**Final Project Report**

**Predicting Churn for bank customers**

**FE 520 Introduction to Python for Financial Applications**

August 25, 2022

Nishitha Reddy Dodda, Payal Motwani

**TABLE OF CONTENTS**

Introduction ...........................................................................................................................3

Data.........................................................................................................................................3

Data processing ......................................................................................................................4

Exploratory data analysis........................................................................................................5

Model approach ..................................................................................................................... 8

Model Accuracy...................................................................................................................... 9

Performance.............................................................................................................................10

Recommendation…………………………………………………………………………….12

References……………………………………………………………………………………12

**INTRODUCTION**

**Customer churn** (also known as customer attrition) occurs when a customer stop using a company's products or services. Customer churn affects profitability, especially in industries where revenues are heavily dependent on subscriptions. It is estimated that acquiring a new customer can cost up to five times more than retaining an existing one. Therefore, customer churn analysis is essential as it can help a business identify problems in its services and make correct strategic decisions that would lead to higher customer satisfaction

The goal of this project is to understand and predict customer churn for a bank which is a classification task. We will initially perform exploratory data analysis to identify and visualize the factors contributing to customer churn. This analysis will later help us build a predictive model to understand whether the customer will churn or not.

This problem is a typical **classification** task. It includes analyzing various performance metrics to use for optimizing our machine learning models.**Recall** can correctly classify elements of the positive class (customers who will churn) is more critical for the bank.

**DATA**

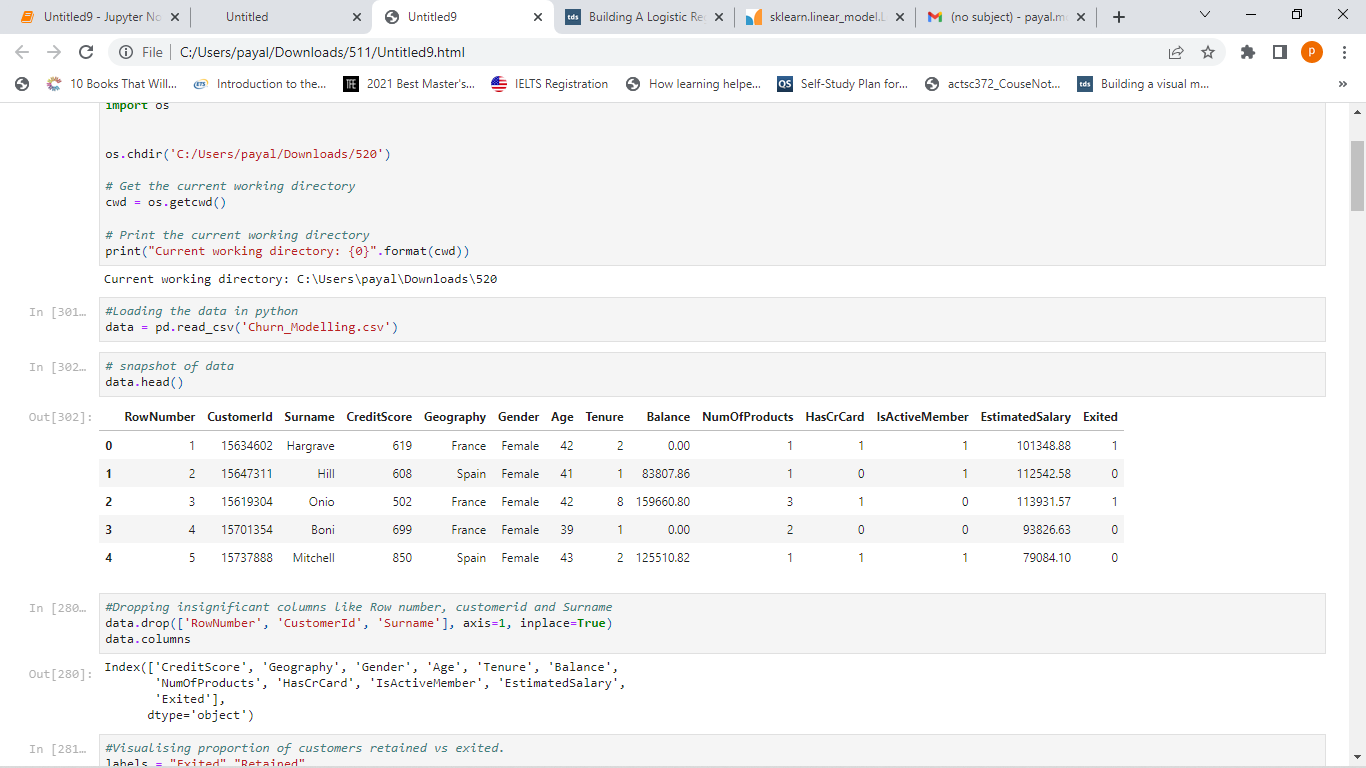
For our analysis, we have used the data from Kaggle. Our data contains 1000 data points of customers of a bank out of which certain customers decided to leave the bank and chose either of the competitor. Our goal is to analyze the factors that affected this change. About 20% of the customers are the ones that turned to another bank.

This dataset contains following 14 variables

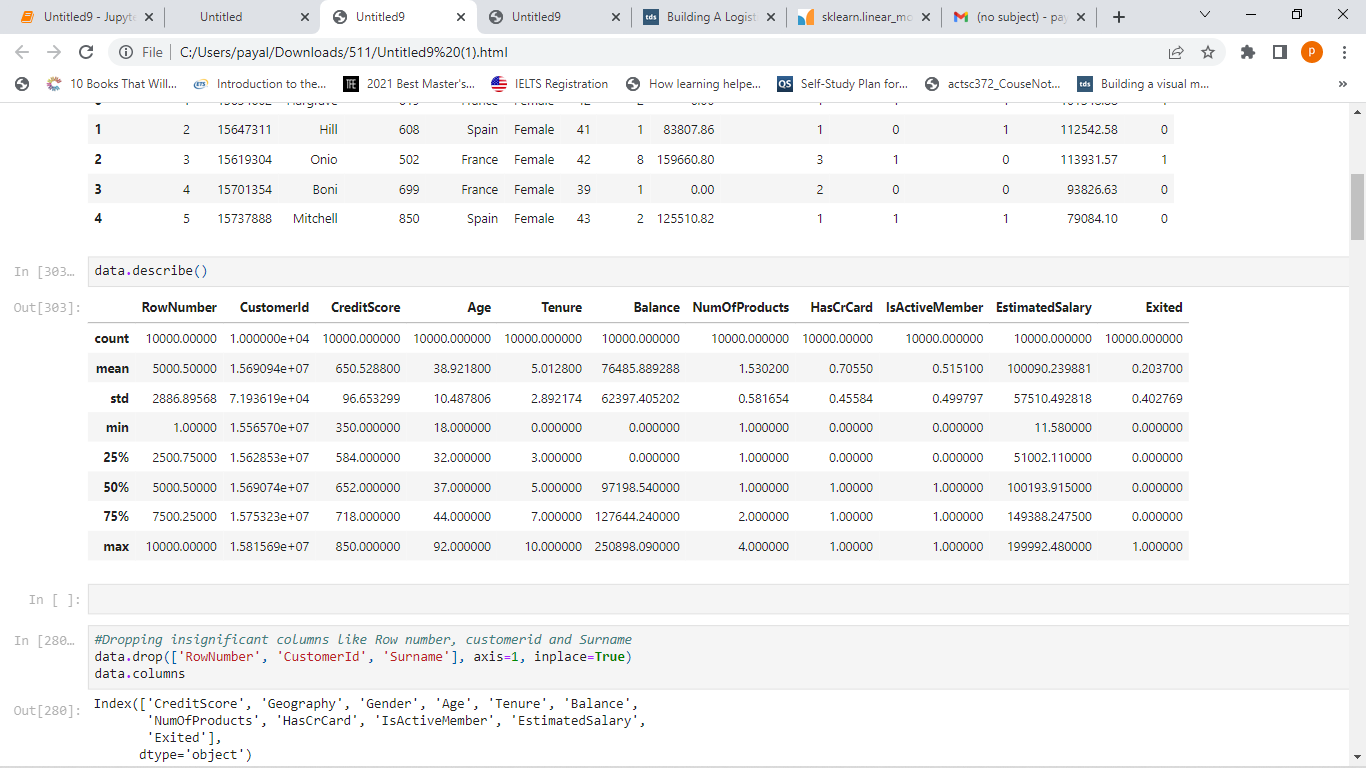
1. Row number
2. Customer id of customers
3. Surname of customers
4. Credit score of the customers
5. Geography, place where customer resides, it is a categorical variable: Germany, Spain and France
6. Age of customer
7. Tenure: years since customer joined Bank
8. Balance: Balance of customer’s account
9. Number of products: Customer’s use of bank’s products like credit card, loans etc
10. Has credit card
11. Is active member: if the account is primary or secondary
12. Estimated salary
13. Exited: categorical variable: 0 means still with bank, 1 means left the bank. This is our target variable.

**DATA PROCESSING**

Each row represents a customer, each column contains customer’s attributes described on the column. A brief snapshot of data as below:



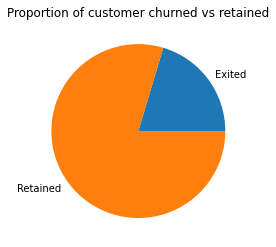
In initial screening of data, we decided to drop Row number, customer id, surname since they represent and id of customer and has no effect on the target variable.



