order to rate customers by their probability to churn based on important features for the model and then set the threshold when to start interventions by the customer success team in order to help clients to stay with bank programs.

Overall performance of the models didn’t show a significant level of accuracy and clear understanding why customers are leaving. Therefore bank can gather additional data for further analysis for example:

- Demographic data, like marital status, education level, number of dependents, employment status, residence;
- Account information: type of accounts and products they are using;
- Transaction history: number of transactions for analyzed period with their frequency and amount;
- Customer interactions: phone, message or visit contacts with bank's representative, their number and some sort of classification;
- External factors, such as changes in the economy or the customer's personal circumstances, can also influence the likelihood of churn. In our case we had information for 3 different countries but not obvious reasons why users from Germany churn more. Additional data and analysis may help to identify the cause of this trend.