University of Hertfordshire

School of Physics, Engineering and Computer Science

MSc. Advanced Computer Science with Advanced Research

MSc. Computer Science Project

(7COM1039)

Interim Progress Report

Project title – An evaluation of unsupervised and supervised learning models to predict and analyse customers’ rate of churn

Name: Simpson Chiwashira

SRN: 21017447

Level: 7

Academic Year: 2022-2023

Supervisor: John Sapsford

**Table of Contents**

[**1. Introduction** 3](#_Toc143294205)

[1.1 Background of Research Area 3](#_Toc143294206)

[1.2. Project’s Aims and Objectives 4](#_Toc143294207)

[1.3. Advanced Project’s Aims 4](#_Toc143294208)

[**2. Project Design** 5](#_Toc143294209)

[2.1 Summary of Progress to Date 5](#_Toc143294210)

[2.2. Project Plan 6](#_Toc143294211)

[2.3. Legal, Ethical, Professional and Social Issues 6](#_Toc143294212)

[**3. References and Bibliography** 7](#_Toc143294213)

[**4. Appendices** 8](#_Toc143294214)

[Appendix 1: Gantt Chart 8](#_Toc143294215)

[Appendix 2: Version Control 8](#_Toc143294216)

[Appendix 3: Evidence of Supervisory Meetings 9](#_Toc143294217)

[Appendix 4: System Specifications 10](#_Toc143294218)

[Appendix 5: Installation of Libraries 10](#_Toc143294219)

[Appendix 6: Relevant Code and Visualisations. 11](#_Toc143294220)

Table of Figures

[Figure 1 9](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Interim%20Progress%20Report%20-%20Masters%20Project.docx#_Toc143297016)

[Figure 2 9](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Interim%20Progress%20Report%20-%20Masters%20Project.docx#_Toc143297017)

[Figure 3 11](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Interim%20Progress%20Report%20-%20Masters%20Project.docx#_Toc143297018)

[Figure 4 12](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Interim%20Progress%20Report%20-%20Masters%20Project.docx#_Toc143297019)

[Figure 5 13](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Interim%20Progress%20Report%20-%20Masters%20Project.docx#_Toc143297020)

# **1. Introduction**

**Keywords**: Customer churn, prediction, supervised learning, and unsupervised learning.

## 1.1 Background of Research Area

As technology advances a variety of sectors have been affected. Technology has become quite instrumental in the decision-making process in the business and marketing industry. The two business sectors which have benefited from this advancement include the telecommunication and banking sectors.

According to Wu et al (2021), the development of 5G technology has created a competitive market where most telecommunication companies would wish to retain their customers. (He, 2022) argues that the banking sector has seen a sizeable impact in the decrease in profit due to customer churn.

Customer churn can be defined as the act of leaving a service or subscription provider for another provider due to many reasons (Dalvi et al, 2016). Authors (Kaur. I and Kaur. J, 2020), define it as customer attrition and one would stop a relationship with a certain brand or company.

One of the reasons why a customer would churn includes having a highly saturated market and hence creating competition. Wu et al (2021) claim that having to look for new customers would prove to be highly expensive than when retaining current customers. Marketing and trying to lure customers into subscribing to a specific service will be very difficult and this is the reason why I justify having to use different retention methods.

According to Coussement, Lessmann and Verstraeten (2016), retention methods can keep track of customers who are at risk of discontinuing their services with the company. Wu et al (2021) suggest that using the proactive approach in identifying the customers before they churn would be key to retaining the current customers. Customer churn predictions would be key in identifying the probability of an individual leaving in the future after analysing the historical record of the churners.

Machine learning models in the past decades have developed exponentially. The use of these tools has seen a wide variety of applications and use. According to Karvana et al (2019), machine learning techniques used in data mining which include classification and prediction should be utilised in customer churn.

Dalvi et al, (2016) on that note discuss how historical evidence in the companies can help give a forecast of who is at risk to leave. We agree with the authors as they even stress how data analysis tools would be used in isolating the characteristics of the customers who have left the business.

In their research, (Rudd, Huo, and Xu, 2021) suggest that we ought to use prediction systems to identify specific churners beforehand. Coussement, Lessmann and Verstraeten (2016) agree with this claim as they suggest the importance of predictive modelling as a method to identify customer churn.

We have justified why there is a need for an evaluation of these different methods for recommendations to be made to users of the churn prediction tools. A literature review on these methods will be conducted and documented as we continue to investigate the methods in question.

## 1.2. Project’s Aims and Objectives

The research aims to find out what methods have been used and what results have they yielded. The goal of the project is to evaluate supervised and unsupervised learning models to predict and analyse the rate of customer churn in a business. The following are the Research Questions (RQ) we wish to answer with this investigation:

*RQ: Which learning model between the supervised and unsupervised produces the best performance in the prediction of customer churn?*

*RQ: Is there a difference between the performances of supervised and unsupervised learning models when predicting the rate at which customers discontinue their service?*

## 1.3. Advanced Project’s Aims

* To use different metrics to measure the performance of the chosen model and try to use a different approach to attain better scores.
* To do a comparison of results and make recommendations as to why the model I have chosen, and the parameters give better results and should be considered by future developers.