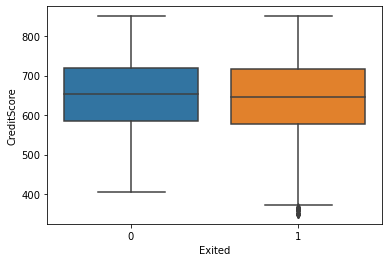
A summary of variables as above. The data captured mean, median, min and max for all the variables. Numerical variables like age, tenure, balance, and number of products, estimated salary were analyzed to find outliers, if any.

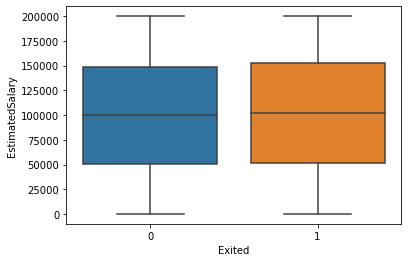
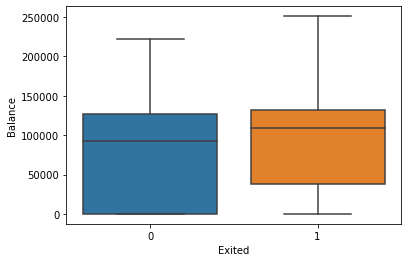
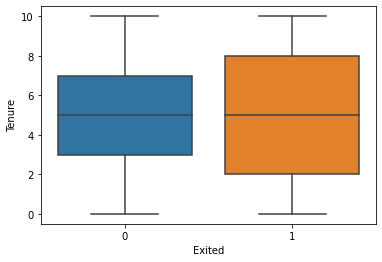
**EXPLORATORY DATA ANALYSIS**

Bank data contains customers who moved from the bank. The pie chart below shows the percentage of customers that retained with bank vs customers who left. From the pie chart we can see that our data is skewed towards retained customers. Around 20% of the customers exited the bank.



**Relationship between numerical variables and customer retention**

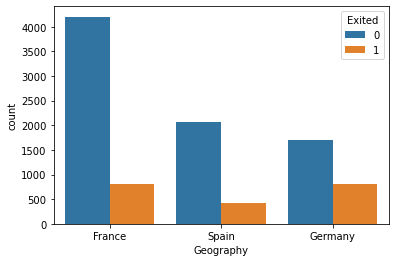
**Chart, box and whisker chart

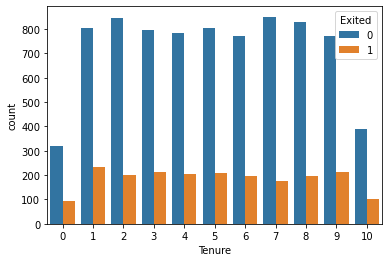
Description automatically generated**

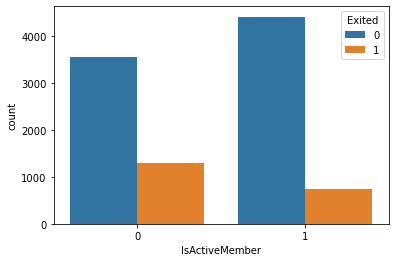
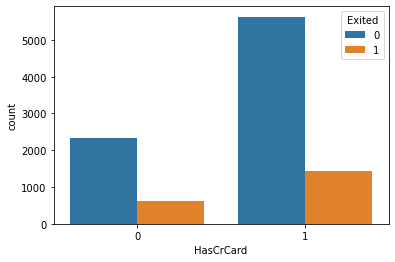
Important points to note from above bar charts

1. Mean credit score of people who left the bank was little less than retained customers. They are also few outliers with low credit score in customers that closed their accounts.
2. Mean age of customers that left the bank was much higher than existing customers.
3. Tenure of customers with bank was more spread in customers that left the bank, indicating that Tenure has less effect on decision to move.
4. Customers those maintained higher balances tend to shift more.
5. Estimated salary’s barchart is almost same indicating salary has no or little effect on the decision to move.

**Relationship between categorical variables and customer retention**

** Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated**

Important points to note from above bar charts

1. Customers of Germany were the ones that exited the bank more comparative to other locations
2. Female customers were more likely to shift
3. Tenure has no effect on the decision to exit
4. People with just one product moved more than customers that were using more products from the bank.
5. Having a credit card led to more attrition and active members were more likely to stay with bank.

**Feature selection:**

'CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'Balance',

'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary',

'Exited'

**MODEL APPROACH**

**Regression Analysis**

Regression is a way of predictive modeling using statistical methods that help us to analyze and understand the relationship between two or more variables of interest. It is a process of getting the response variable as a function of different attributes that matter in its prediction. It is a conglomerative analysis which includes the process right from feature selection, model fitting, Prediction, finding the accuracy of the model which makes it long lasting.

There are different types of regression analysis like linear regression, polynomial regression, logistic regression, discriminant analysis. For this dataset for customer churn, we think that logistic regression best fits the dataset.

**Logistic Regression**

A statistical model typically used to model a binary dependent variable with the help of logistic function. It establishes a relationship between dependent and independent variables. Another name for the logistic function is a sigmoid function and is given by:

This function assists the logistic regression model to squeeze the values from (-k, k) to (0, 1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

**Splitting the data into Training and Test data.**

We split the data randomly into Test and training data to fit a logistic regression model to training set and then check the result on test data. Training data contains 75% of the data points and test data contains remaining 25% customers.

**MODEL ACCURACY**

As an important step of any predictive modelling, we have verified the accuracy of our model using the following techniques. Here is our understanding on the same.

Model accuracy of the logistic regression model is 0.7954

The coefficients and intercept for the equation to formulate the equation to predict the customers churn:

Text

Description automatically generated

**Mean Squared Error**:

The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.

The mean squared error of the logistic regression model to predict bank customer’s churn was 0.2272. The mean squared error is low which reflects on a good accuracy score.

**Confusion Matrix**:

It is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. This consist of four different parts:

* True Positive (TP) - These are cases in which we predicted yes, and that’s actually yes.
* True Negative (TN) - We predicted no, and it’s actually no.
* False Positive (FP) - We predicted yes, but they actually was no.(Type 1 Error)
* False Negative (FN) - We predicted no, but actually it is yes.(Type II Error)

Chart, treemap chart

Description automatically generated

confusion matrix:

[[1910 49]

[ 519 22]]

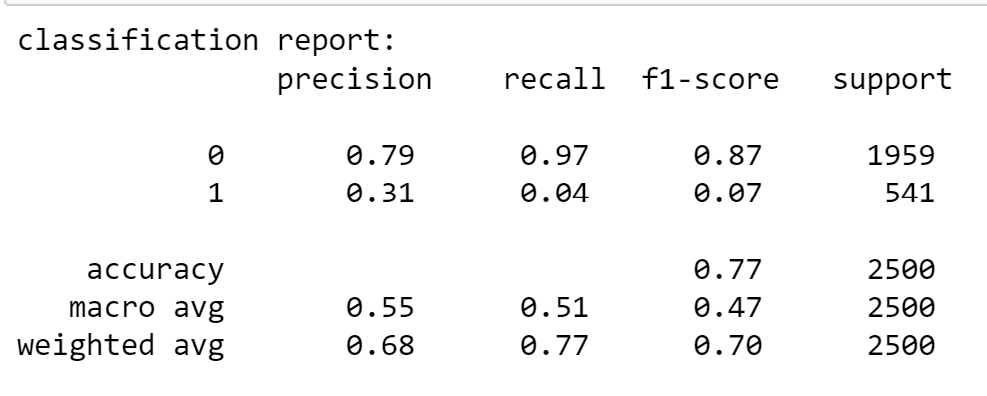
tn: 1910, fp: 49, fn: 519, tp: 22

**MODEL PERFORMANCE**

**Classification report:**

Using the classification report we have got more detailed analysis:

* Precision (79%), this means out of all values of churn we were able to predict 79% times correctly.
* Recall (97%), this means 97% of the actual values are predicted correctly.
* F1 score is 87% which is the average of precision and recall.



**ROC Curve and AUC:**

Diagram

Description automatically generatedAn ROC Curve (receiver operating characteristic curve) is a graph showing the performance of a classification model with the help of True Positive Rate and False Positive Rate at different classification thresholds. AUC stands for Area under the ROC curve measures the entire two-dimensional area underneath the entire ROC curve.

* Using True Positive and False positive rate we have the ROC curve on different classification threshold.
* AUC is 0.66 for this model that means overall we are able to classify and differentiate the customers very well.

Chart, line chart

Description automatically generated

Below figure explains how the Precision of the model varies as the recall increases(True positive rate). This is because it overshadows the true negative and leads to declining precision

Chart, line chart

Description automatically generated

**RECOMMENDATION**

* Finally, we are ready with the equation using the variables (Coefficients and intercept) through which we can easily predict the churn of bank customers.
* This model can be made more useful if we are able to add the cost of false positive and true negative predictions. So we are able to optimize the cost and improve the values of one among them which is less costly using different classification thresholds.

**REFERENCES**

[1] Menard, S., 2002. Applied logistic regression analysis (Vol. 106). Sage.

[2] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression. New York: Springer-Verlag.

[3] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.

.

[4] Narkhede, S., 2018. Understanding auc-roc curve. Towards Data Science, 26, pp.220- 227.

[5] Lewis, K.P., 2004. How important is the statistical approach for analyzing categorical data? A critique using artificial nests. Oikos, 104(2), pp.305-315.

[6] <https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares>