Final_Group_Project_Report

August 15, 2021

GROUP 22 - Final Project Report: Maggie Dong, Sooa Han, Yusef Somani and Martin Wong

PREDICTING ONLINE SHOPPER PURCHASE INTENTION FROM WEBSITE VISIT DATA GATHERED OVER A ONE YEAR PERIOD

Introduction

Over the past decade, eCommerce has seen a meteoric rise in popularity, with the total share of eCommerce retail sales projected to top 21% in 2024, representing a 3-fold increase from its total share of approximately 7% in 2015 (Clement, 2019). This trend has been accelerated by the recent surge in online shopping over the current global pandemic. In Canada alone, eCommerce sales hit an estimated 3.9 billion CAD this past May, representing a 110% increase from the same month the year before (Statistics Canada, 2021). It is now estimated that Canadians purchase up to 10% of their products online, a market share, which is only expected to grow over the present decade (Canadian Broadcast Corperation, 2021). Indeed, this increase in online shopping has driven many retailers to create and expand their own digital platforms, a fact that has been reflected in the near 35% increase to the market value of Shopify.com, a Canadian multinational eCommerce company, which provides an eCommerce platform for online stores, over the past year and a massive 3500% increase over the past 5 years (Google Finance, 2021).

Consequently, the ability to predict the online shopping behaviours and purchasing intentions of prospective clients has become of critical importance to the economic success of most retail outlets, with past studies correlating perceived online transaction inconvenience, information, system and service quality, as well as website design with the rate of digital shopping cart abandonment and failure to secure online sales (Rajamma, et al., 2009; Kaushik, et al., 2017). Therefore, the current study will seek to predict the purchase intention of online shoppers based on website visit information obtained by Sakar et al. (2019) and aggregated in their online shopper data set, currently stored on the UCI repository (retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset). More specifically, the following analysis will attempt to answer the question, can the number of visits to specific webpages and the time spent on these webpages be used to predict whether or not a purchase will be made during an online shopping session?

[1]:
```r
#demonstrate that the data set can be read from web into R
install.packages("formattable")
```

```
Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done
```

[2]:

```r
#load packages
library(tidyverse)
library(tidymodels)
library(formattable)
library(knitr)
library(caret)
library(RColorBrewer)
```

```
  Attaching packages                                      tidyverse
1.3.0

 ggplot2 3.3.2       purrr   0.3.4
 tibble  3.0.3       dplyr   1.0.2
 tidyr   1.1.2       stringr 1.4.0
 readr   1.3.1       forcats 0.5.0

Warning message:
"package ‘ggplot2’ was built under R version 4.0.1"
Warning message:
"package ‘tibble’ was built under R version 4.0.2"
Warning message:
"package ‘tidyr’ was built under R version 4.0.2"
Warning message:
"package ‘dplyr’ was built under R version 4.0.2"
  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()

Warning message:
"package ‘tidymodels’ was built under R version 4.0.2"
  Attaching packages                                     tidymodels
0.1.1

 broom     0.7.0       recipes
0.1.13
 dials     0.0.9       rsample   0.0.7
 infer     0.5.4       tune      0.1.1
 modeldata 0.0.2       workflows 0.2.0
 parsnip   0.1.3       yardstick 0.0.7

Warning message:
"package ‘broom’ was built under R version 4.0.2"
Warning message:
"package ‘dials’ was built under R version 4.0.2"
Warning message:
```

```
"package ‘infer’ was built under R version 4.0.3"
Warning message:
"package ‘modeldata’ was built under R version 4.0.1"
Warning message:
"package ‘parsnip’ was built under R version 4.0.2"
Warning message:
"package ‘recipes’ was built under R version 4.0.1"
Warning message:
"package ‘tune’ was built under R version 4.0.2"
Warning message:
"package ‘workflows’ was built under R version 4.0.2"
Warning message:
"package ‘yardstick’ was built under R version 4.0.2"
  Conflicts
tidymodels_conflicts()
 scales::discard() masks
purrr::discard()
 dplyr::filter()   masks
stats::filter()
 recipes::fixed()  masks
stringr::fixed()
 dplyr::lag()      masks stats::lag()
 yardstick::spec() masks readr::spec()
 recipes::step()   masks stats::step()


Attaching package: ‘formattable’


The following objects are masked from ‘package:scales’:

    comma, percent, scientific


Warning message:
"package ‘knitr’ was built under R version 4.0.1"
Loading required package: lattice


Attaching package: ‘caret’


The following objects are masked from ‘package:yardstick’:

    precision, recall, sensitivity, specificity


The following object is masked from ‘package:purrr’:
```

```
lift
```

[3]:

```
#load data
shoppers <- read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/
↪00468/online_shoppers_intention.csv")
```

```
Parsed with column specification:
cols(
  Administrative = col_double(),
  Administrative_Duration = col_double(),
  Informational = col_double(),
  Informational_Duration = col_double(),
  ProductRelated = col_double(),
  ProductRelated_Duration = col_double(),
  BounceRates = col_double(),
  ExitRates = col_double(),
  PageValues = col_double(),
  SpecialDay = col_double(),
  Month = col_character(),
  OperatingSystems = col_double(),
  Browser = col_double(),
  Region = col_double(),
  TrafficType = col_double(),
  VisitorType = col_character(),
  Weekend = col_logical(),
  Revenue = col_logical()
)
```

[4]:

```
# we will use the logical Revenue variable as the target variable, and convert␣
↪it to the factor datatype
shoppers <- shoppers %>%
    mutate(Revenue = as_factor(Revenue))
head(shoppers)
```

A tibble: 6 × 18

Administrative <dbl>	Administrative_Duration <dbl>	Informational <dbl>	Informational_Duration <dbl>	Produc <dbl>
0	0	0	0	1
0	0	0	0	2
0	0	0	0	1
0	0	0	0	2
0	0	0	0	10
0	0	0	0	19

Table 1: The data set was formed from 12,330 website visit session, with each session being a visit from a unique individual to the website over a 1 year period. The variables listed in the data

frame represent different page ‘types’ visited on each website by each user during a single session. They were collected by Sakar et al. from the URL information of the web pages visted by different individuals, as well as Google Analytics for ‘Bounce and Exit Rate’, as well as ‘Page Value’.

