

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 were applied and evaluated using a public domain dataset from the telecommunications industry.

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

Table of Contents

List of Figures	3
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
4.1 Data Transformation	11
4.2 Exploratory Data Analysis and Feature Selection	14
4.3 Supervised learning steps	17
4.2.1 Binary Logistic Regression	17
4.2.2 Decision Tree	23
4.2.3 Random Forest	26
4.2.3.1 Tunning the Random Forest model	27
4.2.4 Summary	30
5. Skills acquire in the internship	31
6. References	32

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	12
Figure 5: Recoding No phone service	12
Figure 6: Tenure grouped by months	13
Figure 7: Dataset after data wrangling	13
Figure 8: Correlation Matrix for the numerical variables	14
Figure 9: Distribution of Numerical Variables	14
Figure 10: Customer Demographic Data	15
Figure 11: Subscribed Services	15
Figure 12: Customer Account Information	16
Figure 13: Summary of Logistic Regression model	19
Figure 14: Analysis of Deviance ANOVA	20
Figure 15: Confusion Matrix and Accuracy for the Logistic Regression model	21
Figure 16: Odds Ratio	22
Figure 17: Decision Tree model	24
Figure 18: Confusion Matrix and Accuracy for Decision Tree model	25
Figure 19: Random Forest model	26

Figure 20: Confusion Matrix and Accuracy for the Random Forest model..... 27

Figure 21: Random Forest Error Rate 28

Figure 22: OOB error by the number of mtry..... 28

Figure 23: Tuned Random Forest 29

Figure 24: Confusion Matrix and Accuracy of the Tuned Random Forest 29

Figure 25: Random Forest Feature Importance 30

Figure 26: Models Comparison..... 30

1. Analysis of the Company

The internship has been doing at *Nunatac*, 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, Telco and Large Companies in other sectors.

The main characteristics 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 your 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*, ie 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.

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 objective of the internship is the development of data analysis skills and the elaboration of statistical models with the aid of computer systems.

The internship-related activities are 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 is 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). Marketing analytics 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.

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.

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 (GDPR) guidelines, the model developed for this thesis will use a generic database.

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

(SAS - https://www.sas.com/en_us/insights/marketing/marketing-analytics.html).

(source: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation)

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 convince existing customers to buy again and stay loyal to, in this case, to the telecom company. Basically, it is a measure to ensure that customers will not decide to migrate to another company.

In this study case, we will model contractual churn in the telecom business model, where customers can have multiple services with a telecommunications company under one master agreement which defines whether customer is still active, or has churned, which means they have terminated their contract.

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 commonly 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.

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 having to do with 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.

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.

(https://en.wikipedia.org/wiki/Churn_rate)

(<https://www.investopedia.com/terms/c/churnrate.asp>)

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%.

(<https://www.investopedia.com/terms/c/churnrate.asp>)

4. Description of the research methods and analysis

Predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk 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 tool used to do the analysis, modeling and visualization was **RStudio**, an integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

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 ...
 $ 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 3 1 1 3 ...
 $ 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**
- **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 — numeric;
- **TotalCharges**: The total amount charged to the customer — numeric;
- **Churn**: Whether the customer churned or not (Yes or No).

The features in this dataset include the following:

- **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;

In this dataset of over 7000 customers, around 26% of them has left in the last month. This is critical to business because it is often more expensive to acquire new customers than to keep existing ones.

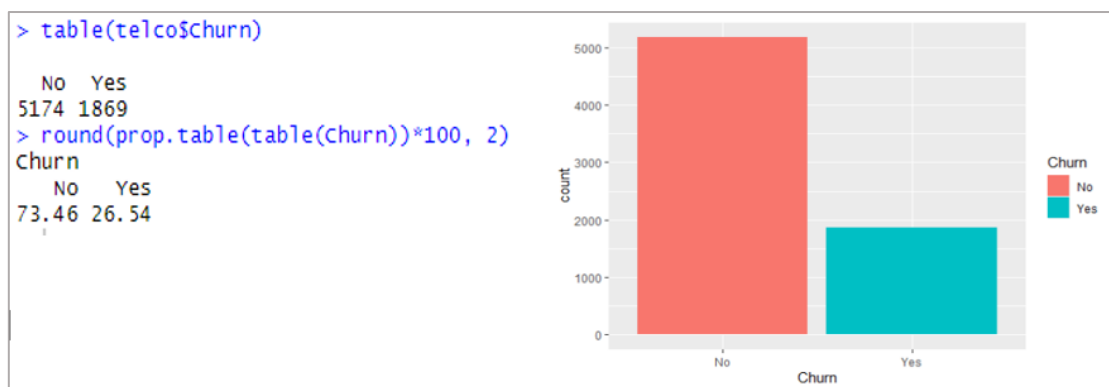


FIGURE 2: EXPLORING CHURN DISTRIBUTION

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 we can see that there are

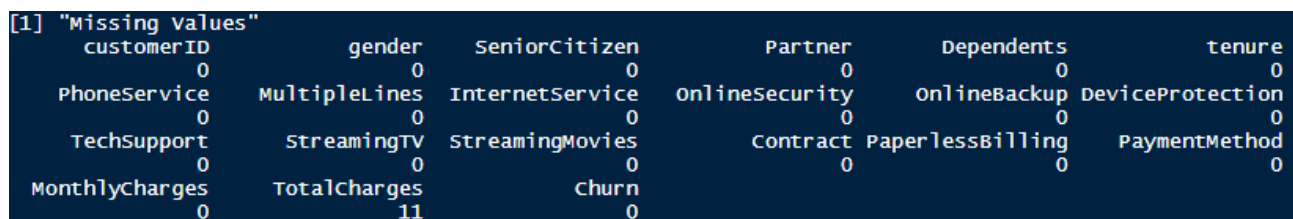
over 26% churned customers and over 73% non-churned customers. There is some class imbalance, but not a severe one.

Typically, if the minority class is less than 5% then we should worry and explore computational ways to increase the minority class or decrease the majority class with oversampling or undersampling techniques.

