JUNE 1

Authored by:

Divya Ulaganathan
(s3759465@student.rmit.edu.au)

Bodiyabaduge Dewsri Lalithi Perera
(s3762890@student.rmit.edu.au)

Table of Contents

Abstract

The aim of this report was to understand the shopper's behaviour on e-commerce website to predict the likelihood of purchase and shopping cart abandonment. To help with the investigation, the "Online Shoppers Purchasing Intention "dataset containing various user's activities on an e-commerce site and other likely influencing factors for a purchase has been used. K- Nearest Neighbours and Decision tree classifiers were used to distinguish the users based on their session activities.

Overall, the results indicate that it is possible to predict a user's purchase intention based on his session activities with approximately 89% accuracy.

1 Introduction

Over the past decade, online shopping industry gained considerable popularity among people. Even though, the growth of e-commerce website users increased hugely, the revenue achieved by online retail stores was considerably low in comparison to physical retail stores. Thus, many e-commerce companies have begun to study the behavior of users in shopping websites and identify the potential causes for a shopping cart abandonment. Several studies and researches are being conducted to incorporate the real time purchase experience in a virtual environment to influence the user's purchase decision.

This report will discuss the behavior of users on an ecommerce website and aims at the possibility of predicting a user's purchase intention based on their activities.

2 Methodology

First according to our aim, a satisfiable dataset was searched through the UCI repository. Then the data set was retrieved and used to conduct our study based on Online customer purchasing intentions. The retrieved data set contained a suitable class attribute 'Revenue', and hence the data modelling task was treated to be a classification task. First the data set will be cleaned for the exploration and modelling tasks.

During the data exploring task, most important attributes were individually graphed to observe their distribution and to identify special patterns. Then a pairwise analysis was conducted to extract the most related features for the data modelling phase and to observe the relationship among these features.

Two approaches were used in data modelling phase, K- Nearest Neighbors classifier and Decision tree classifier since the data set contained both quantitative and categorical attributes. The two classifiers tuned with appropriate argument values and then applied to 3 different training and testing splits of the cleaned data set. For KNN classifier there were 24 models and for Decision tree classifier there were 9 models. Finally, the accuracy rates, precision rates, etc. of the fitted classification models were compared and identified the model which scored the highest accuracy rate as the best model.

3 Data set description

The ‘Online Shoppers Purchasing Intention’ dataset used in this study is from UCI repository. It contains feature vectors belonging to 12,330 sessions from various users on an e-commerce website. In order to avoid any tendency towards a special day, specific campaign, period or specific user preference, the data has been collected over a period of one year. Out of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping, the results were noted on “Revenue” attribute for each user. The dataset contains 10 numerical values and 8 categorical values. The amount of time spent on the website and actions performed were recorded on attributes like "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration". The average pages visited and exited by the users are measured using Google Analytics and noted down on features like "Bounce Rate", "Exit Rate" and "Page Value". Data on visitor types and the period on which the site has been visited are recorded on features like “Visitor Type”, "Special Day”, “Month” and “Weekend”. The user’s “Network type”, “Browser”, “Traffic type” and “Region” are also recorded to determine the most influencing factor for a user’s purchase.

3.1 Data Preparation

Initially the provided dataset ‘online_shoppers_intention.xlsx’ was loaded onto osi_df data frame. The dataset’s information like data type, shape (number of rows and columns), mean, standard deviation, minimum value and maximum values were explored.

The initial data set had 7 float data types, 7 integer data types, 2 object data types and 2 Boolean data types constituting to 18 data columns and 12,330 entries.

The data frame osi_df was copied to fake_osi_df to perform all data cleaning and manipulation activities in order to preserve the original data on osi_df and to resolve any mis-handled data errors in future.

The data set was tested for missing values, outliers, spelling errors and whitespaces during initial exploration steps. Around 125 duplicate records were observed and removed from fake_osi_df as the redundant session details might affect the data model. Out of 18 attributes 8 mis-represented categorical attributes were assigned appropriate data types.

4 Data Exploration

The dataset has been explored with the aim of finding the features that affects the purchase intention of a user. As mentioned earlier, the given dataset has the greatest number of users who didn’t make a purchase. Hence for comparison between two attributes, cross tabulation combined with percentage values were used instead of absolute numbers.

Spread of numerical data

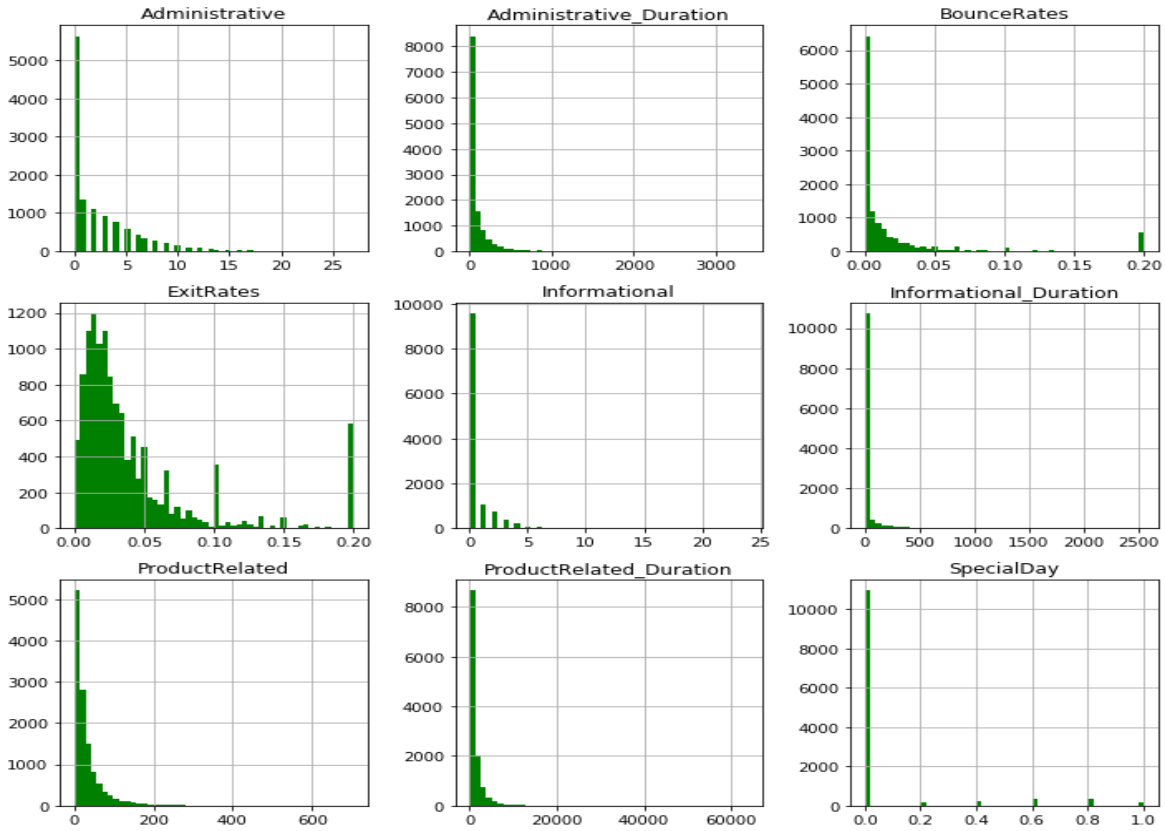

Figure 1 : Frequency distribution of numerical attributes

The insights from the data exploration process are listed below.

- Most users accessed the website during the months of May, November, March and December. Among these November months had the highest amount of sales while compared to the rest.
- Around 86 % of visitors to the site were returning visitors. The amount of shopping cart abandonment was high compared to sales on returning visitors. Alternatively, new visitors had high number of sales compared to shopping cart abandonment.
- The common region among most users were region number 1, the most used browser was browser type 2 and the highly observed traffic type was type 2.
- Furthermore, the sale rates were observed to be equal during weekdays and weekends.

- The number of special days was high on February and May months. But the revenue wasn't affected on special days.
- The users who spent more time on viewing product related page mostly made a purchase.
- The time spent on viewing product related pages are high compared to administrative and information pages.
- Bounce rates and exit rates were correlated. As were administrative pages to administrative duration, informational pages to informational duration and product related pages to product related duration.
- Few outliers that were observed during data exploration weren't removed, to preserve the data and avoid overfitting.

