Predict Online Purchasing Behaviour - Professional Certificate in Data Science by HarvardX Capstone Project Own Submission

Velko Kamenov

September 19, 2020

Contents

1. Introduction **1**

2. Analysis **2**
2.1. Data Exploration . . . 2
2.2. Feature Transformations . . . 3
2.3. Features Relationship to Target . . . 4
2.3.1. Numeric Features vs. the Target . . . 4
2.3.2. Categorical Features vs. the Target . . . 8
2.4. Correlations . . . 12
2.5. Train/Test Split . . . 14
2.6. Modelling . . . 14
2.6.1. Logistic Regression . . . 14
2.6.2. Elastic Net . . . 15
2.6.3. Decision Tree . . . 16
2.6.4. Random Forest . . . 18

3. Results **19**

4. Conclusion **19**

1. Introduction

The aim of this report is to examine the Online Shoppers Purchasing Intention Dataset and to develop and evaluate the best machine learning classification model for predicting whether an online customer will make a purchase or not based on the predictor features in the dataset. This project is built under the Professional Certificate in Data Science by HarvardX program as a Capstone Project on a dataset of student's choice.

The dataset used is downloaded from UCI Machine Learning Repository and was created and uploaded by **Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)**.

The raw dataset is available on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/00468/

Four classification machine learning techniques are tested and their results compared to one another - Logistic Regression, Elastic Net, Decision Tree and Random Forest. The algorithms are implemented via the caret package in R and some of their parameters are tuned via cross-validation on the train set.

The algorithms performance is compared based on AUC, Sensitivity, Specificity and Precision on the test set.

The following 3 sections present the analysis, results and conclusions from the modelling.

2. Analysis

In this section of the report are presented the data exploration, data preprocessing, feature engineering, feature relationships analysis as well as the modelling techniques used to generate the final predictive model.

Since the dataset requires modelling on binary classificatin problem - the numeric predictor features are visually examined with the target via box plots while the categorical predictor features are visually examined with the target via segmented bar charts.

2.1. Data Exploration

The raw dataset has 12330 observations and 18 features.

Table 1: Online Purchasing Dataset Variables Summary

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Administrative	5768	46.78	0	0	0	0	integer	27
Administrative_Duration	5903	47.88	0	0	0	0	numeric	3335
Informational	9699	78.66	0	0	0	0	integer	17
Informational_Duration	9925	80.49	0	0	0	0	numeric	1258
ProductRelated	38	0.31	0	0	0	0	integer	311
ProductRelated_Duration	755	6.12	0	0	0	0	numeric	9551
BounceRates	5518	44.75	0	0	0	0	numeric	1872
ExitRates	76	0.62	0	0	0	0	numeric	4777
PageValues	9600	77.86	0	0	0	0	numeric	2704
SpecialDay	11079	89.85	0	0	0	0	numeric	6
Month	0	0.00	0	0	0	0	character	10
OperatingSystems	0	0.00	0	0	0	0	integer	8
Browser	0	0.00	0	0	0	0	integer	13
Region	0	0.00	0	0	0	0	integer	9
TrafficType	0	0.00	0	0	0	0	integer	20
VisitorType	0	0.00	0	0	0	0	character	3
Weekend	9462	76.74	0	0	0	0	logical	2
Revenue	10422	84.53	0	0	0	0	logical	2

We see that there are no missing values accross the variables in the dataset.

Here are the features meanings given by the dataset providers:

- **Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration** - number of different types of pages visited by the visitor in that session and total time spent in each of these page categories
- **Bounce Rate** - Google Analytics Metric. 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
- **Exit Rate** - Google Analytics Metric. Feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session
- **Page Value** - the average value for a web page that a user visited before completing an e-commerce transaction.
- **Special Day** - 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.
- **Operating system** - Operating System of the user.

- **Browser** - Browser of the user.
- **Region** - Region of the user.
- **Month** - month of the transaction
- **Weekend** - Flag if the transactions was made during the weekend.
- **Traffic type** - Traffic Type of the user
- **Visitor type** - New or Returning Visitor
- **Revenue** - The target column. TRUE values mean a purchase was made and FALSE mean a purchase was not made. This is the target variable we are going to build a classification model to forecast.

Table 2: Outcome Variable Summary

Revenue	Frequency
FALSE	10422
TRUE	1908

From the distribution of the the two possible outcomes - True or False we see that a lot more clients do not make a purchase. The number of observations with no purchases is 10422 while the number of observations with purchase is 1908. We must keep this in mind when dividing the data in Train/Test samples and to do the sample split with **stratification** - i.e. with keeping the proportion of True and False outcomes in the Train and Test Sample relatively close.

2.2. Feature Transformations

We see that some column data types do not correspond correctly to the statistical data type of the variables behind them and some column names can be changed in order to make the interpretation easier and more intuitive. That is why some of the feature types and column names are changed.

Here is the summary of the dataset after transformations of column types:

Table 3: Online Purchasing Transformed Dataset Variables Summary

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Administrative	5768	46.78	0	0	0	0	integer	27
Administrative_Duration	5903	47.88	0	0	0	0	numeric	3335
Informational	9699	78.66	0	0	0	0	integer	17
Informational_Duration	9925	80.49	0	0	0	0	numeric	1258
ProductRelated	38	0.31	0	0	0	0	integer	311
ProductRelated_Duration	755	6.12	0	0	0	0	numeric	9551
BounceRates	5518	44.75	0	0	0	0	numeric	1872
ExitRates	76	0.62	0	0	0	0	numeric	4777
PageValues	9600	77.86	0	0	0	0	numeric	2704
SpecialDay	11079	89.85	0	0	0	0	factor	6
Month	0	0.00	0	0	0	0	factor	10
OperatingSystems	0	0.00	0	0	0	0	factor	8
Browser	0	0.00	0	0	0	0	factor	13
Region	0	0.00	0	0	0	0	factor	9
TrafficType	0	0.00	0	0	0	0	factor	20
VisitorType	0	0.00	0	0	0	0	factor	3
Weekend	0	0.00	0	0	0	0	factor	2

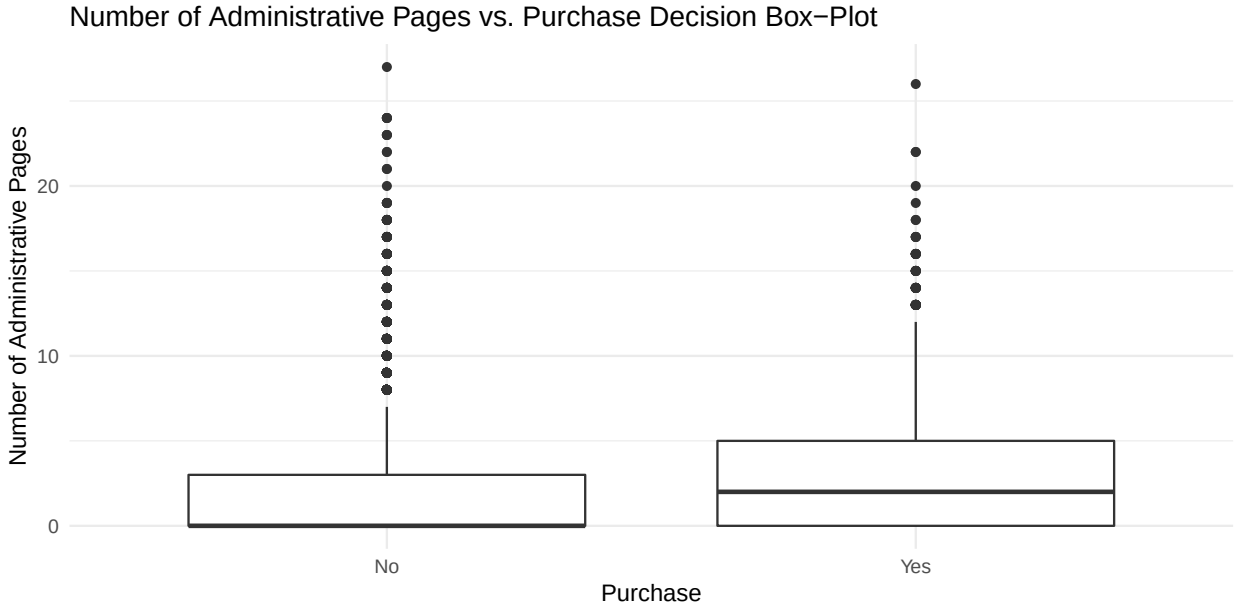
variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Purchase	10422	84.53	0	0	0	0	factor	2
Purchase_Yes_No	0	0.00	0	0	0	0	character	2

2.3. Features Relationship to Target

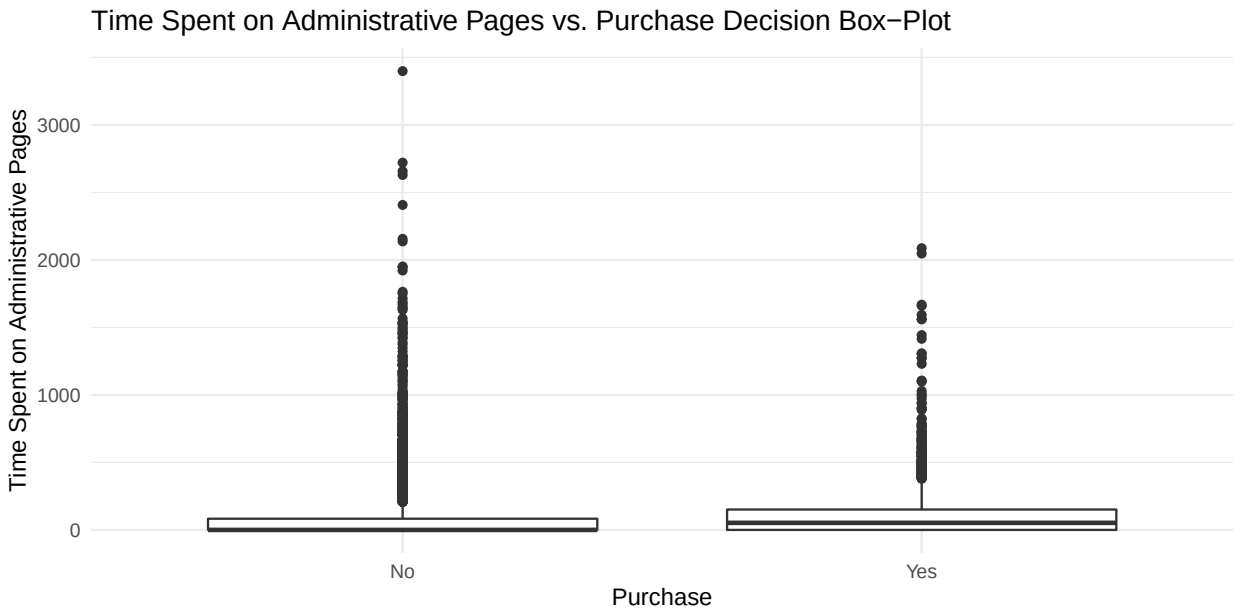
In this section we examine visually the features relationships to target. For the numeric predictor variables we use box plots and for categorical predictor geatures we use segmented 100% bar charts.

