Capstone Report

Capstone Project 1

Churn Analysis of Bank Customers

Vaibhav Agarwal
September 26, 2019

Contents

1. Introduction
2. Data Cleaning and Manipulation
3. Data Exploration & Hypothesis testing
4. Modeling
    A. Data Pre-processing
    B. Evaluation Metrics
    C. Logistic Regression
    D. Gaussian Naive Bayes
    E. k-Neighbors Classifier
    F. Support Vector Machine
    G. Random Forest
    H. Extreme Gradient Boosting
5. Assumptions and Limitations
6. Conclusion

1. Introduction:

The project is churn analysis on bank customers. The project is to try to analyze and understand the customers who will churn.
The client is a Bank. This project will help them as the bank will be able to distinguish between customers who will retain or will not retain or who are uncertain. This distinction will help them to specifically target customers. Ex- The bank will be able to specifically target the uncertain customers to make them certain.
The data used in this project is taken from kaggle. Here we have a dataset of a hypothetical bank containing certain details of 10k customers who are active, non-active or left the bank. The dataset is divided into two parts, one containing customer personal details and other customer bank account related details.
The data includes but not limited to following rows:

- Age
- Gender
- Credit score
- Country
- Exited

2. Data Cleaning and Manipulation

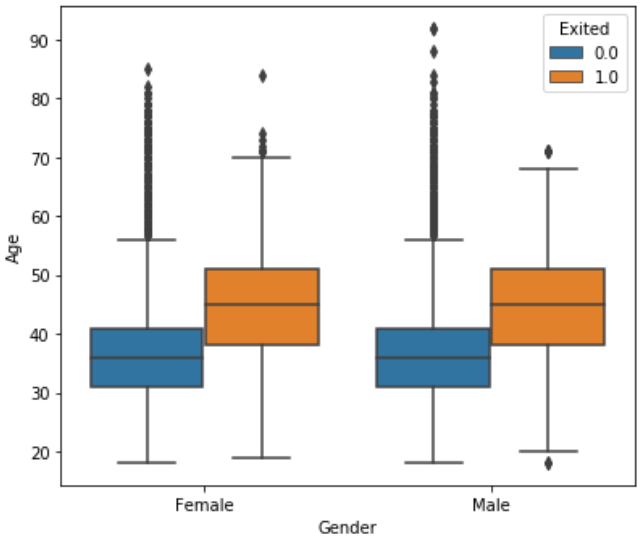
The steps taken to clean the data and prepare for data exploration are below.

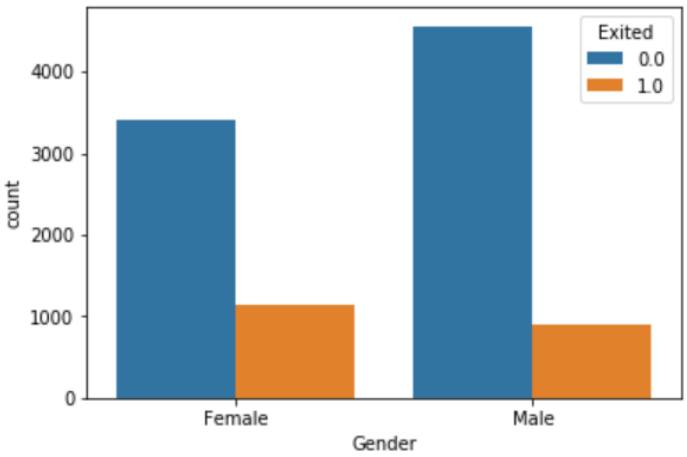
1. The two CSV files are imported in jupyter notebook using pandas.
2. The file's data are then converted into pandas dataframes and checked for head and info.
3. The info suggests that there are NaN values. Also, the data type of columns suggests that there are no string values where numbers are present.
4. The two data files contain one similar column *CustomerId* on which the merge has to be made. The *CustomerId* column of both files is checked if they have all unique values.
5. Both the dataframes are then *outer* merged on *CustomerId*.
6. The index has been set to column *CustomerId*.
7. The info is checked for the merged dataframe. There is the same number of rows as before. So, no extra *CustomerId* in any file.
8. Now, missing data is imputed. The missing values in Surname are filled with *noname* using fillna() function. The missing values in *CreditScore* are filled with the mean value.
9. To fill the Geography and other critical columns, the data is analyzed for patterns.
10. The missing values are filled accordingly.
11. Other missing values are filled similarly. Please check the code for details.
12. To understand more about the values in the columns ‘describe()’ is used.
13. Checking the min and max of columns gives an idea if the values are in the desired range and that there are no negative values. Columns *CreditScore, Age* is fine.
14. Other columns like those having few different values like Gender are checked using value_counts function that they don’t have any irregular values.

