CUSTOMER CHURN PREDICTION IN BANKING INDUSTRY

TABLE OF CONTENTS:

1. BACKGROUND OF THE PROBLEM:

Customer Churn has become a major issue in banking industry since it leads to the loss in the revenue generated by the bank. In current scenario, acquiring new customers has become hard and expensive since there are many different banking offering a variety of offers. So they say that "**The best way to avoid customer churn is to know your customers, and the best way to know your customer is through historical and new customer data"**. Thus, customer churn has become one of the biggest challenges in any industry. So, reducing the customer churn became the primary business goal and it needs to be solved as soon as we can.

With the help of machine learning models, we can resolve the banking problems by finding the regularities. In this project, we are going to use Support Vector Machine model and Logistic Regression model to train our model and evaluate it since our predictor variable is categorical (Dependent variable – Exited).

2. MOTIVATION FOR SOLVING THE PROBLEM:

Customer Churn is defined as the percentage of customers that has stopped using a particular company's product or service during a certain time frame and transitioned to another company. Churn rate can be calculated by dividing the number of customers lost by number of customers that we had during the beginning. The Customer Churn control is most important for any organization because the cost required to attract new customers is equal to 6 to 7 times greater than the cost required to retain the existing customers.

By improving the customer service based on the feedback obtained from the customers of the bank ensures that we are providing the necessary inputs from our side. According to Customer Success Consultant Lincoln Murphy, a good month-to-month churn rate is anywhere between 0.42 and 0.58 percent. With those figures, a yearly rate of 5% to 7%

would be appropriate. Thus, if your month-to-month or annual churn rates exceed these figures, it's a clear indication that steps should be taken to improve both the value and customer service you provide to your members.

3. SOLUTION METHODOLOGY AND EVALUATION METRICS:

The primary goal of the project is to reduce the customer churn, stabilize the banking business and increase the profits in the banking industry. Researchers began to identify elements that contributed to client attrition. The underlying popular machine learning methods naturally showed these patterns. Furthermore, a customer turnover experiment found high-value clients who were potentially dangerous. Proactive measures taken at regular periods help to ensure that such key consumers are retained before they leave. The steps involved to reduce the customer churn rate is as follows:

- Understanding the problem.
- Get the data and process the data by splitting the data to 80% trained and 20% test data.
- Choose the best algorithm that suits the data.
- Evaluating the model and analyzing the results.

4. DESCRIPTION OF DATA SET:

Column	Variable Type	Description
RowNumber	**Continuous**	**Serial Number of customer**
CustomerId	**Continuous**	**ID of customer**
Surname	**Categorical**	**Customer LastName**
CreditScore	**Continuous**	**Credit Score of customer**
Geography	**Categorical**	**Geography of customer**
Gender	**Categorical**	**Gender of customer**
Age	**Continuous**	**Age of customer**
Tenure	**Continuous**	**Tenure of customer in Bank**

Balance	**Continuous**	**Balance of customer**
NumOfProducts	**Categorical**	**Products owned by customer**
HasCrCard	**Categorical**	**if customer owns credit cards**
IsActiveMember	**Categorical**	**Is Active Member of Bank**
EstimatedSalary	**Continuous**	**Salary of customer**
Exited	**Categorical**	**If customer has exited the bank**

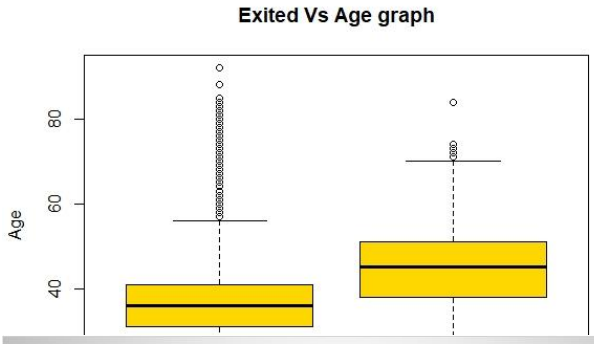
DATASET SAMPLE:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
2	1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
3	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
4	3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
5	4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
6	5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0
7	6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.71	1
8	7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
9	8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.88	1
10	9	15792365	He	501	France	Male	44	4	142051.1	2	0	1	74940.5	0
11	10	15592389	H?	684	France	Male	27	2	134603.9	1	1	1	71725.73	0
12	11	15767821	Bearce	528	France	Male	31	6	102016.7	2	0	0	80181.12	0

