UNIVERSITÀ degli STUDI di CATANIA
Master’s Degree in
Data Science for Management
A.Y. 2019/2020

Data Analysis and Statistical Learning
“Statistical Learning” module
Prof. S. Ingrassia

Ylenia Messina
1000008815

REPORT

Index

1. INTRODUCTION 1
   1.1. DATASET: Online Shoppers Purchase Intention 1

2. DESCRIPTIVE ANALYSIS 2
   2.1. UNIVARIATE ANALYSIS 2
   2.2. BIVARIATE ANALYSIS 5

3. PREDICTION ANALYSIS 7
   3.1. LOGISTIC REGRESSION 7
   3.2. SUPPORT VECTOR MACHINES 11
   3.3. NEURAL NETWORKS 13
   3.4. BEST MODEL & PREDICTION ON TEST SET 14

4. PREDICTION RESULTS ACCORDING TO VARIABLE “MONTH” 15

1. INTRODUCTION

The work in this report is aimed at building a predictive model in the context of a classification problem: specifically, it is about predicting the purchase intention of e-commerce web pages' visitors, based on their behavior during the browsing session.
For this purpose, once the predictor variables have been identified, three main statistical learning methods are considered:

1. Logistic Regression;
2. Support Vector Machines;
3. Neural Networks.

Each model is trained and evaluated. Then the performance of the defined models is compared on the basis of their misclassification rate and finally the best of them is tested to make predictions of the response variable Revenue.

1.1. DATASET: Online Shoppers Purchasing Intention.

The original dataset[1] consists of feature vectors belonging to 12330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

The dataset consists of 10 numerical and 8 categorical attributes.
The "Revenue" attribute can be used as the class label.

The values of "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by *Google Analytics* for each page in the e-commerce site.

The value of "Special Day" is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.

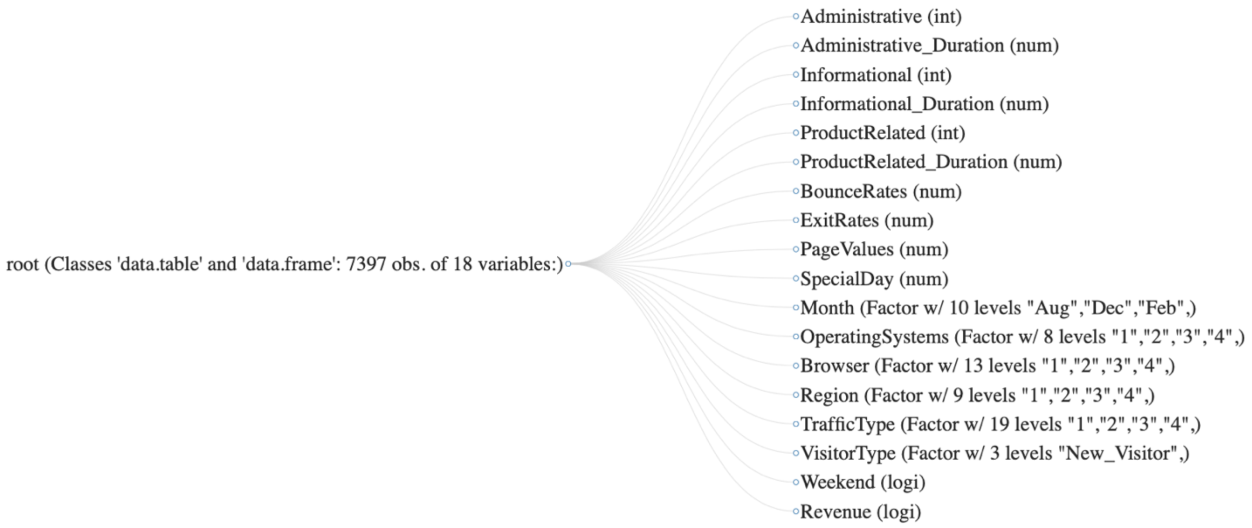
For the predictive analysis purposes, the original dataset is split into three separate sets used in different stages:

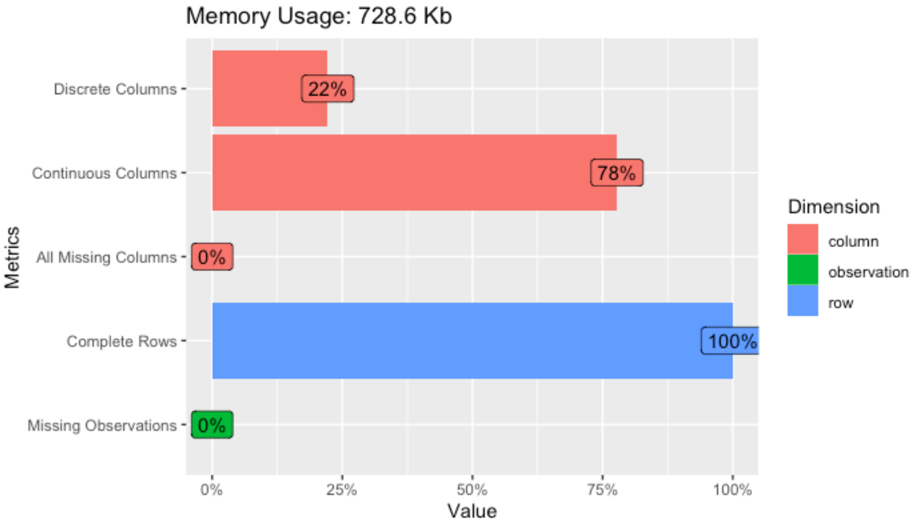
- *training set*, which contains about the 60% of units of the original dataset, used in the initial model fitting stage;
- *validation set*, which contains about the 20% of units of the original dataset, used for model selection;
- *test set,* which contains about the 20% of units of the original dataset, used ultimately to make predictions and for model assessment.

