Customer churn prediction in Banking Industry using Machine Learning.

Data Mining| Prof. Dr. Kiran Garimella

"To lose a customer and replace it is very expensive. It's critical that we have low churn... The investment we make in having a live person pays for itself in the long term."

ANUSHA SALIAN, ASHOK KUMAR KOLLI, MEGHLA SARKAR, SHREYA UDAY, SRI VENKATA VENU GOPAL GUDDATI, UMAKANTH SAI BALGURI

Table of Contents

1. Background of problem

Customer Churn is a huge problem for banking industry as it contributes to a reduction in the revenue. In current situation, acquiring customers is hard and expensive. Customer acquisition costs will further put a mark to the overall revenue of the company.

As part of our observation, we can see more issues and challenges are being reported. As a result, customer churn has become one of the top challenges for most of the banks. In banking sector **reducing customer churn is a key business strategy** and it needs to be addressed with urgency. Machine learning can help to resolve banking problems by finding some regularity, causality, and correlation to business information.

So, we will dive deeper in this paper that lays out the need, problem description, comparison of two algorithm followed by experiment, results and, benefits for addressing. Furthermore, the bank wants to understand the influencing factors for churn so they can be more proactive towards addressing such challenges by analysing customer behaviour. The experimentation was conducted on the churn modelling dataset from Kaggle. In this paper, we will predict the customer churn using 'exited' as our dependent variable. Random Forest, Logistic Regression and Support Vector Machine models are used to train our model as the predictor variable is categorical.

2. Motivation for Solving the Problem

- Churn is characterized as the movement of client starting from one organization to the next. This is extremely important for the bank because, the total cost of attracting new set of customers could be five to six folds more than retaining the existing customers in the bank.
- Usually in the banking industry, the longstanding customers tend to become less expensive to serve. These customers spawn very high profits, and they also could potentially provide new referrals. Loss of a customer will pave a way for loss in profits for the bank. "The utilization of client information or the feedback to gauge the probability of a client or a group of customers ending their membership later on" is the very definition of Customer Churn Forecast by SaaS organization. Being able to precisely foresee the future churn rates are fundamental since it helps any business acquire a superior comprehension of future anticipated income.
- **Use Cases:** You can utilize churn prediction in a wide range of approaches to improve your business. To make it work, nonetheless, you'll need to focus in first on improving three principal regions:

  - Customer/client outreach
  - Client service / customer care
  - Customer value/worth

- Improving the service provided to the customers is significant on the grounds that it guarantees that we are producing the appropriate feedback to precisely anticipate the churn rates. Also, when planned effectively, you can utilize your effort to decide precisely which subscribers recognize as doubters and passives. What's more, from that point, you can do whatever it takes to run after transforming those subscribers to promoters.
- As indicated by Customer Success Consultant Lincoln Murphy, an adequate month to month churn rate would be some place in the 0.42 – 0.58% territory. Those numbers would mean 5% – 7% being a worthy yearly rate. Thus, in case you're finding that your month to month or yearly churn rates are surpassing these numbers, that is an obvious sign that means should be taken to improve both the worth and client support you're giving to the subscribers.

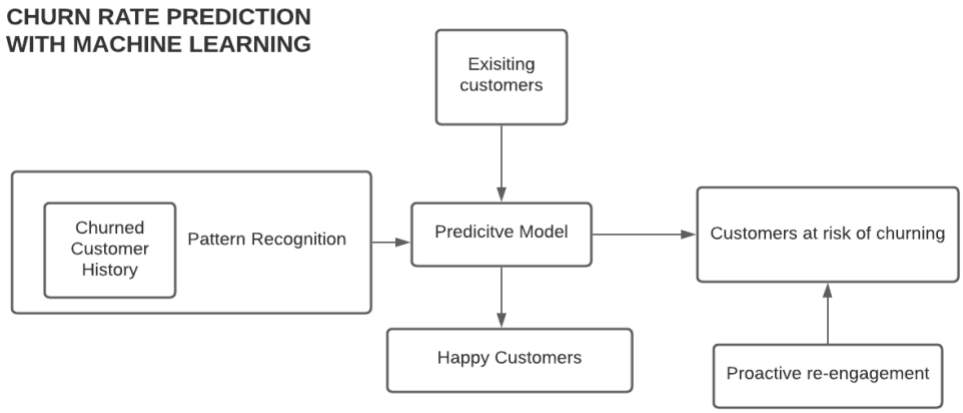
3. Solution methodology and evaluation metrics.