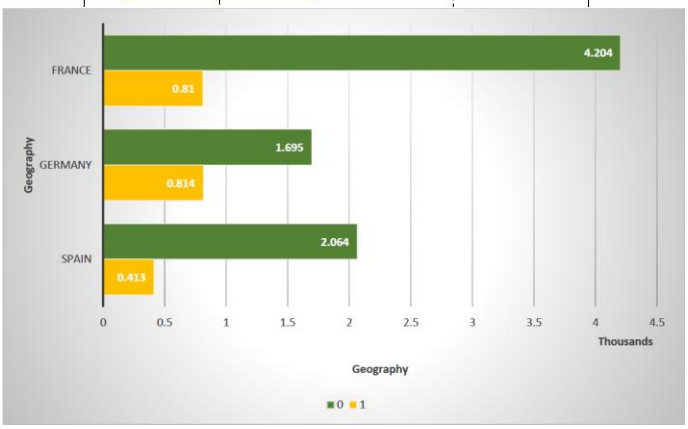
The above picture shown is our dataset, which consists of 10000 rows and 14 columns. The rows consists of the customer details like CustomerID, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumberofProducts, Hascrcard, Isactivemember, EstimatedSalary, Exited.

We are going to predict the customer churn using ‘Exited’ as our dependant variable. We are not considering the columns rownumber, customerid and surname while training the model since they are not predictor variables. Since the ‘Exited’ variable is categorical, we are going to train and evaluate the model using Support Vector Machine and Logistic Regression models.

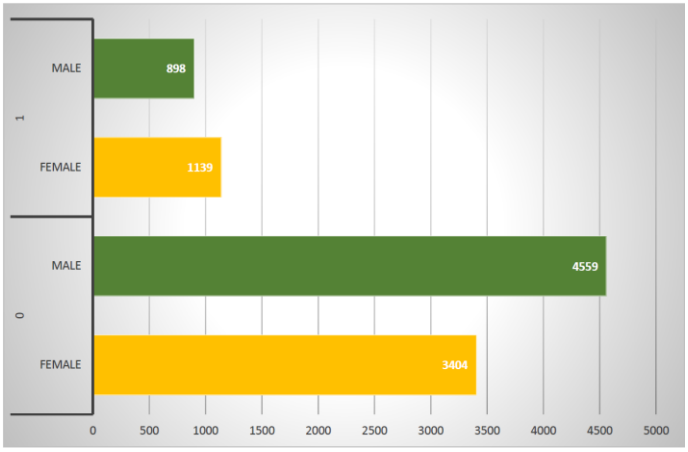
We have used RSTUDIO and MS-EXCEL for data visualization and displaying the results. We have found out the customer churn rate based on gender (male-female ratio), Age group and Geography of the customers.



Age wise exit status for the three regions.



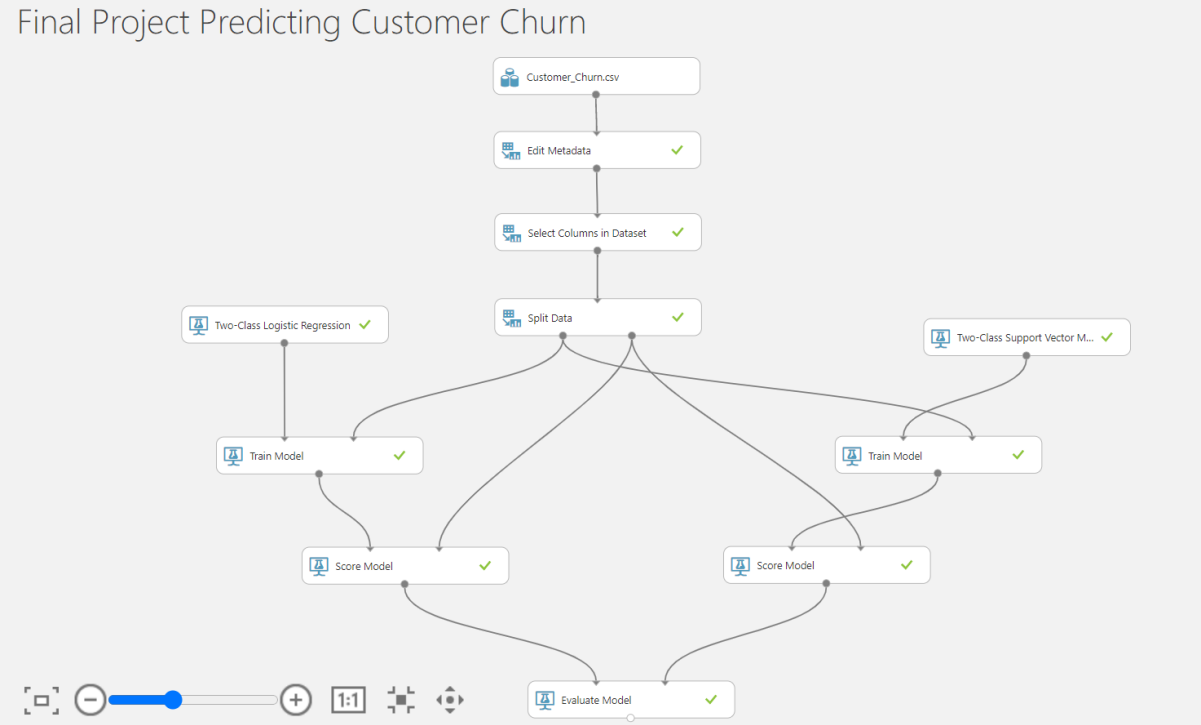
Geography wise exit counts for the three regions.



