University of Hertfordshire

School of Physics, Engineering and Computer Science

MSc. Advanced Computer Science with Advanced Research

MSc. Computer Science Project

(7COM1039)

Final Progress Report

An evaluation of supervised learning models in the prediction and analysis of customer churn

Name: Simpson Chiwashira

SRN: 21017447

Level: 7

Academic Year: 2022-2023

Supervisor: John Sapsford

# **Abstract**

# **Acknowledgements**

**Table of Contents**

[**Abstract** 2](#_Toc147108652)

[**Acknowledgements** 3](#_Toc147108653)

[**1. Introduction** 7](#_Toc147108654)

[1.1 Background of Research Area 7](#_Toc147108655)

[1.2. Project’s Aims and Objectives 8](#_Toc147108656)

[1.3. Advanced Project’s Aims 8](#_Toc147108657)

[1.4. Summary of Progress to Date - IPR 8](#_Toc147108658)

[1.5. Project Plan 9](#_Toc147108659)

[1.6. Resources 9](#_Toc147108660)

[1.7. Legal, Ethical, Professional and Social Issues 10](#_Toc147108661)

[**2. Literature Review** 11](#_Toc147108662)

[2.1 Supervised Learning Models 11](#_Toc147108663)

[2.2. Algorithm Selection 11](#_Toc147108664)

[**3. Methodology** 15](#_Toc147108665)

[3.1. Data Selection 15](#_Toc147108666)

[3.2. Datasets 15](#_Toc147108667)

[3.2.1. Banking and Marketing Dataset 15](#_Toc147108668)

[3.2.2. Telco Dataset 17](#_Toc147108669)

[3.3. Data Pre-Processing 19](#_Toc147108670)

[**References and Bibliography** 22](#_Toc147108671)

[**4. Appendices** 24](#_Toc147108672)

Table of Figures

[Figure 1: Decision Boundary for SVM Classifier (Pedregosa et al, 2011) 11](https://herts365-my.sharepoint.com/personal/sc21aea_herts_ac_uk/Documents/Final-Progress-Report-Customer-Churn.docx#_Toc145143786)

[Table 1: Feature description of the bank and marketing dataset. 16](#_Toc146796122)

654

# **1. Introduction**

**Keywords**: Customer churn, prediction, and supervised 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. Below is the equation 1 to calculate the rate of churn.

CustomerChurnRate = x 100 (Equation 1)

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 more expensive than 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 usage. 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 of leaving. 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 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 supervised learning model produces the best performance in the prediction of customer churn?*

*RQ: Is there a difference in the performances of the supervised learning models when predicting customer churn?*

## 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.

## 1.4. Summary of Progress to Date - IPR

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 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 pre-processing 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 made and the sampling technique will be applied. In the appendices section, there are screenshots of the codes, loaded libraries and output.

## 1.5. Project Plan

The next stage of this project is to conduct a literature review of supervised 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 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 that 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 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 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 plans are all ready, we can now continue with the documentation of the journey of the project.

## 1.6. Resources

* Python IDE / Jupyter Notebook.
* Computer with Windows 11 Operating System (see Appendix 4 for full system specifications).
* Datasets to be used in the experiment (Bank and marketing dataset and Telco dataset).
* Microsoft Word and Excel.
* Github Repository and Bitbucket Repository (For version Control).
* VS Code (Text Editor).
* Team Gantt Software (Project Management Software tool).
* Heroku (Platform to build, host and run our application in the cloud).

## 1.7. 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 public platforms 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.

# **2. Literature Review**

Machine Learning (ML) is a subcategory field of the Artificial Intelligence (AI) domain that has evolved in the past decades and more changes have come with this development. According to Sharma, Kaur, and Semwal (2022), ML makes use of past data i.e., historical data to forecast future results. The authors also make note of how ML has had impacts in different fields like business intelligence. This study is focused on the customer churn on the business side of a Telco Company and the Banking and Marketing sector.

## 2.1 Supervised Learning Models

Supervised learning is a type of ML technique where new data is classified based on the training data (Gaur et al., 2021). The learning process is done under supervision which means the training data used will be labelled and put into classes to be observed. The input data will be added into the ML model and the weighted data will be accurately fit into the model. In their study, (Sharma, Kaur, and Semwal, 2022), the authors argue that there are two types of supervised learning which are classification and regression.

According to (IBM, 2023: Karvana et al., 2019: and Dalvi et al.,2016) classification is one of the data mining techniques where the chosen algorithm tries to distinguish classes based on the training data provided. The model classifier will predict or forecast which category or class the test data belongs to. Algorithms that fall under this category and are going to be examined for our case include Decision Trees, Support Vector Machines (SVM), Naïve Bayes, Random Forests and Artificial Neural Networks (ANN).

Regression is a ML model that tries to map the relationship between dependent and independent variables and the goal of this ML algorithm is to predict if the values are near or closer to the actual output (Gosha et al.,2021; IBM, 2023; Sharma, Kaur, and Semwal, 2022). It is mainly applicable in market prediction, price prediction etc. Algorithms that are used in this type include linear regression and logistic regression.

## 2.2. Algorithm Selection

According to our literature search of the different algorithms and models, we can note that most of the algorithms which appeared were supervised learning models and these were mostly deployed by researchers in their studies. According to Colreavy and Lewandowsky, (2008) in their empirical work, they uncovered the differences between supervised and unsupervised learning, and it all came down to a very small learning rate.

However, (Love, Medin and Gureckis, 2004) make a strong argument in which they claim that supervised learning is easier to understand but unsupervised learning is what most researchers call, “*The Blackbox*” because there is very little explanation of how the model arrived at the outcome. In the research that I have conducted, I can note the difference between these two and why most of the algorithms were selected in the cases of customer churn analysis. This is because supervised learning is mostly used for pattern recognition and customer churn analysis is all about recognising the patterns of the customers that churn whom we intend to forecast. A customer who shows the same behaviour is most likely going to be classified as a potential churner.

However, with the unsupervised, the model does not know who churned and who did not. Since it has no prior knowledge of the model will cluster data points which are similar considering the specific features which are potential predictors. Therefore, we can argue that these two models will perform well considering the application of the model. We have decided to look at supervised learning models and uncover the best which will help us answer our RQ and validate or refute the hypothesis.

A diagram of a graph

Description automatically generated with medium confidenceOne of the most popular data mining algorithms which was used by researchers in this field is the Support Vector Machine (SVM). It is part of a set of supervised learning models mostly used in instances of classification, and regression and to identify any outliers in the dataset (Pedregosa et al, 2011). The ability of this ML technique to generalise and solve both linear and non-linear data in high-dimensional feature spaces makes it ideal. According to (Piri, Delen and Liu, 2017), SVMs try to produce a decision boundary on the hyperplane as it cuts across the data points and separates them in data space. The classifier aims to minimise any errors while classifying the data and to maximise the margin of the decision boundary.

Figure : Decision Boundary for SVM Classifier (Pedregosa et al, 2011)

As shown in Figure 1, the dashed lines denote the margin of the decision boundary, and the middle solid line is the decision line or hyperplane. The two brown dots on the dashed lines and the blue dot on the dashed line are what are called the support vectors from which the name of the model is derived.

In this discussion, we have made it clear that the SVMs solve problems of linear and non-linear data. This is made possible using Kernels. These enable the SVM to be mapped to a feature space. This is achieved using the mapping function which can be denoted by (Hanyue He, 2022). The non-linear SVM makes use of convex quadratic programming to solve problems. The formulation of the decision boundary is denoted by the mathematical formula below:

(Equation 2)

This ML algorithm was proposed by the scientist Vapnik in 1992. Since the introduction of this classifier, we have seen different variations of the same model being applied in different areas. According to Cortes and Vapnik (1995), the classifier has performed better than most of the learning models available. We can note that (Hanyue He, 2022) has also used two variations of the SVM model and compared them against each other in the analysis and prediction of bank customer churn.

It is key to note that after applying a 50:50 resampling split, the algorithm had the highest recall metric score which was a staggering 73.24%. The authors concluded that class sampling of the dataset is key to the model's performance (Kaur. I and Kaur. J., 2020; Karvana et al., 2019). We can see that the SVM classifier performs better when the data is balanced. In their report, Piri, Delen and Liu (2017) define the performance of the model when the data is imbalanced to have a “*dramatic deterioration*” in performance. This is why these authors, (Karvana et al., 2019; Wu and Chang, 2003) agree on the need to have balanced data split when training the model. In the next section, I will touch on the approaches we wish to employ because the current literature agrees with balancing the data, but other techniques can be used, and we will evaluate these techniques.

# **3. Methodology**

In this investigation, we are going to be using quantitative or objective methods to conduct the research whilst answering the posed RQ. We want to find out how much the performances of the chosen models will differ considering the metrics of evaluation that will be discussed further in the report. The justification for the use of quantitative methods is that we will be using numerical figures to evaluate our models using metrics such as F1-score, Recall, AUC etc.

## 3.1. Data Selection

The project plan was to use only one dataset for the experiment. This was the telecommunications dataset obtained from Kaggle. However, after one supervisor meeting, the notion of utilising two datasets i.e. from the banking and marketing sector and the telecommunications sector was adopted. The initial dataset used during the interim progress report was a dataset from Kaggle and it had 70,000 instances with 172 attributes (see Appendix 6). The column names were encoded, and a dictionary was attached to the folder for reference. The majority of the data was encoded to binary (refer to Appendix 6) and there were missing values as well. The initial dataset used in the IPR was not used because two datasets were used instead, and they have a reasonable amount of data that can be modelled and trained on.

## 3.2. Datasets

### 3.2.1. Banking and Marketing Dataset

This dataset was obtained from the UCI Machine Learning Repository. The dataset can be easily accessed from the [link](https://archive.ics.uci.edu/dataset/222/bank+marketing) embedded in this document. According to the creators of this dataset Moro, Rita and Cortez (2012), the dataset was collected from the data extracted by a Portuguese banking institution. The initial plan was to collect data to be used for classification purposes. The dataset repository consisted of at least four datasets with data from May 2008 up to November 2010. The goal of this dataset was to use ML algorithms to predict whether a client would subscribe to a term deposit. However, this dataset has been cited by many researchers for different investigations and experiments than what it was originally set up for. We will be using the dataset to predict and analyse customer churn.

#### Exploratory Data Analysis (EDA)

We know the metadata for the dataset and that means we can begin the exploration of our data. In this investigation, the Pandas library was utilised to allow us to read the data which is in a Comma Delimited Value (CSV) format. The relevant code to load this dataset is in the appendices (refer to Appendix 6). We would like to understand the data types we have, the unique values we have and the shape of the data. After the data is loaded into the Data Frame the df.shape code snippet resulted in us knowing that the data has 45,211 instances and 17 features.

After loading the data, we want to know what the data looks like and how many unique values appear in a column. So the best option is to use the df.head() and that resulted in Figure 4 as shown below;

A table with black text

Description automatically generatedfigure 4

We are aware of the shape of the data and have an idea of what it looks like. But what does the data mean? It is ideal to know the units of measurement and what the column names mean. We also want to know how many unique values appear in each feature. The table below shows the summary of each feature and the unit of measurement.

|  |  |  |
| --- | --- | --- |
| Column Name | Data type | Description |
| age | Integer | The length of time the customer has lived |
| job | Categorical | If someone is employed or not or a student |
| marital | Categorical | If the client is widowed, single or married, etc |
| education | Categorical | How educated the client is |
| default | Binary | Does the customer have any credit? |
| balance | Floating point | Yearly mean balance in Euros |
| loan | Binary | Does the customer have any personal loans active? |
| housing | Binary | Is the customer on any housing loan? |
| contact | Categorical | How is the customer contacted by the bank? |
| day | Integer | Last day customer was contacted |
| month | Categorical | Last month client was contacted |
| duration | Integer | Time the call lasted for in seconds. |
| campaign | Integer | Number of contacts made during the campaign |
| pdays | Integer | Number of days since the last contact |
| previous | integer | Number of contacts made prior |
| poutcome | Categorical | Was the campaign successful or not |
| Y (desired target) | Binary | The desired target |

Table : Feature description of the bank and marketing dataset.

According to Moro, Rita and Cortez (2012), features like duration were measured in seconds and it affects our desired target. For instance, when the duration is zero value then we know the output is going to be no. The prediction of the model is most likely not going to be accurate and therefore we can dismiss the duration as part of the potential predictor. However, we will be employing a correlation heatmap to find any highly correlated features, and these could potentially be our predictors.

According to the data, as shown in Figure 5, we can see the proportion of the customers who churned against non-churners. In a real-world setup, there are going to be fewer customers who churn. The data observed under the exploratory analysis backs that claim as we have 5, 289 which is 11.7% of customers who churned and 39,922 customers which is 88.7% of customers who did not churn after we applied equation 1. However, there exists a data imbalance between the two classes which will be discussed in the pre-processing section in detail.

A blue and orange pie chart

Description automatically generated

Figure 5

### 3.2.2. Telco Dataset

Telco dataset was obtained from Kaggle a public repository which has several datasets. The link to the repository is available [here](https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data). Originally this dataset belonged to IBM sample datasets and has been available online since 2017 and has seen several citations and references of this dataset in the data science community. The main aim of this dataset was for the development of customer retention models to be used to analyse customer behaviour and to build predictive models accordingly.

The repository contains one dataset with 7,043 instances and 21 features. The column labelled churn is our desired target column for classification. The rows of the data represent an individual’s attributes. The dataset can be divided into three categories which are demographics, customer information and services. Services tell us which services the customer is subscribed to for instance streaming, tech support etc. The demographics tell us about information like gender, age etc. The customer information covers the contracts, payment information etc.

Exploratory Data Analysis

## 3.3. Data Pre-Processing

# **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

A close-up of a letter

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

The above Figure 2 shows 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 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 shows the loaded libraries and the head of the dataset. Also, we can see the shape of the dataset.

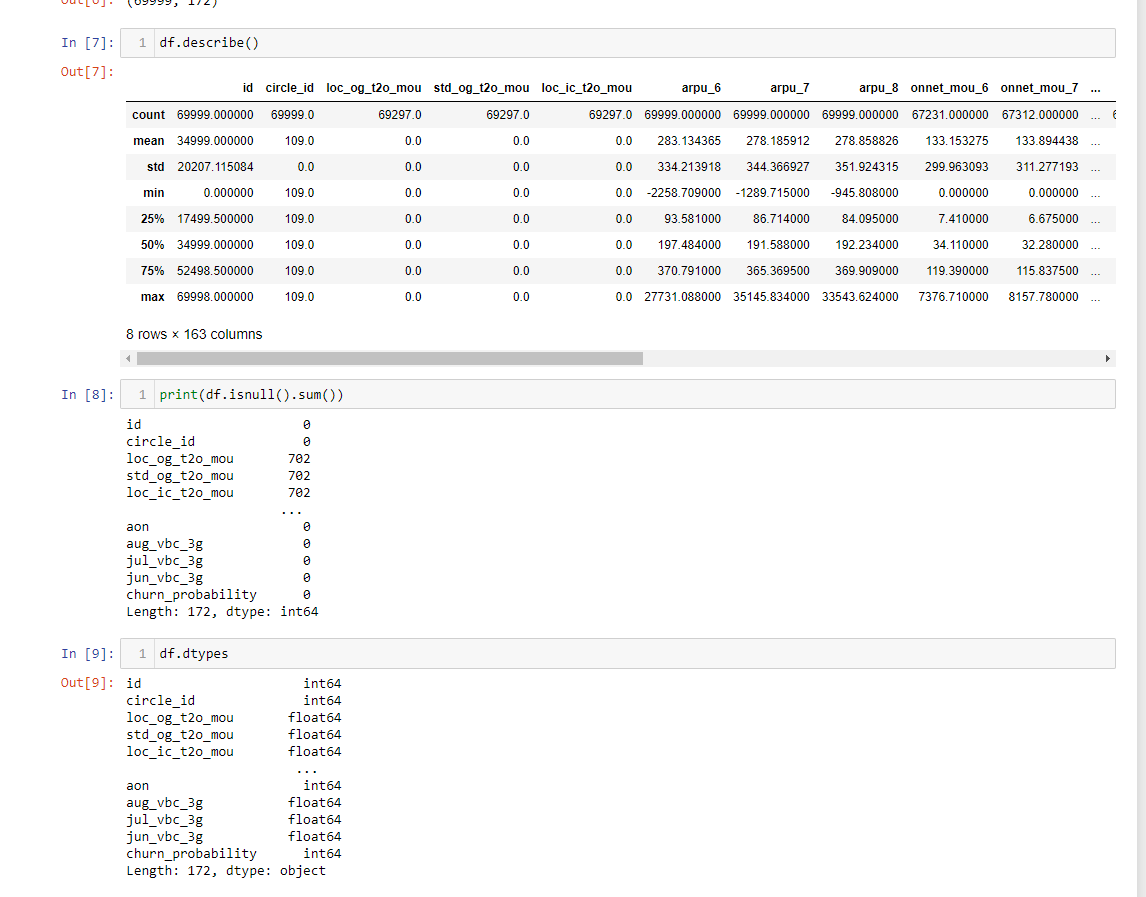


Figure 4 shows the summary of the dataset, and we can get to see what it looks like and the types of data we have in our dataset.

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

As we can see Figure 5 shows 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.