The main purpose is to reduce customer churn, stabilize the business and increase profits in banking sector. Researchers started discovering characteristics that caused customer churn. These patterns were revealed automatically by the underlying popular machine learning

algorithms. Furthermore, customer churn experiment identified high-worth risky customers. Proactive measure at regular intervals helps to ensure that they can retain such valuable customers before they leave.

The general extent of work the data researchers complete to assemble ML-based frameworks able to conjecture customer churn rate down may resemble the following:

1. Understanding an issue and the final objective
2. Information collection: Finding a relevant dataset pertaining to the problem.
3. Data planning and pre-processing
4. Data Modelling and testing
5. Deployment of the model and observing the results.



4. Description of Datasets:

4.1 Variables Table:

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

Sample Dataset:

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.88	1

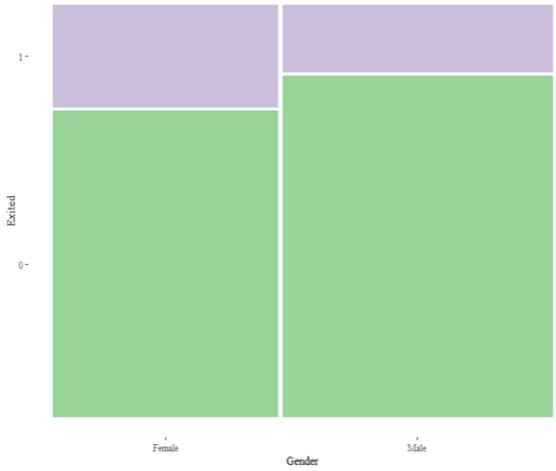
We have 14 columns in the dataset, row number captures serial number of customers in the dataset. For every customer we capture customerId, surname, credit score, geography, gender, age, tenure, balance (account balance), number of products, hascrcard (if customer owns credit card), isactivemember (if customer is active member of the bank), Estimated Salary(salary of customer), Exited(if customer exited the bank or not).

We will predict the customer churn using ‘exited’ as our dependent variable. Columns like rownumber, customerId, surname are excluded while training the model since they are the id’s for each column and not predictors. As the predictor variable is a categorical, we plan to train our model using Random Forest, Logistic Regression and Support Vector Machine models.

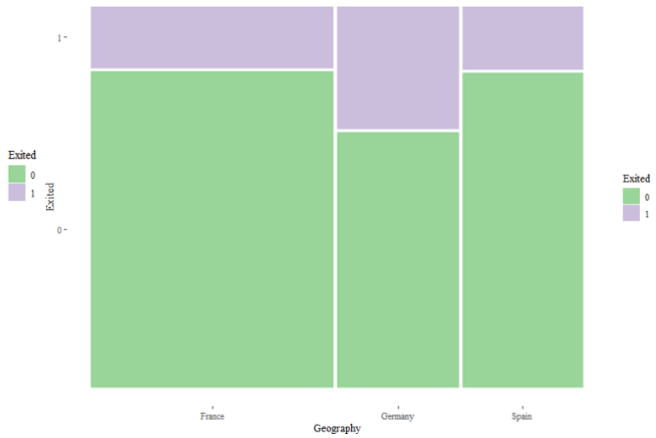
4.2 Exploratory Data Analysis:

We used Jupyter Notebook and python libraries(Scikit learn) for data visualization. Firstly, we found out impact of customer churn out on gender and found out the ratio of churn out was higher for females over males.

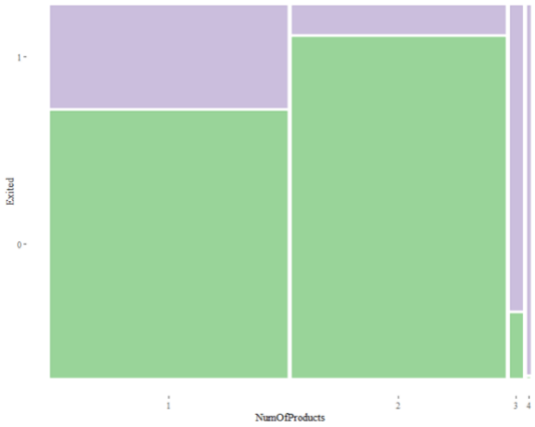
Gender



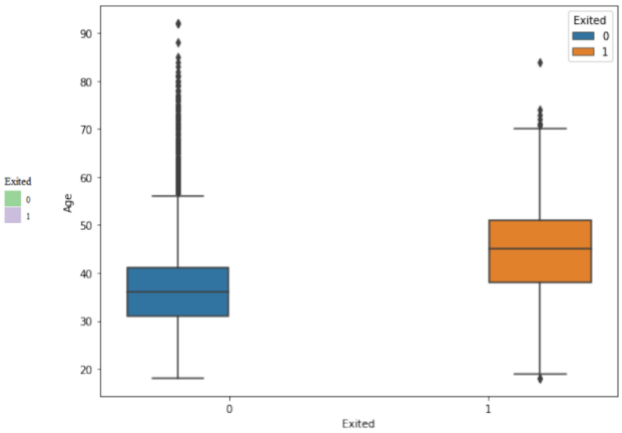
Geography



No of Products



Age



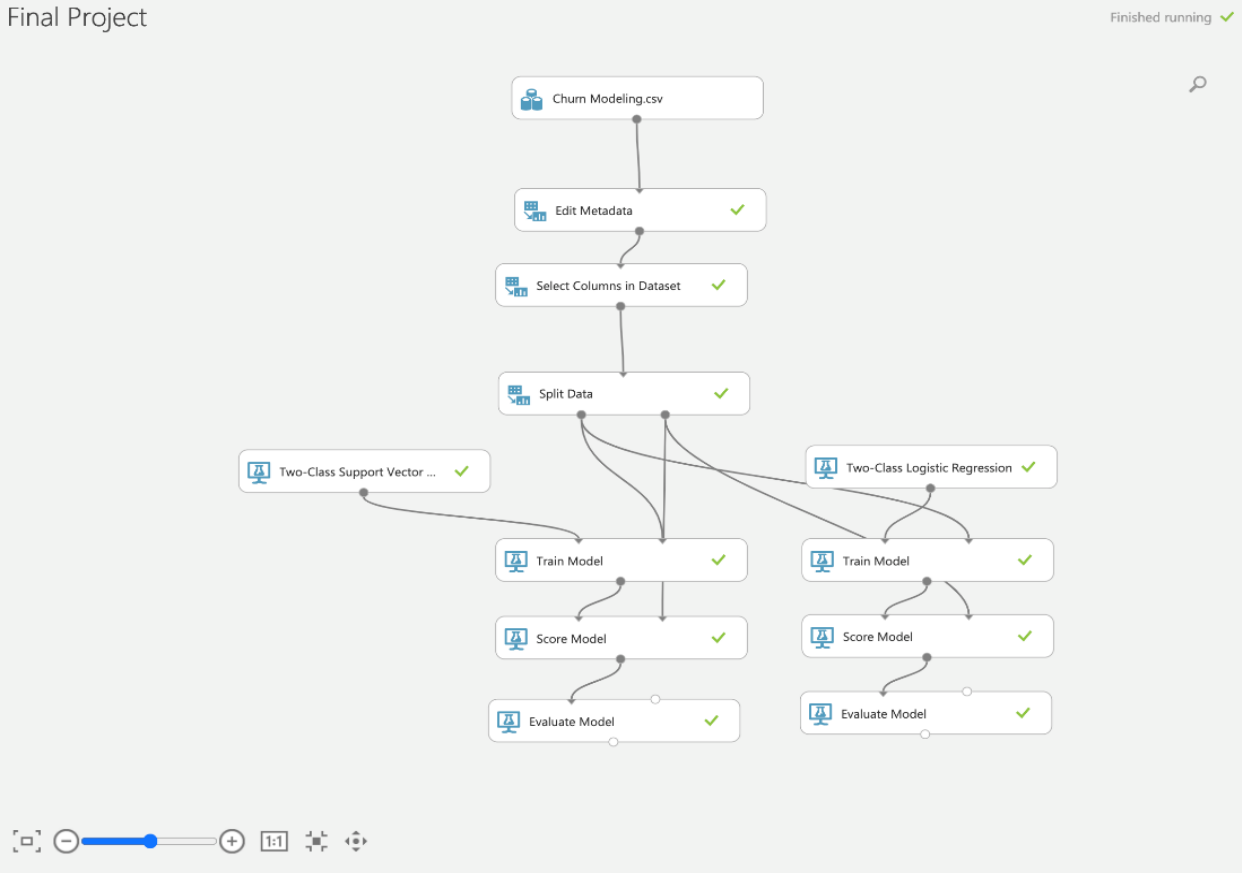
5. Comparison of two algorithms & followed by experiments.

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:

1. Data transformations were not necessary as the data intended was from a single source and had all the top predictors for the customer churn.
2. Using “Edit Metadata” editor, coloums, Geography, Gender, HasCrCard, IsActiveMember, NumOfProducts have been converted to catogerical
3. Using “Coloum Selector” , RowNumber, CustomerId and Surname were excluded from the dataset as they are ID’s of the unique entries.

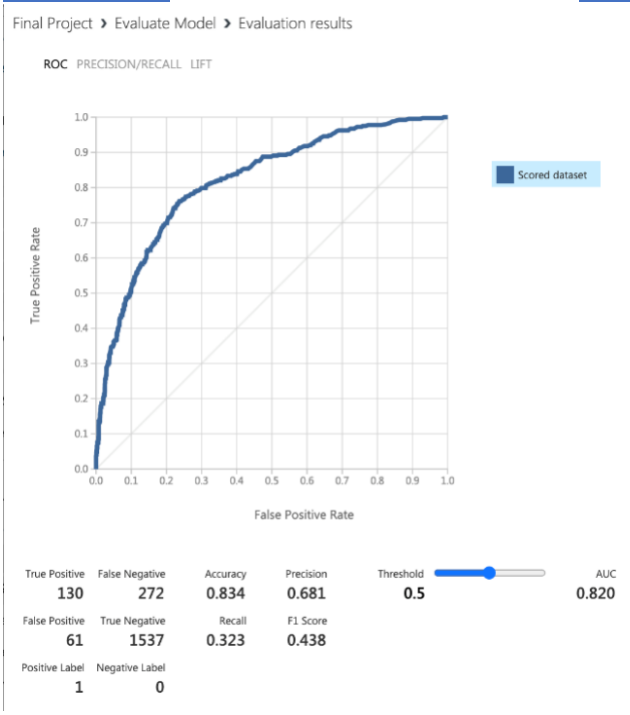
4. Using Split Data, 80 percent of the data has been split to train the model, 20% of the data to test the model. Seed number, 6136 has been used to obtain controlled results each time.



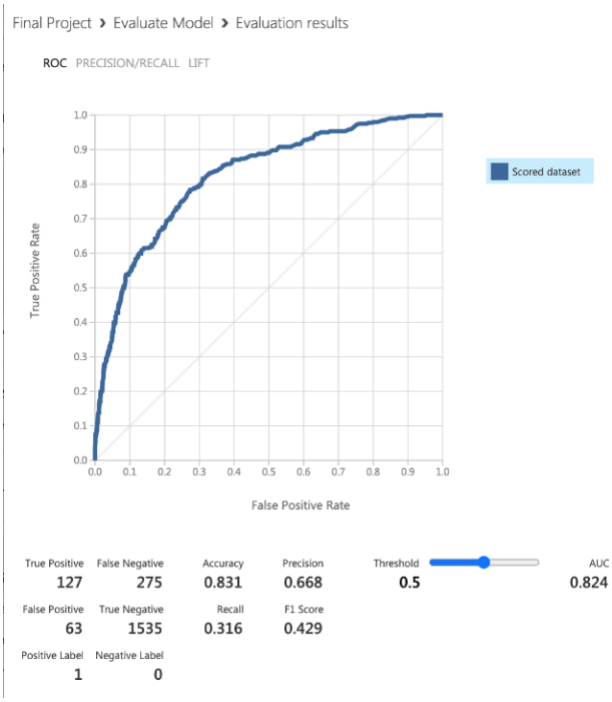
Two-Class Support Vector		
Training Experiment Number	**1**	**2**
Number of iterations	1	2
Lambda	0.001	0.001
Normalize features	Yes	Yes
Random number seed	6136	6136

Results:

Experiment 1:

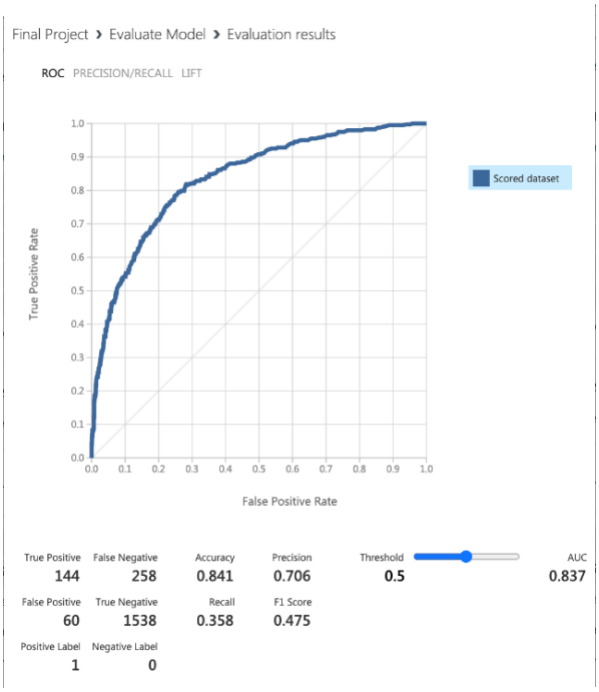


Experiment 2:

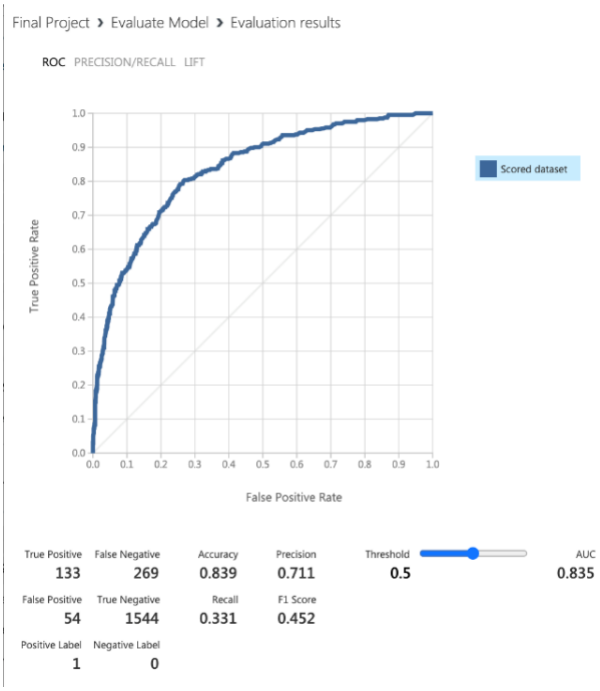


Two-Class Logistic Regression		
Training Experiment	**1**	**2**
Optimization tolerance	1E-07	1E-07
L1 regularization weight	1	3
L2 regularization weight	1	4
Memory size for L-BFGS	20	20
Random number seed	6136	6136