Administrative: number of pages visited for account managment

Administrative duration: total time spent on pages for account management

Informational: number of pages visited for information about the website, contact details, or physical address of the shopping location

Informational duration: total time spent on informational pages

Product Related: Number of product related pages visited

Product-related duration: total time spent on produce related pages

Bounce Rate: Average bounce rate - percentage of visits that enterred the site at that page and then left without performing any other actions on that page (I.e. visited website and immediately left again)

Exit Rate: Average exit rate - percentage of website visits, where the specifc web page was the last one visted on the website

Page Value: average value of a webpage product visited before a transaction was completed

Special Day: Closeness of site visit to a special day

Revenue: Whether or not a purchase was made

Preliminary Exploratory Data Analysis Methods and Results

[5]:

```
# numerical variables used in the analysis model
numeric_shoppers <- shoppers %>%
    select(where(is.numeric)) %>%
    select(-c(OperatingSystems,Browser,Region,TrafficType))


num_shoppers_count <- numeric_shoppers %>%
    nrow() # number of observations in each class
shoppers_mean <- map_df(numeric_shoppers, mean, na.rm = TRUE) # means of the␣
 ↪predictor variables
shoppers_min <- map_df(numeric_shoppers, min, na.rm = TRUE) # mins of the␣
 ↪predictor variables
shoppers_max <- map_df(numeric_shoppers, max, na.rm = TRUE) # maxs of the␣
 ↪predictor variables
num_missing_data1 <- colSums(is.na(numeric_shoppers)| is.
 ↪null(numeric_shoppers)) # numbers of missing data in predictor variables

# make a dataframe consisting of the numerical variables used in the analysis
t1 <-␣
 ↪rbind(num_shoppers_count,shoppers_mean,shoppers_min,shoppers_max,num_missing_data1)
rownames(t1)<-c("count","mean","min","max","# missing data")
```

```
numeric_table <- formattable(t1)
numeric_table
```

```
Warning message:
"Setting row names on a tibble is deprecated."
```

A formattable: 5 × 10

	Administrative <dbl>	Administrative_Duration <dbl>	Informational <dbl>	Informatio <dbl>
count	12330.000000	12330.00000	1.233000e+04	12330.000
mean	2.315166	80.81861	5.035685e-01	34.4724
min	0.000000	0.00000	0.000000e+00	0.0000
max	27.000000	3398.75000	2.400000e+01	2549.3750
# missing data	0.000000	0.00000	0.000000e+00	0.0000

Table 2: All incomplete observations (observations missing numerical values) were removed and the data set was summarized into observation counts, mean, min and max for each variable and displayed in tabular form above. In terms of total time spent on specific webpage types, `ProductRelated_Duraton` was found to have the highest mean and max metrics, followed by `Administrative_Duration`. Similarly, when analyzing the data for the webpage types that were most visited, `ProductRelated` and `Administrative` webpages had the highest average number of visits, with means of 31.5 and 2.2 visits respectively.

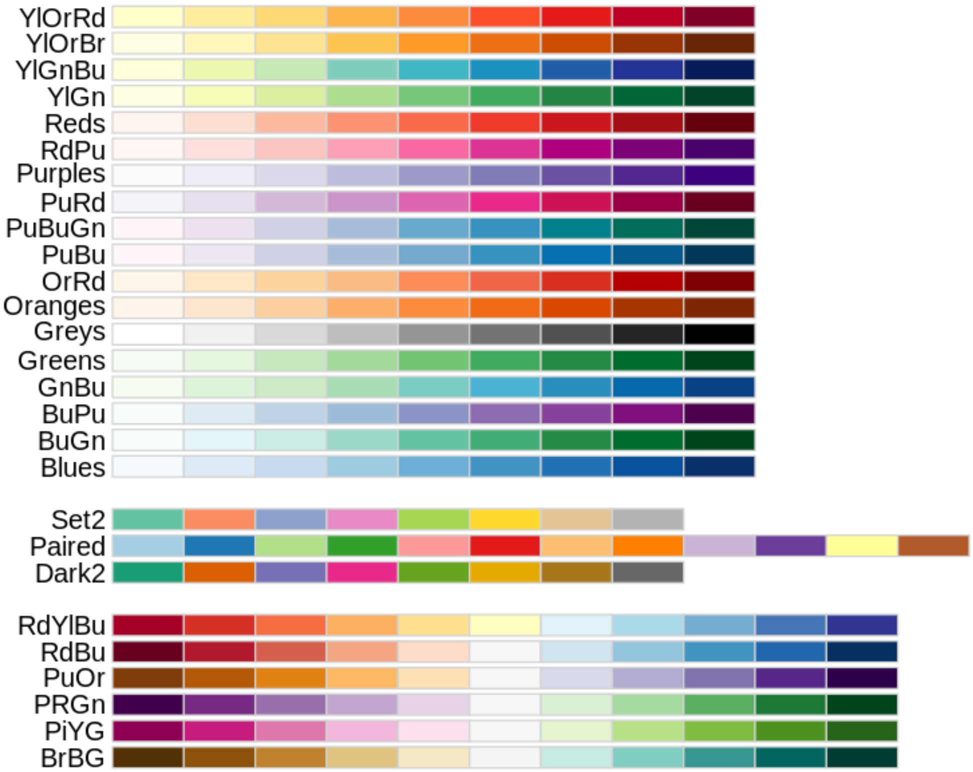
Given these results, as well as the project goal to predict the purchase intentions of online shoppers, we hypothesized that the `ProductRelated` and `ProductRelated_Duration` would be the most likely predictors to return accurate estimates regarding whether or not an online shopper would make a purchase, with both Administrative-type variables possibly playing a lesser role. Consequently, all 4 of these variables were subjected to further exploratory data analysis through graphical plots of visit duration against webpage type within and between `ProductRelated` and `Administrative` webpages.

[6]:

```
#re-organizing the data frame to allow for easier comparisons between absolute␣
 ↪number of visits and the duration of visits between possible predictor␣
 ↪variables
shopper_webtypes <- shoppers %>%
    select(Administrative, ProductRelated, Informational, Revenue) %>%
    pivot_longer(cols = Administrative:Informational, names_to =␣
 ↪"Webpage_Type", values_to = "Number_Visits")

shopper_webtimes <- shoppers %>%
    select(Administrative_Duration, ProductRelated_Duration,␣
 ↪Informational_Duration, Revenue) %>%
    pivot_longer(cols = Administrative_Duration:Informational_Duration,␣
 ↪names_to = "Webpage_Type", values_to = "Visit_Duration")
```

[7]:

```
#examine RColorBrewer colour palettes in order to select colour-blind friendly␣
 ↪colours
display.brewer.all(colorblindFriendly = T)
```



[8]:

```
options(repr.plot.width = 12, repr.plot.height = 10)

#plotting number for each type of webpage
webtype_plot <- shopper_webtypes %>%
    ggplot(aes(x = Webpage_Type, y = Number_Visits, fill = Revenue)) +
    geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set2") +
    labs(x = "Webpage Type", y = "Number of Visits", fill = "Purchase Made") +
    ggtitle("Proportion of Webpages Visted Resulting in Purchases Made") +
    theme(text = element_text(size = 20))

webtype_plot
```

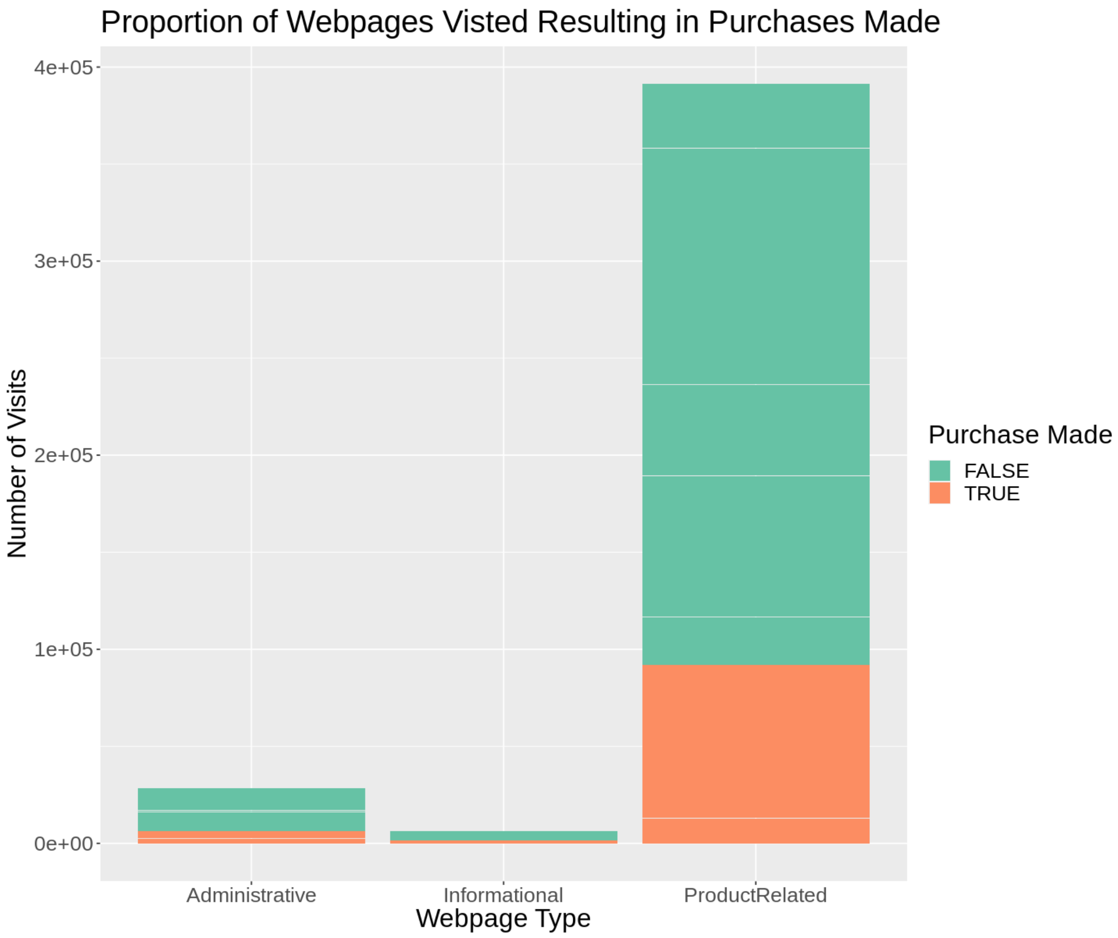


Figure 1: Plot of the types of web pages visited by individual online patrons and the proportion of visits that resulted in a purchase being made by the customer.

[9]:

```
#plotting duration visits for each type of webpage
webtime_plot <- shopper_webtimes %>%
     ggplot(aes(x = Webpage_Type, y = Visit_Duration, fill = Revenue)) +
        geom_bar(stat = "identity") +
        scale_fill_brewer(palette = "Set2") +
        labs(x = "Webpage Type", y = "Duration of Visits", fill = "Purchase␣
 ↪Made") +
        ggtitle("Webpage Visit Duration and Proportions of Purchases Made") +
        theme(text = element_text(size = 20))

webtime_plot
```