2.3.1. Numeric Features vs. the Target

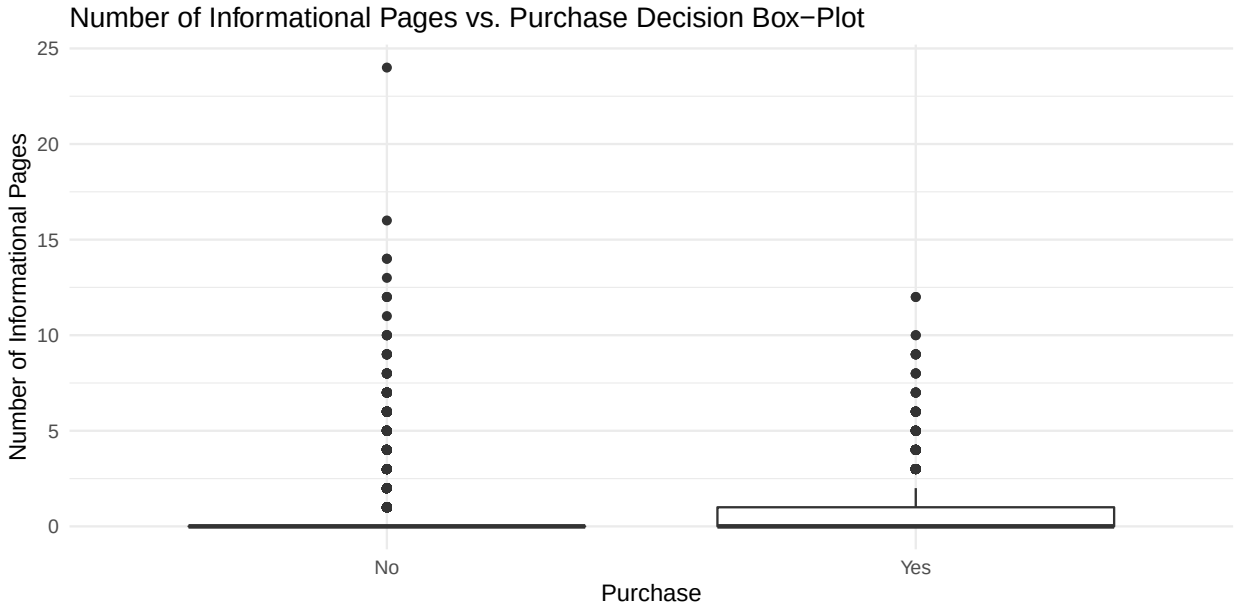
The higher the Number of Administrative Pages the higher the probability that a transaction will end up with purchase.



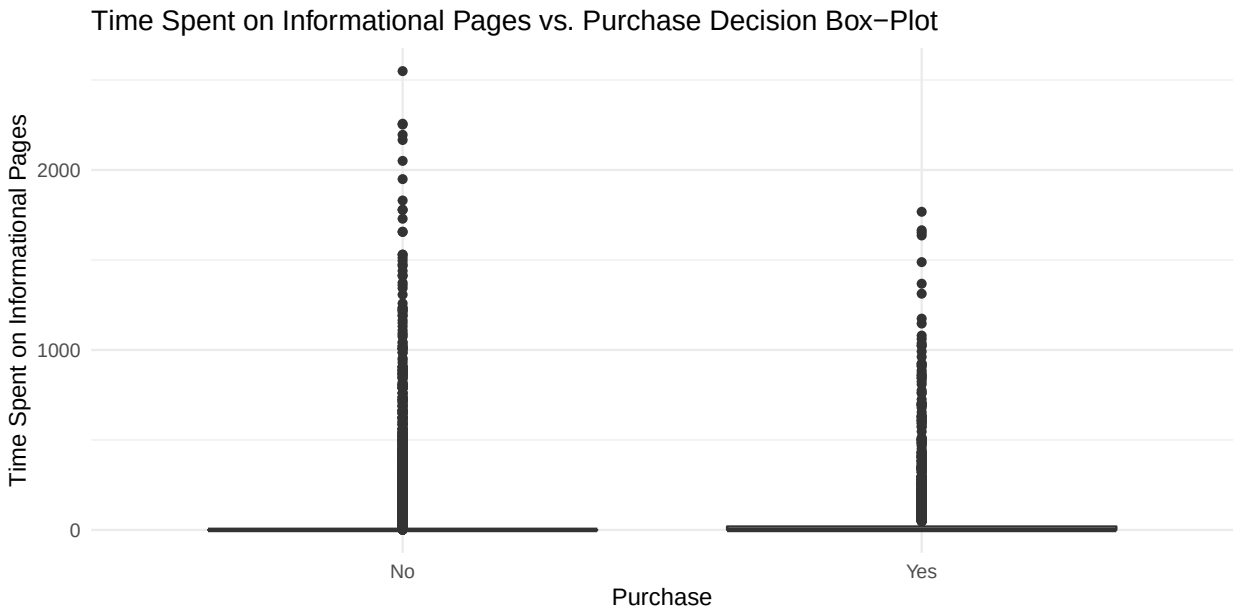
The higher the Time Spent on Administrative Pages the higher the probability that a transaction will end up with purchase.



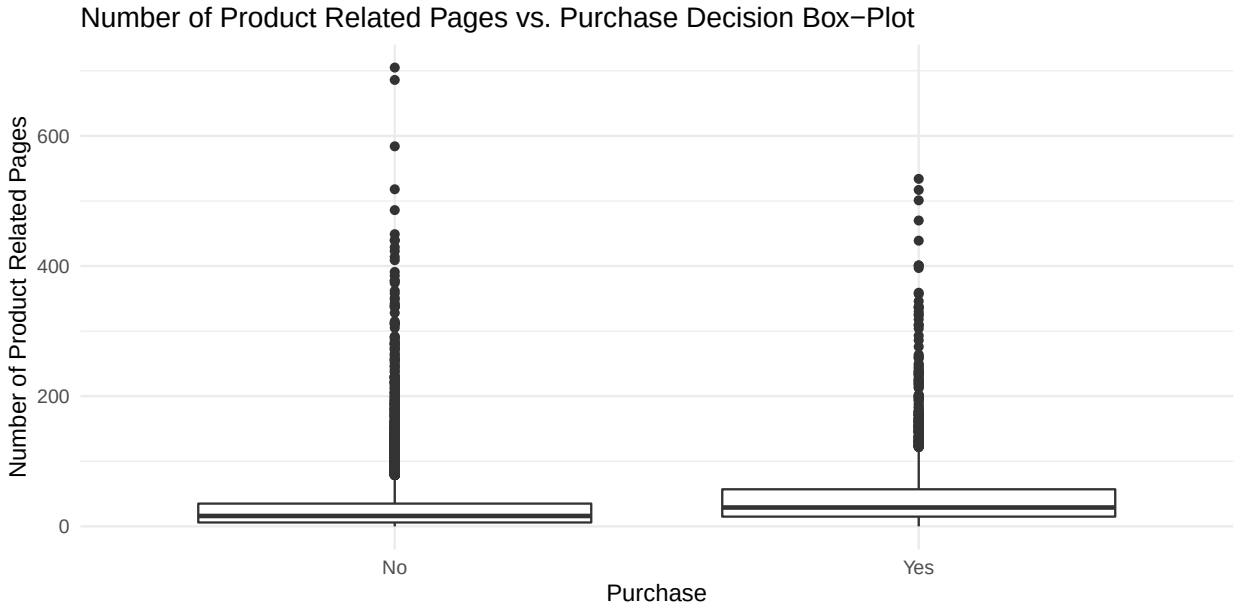
No clear relationship between the Number of Informational Pages visited and the probability that a transaction will end up with purchase can be observed.



