**Churn Modelling**

Omar Saeed, Moataz ashraf, Omar Moataz

**Introduction**

The dataset contains 13 features about customer demographics, Geo data and banking personal data such as Tenure, balance, credit score and number of products purchased from the bank. The dataset also contains a binary target variable to determine whether the customer is churning.

The purpose of this project is to use the given customers’ features to predict whether a customer is likely to Churn or not. It is a classification problem where the following classifiers have been modeled for this prediction.

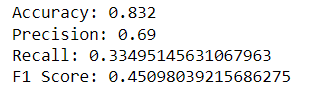
* Logistic Regression
* Random Forest
* K-NN
* SVM

The target variable is slightly unbalanced with 20% of our customers churning. Using class weights to account for this problem, our experimental results show that **Random forest** outperforms a 0.62 F1 score and accuracy of 0.81, **Logistic regression** with a 0.45 F1 score and accuracy of 0.83, **KNN** with 0.38 F1 score and accuracy of 0.78, **SVM** with 0.50 F1 score and accuracy of 0.70.

**Dataset**

The Churn dataset contains 14 Features about Customer Demographics and Banking personal information

* Customer ID
* Surname
* Gender
* Age: Age of the customer
* Geography: The country from which the customer belongs
* Creditscore : Credit score of the customer
* Tenure: Number of years for which the customer has been with the bank
* Balance: Bank balance of the customer
* Num of products: Number of bank products the customer is utilising
* HasCrCard: Binary Flag for whether the customer holds a credit card with the bank or not
* IsActiveMember: Binary Flag for whether the customer is an active member with the bank or not
* Estimated Salary: Estimated salary of the customer in Dollars
* Exited: Binary flag 1 if the customer closed account with bank and 0 if the customer is retained

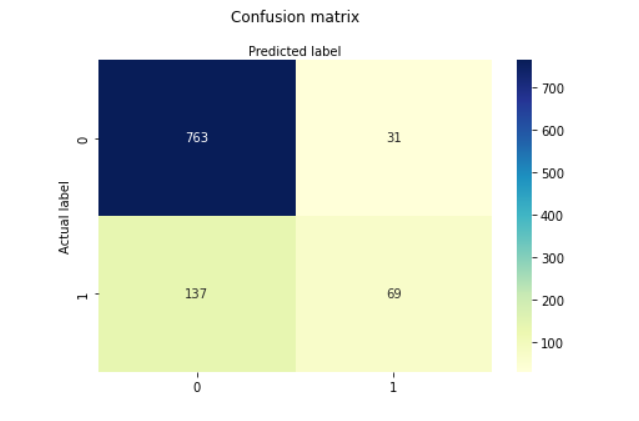
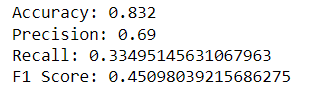


Data Preprocessing:

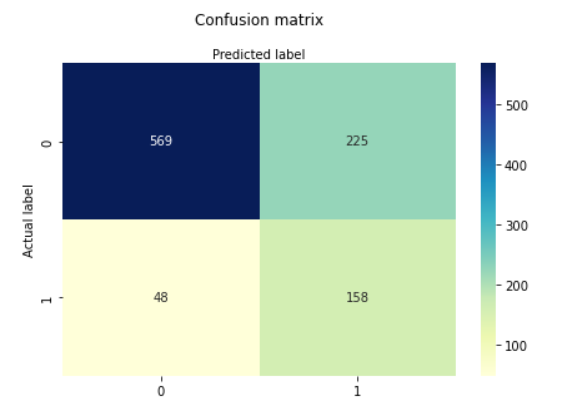
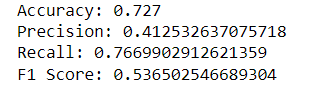
* Age is -binned into 6 Groups
  + 18-30
  + 31-40
  + 41-50
  + 51-60
  + 61-70
  + >70
* Categorical variables has been transformed into dummy variables
* Numerical variables (Credit score, Estimated Salary and Balance) have been normalized.

**Experimental Results:**

**Logistic Regression:**

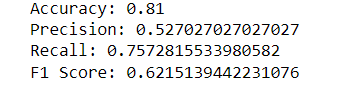
****

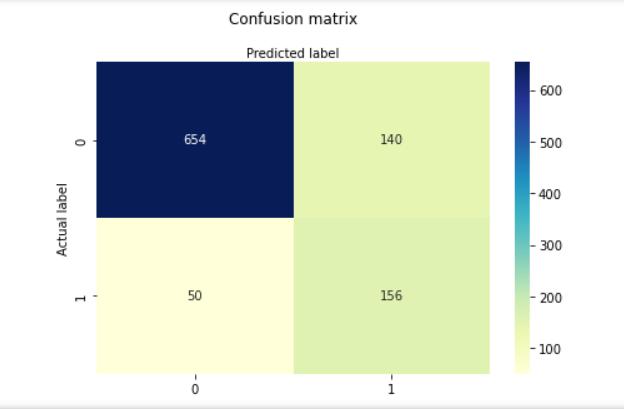
Results without using class weight show that Logistic regression outperforms of 83% of accuracy and 45% of f1 score.