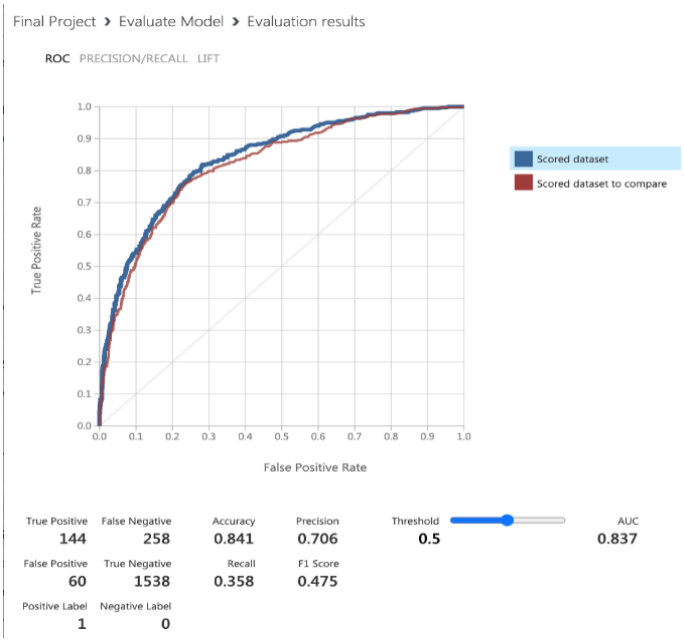
Experiment 1:



Experiment 2:



6. Model Summary:



Algorithm	Exp No	TP	TN	FP	FN	TP+TN	Total=TP+TN+FP+FN	Accuracy	Misclassification Rate	Precision	Recall	Specificity	F1 Score	AUC
Two-class Support Vector	Expirement 1	130	1537	61	272	1667	2000	0.8335	0.1665	0.681	0.323	0.962	0.438	0.820
	Expirement 2	127	1535	63	275	1662	2000	0.831	0.169	0.668	0.316	0.961	0.429	0.824
Two-class Logistic Regression	Expirement 1	144	1538	60	258	1682	2000	0.841	0.159	0.706	0.358	0.962	0.475	0.837
	Expirement 2	133	1544	54	269	1677	2000	0.8385	0.1615	0.711	0.331	0.966	0.452	0.835

In both the models, two class logistic regression has highest true positives, true negatives, and accuracy score of 0.835, AUC of 0.837 making is best suitable for predicting the churn of customers. The blue curve of two class logistic regression more ideal than the red curve, the two-class support vector.

7. Conclusion and further recommendations:

In Banking sector, the stimulating factors for retained customers are customer satisfaction, services and schemes, brand value and transparency of the banks. After exploring the data and modeling, here are our findings and recommendations that will refrain the customers from churning out:

- As our logistic model has 83% accuracy, we recommend the bank to use this model to predict if the customer churns or not.
- Increasing the brand value in Germany is crucial for this bank as the churning rate of the customers is high compared to other two countries.
- Age was one of the significant contributing factors for churn of the customers; Customers with age around 40-50 are mostly churning. This particular age group might have additional responsibilities of their families or retirement purpose and hence they are switching to other banks with better policies which 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.
- The bank has very few customers who have opted for >2 products. Of those, most are likely to churn out, so we recommend banks to investigate the product categories, in terms of the benefits to customer.
- In general, banks should keep reviewing and taking continuous feedbacks on its policies, reward programs and loan interest rates from the customers.