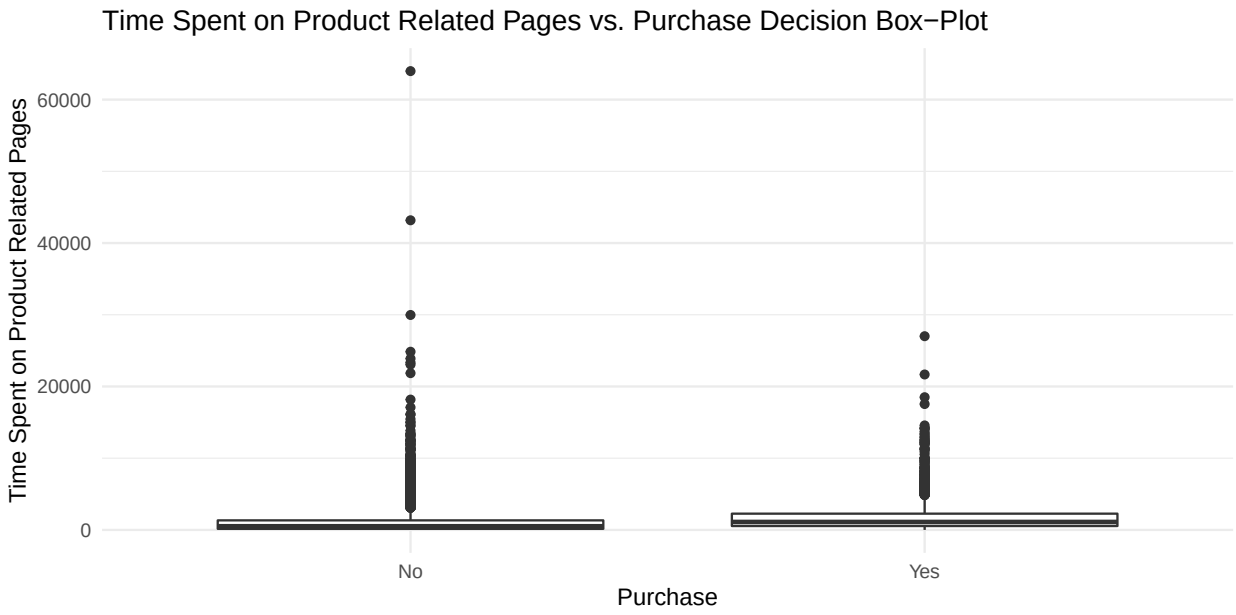
No clear relationship between the Time Spent on Informational Pages visited and the probability that a transaction will end up with purchase can be observed.



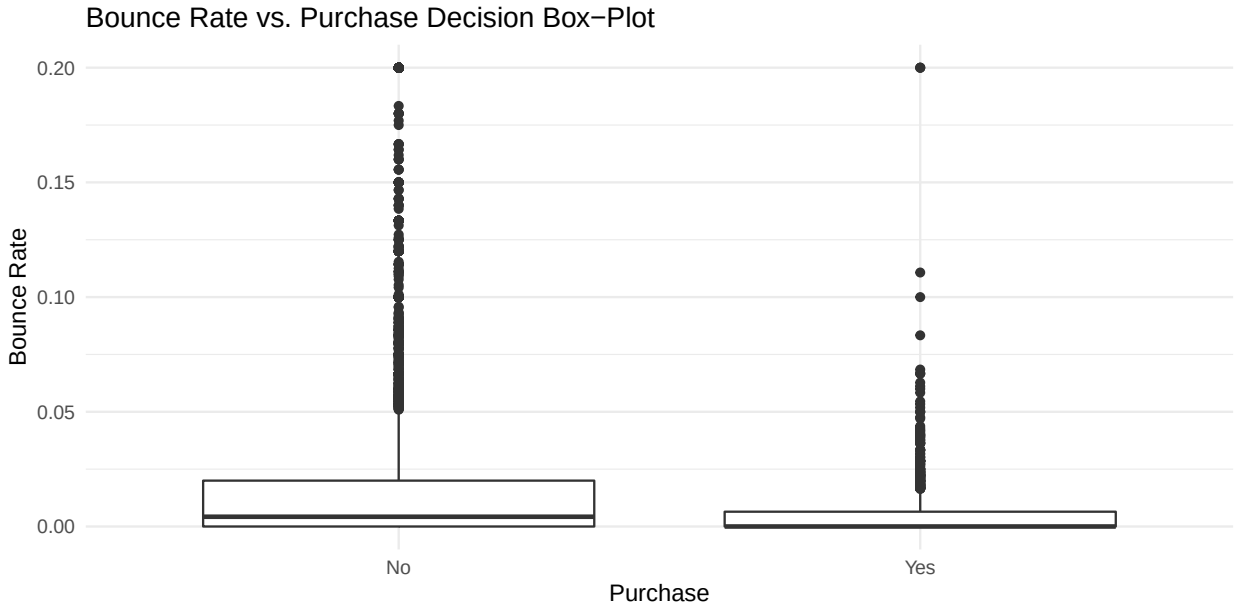
The higher the Number of Product Related Pages the higher the probability that a transaction will end up with purchase.



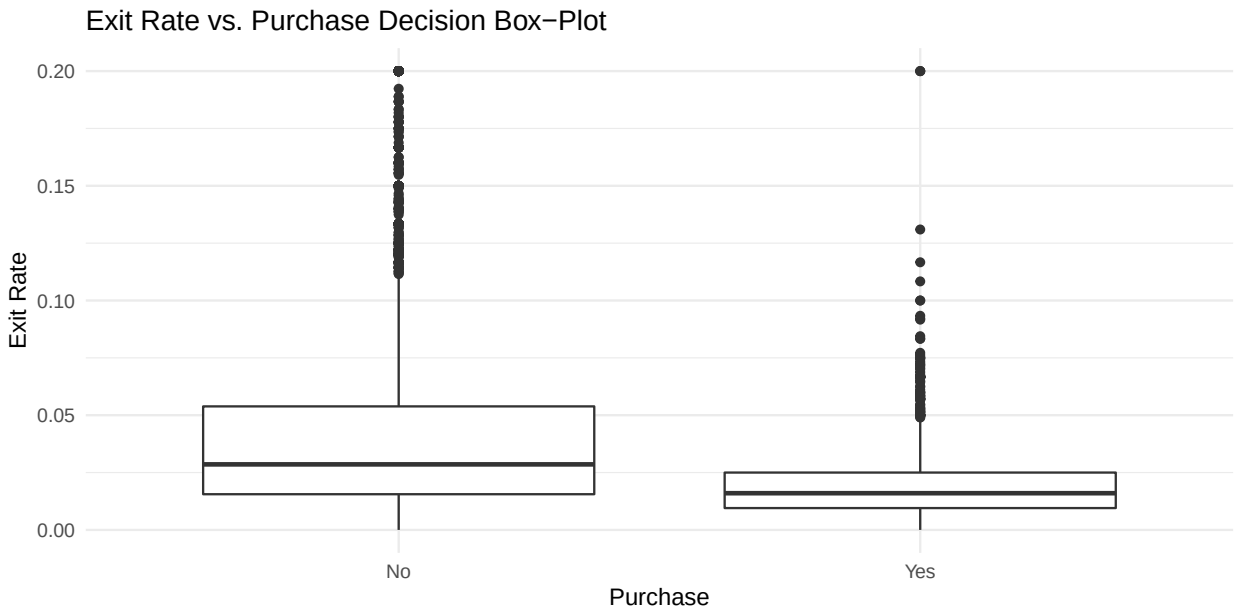
No clear relationship between the Time Spent on Product Related Pages visited and the probability that a transaction will end up with purchase can be observed.



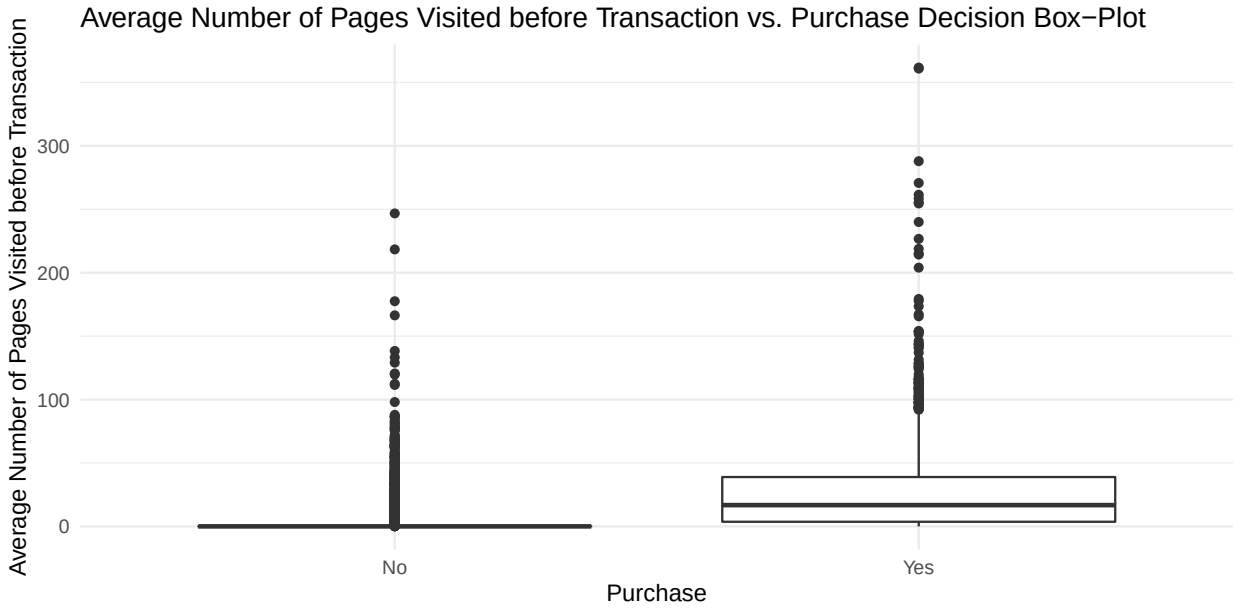
The higher the Bounce Rate the higher the lower probability that a transaction will end up with purchase.



The higher the Exit Rate the higher the lower probability that a transaction will end up with purchase.

