**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Mijail Fausto Blanco Vargas (2023012)  Emily Cristina Herbas Luizaga (2023100) |
| **Student Number:** | 2023012  2023100 |
| **Assessment Due Date:** | 15/11/2023 |
| **Date of Submission:** | 12/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

TELECOM CHURN

Mijail Fausto Blanco Vargas – 2023012

Emily Cristina Herbas Luizaga - 2023100

Content

[1. Introduction 5](#_Toc150547005)

[2. Motivation 5](#_Toc150547006)

[3. Business Understanding 5](#_Toc150547007)

[4. Business Description 5](#_Toc150547008)

[4.1 Research Question 5](#_Toc150547009)

[4.2 General Goal 5](#_Toc150547010)

[4.3 Success criteria/indicators 5](#_Toc150547011)

[4.4 Causes of Customer Churn 5](#_Toc150547012)

[4.5 Types of Customer Churn 6](#_Toc150547013)

[5. Technologies Used 6](#_Toc150547014)

[5.1 Machine Learning Models and Algorithms 6](#_Toc150547015)

[5.2 Libraries 6](#_Toc150547016)

[5.3 Methodology used for Project Management 6](#_Toc150547017)

[6. Data 6](#_Toc150547018)

[6.1 Acomplishment Data 6](#_Toc150547019)

[6.2 Source 6](#_Toc150547020)

[6.3 Data Dictionary 6](#_Toc150547021)

[6.4 Characterization of the Dataset 7](#_Toc150547022)

[6.4.1 Attributes 7](#_Toc150547023)

[6.4.2 Dimensions 7](#_Toc150547024)

[6.4.3 Descriptive Statistics 8](#_Toc150547025)

[6.5 Data Visualization 9](#_Toc150547026)

[6.5.1 Churn Vs Gender 9](#_Toc150547027)

[6.5.2 Churn vs Partner 9](#_Toc150547028)

[6.5.3 Churn vs Dependents 9](#_Toc150547029)

[6.5.4 Churn vs Phone Service 10](#_Toc150547030)

[6.5.5 Churn vs Multiple Lines 11](#_Toc150547031)

[6.5.6 Churn vs Internet Service 11](#_Toc150547032)

[6.5.7 Churn vs Online Security 11](#_Toc150547033)

[6.5.8 Churn vs Online Backup 12](#_Toc150547034)

[6.5.9 Churn vs Device Protection 13](#_Toc150547035)

[6.5.10 Churn vs Device Tech Support 13](#_Toc150547036)

[6.5.11 Churn vs Streaming TV 14](#_Toc150547037)

[6.5.12 Churn vs Streaming Movies 14](#_Toc150547038)

[6.5.13 Churn vs Contract 15](#_Toc150547039)

[6.5.14 Churn vs Paperless Billing 15](#_Toc150547040)

[6.5.14 Churn vs Payment Method 15](#_Toc150547041)

[6.5.15 Tenure 17](#_Toc150547042)

[7. Data Cleaning and Feature Engineering 17](#_Toc150547043)

[7.1 Analysis of Null values 17](#_Toc150547044)

[7.2 Dropping duplicates and columns 18](#_Toc150547045)

[7.3 Integer encoding 18](#_Toc150547046)

[8. Correlation Analysis 19](#_Toc150547047)

[8.1 Correlation Matrix 19](#_Toc150547048)

[8.2 ANOVA Test for Correlation Analysis 20](#_Toc150547049)

[8.3 Chi-squeared Test for Correlation Analysis 21](#_Toc150547050)

[9. PCA 22](#_Toc150547051)

[10. Modeling 22](#_Toc150547052)

[10.1 Base Line Modelling 22](#_Toc150547053)

[10.2 Model Tuning 24](#_Toc150547054)

[10.3 Model Tuning balancing the data 25](#_Toc150547055)

[10.4 Hyperparameter Tuning 26](#_Toc150547056)

[10.4.1 Hyperparameter tuning with Logistic Regression (LR) 26](#_Toc150547057)

[10.4.2 Hyperparameter tuning with Linear Discriminant Analysis (LDA 27](#_Toc150547058)

[10.4.3 Hyperparameter tuning with KNeighbors Classifier (KNN) 27](#_Toc150547059)

[11. Results 28](#_Toc150547060)

[12. Conclusion 28](#_Toc150547061)

[13. Appendix 28](#_Toc150547062)

[14. References 28](#_Toc150547063)

[11. Bibliography 28](#_Toc150547064)

# Introduction

Churn is a common problem in the telecommunications business and refers to the percentage of customers who cancel or do not renew their contract with the company in a given period and it is a very important indicator since it is much more expensive to attract new customers than to retain existing ones, so the analysis of company data can help predict and prevent churn.

To understand why a customer churn from a company is very important since we always need to improve inside the company, offer deals, see what is the reason they are leaving our company; considering that each client brings profit to a company if they use our services and if we give a good service, we keep clients happy with us.

The dataset to be analyzed contains information about the services that our customers get in our company like phone service, internet service, online security, among other variable which will help us determine if a customer is going to churn or not.

# Motivation

The principal motivation for this analysis is to apply what we have been learning in Strategic Thinking into the analysis of Telecom Churn,

Making an analysis and prediction of future churn cases would greatly assist this company in making short-term and long-term decision to prevent the growth of churn cases.

# Business Understanding

To predict Chun cases, we are going to implement different Machine Learning Models which were applied last semester, but also, we are going to improve the models analyzing the recall and applying hyperparameters to have certain results in which we can trust to make good predictions.

# Business Description

## Research Question

Build 3 different Machine Learning models with different test and train splits to predict the customers that are going to churn in the company.

## General Goal

To predict churn cases using as a baseline the models used last semester, improving the analysis of the Machine Learning models to be applied for this prediction.

## Success criteria/indicators

We are going to apply different Machine Learning models improving them using grid search, hyperparameters focusing in the improvement of the accuracy and specially the recall to have good predictions of churned cases in the company.

## Causes of Customer Churn

Customer churn (or customer attrition) refers to the loss of customers or subscribers for any reason at all. Businesses measure and track churn as a percentage of lost customers compared to total number of customers over a given time period. This metric is usually tracked monthly and reported at the end of the month. It's important to note that churn rates vary by industry and knowing your market is key to reducing churn with more precision. (www.paddle.com, n.d.)

* Price: If customers find a more cost-effective solution to the problem they want to solve, they may churn.
* Product/Market Fit: When the client realizes that they cannot achieve their goals with our solution.
* User Experience: If the user experience with the product or application is buggy, and glitchy, for them, they will be less likely to use it on a regular basis and build expertise with it.
* Customer experience – If a customer's experience connecting with other aspects of the company, such as customer service, executives, technical support, and installation service, is not positive, the likelihood of churn could increase.

## Types of Customer Churn

* Competitor Intervention: maybe the competition has better deals or the network has a greater reach.
* Unsuccessful Onboarding: when executives focus only on the sale and not on the right solution for the client.
* Desired Feature or Functionality: when we offer all customers the same product, and we do not understand that the product must be adapted to the customer and not the customer to the product.

# Technologies Used

## Machine Learning Models and Algorithms

Different Machine Learning Models for classification were applied since we are trying to predict if a customer will churn or not in our company and for that analysis we are going to apply first in Logistic Regression, Linear Discrimination Analysis, K-Neighbors Classifier, Gaussian NB, MLP Classifier, and Random Forest, and after defining the best 3 models we are going to go indeed improving the models applying different resources learned in Strategic Thinking lessons.

## Libraries

We are going to use different libraries for this analysis for example, pandas, ,numpy, seaborn, spicy.stats, sklearn for different machine learning models and model selection for splitting into train and test, among others.

## Methodology used for Project Management

As Wijaya states (Medium, 2021) The cross-industry standard process for data mining or CRISP-DM is an open standard process framework model for data mining project planning. This is a framework that many have used in many industrial projects and proven successful in the application.

CRISP-DM methodology was implemented in this project and all the steps of the methodology was evaluated in an excel file to monitor the progress of each state like the Business Understanding, Data Understanding, Data Preparation, Modelling, and Results (See Appendix 1).

# Data

## Acomplishment Data

The Churn dataset has different numerical and categorical variables to be analyzed for the model prediction in which we have 21 variables and 7043 rows for its performance.

## Source

The data was taken from Kaggle in the following link: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn> (Kaggle, 2017).

