**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Mijail Fausto Blanco Vargas  Emily Cristina Herbas Luizaga |
| **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 3](#_Toc134648543)

[2. Business Understanding 3](#_Toc134648544)

[2.1 Causes of Customer Churn 3](#_Toc134648545)

[2.2 Types of Customer Churn 3](#_Toc134648546)

[3. Data Understanding 3](#_Toc134648547)

[3.1 Analysis Plots 6](#_Toc134648548)

[3.1.1 Customer Vs Churn 6](#_Toc134648549)

[3.1.2 Gender Vs Churn 6](#_Toc134648550)

[3.1.3 Senior Citizen Vs Churn 7](#_Toc134648551)

[3.1.4 Partner Vs Churn 7](#_Toc134648552)

[3.1.5 Dependents Vs Churn 8](#_Toc134648553)

[3.1.6 Phone Service and Multiple Lines Vs Churn 8](#_Toc134648554)

[3.1.7 Internet Service Vs Churn 9](#_Toc134648555)

[3.1.8 Online Security Vs Churn 9](#_Toc134648556)

[3.1.9 Online Backup Vs Churn 10](#_Toc134648557)

[3.1.10 Device Protection Vs Churn 10](#_Toc134648558)

[3.1.11 Tech Support Vs Churn 11](#_Toc134648559)

[3.1.12 Streaming TV and Streaming Movie Vs Churn 12](#_Toc134648560)

[3.1.13 Contract and Tenure Vs Churn 13](#_Toc134648561)

[3.1.14 Paperless Billing and Payment Method Vs Churn 14](#_Toc134648562)

[3.1.15 Monthly Charges Vs Churn 15](#_Toc134648563)

[4. Data Cleaning and Normalization 15](#_Toc134648564)

[5. Correlation Analysis 17](#_Toc134648565)

[5.1 Conclusion of Correlation 19](#_Toc134648566)

[6. Modeling 20](#_Toc134648567)

[6.1 Models with 10% test and 90% training 20](#_Toc134648568)

[6.2 Models with 20% test and 80% training 22](#_Toc134648569)

[6.3 Models with 30% test and 70% training 23](#_Toc134648570)

[6.4 Logistic Regression model with 20% testing and 80% training using the SMOTE technique 24](#_Toc134648571)

[6.5 Logistic Regression model with 20% testing and 80% train ing using the NearMiss technique 25](#_Toc134648572)

[6.6 Logistic Regression model with 20% testing and 80% training using the PCA 26](#_Toc134648573)

[6.7 Observation 29](#_Toc134648574)

[7. Conclusion 30](#_Toc134648575)

[8. Bibliography 31](#_Toc134648576)

**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.

Churn is a very important indicator for telecommunications companies 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.

# Business Understanding

## 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.

# Data Understanding

In this project we will analyze the churn data in a telecommunications company, the shape of our data frame has 7043 rows and 21 columns.

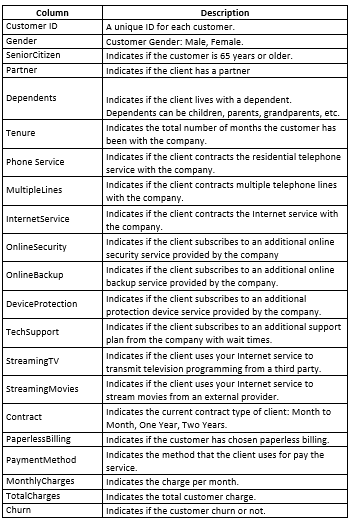
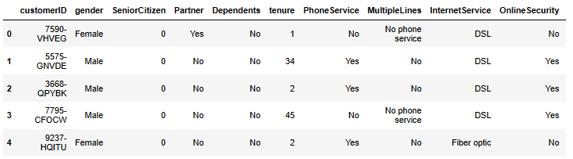


Figure 1 – Dictionary of variables

The variables in our dataset are both numerical and categorical.



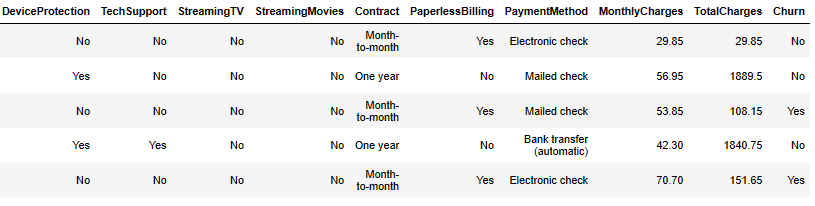


Figure 2 - First 5 rows of the dataset studied.

For obtain descriptive statistics information of our dataset we use the code df\_churn.describe(), this will return us a summary of the statistics of all the columns of the dataset.

In this case, we will analyze the 'churn' column which values were replaced by '0' and '1’.

## Analysis Plots

### 3.1.1 Customers Churned

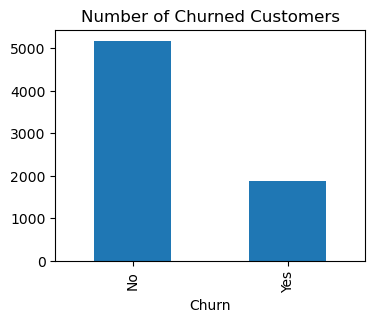


Figure 3 – Number of Churned Customers

We can see that the number of clients without churn is less than the clients with churn.

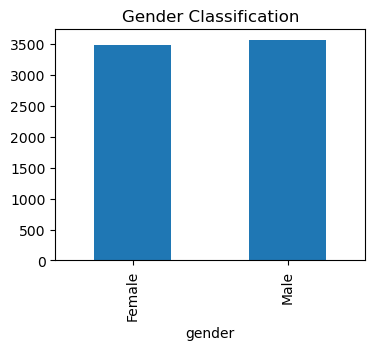


Figure 4 – Number of Clients by Gender

We can see that the company has almost the same number in female clients than male clients.

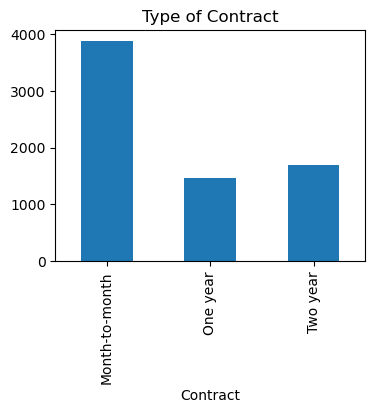


Figure 5 – Type of Contract

We can observe that the company has more month to month contracts customers, se 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.

# Data Cleaning and Normalization

We used the function “df\_churn.dtypes” to verify the data type of each column in our data frame.

This procedure is essential because it ensures the data is being read and processed correctly before applying further analytics methods. It also gives insights when comparing expected values to values found on whether there is any column with “dirty” information, that needs to be cleaned.

The function results in Figure 1.

Table

Description automatically generated

Figure 30 - df\_churn.dtypes results

The column customerID is composed by an ID of each customer, so it is not useful for any further analysis, and therefore will be dropped.

The column TotalCharges catches our attention promptly. Once it contains the total amount paid by customers throughout its contract, it is expected to be a numerical variable. However, the function returns a value that indicate a categorial variable. This indicates that the column may have either null values or strings among numbers in this column.

The function “df\_churn.isna().sum()” is used to verify missing values on the Data Frame. It analyses value by value on the dataset and sums up how many there are in each column.

Table

Description automatically generated

Figure 31 - Verifying null values within the dataset.

It is possible to verify that the dataset does not contain any null value according to Figure 2, therefore it should contain some strings among numerical variables. The list ["n.a.", "?", "NA", "n/a", "na", "--" , " "] contain strings commonly typed by users to represent missing values. The dataset will be verified using this list, and any value common to this list found in the dataset will be replaced by np.nan, which is recognized as null by python, and can be handled uniformly.

The column OnlineBackup was also recognized as an object, whereas it should be an integer. Therefore, using the function “pd.tonumeric(df\_churn[‘OnlineBackup’], errors=’coerce’]”, we converted the variable in integer, and all the errors were converted to null values. Using the same function used before again to count null values, we got the results on Figure 3

Table

Description automatically generated with low confidence

Figure 32 - Verification of null values after handling inconsistences on the dataset.

All the null columns of TotalCharges will be replaced by the product of tenure and MonthlyCharges.

When verifying the counting of values on the column OnlineBackup, It was identified it has 3088 values 0 and 2429 values equals to 1. Once this variable is not among the critical variables verified in our previous analysis, 79% of NA will be replaced by 1 and the remaining by 0, so that this variable will be more balanced.

After these 2 procedures, the dataset has no longer null values or strings among number, therefore it is ready to be entirely converted to “float”. This format allows the use of continuous values, what is especially useful for algorithms that involve distance computation, and make possible use of fractional values, what may improve accuracy.

# Correlation Analysis

Correlation measures the degree of linear relationship between two variables. It is important to highlight, though, that a high correlation does not imply necessarily causation. Therefore, to carry on a proper correlation analysis and draw conclusions from it, it is important first to understand the business, the data and its context, and to identify the objectives of the work. Afterwards, carry out more in-depth studies is necessary to understand different factors that may influence correlations and causation among variables.

First, let us verify the correlation matrix for the dataset. The correlation matrix is in Figure 4.

As it can be verified by the graph, the highest positive correlation of churn is with “InternetService”, whereas the highest negative correlations of churn are with “tenure” and “Contract” variables. However, it is important to notice that we cannot identify from which internet service comes the highest positive correlation, likewise we cannot identify which kind of contract within “Contract” carries the highest negative correlation. Moreover, we identified that the payment method with highest churn was “Electronic Check”, what we cannot observe with the correlation matrix structured as it is.

Chart

Description automatically generated

Figure 33 - Correlation Matrix

The solution to solve this situation and look more properly to the properties of our dataset is encoding our data. Encoding are methods used to convert categorical variables to numerical values. Once we do not have arguments to stand an order to our categorical data, a Dummy Variable Encoding will be used.

Below, Figure 5 bring a piece of codes used to encode the critical variables.

Graphical user interface, text, application

Description automatically generated

Figure 34 - Encoding the variable "InternetService".

The same procedure was repeated to “PaymentMethod”, “Contract” and “PhoneService”.

Table

Description automatically generated

Figure 35 - Results obtained from getting dummies variables.

Figure 6 brings the results obtained from the encoding method. Figure 7 bellow brings the correlation matrix updated with the dummies variables.

|  |  |
| --- | --- |
| Chart  Description automatically generated  Figure 36 - Correlation Matrix including Dummies Variables | Table  Description automatically generated with low confidence  Figure 37 - Correlation with churn |

Printing the correlation Matrix for only the variable Churn, we obtain the values presented on Figure 8. It is possible now to verify that our initial analysis was accurate according to the correlation matrix.

## Conclusion of Correlation

Although there is no rule setting specific when correlation should be considered high, since it depends on the context, by and large, correlations with absolute value over 0.7 are considered high. Nevertheless, in our example, the Churn of the company barely overcomes 25%, therefore we consider correlations with absolute values over 0.3 as high. Based on this and using the correlation matrix, we filtered and sorted the highest pair of correlations over 0.3, which will be used to draw conclusions. The pairs and its correlations are presented in Figure 9.

Text

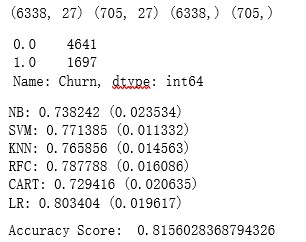
Description automatically generated

Figure 38 - Highest correlations sorted.

* The correlation of electronic check and churn of 0.30 indicates that customers utilizing the payment way are more likely to churn, therefore other payment methods should be recommended rather them this.
* The negative correlation between Churn and tenure of -0.35indicate that last longing customers are less likely to churn, therefore it is recommended promote long-term customers loyalty through marketing strategies.
* The positive high correlation between churn and contract month-to-month are another factor the suggests that long-term contract should be promoted.
* The positive correlation between churn and Internet Service Fiber Optic suggests that customers with this service are likely to churn. This indicates a problem with this service, and it is recommended to look for opportunities to improvement on quality of this service.
* Another important information that should be highlighted is the fact that Fiber Optic service is only available to customers with Phone Service. This is not found on the correlation analysis, but it is importance once Churn has a high correlation with Fiber Optic Service.

# Modeling

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



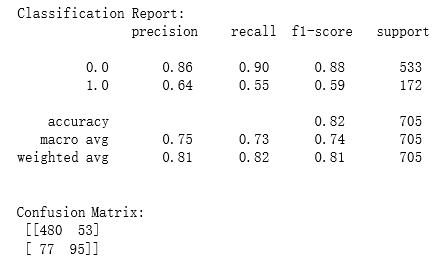
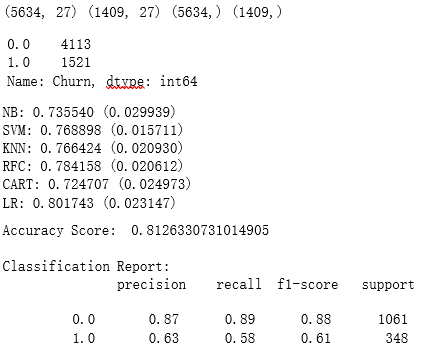
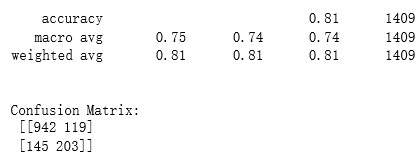




Figure 39 – Resume of ML models results applied for dataset split in 10% test and 90% training.

## Models with 20% test and 80% training





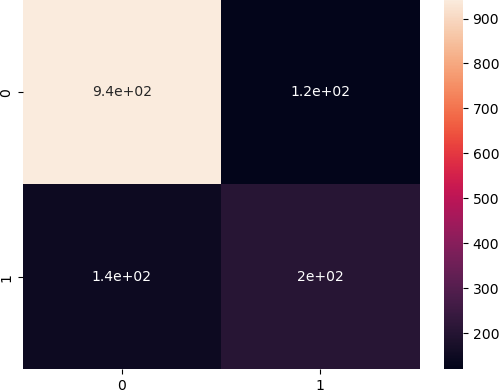
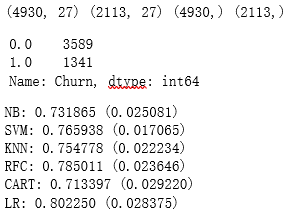
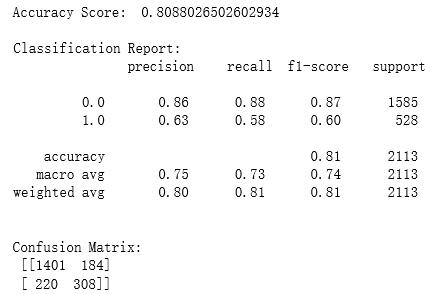


Figure 40 - Resume of ML models results applied for dataset split in 20% test and 80% training.

## Models with 30% test and 70% training





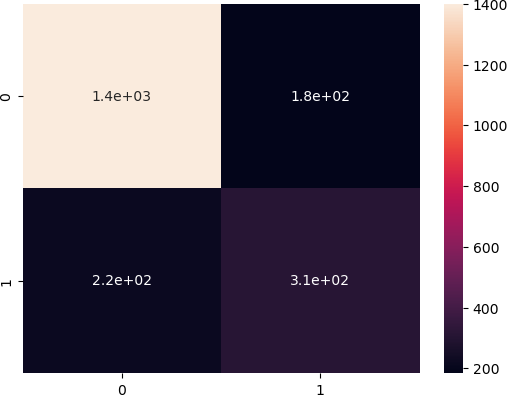
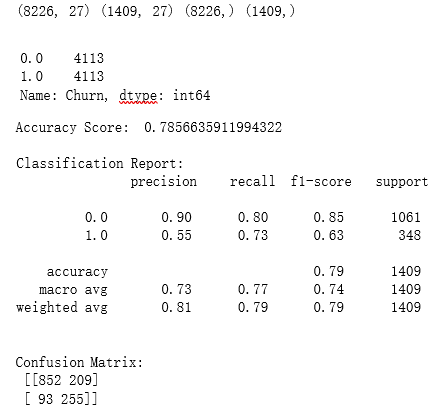


Figure 41 – Resume of ML models results applied for dataset split in 30% test and 70% training.

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



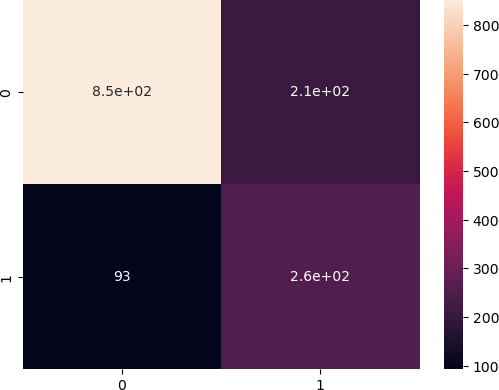
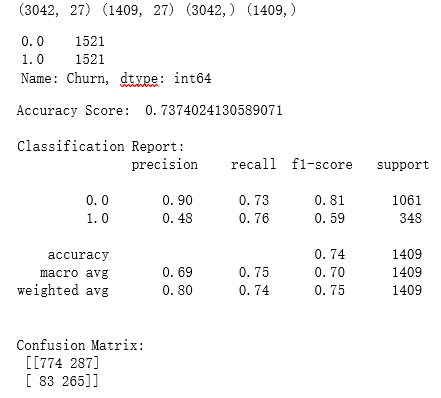


Figure 42 - Resume of ML models results applied for dataset split in 20% test and 80% training using SMOTE technique

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



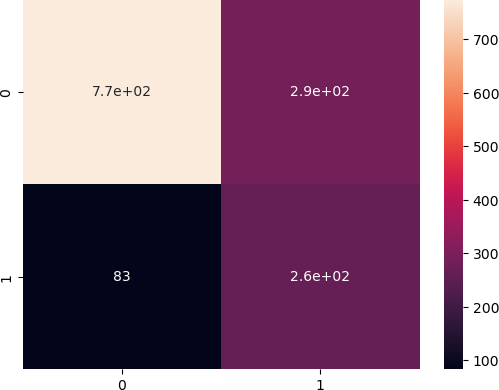
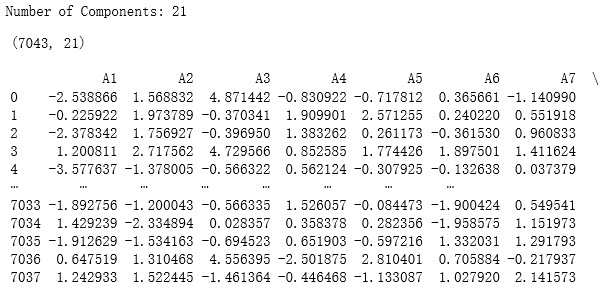


Figure 43 - Resume of ML models results applied for dataset split in 20% test and 80% training using Near Miss technique.

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



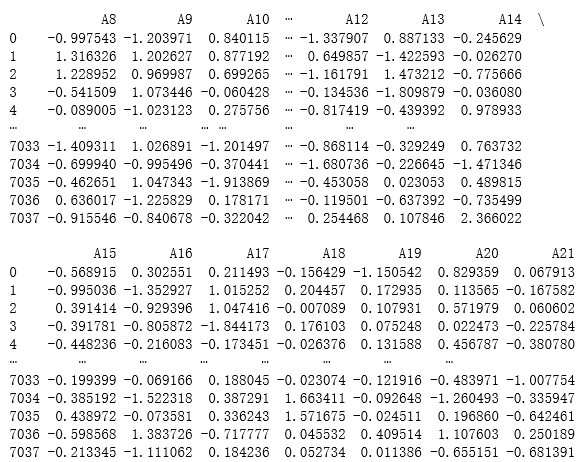
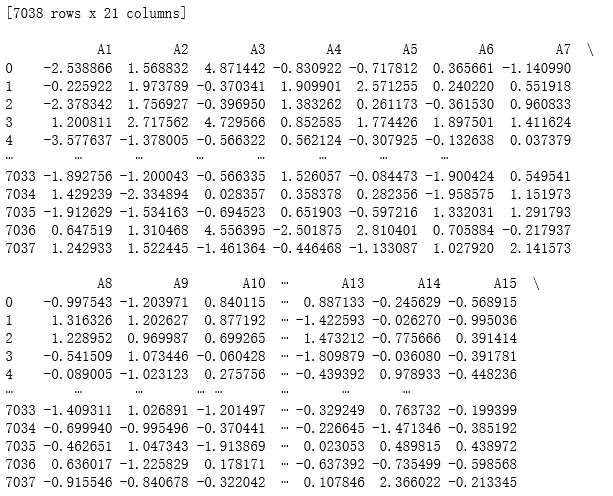


Figure 44 – First and last 5 rows of dataset after being transformed using PCA.



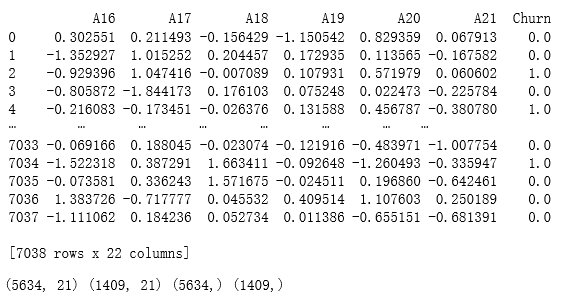
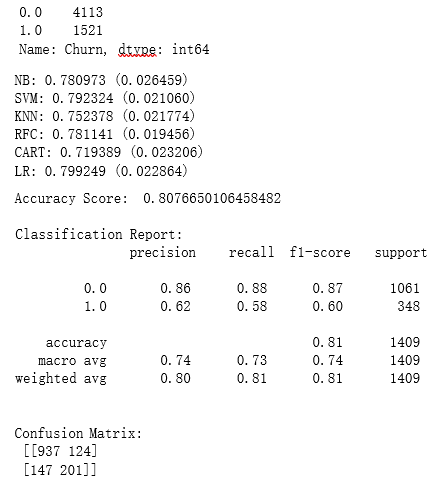


Figure 45 – First and last 5 rows of dataset after being transformed with PCA, including churn.



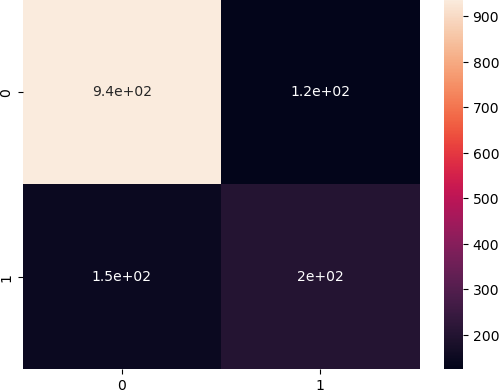


Figure 47 - Resume of ML models results applied for dataset split in 20% test and 80% training using PCA.

## 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:

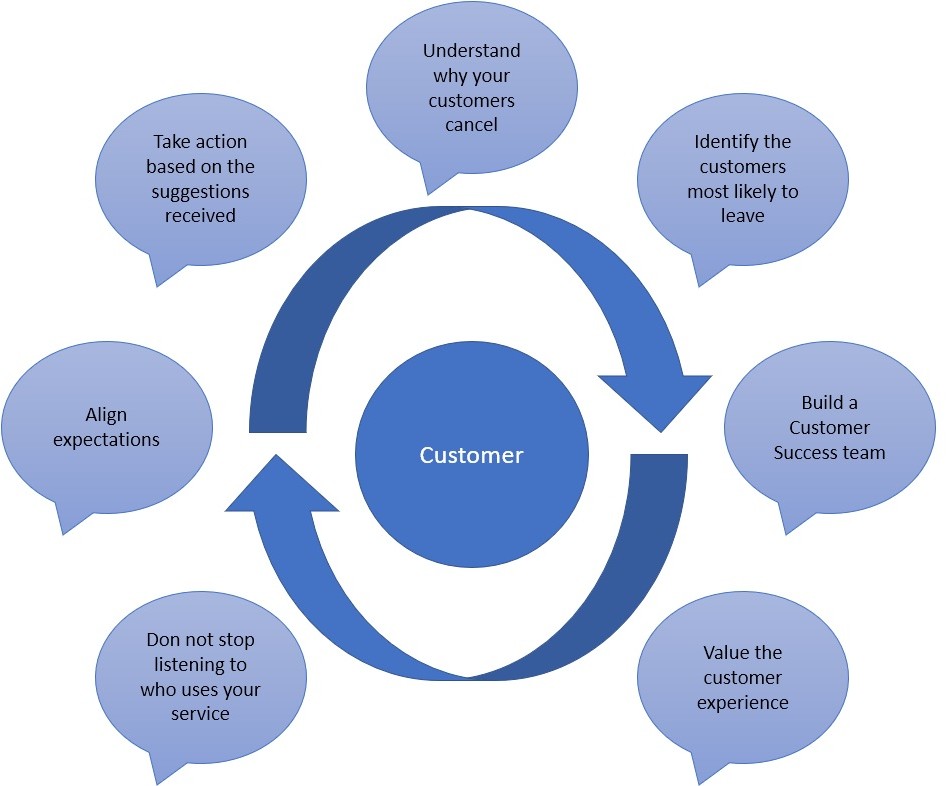


Figure 48 – Cycle of recommendations to reduce Churn and increase revenue.

# 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)