# **2. Project Design**

## 2.1 Summary of Progress to Date

A Gantt chart was designed, and it will be used as the project management tool responsible for important major tasks and milestones of the project. It is key to the completion and success of the project as it gives an overview of the project and the timeline by which the tasks and milestones are expected to be accomplished. A copy of the Gantt chart is available in the appendices section (see Appendix 1). The project management tool used is the Team Gannt software which is free and readily available for use to anyone.

For us to begin our secondary study of the literature review, the University’s digital library was utilised for the review protocol where exclusion and inclusion criteria of the sources were applied. A literature review on the background of the research area was conducted and was able to make decisions based on it on how to move forward with the project.

After executing the review protocol, two major businesses that are affected by this problem identified are the banking and marketing sector and telecommunication companies. A dataset which consists of 3 months' worth of data for 70,000 instances and 172 attributes was chosen. This dataset belongs to a telecommunication company and was collected in 2014 and it is publicly available for use at the Kaggle platform. The banking sector and marketing dataset was identified but it needs further preprocessing before it can be loaded and used. It should be noted that the original plan was to use one dataset and after a meeting with the supervisor the idea of using two datasets for comparison reasons was adopted and hence the reason to use the datasets chosen.

The hardware and software requirements were met, and the environments needed for the first stage of data pre-processing. Anaconda is an open-source platform that allows you to write code in a jupyter notebook using the Python language. This platform has been installed in the workstation to be used and the required libraries for us to begin the data exploration and any related pre-processing were installed.

We have begun to integrate the data that was separated in three months to aggregate it and work to reduce the feature size. The shape and size of the data frame are still big considering the number of instances and this is why we have decided to drop some features that will not affect the classification. Also, all attributes that have zero values are being dropped. When the feature size is reduced, we can now begin dealing with the missing data and after completing the literature review a decision on the sampling method will be done and the sampling technique will be applied. In the appendices section, there are screenshots of the codes, loaded libraries and output.

## 2.2. Project Plan

The next stage of this project is to conduct a literature review of supervised and unsupervised learning models. A review of what they are and their uses in machine learning. We will have a look at the different algorithms which fall under the supervised and unsupervised learning models. The current literature on how they are used in the telecommunication and banking industry. The review will focus on what the researchers did and if it was successful and a closer look at their architecture. We can select one algorithm which has performed better in each learning model, and we can commence our investigations.

We have begun the data exploration and some pre-processing. The next will be to commence the next stage of analysing the data if it is balanced or not. Missing values and unbalanced can be dealt with using the appropriate sampling techniques identified from the literature. We will select the features that are key predictors of the models and save the data frame and ready it for the next stage of modelling and training the model.

The next stage will be to select the chosen algorithm and apply it to the model and begin the training of our models. We can start analysing the results and fine-tuning the hyper-parameters of the model. We can use the confusion matrix and make sure we reduce the False Positives and False Negatives.

Will utilise the pickle function to save the trained model and it will be loaded onto a system that we will design to visualise the data and we can enter some details of current customers and predict if they will churn. After the systems are all ready, we can now continue with the documentation of the journey of the project.

## 2.3. Legal, Ethical, Professional and Social Issues

In this project, the experiments that are being done do not involve using human subjects. This means we do not need ethical approval from the University of Hertfordshire (UH) Ethics Committee. The datasets being used in this project were obtained from Kaggle which is a public platform and that means the data is available for anyone to use.

However, it should be noted that we would like to believe the data when it was collected from the customers there was an ethical approval before it was collected. In our project, we are going to evaluate all the data features and work to try and de-identify any customer’s details if available.

# 

# **3. References and Bibliography**

1. Coussement. K, Lessmann. S and Verstraeten. G (2017) “A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry”, *Decision Support Systems*, 95, pp.27–36, doi: https://doi.org/10.1016/j.dss.2016.11.007.
2. Dalvi. P, Khandge. S, Deomor. A, Bankar. A and Kanade. V (2016). *Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression*.
3. Rudd. D, Huo. H and Xu. G (2022) “*Causal Analysis of Customer Churn Using Deep Learning”*.
4. He. H (2022) “Bank Customer Churn Prediction Analysis Based on Improved WOA-SVM”, *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, doi: https://doi.org/10.1109/iwecai55315.2022.00093.
5. Karvana. K, Yazid. S, Syalim. A and Mursanto. P (2019) “*Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry*.”, IEEE.
6. Kaur. I and Kaur. J (2020) “Customer Churn Analysis and Prediction in Banking Industry using Machine Learning.”, *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, doi: https://doi.org/10.1109/pdgc50313.2020.9315761.
7. Wu. S, Yau. W-C, Ong. T-S and Chong. S-C (2021) “Integrated Churn Prediction and Customer Segmentation Framework for Telco Business.” *IEEE Access*, 9, pp.62118–62136, doi: https://doi.org/10.1109/access.2021.3073776.

# **4. Appendices**

## A screenshot of a project Description automatically generatedAppendix 1: Gantt Chart

## A screenshot of a computer program Description automatically generatedAppendix 2: Version Control

## A close-up of a document Description automatically generatedAppendix 3: Evidence of Supervisory Meetings

Figure 1

A close-up of a letter

Description automatically generatedThe above figure 1 was an introductory meeting and the meeting notes. This is where we discussed what is expected of the project and how we will be working.

Figure 2

The above figure 2 was another meeting with the supervisor.

## Appendix 4: System Specifications

A screenshot of a computer

Description automatically generated

## A screenshot of a computer program Description automatically generatedAppendix 5: Installation of Libraries

## Appendix 6: Relevant Code and Visualisations.

Because there are many lines of code in the notebook, I am attaching the repository’s URL for bitbucket and cloning it and previewing the code and output.

A screenshot of a computer

Description automatically generated(<https://sc21aea@bitbucket.org/simpsonchiwashira/msc-acs-project.git>)

Figure 3

Figure 3 shows the loaded libraries and the head of the dataset. Also, we can see the shape of the dataset.

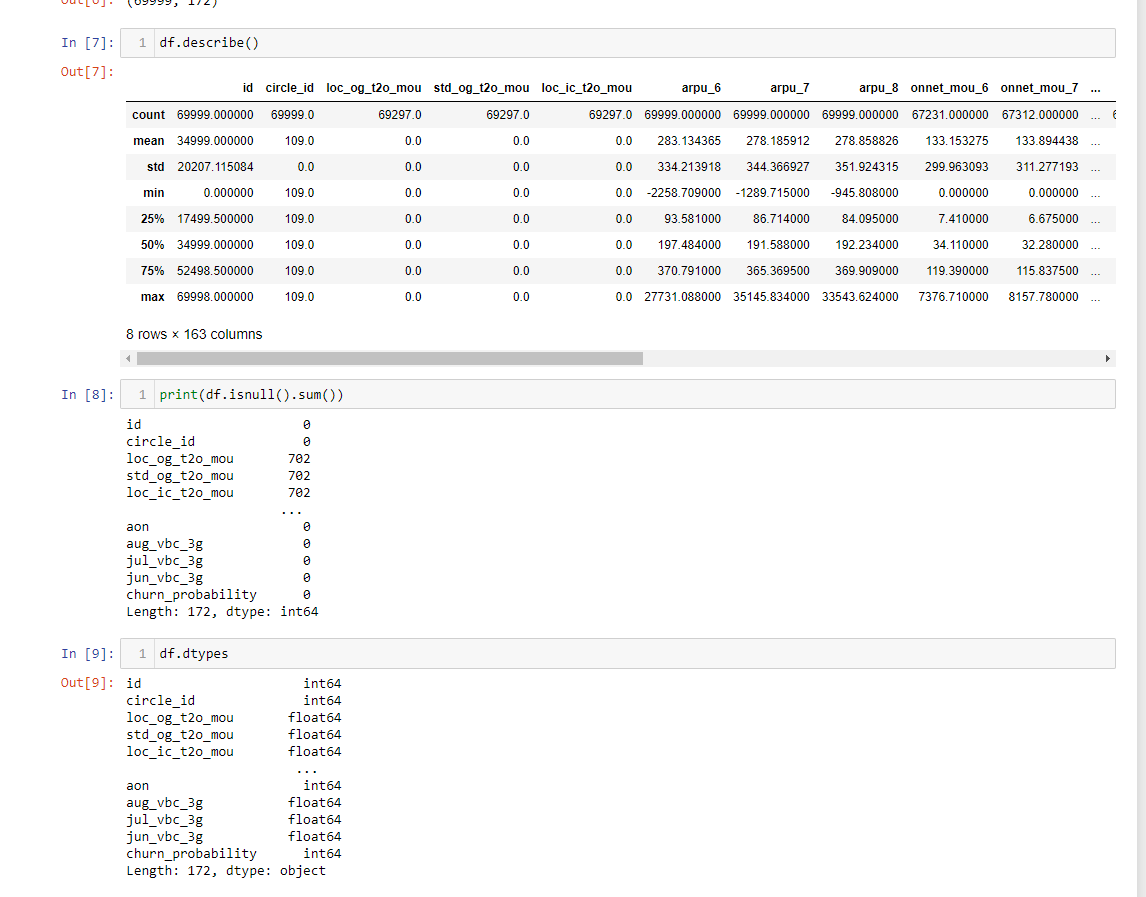


Figure 4

Figure 4 shows the summary of the dataset, and we can get to see what it looks like and the types of data we have in our dataset.

A screenshot of a computer

Description automatically generated

We can see the data is highly skewed towards non-churners. We need to balance it out.

A screenshot of a computer program

Description automatically generated

Figure 5

As we can see figure 5 is showing the code where we aggregate some data and create a new aggregated column and the shape has changed from 172 attributes to 155. The goal is to reduce the size of the data frame as much as possible.