Computer Science Department

Capstone Project

***Prediction of Customer churn in a Bank***

CSC 521

May 2022

By

*SAMUEL KAWUMA*

CSC 521 Project Supervisor

*FATEMA NAFA*

**Contents**

1. Student Objectives x
2. Problem Specification x
3. Solution Process and Design
4. Requirement Collection and Analysis

1. Data Sequence Diagram x

1. High level Architectural Diagram x

1. UI Mockups x

1. Benchmark Specification x

1. Tool List x

1. Time Schedule x

1. Grading Scheme for CSC521 x

1. Deliverables x
2. References x

# Student Objectives

The main objective of my capstone course is to enable me to integrate the knowledge I have gained as a result of pursuing my computer science degree program in college. I plan to integrate the skills and concepts learned systematically during my semester

As a result, I should be able to integrate all aspects of the course which includes the theory, practical skills, and communication skills. I plan to combine the diverse skills acquired in class and apply them work environment. The capstone project thus should act as a self-assessment mechanism that reminds me of the expected learning outcomes.

My Capstone project should, therefore, focus on helping me to perfect technical skills in handling a given task. This requires me to have hands-on experience in the fields of specialization that I have chosen to undertake

**Problem Specification**

My project as an aspect of Data Science analyses private data of Bank Customers with goals to discover key insights from the bank customers database and study the customers’ demographics such as customer (gender, age, and location).

Also, I incline to understand the company’s product and customer’s financial history such as customer (credit score, estimated salary, balance, tenure, credit card possession, etc.). Lastly, how variable such as customers demographics and financial history affects the customers churn rate. The web application ia aimed at performing analysis and developing a prediction model for the bank customer churn.

# Solution Processes and Design

1. **Requirements and Analysis**

Stake holders

**Banks :** Since customers are the most valuable assets of most banking institutions, It is advantageous for banks to know what leads a client towards the decision to leave the company. Churn prevention allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible.The input to the application is a .csv file served as the raw data for the application. The raw data consists of many attributes. Here is a screenshot of the titles of the .csv file.

**Customer Id**

contains random values that are independent and unique has no impact.

**Surname**

the surname of a customer is independent value and hence has no impact

CreditScore—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.

**Geography—a customer’s location can affect their decision to leave the bank. Gender—it’s interesting to explore whether gender plays a role in a customer leaving the bank.**

**Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.**

**Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.**

**Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.**

**NumOfProducts—refers to the number of products that a customer has purchased through the bank.**

**HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.**

**IsActiveMember—active customers are less likely to leave the bank.**

**EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.**

**Exited—whether or not the customer left the bank.**

**METHODOLOGY**

**In this project I will use CRISP-DM to build the banck customer churn prediction model in this methodology , a 5– phase technique was used ;**

1. **Data Understanding**
2. **Data Understanding**
3. **Data Preprocessing**
4. **Modelling and Evaluation**
5. **Model Deployment**

**1.Data Collection;**

**The data used in the project to perform the analysis and predictive meodelling of bank customer churn was sourced from Kaggle.**

**2.Data Understanding;**

**Data understanding focusses on the identification , collecting and analysis of the data that will essentially help in accomplishing the project goals. This phase will entail four tasks:**

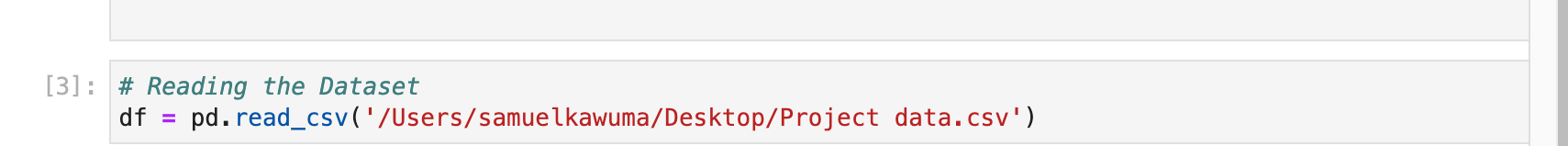
1. **Collect initial data: Acquire the necessary data and (if necessary) load it into your analysis tool.**
2. **Describe data: Examine the data and document its surface properties like data format, number of records, or field identities.**
3. **Explore data: Dig deeper into the data. Query it, visualize it, and identify relationships among the data.**
4. **Verify data quality: How clean/dirty is the data? Document any quality issues.**

