BANK CUSTOMER CHURN PREDICTION

1. INTRODUCTION

The customer churn, also known as customer attrition, refers to the phenomenon whereby a customer leaves a company. Some studies confirmed that acquiring new customers could cost five times more than satisfying and retaining existing customers. Actually, many benefits encourage the tracking of the customer churn rate, for example:

- 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;
- etc.

In this project, our goal is to predict the probability of a customer is likely to churn using machine learning techniques.

2. DATA

Based on definition of our problem, factors that will influence our decission are:

- Credit Score;
- Geography;
- Gender;
- Age;
- Tenure;
- Balance;
- Num Of Products;
- Has Credit Card;
- Is an Active Member;
- Estimated Salary.

Following data sources was extracted from the kaggle database:

The dataset is Churn for Bank Customers and the content is:

- RowNumber—corresponds to the record (row) number and has no effect on the output.
- CustomerId—contains random values and has no effect on customer leaving the bank.
- Surname—the surname of a customer has no impact on their decision to leave the bank.
- CreditScore—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
- Geography—a customer’s location can affect their decision to leave the bank.
- Gender—it is interesting to explore whether gender plays a role in a customer leaving the bank.
- Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.

- Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
- Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
- NumOfProducts—refers to the number of products that a customer has purchased through the bank.
- HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
- IsActiveMember—active customers are less likely to leave the bank.
- EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- Exited—whether or not the customer left the bank.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 1 - first 5 lines of the dataset

	count	mean	std	min	25%	50%	75%	max
RowNumber	10000.0	5.000500e+03	2886.895680	1.00	2500.75	5.000500e+03	7.500250e+03	10000.00
CustomerId	10000.0	1.569094e+07	71936.186123	15565701.00	15628528.25	1.569074e+07	1.575323e+07	15815690.00
CreditScore	10000.0	6.505288e+02	96.653299	350.00	584.00	6.520000e+02	7.180000e+02	850.00
Age	10000.0	3.892180e+01	10.487806	18.00	32.00	3.700000e+01	4.400000e+01	92.00
Tenure	10000.0	5.012800e+00	2.892174	0.00	3.00	5.000000e+00	7.000000e+00	10.00
Balance	10000.0	7.648589e+04	62397.405202	0.00	0.00	9.719854e+04	1.276442e+05	250898.09
NumOfProducts	10000.0	1.530200e+00	0.581654	1.00	1.00	1.000000e+00	2.000000e+00	4.00
HasCrCard	10000.0	7.055000e-01	0.455840	0.00	0.00	1.000000e+00	1.000000e+00	1.00
IsActiveMember	10000.0	5.151000e-01	0.499797	0.00	0.00	1.000000e+00	1.000000e+00	1.00
EstimatedSalary	10000.0	1.000902e+05	57510.492818	11.58	51002.11	1.001939e+05	1.493882e+05	199992.48
Exited	10000.0	2.037000e-01	0.402769	0.00	0.00	0.000000e+00	0.000000e+00	1.00

Figure 2 - Dataset description.

3. METHODOLOGY

As the problem is a classification problem, whether or not the customer will leave the bank, the following models were considered:

- Logistic Regression
- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Kernel SVM
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification

Each of them will be applied to the data set (75% for training and 25% for testing), so that the precision and processing speed of each one of them was evaluated to choose the one that best solves the problem.

4. ANALYSIS

For the analysis, the data were initially pre-processed. Coding the client's gender and country so that they are numeric and could be used in tests. A feature scaling was made, and after that, it was divided into training and test data.

The analysis was made in the sequence showed in the previous section: logistic regression, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Kernel SVM, Naive Bayes, Decision Tree Classification and Random Forest Classification.

5. RESULTS AND DISCUSSION

Assuming the classification algorithms, it was noticed that they have an accuracy between 0.7964 and 0.8624, which is not a bad value for the proposed problem. In this way, regardless of the algorithm used, the bank would have a good hit rate to direct measures that convince customers to stay at the branch, which would reduce spending if these strategies were used for all customers.

6. CONCLUSION

Finally, the prediction will be indicated using the SVM Kernel algorithm, which obtained an accuracy of 86.24%. The rate of false positives was 12.92% and false negatives 20.43%, which was the lowest among the tested algorithms.