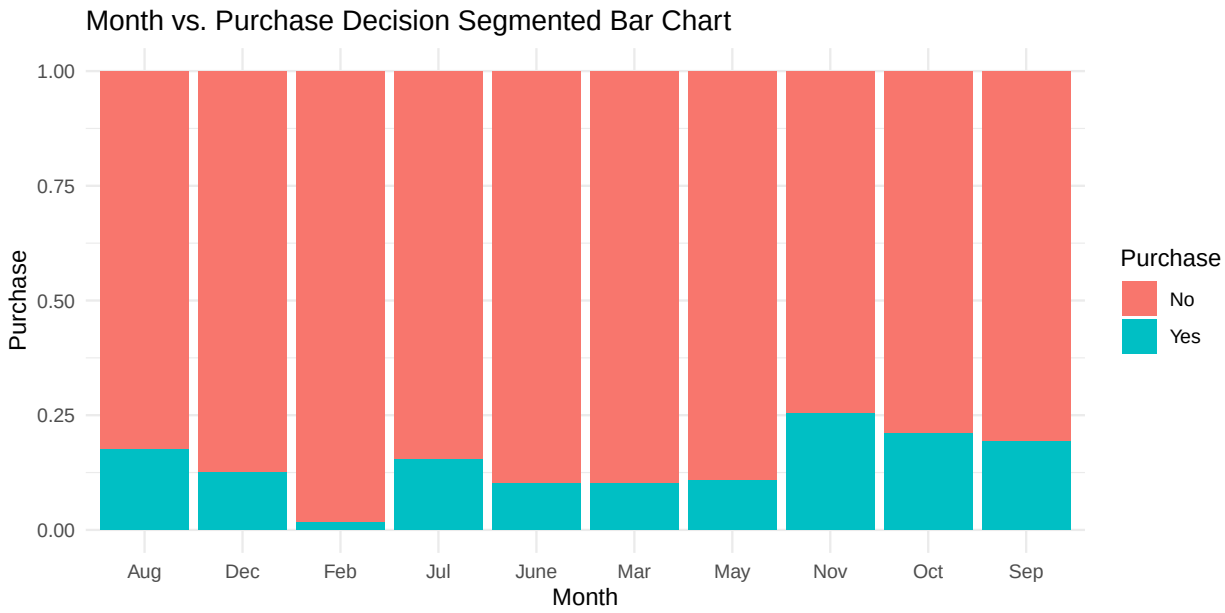


The higher Average Number of Pages Visited before Transaction the higher the probability that a transaction will end up with purchase.

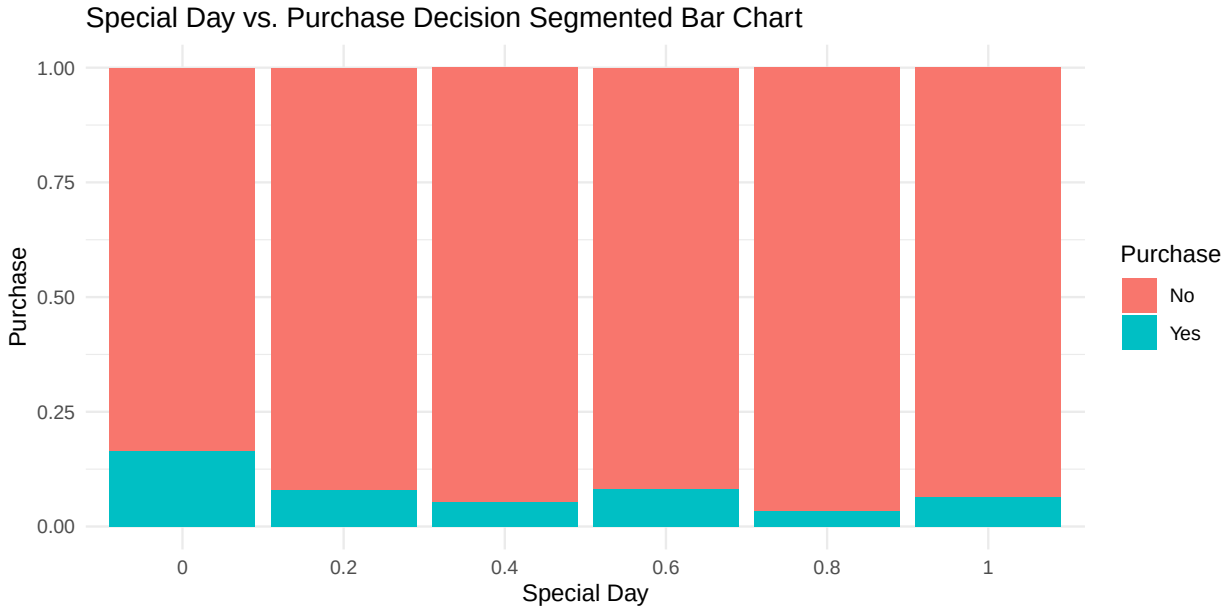


2.3.2. Categorical Features vs. the Target

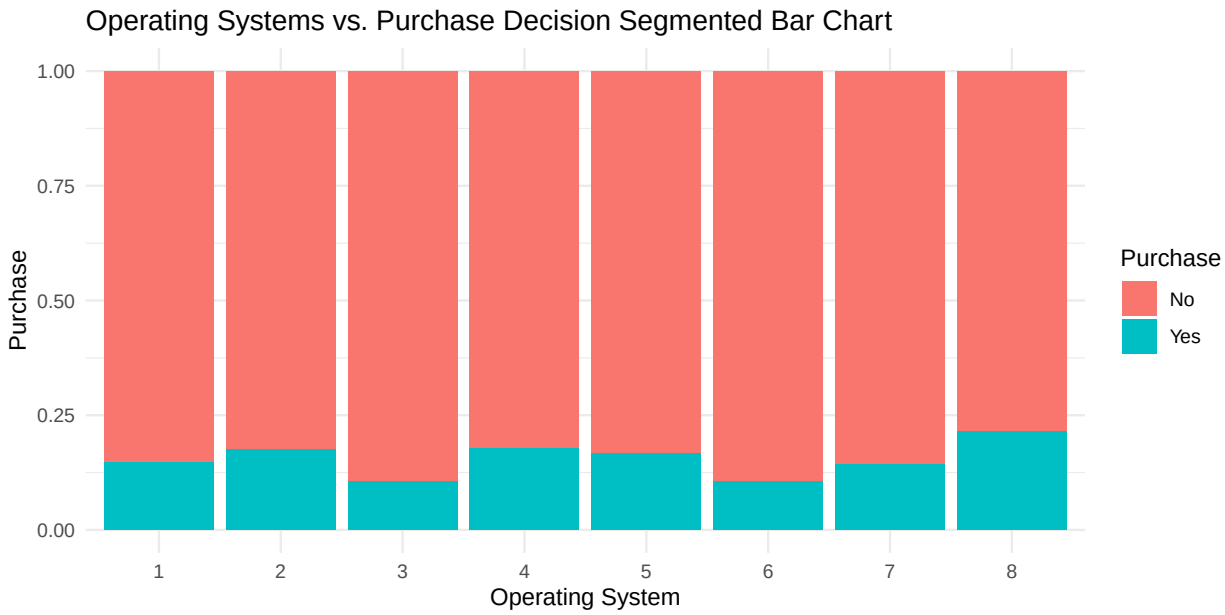
The months November, October, September, August, July are months in which it is more likely for a transaction to end up as a purchase.



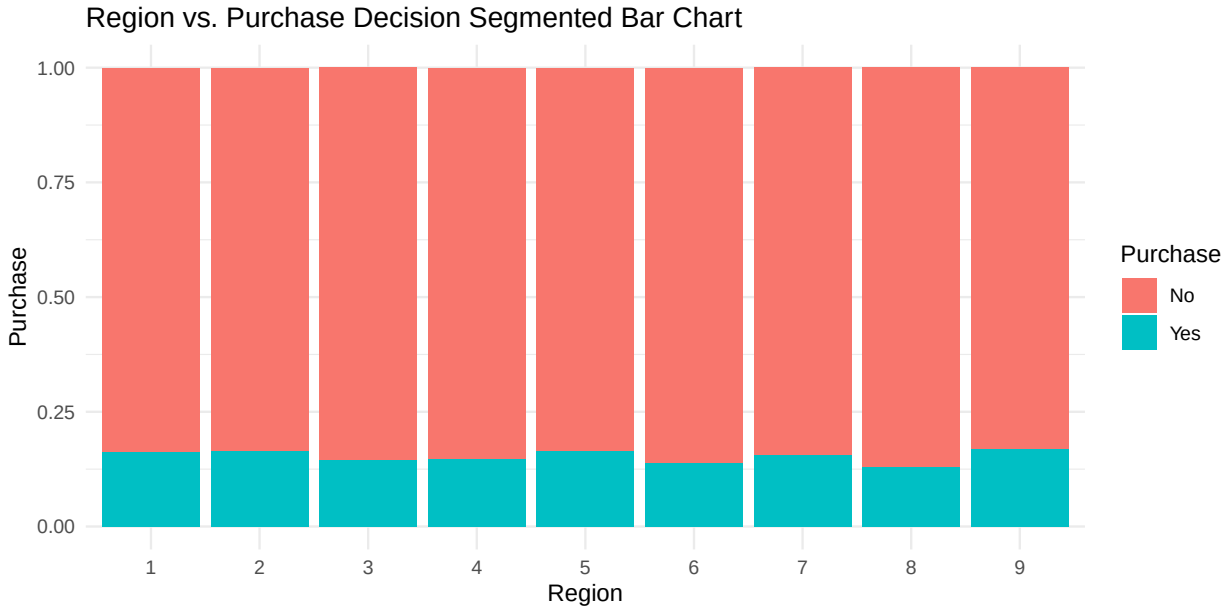
Special Day values of 0 are more likely to end up with a purchase.



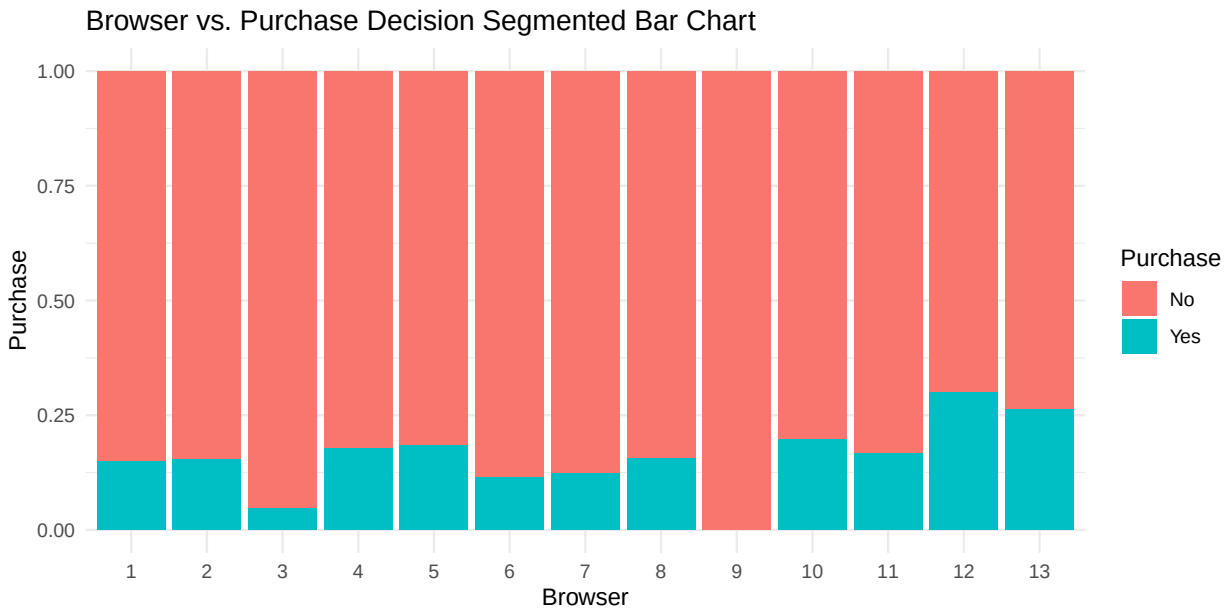
Operating Systems 2,4,5,7,8 are more likely to end up with a purchase.



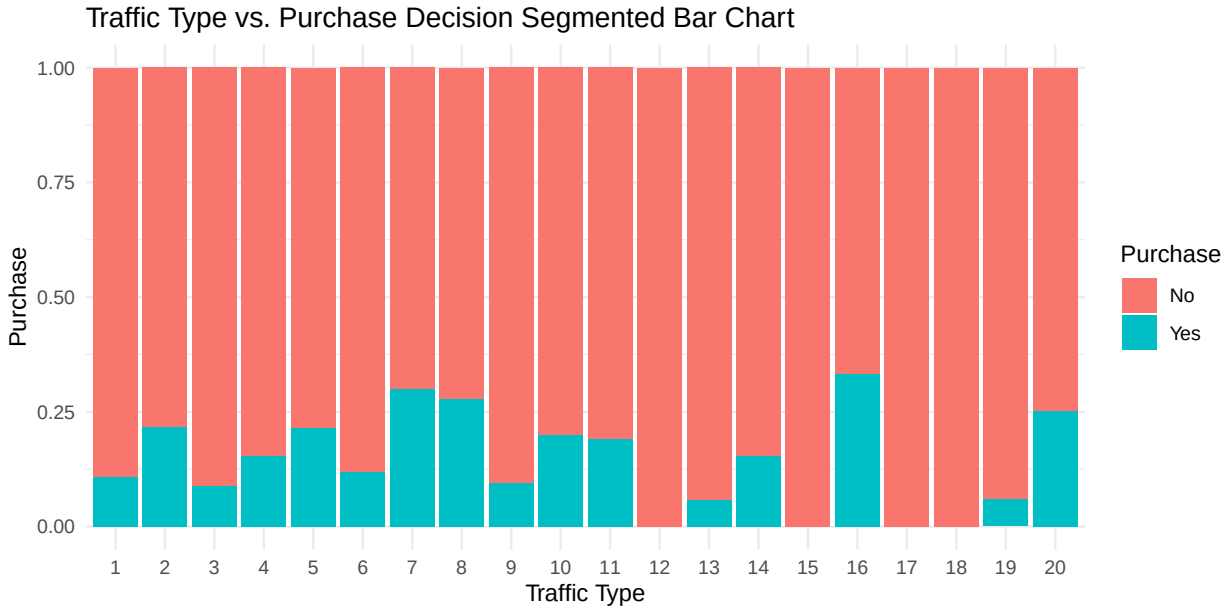
No clear relationship between region and purchase behaviour can be observed.



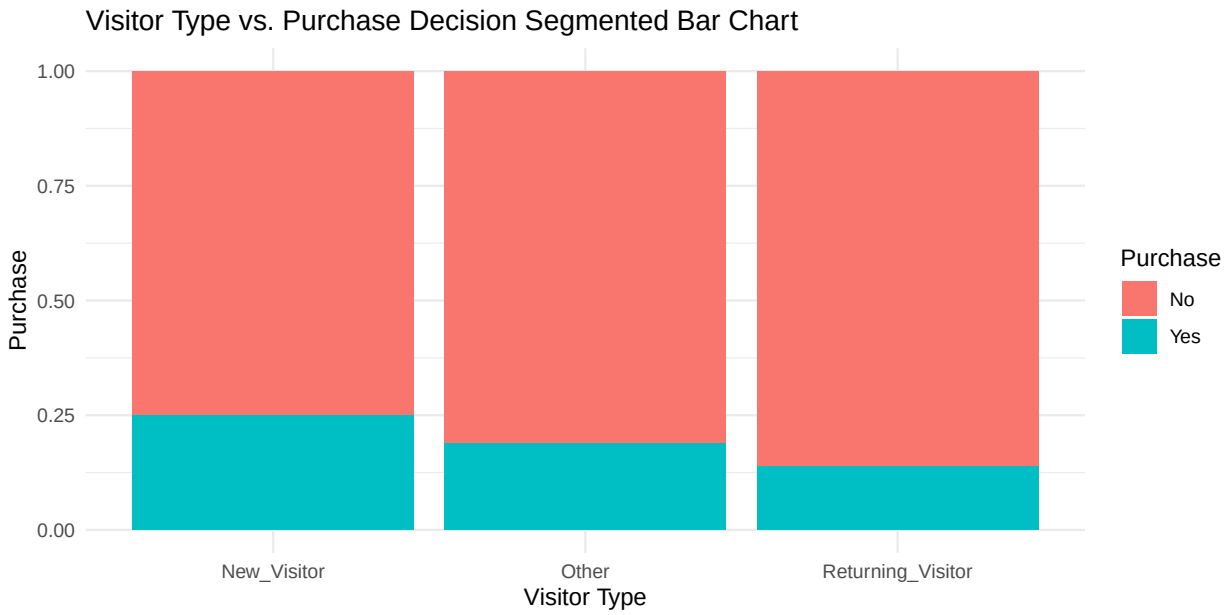
Traffic coming from browsers 4,5,10,11,12,13 is more likely to lead to a purchase.



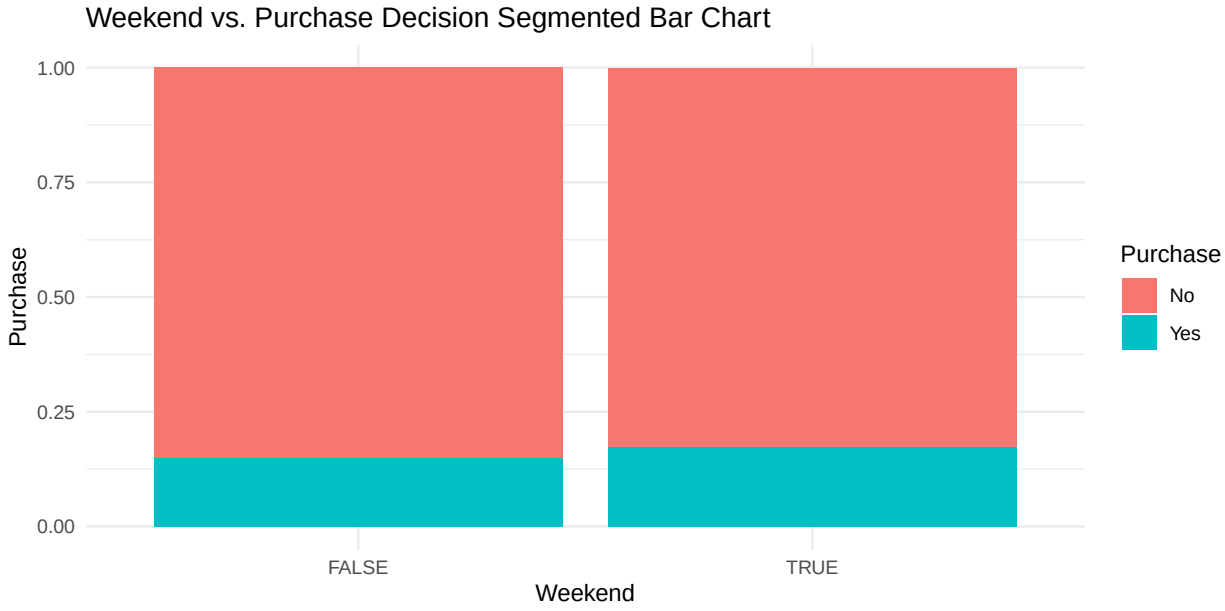
Traffic type 2,4,5,7,8,10,11,16 and 20 are more likely to lead to a purchase.



New visitors are more likely to make a purchase.



No clear relationship between weekend or weekday and purchase behaviour can be observed.



2.4. Correlations

In this section we examine the correlations among numeric and categorical variables. For linear models like Logistic Regression we know it is not good to include correlated variables in the final models.

We see that some variables have strong correlations between them. The correlation between Exit Rate and Bounce Rate is 0.91. The correleation between Product Related and Product Related Duration is 0.86. The correlation between Administrative and Administrative Duration is 0.6. The correlation between Informational, and Informational, Duration is 0.62.

For the numeric variables we calculate the Pearson Correlation Coefficient. This is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and ???1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and ???1 is total negative linear correlation.

For better visibility the following naming convetion is used for the correlation coefficients:

- Administrative - A
- Administrative Duration - AD
- Informational - I
- Informational Duration - ID
- Product Related - PR
- Product Related Duration - PRD
- Bounce Rates - BR
- Exit Rates - ER
- Page Value - PV
- Special Day - SD
- Month - M
- Operating System - OS
- Browser - B
- Region - R
- Trafic Type - TT
- Visitor Type - VT
- Weekeng - W

Table 4: Pearson correlation coefficient among numeric variables

variable	A	AD	I	ID	PR	PRD	BR	ER	PV
A	1.00	0.60	0.38	0.26	0.43	0.37	-0.22	-0.32	0.10
AD	0.60	1.00	0.30	0.24	0.29	0.36	-0.14	-0.21	0.07
I	0.38	0.30	1.00	0.62	0.37	0.39	-0.12	-0.16	0.05
ID	0.26	0.24	0.62	1.00	0.28	0.35	-0.07	-0.11	0.03
PR	0.43	0.29	0.37	0.28	1.00	0.86	-0.20	-0.29	0.06
PRD	0.37	0.36	0.39	0.35	0.86	1.00	-0.18	-0.25	0.05
BR	-0.22	-0.14	-0.12	-0.07	-0.20	-0.18	1.00	0.91	-0.12
ER	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1.00	-0.17
PV	0.10	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1.00

For the categorical variables we calculate the Spearman Correlation Coefficient.The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables. Spearman's correlation assesses monotonic relationships (whether linear or not).

Table 5: Spearman correlation coefficient among numeric and categorical variables

variable	A	AD	I	ID	PR	PRD	BR	ER	PV	SD	M	OS	B	R	TT	VT	W
A	1.0	0.9	0.4	0.4	0.5	0.4	-0.2	-0.4	0.3	-0.1	0.1	0.0	0.0	0.0	0.0	-0.1	0.0
AD	0.9	1.0	0.4	0.4	0.4	0.4	-0.2	-0.4	0.3	-0.1	0.1	0.0	0.0	0.0	0.0	-0.1	0.0
I	0.4	0.4	1.0	1.0	0.4	0.4	0.0	-0.2	0.2	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
ID	0.4	0.4	1.0	1.0	0.4	0.4	0.0	-0.2	0.2	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
PR	0.5	0.4	0.4	0.4	1.0	0.9	-0.1	-0.5	0.3	0.0	0.1	0.0	0.0	0.0	-0.1	0.1	0.0
PRD	0.4	0.4	0.4	0.4	0.9	1.0	-0.1	-0.5	0.4	0.0	0.1	0.0	0.0	0.0	-0.1	0.1	0.0
BR	-0.2	-0.2	0.0	0.0	-0.1	-0.1	1.0	0.6	-0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.3	0.0
ER	-0.4	-0.4	-0.2	-0.2	-0.5	-0.5	0.6	1.0	-0.3	0.2	-0.1	0.0	0.0	0.0	0.0	0.3	-0.1
PV	0.3	0.3	0.2	0.2	0.3	0.4	-0.1	-0.3	1.0	-0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
SD	-0.1	-0.1	-0.1	-0.1	0.0	0.0	0.1	0.2	-0.1	1.0	0.0	0.0	0.0	0.0	0.1	0.1	-0.1
M	0.1	0.1	0.0	0.0	0.1	0.1	0.0	-0.1	0.1	0.0	1.0	0.0	0.0	0.0	0.1	0.0	0.0
OS	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	1.0	0.4	0.0	0.1	0.0	0.0

variable	A	AD	I	ID	PR	PRD	BR	ER	PV	SD	M	OS	B	R	TT	VT	W
B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.0	0.1	0.0	0.0	-0.1
R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.0	0.0	0.0
TT	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	1.0	-0.1	0.0
VT	-0.1	-0.1	0.1	0.1	0.1	0.1	0.3	0.3	0.0	0.1	0.0	0.0	0.0	0.0	-0.1	1.0	0.0
W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	-0.1	0.0	0.0	-0.1	0.0	0.0	0.0	1.0

2.5. Train/Test Split

We divide the sample in train/test sets in proportions 70%/30%. These proportions are chosen as the best balance between getting enough data for training and testing given the total number of observations in the dataset. If the dataset consisted of more observations a ratio of 80%/20% or even 90%/10% would have been suitable. But since the observations in the analysed dataset are just 12330 it is good to set 30% aside as test set in order to get enough observations to make valid conclusions.

The split is done with the createDataPartition function from the caret package which makes the split stratified. This is needed because the target variable is imbalanced.

We get the following samples:

- Train Sample with 8632 observations and 15.48% of purchases made.
- Test Sample with 3698 observations and 15.47% of purchases made.

2.6. Modelling

This section presents the modelling techniques used on the data as well as comparison between the results from the predictive models. Given the fact that the outcome has two classes that are highly imbalanced the Area Under the Curve (AUC), Sensitivity, Specificity and Precision are going to be used as the main metrics to evaluate the quality of the predictive models. The accuracy metric is going to be biased as a model predicting no purchases would result in 85% accuracy.

The chosen metric for model evaluation are explained here:

- **AUC** - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.
- **Sensitivity** - measures the proportion of positives that are correctly identified.
- **Specificity** - measures the proportion of negatives that are correctly identified.
- **Precision** - is the fraction of relevant instances among the retrieved instances.

Because the problem we are trying to solve aims at predicting if a purchase is going to be made or not - **the Precision metric can be viewed as one with special importance along with AUC because it is expected that a strategy and some costs related to clients predicted to make a purchase are going to be persued.**

2.6.1. Logistic Regression

The Logistic Regression Model is the first which predictive power is tested. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is a linear model in terms of log of odds and hence represent a simplistic approach good to be tested as benchmark for comparison with more complicated models.

Only variables with no high correlations with other variables are included in the final model and also only statistically significant variables and categories are included in the final model.

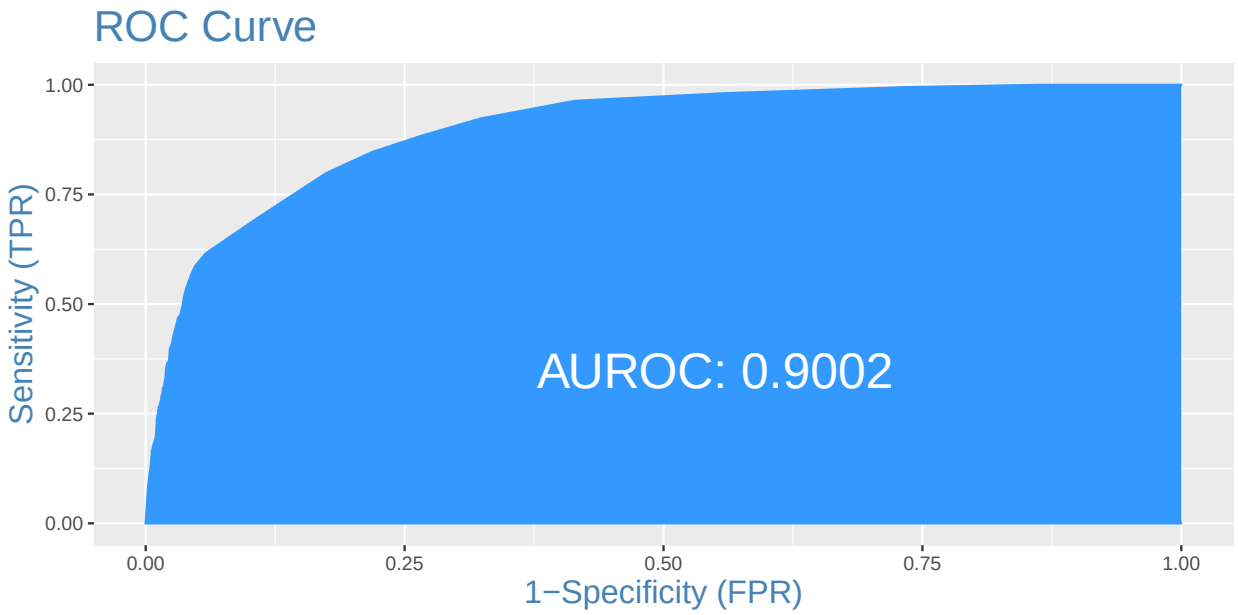
The Logistic Regression Model gives the following results on the test set:

- AUC: 0.901
- Sensitivity: 75%
- Specificity: 90%
- Precision: 38%

They are achived by the following model:

Table 6: Logistic Regression Model Summary

Variable	Coefficient	Standard_Error	z_value	p_value
(Intercept)	-2.427	0.092	-26.287	0.0000
ExitRates	-17.938	1.858	-9.655	0.0000
PageValues	0.081	0.003	28.395	0.0000
`MonthNov-Oct-Sep-Aug-Jul`	0.973	0.075	12.988	0.0000
`OperatingSystems3-6`	-0.305	0.099	-3.078	0.0021
`TrafficType2-5-7-8-10-11-16-20`	0.362	0.075	4.806	0.0000



- **!N.B.** Please note that the AUC values in the plots differ slightly from the calculated in the code. This is done because of differences in calculations of the integrals needed for the AUC estimate!

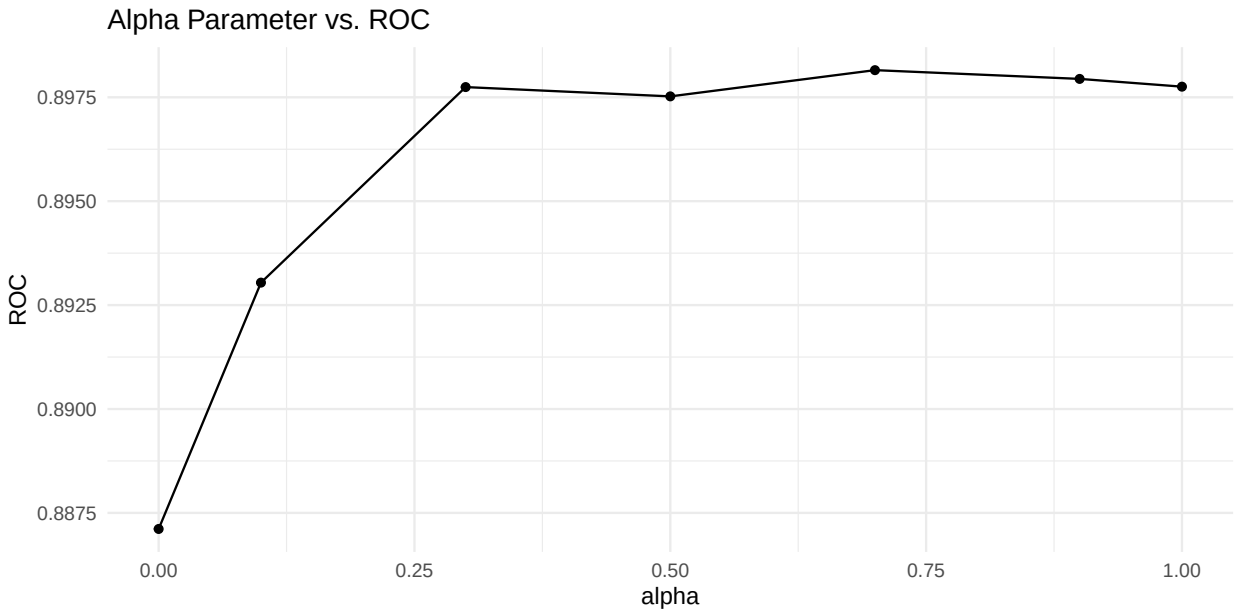
2.6.2. Elastic Net

The elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. It shrinks the coefficients of correlated variables to obtain stable model.

Two parameters of the Elastic Net model are tuned via 5-fold cross validation in the train set because they can have the biggest effect on the model. The 5-fold corss validation for an elastic net model is stable and not much computantional power is needed:

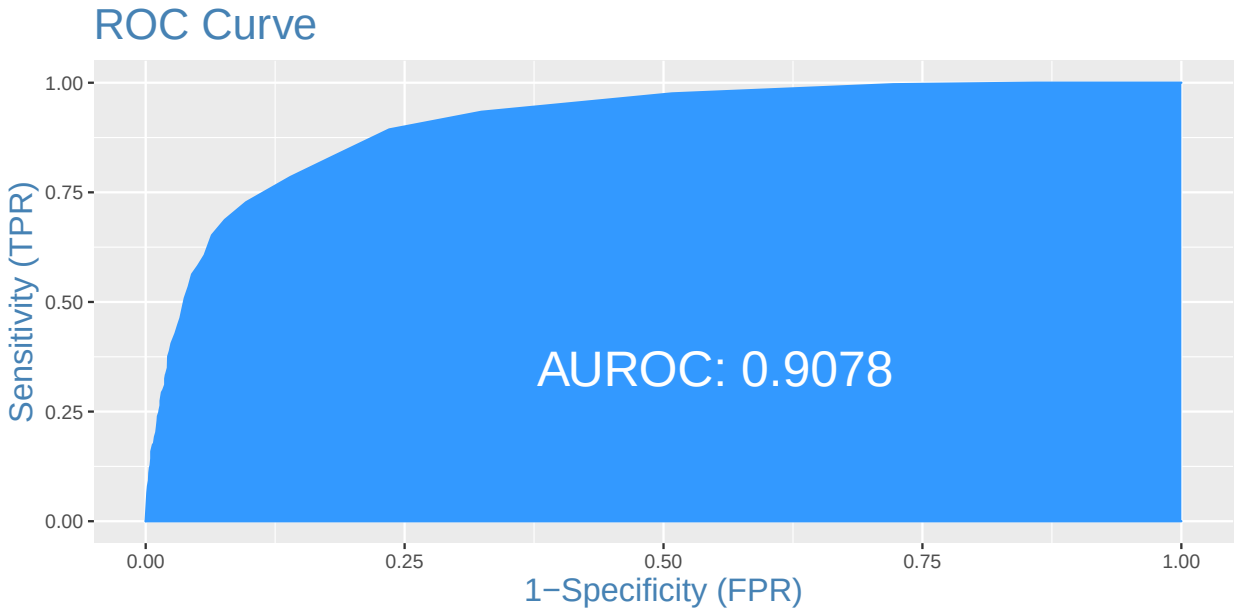
- **alpha** - the degree of mixing between ridge regression and lasso regression. The closer to 0 - the closer the model to ridge regression. The closer to 1 - the closer the model to lasso regression.
- **lambda** - the shrinkage parameter. When equal to 0 - no shrinkage is performed. The higher than zero - the more shrinkage in coefficients.

From the ROC plot we see that the optimal value for alpha parameter is 0.7 and for the lambda parameter is 0.01.



The Elastic Model gives the following results on the test set:

- AUC: 0.909
- Sensitivity: 76%
- Specificity: 89%
- Precision: 34%



2.6.3. Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf

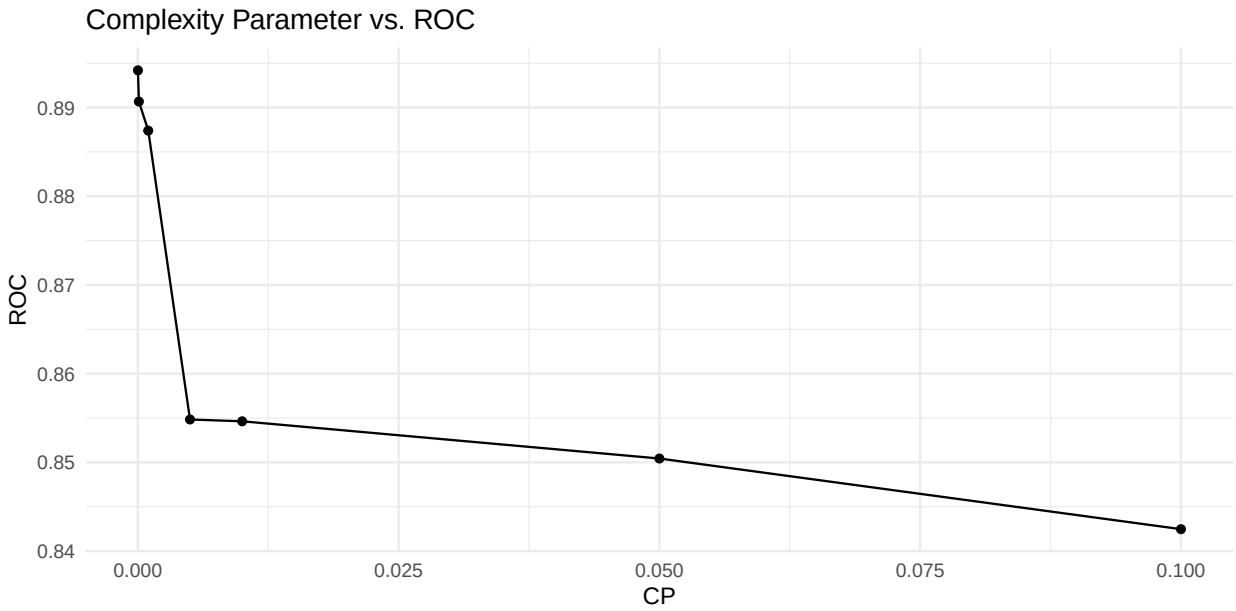
represent classification rules.

The **complexity parameter (cp)** is tuned. This parameter is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then tree building does not continue.

This is the most important parameter in the decision tree model and the caret package allows for tuning.

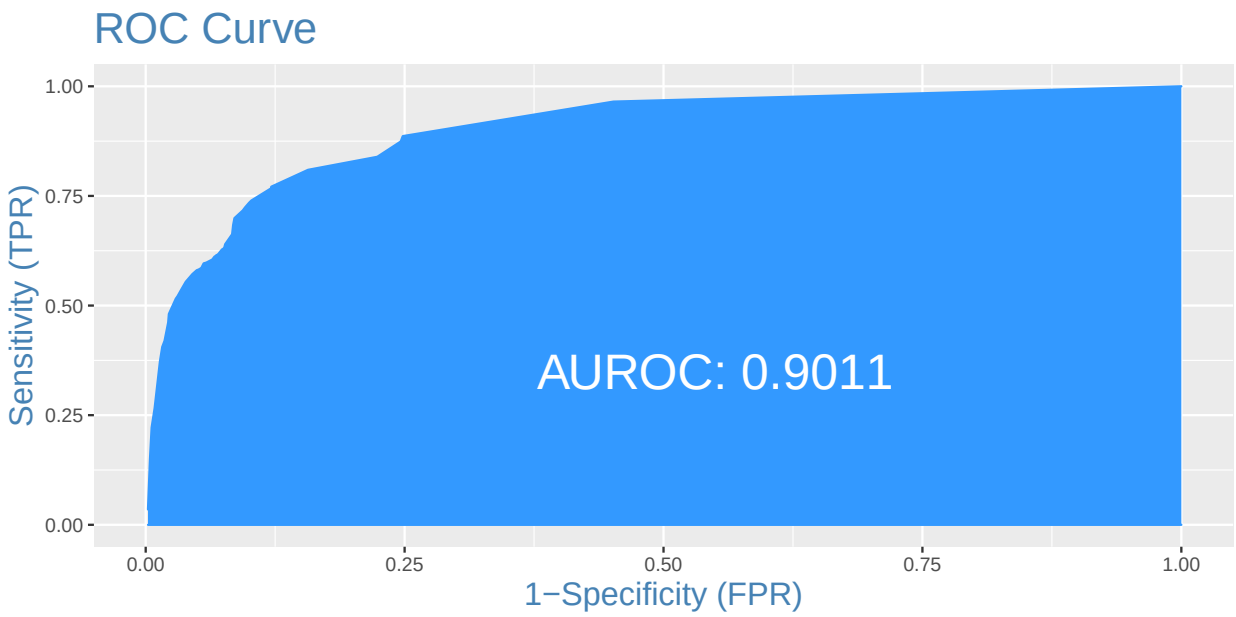
The tuning is done via 5-fold cross valition. The 5-fold corss validation for a decision tree model is stable and not much computantional power is needed to perform it.

From the ROC plot we see that the optimal value for the cp parameter is 0.00001:



The Decision Tree Model gives the following results on the test set:

- AUC: 0.898
- Sensitivity: 63%
- Specificity: 93%
- Precision: 61%



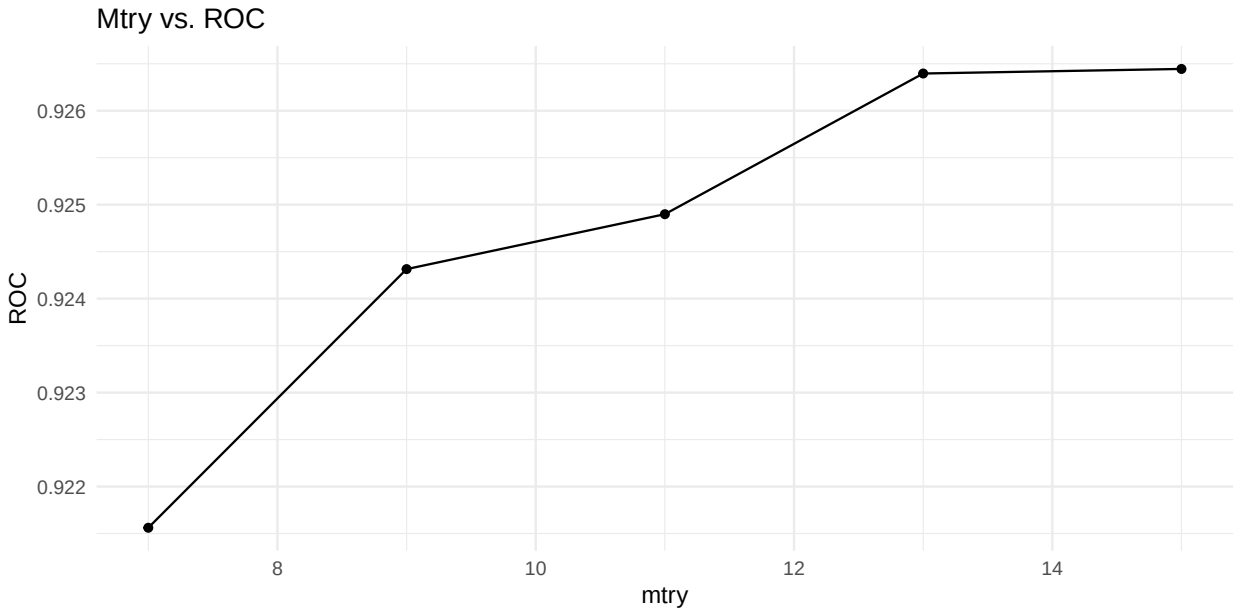
2.6.4. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate 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.

The **mtry** parameter is tuned. This is the parameter which sets the number of variables available for splitting at each tree node. It is the most important paramter in Random Forest model and can be tuned in caret.

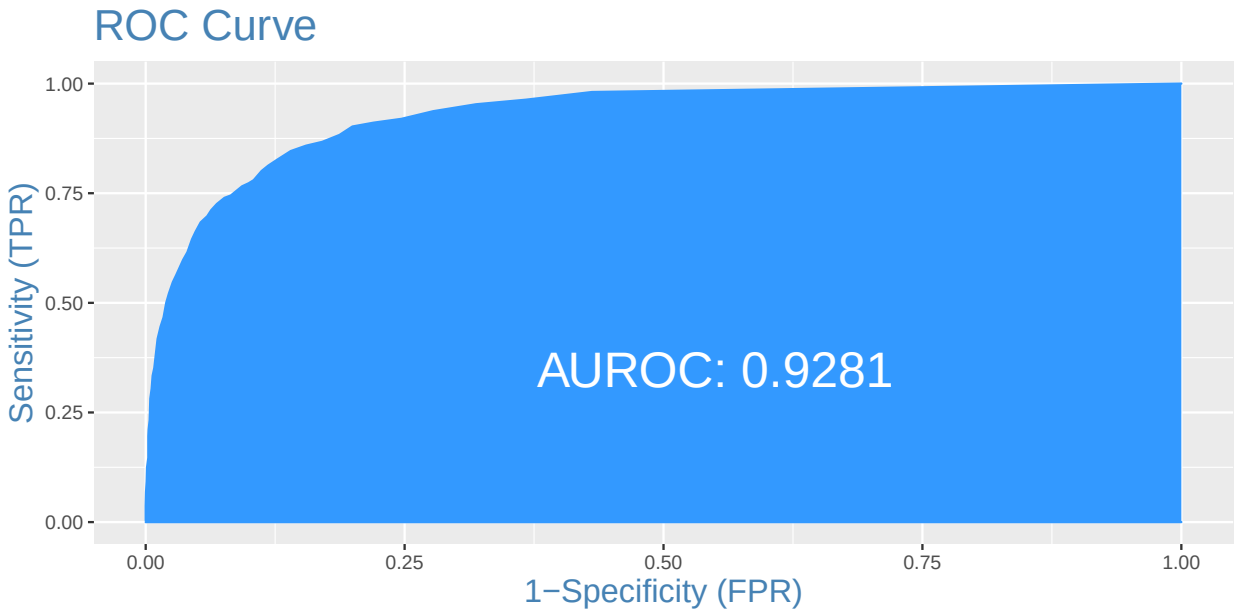
A 3-fold cross validation on train set is used. The value of 3 is chosen in order to speed up the calculations. Higehr values of k would need more computational time.

From the ROC plot we see that the optimal value for the mtry parameter is 15:



The Random Forest Model gives the following results on the test set:

- AUC: 0.931
- Sensitivity: 72%
- Specificity: 94%
- Precision: 66%



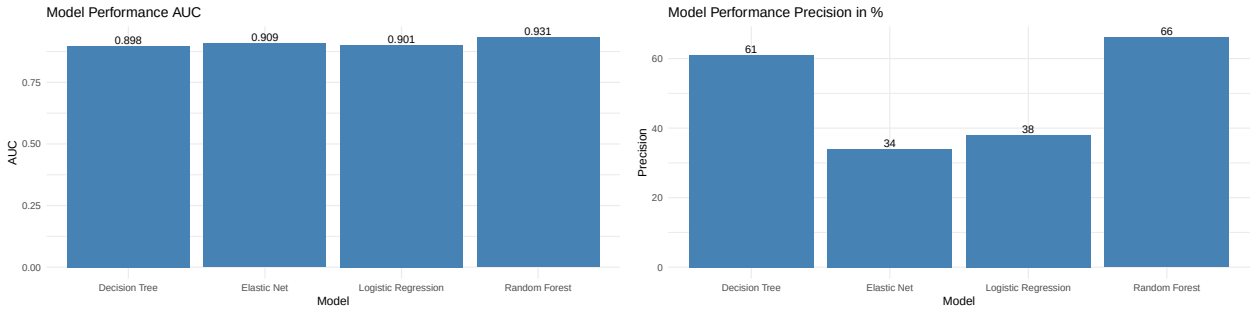
3. Results

Based on the two most important metrics for the model evaluation - AUC and Precision - there is a clear winner among the 4 models tuned and tested and this is the Random Forest Model.

The final model is the Random Forest with a tuned value of mtry parameter of 15.

The 66% precision obtained by the random forest means that from all customers predicted to make a purchase 66% would indeed make a purchase. Moreover the 72% value of sensitivity means that 72% of all purchasing clients are going to be identified by the model.

This gives very good opportunities to build optimized strategies for profit maximization.



4. Conclusion

The Online Shoppers Purchasing Intention Dataset provided very good data for predictive purposes whether an online shopper would make a purchase or not. Because the outcome column indicating if a transactions ends with a purchase or not is very imbalanced - around 15% of the transactions end with a purchase - the

AUC and precision metrics were chosen as the most important ones for model evaluation. The accuracy metric is very biases in cases of binary classification when one of the predicted classes is overrepresented.

Four models were tested and evaluated on test set - Logistic Regression, Elastic Net, Decision Tree and Random Forest. Because of the non-linear fashion of the data the Random Forest Model gave the best results as it comes to AUC and Precision. The best Random Forest model gives AUC of 0.931 and Precision of 66%.

These values provide very good opportunities to build optimized strategies for profit maximization by better targeting purchasing clients.

There is room for future improvement of the model by testing other machine learning models suitable for non-linear data such as xgBoost. The Random Forest model can be further developed if more parameters go through a tuning procedure. The results presented in this paper rely on tuning of just the mtry parameter.