Gender wise exit counts for the three regions.

5. COMPARISION OF TWO ALGORITHMS:

- A two-class support vector and two-class logistic regression models were used to predict the y variable (banking customer churn).
- The following steps were used to clean the data for the modelling:
  - Data transformations were not necessary as the data intended was from a single source and had all the top predictors for the customer churn.
  - Using “Edit Metadata” editor, columns, Geography, Gender, HasCrCard, IsActiveMember, NumOfProducts have been converted to catogorical
  - Using “Column Selector” , RowNumber, CustomerId and Surname were excluded from the dataset as they are ID’s of the unique entries.

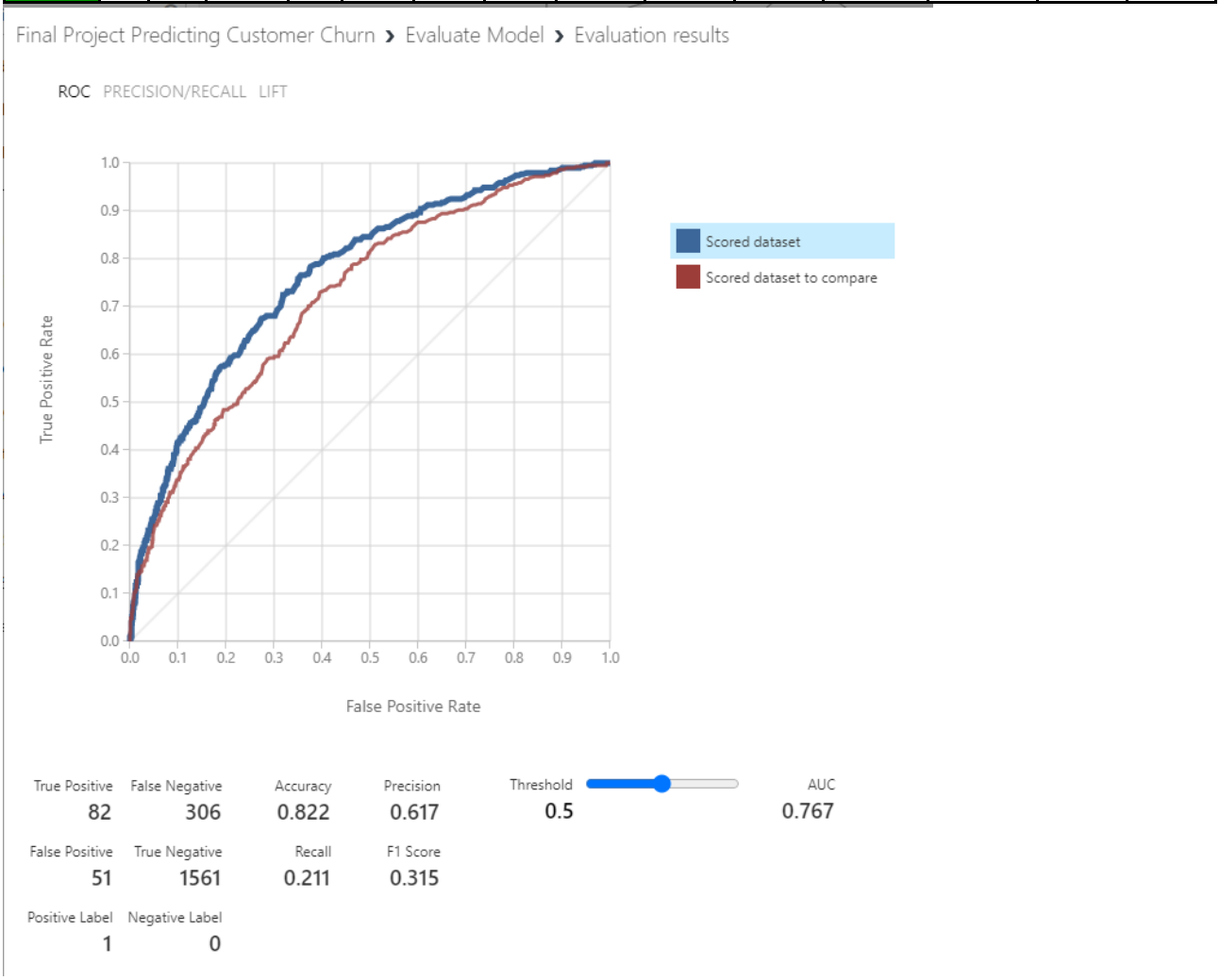


Two-Class Support Vector Machine		
Training Experiment Number	1	2
Number of iterations	1	2
Lambda	0.001	0.001
Normalize features	Yes	Yes
Random number seed	54902	54902

Two-Class Logistic Regression		
Training Experiment	1	2
Optimization tolerance	1.00E-07	1.00E-07
L1 regularization weight	1	3
L2 regularization weight	1	2
Memory size for L-BFGS	20	20
Random number seed	54902	54902

6. SUMMARY SHEET SHOWING THE RESULTS OF ALL EXPERIMENTS:

Algorithm	Exp No	TP	TN	FP	FN	TP+TN	Total = TP+TN+FP+FN	Accuracy	Misclassification Rate	Precision	Recall	Specificity	F1 Score	AUC
Logistic Regression	Exp 1	82	1561	51	306	1643	2000	0.822	0.1785	0.617	0.211	0.968362283	0.315	0.767
	Exp 2	70	1573	39	318	1643	2000	0.822	0.1785	0.642	0.18	0.975806452	0.282	0.763
Two Class Support Vector	Exp 1	62	1565	47	326	1627	2000	0.814	0.1865	0.569	0.16	0.970843672	0.249	0.723
	Exp 2	66	1563	49	322	1629	2000	0.815	0.1855	0.574	0.17	0.969602978	0.262	0.743



