CETM46 Assignment 2:

Literature Review Report

Prepared by: Fung Yue (Lewis)

Date Prepared: July 1, 2020

Document Information

|  |  |  |  |
| --- | --- | --- | --- |
| Assignment 1: | Assignment 2 | | |
| Prepared By: | Fung Yue (Lewis) | Document Version No: | 1.0 |
| Title: |  | Document Version Date: | 1-July-2020 |
| Reviewed By: |  | Review Date: |  |

Distribution List

| From | Date | Phone/Email |
| --- | --- | --- |
| Fung Yue (Lewis) | July 1, 2020 | Bh63qb@student.sunderland.ac.uk |

Version History

| Ver. No. | Ver. Date | Revised By | Description |
| --- | --- | --- | --- |
| 1.3 | 1-July-2020 | Fung Yue(Lewis) | CETM 46 |

Table of Contents

1 Abstract 4

2 Introduction 4

3 Mission of the assignment 4

4 Design the product and what we found in dataset? 5

5 Manipulation of dataset 10

6 Create Web Application 11

7 Testing before WebApp Prototype launch 11

8 How to Run It? 11

9 Conclusion and future work 12

10 References 13

# Abstract

We use the UCI dataset "Online Shoppers Purchasing Intention" regarding customer visits to an online store and their decision to purchase a product(s). The objective of the analysis is to use clustering algorithms and classification algorithms to create **predictive** **models** around the intentions of those online shoppers so that we will be able to build an interactive data product accordingly.

# Introduction

The public dataset used is based on the UCI dataset "Online Shoppers Purchasing Intention" (detailed overview at: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#)

The data has to do with the behaviour of the shoppers as they visit a website. The "session" activity is measured in 17 dimensions and we note the revenue generated by the visitor (FALSE OR TRUE)

# Mission of the assignment

In the assignment, we will design a digital product that use and visualise the Online Shoppers Purchasing Intention to support the decision making of increasing webstore traffic and revenue. But how we develop visual interface from the dataset which allow user easy to use it precisely? We will quickly come across a situation in which a difficult issue has to be resolved. Let’s call this digital product as Shopper Intention Analysis (SIA), the considering points are: -

1. Data Scientist and product designer need a deep insight into the data structure, user interface need to be easy (UX design) for user to access it.
2. Select the best Machine Learning algorithms for the product
3. SIA are aimed and experts in online store fields, the SIA thus intervenes deeply in such business domain and complexity.

To be able to estimate the upcoming online shopper intention of sales takes experience. The impact of the dataset consists 10 numerical and 8 categorical attributes will be apparent. This analyses aim to provide the shopper 's actions with an accurate, feasible recommendation algorithm and then provide an interactive data product for the online store to review and act of increasing their sales revenue. The goal value is the buying decision conditional 'FALSE' or 'TRUE' with respect to the web store customers. The plan is to use clustering techniques and perhaps classification techniques to make predictions about the intentions of the shoppers.

Firstly, the “Page Value” – the average value of a web page accessed by a customer before completing an e-commerce transaction. It tells us which pages on the Website have the highest value. Ecommerce product page is typically higher than a resource page. Secondly, “Special Day”- feature means that visitors to the site are more likely to conclude their sessions with a specific date (e.g., Mother’s day, St. Valentine's Day). Through analysing Ecommerce variables such as the time spent between the order date and delivery date, the value of this feature is calculated, this value takes a non-zero value from Feb. 2 to Feb. 12 and zero during, during, and after the day of Valentine, except on, and its maximum, on Feb. 8, is similar to another special day. The dataset includes also the OS, browser, location, traffic, visitor type or the Boolean type as return or new visitor.

Those type of relationship and questions are hard to answer without data analysis. In this assignment, we will use appropriate data analysis and machine learning (ML) algorithms for the relationship between all variables which can be utilized in making good predictions of upcoming sales and select the best ML with a website enabled application (Webapp) for users to adjust the goal of predictions on revenue outcome.

# Design the product and what we found in dataset?

Before jumping to analyse the dataset, we should have below points to consider: -

1. Conceptualizing SIA - We must identify the function that meets a need from the store owner or management. This introductory step should take place before data acquisition. It requires conceptualizing the SIA.
2. Refinement – Data refining can be achieved with tools, ML and algorithmic processing of the dataset can categorise, correlate, personalise, profile and search data quickly to create meaningful models that have significate value for store owner.
3. Hosting – Data storage, retrieval and processing are taking place in cloud services provider (eg. Google and Amazon) rather than on own site premises. As we are creating data product prototype, we consider a free hosting cloud services provider, eg. Rshiny.io
4. Distribution – SIA should be able to distributed via mobile devices, if Web Application content formats support, it also needs to be designed for supporting both PC and mobile browser.
5. Presentation – The user interface mattered, the easier SIA was to use, the more valuable it was.
6. Testing – Function and features must be tested in a matter of hours or multivariate online testing approaches.

**What we found on the UCI dataset "Online Shoppers Purchasing Intention"?**

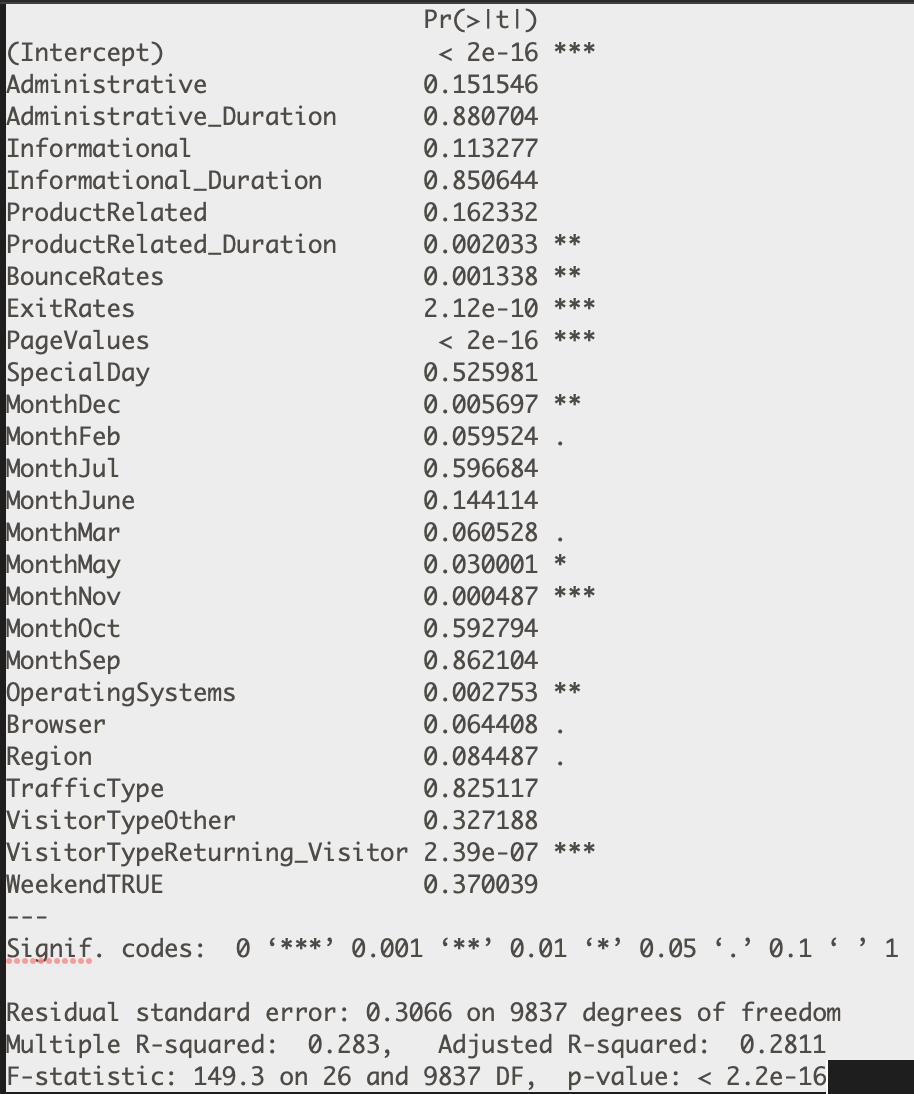
Firstly, we need to check the csv file of the dataset which include: -

1. Rows and Columns
2. Data structure with heading, string, etc…
3. Distribution of target variable
4. Missing values

There are a total consists 10 numerical and 8 categorical attributes, “Revenue” seems a dependent attribute which provide a binary: ‘Ture’, ‘False’).

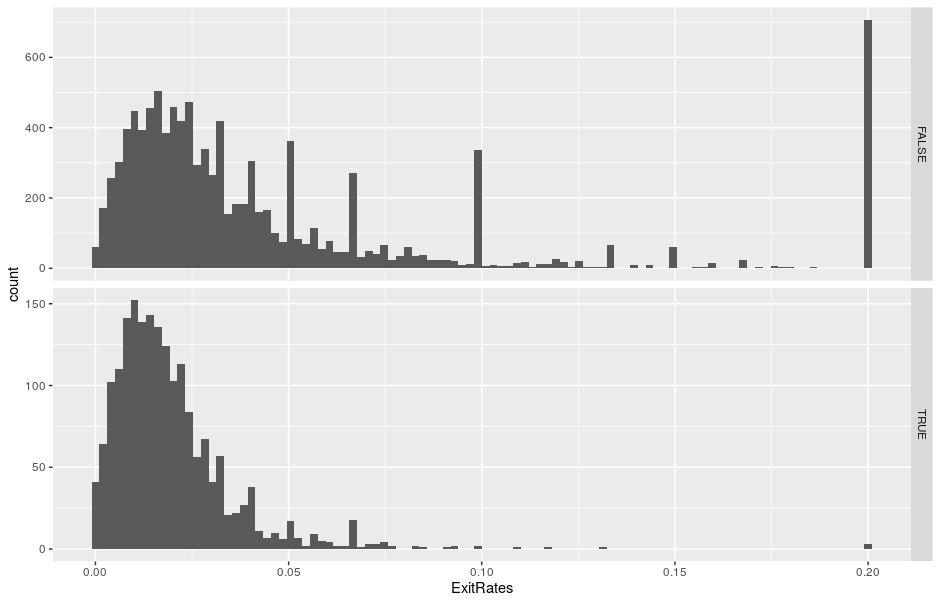
When starting to select different algorithms but the dataset is distributed the number of missing or duplicated values, we need to first identify those values. We will use Rstudio setup and need to leverage different package from R, eg, tidyr, dplyr, ggplots, tm, forcats, caTools, data.table, rpart, corrplot, cluster, gmodels, ggmosaic etc..

From the quick interpretation summary of the dataset, we found the significate codes at ‘\*\*\*’ in ExitRates, PageValues, MonthNov and VisitorTypeReturning\_Visitor.

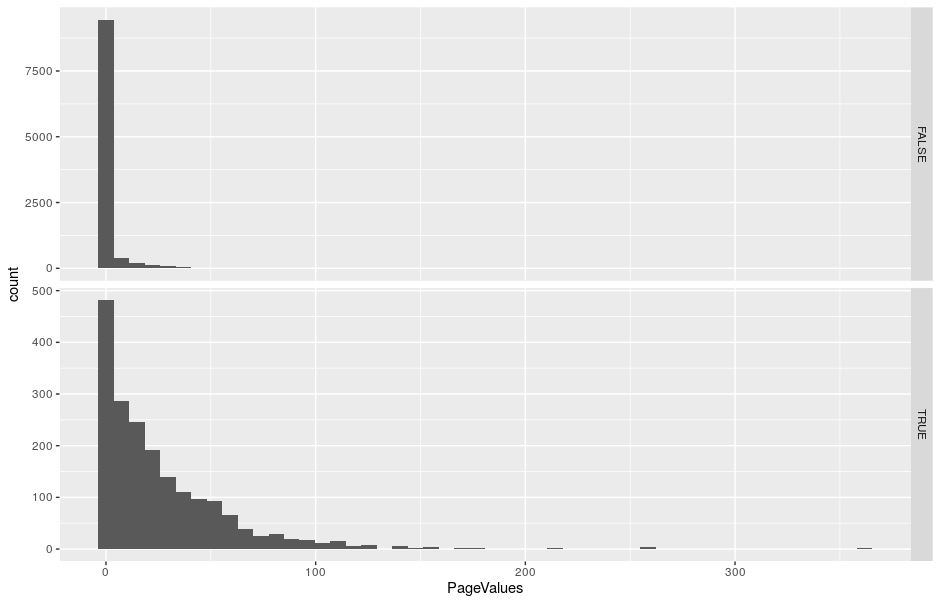


Let’s visualize and analyse the dataset to see if there are any visual difference between the shoppers and non-shoppers.

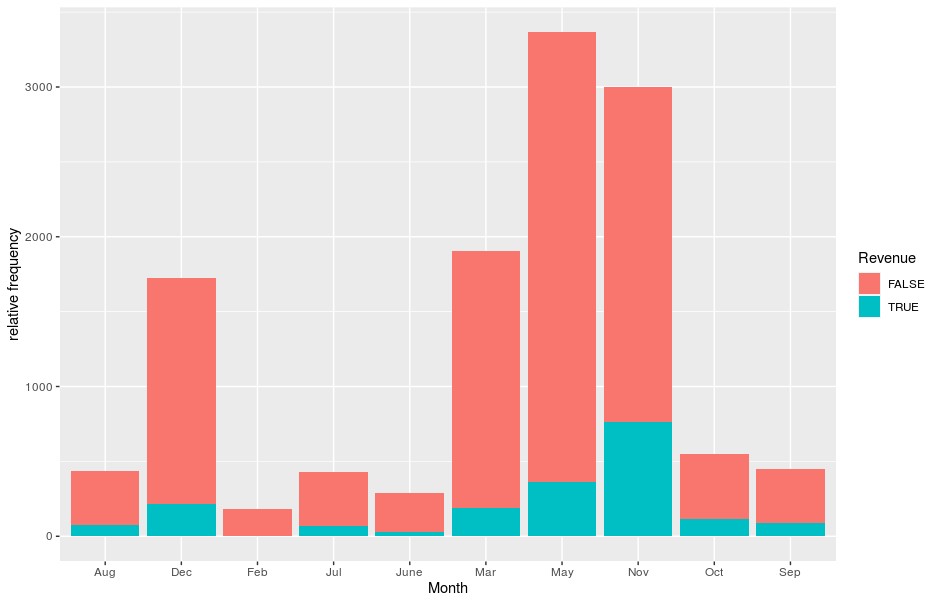
ExitRates: -

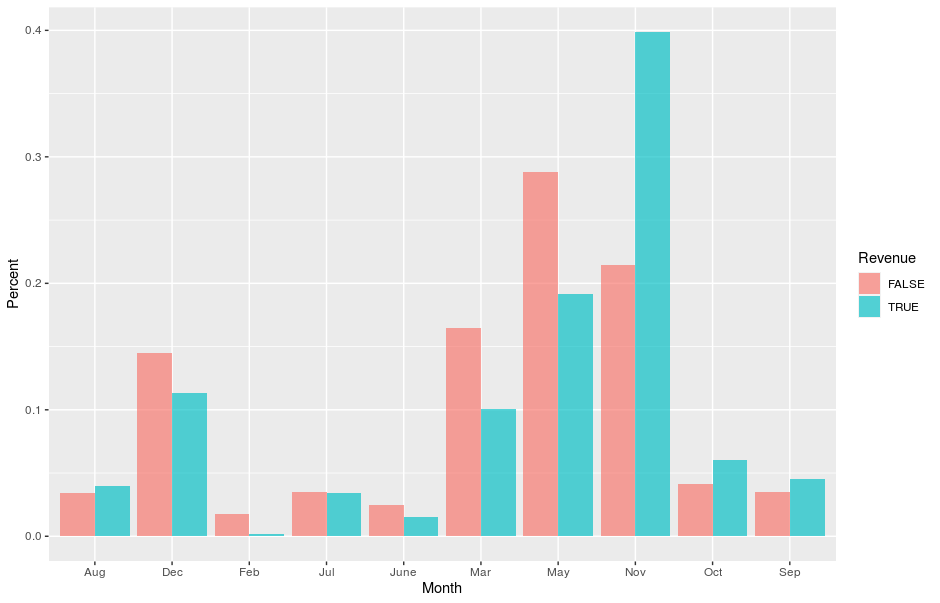


PageValues: -



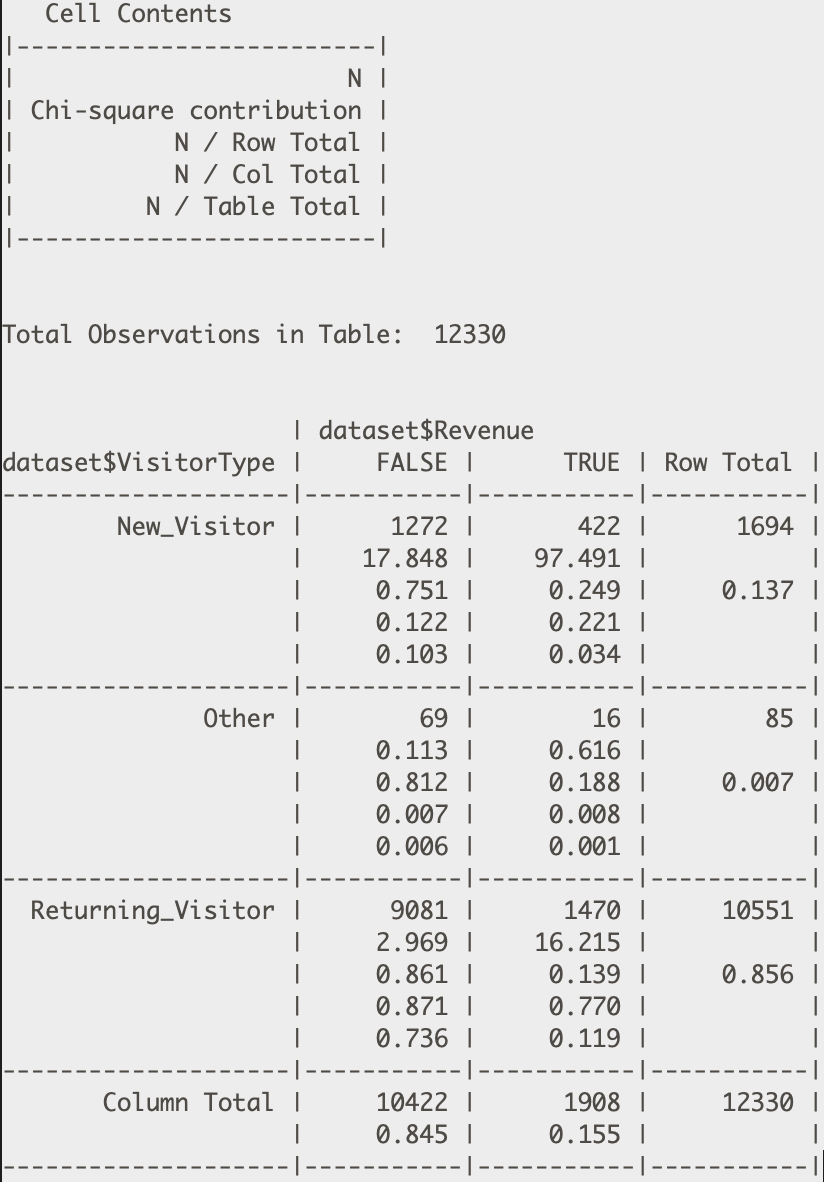
Why Nov make a difference? Let’s check the relationship with Revenue: -

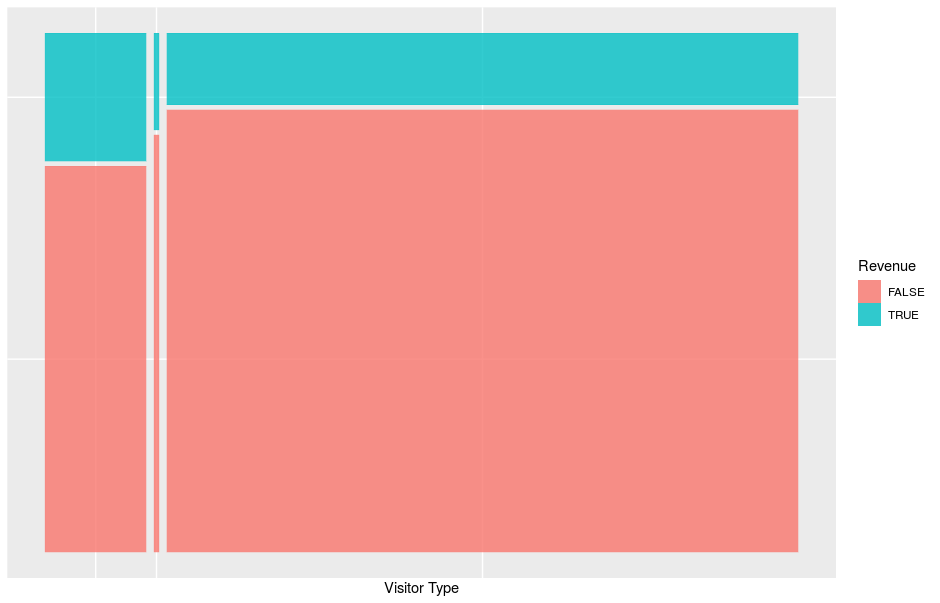




The shopping rates in Sept, Oct and Nov that are high shopping season which correspond in North America and May traffic is very high but only visiting the web store.

Visitor Type, does the visitor also a critical factor?





Here are some interesting findings. Most of our visitors come back to the nation (85.6%) compared with just 13.7%. However, new visitors are more likely to buy a product (24.9 percent) than returning visitors generate revenue only 13.9 percent.

In summarizing our numerical considerations, we see very little variation and slight variability in our group. Based on this, we will seek to group the decision tree algorithm and the cluster by the k-means algorithm. Such methods include "median" imputation, "most common" imputation, or predictive imputation. Prediction Imputation is also another big technique that is suggested.

In addition, this strategy involves classification, a strategy that is even better than average imputation, to take the feature column that contains the missing values and to set this feature column as a dependent variable, while setting the other columns as independent variables, and to divide the data set into a training set and test set where the training set contains all the observations where the feat occurs. Then we run a classification process and decision tree algorithm to predict the missing values in the test set and finally substitute the missing values with the results predictions.

# Manipulation of dataset

The aim of our initiatives is to end up with a model that can be used within an online store to solve a specific problem, which can be shared as quickly and efficiently as possible. The web app (SIA) is accessible almost everywhere and by several users. In addition, changes can be made to the standard web application deployment process. We will introduce R and R Shiny, a web development framework and application server for the R language. Shiny can make data analysis into interactive web apps. In the end, we will end up building a web app SIA.

**Create a classifier**

The classifier that we are going to make here has a peculiarity. Typically, we use a training dataset to train our model and use the resulting model to predict any new input data. In the case of a classifier, this means that the model will agree on the most suitable terms for constructing the classifier from those in the training set. We will split the data into training and test set (80/20) for classification. Then, dataset may visualize suggests that there are no clear distribution patterns among the variables and hence, clustering might a good sorting algorithm for our use. It will also look at the dataset and try to find groupings. We will ask the model to group our data into 2 groups to be able to predict ‘Ture’ and ‘False’ Revenue.

# Create Web Application

Our R Shiny Web application will be composed of two sections. One defined the user interface, and second one defines how the server deals with user interface and executes any required R code.

The planning of user interface will contain four different areas: -

1. Title area which show our application title
2. Control Panel area located at the left-hand side of the page central area. There is where the user will indicate what needs to be done and how.
3. An overview area to the right of the panel. In the form of the distribution map, the user will get an overall impression of the results.
4. A result detail area will be in the bottom of the page. There the user will see the classification.

# Testing before WebApp Prototype launch

Firstly, we should started with the **Unit Tests** which consist with app.R, html files and png links. Secondly, we verify different models and libraries are installed on premise of Rstudio, it was an **Integration Tests** used by our application and libraries work well together. Then we started to have **Functional Tests** which focus on assignment requirement about visualisation and machine learning results. As this is prototype, we assumed the 3 types of testing practice are good enough. For the best practice of data product hand over to user, we should go through more test stages, such as End-to-end test, User acceptance test, Performance test, Smoke test and Exploratory test.

# How to Run It?

Both the ui. R and the server. R code can be found in the App.R. The other file we need is training data. We will do two things to use the web app we created. The app run locally using RStudio and share it with all at Shinyapps.io. The IDE will detect that we are creating a Shiny app when the App. r files are present and detail instructions are reference in Readme.pdf

# Conclusion and future work

This assignment introduces a conceptual analysis of the online store in the sense of modelling the data product. In our dataset of variables, we have found that due to the limitations of our datasets, the decision tree was better able to predict the buying intention of the shoppers than the Clustering models.

The Decision Tree had a higher F-score of 0.94, given the small number of observations and variables, the Clustering model had an F-score of 0.72. We were able to predict that with the Decision Tree; the shopper was more likely to make a purchase between October and November.

In the future, I will look at a number of these methodologies and see how they fit into a larger dataset with more dependent or independent variables comparing different model performance, which will become a typical workflow in data product design and to discover what the good model results of our dataset look like relative to the best outcome of forecasting.

# References

Afsari, Kereshmeh, Charles M Eastman, and Daniel Castro-Lacouture. "Javascript Object Notation (Json) Data Serialization for Ifc Schema in Web-Based Bim Data Exchange." *Automation in Construction* 77 (2017): 24-51.

Briand, Lionel C, Jie Feng, and Yvan Labiche. "Using Genetic Algorithms and Coupling Measures to Devise Optimal Integration Test Orders." Paper presented at the Proceedings of the 14th international conference on Software engineering and knowledge engineering, 2002.

Broda, Tal, Darrell Esau, and Michael Hemmert. "Active Waterfall Charts for Continuous, Real-Time Visualization of Website Performance Data." Google Patents, 2016.

Fandango, Armando. *Python Data Analysis.* Packt Publishing Ltd, 2017.

Grabczewski, Krzysztof, and Norbert Jankowski. "Feature Selection with Decision Tree Criterion." Paper presented at the Fifth International Conference on Hybrid Intelligent Systems (HIS'05), 2005.

Hamill, Paul. *Unit Test Frameworks: Tools for High-Quality Software Development.* " O'Reilly Media, Inc.", 2004.

Huang, Xiaohui, Yunming Ye, Liyan Xiong, Raymond YK Lau, Nan Jiang, and Shaokai Wang. "Time Series K-Means: A New K-Means Type Smooth Subspace Clustering for Time Series Data." *Information Sciences* 367 (2016): 1-13.

Lamancha, Beatriz Pérez, Macario Polo Usaola, and Mario Piattini Velthius. "Software Product Line Testing." *A Systematic Review. ICSOFT (1)*  (2009): 23-30.

Nair, Lekha, Sujala Shetty, and Siddhanth Shetty. "Interactive Visual Analytics on Big Data: Tableau Vs D3. Js." *Journal of e-Learning and Knowledge Society* 12, no. 4 (2016).

Park, Hae-Sang, and Chi-Hyuck Jun. "A Simple and Fast Algorithm for K-Medoids Clustering." *Expert systems with applications* 36, no. 2 (2009): 3336-41.

Peker, Musa. "A Decision Support System to Improve Medical Diagnosis Using a Combination of K-Medoids Clustering Based Attribute Weighting and Svm." *Journal of medical systems* 40, no. 5 (2016): 116.

Safavian, S Rasoul, and David Landgrebe. "A Survey of Decision Tree Classifier Methodology." *IEEE transactions on systems, man, and cybernetics* 21, no. 3 (1991): 660-74.

Sievert, Carson. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny.* CRC Press, 2020.

Srinivasan, Jayakanth, and Kristina Lundqvist. "Using Agile Methods in Software Product Development: A Case Study." Paper presented at the 2009 Sixth International Conference on Information Technology: New Generations, 2009.

Velmurugan, T, and T Santhanam. "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points." *Journal of computer science* 6, no. 3 (2010): 363.

Weinberg, Amir, and Michael Pogrebisky. "Visualization of Web Sites and Hierarchical Data Structures." Google Patents, 2000.