ONLINE SALES PREDICTIONS

Naive Bayes & Decision Tree

PREPARED BY

Jaclyn Glosson

August 24, 2021

EXECUTIVE SUMMARY

Decision Tree and Naive Bayes methods were used to build a model predicting purchasing behavior of website visitors. Data included over 12,000 observations with 18 numeric and categorical variables. The data included redundant variables, irrelevant variables, and large amounts of variation and noise. The Decision Tree method outperformed the Naive Bayes method. The final recommended Decision Tree model was able to correctly predict who would purchase from the company at a rate of approximately 80%, as well as who would not purchase from the company at a rate of approximately 86%. PageValue was identified as the most predictive variable, thus the company should prioritize collecting data on this variable.

INTRODUCTION

The objective of the business case is to use given data regarding online traffic to predict which customers will make a purchase. The company will tailor their advertising based on these predictions, and advertising will be targeted to those customers who have already visited the company website, as the predictive model will require online traffic data for input. When the model results in an inaccurate revenue prediction, the company needlessly spends advertising revenue. Both Naive Bayes and Decision Tree methods are used.

Variables

After removing 125 duplicate observations, the data consisted of 12,205 observations with 18 variables. The dataset includes historical website data for an entire calendar year, excluding the months of January and April. The target variable, 'Revenue', is binary and describes if a website visit resulted in a purchase or not. Other variables are described in *Figure 1.1*

Numerical Variables	**Description**
Administrative	number of administrative pages visited by the site visitor
Administrative Duration	time spent on administrative pages
Informational	number of informational pages visited by the site visitor
Informational Duration	time spent on informational pages
BounceRates	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
ProductRelated	number of product related pages visited by the site visitor
ProductRelated Duration	time spent on product related pages
Exit Rates	calculated as for all page views to the page, the percentage that were the last in the session
PageValue	represents the average value for a web page that a user visited before completing an ecommerce transaction
SpecialDay	feature indicates the closeness of the site visiting time to a specific special day (e.g.Mother's Day, Valentine's Day)
Nominal Variables	**Description**
Month	month of visit
Operating System	operating system of visitor
Browser	browser type of visitor
Region	regional location of visitor
TrafficType	type of web traffic to the website
VisitorType	indicates whether the visitor is a new or returning visitor
Weekend	indicates if the visit occurs on a weekend

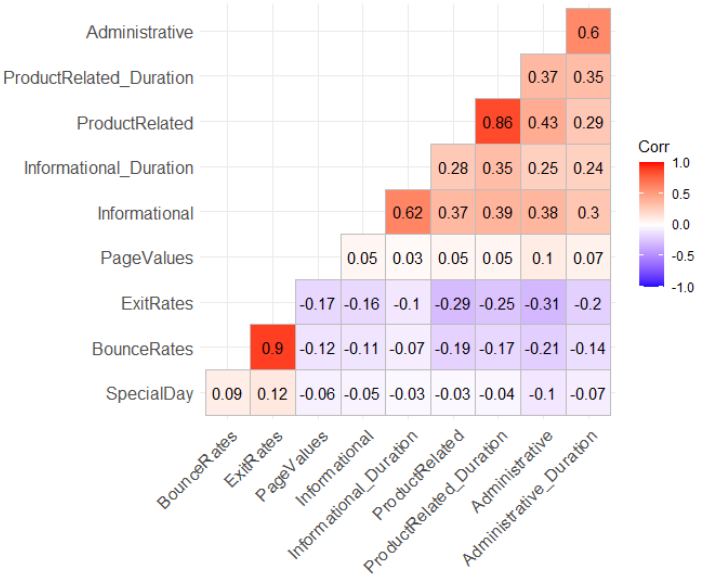
Figure 1.1 Input Variables

Data Quality

There were no missing values in the data set, and 125 duplicate observations were identified and removed. The majority of numeric variables have a large prevalence of meaningful zero values (*Figure 1.2*). For instance; a zero in Administrative, Informational, and Product Related indicate the website visitor did not visit those respective sites. When this occurs, the time spent on the website page (Administrative Duration, Informational Duration, ProductRelated Duration) will likewise be zero. Due to the prevalences of zeros in the data, nearly all numeric distributions are right tailed (*Appendix 2.1*). In general, there is wide variation across all numeric variables, in part due to the prevalence of zero values in each variable (*Appendix 2.2 & 2.3*). All numeric variables contain a large number of outliers, as indicated by an observation with a Z-score greater than 3 (*Figure 1.2*). Therefore, the dataset as a whole contains large amounts of noise.

Variable	Percent "0" value	Number of Outliers
Administrative	46.2%	213
Administrative Duration	47.3%	230
Informational	78.4%	260
Informational Duration	80.3%	229
BounceRates	45.2%	593
ProductRelated	0.3%	236
ProductRelated Duration	5.2%	217
Exit Rates	0.6%	599
PageValue	77.6%	257
SpecialDay	90%	478

Figure 1.2 Prevalence of zeros and outliers



Correlations

High correlations are observed between a webpage visit and visit duration. The number of Product Related web page visits is strongly associated with duration spent on Product Related web pages (*Figure 1.3*). Indeed, all webpage visits and durations are positively correlated with one another, such that visiting and spending time on one type of webpage is associated with visiting and spending time on another type of webpage. Bounce Rates and Exit Rates are strongly associated with each other as well.

Descriptives

The majority (84.4%) of all observed website visits did not result in a purchase, while 15.6% of website visits did result in a purchase *(Figure 1.4*). Most website visits were made by returning website visitors (85.4%), while 13.9% of all visits were new visitors *(Figure 1.5)*.

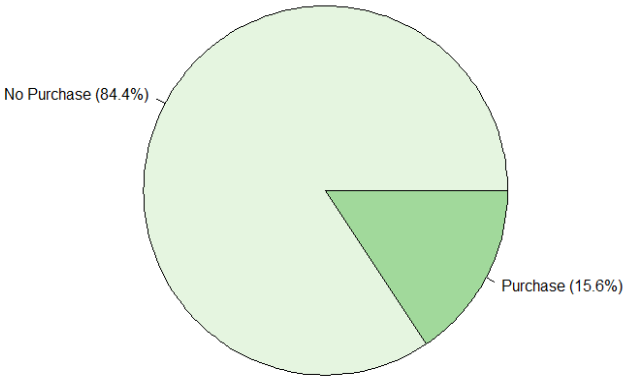


Figure 1.4 Percent of Website Visitors Who Made a Purchase

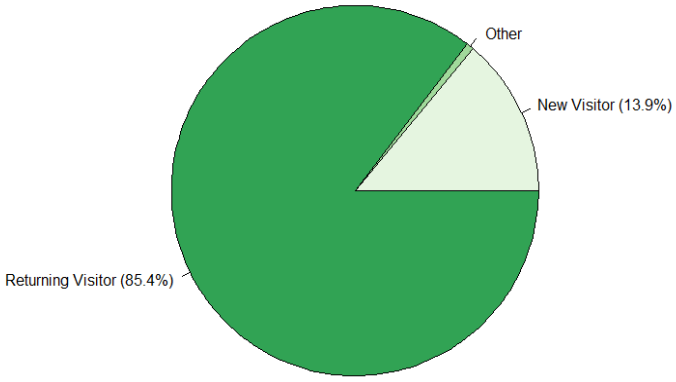


Figure 1.5 Website Visitor Type

Purchases were time variant. The number of website visitors who made a purchase sharply increased in March, May, November, and December. These months also saw the greatest amount of website foot traffic (*Figure 1.6*).

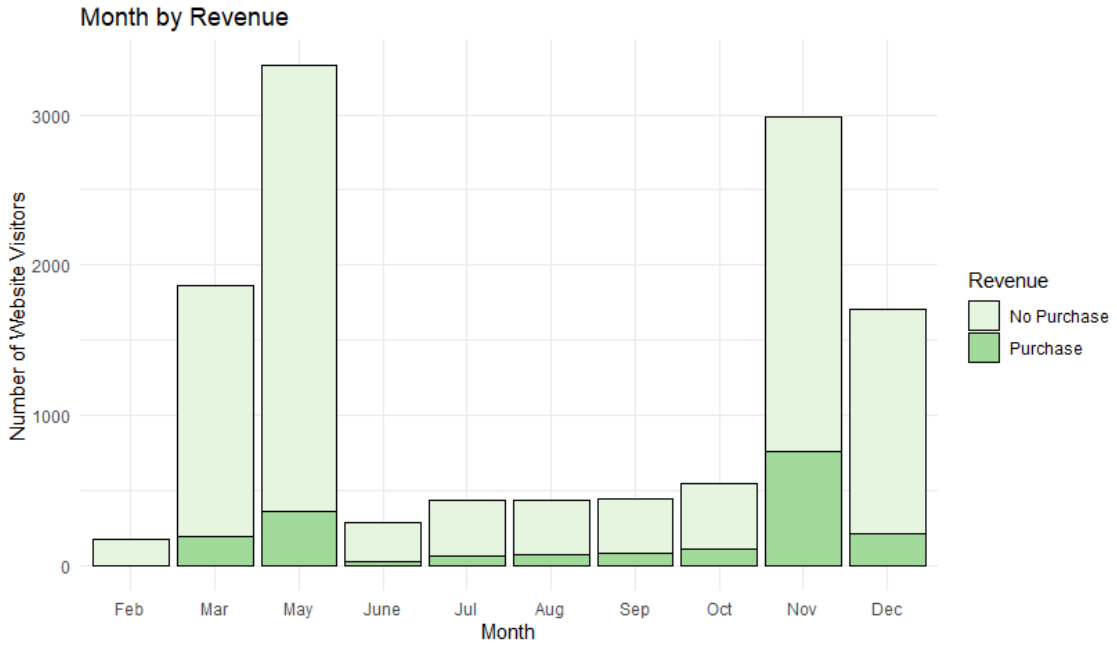


Figure 1.6 Purchases and Website Visits per Month

Comparisons of Group Averages

Those who purchased visited, on average, more company web pages. The largest difference in the number of pages visited was for product related pages (*Figure 1.7*). Those who purchased spent more time, on average, on the company's web pages. The largest difference in time between those who did and did not purchase was observed in time spent on product related pages.

Those who purchased entered the company website from a page with a lower average Bounce Rate. Those who purchased exited the company website from a page with a lower average Exit Rate.Those who purchased had, on average, a higher PageValue than those who did not. Those who purchased visited the site, on average, closer to a specific special day.

Figure 1.7 Comparison of Averages Between Groups

Variable	Purchased	No Purchase
Administrative	3.4	2.1

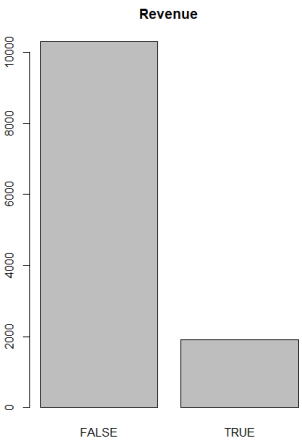
Informational	.8	.5
ProductRelated	48	29
ProductRelated_Duration	1,876 seconds	1,082 seconds
Administrative_Duration	119 seconds	75 seconds
Information_Duration	58 seconds	30 seconds
Bounce Rate	.5%	2.3%
Exit Rate	2%	4.6%
Page Values	27	2
Special Day	0.02	0.070

DECISION TREE ANALYSIS

The Decision Tree method was selected for analysis due to robustness against redundant features, noise, and irrelevant attributes- all of which are present in the current dataset. Within the current data set, high correlations between website visits and website duration variables suggest redundancy. Furthermore, outliers are prevalent throughout the dataset. The decision tree method was chosen due to the prevalence of redundancy and noise in the data set. Two models were run to compare model fit. The first model had no additional transformations, while the second model corrected for class imbalance.. The second model outperformed the first and is described below.

Data Pre-Processing and Transformation

As decision tree models do not require standardization, no normalization transformation was performed. Class imbalance was present in the “Revenue” variable (*Figure 1.8*) and was corrected using case weighting in order to increase specificity.



Eighty-five percent of data was used for model training with the remaining fifteen percent used for model testing and validation. Complexity Parameter (cp) was utilized for hyperparameter tuning. The cp value imposes a penalty to the tree for having too many splits, setting a minimum improvement value that an additional split must add to be included in the tree. In this analysis, the cp associated with the highest accuracy was found to be 0.005. A 10-fold cross validation was utilized. The most important variables identified for predicting Revenue were PageValues, ExitRates,

ProductRelated_Duration, BounceRates, and ProductRelated, with PageValues being the primary variable utilized in the decision tree (*Figure 1.9*).

Figure 1.8 Class Imbalance

Validation and Performance Measures

The model resulted in a balanced goodness of fit between training and testing performance (*Appendix 2.5*). The trained model was able to accurately predict 1,362 customers would not generate revenue, and that 239 customers would generate revenue. The model inaccurately predicted that 182 customers would generate revenue, and that 47 customers would not generate revenue. The model resulted in an accuracy of 87.5%, a moderate Kappa of 60.2%, a Sensitivity of 85%, and a Specificity of 88.2% (*Appendix 2.5*). The model was able to correctly predict who would purchase from the company at an approximate rate of 84%, and who would not purchase from the company at an approximate rate of 88%.

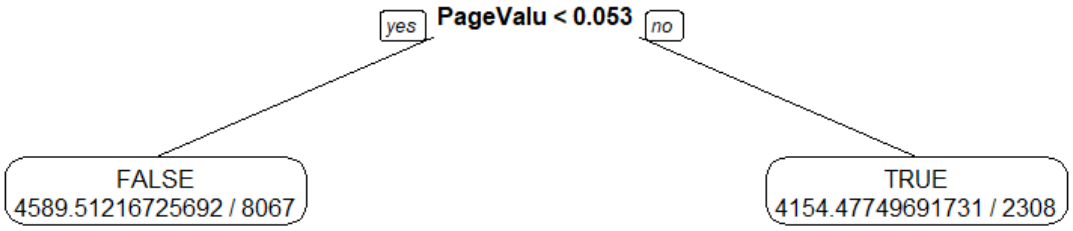


Figure 1.9 Decision Tree Model

NAIVE BAYES ANALYSIS

The Naive Bayes method was selected for analysis due to its ability to incorporate categorical variables, which are present in the dataset. For this method, redundant variables must be removed and numeric variables should approximate normal distribution. Seven variables were identified as potentially redundant variables. Redundancy was handled using variable transformation and/or removal. Three models were run to compare model fit. The first model removed “ProductRelated_Duration” and “ExitRates” due to redundancy. The second model transformed the six website duration and visit variables into three variables representing average visit duration for each web page type. The third model attempted to correct for class imbalance using under and overfitting. The second model outperformed the others, and is described below.

Data Pre-Processing and Transformation

Redundant variables were identified using a correlation matrix. The following seven variables were identified as highly correlated: Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, ExitRates. As the variables website visits and website visit duration were highly correlated, they were combined by dividing the visit duration by the number of visits to achieve an average visit duration. This transformation results in three new variables: “Avg_ProductRelated”, “Avg_Informational”, and “Avg_Administrative”. The variable “ExitRates” was removed due to a high correlation with “BounceRates”. The final dataset included 15 variables. As the Naive Bayes method requires normal distribution, the data was standardized and transformed using YeoJohnson Transformation. Eighty-five percent of data was used for model training with the remaining fifteen percent used for model testing and validation. Laplace Smoothing was applied to prevent model distortion.

Validation and Performance Measures

The model resulted in a balanced goodness of fit between training and testing performance (*Appendix 2.6*). The trained model was able to accurately predict 1,331 customers would not generate revenue, and that 228 customers would generate revenue. The model inaccurately predicted that 213 customers would generate revenue, and that 58 customers would not generate revenue. The Naive Bayes model resulted in an accuracy of 85.2%, a moderate Kappa of 54%, and a Sensitivity of 80% (*Appendix 2.6*). The model was able to correctly predict who would purchase from the company at an approximate rate of 80%. The model was able to correctly predict who would not purchase from the company at an approximate rate of 86%.

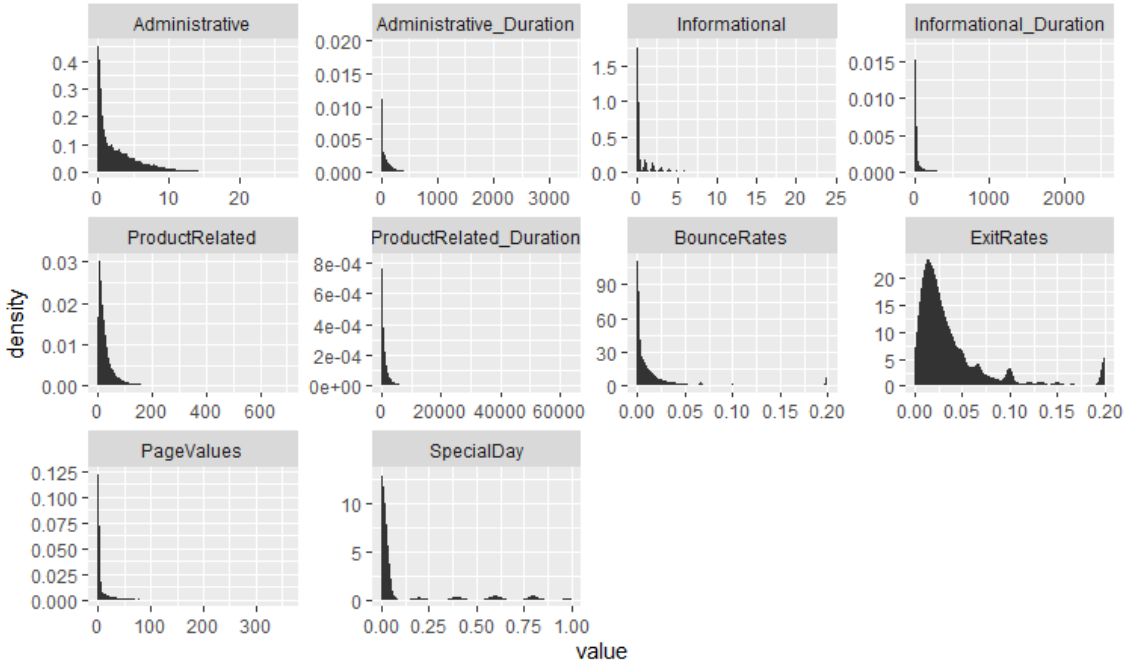
DISCUSSION AND CONCLUSION

In all models run, the decision tree outperformed the Naive Bayes and is therefore recommended for use. The decision tree model will be able to correctly predict who will purchase from the company at an approximate rate of 84%, and who will not purchase from the company at an approximate rate of 88%. The business will be able to use these predictions to target their marketing audience more accurately, and in doing so will avoid spending marketing resources on those unlikely to purchase. The analysis revealed the most important variables for the business to continue collecting data on, the most vital of which was the PageValue variable. The business should prioritize collecting this type of data, as well as the other variables of ExitRates, ProductRelated_Duration, BounceRates, and ProductRelated. All other variables do not need to be collected, thus the company can save resources and time in avoiding unnecessary data collection.

References

Kabacoff, R. (2020, December 1). *Data Visualization With R*. https://rkabacoff.github.io/datavis/Models.html#Corrplot

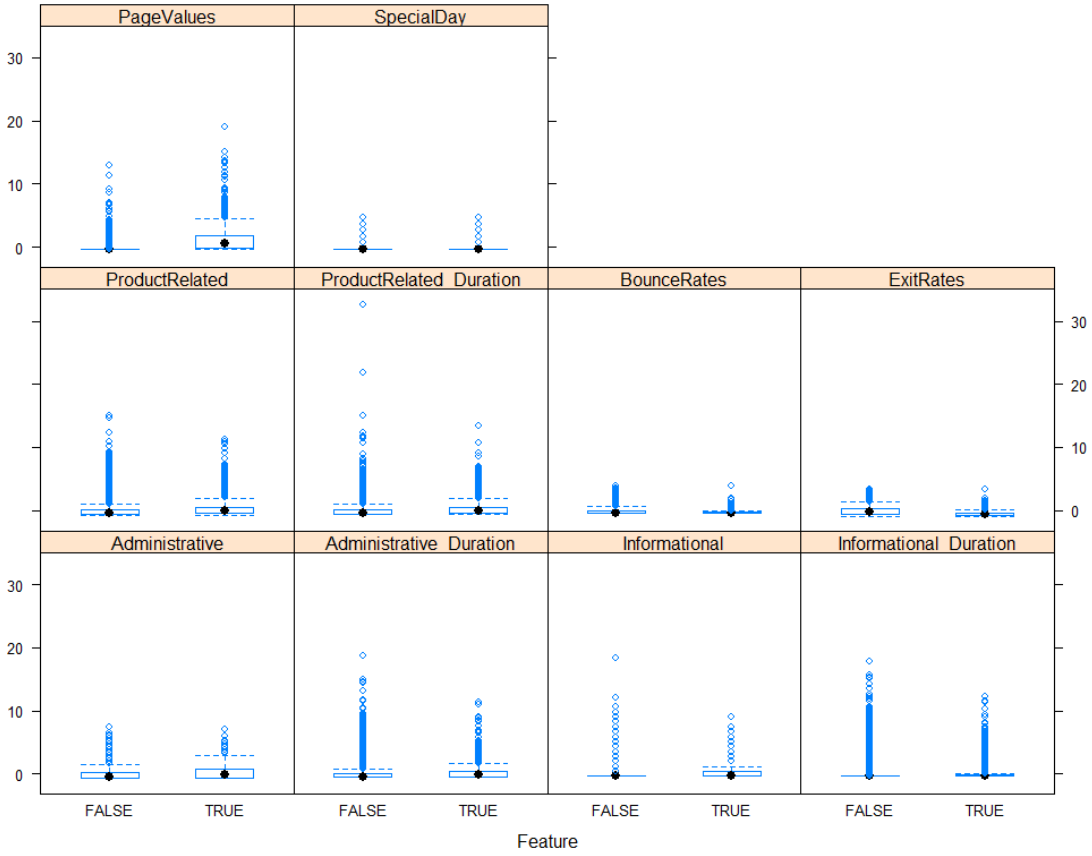
Appendix



Appendix 2.1 Distribution Numeric Variables

Variable	Variance	Standard Deviation
ProductRelated_Duration	3684870	1,920
AdministrativeDuration	31503	177
InformationalDuration	20001	1
ProductRelated	1989	45
PageValues	348	19

Appendix 2.2 Variance and Standard Deviations of Numeric Variables



Appendix 2.3 Box Plots of all Numeric Variables

	Training	Testing
Accuracy	0.87	0.87
Kappa	0.58	0.60
AccuracyLower	0.87	0.86
AccuracyUpper	0.88	0.89
AccuracyNull	0.84	0.84
AccuracyPValue	0.00	0.00
McnemarPValue	0.00	0.00

	Training	Testing
Sensitivity	0.80	0.84
Specificity	0.88	0.88
Pos Pred Value	0.56	0.57
Neg Pred Value	0.96	0.97
Precision	0.56	0.57
Recall	0.80	0.84
F1	0.66	0.68
Prevalence	0.16	0.16
Detection Rate	0.13	0.13
Detection Prevalence	0.22	0.23
Balanced Accuracy	0.84	0.86

Appendix 2.5 Decision Tree Model Performance

	Training2	Testing2
Accuracy	0.85	0.85
Kappa	0.53	0.54
AccuracyLower	0.84	0.83
AccuracyUpper	0.85	0.87
AccuracyNull	0.84	0.84
AccuracyPValue	0.16	0.18
McnemarPValue	0.00	0.00

	Training2	Testing2
Sensitivity	0.80	0.80
Specificity	0.86	0.86
Pos Pred Value	0.51	0.52
Neg Pred Value	0.96	0.96
Precision	0.51	0.52
Recall	0.80	0.80
F1	0.62	0.63
Prevalence	0.16	0.16
Detection Rate	0.12	0.12
Detection Prevalence	0.25	0.24
Balanced Accuracy	0.83	0.83

Appendix 2.6 Naive Bayes Model Performance