Considering both models, we can see that Two Class Logistic Regression model has the highest number of true positives and true negatives, and accuracy of 0.767 and 0.763 which is best suited for the prediction of the customer churn in banking industry. From the above graph, we can also conclude that BLUE coloured curve (Logistic Regression) is more ideal than RED coloured curve (Support Vector Machine) based on the area under curve (AUC).

7. CONCLUSIONS:

In the banking sector, the deciding factors for retaining customers are the services, schemes, brand value, hidden charges, customer satisfaction. Based on the data model we used these are the findings and our recommendation.

- The two-class logistic regression model has a greater accuracy at 82% and is the recommended model for use by the banks.

- The Brand value and schemes need to be upgraded in Germany as the churn rate is the highest among the three countries.

- Age was one of the significant contributing factors for churn of the customers. Customers in the age group of 40-50 years are mostly churning. This group might have investment requirements that are higher than the younger generations for retirement purposes and hence they are switching to other banks with better policies that benefit them. So, we recommend introducing new schemes or policies for customers with age group of 40-50 like family schemes and updated retirement policies or better rates of interest or investment plans.

- Banks should take continuous feedbacks on their policies, reward programs and loan interest rates from the customers on a regular basis.

- Gender wise we can see; more women have churned compared to men. So, banks need to have better plans designed for women to persuade them to stay with the bank. This could be in the form of cash rewards, gift cards, women savings account with better interest rates to name a few.

- We can see the customers who have more than 2 products from a bank have largely churned. So, the banks need to look at these specific customers and decide what products to design for their needs.