## Data Dictionary

In the original dataset we have different columns and Table 1 shows their description of each one:

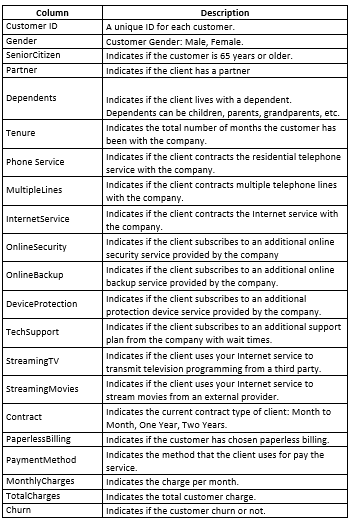


Figure 1: Data Dictionary of Telco Churn Dataset

## Characterization of the Dataset

### 6.4.1 Attributes

In total we have 21 variables in which “Churn” is going to be our target variable and the other 20 are independent variables which will help us build the Models prediction.

### 6.4.2 Dimensions

`

The dimension of Telco Churn Dataset is 7043 rows and 21 columns

### 6.4.3 Descriptive Statistics

Figure 2 shows the statistics of the numerical variables

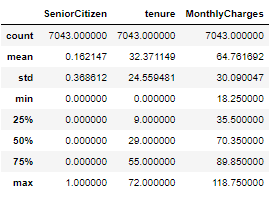


Figure 2 – Statistics of Numerical Values

After analysing the descriptive statistics of the numerical values, we appreciate the number of values we have in each row, mean, standard deviation, minimum and maximum values, and the quartiles divided in 25%, 50%, and 75%, from there we can deduct the following points

* In each column we have 7043 rows.
* The mean of Senior Citizen is 0.16 which column is binary telling us that tends to 0 (which means No) giving us information that the majority of the people are not Senior Citizens, Tenure's mean is 32.37, but we have to note that tenure is measured monthly, and the mean of monthly charges is around 64.76.
* We can appreciate the different standard deviation values which is a measure of dispersion for explaining variability in the dataset.
* we have the min, max and quartile values (25th, 50th and 75th percentiles of the data), denoting that the 50th percentile is the median of the data set.

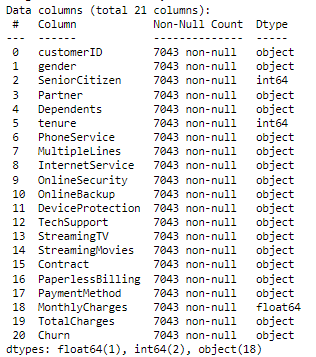


Figure 3 – Data Type of each variable

Figure 3 shows that some variables which would be numerical are shown as object: for example, "TotalCharges" and "OnlineBackup", this will be analyzed in the Data Cleaning part.

## Data Visualization

### 6.5.1 Churn Vs Gender

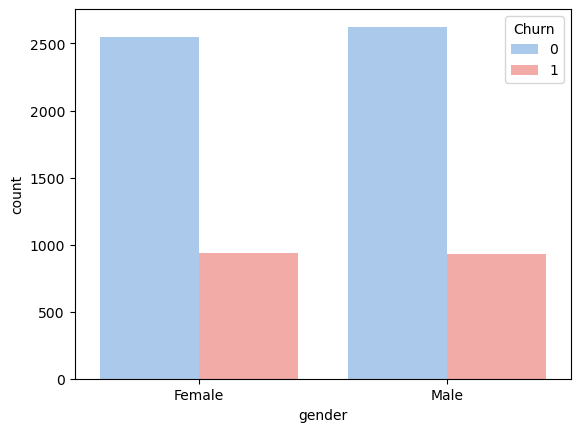


Figure 4 – Churn Vs Gender

We can observe that the company has almost the same quantity of female customers than male customers and in both cases the churn is the same.

Also we this chart we can observe that the number that customers that decided churn is not as big as the clients that don’t decided to churn.

### 6.5.2 Churn vs Partner

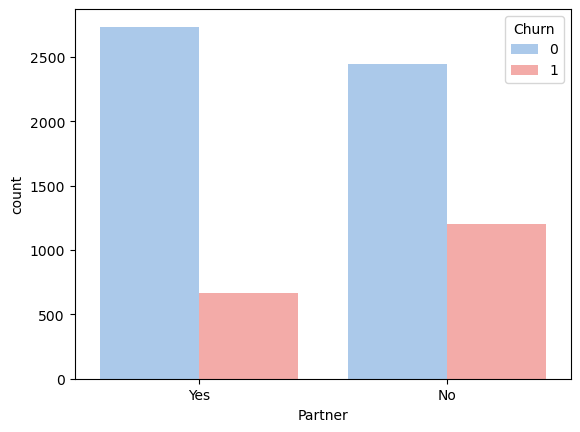


Figure 5 – Churn Vs Partner

We can observe that the company has more customers with partners than those without. The difference between both cases is not significant but customers without partners show a higher tendency to churn.

### 6.5.3 Churn vs Dependents

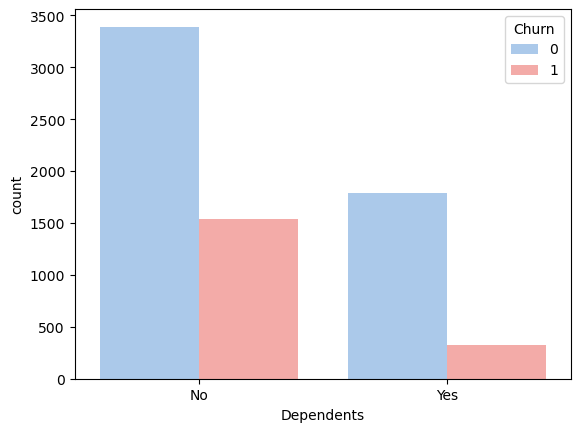


Figure 6 – Churn Vs Dependents

We can observe that the company has more customers without dependents and the level of churn in this case us higher than the customers that have dependents.

This chart could be an opportunity for the company to analyze a solution, perhaps by creating accessible packages where customers with dependents in the service receive advantages or discounts. This approach could help reduce the percentage of customers who churn.

### 6.5.4 Churn vs Phone Service

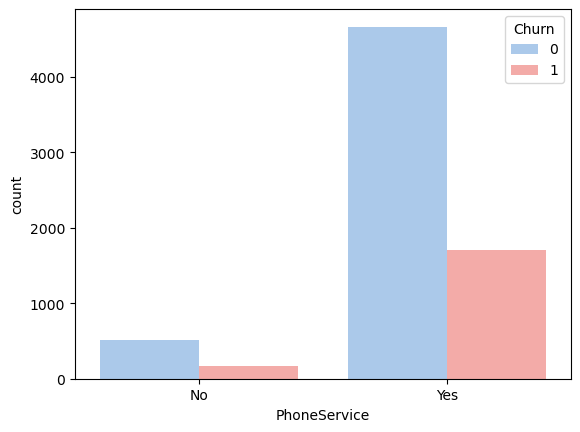


Figure 7 – Churn Vs Phone Service

We can observe the company has more customers with phone Service but also the number of customers who do churn have this service, which means that there is a problem with the service, such as the quality of the product, the quality of the service or the price.

### 6.5.5 Churn vs Multiple Lines

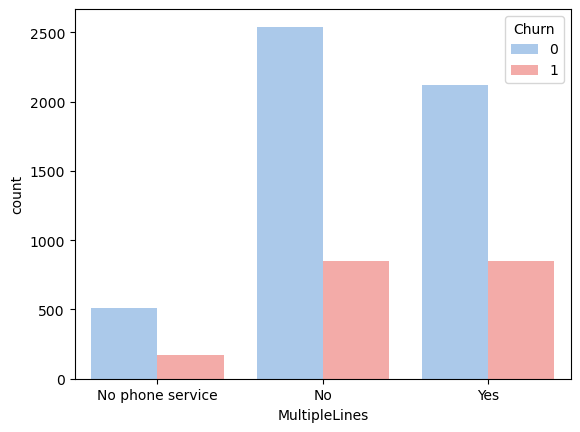


Figure 8 – Churn Vs Multiple Lines

We can observe that client has Multiple Lines or not is indifferent to the churn.

### 6.5.6 Churn vs Internet Service

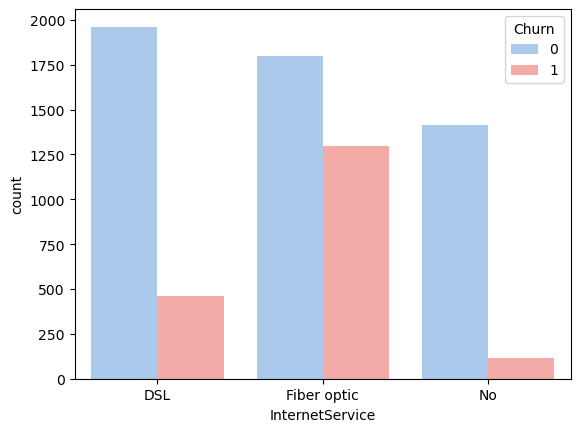


Figure 9 – Churn Vs Internet Service.