**The first step is to import all the necessary libraries needed for analysis and modelling**

**Graphical user interface, text, application, email

Description automatically generated**

**Since the data is in csv format , use . read\_ csv(). format to read the data.**

****

**To view the first 5 columns of the data,** **we use .head() function. Here is an overview of what the dataset looks like by calling the name of the file using .head() function.**

**Table

Description automatically generated with medium confidence**

**To get the statistical overview of the data , I used . describe()**

**Table

Description automatically generated**

**The .info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame. Check the image below to view basic information about the data.**

**Graphical user interface, application

Description automatically generated**

**From the above, there are 10000 observations and 14 variables in the data set and there were no missing values. Since there are no missing values let’s perform basic visualization to understand how the data is distributed.**

**We will start by visualizing the target variable Exited**

**Chart, bar chart

Description automatically generated**

**From the visualization above, the number of customers that exited the bank is lower compared to the number of customers that didn’t leave the bank. Let’s visualize the relationship between the target variable (exited) and the categorical and numerical variables.**

**Gender Distribution**

**Chart, bar chart

Description automatically generated**

**From the visualization above, Female customers left the bank more often compared to the Male customers.**

**Distribution of ‘Geography’ and the target variable (Exited).**

**Chart, bar chart

Description automatically generated**

**From the visualization above, the average loss of customers is highest in Germany followed by France and the least in Spain.**

**Distribution of ‘NumOfProducts’ and the target variable (Exited).**

**Chart, bar chart

Description automatically generated**

**From the visualization above, it is observed that customers who buy more than 2 products have a high rate of loss, but let’s not forget that our data is unstable. All of the customers (60 people) who bought 4 products left the bank. I believe there might be something unexplained in the data here. Perhaps it is because the bank used to have more products but now it doesn’t, and older customers, with greater tenure, that have been with them for a long time, benefited from different products/services that are no longer available.**

**Age and the target variable (Exited).**

**Chart

Description automatically generated**

**From the visualization, exited customers are older, on average, than those still active. This kind of makes sense, as clients who have left must have been with the bank some time. The young ones have not really had the reason or the opportunity to yet leave. The bank should look out for middle aged clients who might be looking for alternatives.**

**Is Active Member’ with the target variable (Exited)**

**Chart, bar chart

Description automatically generated**

**From the visualization above, customers who do not actively use the bank leave the bank more. This is a sure sign of not sticking with the bank much.**

**The analysis and visualization of the dataset above show the dataset is unstable/imbalanced. This means that the number of data points available for the classes is different. For example, the number of exited customers is lower than the number of customers that didn’t exit the bank.**

**3.Data Preprocessing**

**This stage refers to data preparation. It prepares the final data for modelling. It involves data cleaning, feature engineering, feature scaling, data formatting, etc. Firstly, I dropped the “RowNumber”, “CustomerId”, and “Surname” columns because they are not needed in this analysis i.e they don’t have any effect on the problem to be solved. Below is a code snippet to show how these columns are dropped**

**Graphical user interface, text

Description automatically generated with medium confidence**

**To detect the presence of outliers in the datasets, I performed basic visualization using a boxplot of the seaborn library to detect outliers.**

**Chart, box and whisker chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**From the above visualization, there is the presence of outliers in columns such as “CreditScore”, “Age”, “NumOfProducts”. To remove and clean the outliers, I created a function to remove the outliers and I used the pandas library to clean the data.**

