**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 1](#_Toc150522390)

[2. Motivation 1](#_Toc150522391)

[3. Business Understanding 1](#_Toc150522392)

[4. Business Description 1](#_Toc150522393)

[4.1 Research Question 1](#_Toc150522394)

[4.2 General Goal 1](#_Toc150522395)

[4.3 Success criteria/indicators 1](#_Toc150522396)

[4.4 Causes of Customer Churn 1](#_Toc150522397)

[4.5 Types of Customer Churn 2](#_Toc150522398)

[5. Technologies Used 2](#_Toc150522399)

[5.1 Machine Learning Models and Algorithms 2](#_Toc150522400)

[5.2 Libraries 2](#_Toc150522401)

[5.3 Methodology used for Project Management 2](#_Toc150522402)

[6. Data 2](#_Toc150522403)

[6.1 Acomplishment Data 2](#_Toc150522404)

[6.2 Source 2](#_Toc150522405)

[6.3 Data Dictionary 2](#_Toc150522406)

[6.4 Characterization of the Dataset 3](#_Toc150522407)

[6.5 Data Visualization 5](#_Toc150522408)

[7. Data Cleaning and Feature Engineering 13](#_Toc150522409)

[7.1 Analysis of Null values 13](#_Toc150522410)

[7.2 Dropping duplicates and columns 14](#_Toc150522411)

[7.3 Integer encoding 14](#_Toc150522412)

[8. Correlation Analysis 15](#_Toc150522413)

[8.1 Correlation Matrix 15](#_Toc150522414)

[8.2 ANOVA Test for Correlation Analysis 16](#_Toc150522415)

[8.3 Chi-squeared Test for Correlation Analysis 17](#_Toc150522416)

[9. Modeling 18](#_Toc150522417)

[9.1 Models with 10% test and 90% training. 18](#_Toc150522418)

[9.2 Models with 20% test and 80% training 20](#_Toc150522419)

[9.3 Models with 30% test and 70% training 21](#_Toc150522420)

[9.4 Logistic Regression model with 20% testing and 80% training using the SMOTE technique 22](#_Toc150522421)

[9.5 Logistic Regression model with 20% testing and 80% train ing using the NearMiss technique 23](#_Toc150522422)

[9.6 Logistic Regression model with 20% testing and 80% training using the PCA 24](#_Toc150522423)

[9.7 Observation 27](#_Toc150522424)

[10. Conclusion 28](#_Toc150522425)

[11. Bibliography 30](#_Toc150522426)

**Table of figures**

[Figure 1 – Dictionary of variables 5](#_Toc134659344)

[Figure 2 - First 5 rows of the dataset studied. 6](#_Toc134659345)

[Figure 3 – df\_churn.describe() 6](#_Toc134659346)

[Figure 4 - Percentage of churning customers within the dataset. 7](#_Toc134659347)

[Figure 5 - Percentages of genders within churn. 7](#_Toc134659348)

[Figure 6 - Percentages of Senior Citizens within churn. 7](#_Toc134659349)

[Figure 7 - Percentages of people with and without partner within churn. 8](#_Toc134659350)

[Figure 8 - Percentages of people with and without dependents within churn. 8](#_Toc134659351)

[Figure 9 - Percentages of customers with and without phone service within churn. 8](#_Toc134659352)

[Figure 10 - Data of multiple lines of customers within churn. 9](#_Toc134659353)

[Figure 11 – Data of customers with different internet service within churn. 9](#_Toc134659354)

[Figure 12 - Data of customer with Online Security service within churn. 10](#_Toc134659355)

[Figure 13 - Combination of OnlineSecurity and different internet services for customers within churn. 10](#_Toc134659356)

[Figure 14 - Data of customer with OnlineBackup service within churn. 10](#_Toc134659357)

[Figure 15 - Combination of OnlineBackup and different internet services for customers within churn. 11](#_Toc134659358)

[Figure 16 - Data of customer with DeviceProtection service within churn. 11](#_Toc134659359)

[Figure 17 - Combination of Device Protection and different internet services for customers within churn. 11](#_Toc134659360)

[Figure 18 - Data of customer with Tech Support service within churn. 12](#_Toc134659361)

[Figure 19 - Combination of Tech Support and different internet services for customers within churn. 12](#_Toc134659362)

[Figure 20 - Data of customer with Streaming TV service within churn. 12](#_Toc134659363)

[Figure 21 - Combination of Streaming TV and different internet services for customers within churn. 13](#_Toc134659364)

[Figure 22 - Data of customer with Streaming Movies service within churn. 13](#_Toc134659365)

[Figure 23 - Combination of Streaming Movies and different internet services for customers within churn. 13](#_Toc134659366)

[Figure 24 – Contracts with different durations within churn. 14](#_Toc134659367)

[Figure 25 – Tenure of customers within churn. 14](#_Toc134659368)

[Figure 26 – Paperless billing preference for customers within churn. 15](#_Toc134659369)

[Figure 27 – Payment method preference for customers with paperless billing within churn. 15](#_Toc134659370)

[Figure 28 – Payment method preference for customers with paper billing within churn. 15](#_Toc134659371)

[Figure 29 – Sum of monthly charges for customers with critical services within churn. 16](#_Toc134659372)

[Figure 30 - df\_churn.dtypes results 16](#_Toc134659373)

[Figure 31 - Verifying null values within the dataset. 17](#_Toc134659374)

[Figure 32 - Verification of null values after handling inconsistences on the dataset. 17](#_Toc134659375)

[Figure 33 - Correlation Matrix 18](#_Toc134659376)

[Figure 34 - Encoding the variable "InternetService". 19](#_Toc134659377)

[Figure 35 - Results obtained from getting dummies variables. 19](#_Toc134659378)

[Figure 36 - Correlation Matrix including Dummies Variables 19](#_Toc134659379)

[Figure 37 - Correlation with churn 19](#_Toc134659380)

[Figure 38 - Highest correlations sorted. 20](#_Toc134659381)

[Figure 39 – Resume of ML models results applied for dataset split in 10% test and 90% training. 21](#_Toc134659382)

[Figure 40 - Resume of ML models results applied for dataset split in 20% test and 80% training. 23](#_Toc134659383)

[Figure 41 – Resume of ML models results applied for dataset split in 30% test and 70% training. 24](#_Toc134659384)

[Figure 42 - Resume of ML models results applied for dataset split in 20% test and 80% training using SMOTE technique 25](#_Toc134659385)