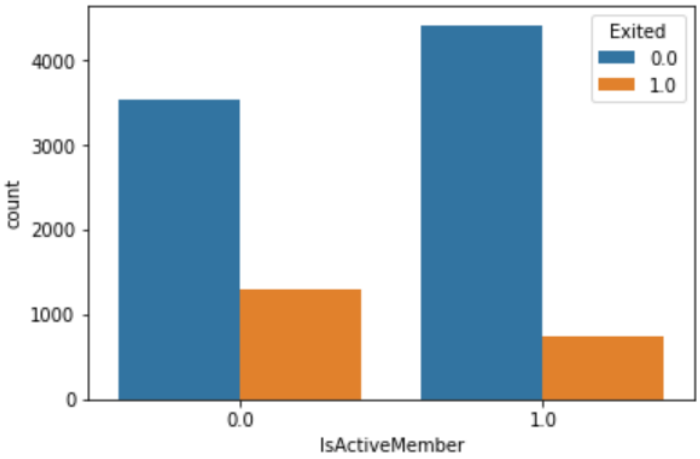
3. Data Exploration & Hypothesis testing

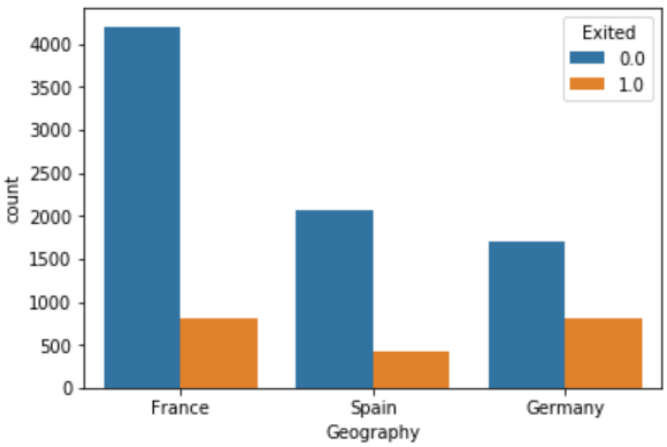
Our main concern is to find factors that lead to customer churning. So there are few questions that should be asked before stating Exploratory Data Analysis and I will answer them at the end.
Questions:

1. What age group are the most leaving the bank?
2. Was it male or female for a particular age group who churned the most?
3. How does the activeness of members in bank affect the churning?
4. Most customers left of which country?
5. Anything different in range or values of other variables that affect churning?









We can now conclude our findings:

1. The median age of people whom the bank retained is 37 with most values between 30-40. The median age of people who exited the bank is 45 with most values between 40-50 and is

normally distributed. To improve on this the bank has to provide special services to senior citizens.

2. It is female who left the bank more than the male. The percentage of the female who left the bank is 25%, whereas for males it is 16%. We can conclude that the bank does not make females comfortable. The bank will have to provide special female services.
3. The inactive members churned more with 27%. Only 14% of active members churned. But active members churning shows that customers must have been unsatisfied for any new policy or some other bank provided a scheme with much better services
4. Most churning of customers happened in Germany with almost 32% of customers leaving the bank. Customers of Germany have the most salary and Balance. This means the bank is not able to cater to the elite class with elite services.
5. In Spain around half of the females have a balance of zero.

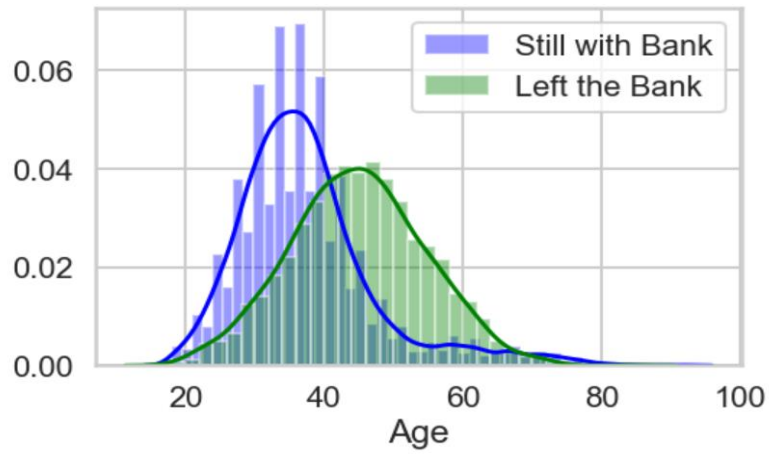
The hypothesis to be tested:

1. Null: Age of people who left the bank and who did not are similar. Alternative: Not similar.
2. Null: Credit score of people who left the bank and who did not are similar. Alternative: Not similar.
3. Null: Balance of people who left the bank and who did not are similar. Alternative: Not similar.
4. Null: Estimated Salary of people who left the bank and who did not are similar. Alternative: Not similar.

The variables have been plotted against each other for every hypothesis and then the frequentist test is applied to them. In some cases, the bootstrap test is also applied.

