Mark Your Customers

BUAN 6356.002 - Group 6

# Contents

1. Executive Summary
2. Background
3. Data Description
4. Descriptive Statistics
5. Classification Tree
6. Logistic Regression
7. Conclusion
8. Recommendations
9. References

# Executive Summary

In this academic project, we worked on a business case of an eCommerce website by applying business analytics models to mine the data retrieved from the website. The data shows that of all the customers visiting the website, only a few customers actually make purchases on the website. Our aim here is to find factors that could help the company in increasing customer purchasing intention.

# Background

Ecommerce businesses are popular these days. By taking their businesses online, companies have realized that they could reach their customers beyond geographic limits. Even customers find it convenient to look up various items on eCommerce websites and make purchases anytime without even stepping a foot out of their doors.

However, delivering the best customer experience on their eCommerce websites is still a challenge many companies face. Ideally, the customer experience should be designed in such a way that a customer faces zero obstacles in making purchases. Solutions to improve customer experience in brick and mortar shopping stores are many. Moreover, in case a customer faces any difficulty in making a purchase in a physical shopping store, he or she can always seek help from any staff present at the store and complete the purchase. However, this situation would work very differently if it were an ecommerce store. Customer experience in an ecommerce store would ensure readily accessible company information, clearly described products, various buying methods, free delivery options, and even easy return options. Also, a number of ecommerce businesses these days are setting up chat boxes on their websites to respond to customers looking for assistance, emulating the customer experience of buying at a local store. Improving customer experience on ecommerce websites is critical because it plays a major role in determining whether the customer would make purchases on the website at all.

In this project, we’re dealing with such an ecommerce business case. The dataset that we adopted has record of customer activity in a real business situation. For privacy purposes, however, the actual names and other details have not be revealed by the database repository, which hosts this dataset. The web activity, nevertheless, represents actual customer experiences on the website as it records the activity of customers on company information pages, account management pages, and product related pages. As a group of business analysts, our role is to analyse this data and return valuable insights for the ecommerce website to improve its customer experience.

# Data Mining Objective

Our dataset contains information about the browsing activity of the many visitors who visit a particular e-commerce website from which the data is sourced. Among all the visitors, some visitors end up buying a few products and add to the revenue of the e-commerce business, while others do browse through several pages on the website but do not make any transactions during their recorded activity. Using this dataset in our project, we aim to create a prediction model which could identify the users who would complete one or many transactions during their browsing activity on the website.

For an e-commerce business, such a prediction model would help its team to focus its marketing efforts on the portion of users who could become its potential buyers. It would be better for the business if it caters its content mostly to those users who would really make purchases on the website. In the project, our team would create various prediction models from the dataset. We will then make performance evaluations to determine which model would best suit our e-commerce business case.

Finally using the insights gained from our data mining work, our team would be able to offer recommendations to the business at the end of the project.

# Description of the data

The dataset contains data of the web activity of an e-commerce website. The recorded data over a period of one complete year, so it captures an entire cycle of special days in a year of business. An ecommerce business usually experiences greater activity before, after and during the special days in the year. It would be vital to include all such variations in user activities while creating a prediction model for the business.

Each record in the dataset is a consolidated session of a unique user who visited the website during the one-year period. A user may have visited and purchased items on the website multiple times, but the activities for a certain user is consolidated in a single record.

Any user visiting the e-commerce website is required to register his or her account. The user information is recorded in the account management webpages. The number of pages a user browsers in the account management section is collected in the variable ‘Administrative’. Similarly, the information about the e-commerce business, such as its physical address, communication, and registered name of the company is stored in the informational pages, while the products sold on the website are listed on the various products pages of the website. Number of informational pages and product pages visited by a user is recorded in the variables Informational and ProductRelated. The duration of sessions on these pages for each user are recorded in the variable Informational\_Duration and ProductRelated\_Duration.

The other variables such as PageValues, ExitRates, and BounceRates are rendered by Google Analytics for website performance. These metrics are defined for each webpage and sessions of all users on the website. However, when variables are recorded for each user in the dataset, the record contains averages of the concerned metric of the pages visited by a particular user.

The variable SpecialDay records the activity of the users. The variable SpecialDay records the activity of the users before, after and during special days when users are more likely to make purchases. The values in records are derived from the webpage performance near those special days and consolidated for each user.

Lastly, the category variables OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend and Revenue stores the information specific to the activity of each user. To create a prediction model, our target variable would be Revenue as it records whether the user finally made a purchase or not on the website.

Calendar

Description automatically generated with low confidence

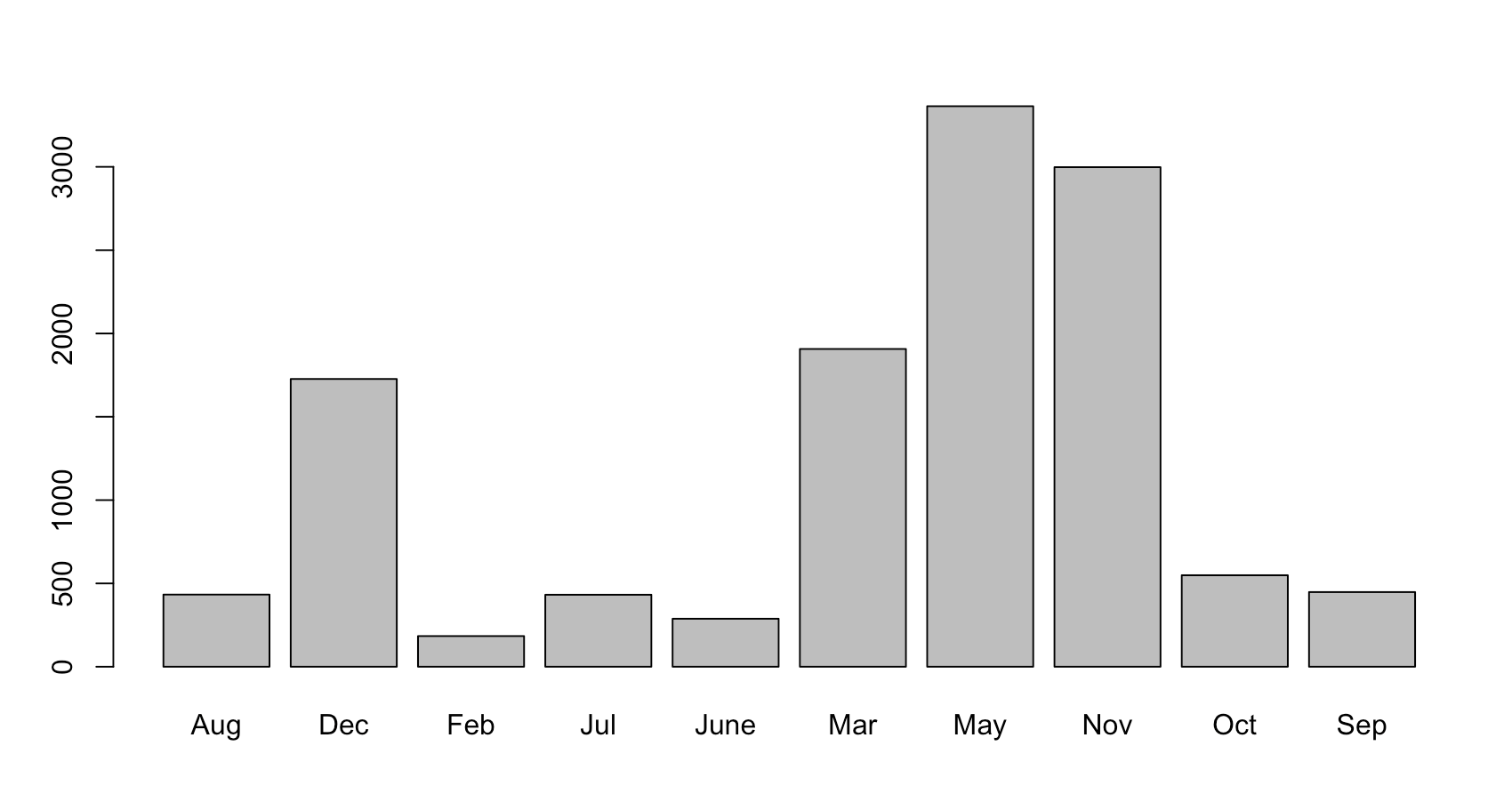
|  |  |
| --- | --- |
| Variable | Description |
| **Numeric variables** | |
| Administrative | No. of pages visited by the user about account management |
| Administrative\_Duration | Seconds spent by the user on account management related pages |
| Informational | No. of informational pages visited by the user |
| Informational\_Duration | Seconds spent by the user on informational pages |
| ProductRelated | No. of pages visited by the user on product related pages |
| ProductRelated\_Duration | Seconds spent by the user on product related pages |
| BounceRates | Average bounce rate value of the pages visited by the user |
| ExitRates | Average exit rate value of the pages visited by the user |
| PageValues | Average page value of the pages visited by the user |
| SpecialDay | Closeness of the site visiting time to a special day |
| **Category Variables** | |
| OperatingSystems | Operating system of the user |
| Browser | Browser of the user |
| Region | Geographic region from which the session has been started by the user |
| TrafficType | Traffic source (e.g., banner, SMS, direct) |
| VisitorType | Visitor type as “New Visitor,” “Returning Visitor,” and “Other” |
| Weekend | Whether the date of the visit is weekend |
| Month | Month value of the visit date |
| Revenue | Whether the visit has been finalized with a transaction |

# Descriptive Statistics

Chart, histogram

Description automatically generatedLooking up central tendencies and distributions of the data on various dimensions was the first step in our analysis. For instance, the distribution of the amount of time customers spent on account management pages (Administrative\_Duration) revealed that most of their time was spent the first time they visited one of these pages. Well, this is intuitive because customers would have to input all their information the first time they create their account on the website. Following is the distribution.

We also created frequency tables and cross-frequency tables to delineate variables that stand out among other variables in the data. These variable would reveal some common patterns in customer activities. For instance, we generated frequency table for the variable Month. The month of May and November experienced greater activity than the other months. Furthermore, from the cross table of Month and the target variable Revenue, we again observed that the number of purchases in these two months was the greatest among those in other months.



A picture containing text

Description automatically generated

We were also interested to see what type of visitors visit the website on weekend the most. Since weekends experienced more customer activity than other days of the week, we wanted to see whether customer looking up our website are first time visitors or are returning after they visited during the weekdays. This revealed that generally new visitors visit on weekends more than returning visitors.

A picture containing text

Description automatically generated

Next we also plotted correlations among the variables and conducted principal component analysis (PCA). The objective of this analysis whether some variables whether were high correlated. Indeed, the graph showed that in general the number of pages visited by the customers were moderately or highly positively correlated with the amount of time these visitors spent on those webpages. This is, however, self-explainable. If a customer browses more pages in a category, he or she would of course spend more time on that category. The PCA on the other hand showed us that variances are spread over almost all the variables in the dataset.

Following is the summary of the information we collected from descriptive statistical analysis.

Chart, scatter chart

Description automatically generated

# A screenshot of a computer Description automatically generated with medium confidence

# Insights from descriptive statistics

* Of the 12330 people who visited the website 10442 ended up not buying any product and the remaining 1908 did purchase from the website.
* There is high correlation between Administrative and Administrative\_duration, Product related and Product related duration, Informational and Informational\_duration, Exit rates and Bounce rates.
* There are 125 duplicate values in the dataset.
* PCA shows that the first 11 variables contain 96% of the information from all the variables of the dataset.
* Decision tree is constructed and a classification of the data is obtained with 90.7% accuracy rate.
* Administrative\_Duration, Informational\_Duration, ProductRelated\_Duration are highly skewed variable.
* May and November are potential months that sessions are more likely to be finalized with transaction.
* Probability of new visitor visiting on weekends = 0.28. Probability of old visitor visiting on weekends = 0.22.
* Probability of new visitor purchasing = 0.24. Probability of old visitor purchasing = 0.13.

# Predictive statistics (Supervised Machine Learning)

We went further in our analysis to observe which factor or combination of factors could determine whether a customer would make purchases at the end of his or her customer journey on the website. We applied two models readily applicable in such a business case – classification tree or decision tree model, and logistic regression.

# Classification Tree Model

*Data Preparation*

Balancing of the data:

* The dataset contained a significantly larger number of customers who did not make purchases on the website than those who did make purchases. We had to handle this case carefully before building our model.
* To balance the dataset, we removed many of the instances in which customers did not actually make purchases. The remaining dataset then had 50% of customers who did not make a purchase and the other 50% who did.
* Also, we have removed several columns like bounce rate, exit rate, page values, OS, browser etc. as they don’t make much difference in predicting the model and also these variables are generated by google analytics and are not clearly explained how they were derived by Google.

Validation of balanced data:

* The data was split into 50% each. Training data consists of 50% of the dataset and validation data consist of other 50% dataset.
* The accuracy of the training dataset is 72.7%.

1. Accuracy of training dataset is 72.7% with a possible accuracy rate of customer agree to buy is 66.08% and possible accuracy rate of customer disagree to buy is 79.34%.

* The accuracy of the validation dataset is 69.9%.

1. Accuracy of training dataset is 69.9% with a possible accuracy rate of customer agree to buy is 66.13% and possible accuracy rate of customer disagree to buy is 73.4%.
2. The area of the ROC curve or can be referred as the AUC value is 76.7%

* The accuracy looks different from the results in the training data. So, we can suspect that there is overfitting in the result of these prediction models. Changing the comparison of the amount of training data and validation data might be one of solutions to resolve these kinds of problems. But in this project, in handling this matter, no further analysis was carried out and we have assumed that overfitting hasn’t occurred.

Validation of whole dataset:

* The accuracy of the whole dataset is 60% which is less than training dataset and validation dataset accuracy.
  1. Accuracy of whole dataset is 60% with a possible accuracy that the model correctly predicts that a customer will make a purchase 72% of the time and that the model correctly predicts that a customer will not make a purchase 57% of the time.
  2. The area of the ROC curve or can be referred as the AUC value is 67.7%

Interpreting the decision tree:

* The root node we are going to consider is product\_related\_duration as it has the lower p value and visitor\_type, product\_related, administrative\_duration are the child nodes.
* Product\_related child node is further split into multiple nodes as it has the high chi-square statistic and lower p value. This we are going to find using entropy and information gain. Attribute with highest information gain is selected for a split.

Performance review of the decision tree:

* Our predictive model suggests that product\_related\_duration lesser than 272 lead to a TRUE 61%. On top of this, product\_related below 69 improves our TRUE to 56% and as an added point of Administrative\_duration below 15 results in a TRUE 62%.

# Logistic regression

We use general linear model with family = binomial to get a logistic regression and turn off scientific notations.

Preparation – categorical into numeric:

* We need to convert few variables, for example from categoric to numeric or change their format to work with different algorithms.
* We converted the revenue variable to integer by removing the digits after decimal.
* We split the original 3800 values dataset into 2. We used one part as the training dataset and another for validation dataset.

Modelling on original:

* The accuracy of the whole dataset is 60% which is less than training dataset and validation dataset accuracy.
* Accuracy of whole dataset is 60% with a possible accuracy that the model correctly predicts that a customer will make a purchase 63% of the time and that the model correctly predicts that a customer will not make a purchase 61% of the time.
* The area of the ROC curve or can be referred as the AUC value is 67.7%
* Here, we cannot be dependent upon only accuracy of the model as it can be misleading. To assist how well a logistic regression model fits the dataset, we can use sensitivity and specificity of the model to find the true positive and negatives rate.

Modelling on balanced data(validation):

* The accuracy for the training model is 63.3% and the p value is approximately 0 whereas the accuracy for validation training dataset is 51.47%. However, the sensitivity of the training and validation data are high.
* The area under the curve(auc) is 67% for the original data whereas the area under the curve for validation dataset is around 78.73%.
* The accuracy for the original dataset using classification tree is 57.6% and using logistic regression is 63.3%. This gives a suspicion that the models classification results are not perfect enough. However, we can compare training and testing data to overcome this issue.
* The auc for original data and validation data using was approximately same for both classification tree and logistic regression. So, we do not need to worry about the overfitting for this model.

The highest probability of gaining revenue is from new visitors, visitors with low bounce rates, promotions in high month sales(or Feb and May), and visitors who spend more time on product related pages.

Different model predictions are better for different variables. However, logistic regression is better than classification tree for predicting the online shopper’s intention due to better accuracy and area under curve value.

# Conclusions

*Which is the better model between the two?*

* + In our predictive analytical process, we tried two models – decision tree model and logistic regression model. Further, we did consider exploratory data mining as well. We applied a clustering model to the dataset. However, we did not gain much from those results. Applying the clustering model would require some data related to the personal information of the customers. In the given dataset, we did not have access to attributes such as zipcode, age, or gender of the customers. The results that we achieved only showed us how some clusters of customers were more active than the other clusters on the ecommerce website.
  + Nevertheless, we found a significant amount of knowledge from the two models we applied to the dataset. We can offer recommendations to the business owners of this ecommerce websites with the knowledge we gain from our analysis. But before going further, we would like to suggest which model of the two they should consider to deploy to production.
  + Both of these models are a good match in such a use case in which we have only two choices to classify the customers. For these supervised learning algorithms to work we had labelled data and a data set that closely represents the business of the eCommerce website. After building our model, we compared the actual values of the labelled data to the predicted values from the models. We generated confusion matrices for this purpose. Although we had to balance the data before building the model, we were interested to validate the performances of our models on the original data.

By evaluating the performances of the two models, we found that the decision tree model fares better in performance than the logistic regression model. In terms of accuracy, both these models are comparable. The accuracy of prediction in the decision tree model is 59.55% while that in the logistic regression model is 61.83%. However, accuracy can be a misleading metric. So, we went further to take a closer look at the implication these models can have on the business. The idea here is that we needed to minimise the chances that when a customer is actually intending to make a purchase, the model should predict that intention accurately as far as possible. Otherwise, the business would miss a substantial proportion of potential customers who could have made purchases. The sensitivity score on the results show us exactly0 this probability. Although the logistic regression model is better in its accuracy than the decision tree model is, it miss many potential customers who actually purchase. The sensitivity of the logistic regression model is only 63.36%. On the other hand, the decision tree model has a sensitivity score of 72.48%. There is a considerable difference between their sensitivity score. Hence, we can say that the decision tree model does a better job of predicting actual purchasing intent than the logistic regression does. The company has a better chance of identifying purchasing intent among its customers by using our the decision tree model.

Table

Description automatically generated

Figure 1. Confusion matrix of decision tree model

Text

Description automatically generated

Figure 2. Confusion Matrix of Logistic Regression Model

# Recommendations for the company

After we built the decision tree, we used the tree to predict the customer journey on the website. By doing so, we realised that the customers who would have browsed many product related pages and then landed on account management pages did not actually make any purchases. Account management is probably related to signing up on the website. Here we suspect that the website requires customers to sign up before make a purchase. A customer would have to fill in many personal details before he or she could proceed. So, this seems to have become a pain-point for many customers. Therefore, we would suggest to the company that it makes the signing up process easier than currently. Perhaps, it can put the sign up process when the customer first visits the website, or use social media sign up APIs on the website to resolve this issue.

Secondly, weekends offers a great opportunity for the company to make good business. Customers, specially those visiting for the first-time, usually make purchases if they visit the website on weekends. This shows strong purchasing intent on weekends. Therefore, it would be good for the company if it targets its visiting customers especially on weekends by giving them newsletters on emails or attractive discounts.

In case of discounts and clearance offers, the best months would be May and November when customers make a great number of purchases. Besides, the company should plan way ahead for the special days all around the year. As the days approach closer to a particular special day, the number of purchases starts to drop. But a good number of these purchases are made about a week in advance. So, the company should target its customers during those days before the special days arrive.

By following these recommendations, we hope that the company would be able to improve customer experience on its ecommerce website. As the customer experience resolves the pain-points and expectations of customers as we explained, the company would realise greater customer intent and reap greater business from the customers visiting its ecommerce website.

# References

* <https://archive-beta.ics.uci.edu/ml/datasets/online+shoppers+purchasing+intention+dataset>.
* Data Mining for Business Analytics: Concepts, Techniques, and Applications in R, by Galit Shmueli, Peter Bruce, Inbal Yahav, Nitin Patel, and Kenneth Lichtendahl. Wiley, ISBN-10: 1118879368, ISBN-13: 978-1118879368.
* Software we used is [https://www.rstudio.com](https://www.rstudio.com/).