4.1 Data Transformation

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

1. The function *sapply* was used to check the number of missing values in each column. It was found that there are 11 missing values in “TotalCharges” columns.



```
[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” who had less than 1 month of contract when they were collected, and this created the missing values for the variable “TotalCharged”. Once they represent only 0.15% of the sample, they were removed from the dataset.

- The category “No internet service” will be recoded to “No” for six columns, they 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	StreamingMovies
No	3497	3087	3094	3472	2809	
No internet service	1520	1520	1520	1520	1520	1520
Yes	2015	2425	2418	2040	2703	

```
[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

- The category “No phone service” will be recoded to “No” for the column “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

- Since the minimum tenure is 1 month and maximum is 72 months, this variable will be grouped into five tenure groups: “0–12 Month”, “12–24 Month”, “24–48 Months”, “48–60 Month”, “>60 Month”

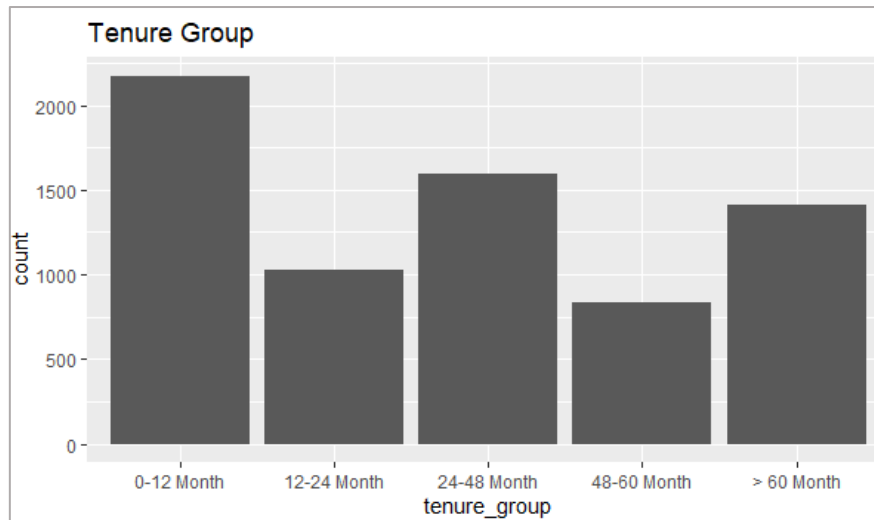


FIGURE 6: TENURE GROUPED BY MONTHS

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

```
[1] "telco dataset after data wrangling"
'data.frame': 7032 obs. of 20 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 1 2 1 ...
 $ 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 ...
 $ 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 ...
 $ tenure_group   : Factor w/ 5 levels "0-12 Month","12-24 Month",...: 1 3 1 3 1 1 2 1 3 5 ...
```

FIGURE 7: DATASET AFTER DATA WRANGLING

4.2 Exploratory Data Analysis and Feature Selection

In the figure 8 the correlation matrix for the numerical variables “MonthlyCharges” and “TotalCharges” have a medium to strong correlation. In order to avoid multicollinearity, one of them will be removed from the dataset for the analysis.