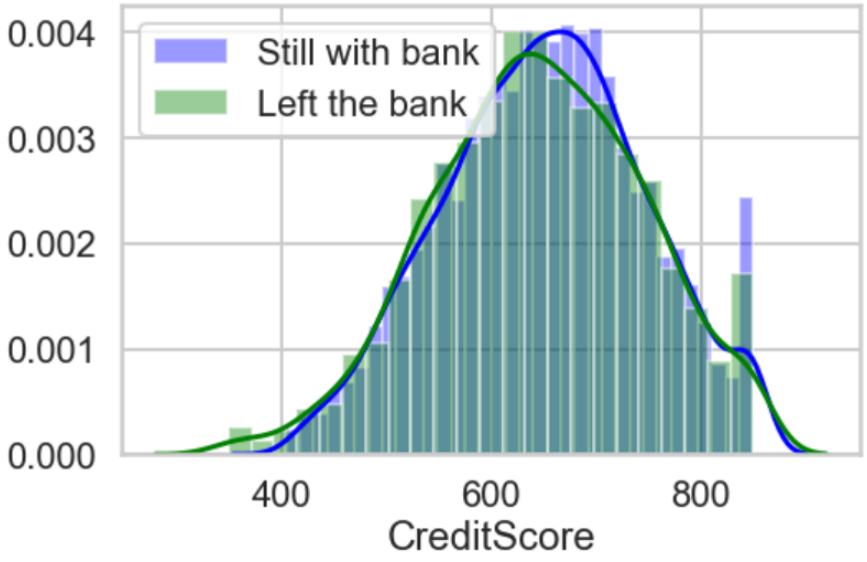
Let's get into each hypothesis-

Hypothesis 1: Age



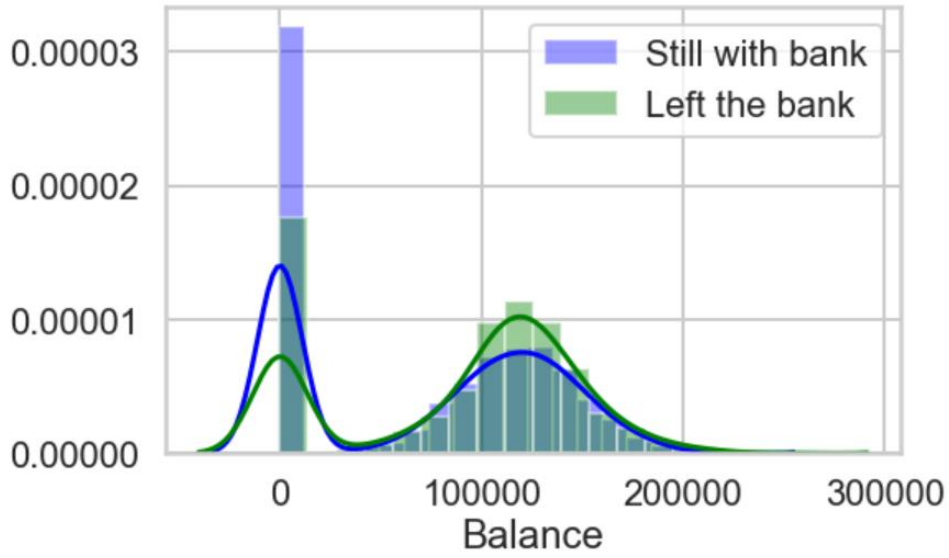
We reject the Null hypothesis. The probability of the null hypothesis is almost zero which is less than the significance level of 0.05.

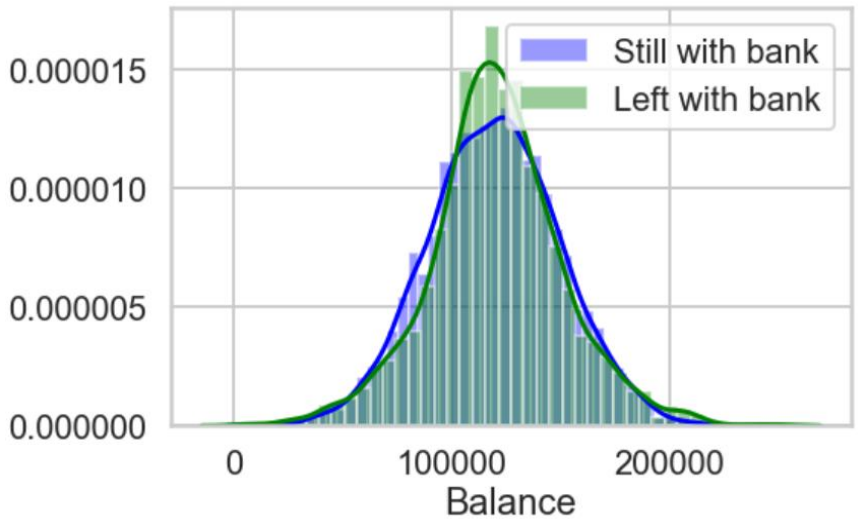
Hypothesis 2: Credit Score



We reject the Null hypothesis. The probability of the null hypothesis is 0.0085 or 0.85 % which is less than the significance level of 0.05.

Hypothesis 3: Balance





The Balances of Zero are too many in the data. When we consider all the data, we reject the Null Hypothesis. When we only remove the Balances which are Zero, the probability of the null hypothesis becomes 19.06% which is significant. Then we reject Alternative Hypothesis.

Hypothesis 4: Estimated Salary

We do not reject the Null hypothesis. The probability of the null hypothesis using the t-test is 0.2416 or 24.16% and using bootstrapping is 0.1222 or 12.22% which is more than the significance level of 0.05.