[1] https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

2. DESCRIPTIVE ANALYSIS

For the descriptive analysis, the *training* dataset will be considered.


root (Classes 'data.table' and 'data.frame': 7397 obs. of 18 variables:)
Administrative (int)
Administrative_Duration (num)
Informational (int)
Informational_Duration (num)
ProductRelated (int)
ProductRelated_Duration (num)
BounceRates (num)
ExitRates (num)
PageValues (num)
SpecialDay (num)
Month (Factor w/ 10 levels "Aug","Dec","Feb",)
OperatingSystems (Factor w/ 8 levels "1","2","3","4",)
Browser (Factor w/ 13 levels "1","2","3","4",)
Region (Factor w/ 9 levels "1","2","3","4",)
TrafficType (Factor w/ 19 levels "1","2","3","4",)
VisitorType (Factor w/ 3 levels "New_Visitor",)
Weekend (logi)
Revenue (logi)



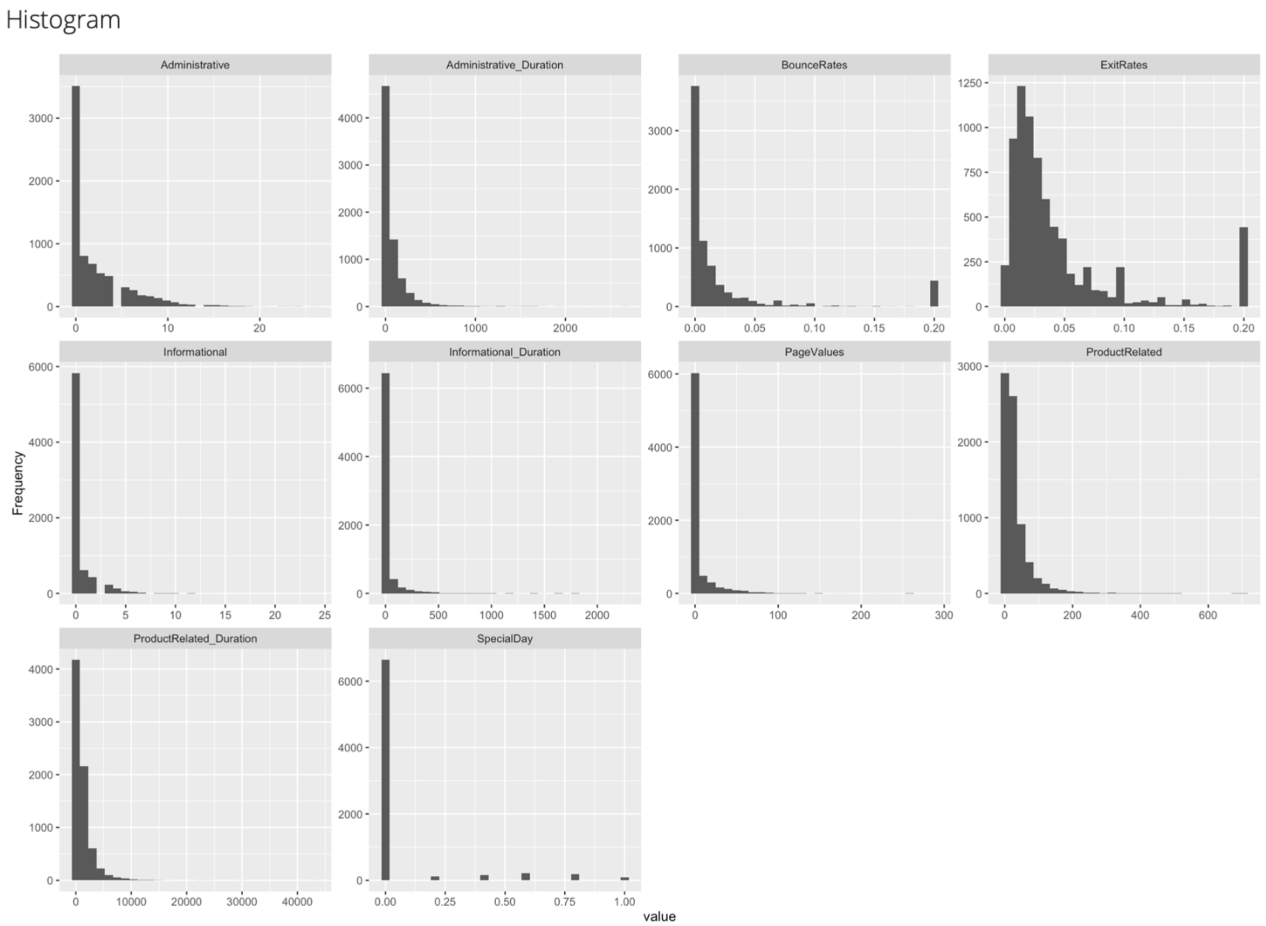
It consists of 7397 observations, each one representing a unique session on shopping web pages which ended with a purchase or not. Each session is described by 10 quantitative and 8 categorical attributes. Such variables and their distributions are better presented below in the context of the univariate analysis. None of them has unknown (*NULL*) values, so the dataset has no missing observations.

2.1. UNIVARIATE ANALYSIS

Quantitative Variables

`Administrative`, `Administrative_Duration`, `Informational`, `Informational_Duration`, `ProductRelated` and `ProductRelated_Duration` represent the number of different types of pages – respectively about account management, communication & address information of the shopping site, and about products – visited by the visitor in that session and the total amount of time (in seconds) spent in each of these page categories.

The “Bounce Rate" feature for a website refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.
The "Exit Rate" feature is the percentage of visitors to a page on the website from which they exit the website to a different website.
The "Page Value" feature represents the average value for a website that a user visited before completing an e-commerce transaction.
The values of `BounceRate`, `ExitRate` and `PageValues` for each session are averages computed on the basis of the pages visited by the visitor.
The "`SpecialDay`" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.