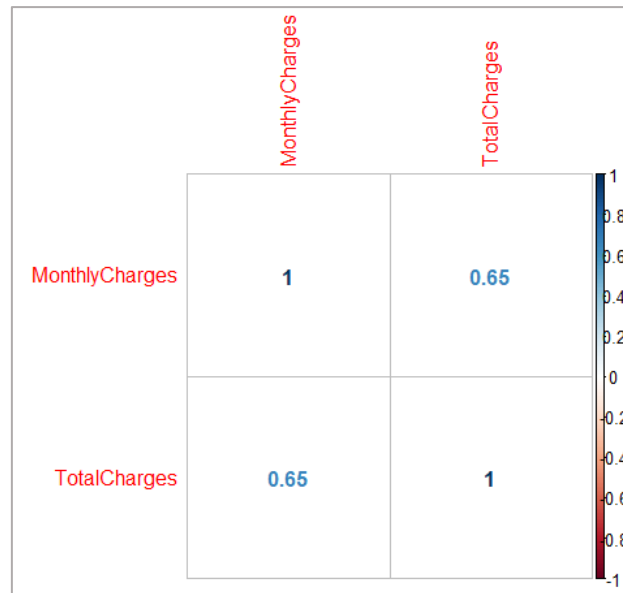


FIGURE 8: CORRELATION MATRIX FOR THE NUMERICAL VARIABLES

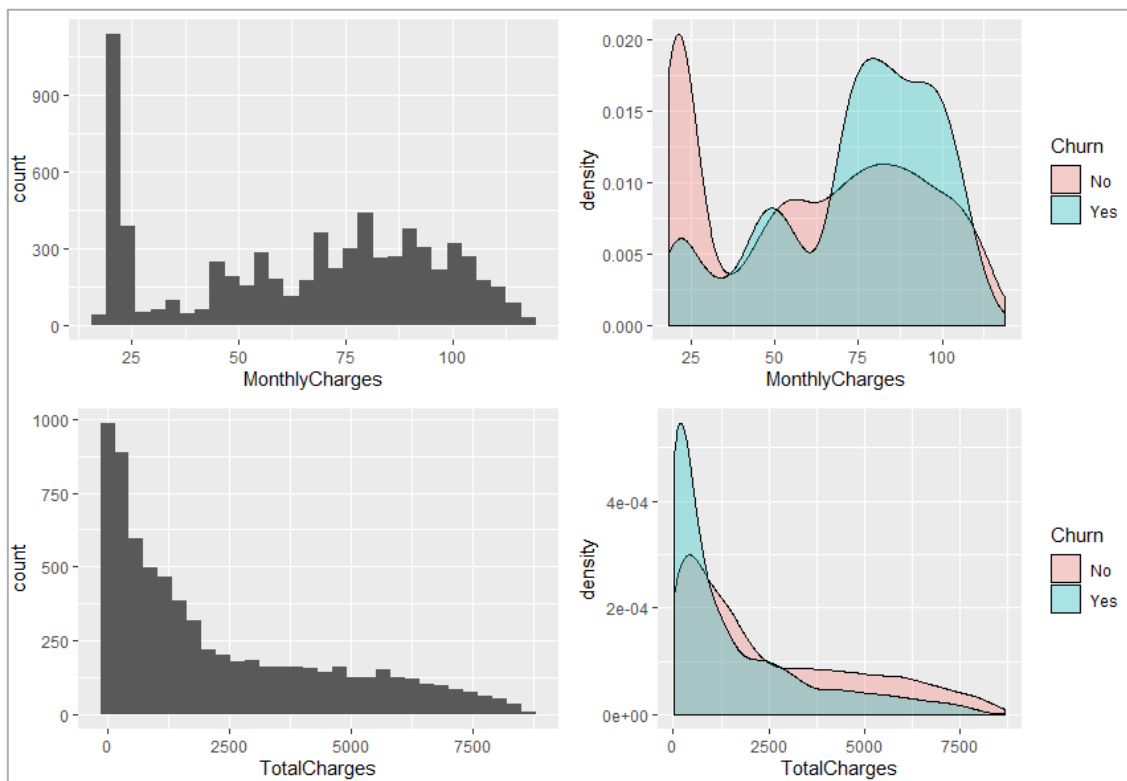


FIGURE 9: DISTRIBUTION OF NUMERICAL VARIABLES

The variable “TotalCharges” will be removed from the dataset.

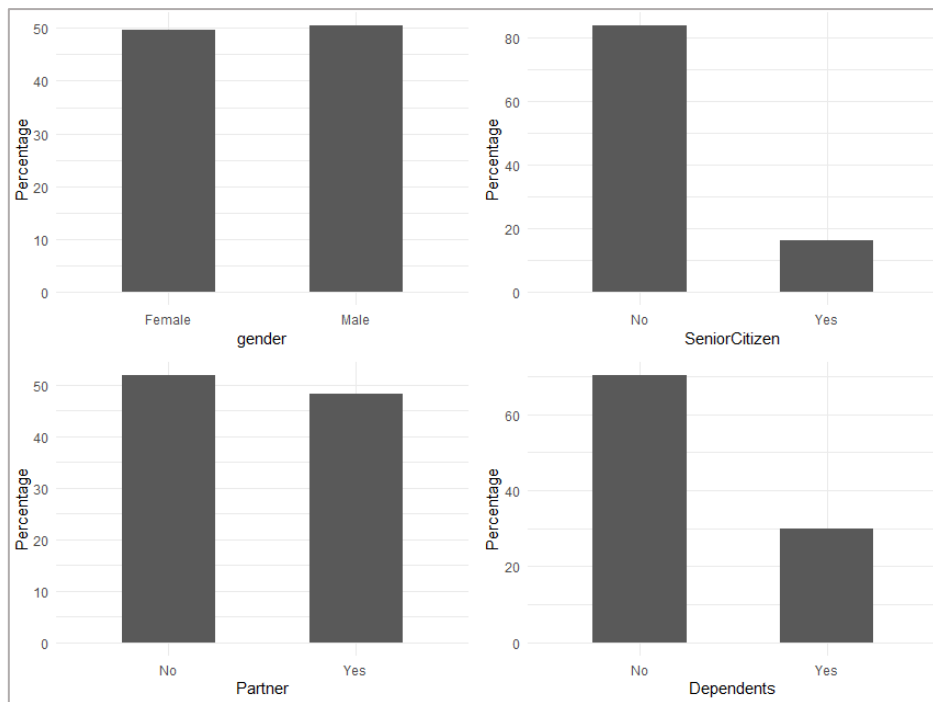


FIGURE 10: CUSTOMER DEMOGRAPHIC DATA

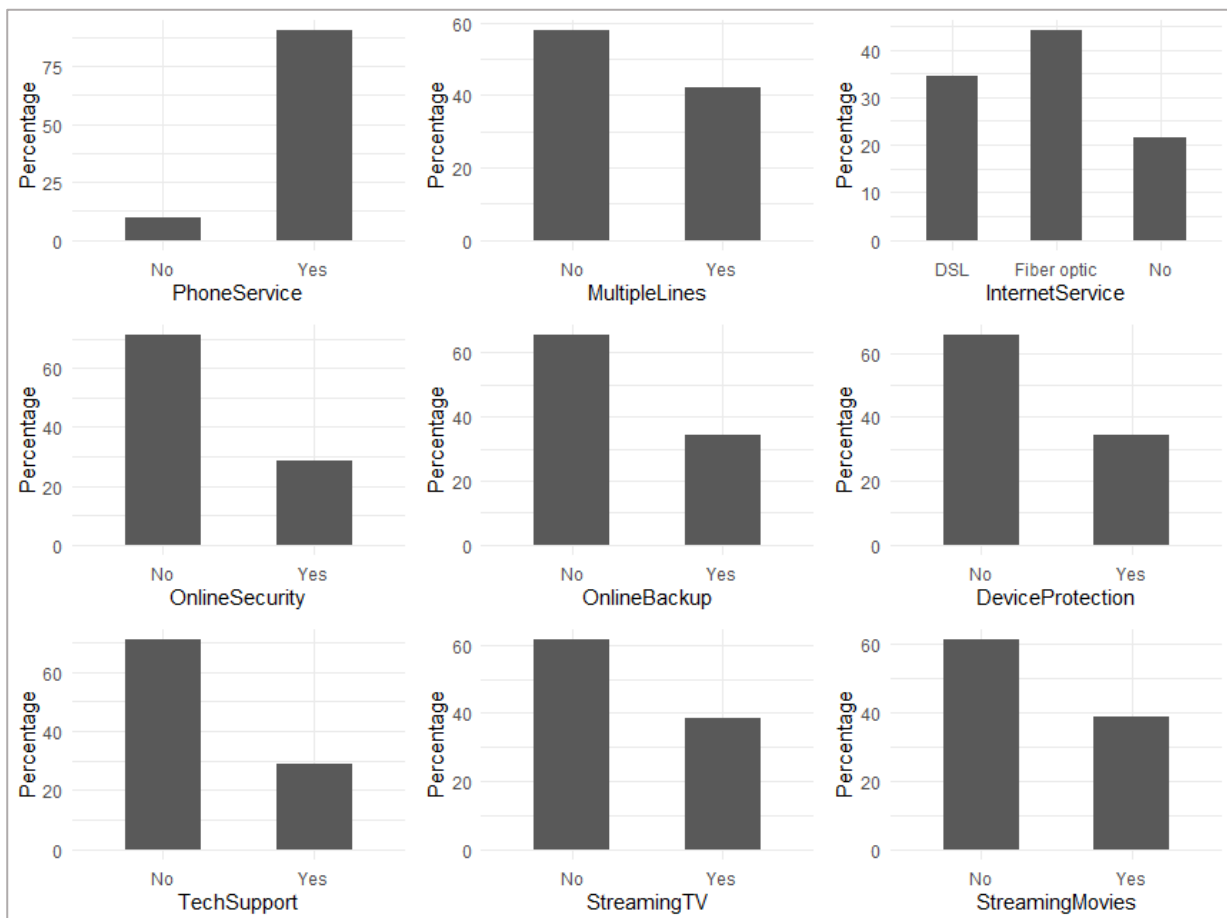


FIGURE 11: SUBSCRIBED SERVICES

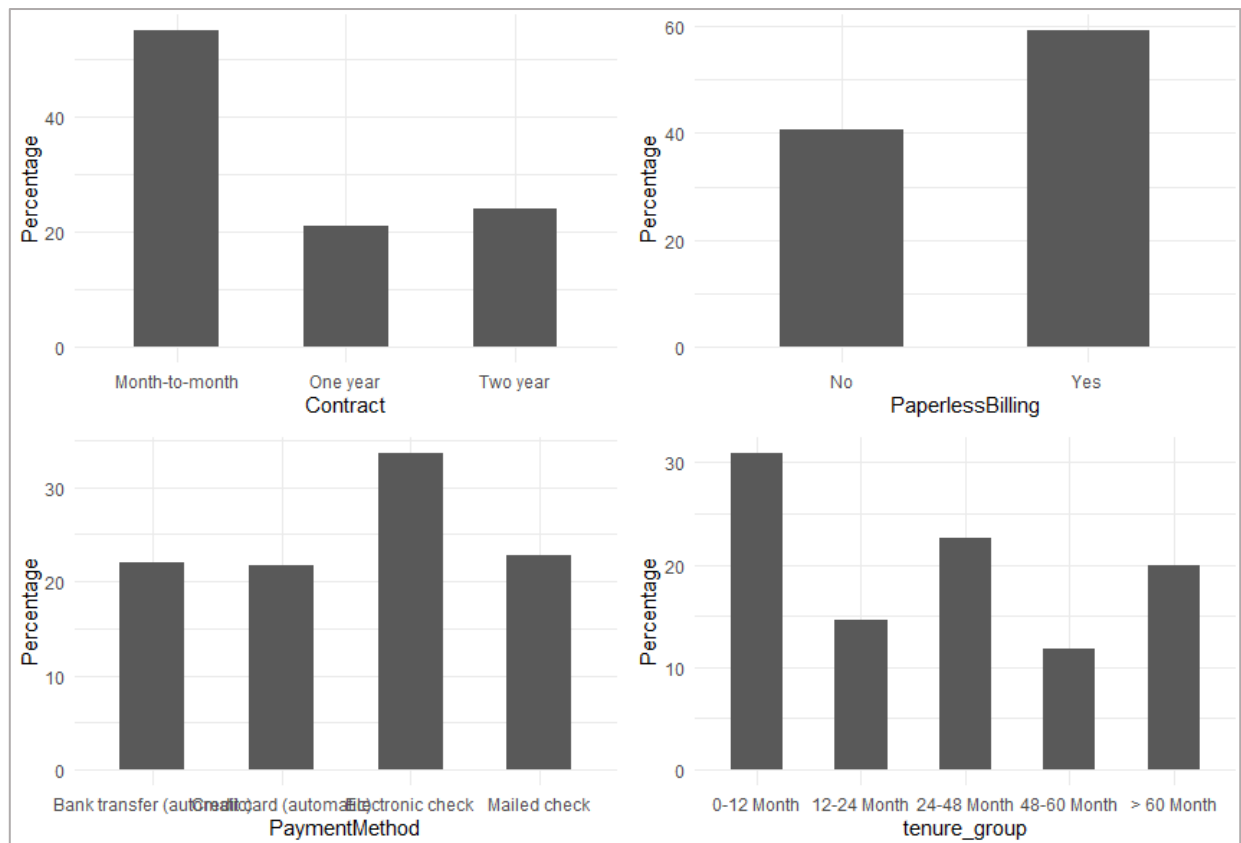


FIGURE 12: CUSTOMER ACCOUNT INFORMATION

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

4.3 Supervised learning steps

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

1. Split the data into Training and Testing: this is important as we want 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.
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

4.2.1 Binary Logistic Regression

A **Binary Logistic Regression** is a supervised learning technique that predicts binary response variables. It models the logarithm of the odds ratio (models log-odds of the probability of the target). Odds is a ratio of the probability of the event occurring divided by the probability of the event not occurring, or p divided by $1 - p$. This approach helps to find the decision boundary between the two classes while keeping the coefficients linearly related to the target variable.

Below is the formula of the *logistic regression equation* based on two input variables and a probability p .

$$1. \text{Probability to Churn: } P(Y = 1)$$

It is not easy to model this probability directly. If we use a linear model, we can end up with non-sensical predictions, like probabilities less than zero or greater than one. What we can model are the log odds that can be seen in the second equation. Removing the log by using the exponential function results in the odds.

$$2. \text{Log Odds: } \log \frac{P(Y = 1)}{P(Y = 0)} = \beta_0 + \sum_{p=1}^p \beta_p x_p$$

Odds is the probability to churn divided by the probability not to churn.

$$3. Odds: \quad \frac{P(Y = 1)}{P(Y = 0)} = e^Z, \text{ with } Z = \beta_0 + \sum_{p=1}^p \beta_p x_p$$

Final model of probability to churn: it gives the probability of the target variable being equal to one, or the probability of a customer churning.

$$4. Probability to Churn: \quad P(Y = 1) = \frac{e^Z}{1 + e^Z}$$

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.

```

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 LOGISTIC REGRESSION MODEL

According to the significance level viewed in the figure 13, the top three most-relevant features are: “Contract”, “tenure_group” and “PaperlessBilling”. 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	
gender	1	0.27	4922	5702.5	0.60149
SeniorCitizen	1	112.04	4921	5590.4	< 2.2e-16 ***
Partner	1	109.03	4920	5481.4	< 2.2e-16 ***
Dependents	1	32.34	4919	5449.1	1.296e-08 ***
PhoneService	1	0.03	4918	5449.0	0.85217
MultipleLines	1	4.09	4917	5444.9	0.04304 *
InternetService	2	446.59	4915	4998.4	< 2.2e-16 ***
OnlineSecurity	1	168.88	4914	4829.5	< 2.2e-16 ***
OnlineBackup	1	70.02	4913	4759.5	< 2.2e-16 ***
DeviceProtection	1	42.07	4912	4717.4	8.800e-11 ***
TechSupport	1	67.03	4911	4650.4	2.670e-16 ***
StreamingTV	1	1.39	4910	4649.0	0.23788
StreamingMovies	1	0.02	4909	4648.9	0.87729
Contract	2	305.11	4907	4343.8	< 2.2e-16 ***
PaperlessBilling	1	16.70	4906	4327.1	4.388e-05 ***
PaymentMethod	3	34.90	4903	4292.2	1.279e-07 ***
MonthlyCharges	1	0.04	4902	4292.2	0.84878
tenure_group	4	150.49	4898	4141.7	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

FIGURE 14: 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 “InternetService”, “Contract”, “OnlineSecurity”, “SeniorCitizen” and “tenure_group” significantly reduces the residual deviance.

The other variables such as “DeviceProtection”, “PaymentMethod” and “Dependents” seem to improve the model less even though they all have low p-values.

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

Step 4: Measure model performance on testing data.

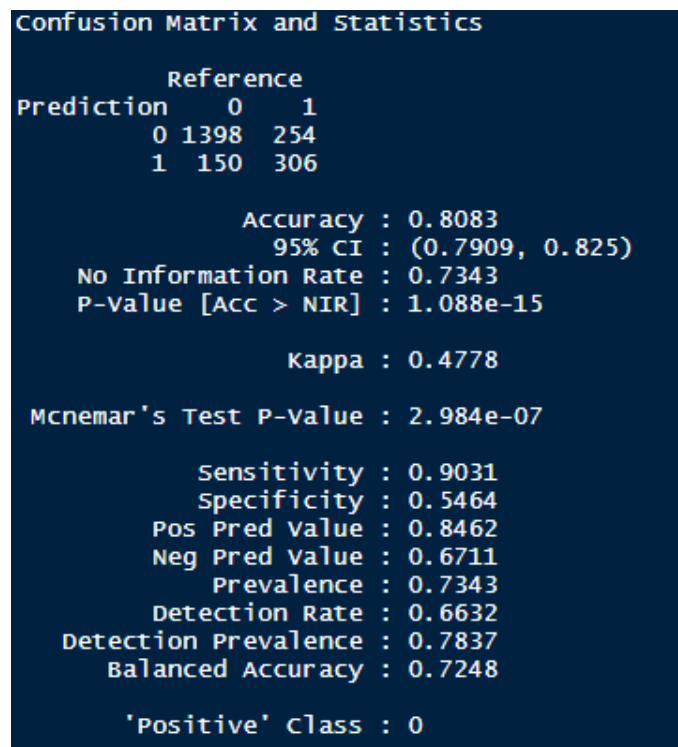


FIGURE 15: CONFUSION MATRIX AND ACCURACY FOR THE LOGISTIC REGRESSION MODEL

The **Accuracy** is the percentage of correctly predicted labels (both Churn and non-Churn). This includes prediction accuracy for both classes combined - both how many churned and non-churned customers the model correctly labeled. This gives the main performance of the model in respect of the class. The accuracy of the Logistic Regression model in the testing dataset is **80.83%** - this means the model has correctly classified around 81% of the customer churn events.

The **Sensitivity** is defined as the proportion of positive results out of the number of samples which were actually positive (**true positive rate**), whereas the **Specificity** is the ability of the model to define the proportion of negative results correctly classified as negative (**true negative rate**).

The **Confusion Matrix** is a matrix of the model's predicted classes versus the actual outcomes. It is called a confusion matrix because it reveals how "confused" the model is between the 2 classes, and highlights instances in which one class is confused for the other. The columns are the true classes, while the rows are the predicted classes. The cells of the confusion matrix are:

- **True positives:** cases where the model correctly predicted yes (churn);
- **False positives:** cases where the model incorrectly predicted yes;
- **False negatives:** cases where the model incorrectly predicted no;
- **True negatives:** cases where the model correctly predicted no (non-churn).

The main diagonal is the cases where the model is correct (**true positives** and **true negatives**) and the second diagonal is the cases where the model is incorrect (**false negatives** and **false positives**).

	OR	2.5 %	97.5 %
(Intercept)	0.6783881	0.10365996	4.4463565
genderMale	0.9661223	0.83069398	1.1235909
SeniorCitizenYes	1.2718332	1.04641906	1.5453958
PartnerYes	1.0119056	0.84464997	1.2125844
DependentsYes	0.9065864	0.73408828	1.1181375
PhoneServiceYes	0.5297402	0.11678807	2.4027637
MultipleLinesYes	1.3006288	0.86481794	1.9570497
InternetServiceFiber optic	2.0618633	0.32156035	13.2694922
InternetServiceNo	0.5387513	0.08224727	3.5259942
OnlineSecurityYes	0.6796676	0.44913030	1.0273476
OnlineBackupYes	0.8758421	0.58201904	1.3178060
DeviceProtectionYes	0.9856909	0.65616277	1.4806341
TechSupportYes	0.7501703	0.49138253	1.1441705
StreamingTVYes	1.3064472	0.61256611	2.7897208
StreamingMoviesYes	1.1596745	0.54282997	2.4798632
ContractOne year	0.4916650	0.38351300	0.6272738
ContractTwo year	0.1884496	0.12191383	0.2835543
PaperlessBillingYes	1.4470708	1.21643270	1.7227685
PaymentMethodCredit card (automatic)	0.9179698	0.70369635	1.1972163
PaymentMethodElectronic check	1.3955844	1.11726881	1.7465274
PaymentMethodMailed check	0.9848192	0.75375838	1.2881289
MonthlyCharges	1.0050903	0.93350602	1.0820832
tenure_group12-24 Month	0.4314007	0.34347603	0.5404239
tenure_group24-48 Month	0.3124252	0.24700011	0.3940109
tenure_group48-60 Month	0.2118881	0.15046609	0.2962211
tenure_group> 60 Month	0.1450637	0.09641040	0.2162935

FIGURE 16: ODDS RATIO

Another performance measurement in logistic regression is the Odds Ratio. The coefficients were extracted with help of the *coef()* function and the *exp()* function to remove the logarithm.

The effect on the odds can be interpreted as follows: looking at the variable “PaperlessBilling”. Since to the $0.3695 = 1.447$, it can be stated that a customer with the “PaperlessBilling = True” increases the odds of churning by the factor of 1.444, so 44.7%, compared to somebody who has the “PaperlessBilling = False”.

4.2.2 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.[2] 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.

(https://en.wikipedia.org/wiki/Decision_tree_learning#cite_note-2)

(<https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/>)

(<https://scikit-learn.org/stable/modules/tree.html>)

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"`.

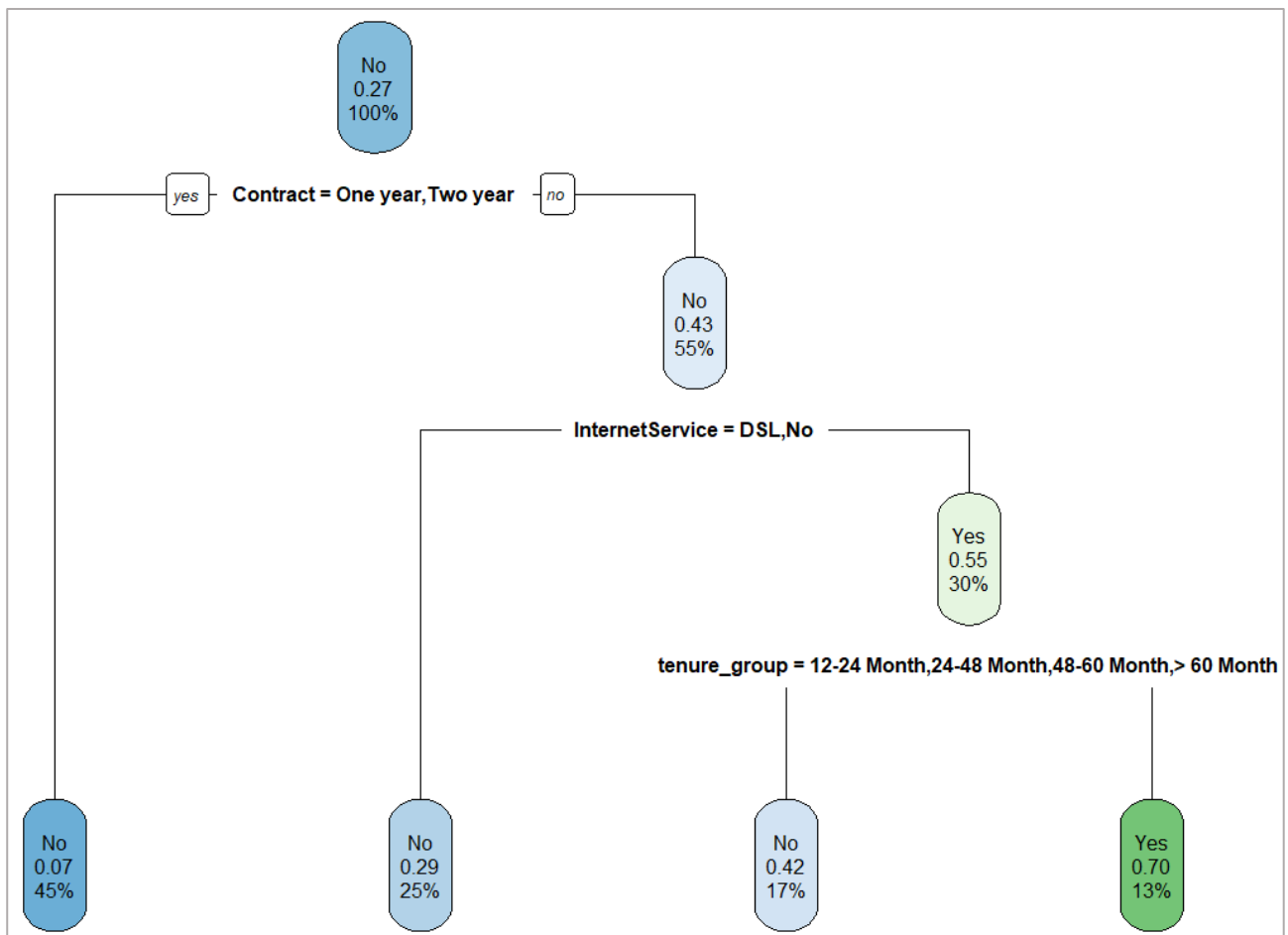


FIGURE 17: DECISION TREE MODEL

Out of the variables used, “Contract” is the most important variable 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.

This graph could be interpreted as follows: the customer with “Contract = Month-to-month”, “InternetService = Fiber optic” and “tenure_group = 0-12 Month” is more likely to churn.

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

Step 4: Measure model performance on testing data.

Confusion Matrix and Statistics		
Prediction	Reference	
	No	Yes
	No 1469	377
Yes	79	183
Accuracy : 0.7837		
95% CI : (0.7655, 0.8011)		
No Information Rate : 0.7343		
P-Value [Acc > NIR] : 9.376e-08		
Kappa : 0.3322		
McNemar's Test P-Value : < 2.2e-16		
Sensitivity : 0.9490		
Specificity : 0.3268		
Pos Pred Value : 0.7958		
Neg Pred Value : 0.6985		
Prevalence : 0.7343		
Detection Rate : 0.6969		
Detection Prevalence : 0.8757		
Balanced Accuracy : 0.6379		
'Positive' Class : No		

FIGURE 18: CONFUSION MATRIX AND ACCURACY FOR DECISION TREE MODEL

The accuracy of the Decision Tree has worsened compared with the accuracy of the Logistic Regression, from 0.8083 to 0.7837. This could be occurred because the decision tree models tend to overfitting.

Overfitting is a modelling error that occurs when a function is too closely or even exactly fit to a particular set of data and may therefore fail to fit additional data or predict future observations reliably. Trees that are grown very deep tend to learn highly irregular patterns: they have low bias, but very high variance.

<https://www.lexico.com/definition/overfitting>

<https://www.investopedia.com/terms/o/overfitting.asp>

4.2.3 Random Forest

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[1][2]

Random decision forests correct for decision trees' habit of overfitting to their training set.[3]:587–588

https://en.wikipedia.org/wiki/Random_forest

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 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: 21.04%
Confusion matrix:
      No Yes class.error
No  3236 379  0.1048409
Yes   657 652  0.5019099
```

FIGURE 19: RANDOM FOREST MODEL

The error rate is relatively low when predicting “No”, and the error rate is much higher when predicting “Yes”.

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

Step 4: Measure model performance on testing data.

```
Confusion Matrix and Statistics

      Reference
Prediction  No  Yes
      No 1388  281
      Yes  160  279

      Accuracy : 0.7908
      95% CI : (0.7728, 0.808)
      No Information Rate : 0.7343
      P-value [Acc > NIR] : 1.062e-09

      Kappa : 0.4241

      Mcnemar's Test P-value : 1.102e-08

      Sensitivity : 0.8966
      Specificity : 0.4982
      Pos Pred Value : 0.8316
      Neg Pred Value : 0.6355
      Prevalence : 0.7343
      Detection Rate : 0.6584
      Detection Prevalence : 0.7917
      Balanced Accuracy : 0.6974

      'Positive' class : No
```

FIGURE 20: CONFUSION MATRIX AND ACCURACY FOR THE RANDOM FOREST MODEL

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

4.2.3.1 Tuning the Random Forest model

The plot below (figure 21) is used to help to determine the number of trees. 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.

https://en.wikipedia.org/wiki/Out-of-bag_error

https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html

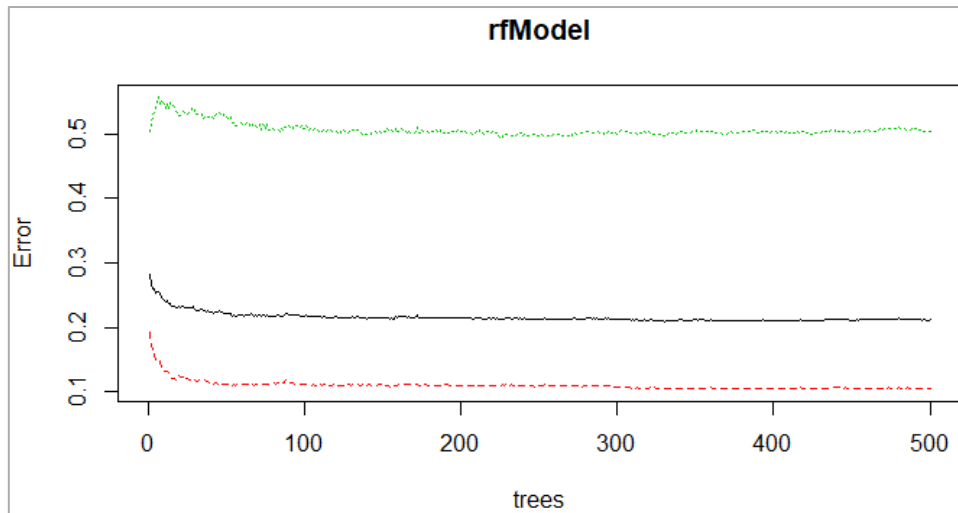


FIGURE 21: RANDOM FOREST ERROR RATE

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 22 is used to define the number of *mtry* to choose.

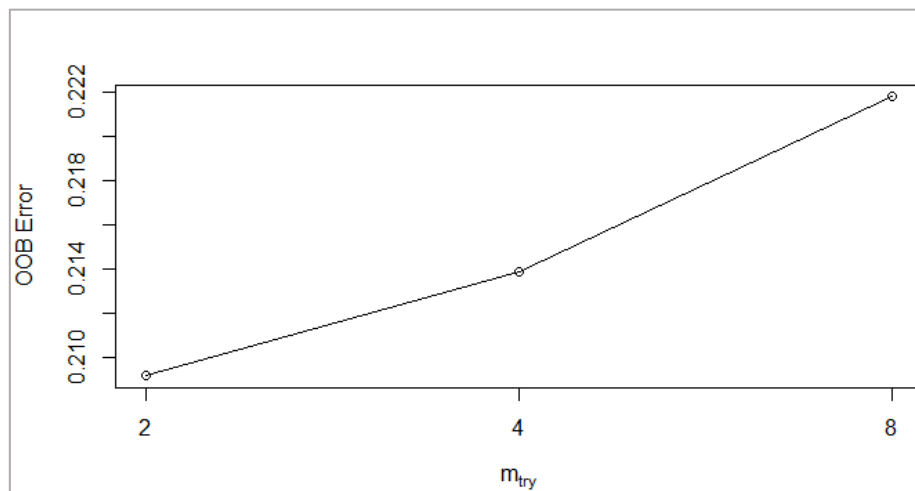


FIGURE 22: OOB ERROR BY THE NUMBER OF MTRY

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 will be repeated.

Step 2: Fitting the Random Forest model after Tuning

```

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: 20.57%
Confusion matrix:
      No Yes class.error
No  3306 309  0.08547718
Yes   704 605  0.53781513

```

FIGURE 23: TUNED RANDOM FOREST

The OOB error rate decreased to 20.57% from 21.04% on figure 19.

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  1421  303
      Yes  127  257

      Accuracy : 0.796
      95% CI : (0.7782, 0.813)
      No Information Rate : 0.7343
      P-value [Acc > NIR] : 2.663e-11

      Kappa : 0.4189

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

      Sensitivity : 0.9180
      Specificity : 0.4589
      Pos Pred value : 0.8242
      Neg Pred value : 0.6693
      Prevalence : 0.7343
      Detection Rate : 0.6741
      Detection Prevalence : 0.8178
      Balanced Accuracy : 0.6884

      'Positive' Class : No

```

FIGURE 24: CONFUSION MATRIX AND ACCURACY OF THE TUNED RANDOM FOREST

For the Tunned Random Forest model, both accuracy and sensitivity improved, however the specificity worsened compared to the Random Forest model.

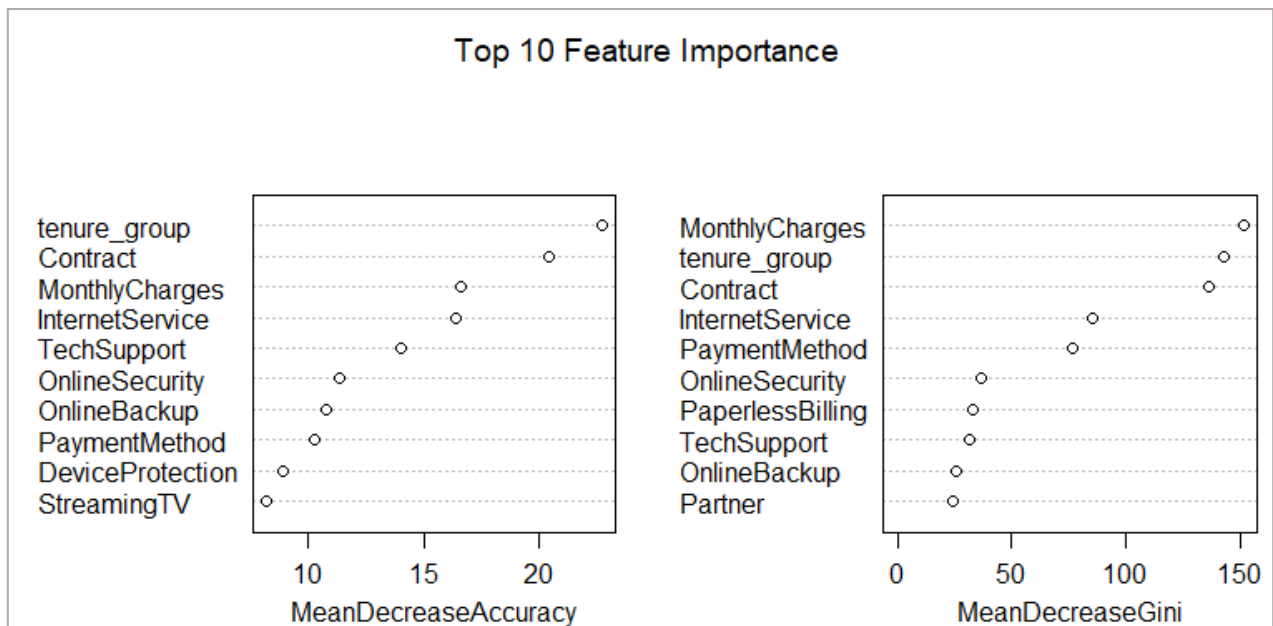


FIGURE 25: RANDOM FOREST FEATURE IMPORTANCE

4.2.4 Summary

```
[1] "Models Comparison"
```

	Accuracy	Sensitivity	Specificity
Logistic_Regression	0.8083	0.9031	0.5464
Decision_Tree	0.7837	0.9490	0.3268
Random_Forest	0.7908	0.8966	0.4982
Tuned_Random_Forest	0.7960	0.9180	0.4589

FIGURE 26: MODELS COMPARISON

From the above example, it is possible to see that **Logistic Regression, Decision Tree and Random Forest models** can be used for customer churn analysis for that particular dataset equally well.

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, that are not using PaperlessBilling, are less likely to churn.

Finally, based on the calculated metrics the Logistic Regression model should be choose, because it has the highest accuracy and specificity and a quite well sensitivity level.

5. Skills acquire in the internship

This internship provided me an invaluable experience. It allowed me to test a great part of the concepts that I have been introduced throughout the masters.

The main skills that I can identify, at this moment, could be:

- Communication: my ability to write and speak effectively in Italian improved a lot during this period, once my manager and colleagues speak majority in their mother tongue.
- Adaptability: also regarding to communication, I had to adapt to the Italian work-culture to succeed in this internship.
- Technical proficiency: during the stage I had the chance to develop my programming skills in SAS and R. In SAS because it was the main language used at the company where I was allocated, and in R because it was the language that I used to perform the analysis for this thesis.

In general, the internship gave me the opportunity to understand my own strengths (gained during my last job, where I worked for a coupled of years) and my weaknesses, these ones that should be impossible to understand without putting myself in this international experience.

6. References