**Graphical user interface, text, application, email

Description automatically generated**

**Feature engineering was used to convert categorical data into numerical data to prepare our data ready for modelling and therefore creating more features in the dataset. Since the column “geography” is a categorical data let’s one-hot encode it by using pandas library (pd.get\_dummies) to create more features from the “geography” column. Also, we create a function to convert the categorical data in “gender” to numerical data. For example, male = 0 while female = 1.**

**Text

Description automatically generated**

**Correlation Matrix for our Variables**

**Graphical user interface, application

Description automatically generated**

**From the above, we observed age has the strongest relation with Exited (0.35). Here we can assume that as the age of the customer increases, the rate of losing the customer increases. (Positive strong relationship). Also, exited and balance variable have a relatively strong relationship (0.12). And lastly, exited and the variable NumOfProducts have a moderately strong relationship (-0.11). They have a strong negative relationship.**

**I perform feature scaling (standardization) on some features using sklearn library (StandardScalar) to scale down features into properties of Standard Normal Distribution where mean = 0 and standard deviation = 1. On realization scaling gave a higher performance in algorithms that involves gradient descent such as Logistic Regression, Support Vector Classifier, and KNN. This means that feature scaling improved the performance of my (Logistic Regression, Support Vector Classifier, and KNN) models.**

**4. Data Modelling and Evaluation**

**This This involves building and developing various models based on several different modelling techniques. In this stage, we determine the algorithm to use for predictive modelling and evaluate which models give the best performance.**

**Pending modelling selection and approach, we might need to split the data into training and test sets using sklearn train\_test\_split library. Since the project is a classification-based project (exited or not exited), I used classification models such as LogisticRegression followed by Random Forest Classifier and XGB Classifier to make a prediction.**

**After performing several operations such as hyper-parameter tuning, cross-validation the highest output was taken with Random Forest Classifier followed by XGB Classifier and LogisticRegression. Check below to see how each classifiers performed.**

**Firstly lets view Random Forest Classifier**

**Graphical user interface, text, application

Description automatically generated**

**Lets View XGB Classifier**

**Graphical user interface, text, application

Description automatically generated**

**And Lastly LogisticRegression**

**Table

Description automatically generated with medium confidence**

**After running the different models , Random Forest obtained the best accuracy score with accuracy of 86.2%.**

**Table

Description automatically generated with medium confidence**

**Model Deployment**

**This is the stage the model goes into production i.e the stage in which clients can access the model results. It involves series of complex operations ranging from deployment plan to plan maintenance and final report production.**

**his final phase has four tasks:**

* **Plan deployment: Develop and document a plan for deploying the model.**
* **Plan monitoring and maintenance: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.**
* **Produce final report: Document a summary of the project which might include a final presentation of data mining results.**
* **Review project: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.**

**I deployed my model into production using flask which is a web application framework written in Python. It is a relatively simple way of building web applications and deploying machine learning models. For the functionality of my application.**

**User case Diagram**

**Diagram

Description automatically generated**

From the above diagram, the System Expert Manage the data set. In data set Management, we notice that data is selected, is filtered, is manipulated and then thereafter ready for visualization. The system user can also perform data pre-processing and Data analysis and finally the system user can perform Visualization of the data

1. High Level Architecture Design



**Data Collection**

The first phase is the data collection phase although my project doesn’t not involve data collection but. Rather. Downloaded data set from [www.kaggle.com](http://www.kaggle.com)

**Data Preprocessing**

The second phase is data pre processing. This is performed to reduce on the number of columns or removal of null values which somewhat may not be needed. During pre-processing data is cleaned and normalized. Normalization is the process of reducing. The number of features which is required or classification. In this case irrelevant columns that are not going to be used and visualized may be. Removed/deleted at this point. The motivation behind this is. To improve thorough output.