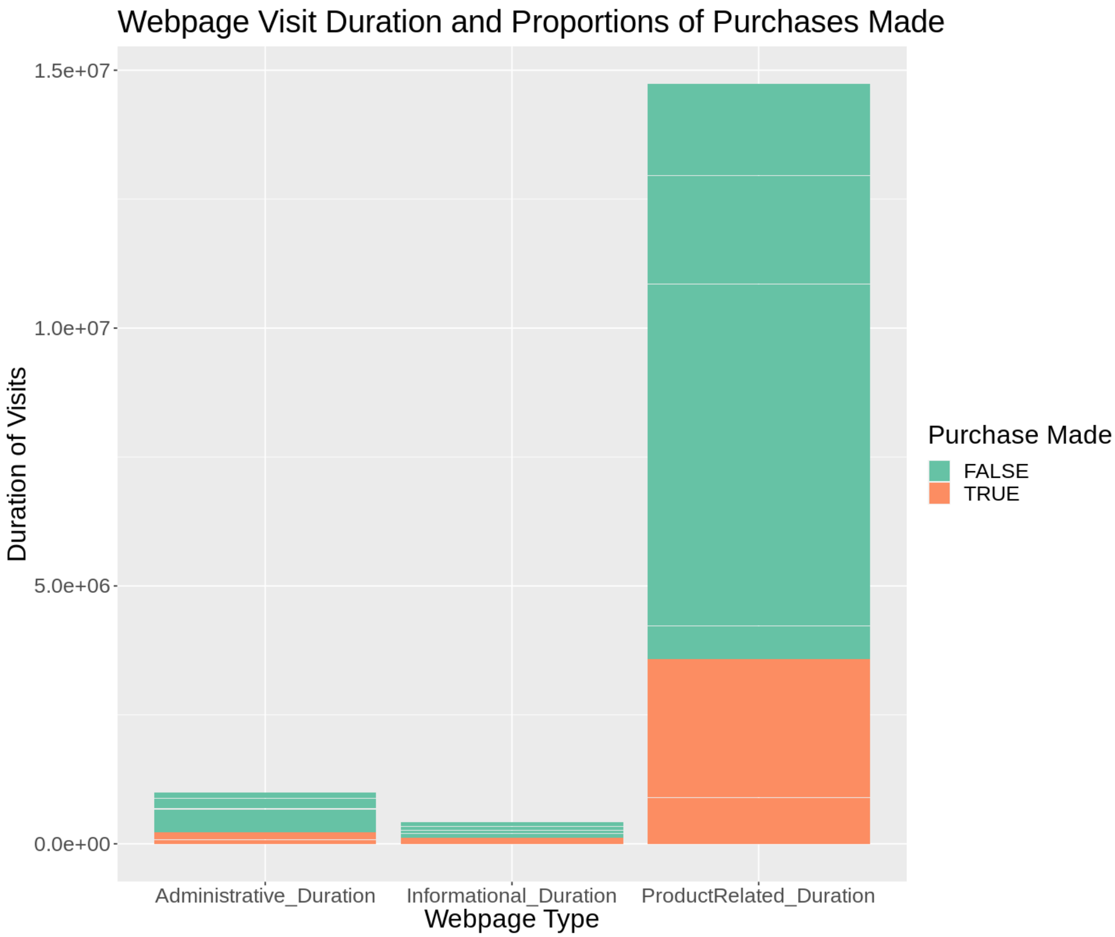


Figure 2: Plot of the time spent on each web page type by individual online patrons and the proportion of purchases made given the web page visited.

As expected, both the absolute number of webpage visits (Figure 1) and time spent on each webpage (Figure 2) were highest for the ProductRelated variables. Likewise, the proportion of webpages visits that resulted in purchases (Revenue = “TRUE”) also appeared to skew in favour of the ProductRelated and ProductRelated_Duration variables (Figures 1 & 2).

Given the high volume of data and substantial proportion of successful sales seen for the ProductRelated and ProductRelated_Duration variables, the two were plotted against each other to investigate whether any trends or relationships would emerge in the data between them and/or the Revenue factor variable.

```
[10]: #plotting ProductRelated webpage visits against duration and colouring for␣
       ↪purchases made
      shoppers_plot <- shoppers %>%
          ggplot(aes(x = ProductRelated_Duration, y = ProductRelated, colour =␣
       ↪Revenue)) +
          geom_point(alpha = 0.5) +
    scale_colour_brewer(palette = "Set2") +
    labs(x = "Time Spent on Product Webpages", y = "Product Webpage Visits",␣
 ↪color = "Purchase Made") +
    ggtitle("Purchase Intention and Product-Specific Webpage Visits") +
    theme(text = element_text(size = 20))

shoppers_plot
```

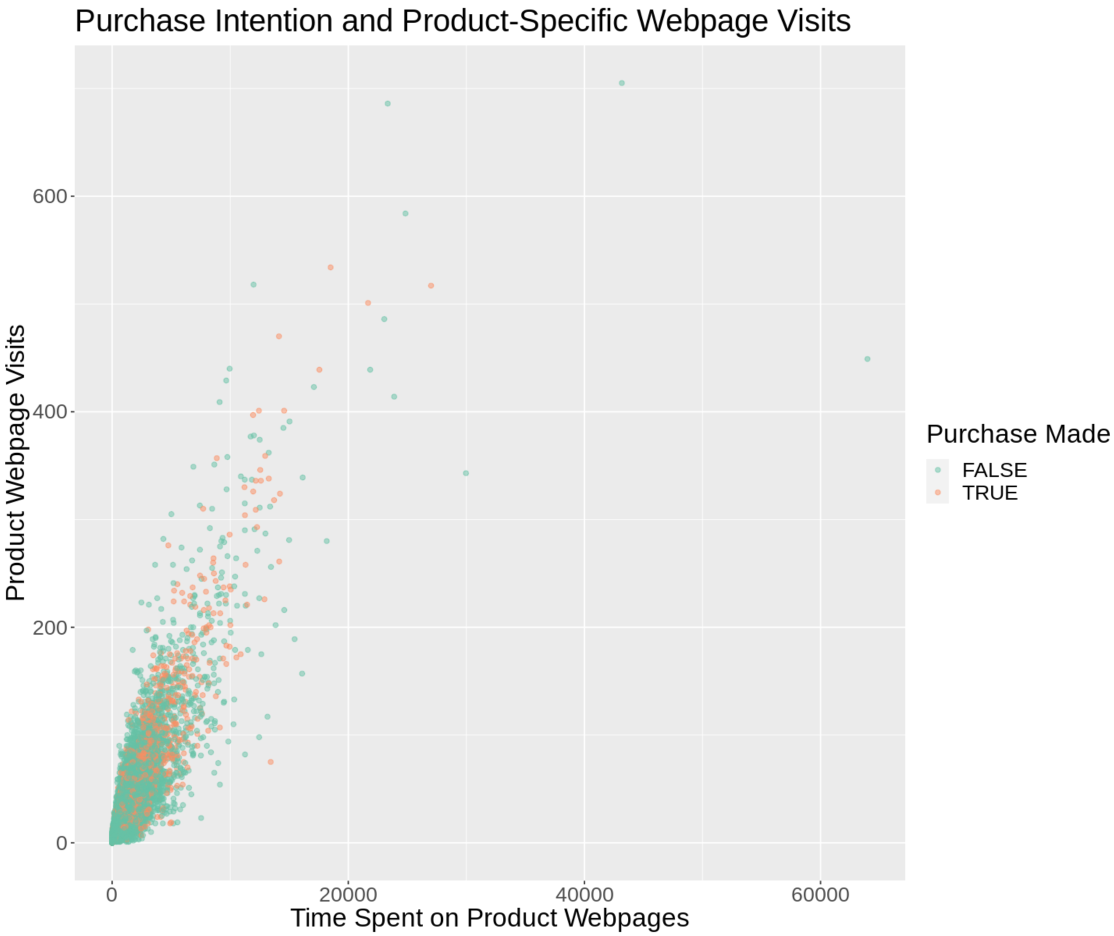


Figure 3: Plot of the number of `ProductRelated` web page visits and time spent on these web pages according to whether or not a purchase was linked to these visits.

The plot generated from the `ProductRelated` and `ProductRelated_Duration` variables failed to show any clear groupings of the observations based on whether or not a purchase was made. However, overall, it did seem that a greater number of successful sales resulted from a higher number of visits to product-related webpages and more time spent on such webpages. Furthermore, both the absolute number of product webpage visits and time spent on the same webpages appeared to demonstrate a strong positive, linear relationship with each other, further suggesting that both variables would likely work well together for use in subsequent predictive analyses. Additionally, it can be expected that the majority of online users who frequent the produce pages of online retailers have, at least to a certain degree, some intention to purchase and it logically follows that more product pages visits, as well as time spent on these webpages, will increase the chance of a successful purchase being made (Wiegran and Koth, 1999).

As a result, both `ProductRelated` and `ProductRelated_Duration` were selected as our predictor/explanatory variables, for designing a classification engine tasked with predicting whether or not a purchase would be made during an online shopping session. More specifically, the following analysis will set out to answer the question: Can the number of `ProductRelated` page visits and time spent on these pages (`ProductRelated_Duration`) be used to predict whether an online shopping session will result in a purchase (`Revenue` = "TRUE") with an acceptably high degree of accuracy?

Given the complex nature of, as well as the many possible factors involved in online shopping behaviour an acceptable prediction accuracy, for the classification engine, was defined as one that is able to correctly predict the class of the `Revenue` factor variable for a new observation with an accuracy $> 60\%$ (Rajamma et al., 2009; Kaushik and Srinivasa, 2017; Sakar et al., 2019). In other words, a predictor that is able to correctly estimate the label of an observation for significantly more than 50% of all new observations presented to it, given that a 50% prediction accuracy would be expected as the baseline for any predictor evaluating a 2-label factor variable, as this is the accuracy expected for any untrained, random chance estimation.

Predictive Analysis Methods

The classification task was performed using the K-nearest neighbours (K-NN) classification method. This method uses the known quantities of specified variables for a new observation to predict the label for an unknown categorical variable for that observation. It does this by assigning the same label to the new observation to which the majority of K-NN data points in a training set belong, based on the straight-line distance of these data points to the new observation as measured using the known quantitative variables.

Specifically, further analysis of the data was conducted using the K-NN classification, with cross validation using 5 chunks to further train our classifier for more accurate prediction results. In this data set, only `ProductRelated` and `ProductRelated_Duration` were used as explanatory variables to predict the selected target variable, `Revenue`. These two explanatory variables were chosen due to their high frequency in online shopping sessions, correlation with each other and perceived importance to the factor variable of interest.

When creating the model specifications for the K-NN classifier cross-validation task, weight_func was set to "rectangular", engine was set to 'kknn', neighbors were set to 'tune()' and the mode was set to 'classification', with the resamples for cross-validation being set to use 5 different chunks of data from the `shoppers_train` data subset. Each validation was then performed using a range of K-NN from 1-25, in sequence.

In order to ensure the accuracy of the classifier, the dataset was split such that 75% of the data was used to train the classifier (training set) and 25% of the data was reserved for testing the classifier (testing set). Typically, a 70:30 split for a dataset of size $n = 10,000$ and 80:20 for $n = 100,000$ has been used in the literature (Tokuç, 2021). As a result, it was decided that a training:testing ratio of 75:25 would return a sufficiently satisfactory performance for the classifier, trading off well between the need to train an accurate model (by using a larger training set), while also producing an accurate evaluation of its performance (by using a larger testing set).

Due to the nature of the K-NN classifier, the validation was assessed using a scatter plot of accuracy against tested K values to select the appropriate K-NN value for our classification. The training set was then fitted to the engine, with the specified K parameter (k = 10) and tested on the testing set, which contained the remaining 25% of our data. The resultant predictions accuracy was then tabulated and a confusion matrix was presented to assess the performance of the classification model.

Predictive Analysis and Results

```
[11]: #remove unnneeded variables from the original data set
      shoppers_clean <- shoppers %>%
          select(ProductRelated:ProductRelated_Duration, Revenue)

      shoppers_clean %>%
          head()
```

A tibble: 6 × 3

ProductRelated <dbl>	ProductRelated_Duration <dbl>	Revenue <fct>
1	0.000000	FALSE
2	64.000000	FALSE
1	0.000000	FALSE
2	2.666667	FALSE
10	627.500000	FALSE
19	154.216667	FALSE

Table 3: Table with the first 6 rows of our data set, with all non-predictor or target variables removed.

```
[12]: #we will use 75% of the data for training and 25% for testing.
      set.seed(1)
      shoppers_split <- initial_split(shoppers, prop = 0.75, strata = Revenue)
      shoppers_train <- training(shoppers_split)
      shoppers_test <- testing(shoppers_split)

      #create 5 cross-validation chunks
      shoppers_vfold <- vfold_cv(shoppers_train, v = 5, strata = Revenue)

      #create standardization recipe and model using training data without a k value
      shoppers_recipe <- recipe(Revenue ~ ProductRelated + ProductRelated_Duration,␣
       ↪data = shoppers_clean) %>%
          step_scale(all_predictors()) %>%
          step_center(all_predictors())

      knn_spec <- nearest_neighbor(weight_func = "rectangular", neighbors = tune())␣
       ↪%>%
          set_engine("kknn") %>%
          set_mode("classification")

      #cross validate in the workflow and find accuracy for each k in a range of 1-25
      gridvals <- tibble(neighbors = seq(1, 25))
```

```r
set.seed(1)

shoppers_results <- workflow() %>%
    add_recipe(shoppers_recipe) %>%
    add_model(knn_spec) %>%
    tune_grid(resamples = shoppers_vfold, grid = gridvals) %>%
    collect_metrics()

k_accuracy <- shoppers_results %>%
    filter(.metric == "accuracy")

#plot accuracy vs k and select k with highest reasonable accuracy
accuracy_vs_k <- k_accuracy %>%
    ggplot(aes(x = neighbors, y = mean)) +
    geom_point() +
    geom_line() +
    labs(x = "Nearest Neighbors (K)",
        y = "Accuracy Estimate") +
    ggtitle("Accuracy Estimate for Each K Value") +
    theme(text = element_text(size = 15))

accuracy_vs_k
```

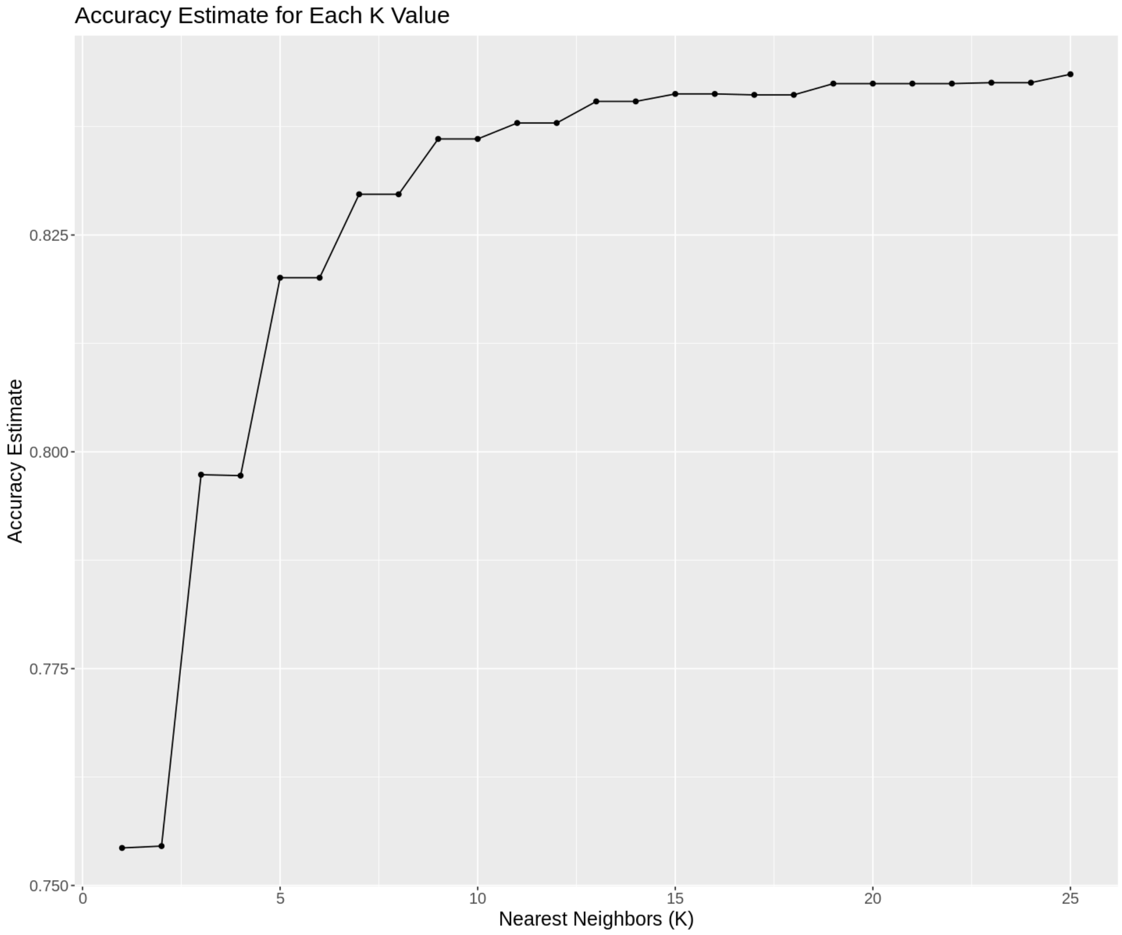


Figure 4: The plot shows the accuracy estimate for each K value produced during the cross-validation of our trained model. In this case, accuracy is defined as the fraction of correct predictions made by the trained classification engine for observations across all validation chunks of the `shoppers_train` data object. Based on these results, a K-NN value of 9 was chosen as the number of neighbouring data points that allowed the engine to make the highest accuracy predictions without beginning to account for too many K-NN data points and, thereby, risking underfitting the classification engine.

[29]:

```r
#retrain the model using the specified k (k = 9) and the training data
set.seed(1)

knn_spec <- nearest_neighbor(weight_func = "rectangular", neighbors = 9) %>%
    set_engine("kknn") %>%
    set_mode("classification")

knn_fit <- workflow() %>%
    add_recipe(shoppers_recipe) %>%
    add_model(knn_spec) %>%
    fit(data = shoppers_train)


#predict the labels for the test set using the trained engine and evaluate␣
 ↪accuracy of prediction
shoppers_test_predictions <- predict(knn_fit, shoppers_test) %>%
    bind_cols(shoppers_test) %>%
    metrics(truth = Revenue, estimate = .pred_class)

shoppers_test_predictions

shoppers_test_conf_mat <- predict(knn_fit, shoppers_test) %>%
    bind_cols(shoppers_test) %>%
    conf_mat(truth = Revenue, estimate = .pred_class)

shoppers_test_conf_mat
```

A tibble: 2 × 3

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
accuracy	binary	0.83484750
kap	binary	0.03108938

```
          Truth
Prediction FALSE TRUE
     FALSE  2554  458
     TRUE     51   19
```

Table 4: Tabulated metrics for the classification engines prediction accuracy for predicting online shopper purchase intention based on the number of `ProductRelated` web pages visited and time spent on those web pages (`ProductRelated_Duration`), as evaluated on a testing data subset containing 25% of the observations in the greater data set. The overall prediction accuracy of our engine was evaluated as being 83% accurate, exceeding our required, 60% threshold, with a total of 2573 correct predictions out of a total of 3082 complete observations (observations not missing numerical values for our variables of interest) being made.

Discussion

The rapid increase in online shopping over the past decade has necessitated that retailers greatly expand their online platforms in order to remain competitive. However, the resulting success of these platforms has been varied and largely dependent on several factors, many of which are influenced by website design, including ease of navigation/ability to find products, transaction convenience, total time required on webpages and perceived financial security risk (Rajamma et al., 2009; Kaushik and Srinivasa, 2017; Araffin et al., 2018). As a result, the ability to predict whether a potential online shopper will make a purchase, based on the specific types of webpages they visit, and the time spent on these webpages, has become a useful tool in optimizing website design to increase the likelihood of a successful sale (Wiegran and Koth, 1999; Kaushik and Srinivasa, 2017; Sakar et al., 2019).

Consequently, the current study set out to answer the question, can the online purchase intention of individual shoppers be predicted from data gathered on the number of product webpages visited

(ProductRelated) and the time spent on such webpages (ProductRelated_Duration)? These 2 variables were chosen during exploratory data analysis due to the high number of complete observations (observation with all required numeric data) available for these 2 variables, the substantial proportion of ProductRelated page visits that resulted in revenue (Revenue = “TRUE”; see Figures 1 & 2), the strong positive relationship between them (Figure 3) and their perceived importance to the target variable (Revenue). It was expected that the number of visits to ProductRelated webpages, as well as the time spent on such pages (ProductRelated_Duration), would function as good predictors for online purchase intention, where a good prediction accuracy was defined as any accuracy exceeding a 60% correct prediction threshold.

The predictive analysis was completed through the design, cross-validation and training of a K-NN classification model using data gathered by Sakar et al. (2019), and stored on the UCI repository, to train and test the model. In order to assess the quality of the K-NN classifier, both a confusion matrix and a metrics data frame for accuracy were generated following application of the classification engine to the testing data set (Table 4). The classifier was found to correctly predict the label of the Revenue variable 83% of the time, fulfilling our required threshold criteria for prediction accuracy.

However, it should be noted that the prediction accuracy fell drastically when only looking at observations for which a purchase was made (Revenue = “TRUE”), with only 19 out of 477 observations (4%) being accurately predicted in this scenario. This may suggest a possible bias in our classification model towards type II error: i.e., the model is better at predicting when the Revenue label is “FALSE”, rather than “TRUE”, which may be the result of an imbalanced data set, containing a disproportionate ratio of Revenue = “FALSE” vs. “TRUE” observations. Indeed, correct predictions for the “FALSE”Revenue label demonstrated a substantially higher accuracy relative to the opposite label (“TRUE”), with the model correctly predicting the negative label 2554 times out of 2605 Revenue = “FALSE” observations, for a 98% prediction accuracy. Therefore, it is likely that the classifier was more inclined towards predicting the majority class, Revenue = “FALSE” observations, simply due to the far greater volume of these observations across the data set during its training. It is suggested that future analyses employ statistical methods for dealing with imbalanced data, such as ‘class confidence weighting’, and possibly look to add other variables to further increase the Revenue = “TRUE” prediction accuracy (Lui and Chawla, 2011).

Other variables of interest in the data set for predicting purchase intention (Revenue) may include, the bounce rate, SpecialDay or even the month of the year (see Supplemental Figure 1 & 2). It may also be wise to analyze the outcome of training the classifier using only the data for the ProductRelated or the ProductRelated_Duration variables alone, as past research has suggested that more time spent on the webpages of an online retailer may actually have a negative impact on positive purchase intention of the customer (Ariffin et al., 2018). This last direction may even help explain the majority class of negative Revenue labels for the data set. Additionally, further analysis could be completed to assess whether any of these variables, particularly the ProductRelated ones can be used to predict each other, as a way of further tuning predictive models for online shoppers purchase intention. Unfortunately, given the shortcomings of K-NN classification algorithms for dealing with imbalanced data, another predictive algorithm may be required that is better able to manage the class imbalance in this or other data sets.

Ultimately, given the 83% accuracy found for the classifier in this study, online stores could potentially make use of this model to increase the proportion of successful purchases, relative to online shoppers, by optimizing marketing strategies or website design to direct more online shoppers toward product-related webpages. An example of one such strategy includes audience-specific marketing, which aims to target adverts and product-placement toward specific audiences, based on browsing data, in order to display an excess of webpages with the likely products of interest to the potential buyer, in order to increase the chance of a successful sale. Future investigations could also be performed to determine whether the shortest possible website path to a product of interest (convenience of navigation to product-related webpages) has any effect on purchase intention and whether specific types of products (further subsetting the product-related pages) are associated with a higher rate of successful sales.

Conclusion

Using the K-nearest neighbours algorithm, with a K-value of 9, the study set out to determine whether the variables `ProductRelated` and `ProductRelated_Duration` could predict the online purchase intention (factor variable `Revenue`) of individual shoppers in a 2019 data set gathered by Sakar et al. and made publically available through the UCI repository. The algorithm was cross-validated and trained on a training data subset, before being tested on a smaller testing set, with both being generated from the same larger data frame. The metrics data produced determined that the classifier could predict purchase intention to an accuracy of 83%, exceeding the expectations set out for the analysis. As a result, this report demonstrated that the quantity of product-related pages visited and the time spent on said pages have potential for use in guiding both marketing strategies and website design.

References

Ajay Kaushik, N., & Potti Srinivasa, R. (2017). Effect of website quality on customer satisfaction and purchase intention in online travel ticket booking websites. Management, 7(5), 168-173.

Ariffin, S. K., Mohan, T., & Goh, Y. N. (2018). Influence of consumers' perceived risk on consumers' online purchase intention. Journal of Research in Interactive Marketing.

Aston, J., Vipond, O., Virgin K. & Youssouf, O. (2020). Retail e-commerce and COVID-19: How online shopping opened doors while many were closing. Retrieved from https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00064-eng.htm on 28 Jun 2021.

Clement, J. (2019). Worldwide e-commerce share of retail sales 2015-2023. Retrieved from https://www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/ on 28 July 2021.

Google Finance: Market Summary - Shopify Inc. Retrieved from https://www.google.com/finance on 12 August 2021.

Liu, W., & Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In Pacific-Asia conference on knowledge discovery and data mining (pp. 345-356). Springer, Berlin, Heidelberg.

Rajamma, R. K., Paswan, A. K., & Hossain, M. M. (2009). Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience. Journal of Product & Brand Management.

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), 6893-6908.

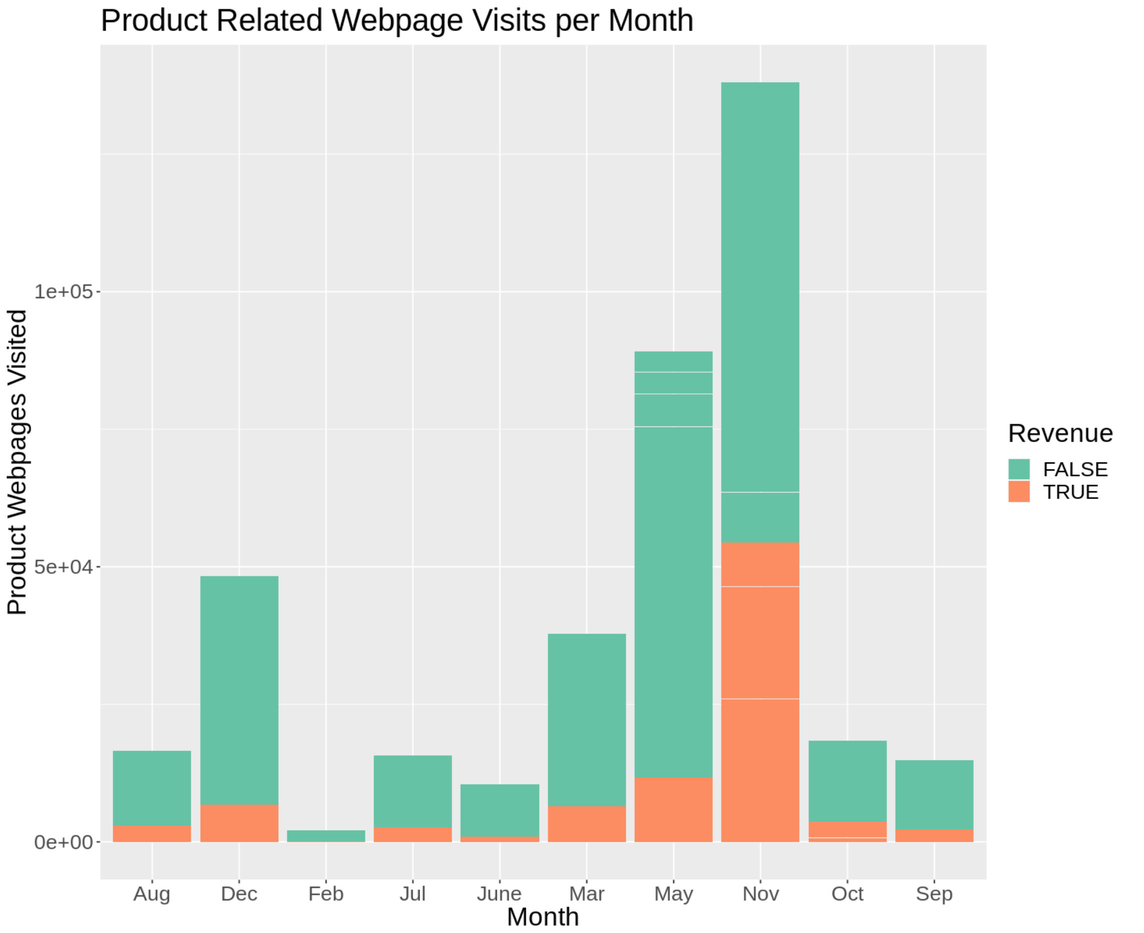
Tokuç, A. Aylin. "Splitting a Dataset into Train and Test Sets | Baeldung on Computer Science." Baeldung on Computer Science, Baeldung, 14 Jan. 2021, https://www.baeldung.com/cs/train-test-datasets-ratio.