We can observe that users with internet service have more cases of churn when they have Fiber Optic service compared to DSL.

### 6.5.7 Churn vs Online Security

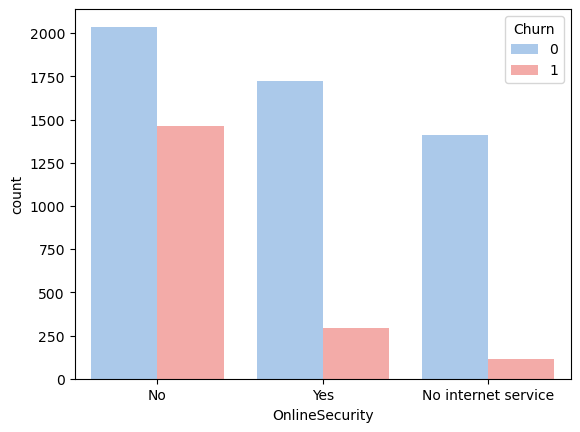


Figure 10 – Churn Vs Online Security

We can see that users without Online Security service are the ones who churned at a higher percentage.

### 6.5.8 Churn vs Online Backup

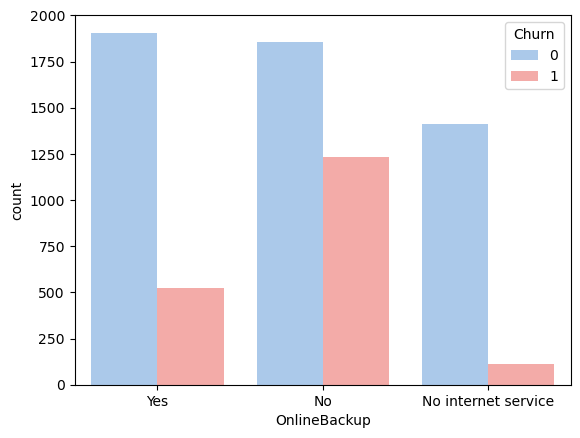


Figure 11 – Churn Vs Online Backup

We can see that users without Online Backup service are the ones who churned at a higher percentage.

### 6.5.9 Churn vs Device Protection

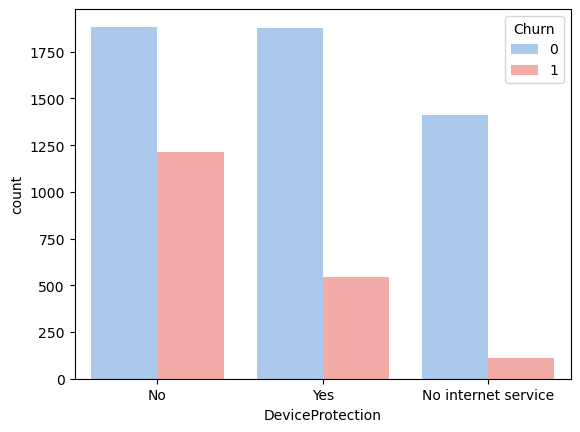


Figure 12 – Churn Vs Device Protection

We can see that users without Device Protection service are the ones who churned at a higher percentage.

### 6.5.10 Churn vs Device Tech Support

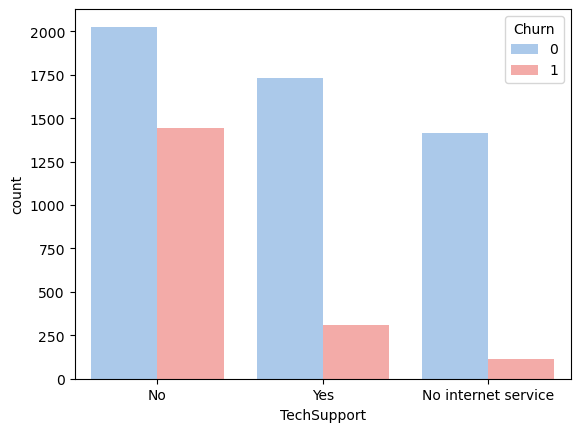


Figure 13 – Churn Vs Tech Support

We can see that users without Tech Support service are the ones who churned at a higher percentage.

### 6.5.11 Churn vs Streaming TV

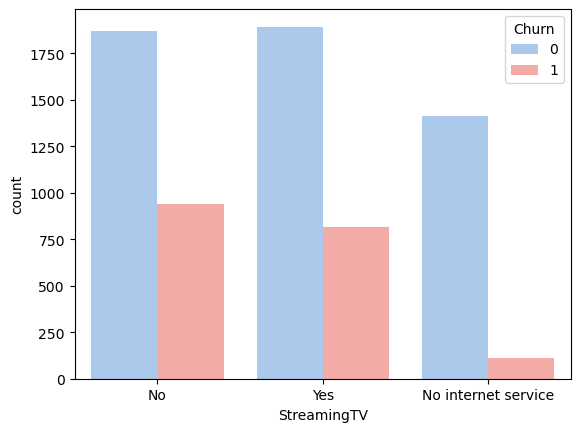


Figure 14 – Churn Vs Streaming TV

We can identify that both the clients that have the service and those that don’t have the service have a close percentage of clients that are within the Churn, so we could indicate that there is a problem with the product, service or its price.

### 6.5.12 Churn vs Streaming Movies

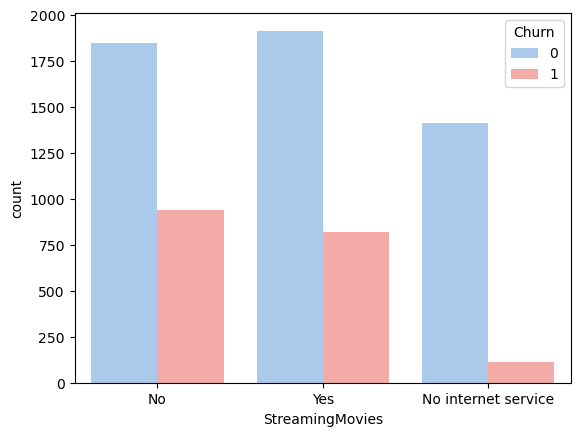


Figure 15 – Churn Vs Streaming Movies

We can identify that both the clients that have the service and those that don’t have the service have a close percentage of clients that are within the Churn, so we could indicate that there is a problem with the product, service or its price.

### 6.5.13 Churn vs Contract

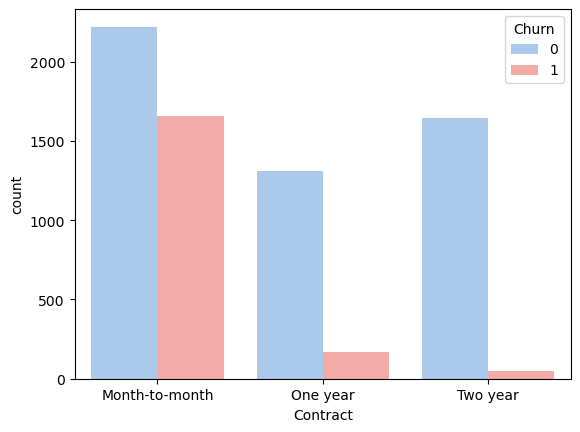


Figure 16 – Churn Vs Contract

We can observe that clients that have a contract Month to Month have highest percentage of churn clients, so we could assume that it is essential for the company to work on a solution to convince these customers to switch from Month-to-Month contracts to longer-term contracts (one year, two years). This is because with short-term contracts, it is easier for customers to churn.

### 6.5.14 Churn vs Paperless Billing

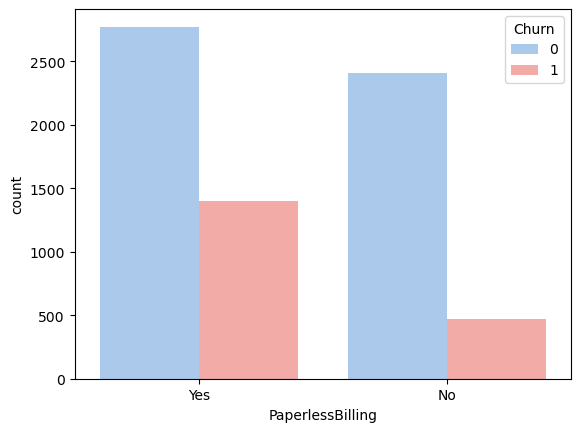


Figure 17 – Churn Vs Paperless Billing

We can see that clients that chose the Paperless Billing Method have highest percentage of churn clients.

### 6.5.14 Churn vs Payment Method

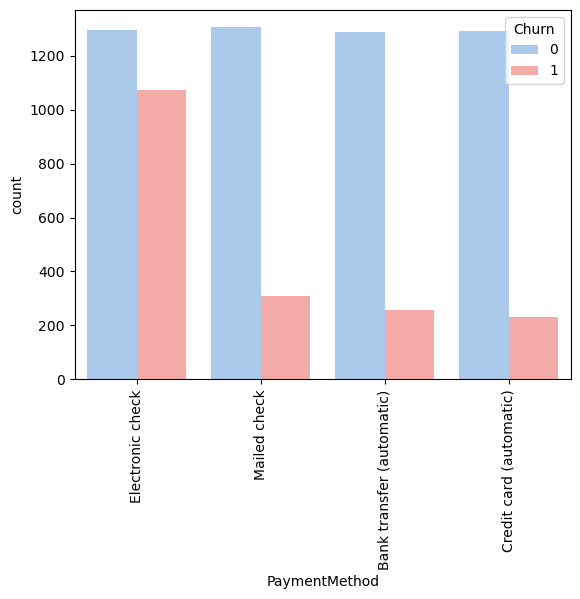


Figure 18 – Churn Vs Payment Method

We can observe that the payment method through Electronic Check presents problems because is with this Payment Method that the company have highest percentage of churn clients.

This payment Method should be reviewed, since it may have service problems, duplicate billing problems or other types of problems that must be identified.

### 6.5.15 Tenure

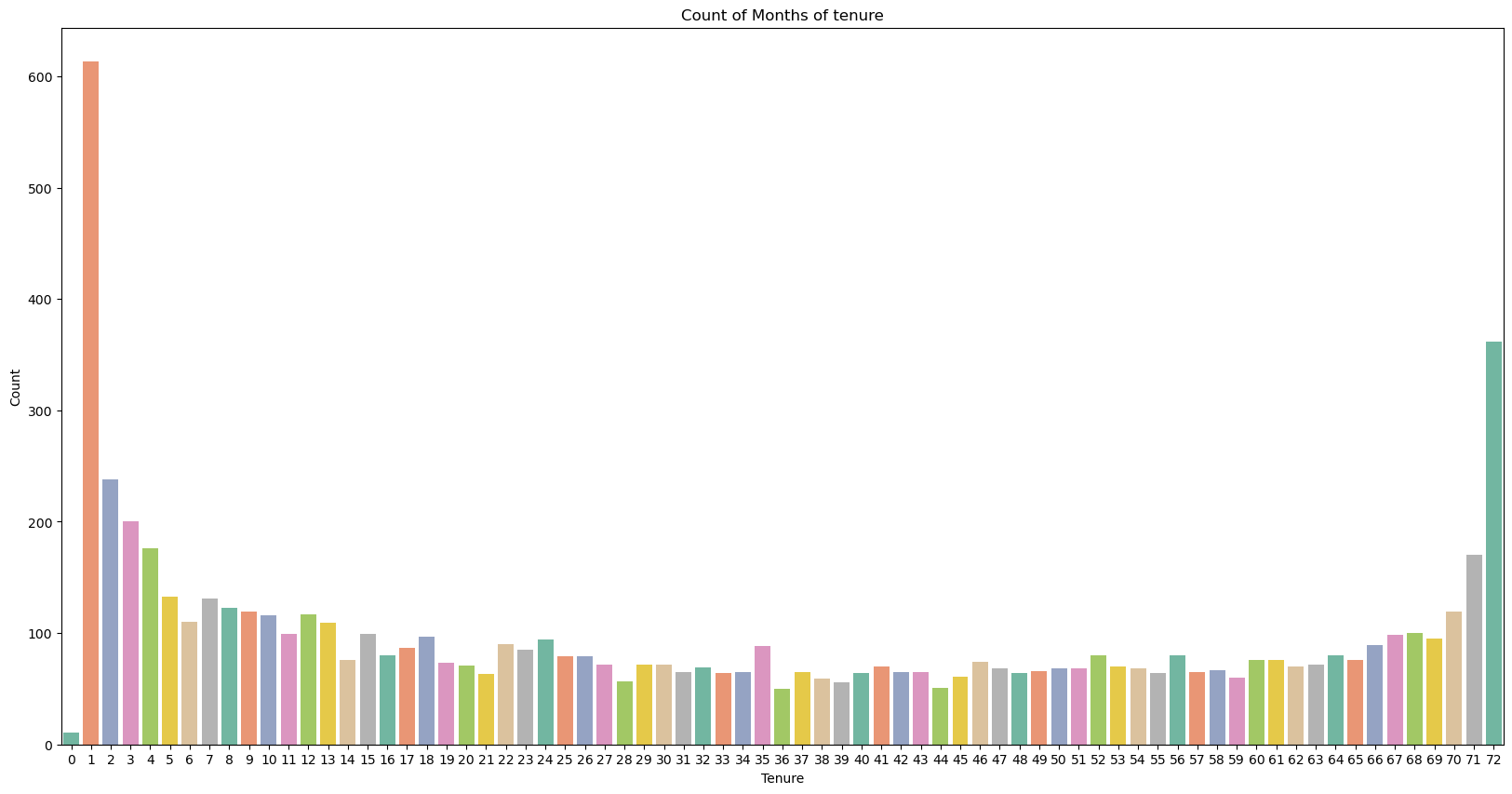


Figure19 – Count of Months of tenure

In this graphic we can see that many customers that stayed with us one month doesn’t get our products for the second month and that is very important to analyse for keeping the customers in the future and making them loyal. On the other hand, we can see the last bar in the bar pot that many customers stayed with us for a long time making them loyal inside our company.

# Data Cleaning and Feature Engineering

## Analysis of Null values

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. (www.ibm.com, 2017)

We started started standardizing null values and we go the next results:

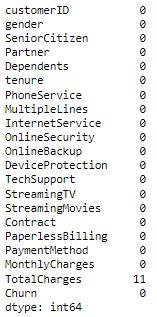


Figure 30 – Number of “Nan” Values in each column.

We can see that 11 “Nan” values of “TotalCharges” column have missing values and to deal with them; the 11 values were replaced by the multiplication of “tenure” and “MonthlyCharges”. We have to notice that this column was detected as an object and it was because of the existence of the null values and to proceed with the analysis as a return we converted the variable type as float since it is a continuous numerical value that tell us how much a customer paid while he is using our services.

In last semester analysis, through the experimentation there was a mistake replacing the values of “Yes” and “No” for numerical values inside the column “OnlineBackup” in which they didn’t replace “No internet service”, and that is why they had 1526 “Nan” values inside this company. However, exploring the column, we detected that these values will be represented as another number.



Figure 31 – Value count of the column “Online Backup”.

In figure 31, 0 means internet service without online backup, 1 represents internet service with online backup, and No internet service was analysed merging it to “internet service without online backup”, but we decided to represent this category with the number 2 which means “No internet service in this column.

## Dropping duplicates and columns

Duplicate data can skew prediction results. Thus, for columns that should contain unique values, it’s important to search for and exclude any duplicate rows to achieve a more general and accurate prediction. (Mage, n.d.)

When we dropped the duplicates, the shape of the data didn’t change (4073 rows, 21 columns) that tell us that we don’t have duplicate values inside our dataset.

The variable “customerID” is a unique value for each customer which gives them a unique representation, but this variable is not relevant for the application of Machine Learning models trying to classify the customers that churn or not and that is why we are going to drop this column.

## Integer encoding

Integer encoding consist in replacing the categories by digits from 1 to n (or 0 to n-1, depending the implementation), where n is the number of distinct categories of the variable. (kunwar, 2020)

We decided to apply integer encoding to represent categorical variables into numerical ones and then being able to continue analyzing the machine learning models, all the values are represented in the next table:

|  |  |
| --- | --- |
| **Variables** | **Number and description** |
| gender | 0:Male |
| 1: Female |
| Partner | 0:No |
| 1: Yes |
| Dependents | 0:No |
| 1: Yes |
| PhoneService | 0:No |
| 1: Yes |
| MultipleLines | 0:No |
| 1: Yes |
| 2: No phone service |
| InternetService | 0:No |
| 1: DSL |
| 2: Fiber optic |
| OnlineSecurity | 0:No |
| 1: No internet service |
| 2: yes |
| DeviceProtection | 0:No |
| 1: No internet service |
| 2: yes |
| TechSupport | 0:No |
| 1: No internet service |
| 2: yes |
| StreamingTV | 0:No |
| 1: No internet service |
| 2: yes |
| StreamingMovies | 0:No |
| 1: No internet service |
| 2: yes |
| Contract | 0: Month-to month |
| 1: Two year |
| 2: One year |
| PaperlessBilling | 0:No |
| 1: Yes |
| PaymentMethod | 0: Electronic check |
| 1: Mailed check |
| 2: Bank transfer (automatic) |
| 3: Credit card (automatic) |
| Churn | 0:No |
| 1: Yes |

Figure 32 – Representation of categorical into numerical values.