[Figure 43 - Resume of ML models results applied for dataset split in 20% test and 80% training using Near Miss technique. 26](#_Toc134659386)

[Figure 44 – First and last 5 rows of dataset after being transformed using PCA. 27](#_Toc134659387)

[Figure 45 – First and last 5 rows of dataset after being transformed with PCA, including churn. 28](#_Toc134659388)

[Figure 47 - Resume of ML models results applied for dataset split in 20% test and 80% training using PCA. 29](#_Toc134659389)

[Figure 48 – Cycle of recommendations to reduce Churn and increase revenue. 31](#_Toc134659390)

# 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

# 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

* 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

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:

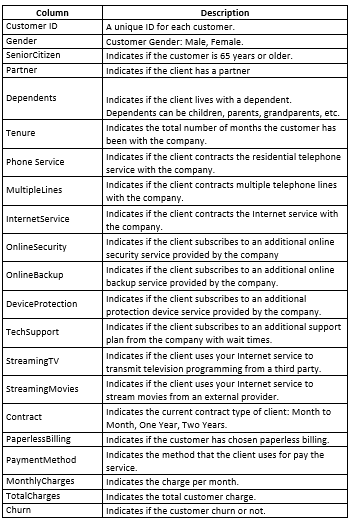


Figure 1: Data Dictionary of Telco Churn Dataset

## 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

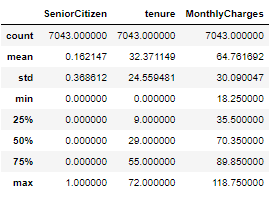


Figure 2 – Statistics of Numerical Values

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.