Conclusion of Hypothesis testing

The variables CreditScore and Age will be most helpful in predicting churning. The variable Balance will also be helpful only in cases where Balance is zero.

4. Modeling

A. Data Pre-processing:

Before feeding the data into any machine learning algorithms, there are some preprocessing steps that must be performed on the data. We outline these steps below.

1. **Dropping insignificant features:** The insignificant features which were understood during storytelling and statistics are dropped. Other features such as name or id which do not provide insights are dropped.
2. **Label encoding:** In the dataset, there are some variables with numerical values, some variables with categories and some variables with binary values (0 and 1). For numerical and binary variables, we do not worry about labeling. However, we perform label encoding for the categorical variables. This step is carried out on the whole dataset.
3. **Scaling:** For some algorithms, it is necessary that we scale the values of all features to lie within a fixed range. We scale features such that all features have values balanced around 0.
4. **Data splitting:** The second step involves splitting the label encoded dataset into train and test datasets. In this project, we split them equally with a 70%-30% ratio. Also, we split them in such a manner that the fractions of both classes remain almost the same in train and test datasets.

B. Evaluation Metrics:

Once the data is pre-processed, we feed them to a classification algorithm to build the model. In order to evaluate the performance of the model, we test the model on the test dataset. One more important consideration while performing cross-validation is the selection of a proper evaluation metric. Especially, for imbalanced data, it is important to be careful about the choice of the evaluation metric. Our aim is to predict the churning likelihood with data containing only about 20.36% of the churned customers. Accuracy is not a good metric for such datasets. We definitely want to have a high true positive rate (or recall) with canceled flight tagged as the positive class. At the same time, we do not want lots of false positives or less precision. Most of the time, the choice of a good metric depends on business needs. In this project, we keep in mind all elements of a confusion matrix.

C. Logistic Regression:

We train the training data from the splitted data with Logistic Regression.
Tuned hyperparameters :'C': 1.0, 'penalty': 'l1'
Accuracy : 0.84
Best Estimator : LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='warn', n_jobs=None, penalty='l1', random_state=None, solver='warn', tol=0.0001, verbose=0, warm_start=False)
Roc Auc Score: 0.8407
Classification Report:

	precision	recall	f1-score	support
-1	0.85	0.96	0.90	2363
1	0.72	0.37	0.49	637
micro avg	0.84	0.84	0.84	3000
macro avg	0.78	0.66	0.69	3000
weighted avg	0.82	0.84	0.81	3000

D. Gaussian Naive Bayes:

We train the training data from the splitted data with Gaussian Naive Bayes.
Accuracy : 0.8226
Roc Auc Score: 0.8137
Classification Report:

	precision	recall	f1-score	support
-1	0.83	0.97	0.90	2363
1	0.71	0.28	0.40	637
micro avg	0.82	0.82	0.82	3000
macro avg	0.77	0.62	0.65	3000
weighted avg	0.81	0.82	0.79	3000

E. kNN:

We train the training data from the splitted data with the k-Neighbors Classifier.
Tuned hyperparameters : 'n_neighbors': 17
Accuracy : 0.8528
Best Estimator : KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=17, p=2, weights='uniform')
Roc Auc Score: 0.8341
Classification Report:

	precision	recall	f1-score	support
-1	0.86	0.97	0.91	2363
1	0.76	0.39	0.52	637
micro avg	0.85	0.85	0.85	3000
macro avg	0.81	0.68	0.71	3000
weighted avg	0.84	0.85	0.83	3000

F. Support Vector Machine:

I. SVM with 'rbf' Kernel:

We train the training data from the splitted data with SVC using ‘rbf’ kernel.
Tuned hyperparameters : 'C': 1, 'gamma': 0.1, 'kernel': 'rbf', 'probability': True
Accuracy : 0.8597
Best Estimator : SVC(C=1, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=0.1, kernel='rbf', max_iter=-1, probability=True, random_state=None, shrinking=True, tol=0.001, verbose=False)
Roc Auc Score: 0.7925

Classification Report:

	precision	recall	f1-score	support
-1	0.85	0.98	0.91	2363
1	0.84	0.37	0.52	637
micro avg	0.85	0.85	0.85	3000
macro avg	0.85	0.68	0.71	3000
weighted avg	0.85	0.85	0.83	3000

II. SVM with 'poly' Kernel:

We train the training data from the splitted data with SVC using ‘poly’ kernel.
Tuned hyperparameters : 'kernel': 'poly', 'gamma': 0.5, 'degree': 2, 'C': 10, 'probability': True
Accuracy : 0.8614
Best Estimator : SVC(C=10, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=2, gamma=0.5, kernel='poly', max_iter=-1, probability=True, random_state=None, shrinking=True, tol=0.001, verbose=False)
Roc Auc Score: 0.8097
Classification Report:

	precision	recall	f1-score	support
-1	0.85	0.98	0.91	2363
1	0.83	0.38	0.52	637
micro avg	0.85	0.85	0.85	3000
macro avg	0.84	0.68	0.71	3000
weighted avg	0.85	0.85	0.83	3000

G. Random Forest:

We train the training data from the splitted data with Random Forest.
Tuned hyperparameters : 'n_estimators': 155, 'min_samples_split': 11, 'max_features': 'auto', 'max_depth': 16
Accuracy : 0.8614
Best Estimator : RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=16, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=11, min_weight_fraction_leaf=0.0, n_estimators=155, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False)
Roc Auc Score: 0.8505
Classification Report:

	precision	recall	f1-score	support
-1	0.87	0.96	0.91	2363
1	0.76	0.46	0.57	637
micro avg	0.85	0.85	0.85	3000
macro avg	0.81	0.71	0.74	3000
weighted avg	0.85	0.85	0.84	3000

H. Extreme Gradient boosting:

We train the training data from the splitted data with XGB Classifier.

Tuned hyperparameters : 'n_estimators': 20, 'min_child_weight': 1, 'max_depth': 5, 'learning_rate': 0.3, 'gamma': 0.001

Accuracy : 0.8653

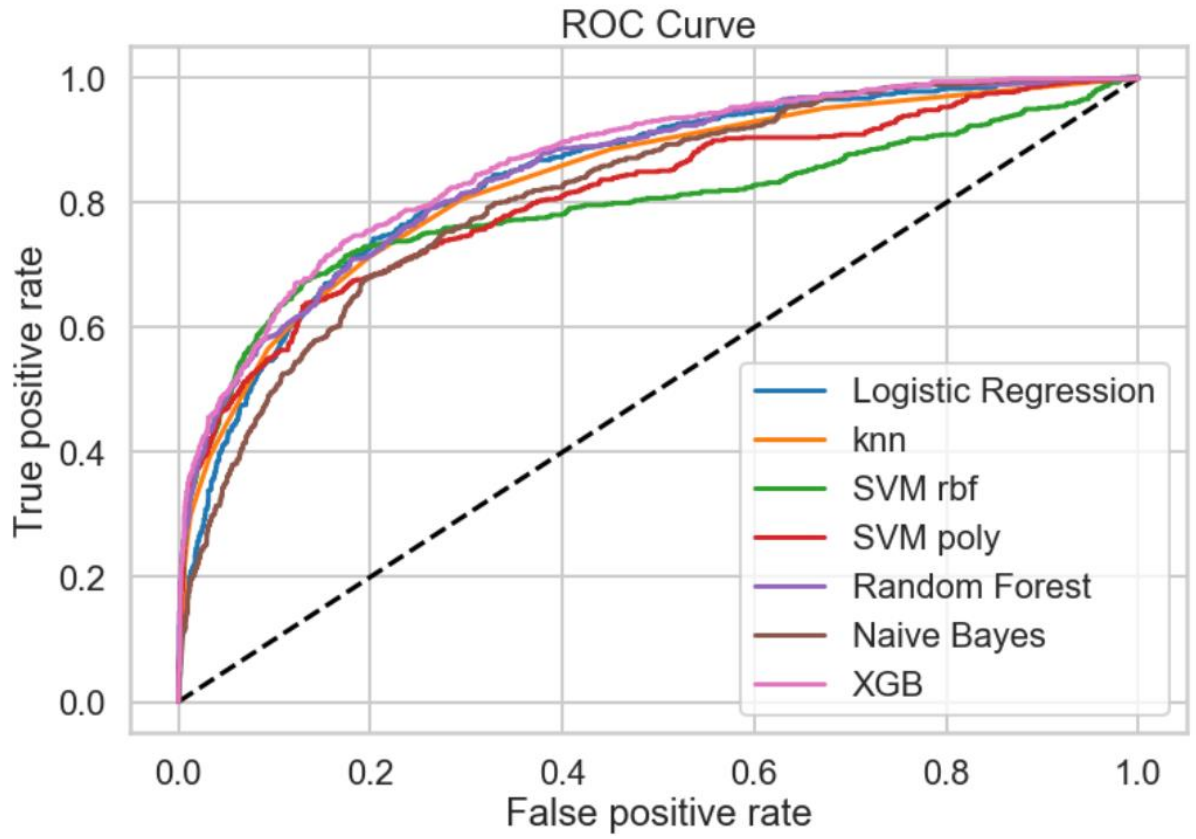
Best Estimator : XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0.001, learning_rate=0.3, max_delta_step=0, max_depth=5, min_child_weight=1, missing=None, n_estimators=20, n_jobs=1, nthread=None, objective='binary:logistic', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=1)

