

UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Facoltà di Economia

Master's in Data Science for Management



Predictive Modelling Applied to Churn Prevention

*Student:* Carlos Henrique da Silva Amaral  
*ID:* 4816927

Academic Year 2018-2019

## Abstract

### Predictive Modelling Applied to Churn Prevention – a case study

Carlos Henrique da Silva Amaral

Here is presented a comparative study on the machine learning methods applied to the challenging problem of customer churning prediction. The models used were Logistic Regression, Decision Tree and Random Forest and they were applied and evaluated using a public domain dataset from the telecommunications industry. The evaluation was made by comparing some key metrics of the results – accuracy, sensitivity and specificity.

Keywords: Churn prediction; Machine learning techniques; Marketing analytics; Predictive analysis.

# Table of Contents

List of Figures.....	4
1. Analysis of the Company .....	5
2. Learning Objectives .....	6
3. Definition of the problem that is being investigated .....	7
3.1 Churn Rate .....	7
3.2 An Example: Telecommunications Industry Churn Rates .....	8
4. Description of the research methods and analysis .....	8
4.1 Dataset Description .....	9
4.2 Data Transformation .....	11
4.3 Exploratory Data Analysis and Feature Selection.....	13
4.4 Supervised learning steps.....	16
4.4.1 Logistic Regression .....	16
4.4.2 Decision Tree .....	22
4.4.3 Random Forest .....	26
4.5 Summary .....	32
5. Skills acquire in the internship .....	33
6. References.....	34

## List of Figures

Figure 1: Data Structure of the Telco data set .....	9
Figure 2: Exploring Churn Distribution .....	10
Figure 3: Missing Values .....	11
Figure 4: Recoding No Internet Service .....	11
Figure 5: Recoding No phone service .....	12
Figure 6: Tenure grouped by months .....	12
Figure 7: Dataset after data wrangling .....	15
Figure 8: Correlation Matrix for the numerical variables .....	13
Figure 9: Distribution of Numerical Variables .....	13
Figure 10: Customer Demographic Data .....	14
Figure 11: Subscribed Services .....	14
Figure 12: Customer Account Information .....	15
Figure 13: Summary of Full Logistic Regression model .....	18
Figure 14: Logistic Regression after feature selection process .....	19
Figure 15: Analysis of Deviance ANOVA .....	20
Figure 16: Model fit statistics .....	20
Figure 17: Confusion Matrix and Accuracy for the Logistic Regression model .....	21
Figure 18: Decision Tree model .....	23
Figure 19: Complexity Parameter Table .....	24
Figure 20: Pruned Decision Tree .....	25
Figure 21: Confusion Matrix and Accuracy for the Decision Tree Model .....	26
Figure 22: Random Forest model .....	27
Figure 23: Confusion Matrix and Accuracy for the Random Forest model .....	28
Figure 24: Random Forest Error Rate .....	29
Figure 25: OOB error by the number of mtry .....	29
Figure 26: Tuned Random Forest .....	30
Figure 27: Confusion Matrix and Accuracy of the Tuned Random Forest .....	30
Figure 28: Random Forest Feature Importance .....	31
Figure 29: Key metrics for models comparison .....	32

# 1. Analysis of the Company

The internship was done at **Nunatac** s.r.l, an Italian company specialized in Business Intelligence and Data Analytics, with twenty-five years of design experience in advanced analytics for complex national and international organizational contexts, specialized in all the areas of Data Warehousing and Data Mining for banks, insurance companies, telecommunications and large companies in other sectors.

The main values of the company are:

- the recognition of the importance and willingness to understand all the real needs of customers, adopting their needs as their own;
- a concrete approach, strongly focused on the feasibility and affordability of the proposed solutions;
- attention to the quality and customization of its deliveries.

Nunatac was founded in 1994 in order to respond to a new market request: to combine a clear understanding of the customer's business needs and the ability to transform available data, structured or unstructured, into analytical processes that integrate in business systems, producing measurable value.

Nunatac is part of Alkemy, an Italian digital enabler, i.e. a consultant and service provider who provides the support needed to identify growth opportunities and other innovative solutions. It is a B2B company that contributes to the growth of its clients, mid-sized to large Italian and international organizations, accompanying them in the digital transformation.

The Alkemy's goal is to help companies redefine strategies, products and services, media and sales, aligned with the evolution of digital technologies and new consumer behavior.

The company works in the core areas of digital transformation: consulting, e-commerce, creativity and brand strategy, UX and design, social media, content, digital transformation, technology and data analysis - the latter managed by Nunatac.

## 2. Learning Objectives

The aim of the internship was the development of data analysis skills and the elaboration of statistical models with the help of computer systems.

The internship-related activities were those that helps the student to manage increasingly complex activities, optimizing time and improving group work skills and relationships with co-workers.

The training objective of the internship was to support an Italian company in the field of Personal Loans, in the area of Digital Marketing Analytics, providing them with first-hand support and assistance in their daily activities.

*Marketing analytics* comprises the processes and technologies that enable marketers to evaluate the success of their marketing initiatives. This is accomplished by measuring performance (e.g., blogging versus social media versus channel communications). It uses important business metrics, such as ROI, marketing attribution and overall marketing effectiveness. In other words, it tells you how your marketing programs are really performing. Marketing analytics gathers data from across all marketing channels and consolidates it into a common marketing view. (SAS Marketing Analytics, 2019)

At this company, the data collection/extraction and the analysis are done by using the *SAS Enterprise Guide* tool, a point-and-click, menu- and wizard-driven tool that helps the users to analyze data and publish their results. It provides fast-track learning for quick data investigations, generating the code for greater productivity, accelerating deployment of analyses and forecasts.

However, since the company data may refer to a certain level of personal data of European Union citizens and, in order to comply with the *General Data Protection Regulation* guidelines (Data protection in the EU, 2019), the model developed for this thesis will use a generic database.

The main idea of this project is to describe the steps necessary to create a model of churn prediction and prevention, whose methodology can be applied in different fields.

### 3. Definition of the problem that is being investigated

By means of data-driven marketing as well as machine learning technology, this paper presents a case study about the prediction of customer churning using an open dataset from a telecommunication company.

The main incentive for a generic business to do churn prevention is to convincing existing customers to buy again and/or stay loyal to, in this case, the telecom company. Basically, it is a measure to ensure that customers will not decide to migrate to another company.

#### 3.1 Churn Rate

The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity. It is most expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period. For a company to expand its number of clients, its growth rate (measured by the number of new customers) must exceed its churn rate (Churn Rate Definition, 2019).

The churn rate, when applied to a customer base, refers to the proportion of contractual customers or subscribers who leave a supplier during a given time period. It is a possible indicator of customer dissatisfaction, cheaper and/or better offers from the competition, more successful sales and/or marketing by the competition, or reasons related to the customer life cycle.

Companies usually make a distinction between **voluntary churn** and **involuntary churn**. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, while involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies' control, such as how billing interactions are handled or how after-sales help is provided (Churn Attrition, 2019).

Telephone service companies, internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer churn analysis and customer churn rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches

which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

### 3.2 An Example: Telecommunications Industry Churn Rates

The churn rate is a particularly useful measurement in the telecommunications industry. This includes cable or satellite television providers, internet providers, and telephone service providers (landline and wireless service providers). As most customers have multiple options from which to choose, the churn rate helps a company determine how it is measuring up to its competitors. If one out of every 20 subscribers to a high-speed Internet service terminated their subscriptions within a year, the annual churn rate for that internet provider would be 5%.

## 4. Description of the research methods and analysis

Marketing analytics teams can use supervised learning models to predict customer churn by assessing their propensity to churn. Since these models generate a small prioritized list of potential deserters, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

Supervised learning models require two key data elements: the first one is the target variable which is what we want to predict. It could be predicting which customers will churn, or which customers will buy again. The second data element are the features that will be used to predict the target variable.

The programming language used to perform this analysis was **R 3.6.2 for Windows** and the integrated development environment (IDE) used was **RStudio Desktop 1.2.5033**.



## 4.1 Dataset Description

The structure of the **Telco Dataset** is as follows. Each row/observation (7.043) represents a customer, and each column/variable (21) contains customer's attributes. The last column "**Churn**" classifies whether a specific customer has churned or not.

```
[1] "Data Structure - Telco Dataset"
'data.frame': 7043 obs. of 21 variables:
 $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",...: 5376 3963 2565 5536 6512 4771 5605 4535 ...
 $ gender      : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Partner     : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure     : int 1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines : Factor w/ 3 levels "No","No phone service",...: 2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",...: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",...: 1 3 3 3 1 1 1 3 1 3 ...
 $ OnlineBackup : Factor w/ 3 levels "No","No internet service",...: 3 1 3 1 1 1 1 3 1 1 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",...: 1 3 1 3 1 3 1 1 3 1 ...
 $ TechSupport : Factor w/ 3 levels "No","No internet service",...: 1 1 1 3 1 1 1 1 3 1 ...
 $ StreamingTV : Factor w/ 3 levels "No","No internet service",...: 1 1 1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",...: 1 1 1 1 1 3 1 1 3 1 ...
 $ Contract     : Factor w/ 3 levels "Month-to-month",...: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",...: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn        : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

FIGURE 1: DATA STRUCTURE OF THE TELCO DATA SET

- **customerID**: client identification;
- **gender**: female, male;
- **SeniorCitizen**: Whether the customer is a senior citizen or not (1, 0);
- **Partner**: Whether the customer has a partner or not (Yes, No);
- **Dependents**: Whether the customer has dependents or not (Yes, No);
- **Tenure**: Number of months the customer has stayed with the company;
- **PhoneService**: Whether the customer has a phone service or not (Yes, No);
- **MultipleLines**: Whether the customer has multiple lines or not (Yes, No, No phone service);
- **InternetService**: Customer's internet service provider (DSL, Fiber optic, No);
- **OnlineSecurity**: Whether the customer has online security or not (Yes, No, No internet service);
- **OnlineBackup**: Whether the customer has online backup or not (Yes, No, No internet service);
- **DeviceProtection**: Whether the customer has device protection or not (Yes, No, No internet service);

- **TechSupport:** Whether the customer has tech support or not (Yes, No, No internet service);
- **streamingTV:** Whether the customer has streaming TV or not (Yes, No, No internet service);
- **streamingMovies:** Whether the customer has streaming movies or not (Yes, No, No internet service);
- **Contract:** The contract term of the customer (Month-to-month, One year, Two year);
- **PaperlessBilling:** Whether the customer has paperless billing or not (Yes, No);
- **PaymentMethod:** The customer's payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic));
- **MonthlyCharges:** The amount charged to the customer monthly;
- **TotalCharges:** The total amount charged to the customer;
- **Churn:** Whether the customer churned or not (Yes or No).

The features in this dataset could be divided as follows:

- **Customer demographic data:** Gender, SeniorCitizen, Partner, Dependents;
- **Subscribed services:** PhoneService, MultipleLine, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies;
- **Customer account information:** CustomerID, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Tenure;

One thing that is important to explore is whether there is a severe class imbalance meaning there are large differences in the number of observations in each class. In this case it is possible to see that there are over 26% churned customers and over 73% non-churned customers. There is some class imbalance, but not a severe one.

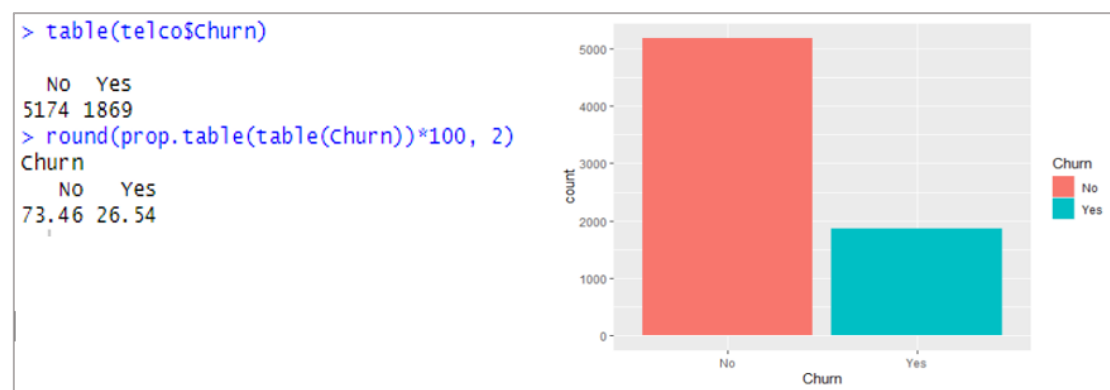


FIGURE 2: CHURN DISTRIBUTION

## 4.2 Data Transformation

The steps below were done with the objective to transform and map the data from raw form to a more appropriate format to the further analysis.

1. Checking the Missing Values: the function *sapply()* was used to check the number of missing values in each column. Eleven NA values were found in the "TotalCharges" column;

```
[1] "Missing values"
customerID      gender  SeniorCitizen      Partner      Dependents      tenure
0              0      0              0              0              0
PhoneService    MultipleLines  InternetService  OnlineSecurity  OnlineBackup  DeviceProtection
0              0      0              0              0              0
TechSupport     StreamingTV    StreamingMovies    Contract  PaperlessBilling  PaymentMethod
0              0      0              0              0              0
MonthlyCharges  TotalCharges    Churn
0              11      0
```

FIGURE 3: MISSING VALUES

These missing values exist because there are 11 observations from the variable "Tenure" which had less than 1 month of contract when they were collected, creating the NA values for the variable "TotalCharges". Once they represented only 0.15% of the total sample, they were removed from the dataset.

2. The category "No internet service" was recoded to "No" for six variables, that are: "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "streamingTV", and "streamingMovies". The function *mapvalues()* was used to recode this category;

```
[1] "Recoding 'No Internet Service' "
OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV
No              3497          3087          3094          3472          2809
No internet service 1520          1520          1520          1520          1520
Yes              2015          2425          2418          2040          2703
StreamingMovies
No              2781
No internet service 1520
Yes              2731
[1] "After recoding"
OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
No              5017          4607          4614          4992          4329          4301
Yes              2015          2425          2418          2040          2703          2731
```

FIGURE 4: RECODING NO INTERNET SERVICE

3. The category “No phone service” was recoded to “No” for the variable “MultipleLines”;