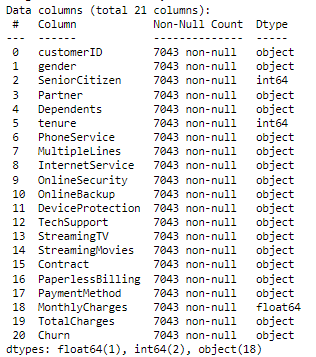


Figure 3 – Data Type of each variable

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

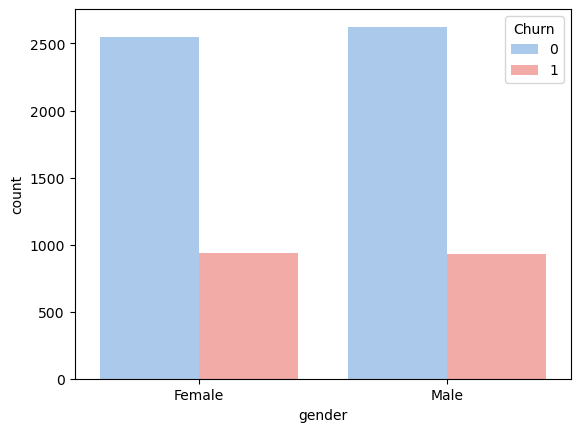


Figure 4 – Churn Vs Gender

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

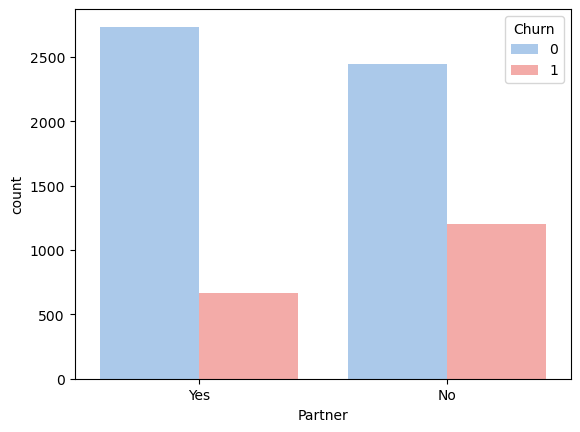


Figure 5 – Churn Vs Partner

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

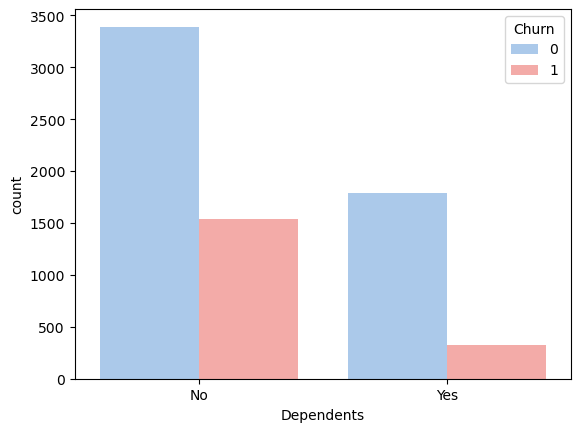


Figure 6 – Churn Vs Dependents

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

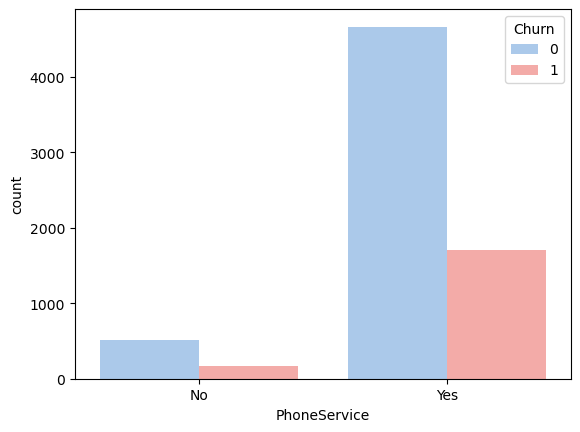


Figure 7 – Churn Vs Phone Service

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

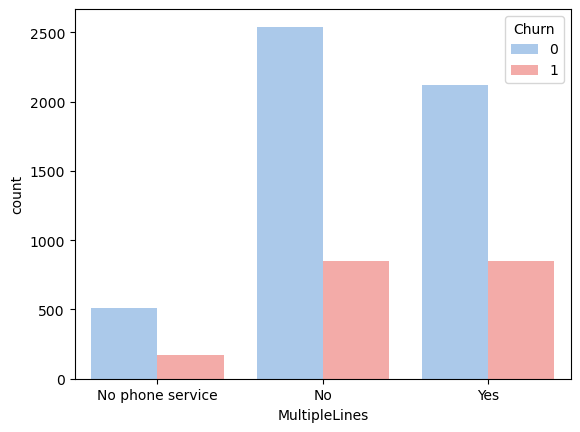


Figure 8 – Churn Vs Multiple Lines

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

6.5.6 Churn vs Internet Service

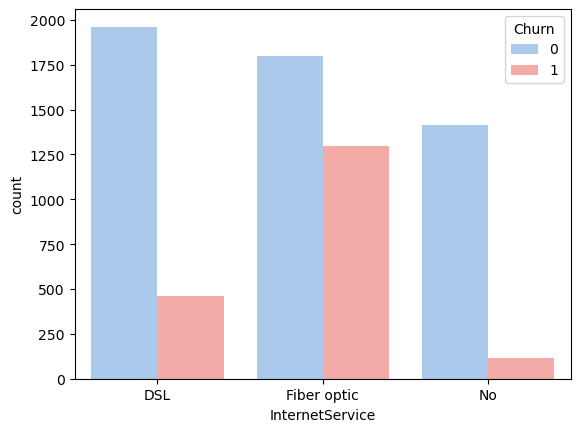


Figure 9 – Churn Vs Internet Service.