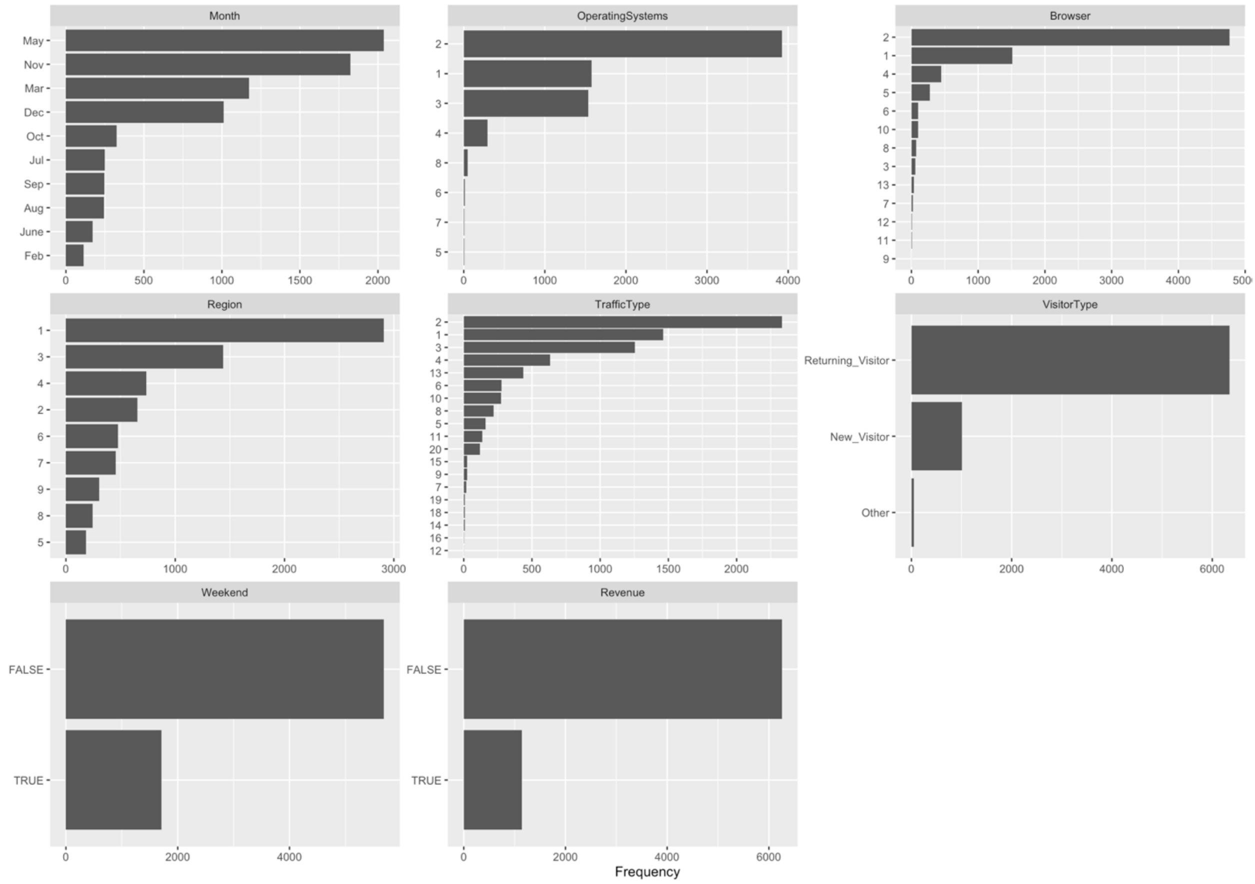
With reference to the *numerical* variables, we can see that they all have approximately the same distribution, that is an asymmetrical positively skewed distribution. Most observations are in fact clustered around values close to 0 (the left tail of the distribution) while the right tail of the distribution is longer.
The majority of visitors to the e-commerce websites, during the browsing session, tend not to explore any page with administrative, informative or product-related content (although a greater percentage tends to visit at least one page related to a product) or very few, but they remain there for a short time. Also, on average the percentage of visitors to a page from which they exit the website, as well as that of visitors who entered the website from that page and then left (“bounce”) without triggering any other request, remain low.

Categorical Variables

- `OperatingSystems:` operating system of the visitor; it can assume 8 values.
- `Browser:` browser of the visitor; it can assume 13 values.
- `Region:` geographic region from which the session has been started by the visitor; it can assume 9 values.
- `TrafficType:` traffic source by which the visitor has arrived at the website (e.g. banner, sms, direct etc.).
- `VisitorType:` visitor type as one of the 3 possible categories: "New Visitor", "Returning Visitor" and "Other".
- `Weekend:` Boolean value indicating whether the date of the visit is weekend.
- `Month:` month value of the visit date; this factor variable has 12 levels but in the training data assumes only 10 of them, which are: Feb, Mar, May, June, July, Aug, Sep, Oct, Nov, Dec.
- `Revenue:` class label indicating whether the visit has been finalized with a transaction.

The categories of the variables OperatingSystems, Browser, Region and TrafficType are coded with numbers, and the original labels are unfortunately not provided.

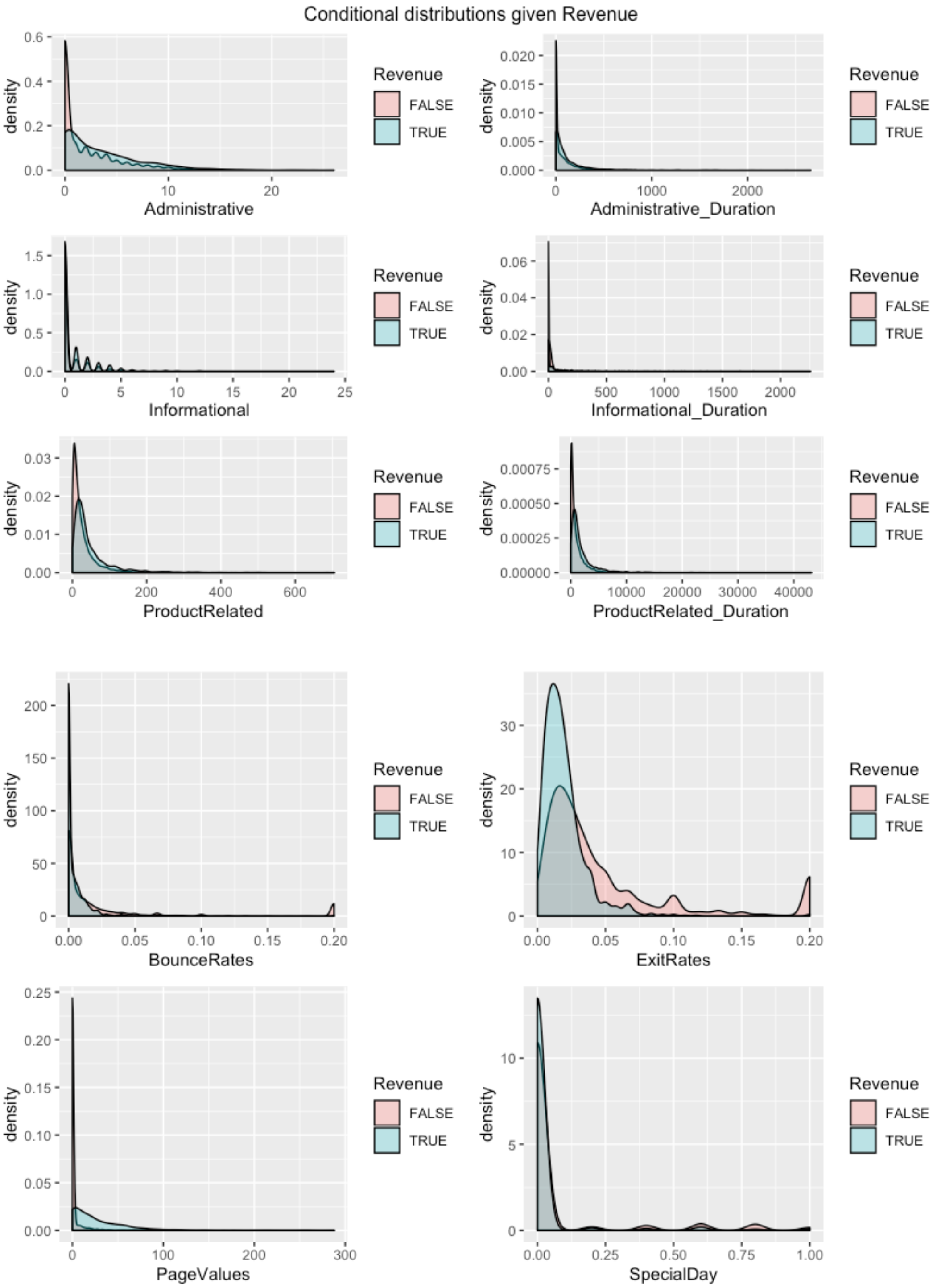
Bar Chart (by frequency)

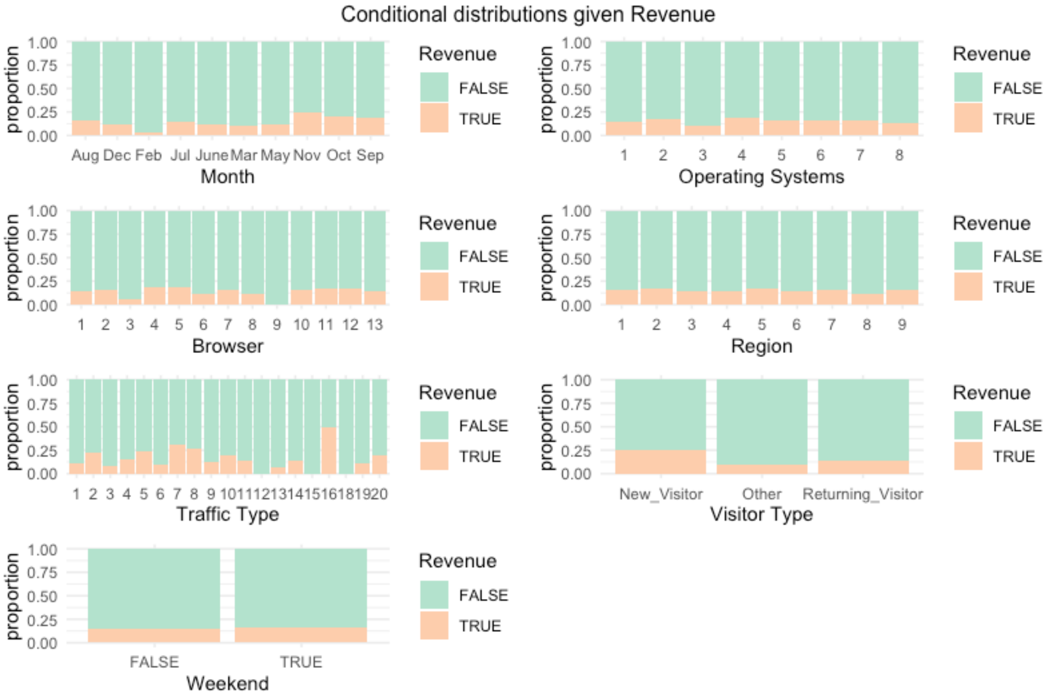


Most of the visits to e-commerce websites were concentrated in November and especially in May, mostly during weekdays, by "returning visitors" – that are users who had previously visited the site at least once. They come from the geographic region coded with number 1, mostly using the operating system 2 and browser 2, and coming from the traffic source 2. Furthermore, most of these visits did not end with a transaction.

2.2. BIVARIATE ANALYSIS

Set Revenue as the target variable, the purpose of the bivariate analysis is to see how the other variables relate to it, and to detect which of them seem to be of greater importance in the perspective of building a model to predict an online user's intention to purchase.





From training data, it results that the sessions on e-commerce sites that ended with a transaction were not particularly influenced by the number of pages with administrative, informative or product-related content, previously visited in the same session, nor by the time spent in consulting the latter. This can be seen from the tendential overlap of the areas of the conditional distributions of these variables given Revenue. The same seems to hold also with regard to the bounce rate, the page values and the proximity of the session date to a special day. On the other hand, the curve of the distribution of the ExitRates of the sessions that ended with a transaction seems not to perfectly match that of the sessions that vice versa did not involve any transaction. It has a wider curve and a longer tail: the sessions in which users visited pages of the e-commerce website with an average exit rate (from the website) of up to 20%, are the ones that mostly resulted in a transaction.

Among the categorical variables, some features such as the geographic region from which the session has been started as well as the day of the week (weekend or not), the browser and the operating system used by the visitor do not seem to make much difference in determining the purchase intention. In fact, the stacked columns of these variables, for which Revenue = TRUE, are mostly the same height.

Instead, the stacked columns of the traffic source by which the visitor has arrived at the website are more diversified in height: in particular the main traffic type of the session finalized with a transaction, is represented by the category that has been associated with the integer number 16.

Finally, most of the purchases on shopping sites were made by users visiting that e-commerce webpage for the first time. Moreover, most of the sessions that involved a transaction took place mostly in November, October and September.

3. PREDICTIVE ANALYSIS
3.1. LOGISTIC REGRESSION

The first statistical learning method used in order to predict the response variable “Revenue”, given the set of both *numerical* and *categorical* variables in the online shopping intention dataset, is the Logistic Regression model – one of the main classifiers to predict a qualitative response.

First step in the process involves determining which predictors among the 17 variables are useful and therefore have to be included in the model during the training phase.

This is done by performing *stepwise logistic regression* which consists of automatically selecting a reduced number of predictor variables for building the best performing logistic regression model by AIC (Akaike Information Criterion).

```
> logR <- step(glm(Revenue~., data = train_data, family = binomial), direction = "both")
Start:  AIC=4374.99
Revenue ~ Administrative + Administrative_Duration + Informational + 
    Informational_Duration + ProductRelated + ProductRelated_Duration + 
    BounceRates + ExitRates + PageValues + SpecialDay + Month + 
    OperatingSystems + Browser + Region + TrafficType + VisitorType + 
    Weekend

                          Df Deviance    AIC
- Browser                 11   4251.3 4363.3
- Region                   8   4251.4 4369.4
- OperatingSystems         6   4249.6 4371.6
- TrafficType             18   4274.4 4372.4
- Weekend                  1   4241.0 4373.0
- SpecialDay               1   4241.0 4373.0
- Informational_Duration   1   4241.0 4373.0
- BounceRates              1   4241.1 4373.1
- ProductRelated           1   4241.1 4373.1
- Administrative_Duration  1   4241.3 4373.3
- Informational            1   4241.7 4373.7
- Administrative           1   4241.9 4373.9
<none>                         4241.0 4375.0
- VisitorType              2   4245.3 4375.3
- ProductRelated_Duration  1   4248.2 4380.2
- ExitRates                1   4275.8 4407.8
- Month                    9   4366.4 4482.4
- PageValues               1   5463.4 5595.4

[...]

Step:  AIC=4337.15
Revenue ~ ProductRelated_Duration + ExitRates + PageValues + 
    Month + VisitorType

                          Df Deviance    AIC
<none>                         4307.1 4337.1
+ Administrative           1   4305.4 4337.4
+ Informational            1   4305.5 4337.5
+ Informational_Duration   1   4306.2 4338.2
+ BounceRates              1   4307.0 4339.0
+ SpecialDay               1   4307.1 4339.1
+ Administrative_Duration  1   4307.1 4339.1
+ Weekend                  1   4307.1 4339.1
+ ProductRelated           1   4307.1 4339.1
+ TrafficType             18   4273.3 4339.3
+ OperatingSystems         7   4297.1 4341.1
- VisitorType              2   4317.5 4343.5
+ Region                   8   4297.9 4343.9
+ Browser                 12   4297.2 4351.2
```

```
- ProductRelated_Duration  1   4344.4 4372.4
- ExitRates                1   4428.5 4456.5
- Month                    9   4454.5 4466.5
- PageValues               1   5586.2 5614.2
```

It results that the best performing logistic regression model is the one that contains only 5 predictors. In particular: ProductionRelated_Duration, ExitRates, PageValues, Month & VisitorType.

After variable selection, coefficients estimates are obtained in order to build the model by which to make predictions.

```
> summary(logR)

Call:
glm(formula = Revenue ~ ProductRelated_Duration + ExitRates + 
    PageValues + Month + VisitorType, family = binomial, data = train_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.1405  -0.4670  -0.3381  -0.1597   3.3792  

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -1.726e+00  2.317e-01  -7.449 9.42e-14 ***
ProductRelated_Duration      1.085e-04  1.765e-05   6.147 7.89e-10 ***
ExitRates                   -1.850e+01  2.138e+00  -8.651  < 2e-16 ***
PageValues                   8.055e-02  3.036e-03  26.535  < 2e-16 ***
MonthDec                    -4.553e-01  2.458e-01  -1.852  0.06400 .  
MonthFeb                    -1.055e+00  6.492e-01  -1.626  0.10402    
MonthJul                     1.885e-01  2.955e-01   0.638  0.52345    
MonthJune                    1.005e-01  3.467e-01   0.290  0.77188    
MonthMar                    -3.310e-01  2.406e-01  -1.376  0.16881    
MonthMay                    -4.590e-01  2.297e-01  -1.999  0.04566 *  
MonthNov                     6.639e-01  2.214e-01   2.999  0.00271 ** 
MonthOct                     4.823e-02  2.715e-01   0.178  0.85902    
MonthSep                     6.677e-02  2.903e-01   0.230  0.81808    
VisitorTypeOther            -1.329e+00  9.705e-01  -1.369  0.17096    
VisitorTypeReturning_Visitor -3.284e-01  1.094e-01  -3.001  0.00269 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6371.8  on 7396  degrees of freedom
Residual deviance: 4307.1  on 7382  degrees of freedom
AIC: 4337.1

Number of Fisher Scoring iterations: 7
```

Looking at the coefficients estimates, we learn that:

- A one-unit increase in ProductRelated_Duration is associated with an increase in the *log-odds* of Revenue by 0.0001085 units. That is, one more second spent by the visitor on product related pages makes the odds of the session finalizing with a transaction increase by $e^{0.0001085}$.
- A one-unit increase in ExitRates is associated with a decrease in the *log-odds* of Revenue by 18.5 units. That is, a one-unit increase in the average exit rate value of the pages visited by the visitor makes the odds of the session finalizing with a transaction decrease by $e^{18.5}$.

- A one-unit increase in PageValues is associated with an increase in the *log-odds* of Revenue by 0.08055 units. That is, a one-unit increase in the average page value of the pages visited by the visitor makes the odds of the session finalizing with a transaction increase by $e^{0.08055}$.

Also, *p-values* of the coefficients related to the numerical variables are very small. Therefore, it is possible to state that there is indeed an association between the predictors ProductRelated_Duration, ExitRates, PageValues and the response Revenue.

Once coefficients have been estimated, it is possible to compute the probabilities of Revenue given certain values of the predictors and make predictions as to whether a visitor's online session on a shopping page will turn into a purchase intention or not.
On the basis of the latter it is possible to construct the confusion matrix, calculate some measures such as the *true positive rate* and *false positive rate*, and make some considerations.

```
> conf_matrix_train
               True Class
Predicted Class FALSE TRUE  Sum
          FALSE  6103  716 6819
          TRUE    150  428  578
          Sum    6253 1144 7397
```

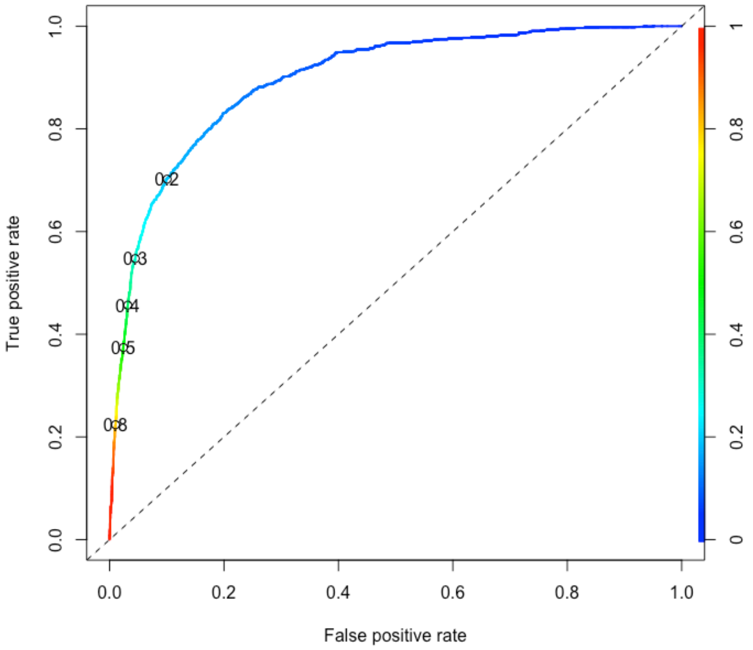
TRUE POSITIVE RATE	FALSE POSITIVE RATE	FALSE NEGATIVE ERROR
37.41 %	2.4 %	62.59 %

On the training set, it results an overall *misclassification rate* based on Bayes classification (threshold of 50% for the posterior probability of Revenue) of 11.71%.
However, the analysis of the confusion matrix shows that:

- A total of 578 visitors of online shopping pages would end up purchasing something and, of these people, 428 actually purchased and 150 did not;
- In particular only 150 out of 6253 of the individuals who did not purchased were incorrectly labeled;
- Of the 1144 who actually purchased, 716 were missed.
- From the perspective of a company that sells its products online and is trying to quantifying its sales potential, a Type I error rate of 2.4% versus a Type II error rate of 62.59% among individuals who purchased may well be acceptable, since a company would be more concerned with sales that did not actually occur rather than those that occurred but have been neglected.

Ultimately, the threshold of 50% for the posterior probability of Revenue seems to be a good decision boundary.
It is possible to evaluate the overall performance of the model, summarized over all possible thresholds, through the area under the ROCR (*Receiving Operating Characteristics*) curve.

The AUC for the logistic regression model that was defined is 0.896, which is quite close to the ideal case “AUC = 1”, suggesting it is a good classifier.



At this point, the model is evaluated on the validation set.

```
> conf_matrix_val
               True Class
Predicted Class FALSE TRUE  Sum
          FALSE  2047  239 2286
          TRUE     38  144  182
          Sum    2085  383 2468

> misc_logR_val = round((conf_matrix_val[1,2] + conf_matrix_val[2,1])/conf_matrix_val[3,3] * 100, 2)
> misc_logR_val
[1] 11.22
```

On the validation set, it results an overall *misclassification rate* of 11.22%, which is slightly lower than the *misclassification rate* on the training set.

3.2. SUPPORT VECTOR MACHINES

Preliminarily, it is appropriate to scale the quantitative variables in the dataset – both in the training and validation set. In this case, it is a matter of scaling only ProductRelated_Duration, ExitRates and PageValues, since among the predictors of interest – selected on the basis of the AIC during the logistic regression fitting – only those three are *numeric.* As it emerged in the univariate analysis, these variables do not have normal distributions, therefore the scaling process takes the form of their normalization using a MinMaxScaler function.

Then, three statistical learning models are fitted: 1) Support Vector Classifier (*linear kernel*); 2) Support Vector Machine using *radial kernel*; 3) Support Vector Machine using *polynomial kernel*.
As for logistic regression, the predictors of interest are ProductRelated_Duration, ExitRates, PageValues, Month & VisitorType, while the response variable is Revenue.

For each of these three models, the `tune()` function is computed to perform *ten-fold cross validation,* in order to compare SVMs using a range of different values of the *cost* parameter (`0.001, 0.01, 0.1, 1, 5, 10, 100`), but also of *gamma* (`0.5, 1, 2, 3, 4`) in the SVM with radial kernel and *degree* (`2, 3, 4, 5, 6, 7, 8, 9, 10`) in the SVM with polynomial kernel.
The best model, which is the one resulting in the lowest cross-validation error rate, is used to predict the class label on the observations of the validation set. On the basis of these predictions the *misclassification error rate* is computed. Ultimately, the misclassification error rates will be used as a term of comparison to select the best SVM model among the three.