# Correlation Analysis

## Correlation Matrix

Bruce, Gedeck. and Bruce. (2020, p.30) believe that exploratory data analysis in many modelling projects (whether in data science or in research) involves examining correlation among predictors and between predictors and a target variable. Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated

Now that all our values are numerical in our dataset, we are going to analyze the correlation matrix to have an idea of which variables influence for the prediction of our target variable “churn”.

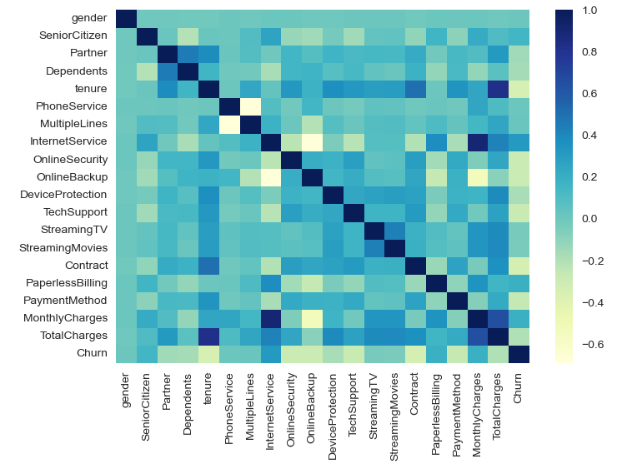


Figure 33 - Correlation Matrix

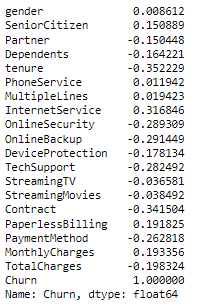


Figure 34 – Correlation of the variables with “churn” variable

According to the matrix of correlation (Figure 33) and the results we have gotten with the target variable "Churn" (Figure 34), we can see that our target variable is very correlated with the column "tenure", "InternetService", "OnlineBackup", among others, but we will apply different tests to see if all the variables are really necessary for our analysis.

## ANOVA Test for Correlation Analysis

According to Andreas C. and Sarah G (2016, p.236)In univariate statistics, we compute whether there is a statistically significant relation‐

ship between each feature and the target. Then the features that are related with the

highest confidence are selected. In the case of classification, this is also known as

analysis of variance (ANOVA)

For selecting the columns to apply ANOVA test, we analyzed the unique values in each column

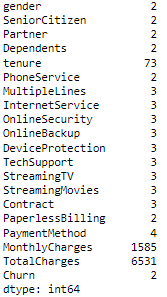


Figure 35 – Unique values in each column

For the columns in which we have less than 30 unique values, we are going to apply Chi-squared test and for the ones that have more or same as 30 unique values, we are going to apply Anova test.

In this case the columns analyzed were 'tenure', 'MonthlyCharges', 'TotalCharges', and the next hypothesis was stated:

H0: There is evidence that the variable is correlated with the target variable "churn"

HA: There is non-evidence that the variable is correlated with the target variable "churn"

The significance level (alpha) stablished by default is 0.05



Figure 36 – Results of ANOVA test

According to our results the 3 variables are correlated with our target variables since the p-value is less than 0.05.

## Chi-squared Test for Correlation Analysis

According to Medium (Medium, 2019)The Chi Square statistic compares the tallies or counts of categorical responses between two (or more) independent groups. (note: Chi square tests can only be used on actual numbers and not on percentages, proportions, means, etc.) Chi-square Test is a method that is used to test if there is any relationship between two categorical variables.

In this case the columns analyzed were 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', and the next hypothesis was stated:

H0: There is evidence that the variable is correlated with the target variable "churn"

HA: There is non-evidence that the variable is correlated with the target variable "churn"

The significance level (alpha) stablished by default is 0.05

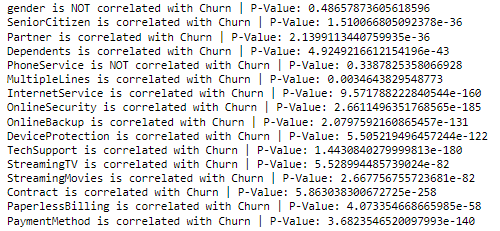


Figure 37 – Results of Chi-squared test

According to the results of this test (Figure 37), we can appreciate that the columns "gender", and "PhoneService" are not correlated because they have a p-value higher than 0.05, and that is why we are going to drop them since they are not relevant for our analysis.

# PCA

As we learn the PCA is used to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.

After applied PCA we could reduce our columns from 21 to 15.

# Modeling

## Base Line Modelling

The target variable “churn” is represented by “y” that is the value we want to predict and all the other variables were considered as independent variables which will help us improve the models.

First, we divided the data into 10%, 20% and 30% splits with a random state of 42.

When we analyze the target variable, we can see that is not balanced



Figure 38 – Counting values in the “Churn” variable

The target variable is not balanced. However, we are going to perform different models and then apply techniques learned in strategic thinking to balance it and improve the models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy (10% test)** | **Accuracy (20% test)** | **Accuracy (30% test)** |
| **Linear Regression (LR)** | 0.801201 | 0.799079 | 0.800813 |
| **Linear Discrimination Analysis (LDA)** | 0.799782 | 0.796240 | 0.800203 |
| **K-Neighbors Classifier(KNN)** | 0.759701 | 0.760029 | 0.761054 |
| **Decision Tree Classifier(CART)** | 0.729250 | 0.716370 | 0.721094 |
| **Gaussian NB (NB)** | 0.788102 | 0.786298 | 0.788439 |
| **MLP Classifier (NN)** | 0.788890 | 0.787541 | 0.783573 |
| **Random Forest Classifier (RF)** | 0.784000 | 0.776361 | 0.777892 |

Figure 39 – Accuracy Results of the Models applying different splits for train and test

As we can see in the previous figure, the best model is logistic regression. However, when we explore indeed the classification report in Logistic Regression, we get the next results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **10% test** | | | **20% test** | | | **30% test** | | |
|  | **Accuracy LR: 0.80** | | | **Accuracy LR: 0.8** | | | **Accuracy LR: 0.8** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **0** | **0.86** | **0.91** | **0.88** | **0.86** | **0.9** | **0.88** | **0.85** | **0.9** | **0.87** |
| **1** | **0.72** | **0.6** | **0.65** | **0.68** | **0.58** | **0.63** | **0.69** | **0.56** | **0.61** |

Figure 40 – Precision, recall and f1-score of Logistic Regression with different splits

In the Classification Report the metrics that we obtained with every model are:

* Precision: that is the ability of the models to correctly identify positive instances, high precision is desirable when minimizing false positives is crucial.
* Recall: that is the ability of the model to capture all positive instances, high recall is desirable when minimizing false positives is crucial.
* F1-Score: balances precision and recall, useful when there is an uneven class distribution or when false positives and false negatives have different consequences.

Although Linear Regression Model has a good accuracy, when we look indeed their precision, recall and f1-score, the values of 1(customers that churned) are not better than 0 (customers that didn’t churn) and specially recall tell us which customers churned the company and their metric tell us the percentage of how many of them were predicted correctly. As shown in the table we have a low recall in the 3 train and test splits.

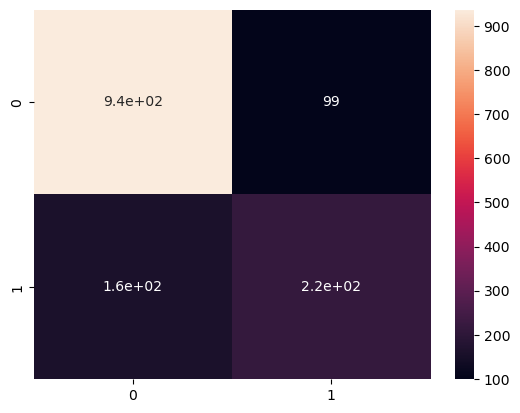


Figure 42 – Initial Confusion Matrix of Logistic Regression test 20% using “Accuracy” as score metric

In Figure (42) we can see that 160 cases of churn were incorrectly predicted from 380 cases in Logistic Regression model were 20% was using for test and 80% was used for training. Those 160 cases churned in reality but they were predicted as false.

## Model Tuning

Model tuning is the experimental process of finding the optimal values of hyperparameters to maximize model performance. Hyperparameters are the set of variables whose values cannot be estimated by the model from the training data. These values control the training process. Model tuning is also known as hyperparameter optimization. (domino.ai, n.d.)

In this part instead of using “accuracy” as our score, we are going to use “recall” since we want to improve the models for predicting the people that are going to churn in our company.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Recall (10% test)** | **Recall (20% test)** | **Recall (30% test)** |
| **Linear Regression (LR)** | 0.531041 | 0.524182 | 0.512283 |
| **Linear Discrimination Analysis (LDA)** | 0.557347 | 0.551650 | 0.547506 |
| **K-Neighbors Classifier(KNN)** | 0.479523 | 0.483889 | 0.495128 |
| **Decision Tree Classifier(CART)** | 0.482039 | 0.492889 | 0.465911 |
| **Gaussian NB (NB)** | 0.568372 | 0.566606 | 0.555909 |
| **MLP Classifier (NN)** | 0.515172 | 0.487946 | 0.501026 |
| **Random Forest Classifier (RF)** | 0.443114 | 0.434240 | 0.421270 |

Figure 39 – Recall Results of the Models applying different splits for train and test

Now the best models using “recall” as our score metric are Gaussian NB, Linear Regression and Linear Discrimination Analysis. Let’s take a look to the precision, recall and F1 score inside Gaussian NB model.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **10% test** | | | **20% test** | | | **30% test** | | |
|  | **Accuracy LR: 0.80** | | | **Accuracy LR: 0.8** | | | **Accuracy LR: 0.8** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **0** | **0.85** | **0.88** | **0.87** | **0.86** | **0.87** | **0.86** | **0.85** | **0.87** | **0.86** |
| **1** | **0.66** | **0.59** | **0.62** | **0.63** | **0.61** | **0.62** | **0.62** | **0.59** | **0.61** |

Figure 41 – Precision, recall and f1-score of Gaussian NB model with different splits

In the previous table we can see that recall improved; however, we can still do it better balancing the data

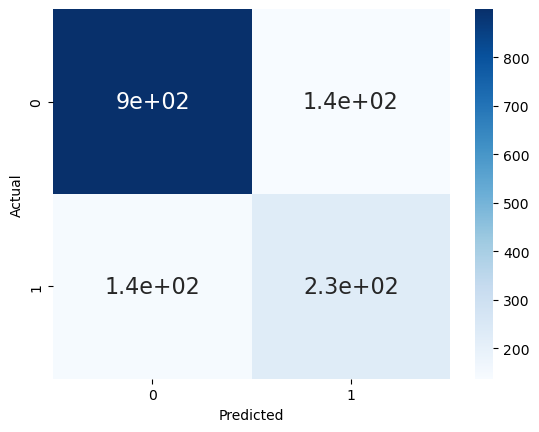


Figure 42 – Confusion Matrix of Logistic Regression test 20% using “Recall” as a score metric

The model has improved a little bit since now we have 140 cases that were wrong predicted from around 370 cases reducing from the last confusion matrix analysed previously where we had 160 wrong predicted churned cases that were true but the model predicted as false

## Model Tuning balancing the data

In this part we decided to equally distribute the data in which each sample will have 1675 cases of churn

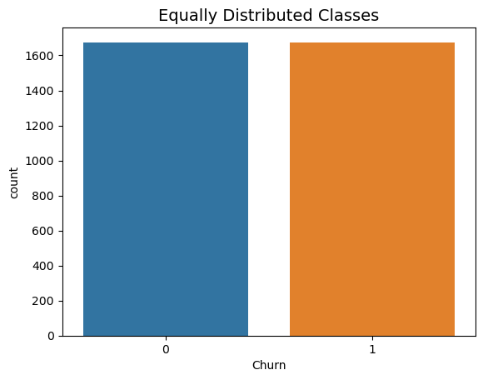


Figure 42 – New equal distribution of churned cases.

Now that the data is balanced and applying again all the previous models, we got the next results using “accuracy” as score metric:

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy (10% test)** | **Accuracy (20% test)** | **Accuracy (30% test)** |
| **Linear Regression (LR)** | 0.756425 | 0.759693 | 0.755985 |
| **Linear Discrimination Analysis (LDA)** | 0.754038 | 0.758022 | 0.750965 |
| **K-Neighbors Classifier(KNN)** | 0.719708 | 0.723608 | 0.726641 |
| **Decision Tree Classifier(CART)** | 0.665697 | 0.661109 | 0.658301 |
| **Gaussian NB (NB)** | 0.747169 | 0.748337 | 0.740927 |
| **MLP Classifier (NN)** | 0.740006 | 0.736636 | 0.740541 |
| **Random Forest Classifier (RF)** | 0.725680 | 0.725279 | 0.737066 |

Figure 43 – Accuracy Results of the Models applying different splits for train and test

Now our models are ready to be analyzed indeed and continue making improvements, we can see that Linear Regression, Linear Discrimination Analysis and KNN and Gaussian NB Model. However, we are going to use KNN as last model to compare since we want to apply different hyperparameters in the modelling part.

## Hyperparameter Tuning

The hyperparameter-tuning technique is use to optimize the accuracy of our model and helps to prevent overfitting or underfitting of the model, for this we applied the hyperparameter tuning with different models for 10%, 20% and 30% of testing, these models were selected for having the best accuracy at the time of modeling.

### 10.4.1 Hyperparameter tuning with Logistic Regression (LR)

Logistic Regression includes a regularization term that prevents overfitting. Tuning the regularization parameter helps find the right balance between fitting the training data well and avoiding overcomplex.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Hyperparameter Tuning** | | |
|  | **Logistic Regression** | | |
|  | **10%** | **20%** | **30%** |
| **Accuracy** | 0,752 | 0,759 | 0,773 |

Figure 44 – Accuracy Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression (LR)** | | | | | | | | |
|  | **10% test** | | | **20% test** | | | **30% test** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **0** | 0,73 | 0,75 | 0,74 | 0,79 | 0,75 | 0,77 | 0,75 | 0,8 | 0,78 |
| **1** | 0,77 | 0,75 | 0,76 | 0,73 | 0,77 | 0,75 | 0,79 | 0,77 | 0,77 |

Figure 45 – Classification Report

10.4.2 Hyperparameter tuning with Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a supervised dimensionality reduction technique used for classification task. The goal of LDA is to project the features in higher dimensional space.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Hyperparameter Tuning** | | |
|  | **Linear Discriminant Analysis** | | |
|  | **10%** | **20%** | **30%** |
| **Accuracy** | 0,743 | 0,742 | 0,772 |

Figure 46 – Accuracy Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Linear Discriminant Analysis (LDA)** | | | | | | | | |
|  | **10% test** | | | **20% test** | | | **30% test** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **0** | 0,73 | 0,75 | 0,74 | 0,78 | 0,72 | 0,75 | 0,76 | 0,78 | 0,77 |
| **1** | 0,77 | 0,75 | 0,76 | 0,71 | 0,77 | 0,74 | 0,78 | 0,77 | 0,77 |

Figure 47 – Classification Report

### 10.4.3 Hyperparameter tuning with KNeighbors Classifier (KNN)

As we know the KNN is a non-parametric method used for regression and classification problems, in which the most important hyperparameter is the number of neighbours

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Hyperparameter Tuning** | | |
|  | **Kneigbours Classifier** | | |
|  | **10%** | **20%** | **30%** |
| **Accuracy** | 0,746 | 0,741 | 0,727 |

Figure 48 – Accuracy Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Kneigbours Classifier** | | | | | | | | |
|  | **10% test** | | | **20% test** | | | **30% test** | | |
|  | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** | **precision** | **recall** | **f1-score** |
| **0** | 0,74 | 0,72 | 0,73 | 0,8 | 0,68 | 0,74 | 0,73 | 0,7 | 0,72 |
| **1** | 0,75 | 0,77 | 0,76 | 0,69 | 0,81 | 0,75 | 0,72 | 0,75 | 0,74 |

Figure 49 – Classification Report

# Results

First, we are going to see the summary table of results of the capstone project applied last semester.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **10%** | | **20%** | | **30%** | |
|  |  |  | **0** | **1** | **0** | **1** | **0** | **1** |
| **Logistic Regression** | **Without Technique** | **precision** | 0,86 | 0,65 | 0,87 | 0,63 | 0,86 | 0,63 |
| **recall** | 0,9 | 0,55 | 0,89 | 0,58 | 0,88 | 0,58 |
| **f1-score** | 0,88 | 0,6 | 0,88 | 0,6 | 0,87 | 0,6 |
| **With Smote** | **precision** |  |  | 0,9 | 0,55 |  |  |
| **recall** |  |  | 0,8 | 0,73 |  |  |
| **f1-score** |  |  | 0,85 | 0,62 |  |  |
| **With NearMiss** | **precision** |  |  | 0,9 | 0,48 |  |  |
| **recall** |  |  | 0,73 | 0,76 |  |  |
| **f1-score** |  |  | 0,81 | 0,59 |  |  |
| **With PCA** | **precision** |  |  | 0,86 | 0,62 |  |  |
| **recall** |  |  | 0,88 | 0,57 |  |  |
| **f1-score** |  |  | 0,87 | 0,6 |  |  |

Figure 50 – Last Semester Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **10%** | **20%** | **30%** |
| **Accuracy- Logistic Regression** | **Without Technique** | 0,82 | 0,81 | 0,81 |
| **Smote** |  | 0,78 |  |
| **NearMiss** |  | 0,74 |  |
| **PCA** |  | 0,81 |  |

Figure 51 – Last Semester Accuracy Results

Last semester Emily’s group analysed different models defining Logistic Regression as the best model because of its accuracy and they applied with this model Smote, Nearmiss, and PCA; all of them separated from each other and PCA was the best model with an accuracy of 81% and the values of recall were high 0.86 when they analysed the 0 (customers that didn’t churn), but in the other side we had a recall that was 0.57 for the value of 1 (customers that churned the company). This model was good; however, we decided to make changes to get new results to improve the model with the base that we want to improve the prediction to know the customers that are going to churn in the company

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **10%** | | **20%** | | **30%** | |
|  |  | **0** | **1** | **0** | **1** | **0** | **1** |
| **PCA and Standard Scaler** | **precision** | 0,86 | 0,72 | 0,86 | 0,68 | 0,85 | 0,69 |
| **recall** | 0,91 | 0,6 | 0,9 | 0,58 | 0,9 | 0,56 |
| **f1-score** | 0,88 | 0,65 | 0,88 | 0,63 | 0,87 | 0,61 |
| **Recall** | **precision** | 0,85 | 0,66 | 0,86 | 0,63 | 0,85 | 0,62 |
| **recall** | 0,88 | 0,59 | 0,87 | 0,61 | 0,87 | 0,59 |
| **f1-score** | 0,87 | 0,62 | 0,86 | 0,62 | 0,86 | 0,61 |
| **Hyperparameter Tuning LR** | **precision** | 0,73 | 0,77 | 0,79 | 0,73 | 0,75 | 0,79 |
| **recall** | 0,75 | 0,75 | 0,75 | 0,77 | 0,8 | 0,75 |
| **f1-score** | 0,74 | 0,76 | 0,77 | 0,75 | 0,78 | 0,77 |
| **Hyperparameter Tuning LDA** | **precision** | 0,73 | 0,77 | 0,78 | 0,71 | 0,76 | 0,78 |
| **recall** | 0,75 | 0,75 | 0,72 | 0,77 | 0,78 | 0,77 |
| **f1-score** | 0,74 | 0,76 | 0,75 | 0,74 | 0,77 | 0,77 |
| **Hyperparameter Tuning KNN** | **precision** | 0,74 | 0,75 | 0,8 | 0,69 | 0,73 | 0,72 |
| **recall** | 0,72 | 0,77 | 0,68 | 0,81 | 0,7 | 0,75 |
| **f1-score** | 0,73 | 0,76 | 0,74 | 0,75 | 0,71 | 0,74 |

Figure 52 – Final Results Improved models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **10%** | **20%** | **30%** |
| **Accuracy** | **PCA/Standard Scaler** | 0,82 | 0,82 | 0,81 |
| **LR** | 0,75 | 0,76 | 0,77 |
| **LDA** | 0,75 | 0,74 | 0,77 |
| **KNN** | 0,75 | 0,74 | 0,73 |

Figure 53 – Accuracy Results Improved models

This is the summary tables of results for the prediction, we decided to apply PCA and standard Scaler, then evaluate the models according to the accuracy and then recall comparing them with the base that we want to improve recall; that is how we decided to balance the data equally implementing distributed Balance Tuning and we decided to analyse Logistic Regression, Linear Discriminant Analysis, and KNeighbors models with 3 different test splits that were 10, 20, 30. In all of them we can see the improvement of recall that was our principal goal since we want a model that predicts exactly what clients are going to churn in the company. The best model of all is Logistic Regression, which has an accuracy of 76%; this accuracy is lower than the previous semester. However, we need to consider that in this case hyperparameters were applied and the data was correctly balance giving results that give us confidence to trust the prediction and recall was improved to 0.81 if we compare with the 20% split that las semester recall had.

# Conclusion

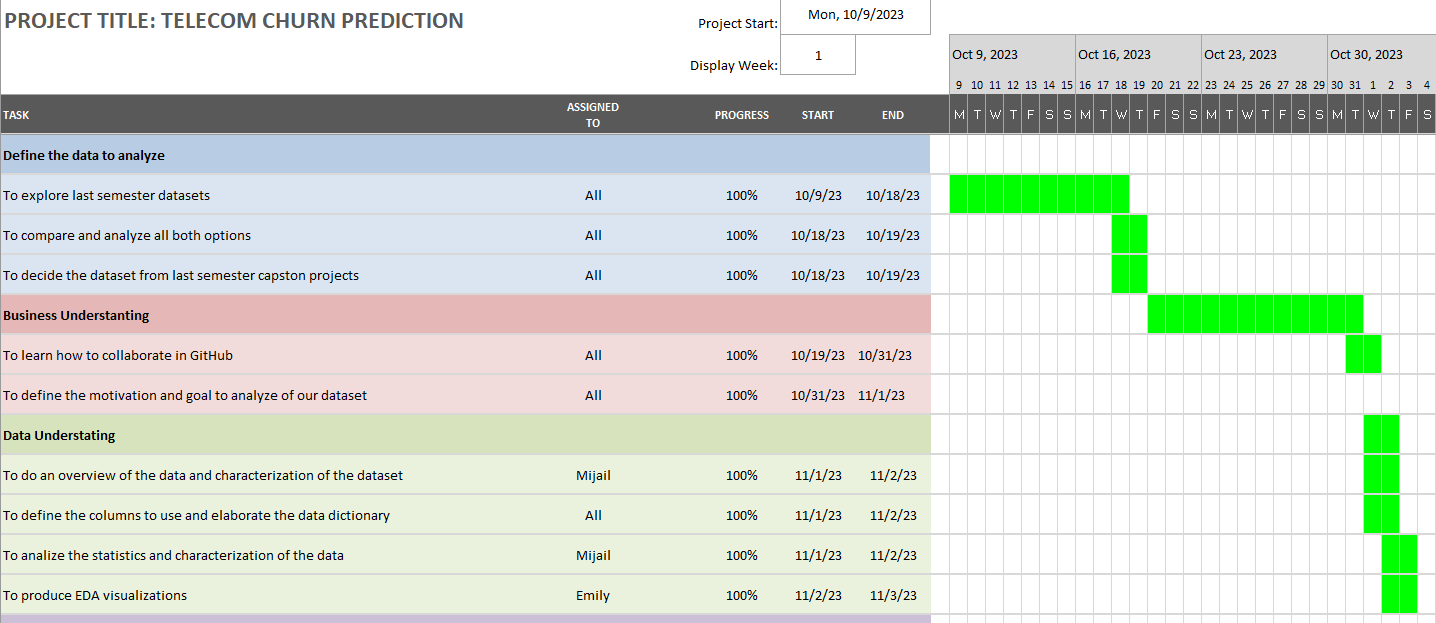
The conclusions are the next:

* The best model applying hyperparameter and balancing the data is Logistic Regression in which we have an accuracy of 76 % in 20% test split and we have a recall of 0.81% improving last semester recall that was 0.57.
* The best 3 models to perform a prediction based in the implementation of hyperparameters are Linear Regression, Linear Discrimination Analysis, and KNeighbors models.
* Last semester Smote and Nearmiss were applied into the models and they introduce bias and variance giving us results in which we can not trust at all, that is why we decided to balance the data before applying hyperparameters.
* Last semester One Hot Encoding was applied to the data which is a good method if we have many rows to analyse, but as our data is limited, we decided to encode the categorical to numerical values with Label Encoding for not having problems with the curse of Dimensionality.
* An indeed exploration was done inside the variables to encode them correctly specially analysing null values where we found that last semester “OnlineBackup” was encoded incorrectly.
* Correlation methods were applied to realize feature engineering with ANOVA and chi-squared test to fundament the dropped variables that were not correlated with the target variable.
* In EDA visualizations we can see that many customers churn after the first month of getting our services and we need to do external studies to know the reason of that since probably other companies get better offers for them when clients are new.
* The company needs to analyse a way to change the monthly contract for longer term contracts since according to the visualization customers that get long term plans stays at our company and they become loyal clients.

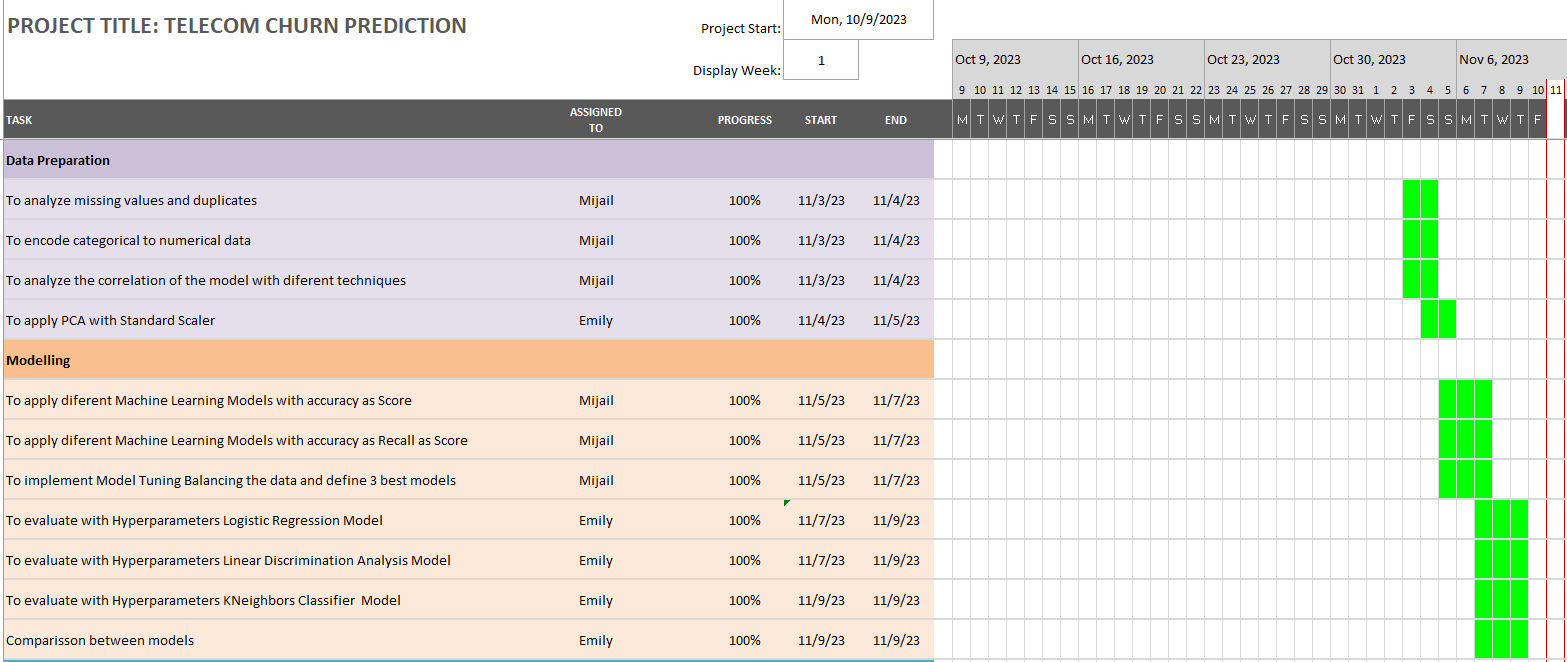
# Appendix

* 1. Appendix 1: CRISP-DM Part 1

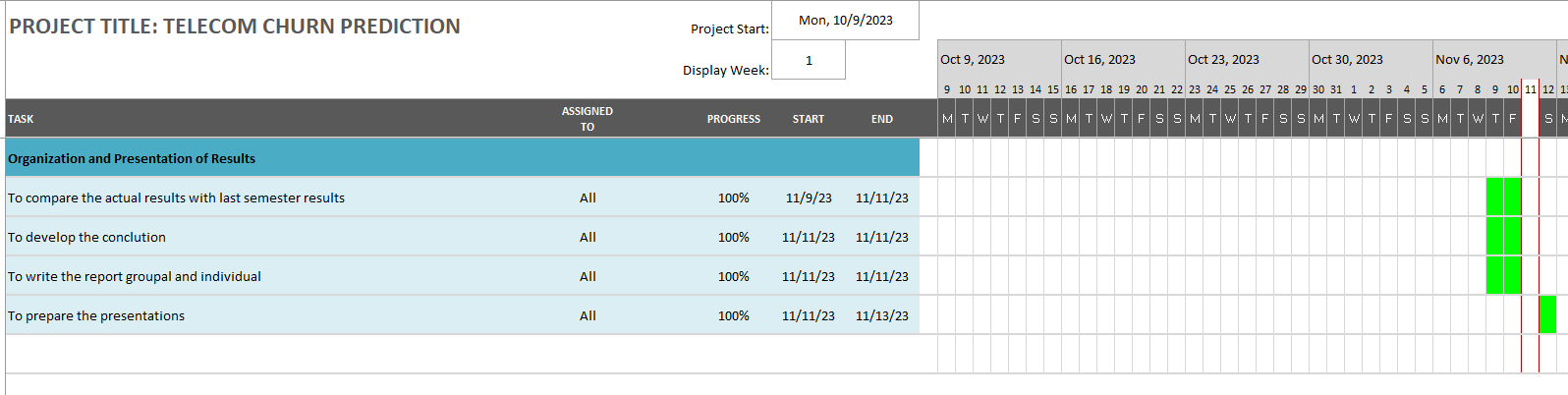
Crisp DM was applied as a Project Management Methodology for this assignment:



* 1. Appendix 1: CRISP-DM Part 2



**13.3 Appendix 1: CRISP-DM Part 3**

****

# Bibliography

* www.kaggle.com. (2017). Telco Customer Churn. [online] Available at: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>.
* Sartorius (2020) *What is principal component analysis (PCA) and how it is used?*, *Sartorius*. Available at: https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186#:~:text=The%20most%20important%20use%20of,variables%2C%20and%20among%20the%20variables. (Accessed: 10 November 2023).
* Seaborn.color\_palette# (n.d.) seaborn.color\_palette - seaborn 0.13.0 documentation. Available at: https://seaborn.pydata.org/generated/seaborn.color\_palette.html (Accessed: 10 November 2023).
* Andreas C. and Sarah G. (2016) Introduction to Machine Learning with Python. United States: O’Reilly Media.Inc.
* Bruce, P.C., Bruce, A. and Gedeck, P. (2020). Practical statistics for data scientists : 50+ essential concepts using R and Python. Sebastopol, Ca: O’reilly Media, Inc
* Tran, N. (2019). Fundamental of The Chi Square in Statistics. [online] Medium. Available at: https://medium.com/@nhan.tran/the-chi-square-statistic-p-1-37a8eb2f27bb#:~:text=(note%3A%20Chi%20square%20tests%20can [Accessed 11 Nov. 2023].
* Wijaya, C.Y. (2021). CRISP-DM Methodology For Your First Data Science Project. [online] Medium. Available at: <https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c>.
* kunwar, A. (2020). Categorical Encoding (Label / Ordinal / Integer encoding) in Feature engineering. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/categorical-encoding-label-ordinal-integer-encoding-in-feature-engineering-1beeaa00f0fa#:~:text=Integer%20encoding%20consist%20in%20replacing [Accessed 11 Nov. 2023].
* www.paddle.com. (n.d.). Customer churn 101: What is it, types of churn, and what to do about it. [online] Available at: <https://www.paddle.com/resources/customer-churn>.
* Mage. (n.d.). Data Cleaning - Remove duplicates. [online] Available at: <https://www.mage.ai/blog/data-cleaning-remove-duplicates>.
* domino.ai. (n.d.). What is Model Tuning for ML Models? | Domino Data Lab. [online] Available at: https://domino.ai/data-science-dictionary/model-tuning [Accessed 11 Nov. 2023].
* www.ibm.com. (2017). Missing Value Analysis. [online] Available at: https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=values-missing-value-analysis.