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

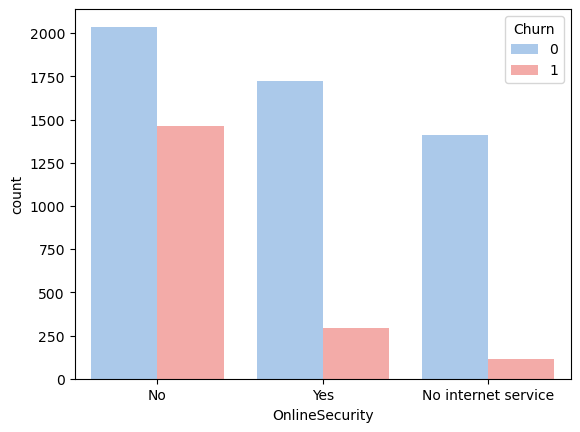


Figure 10 – Churn Vs Online Security

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

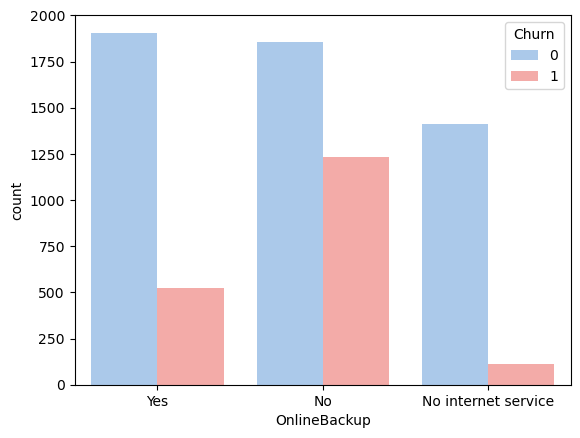


Figure 11 – Churn Vs Online Backup

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

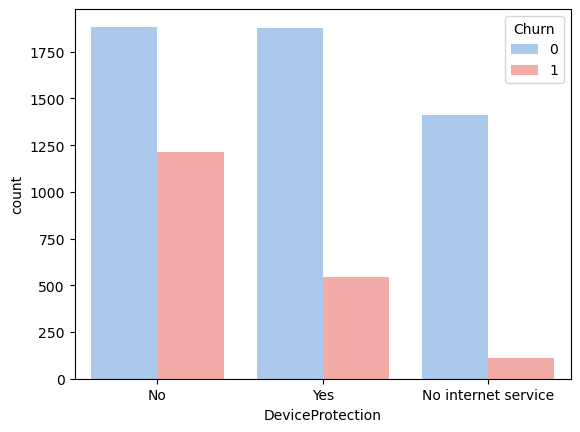


Figure 12 – Churn Vs Device Protection

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

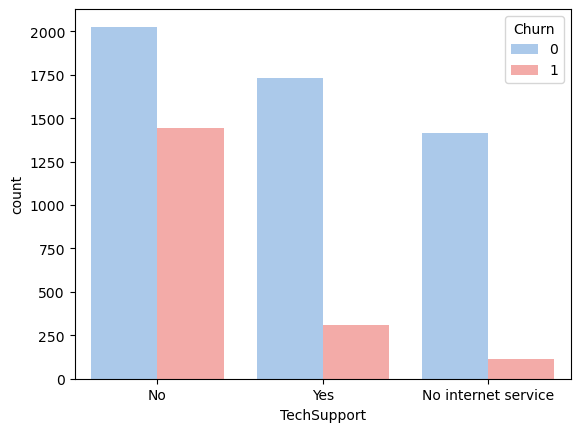


Figure 13 – Churn Vs Tech Support

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

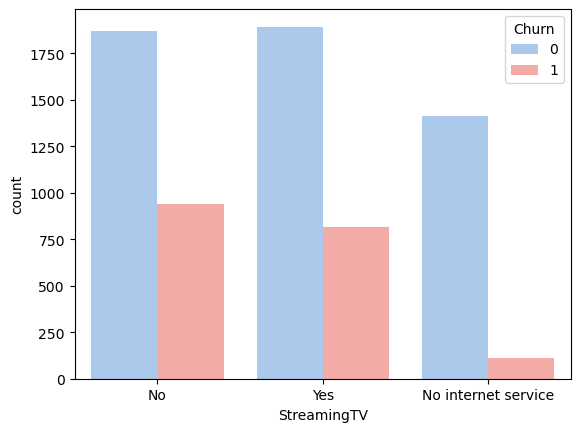


Figure 14 – Churn Vs Streaming TV

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

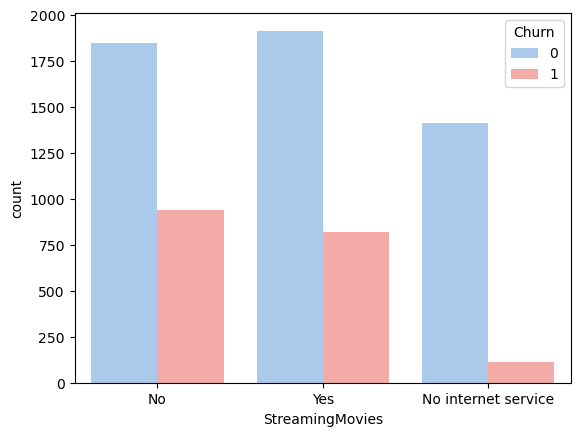


Figure 15 – Churn Vs Streaming Movies

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

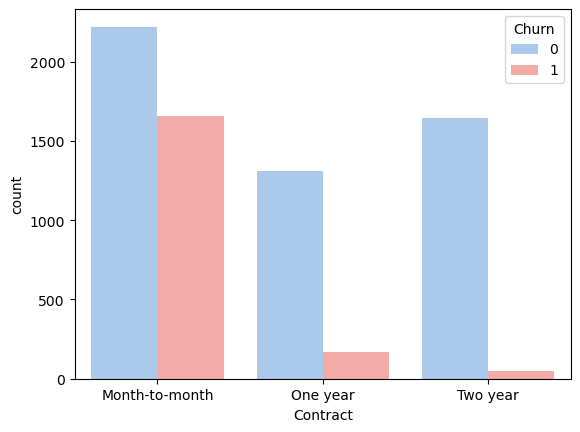


Figure 16 – Churn Vs Contract

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

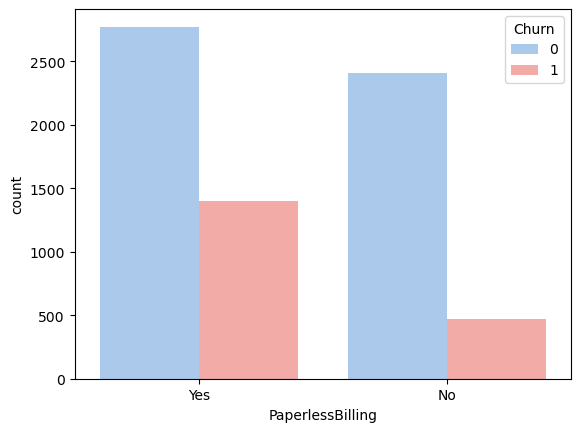


Figure 17 – Churn Vs Paperless Billing

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

6.5.14 Churn vs Payment Method

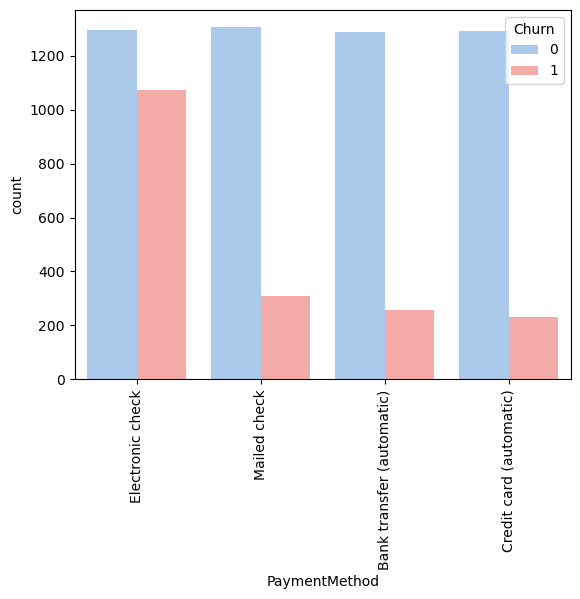


Figure 18 – Churn Vs Payment Method

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

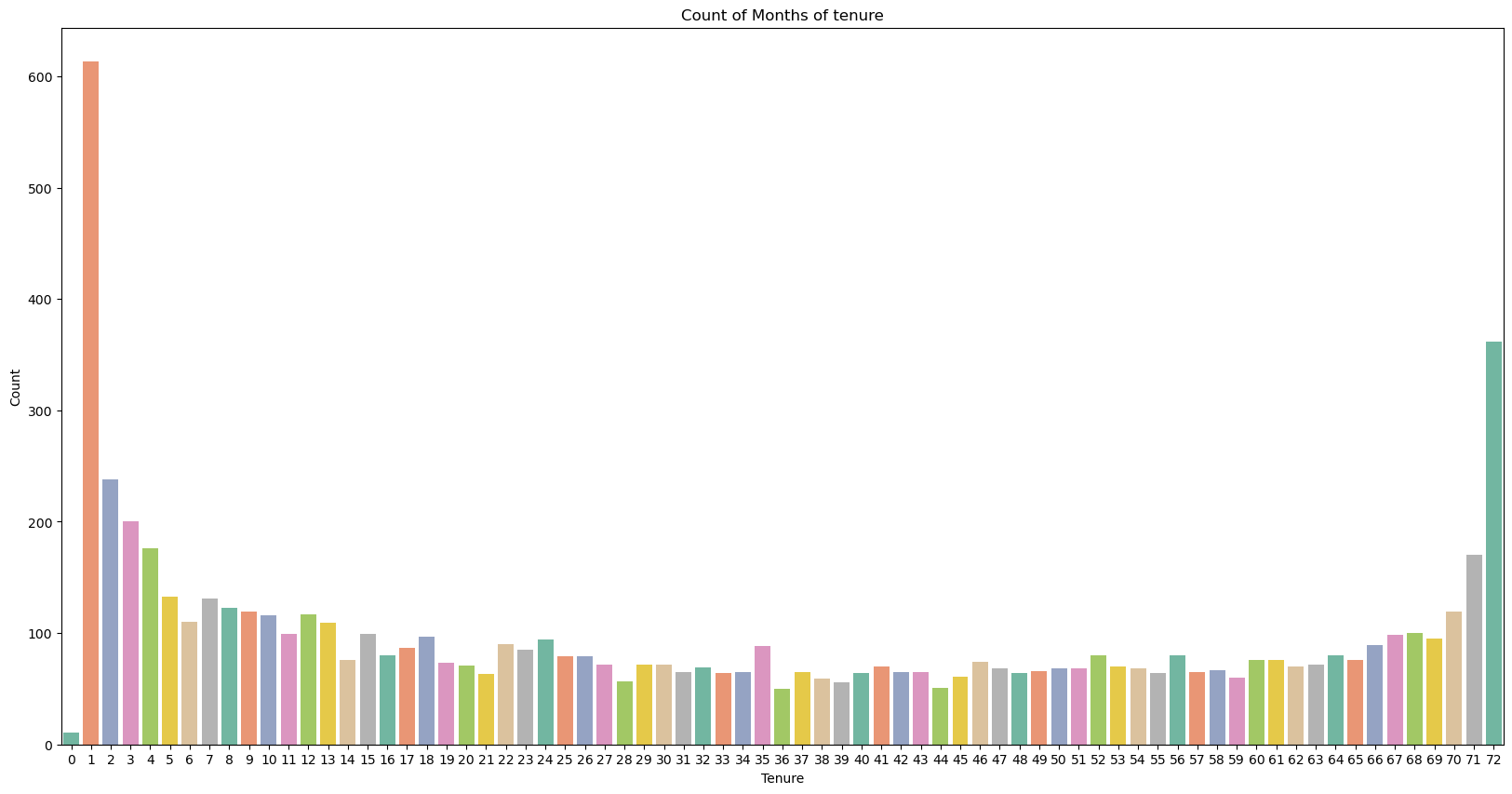


Figure19 – Count of Months of tenure

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

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

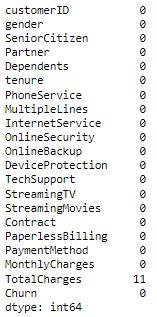


Figure 30 – Number of “Nan” Values in each column.