```
[1] "Recoding 'No phone service' "  
Var1 Freq  
1      No 3385  
2 No phone service 680  
3      Yes 2967  
[1] "After recoding "  
Var1 Freq  
1      No 4065  
2      Yes 2967
```

FIGURE 5: RECODING NO PHONE SERVICE

4. Since the minimum tenure was 1 month and maximum was 72 months, this variable was grouped into five tenure groups: “0–12 Month”, “12–24 Month”, “24–48 Months”, “48–60 Month”, “>60 Month”;

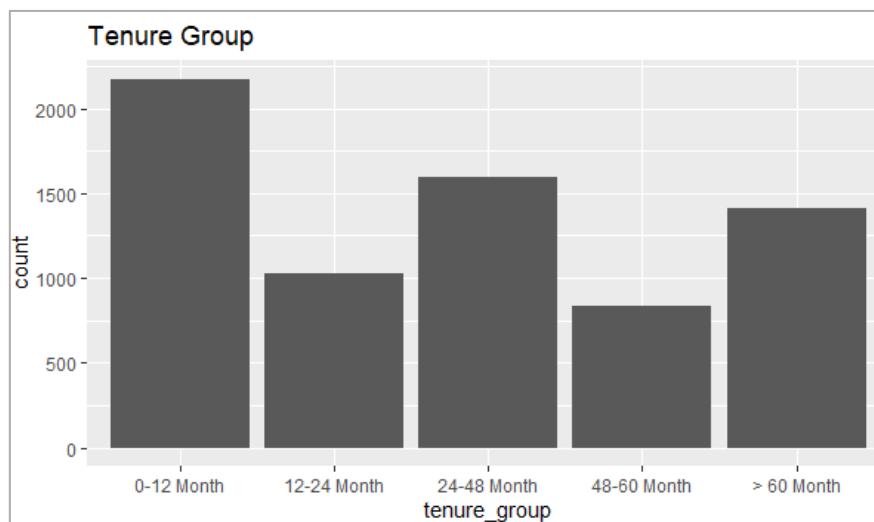


FIGURE 6: TENURE GROUPED BY MONTHS

5. The column “SeniorCitizen” was changed from INT 0 or 1 to factor “Yes” or “No”;
6. Finally, the columns “CustomerID” and “tenure” were removed from the original dataset.

### 4.3 Exploratory Data Analysis and Feature Selection

In Figure 7, the correlation matrix for the numeric variables “MonthlyCharges” and “TotalCharges” shows that these variables have a medium to strong correlation (0.65). To avoid multicollinearity, one of them will be removed from the data set for the analysis.

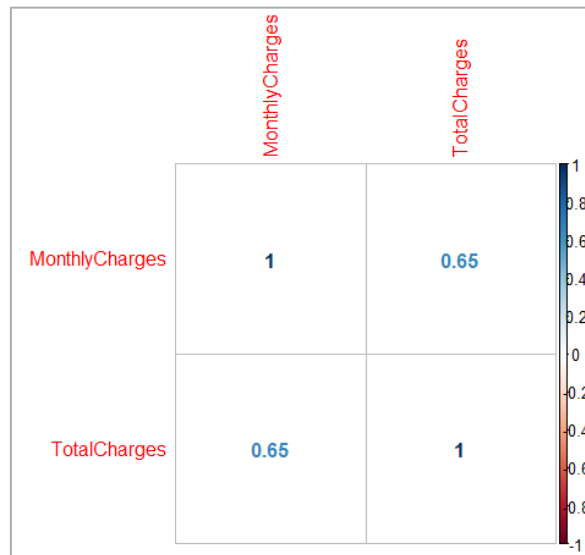


FIGURE 7: CORRELATION MATRIX FOR THE NUMERICAL VARIABLES

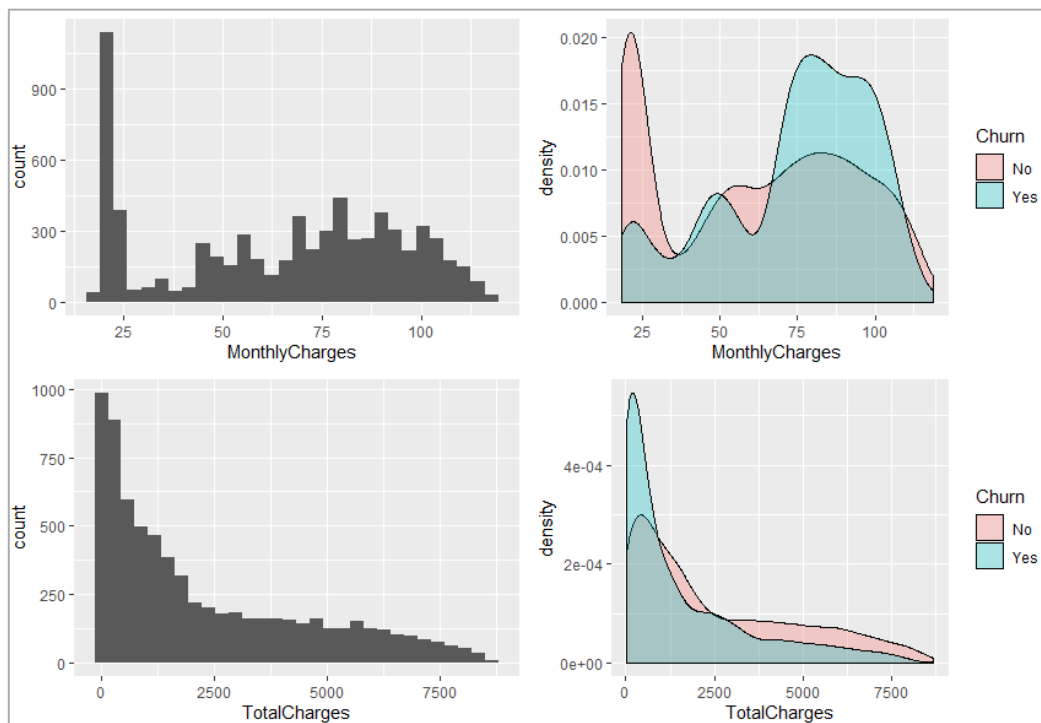


FIGURE 8: DISTRIBUTION OF NUMERICAL VARIABLES

Based on the distribution of the variables, it was decided that the variable “TotalCharges” was removed from the dataset.

The graphs below show the distribution of the categorical variables.

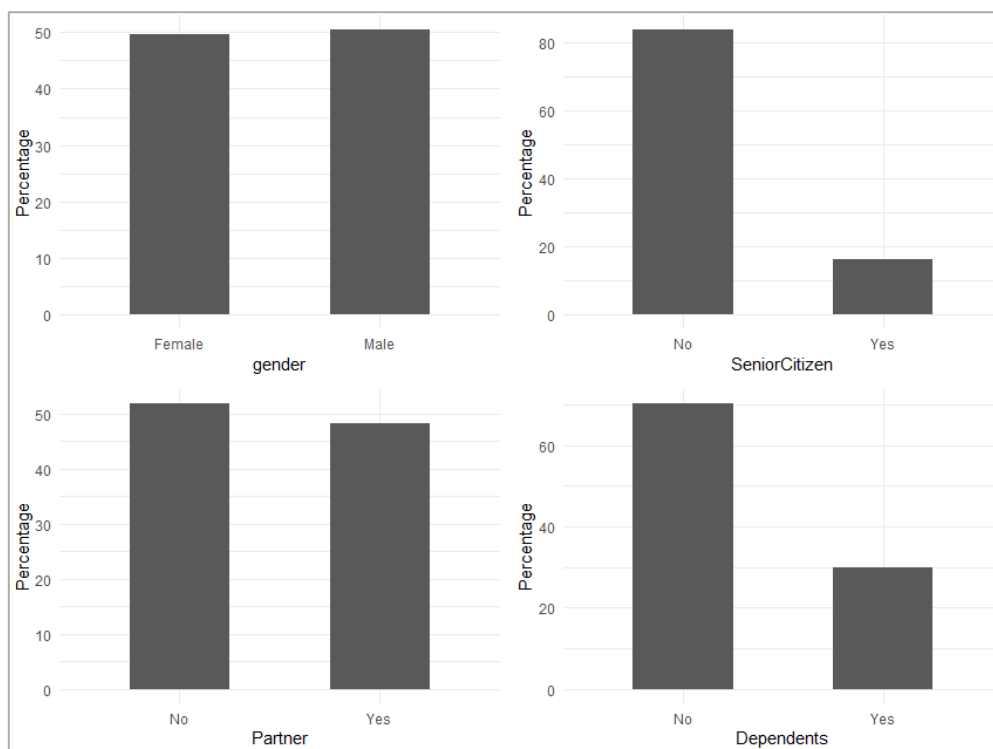


FIGURE 9: CUSTOMER DEMOGRAPHIC DATA

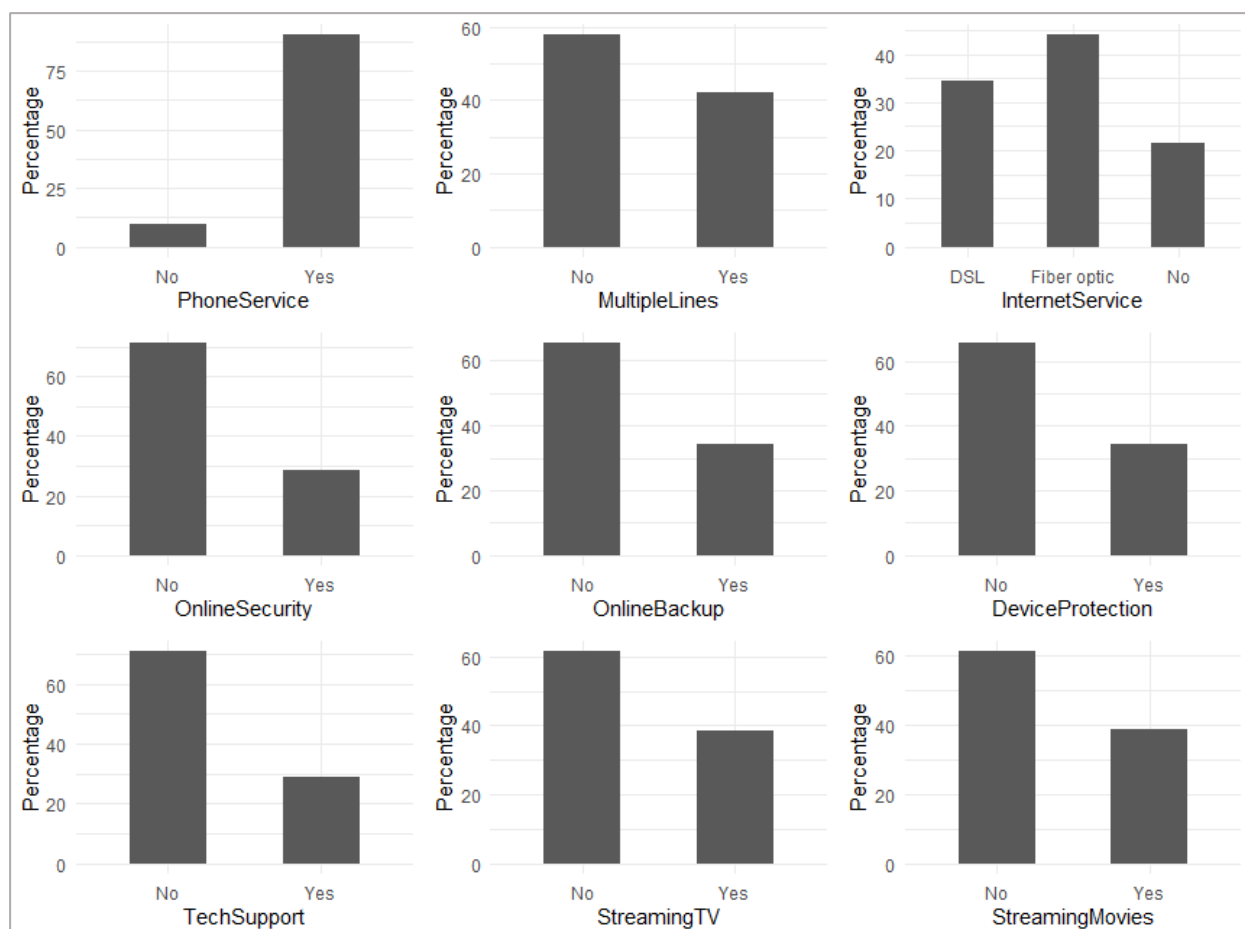


FIGURE 10: SUBSCRIBED SERVICES

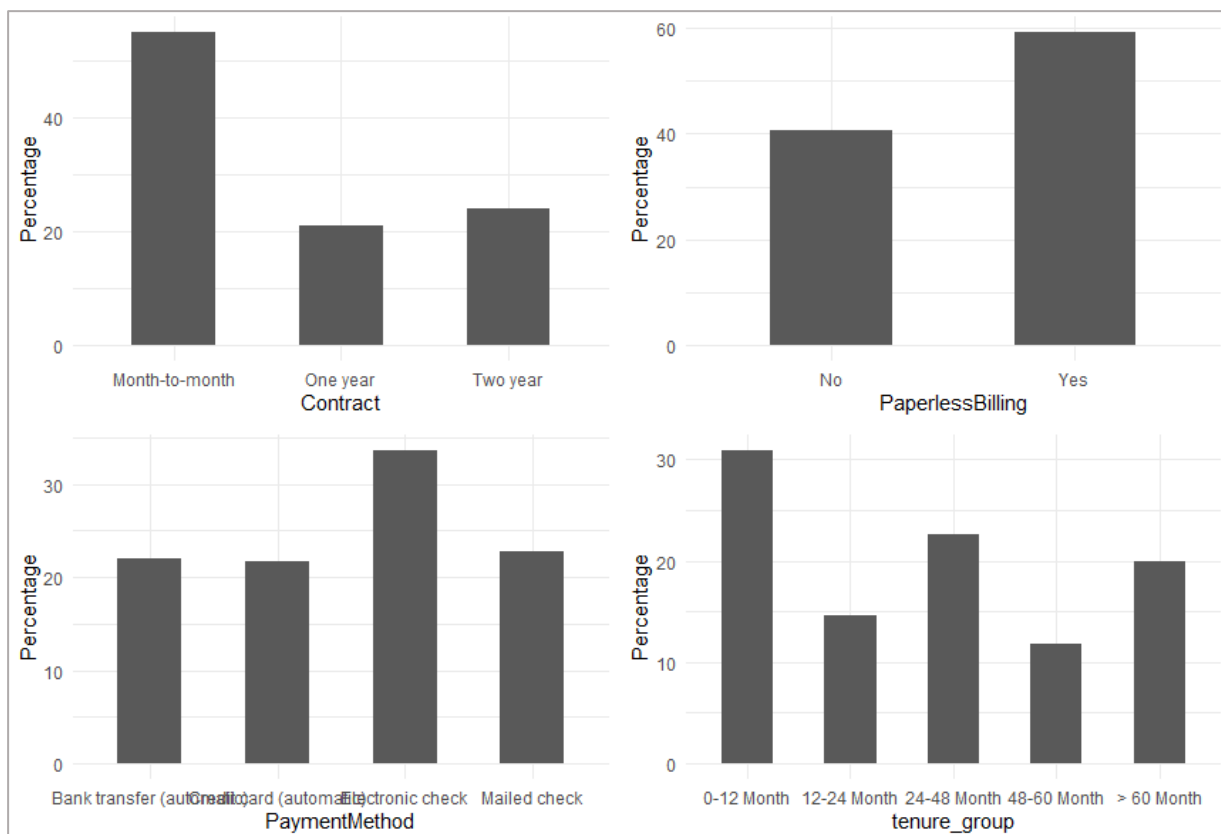


FIGURE 11: CUSTOMER ACCOUNT INFORMATION

All the categorical variables seem to have a reasonably broad distribution, therefore, all of them will be kept for the further analysis.

```
[1] "Telco dataset after data wrangling"
'data.frame': 7032 obs. of 19 variables:
 $ gender      : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner      : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",...: 1 1 1 1 2 2 2 2 1 2 ...
 $ onlineSecurity : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ onlineBackup  : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ TechSupport   : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
 $ StreamingTV   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
 $ Contract      : Factor w/ 3 levels "Month-to-month",...: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",...: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...
 $ churn         : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
 $ tenure_group  : Factor w/ 5 levels "0-12 Month","12-24 Month",...: 1 3 1 3 1 1 2 1 3 5 ...
```

FIGURE 12: TELCO DATASET AFTER DATA WRANGLING AND FEATURE SELECTION

## 4.4 Supervised learning steps

To achieve the results from the Supervised Learning models used in this work, that are **Logistic Regression**, **Decision Tree** and **Random Forest**, four steps will be followed:

1. Split the data into Training and Testing: this is important step as it is necessary to "train" the model on one set of data, and then measure its performance on unseen values or testing dataset to make sure it works well on unseen data. All models used the same datasets.
2. Fit the model on the Training data: also said that the model is being trained.
3. Predict the outcome on the Testing data: the values of unseen data are predicted using the trained model.
4. Measure model performance on testing data: calculate the accuracy score which is the percentage of correctly predicted outcome variables.

### 4.4.1 Logistic Regression

**Logistic Regression** is a supervised learning technique that predicts binary response variables. A **Logistic Regression Model** is to model the conditional probability of  $Y = 1$  given explanatory  $X_1, \dots, X_r$  as a logit function of a linear combination of  $X_1, \dots, X_r$ .

$$P(Y = 1|X_1, \dots, X_r) = \frac{\exp(\alpha + \beta_1 X_1 + \dots + \beta_r X_r)}{1 + \exp(\alpha + \beta_1 X_1 + \dots + \beta_r X_r)}$$

$$P(Y = 0|X_1, \dots, X_r) = 1 - \frac{\exp(\alpha + \beta_1 X_1 + \dots + \beta_r X_r)}{1 + \exp(\alpha + \beta_1 X_1 + \dots + \beta_r X_r)}$$

$$ODDS \rightarrow \text{Log} \left( \frac{P(Y = 1|X_1, \dots, X_r)}{P(Y = 0|X_1, \dots, X_r)} \right) = \alpha + \beta_1 X_1 + \dots + \beta_r X_r$$

The logistic regression models the log-odds of the probability of the target. The *ODDS* is the relative probability of success ( $Y = 1$ ) compared to the probability of failure ( $Y = 0$ ). The model assumes linear relationship between log-odds and predictors and returns coefficients and prediction probability.



## Modeling Steps

### Step 1: Splitting the data into Train and Test.

The final dataset was divided in two parts: Train, with 70% of the data, and Test, with the remaining 30% of the data.

The *set.seed()* function was used in order to ensure the reproducibility of the model.

### Step 2: Fitting the *Logistic Regression* model on the training data.

The *glm()* function (generalized linear models) from the *MASS* package was used to fit the logistic regression model.

All the predictors used in a regression analysis must be numeric, this means that all categorical data must be represented as a number. The *glm* function will automatically *dummy code* any factor type variables used in the model. That is, the model will create a set of binaries (one-zero) variables that represent each category except one that serves as the reference group.

```

Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9628  -0.6889  -0.2929   0.7027   3.1055

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.388036   0.958600  -0.405  0.68563
genderMale    -0.034465   0.077033  -0.447  0.65458
SeniorCitizenYes  0.240459   0.099442   2.418  0.01560 *
PartnerYes     0.011835   0.092220   0.128  0.89788
DependentsYes  -0.098069   0.107307  -0.914  0.36076
PhoneServiceYes -0.635369   0.771228  -0.824  0.41003
MultipleLinesYes  0.262848   0.208277   1.262  0.20694
InternetServiceFiber optic  0.723610   0.948738   0.763  0.44564
InternetServiceNo -0.618501   0.958467  -0.645  0.51873
OnlineSecurityYes -0.386151   0.211018  -1.830  0.06726 .
OnlineBackupYes  -0.132570   0.208419  -0.636  0.52473
DeviceProtectionYes -0.014412   0.207549  -0.069  0.94464
TechSupportYes  -0.287455   0.215557  -1.334  0.18235
StreamingTVYes   0.267311   0.386639   0.691  0.48933
StreamingMoviesYes 0.148139   0.387437   0.382  0.70220
ContractOne year -0.709958   0.125438  -5.660 1.52e-08 ***
ContractTwo year -1.668924   0.214835  -7.768 7.95e-15 ***
PaperlessBillingYes 0.369541   0.088755   4.164 3.13e-05 ***
PaymentMethodCredit card (automatic) -0.085591   0.135496  -0.632  0.52759
PaymentMethodElectronic check  0.333313   0.113915   2.926  0.00343 **
PaymentMethodMailed check -0.015297   0.136651  -0.112  0.91087
MonthlyCharges  0.005077   0.037667   0.135  0.89277
tenure_group12-24 Month -0.840718   0.115587  -7.273 3.50e-13 ***
tenure_group24-48 Month -1.163390   0.119098  -9.768 < 2e-16 ***
tenure_group48-60 Month -1.551697   0.172688  -8.986 < 2e-16 ***
tenure_group> 60 Month -1.930582   0.205985  -9.372 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5702.8  on 4923  degrees of freedom
Residual deviance: 4141.7  on 4898  degrees of freedom
AIC: 4193.7

