University of Hertfordshire

School of Physics, Engineering and Computer Science

MSc. Advanced Computer Science with Advanced Research

Module: 7COM1039 - MSc Advanced Computer Science Project

Detailed Project Proposal

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

Name: Simpson Chiwashira

SRN: 21017447

Level: 7

Academic Year: 2022-2023

Table of Contents

[Introduction 3](#_Toc138949317)

[Project Aims and Objectives 4](#_Toc138949318)

[Project requirements 4](#_Toc138949319)

[**Core project requirements** 4](#_Toc138949320)

[**Advanced project aims.** 4](#_Toc138949321)

[Secondary Research – Literature Review 5](#_Toc138949322)

[Primary Research – Practical Investigation 6](#_Toc138949323)

[Project Resources 7](#_Toc138949324)

[**Hardware requirements** 7](#_Toc138949325)

[**Software requirements** 7](#_Toc138949326)

[Learning Outcomes of the Project 8](#_Toc138949327)

[References and Bibliography 9](#_Toc138949328)

**Keywords**: Customer Churn, Prediction, supervised learning, and unsupervised learning.

# Introduction

Technology has been instrumental in the making of business and marketing decisions. Machine learning models developed have advanced to the state where they can now be used to facilitate us in making informed decisions. Customer churn is the rate at which clients of a certain service discontinue the service and look for an alternative service.

The management is always looking for ways to capitalise on profits and value their customers. There is a need to create a tool that would enable them to know which customers are most likely to be leaving their services. This is a powerful tool that can help in the prediction of the rate at which they are churning, and measures can be put in place to find the best way to improve the services to retain the customers they have. The management and stakeholders can conduct some further research to understand why customers keep leaving their services.

A lot of industries including the finance sectors such as banks, building societies and insurance can adopt this powerful tool. The industries are not just limited to finance but also within the telecommunication industry.

# Project Aims and Objectives

The main aim of this project is to evaluate the unsupervised learning methods and unsupervised learning models to predict and analyse if customers will discontinue service.

The second goal of this project is to investigate which best learning model produces the best performance in terms of f1-Score, Accuracy, and the Geometric Mean metric.

These two goals will allow me to answer the main answer to my research question (RQ) which will be mentioned in the following section.

# Project requirements

## **Core project requirements**

* To select and design one supervised and one unsupervised learning model to analyse and predict to see if a customer will churn.
* To split training data and test the algorithms chosen and gain an understanding of the results.
* To adjust the hyperparameters of these two algorithms and see which algorithm will outperform the other.
* Will look at the confusion matrix and try to reduce the false positives and negatives.
* Choose the best model from the two and save it using the pickle library to load it later.
* Create a friendly user interface the shareholders and management can input data and the saved model will predict if that customer churns. The interface should have some dashboards that have insights into all customers and give recommendations.

## **Advanced project 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.

# Secondary Research – Literature Review

The plan is to use the digital libraries and other sources that are at my disposal to conduct sound and comprehensive research on different supervised and unsupervised learning models. Will define the inclusion and exclusion criteria of sources and find out how credible they are and if they were funded so that we can trust the data.

In the inclusion criteria, I look forward to having research that has an empirical study on the subject. I will look at previous research work that has been carried out and has been peer-reviewed. Will look at the approach they have taken and look at the architecture of the chosen models and how it was built. Will do a comparison of the model against other models that could have been employed instead.

A definition of the exclusion criteria will be made up of research that has been carried out but without an empirical study. Also, it will be research that is related to the subject but not arising from the keywords searched.

The research aims to find out what methods have been used and what results have they yielded, and I will conduct a gap analysis of what I could do to improve the model. This research will allow me to choose the best unsupervised and the best-supervised learning model so that I can further conduct an experiment that will answer my research questions which are:

*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?*

The above are the two potential RQs that must be answered by my research. Will need to obtain some results before the algorithm was tuned and after and see how it performs.

# Primary Research – Practical Investigation

In this phase, we will conduct our investigations to answer the above research questions. The first step is to select the dataset to conduct our experiments. I plan to find a dataset from credible and well-known websites where there are a lot of datasets. I also plan to do some research for datasets on Kaggle, Github, Amazon etc.

I would like to compare the results I achieve, and the results achieved by other researchers who have done similar research. I would take a different approach and hope to get different results. So, it makes it easy to make the comparison.

Pre-processing is the first step in data analysis and most of the time we process the data more than we execute the model. Our data analysis is as good as the input data we give the model. The first step would be to explore the data and see how many attributes I have and the number of rows and columns.

The next step is to find out how much data is missing and if I would decide on balancing the data either by over or under-sampling, Once the data is balanced and all is ready, I can start to plot some visualisations and do some preliminary analysis and try to see any patterns.

When I have my cleaned data, I am going to move forward with testing my chosen models and collect the results obtained and observed. I will fine-tune the models and try to see if a better performance can be achieved. The results will be used to make the comparison and answer the RQs stated.

The next stage is the evaluation of the results and the models and deciding on the best model will be saved and loaded onto the designed interface that will be used for demonstration on how it will predict and analyse using some visualisation. And lastly, the deployment of the model in a real-world environment.

# Project Resources

## **Hardware requirements**

* Laptop – Acer Swift 3, Intel Core i7 EVO 12th generation, 16GB RAM, 1 TB SSD

## **Software requirements**

* Jupyter notebook
* VS Code Editor
* Python 3.10
* Github Repository (Version Control)
* Trello (Project Management tool)

# Learning Outcomes of the Project

* The ability to collect data and pre-process for further examination and execution.
* I would like to learn data visualization and analysis.
* To be able to learn how to do secondary research and select suitable models for the problem and make informed decisions.
* Learning to conduct an evaluation of the obtained results and compare them with other research done in the field and draw a sound conclusion.
* Should be able to deploy the trained model into a real-world piece of software that is usable to shareholders and the management team.
* This project will allow me to pose Research Questions and be able to answer them using appropriate methods and be able to validate or refute the hypothesis.

# References and Bibliography

1. practicaldatascience.co.uk. (2021). *How to save and load machine learning models using Pickle*. [online] Available at: https://practicaldatascience.co.uk/machine-learning/how-to-save-and-load-machine-learning-models-using-pickle [Accessed 29 Jun. 2023].
2. Bedi, M. (2023). *manan-bedi2908/Customer\_Churn-Deployment*. [online] GitHub. Available at: https://github.com/manan-bedi2908/Customer\_Churn-Deployment [Accessed 29 Jun. 2023].
3. Hason, D., 1, R., Huo, H. and Xu, G. (n.d.). *Causal Analysis of Customer Churn Using Deep Learning*.
4. GitHub. (2023). *Deep Learning using Tensorflow 2.0 and Keras*. [online] Available at: https://github.com/codebasics/deep-learning-keras-tutorial [Accessed 29 Jun. 2023].