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

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 clasigy the customers that churn or not and that is why we are going to drop this column.

## Integer encoding

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

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

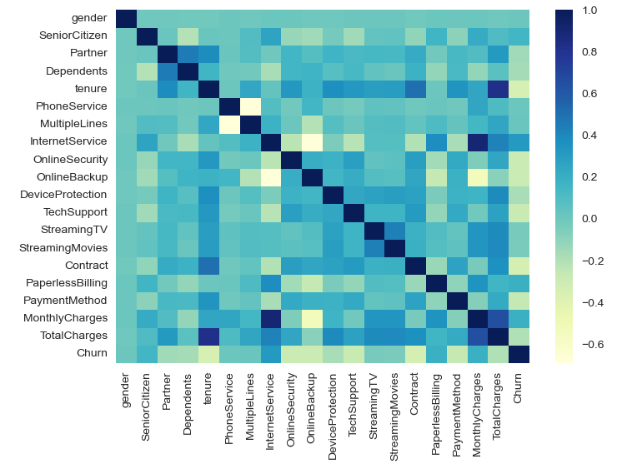


Figure 33 - Correlation Matrix

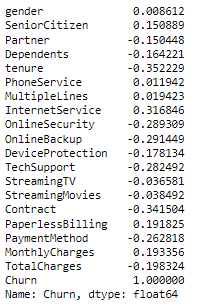


Figure 34 – Correlation of the variables with “churn” variable

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

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

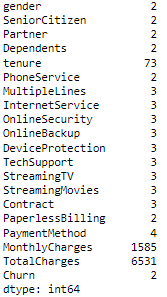


Figure 35 – Unique values in each column

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-squeared Test for Correlation Analysis

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

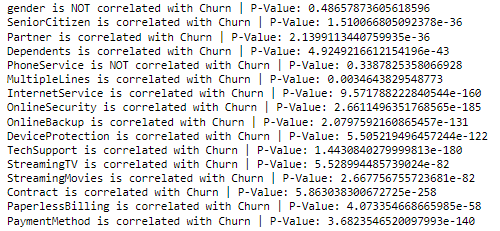


Figure 37 – Results of Chi-squared test

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.

# 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

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.

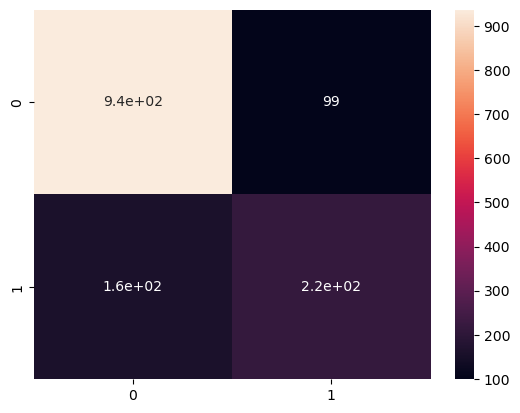


Figure 42 – Initial Confusion Matrix of Logistic Regression test 20% using “Accuracy” as score metric

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

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

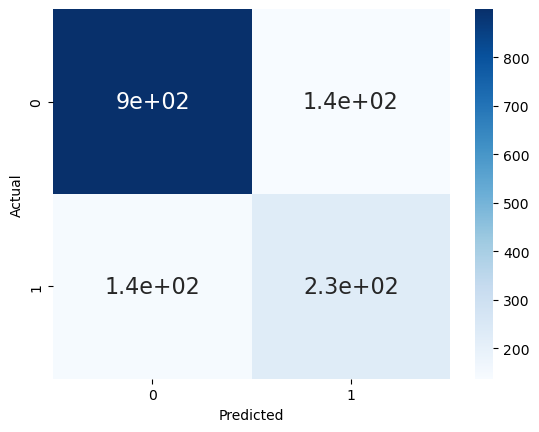


Figure 42 – Confusion Matrix of Logistic Regression test 20% using “Recall” as a score metric

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

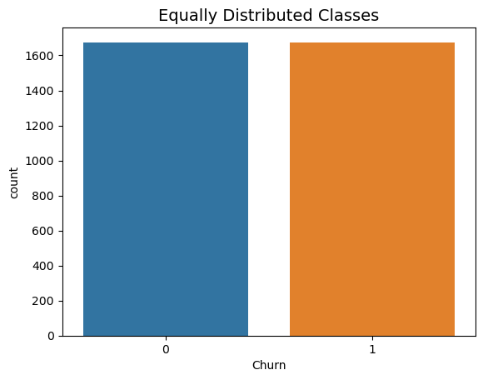


Figure 42 – New equal distribution of churned cases.

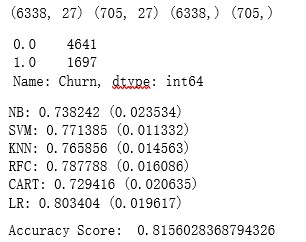
Now that the data is balanced and applying again all the previous models, we got the next results using “accuracy” as score metric:

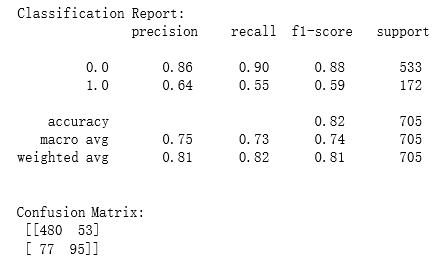
|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy (10% test)** | **Accuracy (20% test)** | **Accuracy (30% test)** |
| **Linear Regression (LR)** | 0.756425 | 0.759693 | 0.755985 |
| **Linear Discrimination Analysis (LDA)** | 0.754038 | 0.758022 | 0.750965 |
| **K-Neighbors Classifier(KNN)** | 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.

## Models with 10% test and 90% training.





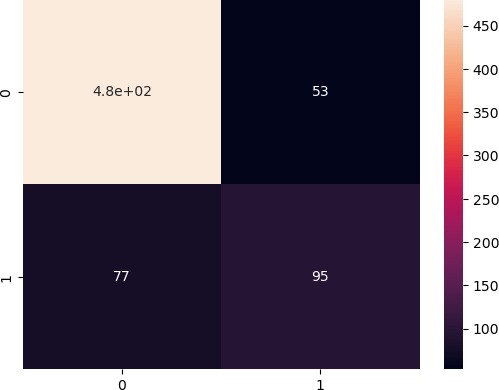
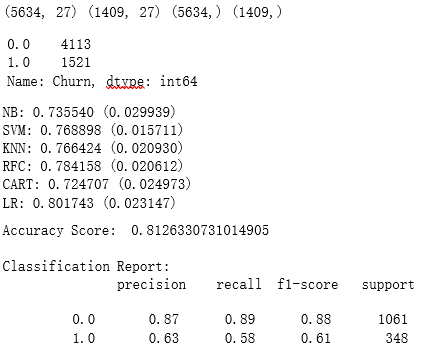
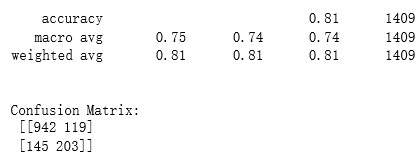


Figure 39 – Resume of ML models results applied for dataset split in 10% test and 90% training.

## Models with 20% test and 80% training





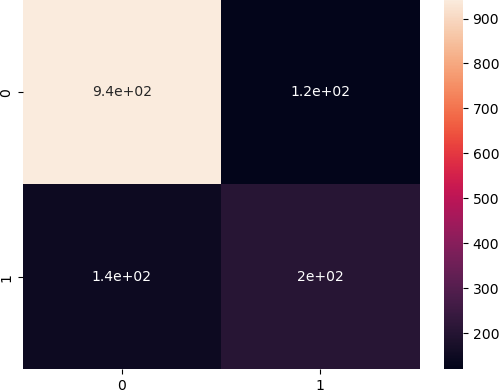
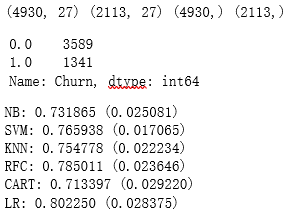
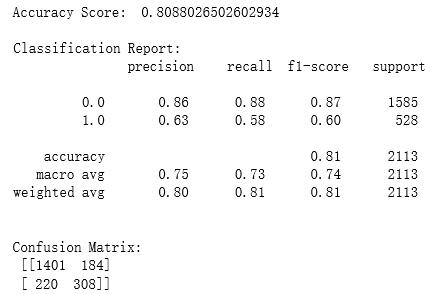


Figure 40 - Resume of ML models results applied for dataset split in 20% test and 80% training.

## Models with 30% test and 70% training





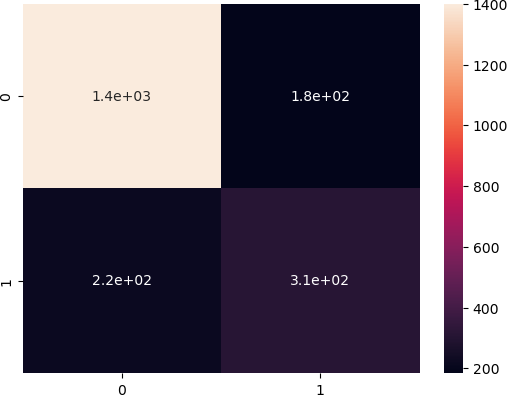
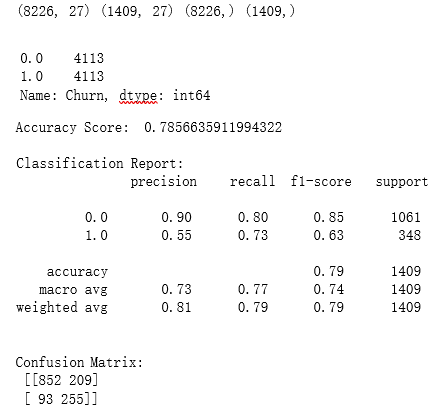


Figure 41 – Resume of ML models results applied for dataset split in 30% test and 70% training.

## Logistic Regression model with 20% testing and 80% training using the SMOTE technique



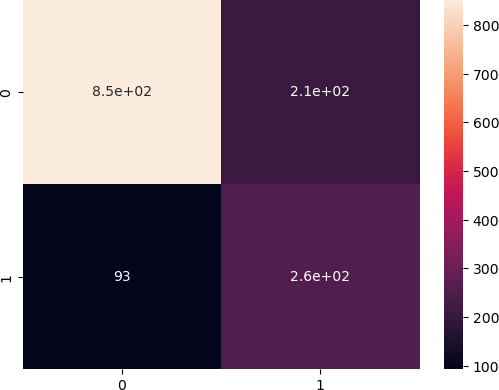
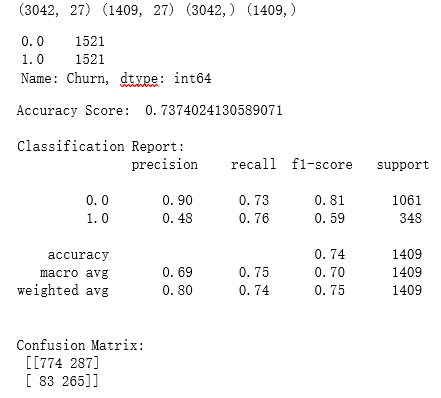


Figure 42 - Resume of ML models results applied for dataset split in 20% test and 80% training using SMOTE technique

## Logistic Regression model with 20% testing and 80% train ing using the NearMiss technique



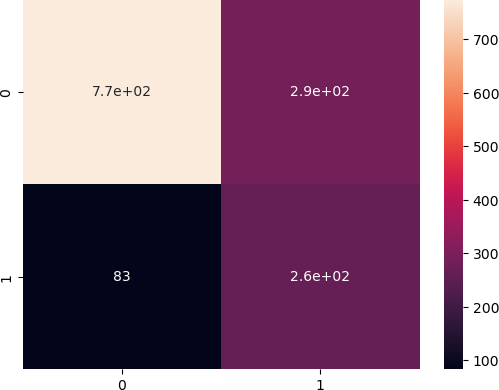
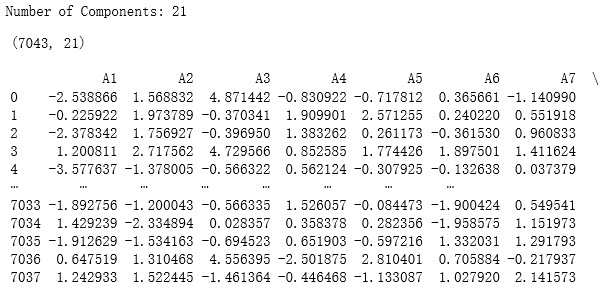


Figure 43 - Resume of ML models results applied for dataset split in 20% test and 80% training using Near Miss technique.

## Logistic Regression model with 20% testing and 80% training using the PCA



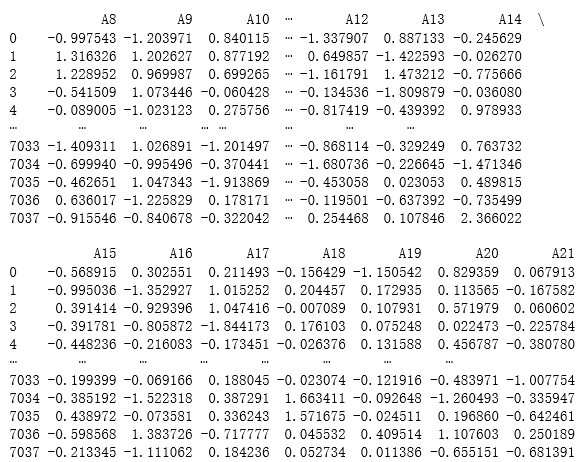
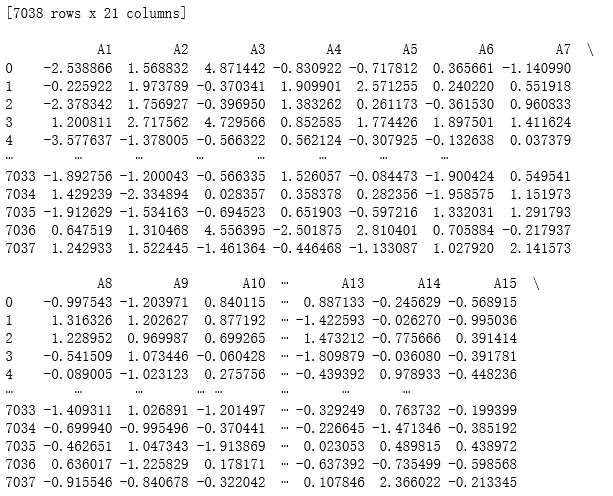


Figure 44 – First and last 5 rows of dataset after being transformed using PCA.



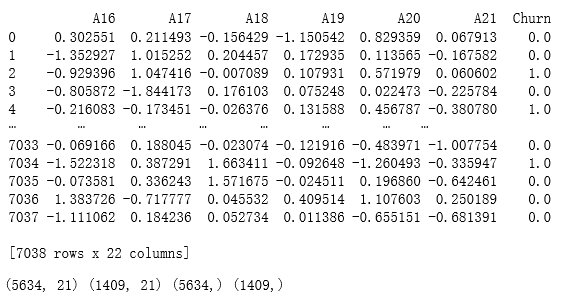
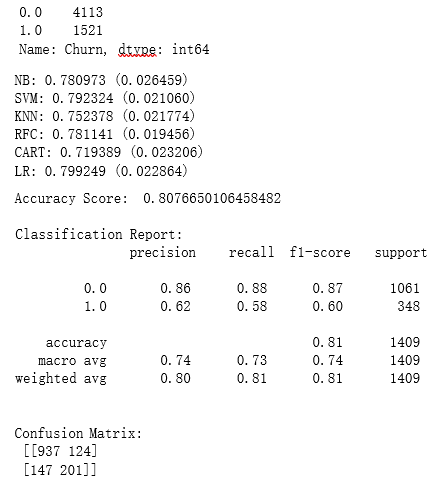


Figure 45 – First and last 5 rows of dataset after being transformed with PCA, including churn.



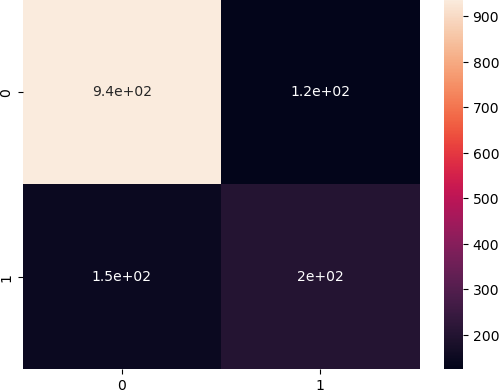


Figure 47 - Resume of ML models results applied for dataset split in 20% test and 80% training using PCA.

## Observation

After having carried out the different models with different percentages in testing and training, we can conclude that the best resulting model is the Logistic Regression with 20% testing and 80% training, which gives us the following information:

* The Accuracy Score of the model is 0.8126, which indicates that the model is capable of correctly predicting 81.26% of the cases.
* The Classification Report shows us the metrics of precision, recall and F1-score for each Churn class (0 and 1).
* Precision tells us how accurate the model is when predicting a given class.
* Recall indicates how well cases of the given class recover.
* The F1-score is a combined measure of accuracy and recovery.
* After having clarified these points we can indicate that:
* The report shows us that the model has a good precision for class 0 (87%) and a moderate precision for class 1 (63%), but the recovery is moderate for class 0 (89%) and low for class 1 (58%). The F1 score also reflects these trends, being highest for class 0 (0.88) and lowest for class 1 (0.61).
* The confusion matrix shows us the number of true positives (TP), false positives (FP), true nega tives (TN) and false negatives (FN) in the model. In this case, the model correctly predicted 942 Churn equals 0 (TN) and 203 Churn equals 1 (TP).
* However, it also incorrectly predicted 119 instances of Churn equal to 0 as Churn equal to 1 (FP) and 145 instances of Churn equal to 1 as Churn equal to 0 (FN).
* The Logistic Regression model has good accuracy in predicting the majority class of Churn equal to 0 but has diﬀiculties correctly predicting the minority class of Churn equal to 1. This is because our data is biased, which mostly has Churn data equal to 0 at 73.46% and Churn data equal to 1 at 26.54%.
* In order to remedy this bias, we apply the Logistic Regression model with 20% testing and 80% training using the SMOTE technique to add synthetic data to our minority variable equivalent to Churn equal to 1.
* In terms of Accuracy Score, the Logistic Regression model without applying the SMOTE technique (0.8126) performs better than the model applying the SMOTE technique (0.7849).
* However, when the Classification Report metrics are analyzed, it’s observed that the model applying the SMOTE technique presents better performance in terms of recall for the minority class (Churn = 1).
* We can see this reflected in the confusion matrix, where it is observed that the model applying the SMOTE technique has fewer false negatives (99) than the model without the technique (145).
* In general, the choice of the model depends on the business objective, for this reason, an initial analysis of the business and our data is of the utmost importance. This way we will be able to detect the importance that is given to each of the metrics that we are evaluating.
* As we already know in the previous analysis, we are looking for a model that has a better performance for the detection of the minority class, in this case, the Clients with Churn (1), knowing this we can opt for the model that uses the SMOTE technique.

# Conclusion

* The products that must be reviewed since they may have some quality and/or price problems are the Phone Service service and the Internet service with Fiber Optic. On the other hand, we can see that the Internet Service through Fiber Optic is directly associated with the Phone Service.
* With this analysis, we can advise executives to observe these services since of the clients that have Churn, these two services are critical, which leads to problems with services derived from the Internet, which are Online Security, Online Backup, Device Protection, Tech Support, Streaming TV and Streaming Movie since they depend directly on the Internet Service.
* We also have observations on the characteristics of the contracts and the form of payment. These services caught our attention since within Churn a large percentage have a month-to-month contract and apparently have problems with the form of payment through Electronic Check.
* We can advise the executives of the company to observe and take action on these products, since in this way the Churn can be reduced and at the same time increase the income of the company and increase customer loyalty.
* The probability that the clients have Churn with the critical services that were detected is 11.20%. At first glance, this probability seems low, but for the customers that are within the Churn with these characteristics, the company has lost a monthly income of €68,282, which is not a minor amount.
* These are the following recommendations that we can deliver to company executives in order to reduce Churn and increase revenue:

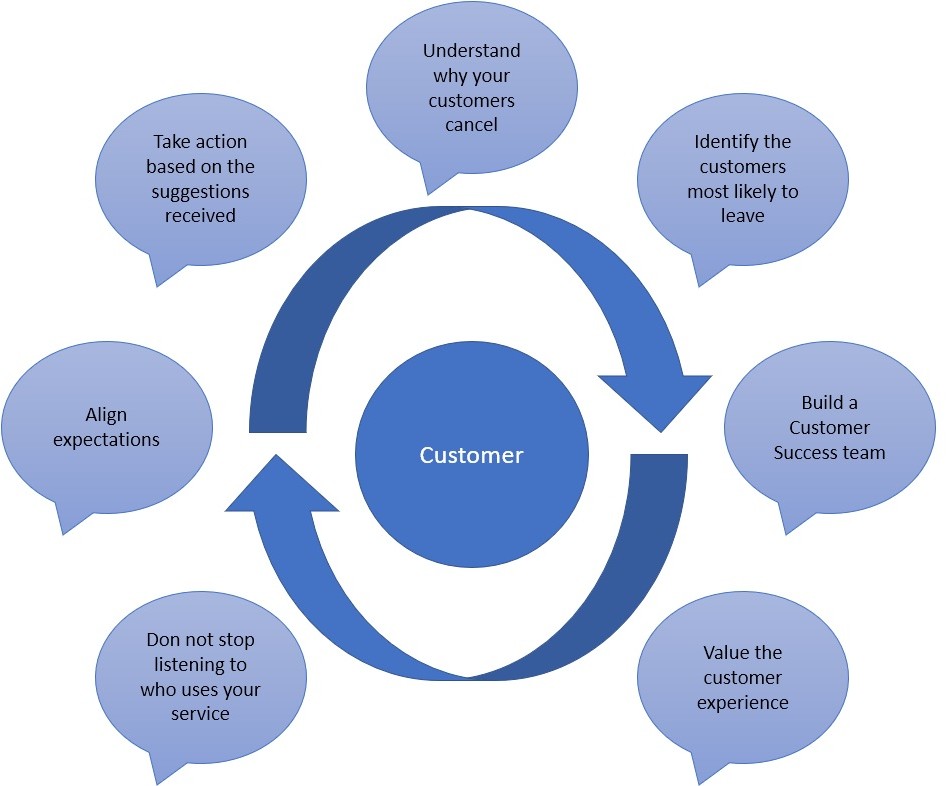


Figure 48 – Cycle of recommendations to reduce Churn and increase revenue.

# Bibliography

* Stack Overflow [online] (May 10, 2021) Available at: https://stackoverflow.com/questions/67474348/how-do-i-create-a-bar-chart-with-percentage- values-in-python-plotly-express
* GeeksforGeeks [online] (16 Jul, 2020) Available at: h[ttps://www.geeksforgeeks.org/p](http://www.geeksforgeeks.org/python-)ython- binomial-distribution/?ref=gcse
* Scribbr [online] Published on November 5, 2020 by Pritha Bhandari. Available at: h[ttps://www.scribbr.com/statistics/standard-normal-distribution/](http://www.scribbr.com/statistics/standard-normal-distribution/)
* nbconvert documentation, Jupyter Development Team [online] Available at: https://nbconvert.readthedocs.io/en/latest/install.html
* Kaggle, Updated 5 Years Ago by Blastchar [online] Available at: h[ttps://www.kaggle.com/datasets/blastchar/telco-customer-c](http://www.kaggle.com/datasets/blastchar/telco-customer-churn)hurn
* W3Schools, Copyright 1999-2023 by Refsnes Data [online] Available at: h[ttps://www.w3sc](http://www.w3schools.com/statistics/statistics_standard_normal_distribution.php)ho[ols.com/statistics/statistics\_standard\_normal\_distribution.php](http://www.w3schools.com/statistics/statistics_standard_normal_distribution.php)

New references with Harvard style

www.kaggle.com. (2017). Telco Customer Churn. [online] Available at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.