Wiegran, G., & Koth, H. (1999). Customer retention in on-line retail. Journal of Internet Banking and Commerce, 4(1), 9909-07.

Supplemental Analysis

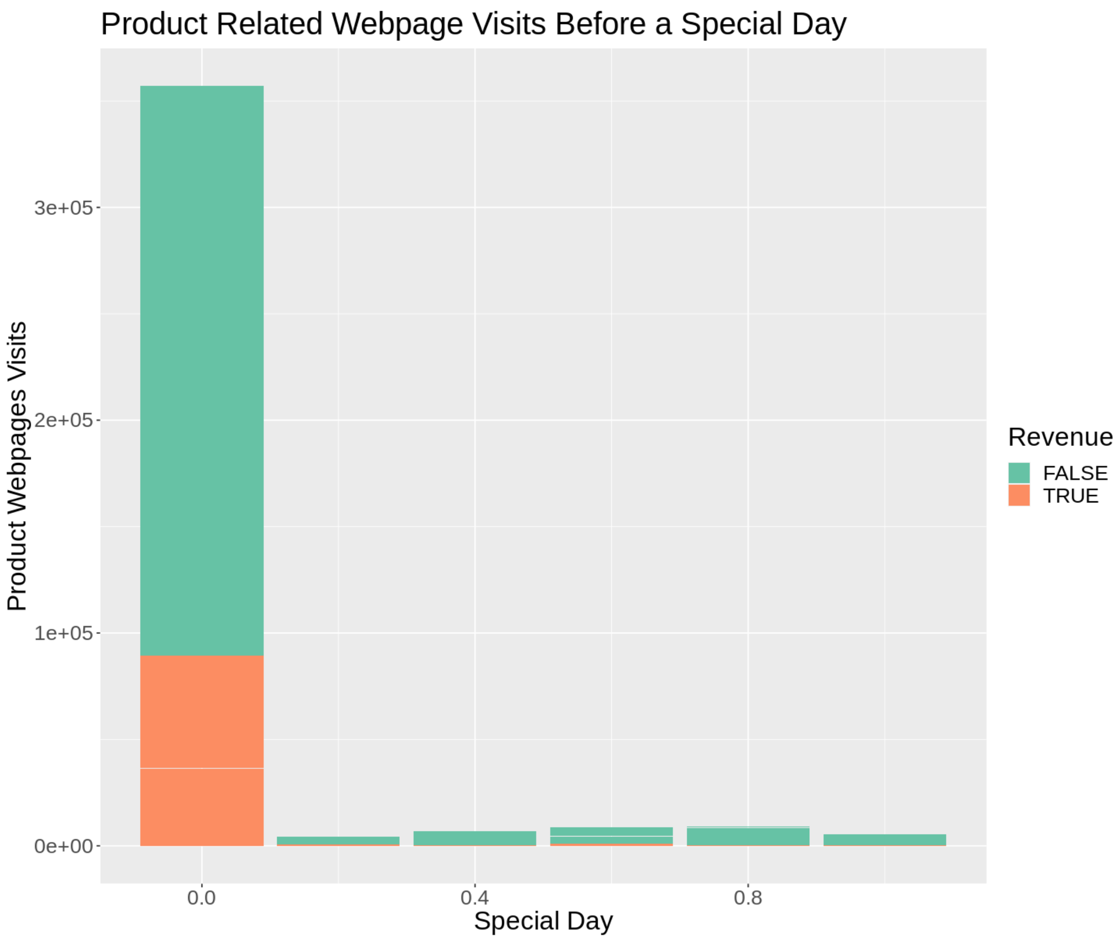
[13]:

```
#generate a supplemental plot of ProductRelated webpages visited per month and␣
 ↪the proportions that generated revenue
month_plot <- shoppers %>%
    ggplot(aes(x = Month, y = ProductRelated, fill = Revenue)) +
    geom_bar(stat = "Identity") +
    labs(x = "Month", y = "Product Webpages Visited", colour = "Revenue") +
    scale_fill_brewer(palette = "Set2") +
    theme(text = element_text(size=20))+
    ggtitle("Product Related Webpage Visits per Month")

month_plot
```



Supplemental Figure 1: The total number of unique product-specific webpage visits by month and the proportion of these visits that resulted in a purchase being made. A preponderance of sales are seen to occur for the months of May, November and December, suggesting the possibility of using the time of year as a predictor for online purchase intention.

[14]:
```r
#generate a supplemental plot of ProductRelated webpage visits made close to a special day and the proportions that generate revenue
specialday_plot <- shoppers %>%
    ggplot(aes(y = ProductRelated, x = SpecialDay, fill = Revenue)) +
    geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set2") +
    xlab("Special Day") +
    ylab("Product Webpages Visits") +
    labs(color = "Revneue") +
    ggtitle("Product Related Webpage Visits Before a Special Day") +
    theme(text = element_text(size = 20))

specialday_plot
```



Supplemental Figure 2: The total number of unique product-specific webpage visits made in the lead up to a 'special day' and the proportion that resulted in a purchase being made. May also be worth further investigation in the context of online purchase intention.

Acknowledgements

We would like to thank Melissa Lee, our group's TA, Ding, and the entire teaching team for their instruction and assistance over the past 6 weeks of the course. The course, without a doubt, achieved its goal in effectively teaching the foundations of the R programming language for statistical purposes to all those enrolled in it. We understand that this is a difficult time for everyone, which makes the quality with which the course was instructed that much more impressive and appreciated. Thank you for providing us with such a great learning experience.

[ ]: