**Purchasing Intentions of Online Shoppers**

Helen Phan, Sean Robertson

STAT 431

Dr. David Chu

University of the Fraser Valley

December 3, 2019

**Table of Contents**

[**Table of Contents 2**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.rmc2h6is8q8j)

[**List of Tables 2**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.iisar9mkq3ez)

[**Introduction 3**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.iisar9mkq3ez)

[**Project Goal 3**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.ym4k72tqlirg)

[**Data description 3**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.ym4k72tqlirg)

[**Methodology 7**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.i7gojgyb8ziu)

[Decision Tree 7](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.fzaxwut63l9o)

[Logistic Regression 9](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.uqv2oxoze08l)

[Neural Networks 11](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.3ykn1i7i6crz)

[Random Forest 11](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.dzs0xw7r1aq2)

[**Conclusion 11**](https://docs.google.com/document/d/1szQReTNKpJtG0JjMWCdt41Cgu8v0oDszVFPSehclnWc/edit#heading=h.kphnnssvivyw)

LIST OF TABLES

Table 1: Numerical variables

Table 2: Categorical variables

Table 3: Details of leaves in the optimal tree

Table 4: Odd ratios in logistic regression model

**Introduction**

Online shopping is used by many people around the world. It is a popular method of finding the items you want without needing to drive to a physical location and wait in a checkout line. Companies which offer online shopping services are often interested in what factors lead to purchases; this report explores these factors and presents multiple models designed to predict whether or not a purchase will occur.

**Project goal**

The main goal of this project is to use the program SAS Enterprise Miner and applications of classification algorithms, including decision trees, regression, neural networks, and random forests to create different types of classification models to predict the shopping intent of visitors to the website columbia.com.tr. With this knowledge, we will be able to classify who is likely to make a purchase and who is not among all visitors to the website.

This project also uses the application of cluster analysis to divide all website visitors into a number of groups which have different online shopping behaviors. Those visitor segments will provide businesses with a deeper understanding of online activities of the website visitors, and will help them to take actions accordingly to improve the purchase conversion rates and increase sales.

**Data description**

For this analysis, we obtained the online shoppers purchasing intention dataset from the website: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset. The dataset recorded 12,330 sessions to the website columbia.com.tr in a period of 10 months from February, March, May to December 2016. The dataset consisted of 10 numerical and 8 categorical variables shown in tables 1 and 2, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable name | Variable description | Max | Mean | SD |
| Administrative | Number of administrative pages visited in the session  Administrative pages refers to account-related pages. | 27 | 2.32 | 3.32 |
| Administrative duration | Total amount of time (in seconds) spent on administrative pages | 3398 | 80.82 | 176.70 |
| Informational | Number of pages visited by the visitor about the Website, communication, address information of the shopping site | 24 | 0.5 | 1.26 |
| Informational duration | Total amount of time (in seconds) spent on informational pages | 2549 | 34.47 | 140.64 |
| Product related | Number of pages visited by the visitor about product related pages | 705 | 31.73 | 44.45 |
| Product related duration | Total amount of time (in seconds) spent on product related pages | 63,973 | 1194.75 | 1912.25 |
| Bounce rate | Average bounce rate value of the pages visited by the visitor | 0.2 | 0.02 | 0.04 |
| Exit rate | Average exit rate value of the pages visited by the visitor | 0.2 | 0.04 | 0.05 |
| Page value | Average page value of the pages visited by the visitor | 361 | 5.89 | 18.55 |
| Special day | Closeness of the site visiting time to a specific special day | 1 | 0.06 | 0.19 |

**Table 1** Numerical variables

|  |  |  |
| --- | --- | --- |
| Variable name | Variable description | Number of values |
| Operating system | Type of operating system used by the visitor (e.g., Mac, PC, Linux) | 8 |
| Month | Month value of the visit date | 10 |
| Browser | Web browser used by the visitor (e.g., Internet Explorer, Google Chrome, Firefox) | 13 |
| Region | Geographic region from which the session has been started by the visitor | 9 |
| Traffic type | Traffic source by which the visitor has arrived at the Website (e.g., banner, advertising, SMS, direct) | 20 |
| Visitor type | Visitor type as New Visitor, Returning Visitor, and Other | 3 |
| Weekend | Boolean value indicating whether the date of the visit is weekend | 2 |
| Revenue | Class label indicating whether the visit has been finalized with a transaction | 2 |

**Table 2** Categorical variables

The data about administrative, administrative duration, informational, informational duration, product related, product related duration shown in Table 1 were derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, for example moving from one page to another. The data about bounce rate, exit rate, page value were measured by the web analytics tool, Google Analytics.

*Bounce rate*

According to Google Analytics (2019), the bounce rate for a specific page represented the percentage of all sessions that started with that page in which visitors viewed only that page and then left the website without moving to another page or doing anything else during that session.

*Exit rate*

Google Analytics (2019) defined the exit rate as the percentage of pageviews for a specific page that were the last view of a session.

*Page value*

Google Analytics (2019) measured the metric page value by the average value for a specific page that a user visited before completing an e-commerce transaction or an online purchase. The equation for calculating page value was shown below.

where

ecommerce revenue specified the total revenue or grand total associated with the transaction, and goal value was a monetary amount that was assigned to the goal conversion. A goal represents a completed activity, called a conversion, that contributes to the success of your business. Examples of goals include making a purchase, completing a game level, or submitting a contact information form.

A unique pageview represented the number of individual users who have loaded a given page per session. Each user is counted only once per session, no matter how many pages are opened by the same user.

*Special day*

In a study of this dataset, Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018) considered the effect of special days (e.g. Mother’s Day, Valentine’s Day) on the purchasing decision of online shoppers by computing the variable special day. An example of calculating the value of special day was explained with the scenario of Valentine’s day. As follows, this value took a nonzero value between February 2 and February 12, zero before and after this date unless it was close to another special day, and its maximum value of 1 on February 8.

*Revenue*

The variable revenue had two class labels: false if the visitor did not make a purchase, and true if the user finalized with a transaction. It was used as the response variable or target variable in classification models.

There were none missing information throughout the dataset. To validate the accuracy of the prediction results, the dataset was divided into two partitions. One with 50% of the dataset was used to fit our classification models, and the other was used to validate how accurately our models predicted the likelihood of making a purchase.

**Methodology**

*Decision Tree*

Decision tree was our first approach to predict the online purchasing intention. Using the function Decision Tree in SAS Enterprise Miner, we obtained the first tree on the training data. Observing the Subtree Assessment Plot, the misclassification rate on the training sample decreased while the misclassification rate on the validation sample increased when number of leaves of the tree rose from 8 to 25. This pointed out an evidence of model overfitting when the tree became more complex. To solve the problem of model overfitting, we chose the eight-leaf tree as our optimal tree where we reached the lowest misclassification rate on the validation data.

After the SAS algorithms about decision tree run, five variables: page values, bounce rates, number of administrative page views (Administrative), month, the amount of time viewing product related pages (Product Related Duration) were selected as classifying conditions to split the optimal tree into 41 nodes and 8 leaves. Table 4 showed the details of 8 leaves with classification rules to obtain each specific leaf, the percentage of sessions in which the visitors finalized with a transaction (Y = True, Y was the target variable, revenue), the percentage of sessions in which the users did not make a purchase (Y = False). In table 4, the purchasing intention was our prediction of the shopping intent of visitors based on the validation data. On the validation data of a specific leaf, if the percentage of sessions in which the visitors finalized with a transaction is higher than the percentage of sessions in which the users did not make a purchase, the value of purchasing intention was no and vice versa.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Leaf** | **Classification Rule** | **% of train data** | | **% of validation data** | | **Purchasing Intention** |
| **Y = False** | **Y = True** | **Y = False** | **Y = True** |
| 1 | Page value < 0.9805 | 96.03 | 3.97 | 95.96 | 4.04 | No |
| 2 | Page value >= 0.9805  Bounce Rates < 0.0001 Administrative <3.5 | 14.55 | 85.45 | 19.72 | 80.28 | Yes |
| 3 | Page value >= 0.9805  Bounce Rates < 0.0001  in Mar, Nov | 43.49 | 56.51 | 38.59 | 61.41 | Yes |
| 4 | Page value >= 0.9805 or <21.9121  Bounce Rates < 0.0001  Administrative >=3.5 | 59.66 | 40.34 | 52.63 | 47.37 | No |
| 5 | Page value >= 21.9121 Bounce Rates < 0.0001 Administrative >=3.5 | 26.04 | 73.96 | 28.07 | 71.93 | Yes |
| 6 | Page value >= 0.9805  Bounce Rates < 0.0001  in May, Jun, Oct, Aug, Jul, Sep  Product Related Duration >= 505.4 | 72.14 | 27.86 | 68.06 | 31.94 | No |
| 7 | Page value >= 0.9805 or <12.174  Bounce Rates < 0.0001  in May, Jun, Oct, Aug, Jul, Sep  Product Related Duration < 505.4 | 71.43 | 28.57 | 64.29 | 35.71 | No |
| 8 | Page value >= 12.174  Bounce Rates < 0.0001  in May, Jun, Oct, Aug, Jul, Sep  Product Related Duration < 505.4 | 11.54 | 88.46 | 28.57 | 71.43 | Yes |

**Table 3** Details of leaves in the optimal tree

The misclassification error of the optimal tree model was 9.312% on the train data and 9.828% on the validation data. The optimal tree appeared to be a strong classifier with a low error rate.

*Logistic Regression*

Among 10 numerical variables shown in Table 1; five following variables including administrative duration, information duration, product related, product related duration, and page value distributed with large variabililities compared to others. Therefore, using the transformation function in SAS Enterprise Miner we did the log transformation for those variables to reduce the variability of data. Additionally, using the function of creating recorded columns in SAS Enterprise Guide, we recoded some of categorical variables to a number of indicator variables based on the corresponding number of categories of the variable. We transformed month to 10 indicator variables, traffic type to 20 dummy variables, visitor type to 3 indicator variables.

Using the regression module in SAS Enterprise Miner, we fitted a logistic regression model with the stepwise method in order to remove insignificant predictors out of the model. The model was produced:

(Nov) -

Where

was the estimaed percentage of visitors who made a purchase

ExitRates was the variable exit rate

Aug was the indicator variable August derived from the categorical variable month

Sept was the indicator variable September derived from the categorical variable month

Oct was the indicator variable October derived from the categorical variable month

Nov was the indicator variable November derived from the categorical variable month

Log\_administrative duration was the value of adminitrative duration after log transformation

Log\_pagevalues was the value of page values after log transformation

New Visitor was the indicator variable new visitor derived from the categorical variable visitor type

Traffic type\_8 was the indicator variable “8” derived from the categorical variable traffic type

Traffic type\_10 was the indicator variable “10” derived from the categorical variable traffic type

Traffic type\_13 was the indicator variable “13” derived from the categorical variable traffic type

Traffic type\_20 was the indicator variable “20” derived from the categorical variable traffic type

From the SAS output, the Chi-square statistic resulting the p-value lower than the 5% significance level suggested the validity of the model. All predictors were useful and significant at 5% significance level. Table 5 showed the odd ratios and the percentage of change in the response variable.

|  |  |  |
| --- | --- | --- |
| Predictor | Odd ratio | Percentage of change |
| ExitRates | < 0.001 | -100% |
| Indicator\_Aug | 0.508 | -49% |
| Indicator\_Nov | 0.283 | -72% |
| Indicator\_Oct | 0.625 | -38% |
| Indicator\_Sept | 0.432 | -57% |
| LOG\_Administrative\_Duration | 0.934 | -7% |
| LOG\_PageValues | 3.114 | 211% |
| New\_Visitor | 0.596 | -40% |
| TrafficType\_10 | 0.546 | -45% |
| TrafficType\_13 | 2.05 | 105% |
| TrafficType\_20 | 0.46 | -54% |
| TrafficType\_8 | 0.434 | -57% |

**Table 4** Odd ratios in logistic regression model

The log of page value had the biggest positive effects on the response variable. If the log of page value increased one unit, it would rise the change of purchasing a thing by 211%. The exit rate had the biggest negative effects in the way that if the exit rate increased one unit, the change of purchasing would decrease by 100%.

*Neural Networks*

The neural-network contains an estimated 187 weights or parameters, and uses 20 iterations to optimize the Average Squared Error and Misclassification Rate. This is a large model, so we also used an auto-neural network.

The auto-neural network has an estimated 94 weights/parameters, and the model has 3 hidden units for both Misclassification Rate and Average Squared Error. The training process used 8 iterations for a one-hidden-unit network, 12 iterations for two units, 8 iterations for three units, and 12 iterations for four units.

*Random Forest*

The results from the high-performance random forest (after running a stepwise regression) indicate that the 5 most important variables were, from most important to least important: Page Values, Month, Product Related Duration, Exit Rates, and Traffic Type.

**Conclusion**

Using a decision tree yielded the lowest misclassification rate for determining whether or not a purchase is made. If the page value was low (< .9805) then it was extremely unlikely that a purchase would be made (about 95% accuracy in validation data). If the page value was higher than .9805, had a low bounce rate (< .0001) and administrative page count was less than 4, a purchase was very likely (about 80% accuracy in validation).

In general, page value was the strongest single predictor for our models. This is logical, as it directly measures the revenue or value generated by a page. Bounce rate and the number of administrative pages visited were the next two most useful predictors. These are also sensible; if many people arrive at a page by mistake and leaves immediately (bouncing) it is much less likely that the page will end in a transaction. Similarly, if a person spends a lot of time on administrative pages (changing their password or email address, for example) they were probably not visiting the website to do shopping, or they may have grown frustrated from needing to reset some of their information and thus refuse to do the shopping they had planned.

**References**

Google Analytics (2019). *Bounce rate*. Retrieved from: <https://support.google.com/analytics/answer/1009409?hl=en>

Google Analytics (2019). *Exit rate*. Retrieved from: <https://support.google.com/analytics/answer/2525491?hl=en>

Google Analytics (2019). *Page Value*. Retrieved from: <https://support.google.com/analytics/answer/2695658?hl=en>

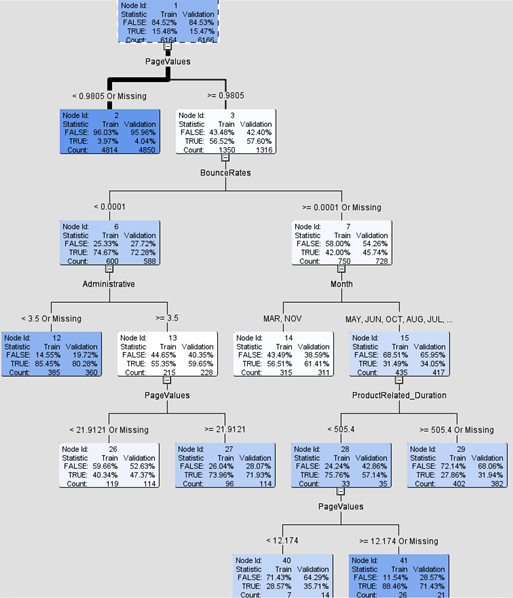
Google Analytics (2019). *Goal Value*. Retrieved from: <https://support.google.com/analytics/answer/1012040?hl=en>

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). *Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks.* Retrieved from:

<https://link.springer.com/article/10.1007%2Fs00521-018-3523-0>

**APPENDIX**

***A: The optimal tree***



***B: The Subtree Assessment Plot***

A screenshot of text

Description automatically generated

***C: The output of Logistic Regression***

***A close up of a receipt

Description automatically generatedA screenshot of a cell phone

Description automatically generated***

***D: The iteration plot of Neural Network model***

***A screenshot of text

Description automatically generated***

***A screenshot of a cell phone

Description automatically generatedE: The iteration plot of Auto Neural Network model***

***F: The weights of Auto Neural Network model***

***A screenshot of a cell phone

Description automatically generated***

***G: The iteration plot of Random Forest Model***

***A screenshot of text

Description automatically generated***