Number of Fisher Scoring iterations: 6

```

FIGURE 13: SUMMARY OF FULL LOGISTIC REGRESSION MODEL

The statistical significance measures a range from  $< 0.0001$  (highly significant) to some values  $> 0.9$  (highly dubious). Results such as this suggest that certain inputs can be eliminated without affecting the predictive power of the model, so an input selection procedure was used.

In order to select the optimal subset of independent variables among all possible subsets, the function `stepAIC()` from *MASS* package was used. This function performs a **backward model selection** by starting from "maximal" model, with all available variables, and then the variables are dropped based on their significance. The process goes on as long as the AIC value decreases and stops when a minimum value is reached.

```

Call:
glm(formula = formulaLogit, family = binomial(link = "logit"),
     data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9962  -0.6906  -0.2901   0.6976   3.1249

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.088870    0.182444  -5.968 2.40e-09 ***
SeniorCitizenYes    0.261143    0.097422   2.681  0.00735 **
PhoneServiceYes   -1.102457    0.152342  -7.237 4.60e-13 ***
MultipleLinesYes    0.142166    0.096216   1.478  0.13952
OnlineSecurityYes  -0.517796    0.098418  -5.261 1.43e-07 ***
OnlineBackupYes    -0.264070    0.092759  -2.847  0.00442 **
DeviceProtectionYes -0.153252    0.095239  -1.609  0.10759
TechSupportYes    -0.434683    0.101546  -4.281 1.86e-05 ***
ContractOne year   -0.733438    0.123438  -5.942 2.82e-09 ***
ContractTwo year   -1.697571    0.212801  -7.977 1.50e-15 ***
PaperlessBillingYes  0.372139    0.088564   4.202 2.65e-05 ***
PaymentMethodCredit card (automatic) -0.084746    0.135323  -0.626  0.53115
PaymentMethodElectronic check  0.336009    0.113820   2.952  0.00316 **
PaymentMethodMailed check -0.021439    0.136133  -0.157  0.87486
MonthlyCharges      0.030845    0.002344  13.158 < 2e-16 ***
tenure_group12-24 Month -0.852746    0.114405  -7.454 9.07e-14 ***
tenure_group24-48 Month -1.180126    0.116610 -10.120 < 2e-16 ***
tenure_group48-60 Month -1.561364    0.168917  -9.243 < 2e-16 ***
tenure_group> 60 Month -1.950897    0.200788  -9.716 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5702.8  on 4923  degrees of freedom
Residual deviance: 4145.1  on 4905  degrees of freedom
AIC: 4183.1

```

FIGURE 14: LOGISTIC REGRESSION AFTER FEATURE SELECTION PROCESS

According to the significance level viewed in the figure 14, the most-relevant features are: “Contract”, “tenure\_group”, “PaperlessBilling”, “PhoneServices”, “OnlineSecurity”, “TechSupport”, and “MonthlyCharges”. It can be assumed that these variables do have a significant effect on the customer churning.

Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			4923	5702.8	
SeniorCitizen	1	111.95	4922	5590.8	< 2.2e-16 ***
PhoneService	1	0.01	4921	5590.8	0.92204
MultipleLines	1	0.58	4920	5590.2	0.44820
OnlineSecurity	1	139.76	4919	5450.5	< 2.2e-16 ***
OnlineBackup	1	16.17	4918	5434.3	5.779e-05 ***
DeviceProtection	1	3.40	4917	5430.9	0.06504 .
TechSupport	1	45.71	4916	5385.2	1.368e-11 ***
Contract	2	757.28	4914	4627.9	< 2.2e-16 ***
PaperlessBilling	1	72.94	4913	4555.0	< 2.2e-16 ***
PaymentMethod	3	88.06	4910	4466.9	< 2.2e-16 ***
MonthlyCharges	1	156.36	4909	4310.5	< 2.2e-16 ***
tenure_group	4	165.47	4905	4145.1	< 2.2e-16 ***

---  
 signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

FIGURE 15: ANALYSIS OF DEVIANCE ANOVA

Analyzing the deviance table, it is possible to see the drop-in deviance when adding each variable one at a time. Adding “Contract”, “MonthlyCharges”, and “tenure\_group” significantly reduces the residual deviance. On the other hand, some variables such as “SeniorCitizen”, “TechSupport” and “OnlineBackup” seem to improve the model less even though they all have low p-values.

Finally, in order to see the explanatory power of the model, the *logRegR2()* function from *descr* package was used. To measure the model fit (like  $R^2$  in the linear regression), there are three so called pseudo  $R^2$  statistics: the McFadden, the Cox & Snell and the Nagelkerke. Here it is possible to see that the explanatory power of the model is reasonable (if  $x > 0.2$ ) to good (if  $x > 0.4$ ).

[1] "Pseudo R2"	
Chi2	1557.694
Df	18
Sig.	0
Cox and Snell Index	0.2711937
Nagelkerke Index	0.3953635
McFadden's R2	0.2731474

FIGURE 16: MODEL FIT STATISTICS

**Step 3:** Predict the outcome on the testing data for the Logistic Regression model.

After creating a model with the training dataset, now the function *predict()* is executed with the testing dataset in order to assess the predictive ability of the Logistic Regression model.

**Step 4:** Measure model performance on testing data.

The *confusionMatrix()* function from *caret* package was executed to show the accuracy of the model. This function calculates a cross-tabulation of observed and predicted classes with associated statistics. The accuracy of the Logistic Regression here is 0.8088, meaning that the model correctly predicted around 81% of the customer churn and non-churn events for this particular dataset.

```
Confusion Matrix and Statistics

      Reference
Prediction  No  Yes
      No  1400  255
      Yes   148  305

      Accuracy : 0.8088
      95% CI : (0.7914, 0.8254)
      No Information Rate : 0.7343
      P-value [Acc > NIR] : 7.089e-16

      Kappa : 0.4782

      Mcnemar's Test P-value : 1.290e-07

      Sensitivity : 0.9044
      Specificity : 0.5446
      Pos Pred Value : 0.8459
      Neg Pred Value : 0.6733
      Prevalence : 0.7343
      Detection Rate : 0.6641
      Detection Prevalence : 0.7851
      Balanced Accuracy : 0.7245

      'Positive' Class : No
```

FIGURE 17: CONFUSION MATRIX AND ACCURACY FOR THE LOGISTIC REGRESSION MODEL

#### 4.4.2 Decision Tree

**Decision Trees** (DTs) are a non-parametric supervised learning method used for regression and classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

The process goes from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called **classification trees**; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

A tree is built by splitting the source set, constituting the root node of the tree, into subsets - which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions (Shalev-Shwartz & Ben-David, 2014).

#### Modeling Steps

**Step 1:** Splitting the data into Train and Test. This step has already been executed in the previous model.

**Step 2:** Fitting the Decision Tree model on the training data.

The function used to fit the decision tree was the *rpart()* (recursive partitioning and regression trees) from the *rpart* package. The model was fitted with all the response variables from the dataset, with type = "Class".

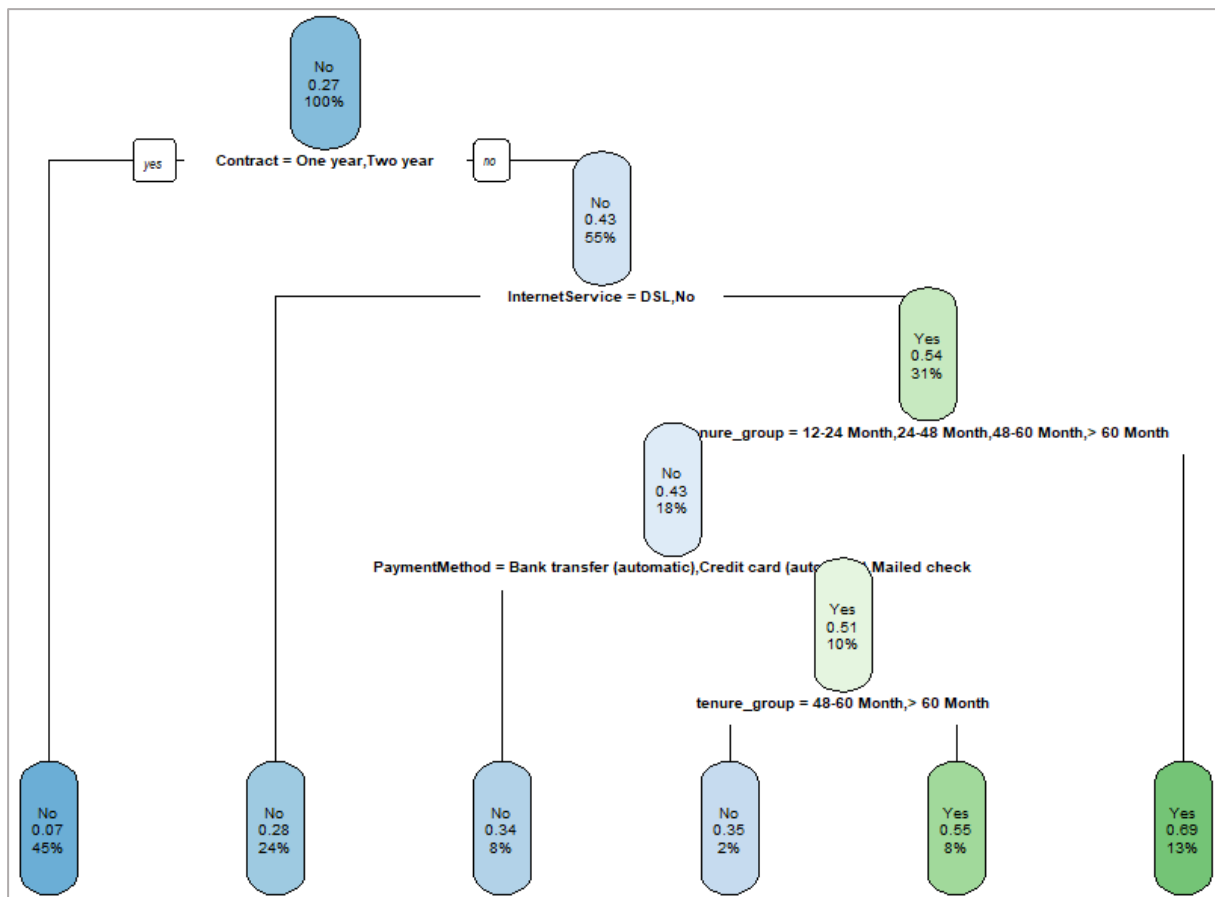
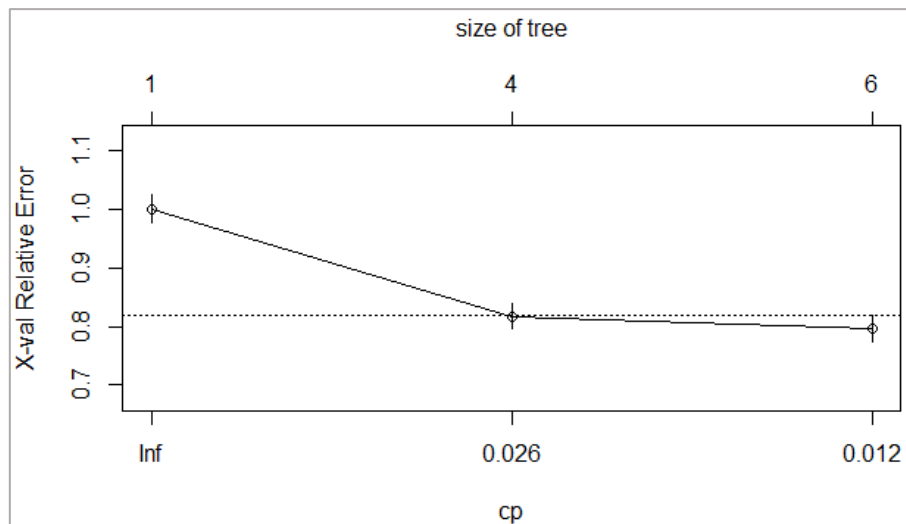


FIGURE 18: DECISION TREE MODEL

The maximal tree represents the most complicated model that could be constructed from a set of training data.

The model printed above (figure 18) may produce good predictions on the training set, but is likely to overfit the data, leading to poor test set performance. This is because the resulting tree might be too complex. A smaller tree with fewer splits might lead to lower variance and better interpretation at the cost of a little bias. The better strategy is to grow a very large tree  $T_0$ , and then prune it back in order to obtain a subtree. The goal is to select a subtree that leads to the lowest test error rate in order to determine the best way to prune the tree. Given a subtree, we can estimate its test error using cross-validation or the validation set approach. (James, Witten, Hastie, & Tibshirani, 2017).

Therefore, a post-pruning method was performed, where nodes and branches with only a minor impact on the tree's overall accuracy are removed after the fact.



**FIGURE 19: COMPLEXITY PARAMETER TABLE**

The *plotcp()* function generates a visualization (figure 19) for the error rate versus model complexity, which provides insights into the optimal cut point for pruning. The horizontal dotted line identifies the point at which the error rate becomes statistically similar to the most complex model.

Thus, a new decision tree is created with the parameter  $cp = 0.026$  and generates the graph below (figure 20).



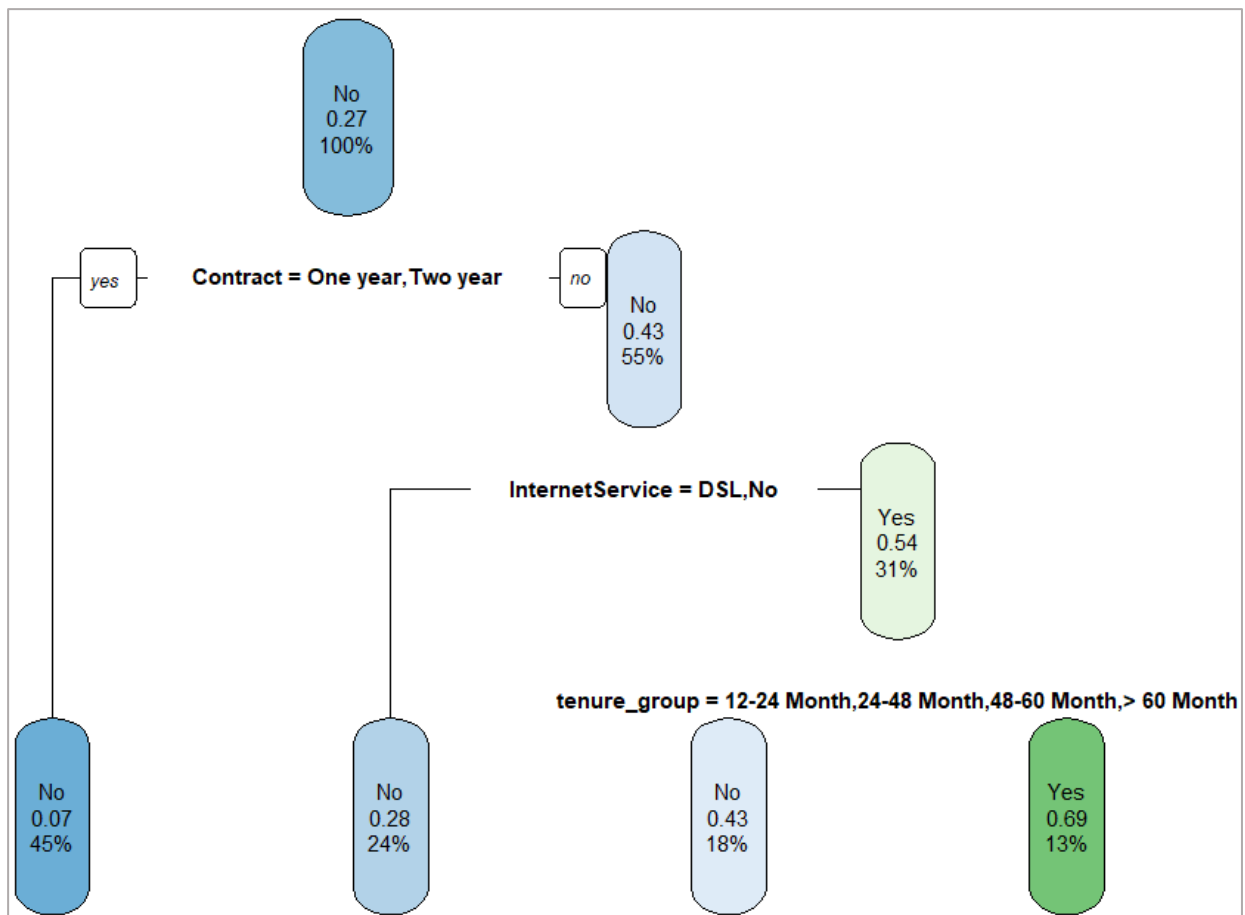


FIGURE 20: PRUNED DECISION TREE

Out of the variables used, “Contract”, “InternetService” and “Tenure\_group” are the most important variables to predict customer churn or not churn. If a customer in a **one-year** or **two-year contract**, he or she is less likely to churn.

By this graph is also possible to see that if a customer has a Contract type different from One or Two-year and Internet Service equals Fiber Optic, he or she has a higher probability to churn. Also, the customer with Tenure over twelve months is less likely to churn.

**Step 3:** Predict the outcome on the testing data for the Decision Tree model.

After creating a model with the training dataset, now the function *predict()* is executed with the testing dataset in order to assess the predictive ability of the Decision Tree model.

**Step 4:** Measure model performance on testing data.

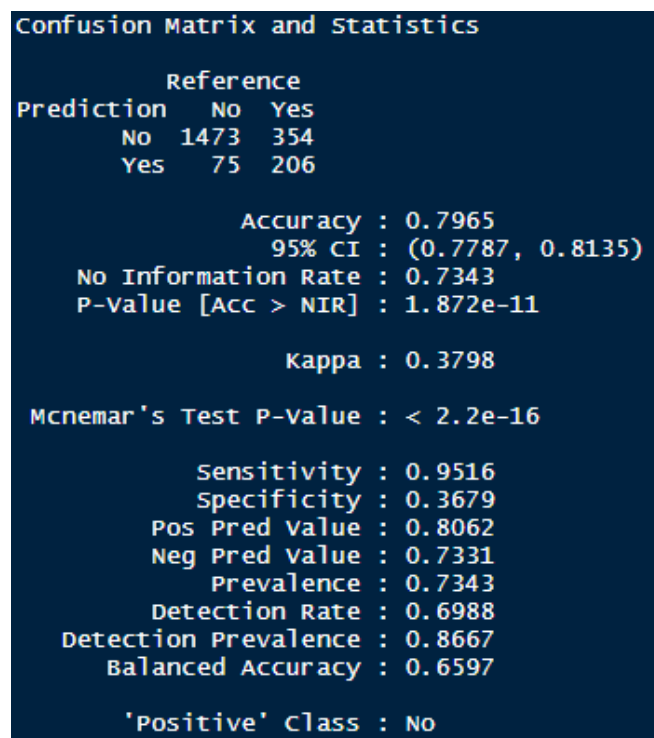


FIGURE 21: CONFUSION MATRIX AND ACCURACY FOR THE DECISION TREE MODEL

The accuracy of the Decision Tree has worsened compared with the accuracy of the Logistic Regression, from 0.8088 to 0.7965.

#### 4.4.3 Random Forest

A **Random forest** is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm  $A$  on the training set  $S$  and an additional random vector,  $\theta$ , where  $\theta$  is sampled independent and identically distributed from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set (Shalev-Shwartz & Ben-David, 2014).

### Modeling Steps

**Step 1:** Splitting the data into Train and Test. This step has already been executed in the first model.

**Step 2:** Fitting the Random Forest model on the training data.

The function *randomForest()* from the **caret** package (Classification and Regression Training) was used to generate the model.

```
Call:
randomForest(formula = Churn ~ ., data = training)
  Type of random forest: classification
    Number of trees: 500
No. of variables tried at each split: 4

OOB estimate of error rate: 21.97%
Confusion matrix:
      No Yes class.error
No  3215 400   0.1106501
Yes   682 627   0.5210084
```

FIGURE 22: RANDOM FOREST MODEL

The error rate is relatively low when predicting “No”, around 11%, but it is much higher when predicting “Yes”, 52%.

**Step 3:** Predict the outcome on the testing data for the Random Forest model.

After creating the first model with the training dataset, the function *predict()* is executed with the testing dataset in order to assess the predictive ability of the Random Forest model.

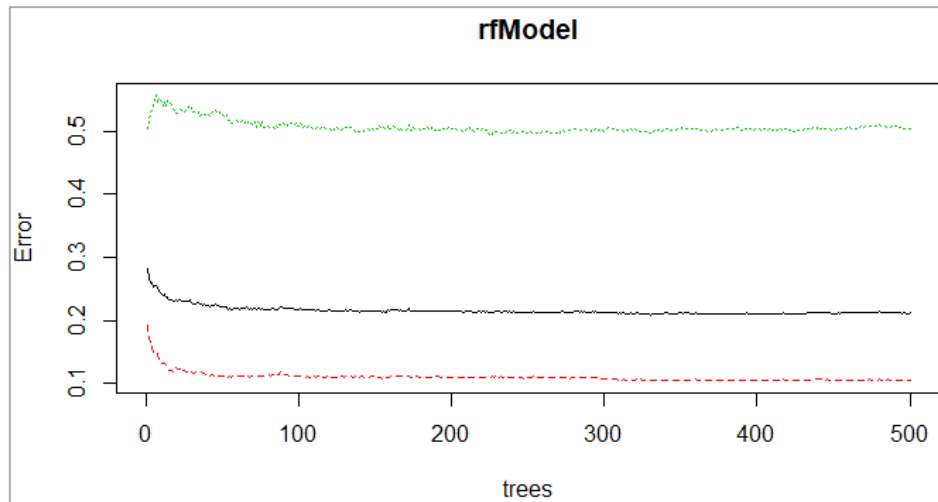
**Step 4:** Measure model performance on testing data.

Confusion Matrix and Statistics		
Prediction	Reference	
	No	Yes
	No 1402	279
Yes	146	281
Accuracy : 0.7984		
95% CI : (0.7806, 0.8153)		
No Information Rate : 0.7343		
P-Value [Acc > NIR] : 4.447e-12		
Kappa : 0.4409		
McNemar's Test P-Value : 1.524e-10		
Sensitivity : 0.9057		
Specificity : 0.5018		
Pos Pred Value : 0.8340		
Neg Pred Value : 0.6581		
Prevalence : 0.7343		
Detection Rate : 0.6651		
Detection Prevalence : 0.7974		
Balanced Accuracy : 0.7037		
'Positive' Class : No		

FIGURE 23: CONFUSION MATRIX AND ACCURACY FOR THE RANDOM FOREST MODEL

The accuracy of the Random Forest model (0.7984) slightly improved from the Decision Tree (0.7965) but is still worse than the Logistic Regression (0.8088).

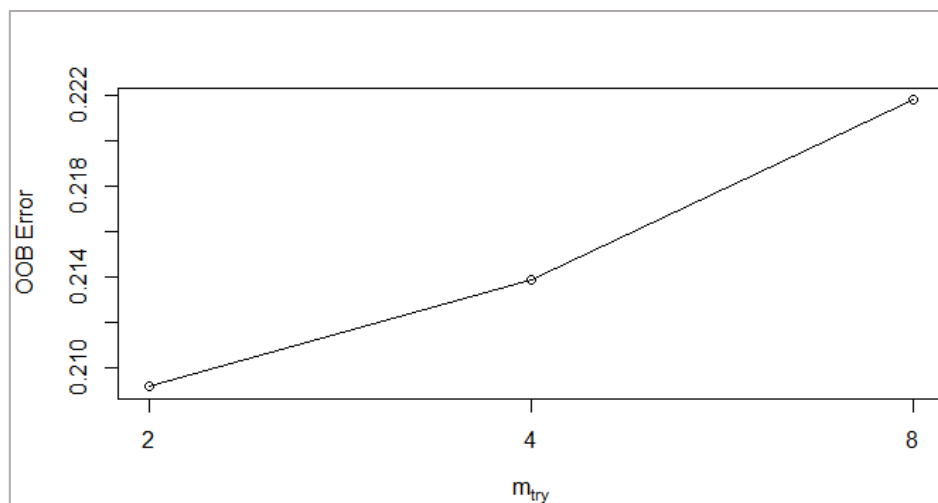
The plot below (figure 24) is used to help to determine the number of trees for the random forest model. As the number of trees increases, the out-of-bag (OOB) error rate decreases, and then becomes almost constant. The **OOB error** is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training (Out of bag error, 2019).



**FIGURE 24: RANDOM FOREST ERROR RATE**

Based on the graph above, it is not possible to decrease the OOB error rate after around 100 to 200 trees.

The *mtry* parameter refers to the number of variables available for splitting at each tree node. The plot of the figure 25, generated by the function *tuneRF()*, is used to define the optimal number of *mtry* to choose.



**FIGURE 25: OOB ERROR BY THE NUMBER OF MTRY**

OOB error rate is at the lowest when *mtry* is equal 2. Therefore, *mtry* = 2 is chosen.

At this point, the Random Forest model is re-executed with the tuned parameters. The steps 2, 3 and 4 were repeated.

**Step 2:** Fitting the Random Forest model after Tuning (ntree = 200, mtry = 2):

```
call:
 randomForest(formula = churn ~ ., data = training, ntree = 200,
 mtry = 2, importance = TRUE, proximity = TRUE)
      Type of random forest: classification
      Number of trees: 200
No. of variables tried at each split: 2

      OOB estimate of  error rate: 21.04%
Confusion matrix:
      No Yes class.error
No  3308 307  0.08492393
Yes   729 580  0.55691367
```

FIGURE 26: TUNED RANDOM FOREST

The OOB error rate decreased to 21.04% from 21.97% on figure 22.

**Step 3:** Predict the outcome on the testing data for the Random Forest after tuning.

**Step 4:** Measure model performance on testing data.

```
Confusion Matrix and Statistics

      Reference
Prediction  No  Yes
      No  1437  298
      Yes   111  262

      Accuracy : 0.806
      95% CI   : (0.7884, 0.8227)
      No Information Rate : 0.7343
      P-value [Acc > NIR] : 8.876e-15

      Kappa : 0.4434

      Mcnemar's Test P-value : < 2.2e-16

      Sensitivity : 0.9283
      Specificity : 0.4679
      Pos Pred Value : 0.8282
      Neg Pred Value : 0.7024
      Prevalence : 0.7343
      Detection Rate : 0.6817
      Detection Prevalence : 0.8231
      Balanced Accuracy : 0.6981

      'Positive' class : No
```

FIGURE 27: CONFUSION MATRIX AND ACCURACY OF THE TUNED RANDOM FOREST

For the tuned Random Forest model, both accuracy (from 0.7984 to 0.806) and sensitivity (from 0.9057 to 0.9283) improved, however the specificity worsened compared to the first Random Forest model (from 0.5018 to 0.4679).

Finally, the function *varImpPlot()* was executed to show the most important variables based on the Mean Decrease Accuracy and Mean Decrease Gini index.

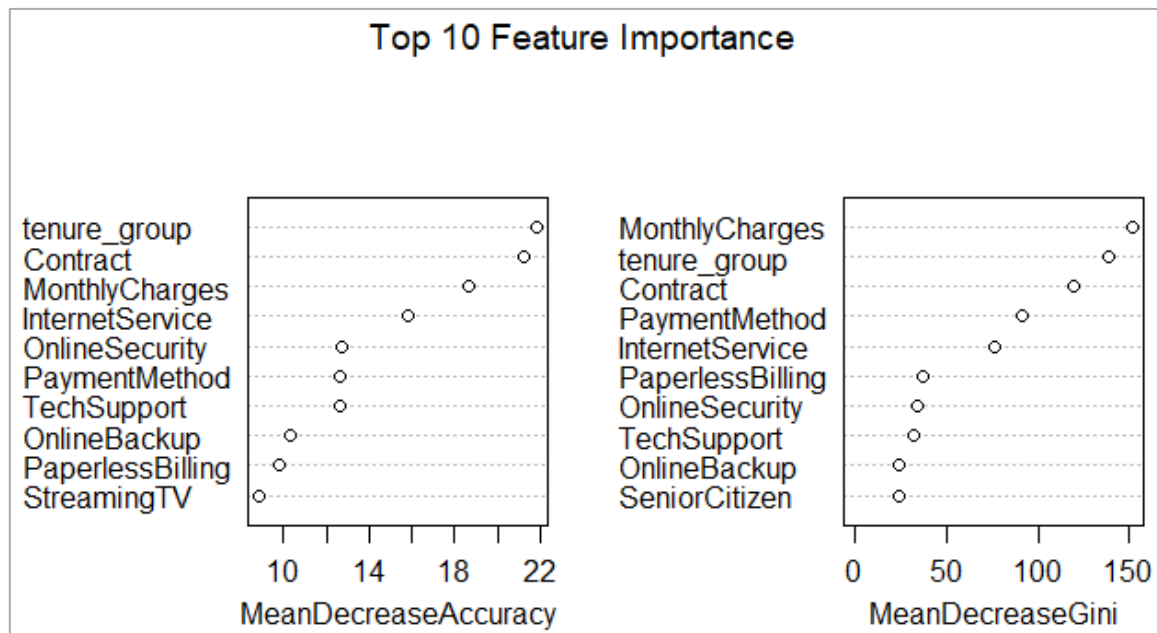
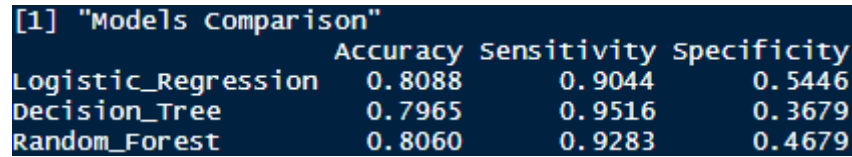


FIGURE 28: RANDOM FOREST FEATURE IMPORTANCE

Here is possible to see that the top tree variables in terms of importance for the random forest model are “tenure\_group”, “Contract” and “MonthlyCharges”.

## 4.5 Summary

Based on the results of the classification models (figures 17, 21, 27), here the key metrics are summarized.



[1] "Models Comparison"			
	Accuracy	Sensitivity	Specificity
Logistic_Regression	0.8088	0.9044	0.5446
Decision_Tree	0.7965	0.9516	0.3679
Random_Forest	0.8060	0.9283	0.4679

FIGURE 29: KEY METRICS FOR MODELS COMPARISON

The Logistic Regression model has a slightly higher **Accuracy** (out of all the classes, how much the model predicted correctly, both churn and no-churn) and **Specificity** (the proportion of negative results out of the number of samples which were actually negative) while the Decision Tree model has the higher value for **Sensitivity** (the proportion of positive results out of the number of samples which were actually positive).

Throughout the analysis, I have learned several important things:

- Features like as “tenure\_group”, “Contract”, “PaperlessBilling”, “MonthlyCharges” and “InternetService” **seem to have a significant influence** on customer churn.
- There **does not seem to be a relationship** between customer churn and the variables “gender”, “SeniorCitizen”, “Partner”, “Dependents”, and “PhoneService”.
- Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are **more likely to churn**; On the other hand, customers with one- or two-year contract, with longer than 12 months tenure, who are not using PaperlessBilling, are less likely to churn.

Finally, from the above example, it is possible to see that the tree models can be used for customer churn analysis for this particular dataset equally fine. However, based on the calculated metrics, the Logistic Regression model should be chosen, because it has the highest accuracy and specificity and a quite well sensitivity level.



## 5. Skills acquire in the internship

The lessons learned that can be pointed out during the internship and the development of this thesis were:

- In this project I was able to reproduce a data analysis methodology with the implementation of machine learning techniques, from the initial analysis, data wrangling, the execution of different binary classification models and finally the comparison of the final metrics.
- There are numerous other machine learning methods and models that could have been applied in this context, as well as various marketing activities (i.e. customer churning, segment analysis, customer experience enhancement, creation of new revenue streams, development of new products and services, etc.) that can be enhanced by properly using these tools.
- With the insights provided by machine learning, companies can tailor their marketing efforts by providing better customer service and ultimately delivering a more personalized experience to their clients.

Finally, I realized that Data Science is a lifelong learning path. This master means a lot to me because it represents a complete change of my career path. The internship opened the door to this new world, showed me some of the challenges faced by a business intelligence company, and now my goal is to dig deeper into the data.

## 6. References

- Churn Attrition*. (2019, 11). Retrieved from Wikipedia:  
[https://en.wikipedia.org/wiki/Customer\\_attrition](https://en.wikipedia.org/wiki/Customer_attrition)
- Churn Rate Definition*. (2019, 11). Retrieved from Investopedia:  
<https://www.investopedia.com/terms/c/churnrate.asp>
- Confusion Matrix*. (2019, 12). Retrieved from rdocumentation:  
<https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix>
- Data protection in the EU*. (2019, 11). Retrieved from European Commission:  
[https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en)
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with applications in R*. Springer.
- Out of bag error*. (2019, 12). Retrieved from Wikipedia: [https://en.wikipedia.org/wiki/Out-of-bag\\_error](https://en.wikipedia.org/wiki/Out-of-bag_error)
- Random forest*. (2019, 12). Retrieved from Wikipedia: [https://en.wikipedia.org/wiki/Random\\_forest](https://en.wikipedia.org/wiki/Random_forest)
- Reichheld, F. (2014, 10). *Prescription for Cutting Costs*. Retrieved from Bain & Company:  
[http://www2.bain.com/Images/BB\\_Prescription\\_cutting\\_costs.pdf](http://www2.bain.com/Images/BB_Prescription_cutting_costs.pdf)
- SAS Marketing Analytics*. (2019, 11). Retrieved from SAS:  
[https://www.sas.com/en\\_us/insights/marketing/marketing-analytics.html](https://www.sas.com/en_us/insights/marketing/marketing-analytics.html)
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.