Correlation Matrix

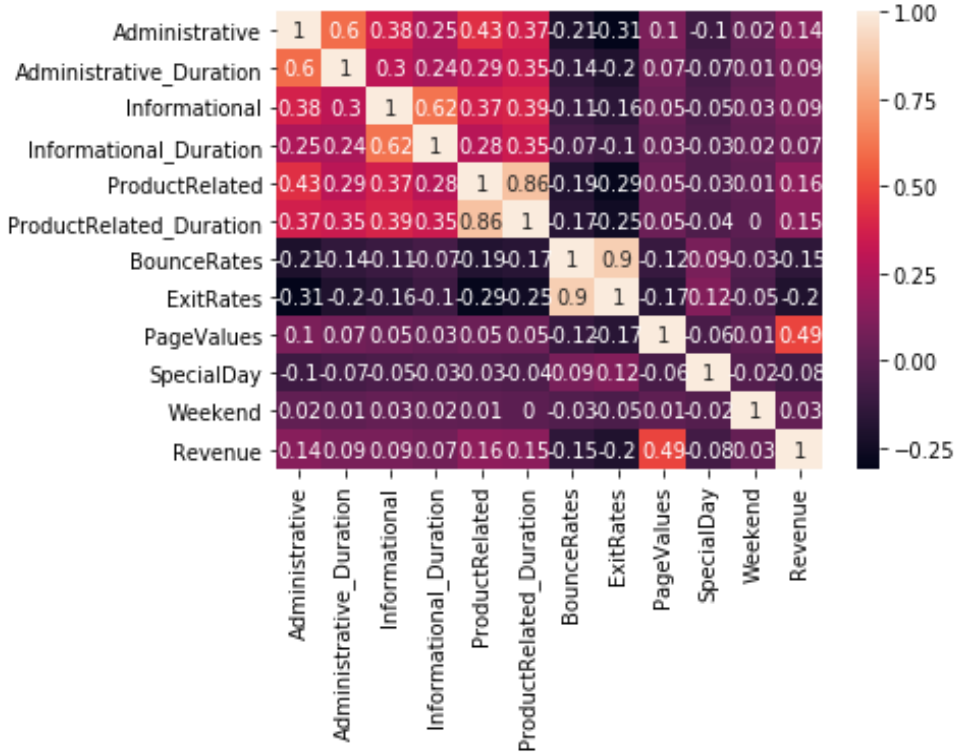

Figure 2 : Heatmap based on correlation between variables

The above heatmap depicts the correlation between revenue and other values. Among these Page Value has the highest amount of correlation to revenue closely followed by

Administrative, Product related view and duration. Surprisingly, attributes like Special day, exit rates and bounce rates have negative correlation towards revenue.

5 Data Modelling

Data modeling was treated as a classification task and was implemented using K nearest neighbors method and using Decision tree classifier. The attribute "Revenue" was considered as the class attribute which has two classes. The two algorithms were executed for three different training & testing splitting ratios as follows,

- 50% for training and 50% for testing;
- 60% for training and 40% for testing;
- 80% for training and 20% for testing;

Models were trained with different argument values of the two algorithms as follows,

KNN: number of neighbors(k=2,3), weights ('uniform', 'distance') and distance measures (Manhattan->p=1, Euclidean-> p=2). Number of Models = 24.

Decision tree: Maximum Depth = 4, 5, 6. Number of models =9.

Accuracy rates of each model was recorded, and the performance of instances were compared at the end to find the best fitted model.

6 Results and Discussions

6.1 K - Nearest Neighbor Classifier

KNN classifier was chosen to model the data set since it is non-parametric, which needs no assumption about the data distribution. In addition to that, only the most relevant features were taken into consideration to enhance the accuracy and to minimize overfitting. Feature set was labelled as (X) and the selected features were No.of product related pages visited, Time spent in product related pages, No.of administrative pages visited, Time spent in administrative pages, No.of informational pages visited, Time spent in informational pages, Bounce rate, Exit rate, Page values and Special day. 'Revenue' attribute was selected as labels (y).
KNN algorithm was imported from sklearn.neighbors. Following table depicts the accuracy rates recorded for each model.

KNN Parameters	training-0.8, testing-0.2	training-0.5, testing-0.5	Training-0.6, testing-0.4
N=2,'distance',p=1	83.65	82.84	82.69
N=2,'distance',p=2	83.37	82.91	82.38

N=2,'uniform',p=1	86.15	86.19	86.36
N=2,'uniform',p=2	86.93	86.42	86.56
N=3,'distance',p=1	85.33	85.25	85.13
N=3,'distance',p=2	85.21	84.79	85.13
N=3,'uniform',p=1	85.33	85.48	85.64
N=3,'uniform',p=2	85.58	85.09	85.54

According to figure 3 the highest accuracy rate which is 86.93% has been recorded for training 80%, testing 20% split and for the parameters (n=2, 'uniform', p=2). When observing the accuracy rates among three splitting sets, in most cases training 80%, testing 20% split shows high accuracy. This result may be due to overfitting of the model. However, the differences between the three splitters are not significantly high. Minimum accuracy was recorded for training 60%, testing 40% split and for the parameters (n=2, 'distance', p=2) which is 82.38%.
From the KNN model it can be concluded that the chosen feature set can predict whether a person will make a purchase with 86.93% accuracy.

6.2 Decision Tree Classifier

Due to the presence of categorical attributes Decision tree classifier was used since it gives high preference to categorical data. Similar to the KNN modeling the three splitting criteria was used and maximum depth of the decision tree was changed to tune the decision tree algorithm.

When selecting the feature set, the variables ‘Month’ and ‘Visitor type’ was removed. ‘Revenue’ attribute was selected as the label set.

Accuracy Rates of Decision tree classifiers

Maximum Depth	Training = 80%, Testing=20%	Training = 60%, Testing=40%	Training = 50%, Testing=50%
4	89.82	89.31	88.92
5	89.19	89.1	88.76
6	89.2	89.02	88.71

Maximum

Minimum

F1-scores of Decision tree classifiers (False)

Maximum Depth	Training = 80%, Testing=20%	Training = 60%, Testing=40%	Training = 50%, Testing=50%
4	0.94	0.94	0.94
5	0.94	0.94	0.96
6	0.94	0.94	0.93

F1-scores of Decision tree classifiers (True)

Maximum Depth	Training = 80%, Testing=20%	Training = 60%, Testing=40%	Training = 50%, Testing=50%
4	0.62	0.6	0.5
5	0.58	0.6	0.47
6	0.61	0.49	0.59

Precision rates of Decision tree classifiers (False)

Maximum Depth	Training = 80%, Testing=20%	Training = 60%, Testing=40%	Training = 50%, Testing=50%
4	92	91	91
5	92	91	91

Precision rates of Decision tree classifiers (True)

Maximum Depth	Training = 80%, Testing=20%	Training = 60%, Testing=40%	Training = 50%, Testing=50%
4	70	72	71
5	69	71	71

6	93	91	91		6	66	72	69

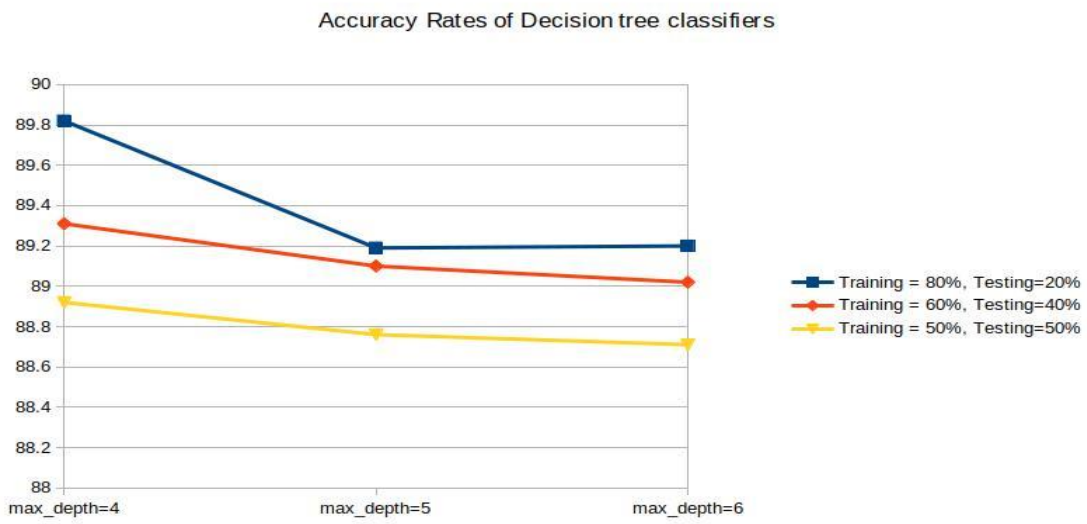

Figure 4 : Decision tree accuracy rates

Figure 4 accuracy rate comparison chart depicts the highest accuracy score of 89.2% for training 80%, testing 20% split and for max_depth=4. F1 scores and precision rates show a significant difference for the two classes 'True' and 'False'. It can be assumed that the low percentage of 'True' instances (15.63%) may have caused for this result. However, the highest accuracy rate is neither supported nor rejected by the F1 scores and the precision rates.

7 Conclusions

When it comes to data modelling, decision tree classifier wins with its highest of 89.2% accuracy over KNN classifier which obtained 86.93% as the highest accuracy. Both of the classifiers gained their highest accuracies for testing 80% and training 20% splitting criteria. How ever the other two splitting conditions also gained almost similar accuracy rates which were differed only by 1 to 2 percent.

Overall results show that a person's purchasing intention can be described or predicted by his/her behavior on e-commerce website with approximately 89% accuracy rate.

8 References

1. Carmona CJ, Ramı´rez-Gallego S, Torres F, Bernal E, del Jesu´s MJ, Garcı´a S (2012) Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. Expert Syst Appl 39(12):11243–11249

2. Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). https://doi.org/10.1007/s00521-018-3523-0

3. Scott Robinson, Decision Trees in Python with Scikit-Learn https://stackabuse.com/decision-trees-in-python-with-scikit-learn/

4. https://pythonprogramming.net/k-nearest-neighbors-application-machine-learning-tutorial/
5. Data source: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset