PREDICTING CHURN OF BANK CUSTOMERS

JOSH EBY-20BCAB08

MUSAIB AHMAD KHANDAY-20BCAB32

VRISHAB THOMAS-20BCAB13

CONTENT

ABSTRACT

Customer retention rate has a strong impact on customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Customer value analysis along with customer churn predictions will help marketing programs target more specific groups of customers. We predict customer churn with random forest techniques and analyse the churning and non-churning customers by using data from a consumer retail banking company. The result of the case study shows that using conventional statistical methods to identify possible churners can be successful.

INTRODUCTION

The customer churn, also known as customer attrition, refers to the phenomenon whereby a customer leaves a company. Banks aim to keep their current consumers while also attempting to attract new ones. In order to discover existing consumers without losing them, develop new goods, and make new strategic decisions for customer retention, it is crucial to analyse customer turnover. The customer churn analysis, a key topic in banks' customer relationship management, is the subject of this study. The management will be able to target consumers with promotions and classify those who are most likely to churn early with the aid of customer churn identification in banks. It will also be possible to gain insight into the elements that should be taken into account when maintaining customers. Utilizing actual banking data, the research aims to estimate the explainable machine learning model and assess these using test data.

PROBLEM STATEMENT

Marketing costs to acquire new customers are high. Therefore, it is important to retain customers so that the initial investment is not wasted. It has a direct impact on the ability to expand the company. In this project we shall predict the probability a customer is likely to churn using machine learning techniques.

METHODOLOGY

i. Data Collection

The dataset has been collected from Kaggle. There are around ten thousand rows and eleven columns.

ii. Exploratory Data Analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summery statistics and geographical representations.

Here we see the dimensions of the dataset and then find the unique values in the data attributes. We can see that the rownumber attribute is similar to the counter of records, the customerid attribute is a unique identifier for a certain customer, and the surname attribute is likewise used in client profiling. As a result, we will eliminate them from our dataset because they do not provide valuable information for the research.

iii. The distribution of categorical variables are shown below graphically

**Geographical distribution in customer attrition**
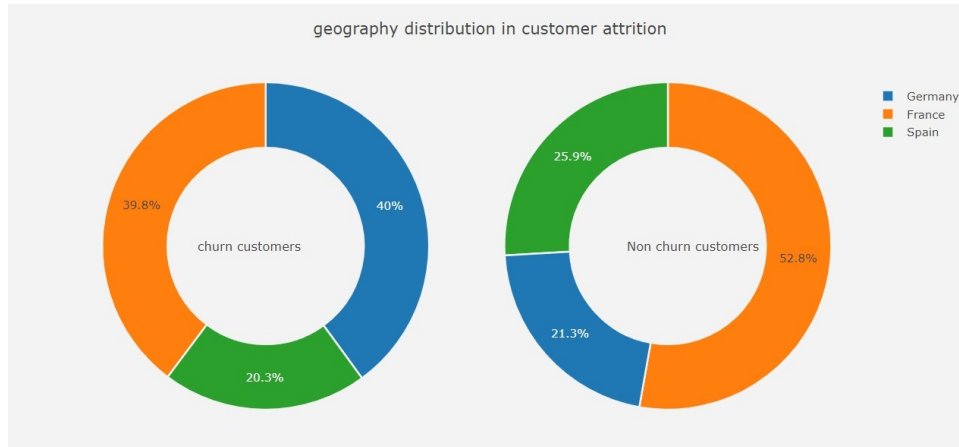
According to the output above, among churned customers, those geographically based in Germany have a high rate of churn with 40%, followed by France with 39.8 percent and Spain with 20.3 percent. France leads the way with 52.8 percent of non-churn clients, followed by Spain with 25.9 percent and Germany with 21.3 percent.

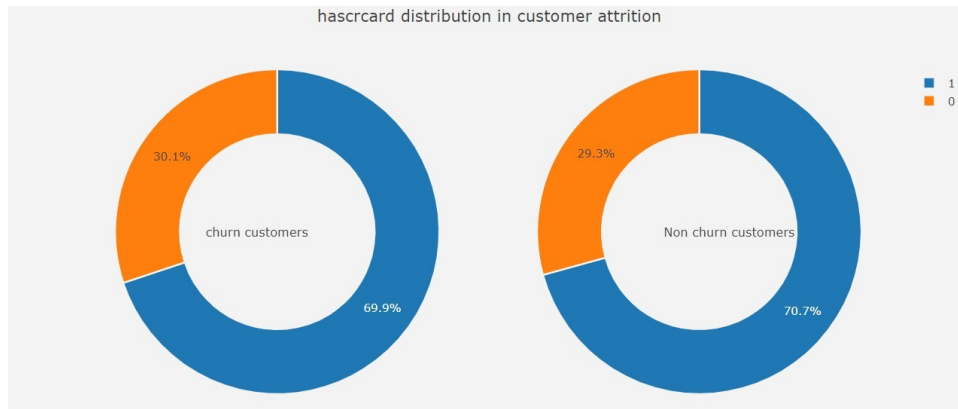**Gender distribution in customer attrition**
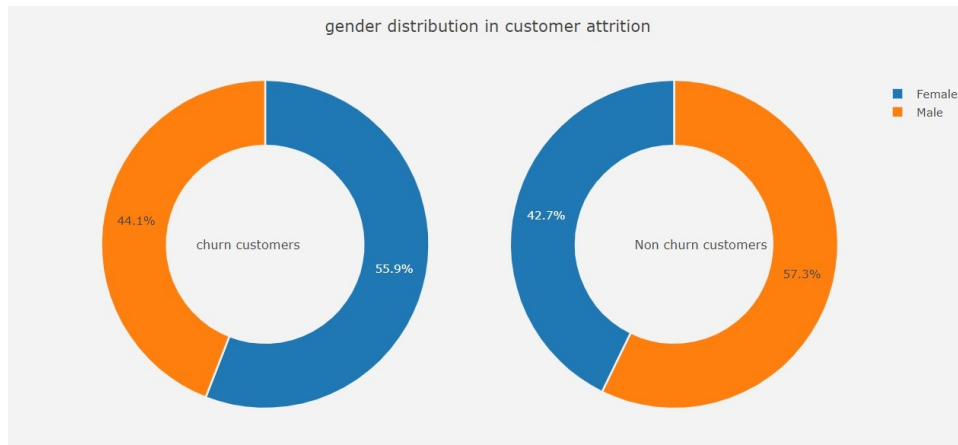
According to the data above, females account for 55.9 percent of turnover consumers, while males account for 44.1 percent. In terms of non-churn clients, 57.3 percent are men and 42.7 percent are women.

**Number of products distributions in customer attrition**
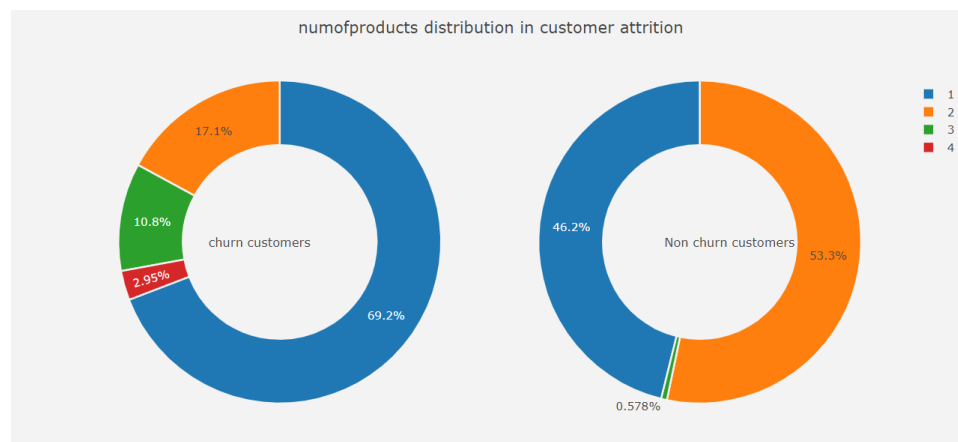
The graph above illustrates that among turnover consumers, those who use one product have the highest rate of 69.2 percent, followed by those who use two goods (17.1 percent), three products (10.8%), and four items (2.95 percent). Customers with two items account for 53.3 percent of non-churn consumers, customers with one product account for 46.2 percent, and customers with three goods account for 0.58 percent.

Has a card distribution in customer attrition

hascrcard distribution in customer attrition

According to the data above, those who have a card account for 69.9 percent of turnover clients, while those who do not have a card account for 30.1 percent. In the case of non-churn consumers, 70.7 percent have a card, whereas 29.3 percent do not.

Is an active member distribution in customer attrition
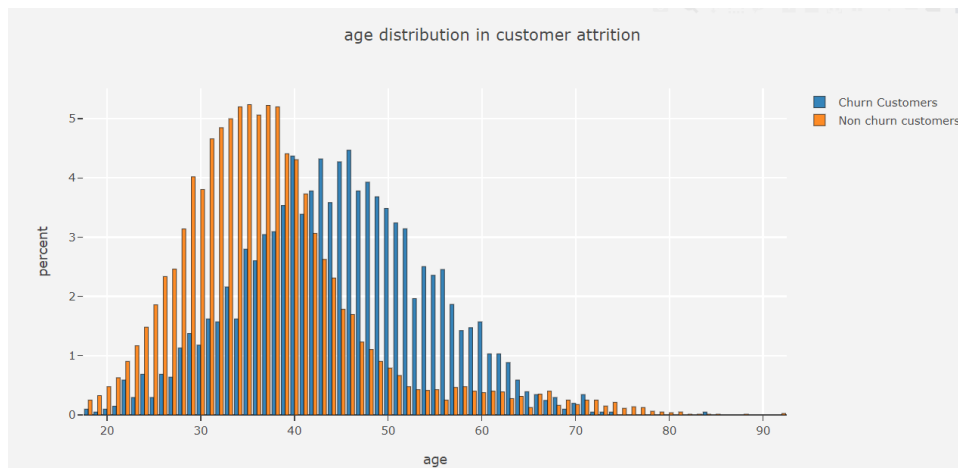

isactivemember distribution in customer attrition
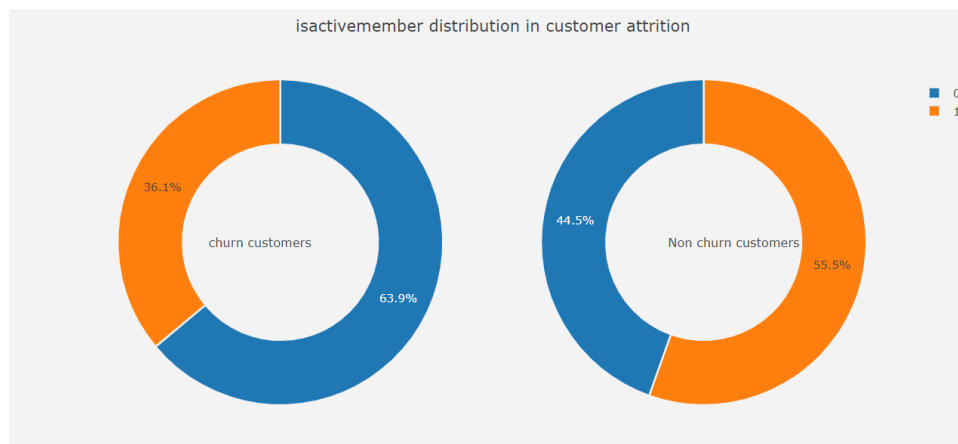
The result above demonstrates that among the churned customers, those who are not active members have a high rate of churn (63.9%), whereas active members have a rate of churn (36.1%). Active members account for 55.5 percent of non-Chun clients, while non-active members account for 44.5 percent.

iv. Distribution of the continuous variables
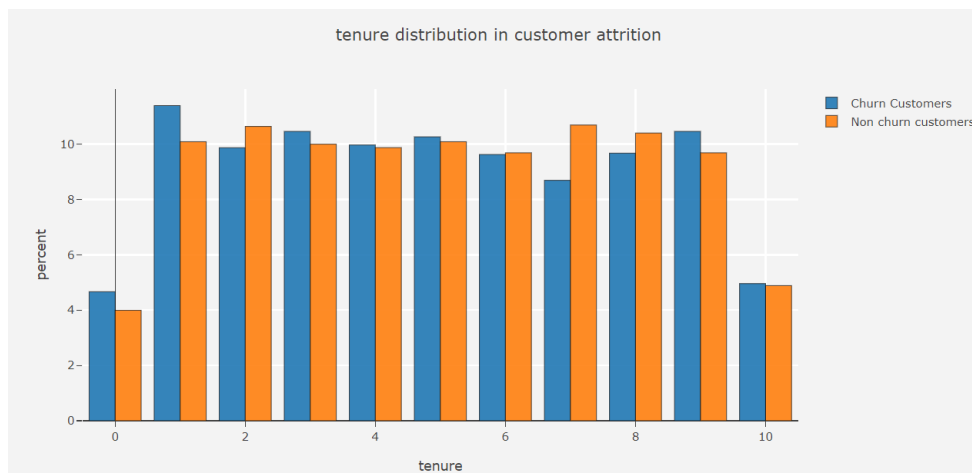
Credit score distribution in customer attrition

Age distribution in customer attrition
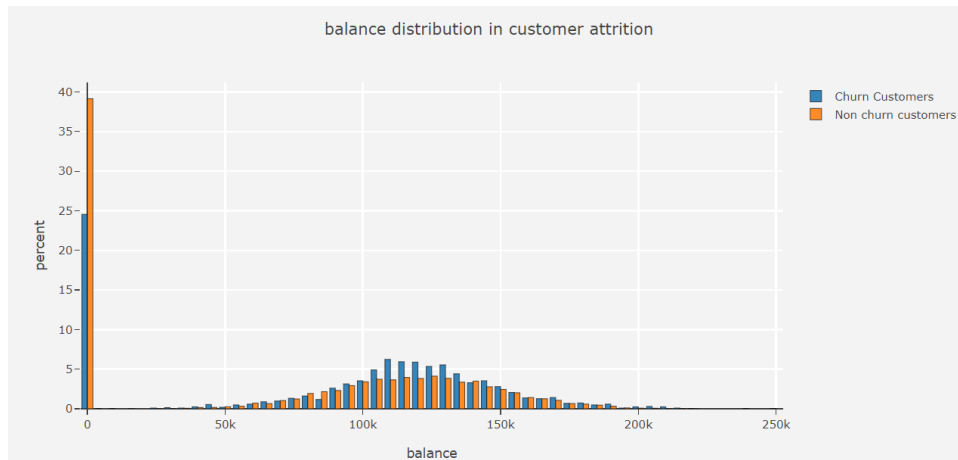
age distribution in customer attrition

According to the graph above, clients between the ages of 46 are the most churned.

Tenure distribution in customer attrition

tenure distribution in customer attrition

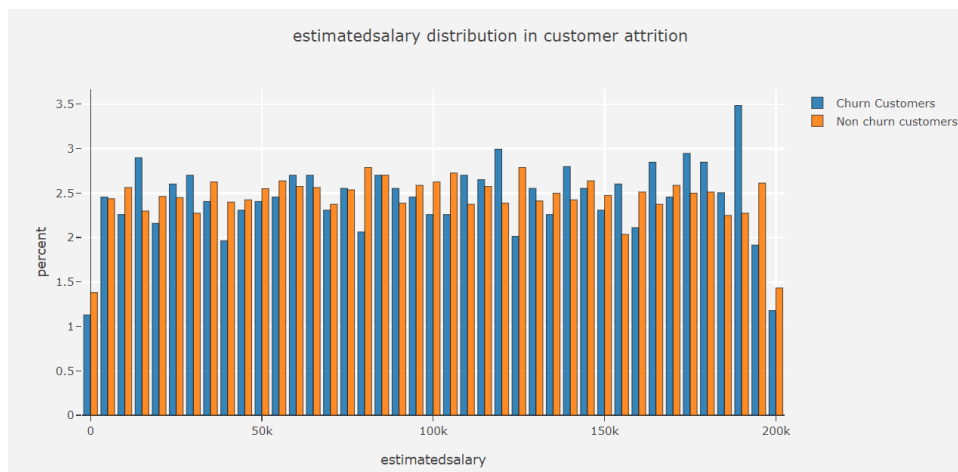**Balance distribution in customer attrition**

balance distribution in customer attrition

Estimated salary distribution in customer attrition

estimatedsalary distribution in customer attrition

v. Correlation

The term correlation refers to a mutual relationship or association between quantities. So, here, we will find out what kind of association is present among the different data attributes.

	creditscore	age	tenure	balance	numofproducts	hascrcard	isactivemember	estimatedsalary	exited
creditscore	1.000000	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.001384	-0.027094
age	-0.003965	1.000000	-0.009997	0.028308	-0.030680	-0.011721	0.085472	-0.007201	0.285323
tenure	0.000842	-0.009997	1.000000	-0.012254	0.013444	0.022583	-0.028362	0.007784	-0.014001
balance	0.006268	0.028308	-0.012254	1.000000	-0.304180	-0.014858	-0.010084	0.012797	0.118533
numofproducts	0.012238	-0.030680	0.013444	-0.304180	1.000000	0.003183	0.009612	0.014204	-0.047820
hascrcard	-0.005458	-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.009933	-0.007138
isactivemember	0.025651	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.011421	-0.156128
estimatedsalary	-0.001384	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.000000	0.012097
exited	-0.027094	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.012097	1.000000

vi. Feature importance

We need to know what the key features are. In order to find out, we will train the model with the Random Forest classifier. After that, we'll have a good notion of what characteristics are crucial to us.

vii. MODELING

Random Forest Model

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.

RESULT

The random forest model got an accuracy of 85.04%

CONCLUSION

In this project we build a model that predict how likely a customer is going to churn. During exploratory data analysis we found out that the female customer are the most likely to churn, customer that are located in Germany are the most churned, and also customer using only one product are the most churned. After creating a random forest model, we got an accuracy of 85.04%. The model can achieve better performance providing a lot of historical data for the training phase.

REFERENCE

https://www.researchgate.net/publication/47749836_Customer_churn_prediction_-_A_case_study_in_ret

https://www.researchgate.net/publication/340855263_Churning_of_Bank_Customers_Using_Supervised_I

https://www.econstor.eu/bitstream/10419/176421/1/10.1186_s40854-016-0029-6.pdf

https://www.diva-portal.org/smash/get/diva2:1435454/FULLTEXT01.pdf