Final Project Proposal (DS105)

Problem Statement:

The project objective is to develop churn prediction model that how likely its current customers will be leaving the bank in near future.

The problem here is a classification problem to classify a customer whether he or she will exit or not based on his or her credit score, region, gender, age, tenure, balance, estimated salary and etc. We will be dealing with binary classification by using 3 algorithms like logistic regression, decision trees and SVM to predict the test and select the best model that able to provide the better accuracy.

Introduction of dataset:

Dataset: Churn_Modelling.csv

The bank decided to collect data from 6 months period to evaluate the problem after noticing increase in the number of customers leaving the bank. This data set contains of 10,000 customers were selected randomly among three countries – France, Germany and Spain. The dataset is well-labelled to explain all its columns and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed account) or continues to be a customer.

There are 14 attributes of the dataset have as per below:

1. RowNumber: corresponds to the record (row) number
2. CustomerId: The customer ID created by the bank
3. Surname: The customer surname
4. CreditScore: The customer credit score`
5. Geography: The country of the customer (Germany/France/Spain)
6. Gender: The gender of the customer (Female/Male)
7. Age: The age of the customer
8. Tenure: The customer's number of years in the bank
9. Balance: The customer's account balance
10. NumOfProducts: The number of bank products that the customer uses
11. HasCrCard: Does the customer has a credit card? (0=No,1=Yes)
12. IsActiveMember: Does the customer has an active membership (0=No,1=Yes)
13. EstimatedSalary: The estimated salary of the customer
14. Exited: Churned or not? (0=No,1=Yes)

The dataset was obtained from the website of Kaggle as per the link below.

https://www.kaggle.com/santoshd3/bank-customers

Sample of the dataset

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedS
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	1013
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	1125
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	1139
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	938
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	790
...	...	...	...	...	...	...	...	...	...	...	...	...	
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	962
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	1016
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	420
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	928
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	381

10000 rows × 14 columns

Challenge or Difficulties:

The challenges that were anticipated during the preparation of the data for modelling which is to identify outliers or unbalanced. We are to decide which unimportant features to drop and the selection of encoding method to prevent curse of dimensionality.

Goals

Some questions that can be lead to the goal from this dataset:

1. Which feature set is suitable, so to achieve predictive accuracy and to avoid the curse of dimensionality?
2. Identify groups of customers who share the same characteristics using different supervised learning techniques.
3. Within each homogeneous group, predict which customers are most likely to churn.
4. Evaluate different combinations of supervised learning techniques with respect to the accuracy of predicting churn.
5. Which encoding method suitable for categorical features.
6. Do we need to do scaling for the dataset?
7. Which performance metrics suitable for our model?

Hope this prediction is able to help the bank on the steps of improve the retention rate and have better planning of their budget and resources for the likelihood of an existing customer leaving before they do so to increase profitability.