Roc Auc Score: 0.8635

Classification Report:

	precision	recall	f1-score	support
-1	0.87	0.97	0.91	2363
1	0.78	0.46	0.58	637
micro avg	0.86	0.86	0.86	3000
macro avg	0.83	0.71	0.75	3000
weighted avg	0.85	0.86	0.84	3000

Combined ROC Curve:



Our main aim is to predict the customers who will possibly churn, so we will have to so they can be put in some sort of scheme to prevent churn hence the recall measures on the 1's is of more importance to me than the overall accuracy score of the model.

Given that in the data we only had 20% of churn, a recall greater than this baseline will already be an improvement but we want to get as high as possible while trying to maintain a high precision so that the bank can train its resources effectively towards clients highlighted by the model without wasting too many resources on false positives.¶

From the review of the fitted models above, the best models that give a decent balance of recall and precision are the SVM, which gives a better precision of 0.84 and Extreme Gradient Boosting, which gives a better recall of 0.46. According to fit on the training set, with a precision score on 1's of 0.84, out of all customers that the model thinks will churn, 84% do actually churn and with the recall score of 0.46 on the 1's, the model is able to highlight 46% of all those who churned.

5. Assumptions and Limitations

Few Assumptions that are taken considering the data.

- This is not continuous data. Thus the assumption is that the data is a snapshot as some point in time e.g. the balance is for a given date.
- There are customers who have exited but still have a balance in their account. The assumption is that the customers had that much of amount when they cleared the account.
- There is an active member but have exited. This should mean that the customer was an active member before they left.
- A break down to the products bought into by a customer could provide more information topping listing of product count

6. Conclusion

In this project, we explored the bank customers' detail dataset to predict the customers who will churn. While doing data exploration we have many interesting points such as the average age of customers who churned is greater than those who did not churn, female churned more than the male, non-active members churned more than the active members, more members of Germany churned than other countries.
We have run multiple algorithms on the dataset such as Logistic Regression, SVM, Random Forest, etc. The accuracy is not very high and the recall is low. This can be improved by adding more data. The continuous data will be very beneficial in improving the prediction. The precision is good, thus the model will at least help the bank to implement a scheme for predicted customers with low loss.