1) Support Vector Classifier (linear kernel)

According to ten-fold cross validation, the best Support Vector Classifier is the one with a *cost* parameter of 0.1. It has 1939 support vectors: in particular 905 of them belong to the FALSE class, while the remaining 1034 belong to the TRUE class.

```
> svc.confM    # Confusion Matrix
               True Class
Predicted Class FALSE TRUE  Sum
          FALSE  2050  262 2312
          TRUE     35  121  156
          Sum    2085  383 2468
```

2) Support Vector Machine (radial kernel)

According to ten-fold cross validation, the best Support Vector Machine with a radial kernel is the one with a *cost* of 1 and a *gamma* equal to 0.5. It has less support vectors than the linear kernel SVC. It has in fact 1905 support vectors: 863 of them belong to the FALSE class, while the remaining 1042 belong to the TRUE class.

```
> svmr.confM   # Confusion Matrix
               True Class
Predicted Class FALSE TRUE  Sum
          FALSE  2031  213 2244
          TRUE     54  170  224
          Sum    2085  383 2468
```

2) Support Vector Machine (polynomial kernel)

According to ten-fold cross validation, the best Support Vector Machine with a polynomial kernel is the one with a *cost* of 5 and a *degree* equal to 2. It has less support vectors than both the linear kernel SVC and the radial kernel SVM. It has in fact 1792 support vectors: 831 of them belong to the FALSE class, while the remaining 961 belong to the TRUE class.

```
> svmp.confM   # Confusion Matrix
               True Class
Predicted Class FALSE TRUE  Sum
          FALSE  2022  209 2231
          TRUE     63  174  237
          Sum    2085  383 2468
```

SVM models in comparison:

SVM models	***Misclassification error on validation set (%)***
Support Vector Classifier	12.03 %
Radial SVM	10.82 %
Polynomial SVM	11.02 %

Comparing the three previous SVM models, it is clear that the one that performs better on validation data in terms of accuracy is the *SVM with radial kernel*. Its misclassification error rate on the validation set – that is 10.29% – is in fact the lowest.

3.3. NEURAL NETWORKS

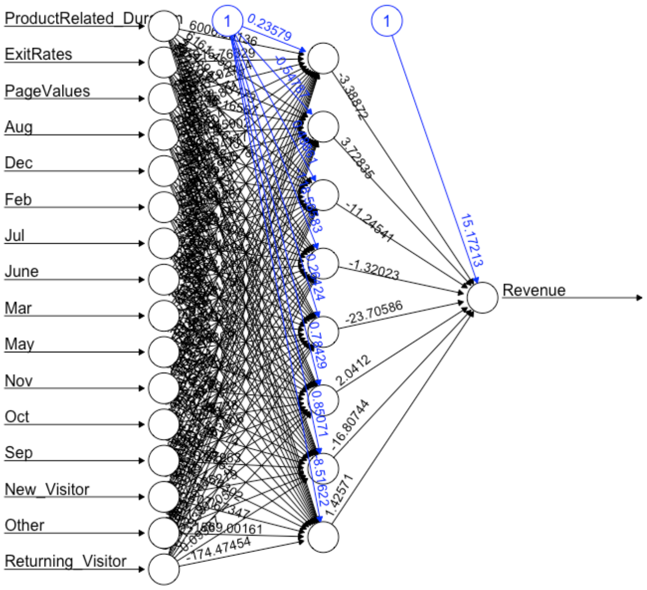
Before creating a machine learning model via `neuralnet()` R function in order to model the target variable, some pre-processing of data is necessary. In particular, Revenue needs to be coded as a binary variable taking in the values 0 (“FALSE”) or 1 (“TRUE”), while each categorical predictor with *k* > 2 categories needs to be transformed into a matrix with *k* dummy variables which are equal to 0 except for the column corresponding to the class. These operations are executed both on training and validation set, after they have been appropriately scaled.

Afterwards, it is possible to create the machine learning model. In particular, here the training phase involved the fitting of 8 neural networks with a single hidden layer, a number of hidden neurons ranging from 1 to 8, and with three repetitions for the neural network’s training (the one with lower error is chosen). For each of them the *misclassification error rate* was computed, on both training and validation set.

```
> nn.overview      # Fitted Neural Networks | Misclassification Error (%)
  NN models Hidden Neurons Miscl. error on training set (%) Miscl. error on validation set (%)
1       nn1              1                            10.34                              11.22
2       nn2              2                            10.04                              10.86
3       nn3              3                             9.76                              10.94
4       nn4              4                             9.75                              10.29
5       nn5              5                             9.54                              10.25
6       nn6              6                             9.15                              10.70
7       nn7              7                             9.21                              10.66
8       nn8              8                             8.52                              10.05
```

Comparing the various models summarized in the data-frame above, the best neural network in terms of accuracy in predicting the target variable Revenue turns out to be the one with one hidden layer and 8 hidden neurons. It scores in fact the lowest misclassification rate on validation set – 10.05%.

Here it is below a plot of the neural network that performed better. It uses 16 input nodes – one for each predictor: ProductRelated_Duration, ExitRates, PageValues and the categorical variables Month & VisitorType that have been split into as many dummy variables as their respective categories –, one hidden layer with 8 hidden neurons and 1 output node which is the target variable Revenue.



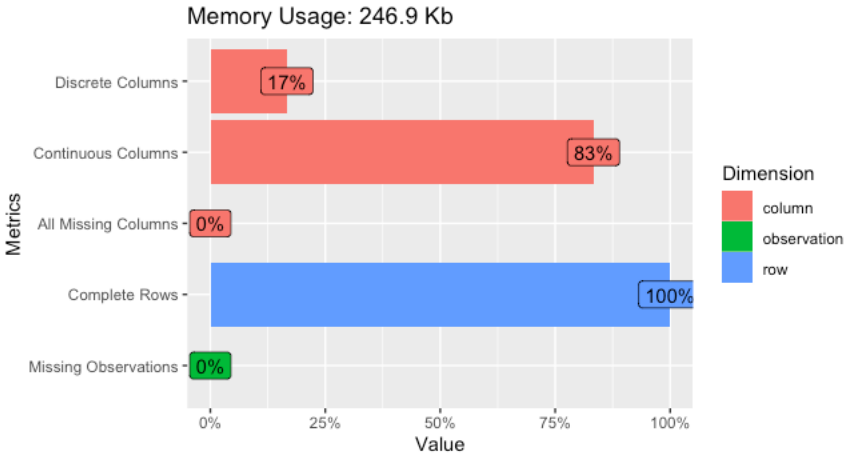
3.4. BEST MODEL & PREDICTION ON TEST SET

A further and definitive comparison between the three statistical learning methods used for the classification problem – that is logistic regression, support vector machine and neural networks – is made in order to determine the best predictive model and finally test it to predict for every visitor (identified by an *id_number*) if he purchased something or not at the end of his online session.

```
> models    # Recap on statistical learning models results
  Stat. Learning Models Miscl. error on validation set (%)
1   Logistic Regression                              11.22
2            Radial SVM                              10.82
3        Neural Network                               8.52
```

Among the logistic regression model, the support vector machine with radial kernel parameterized with *cost* = 1 and *gamma* = 5, and the neural network with one hidden layer and 8 hidden neurons, the better model seems to be the latter one. It performs better in accuracy of about 2% with respect to the other two models.

The test dataset, which consists of 2465 sessions, is finally used to make predictions of the target variable Revenue. It has no missing values, therefore all 2465 are considered to test the model.



```
> table(predictedValues)
predictedValues
FALSE  TRUE
 1950   515

> round(table(predictedValues)/length(predictedValues)*100, 2)
predictedValues
FALSE  TRUE
79.11 20.89
```

The machine learning model defined predicts that among the 2465 online browsing sessions (each of which associated with a unique different user) on shopping web pages, only 20.89% of them will end with an actual purchase.

From these results it is possible to make some assumptions about online consumer behavior: for example, that the majority of user visits on shopping web pages are not motivated a priori by an intention to purchase, or at least not in the first session.

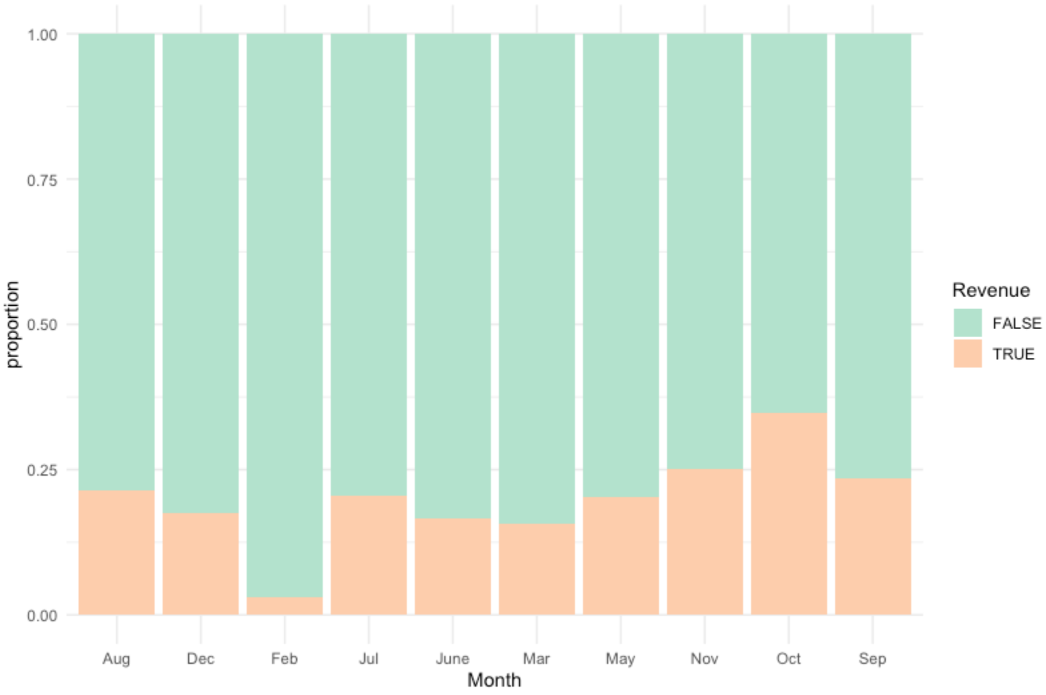
PREDICTION RESULTS ACCORDING TO VARIABLE “MONTH”

```
> counts_table

        Aug Dec Feb Jul June Mar May Nov Oct Sep
  FALSE  73 305  33  70   45 311 533 423  79  78
  TRUE   20  65   1  18    9  58 136 142  42  24
```

```
> proportions_per_month

        Aug Dec Feb Jul June Mar May Nov Oct Sep
  FALSE  78  82  97  80   83  84  80  75  65  76
  TRUE   22  18   3  20   17  16  20  25  35  24
```



The variable “Month” is one of the five variables of interest, taken into consideration in the construction of the different models presented in this report, as predictors of the target variable Revenue.

Analyzing the forecast results on the test set from the point of view of this categorical feature, it is possible to notice a distribution of the *purchase* frequencies that is not particularly inconstant.

In general, the machine learning model built has predicted – on the test set – that the largest number of sessions with purchase will be recorded in May (with 136 transactions). However, if you compare the latter with the number of sessions in the same month that ended without any purchase, you can see that they actually represent only 20%.

The month expected to have the highest proportion of sessions converted into transactions is October (35%). Following the months of September and November, with a lower proportion of 10%. (24-25%).

Conversely, according to the predicted values, the month of February seems to be the least decisive month in influencing the conversion of sessions on eCommerce webpages into purchase intentions by the visitor.