Results using class weight balanced show that Logistic regression outperforms of 72% of accuracy and 53% of f1 score.

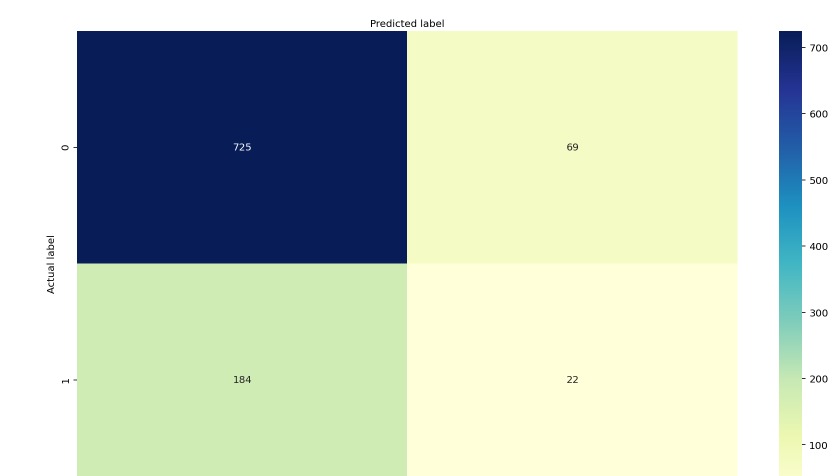
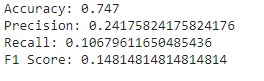
Random Forest:



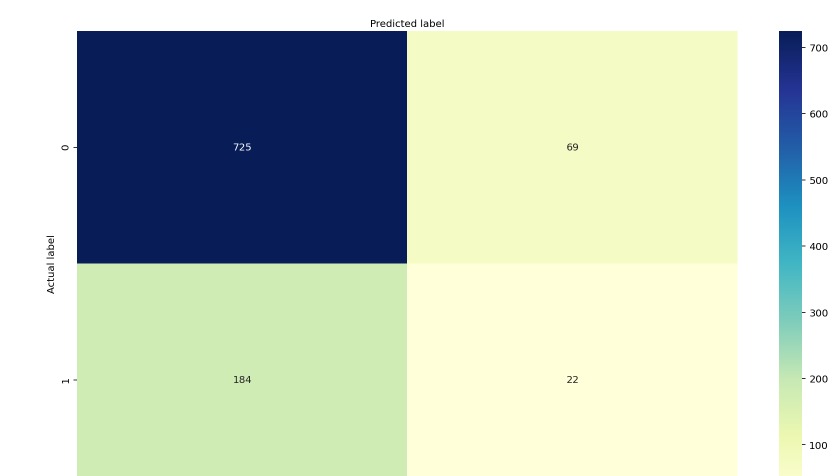
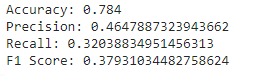


Results show that Random Forest outperforms of 81% of accuracy and 62% of f1 score.

**KNN:**

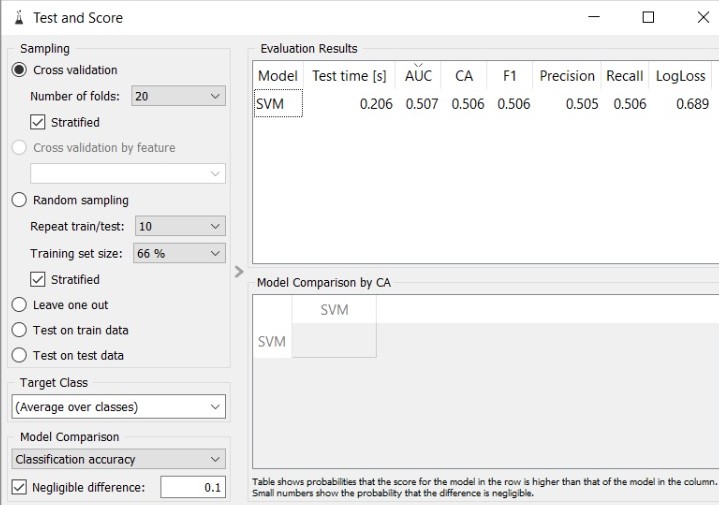


Results show that KNN outperforms of 74% of accuracy and 14% of f1 score.



Results show that KNN with feature scaling outperforms of 78% of accuracy and 37% of f1 score.

**SVM:**

****

Results show that SVM with outperforms of 70% of accuracy and 50% of f1 score.

**References:**

https://www.kaggle.com/shrutimechlearn/churn-modelling