**BUSINESS ANALYTICS WITH R**

**BUAN 6356.004**

**Professor Shujing Sun- Spring 2023**

**PROJECT REPORT**

**Online Shoppers’ Purchasing Intention**

**GROUP- 9**

Ankita Sirurmath PXS220071

Asmita Shetty AXS220272

Shreya Manoj Kolekar SXK220274

Siddhesh Chavan SXC220123

Suma Nednurkar SXN220122

Swapnita Samuel Mane SXM220231

Logo, icon

Description automatically generated

Contents

[INTRODUCTION 3](#_Toc133358987)

[DESCRIPTIVE STATISTICS 3](#_Toc133358988)

[Column Descriptions: 4](#_Toc133358989)

[Insights and Recommendations 8](#_Toc133358990)

[Data Pre-processing 9](#_Toc133358991)

[Classification Models executed by team 10](#_Toc133358992)

[Naive Bayes Classification 10](#_Toc133358993)

[K-NN Classification 11](#_Toc133358994)

[Decision Tree 13](#_Toc133358995)

[Linear SVM and Radial SVM 14](#_Toc133358996)

[Comparison 15](#_Toc133358997)

[Insights 15](#_Toc133358998)

## INTRODUCTION

The question that we are aiming to answer with our analysis is a predictive based question, that is: “Is it possible for us to predict whether a user will complete an online purchase on an e-commerce website using their clickstream and session data?” The growth of online shopping has led to the importance of predicting consumer behaviour to ensure the profitability of e-commerce businesses. The research question of our analysis is focused on predicting whether a user will make a purchase on an e-commerce website based on their clickstream and session data. By utilizing data mining, machine learning algorithms, and various sources of data, such as browser data and page information, we aim to create an accurate model to predict consumer purchase intentions. This information can be used to prompt prospective customers to finish online transactions in real-time, increasing total purchase conversion rates. The study focuses on anticipating online buyers' purchasing decisions and examining consumer buy intentions using empirical data to create a more precise prediction model.

## DESCRIPTIVE STATISTICS

Summary*:*

* The dataset provides information on the intentions of shoppers who want to buy products online.
* The dataset includes a range of attributes and product details that enable the analysis of purchasing trends and the extraction of valuable insights regarding shopper intentions.

About the Data:

* We obtained the data from the following source:

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

* The nature of online shopping varies geographically from country to country. This dataset provides details on the potential level of intention among shoppers to purchase products online. The data was collected through a survey to uncover patterns in shopping behavior. The information in this dataset can be utilized to identify trends in the behavior of online shoppers.
* The data consists of 12,330 records and 18 columns.
* The data doesn’t contain any null values.

### Column Descriptions:

**Administrative:** This refers to the quantity of pages of the administrative category that the user accessed.

**Administrative\_Duration:** This denotes the duration of time that was spent on pages belonging to this category.

**Informational:** This indicates the total count of pages of the informational category that were visited by the user.

**Informational\_Duration:** This refers to the duration of time that was spent on pages belonging to the specific Informational\_Durationcategory.

**ProductRelated:** This indicates the overall number of pages related to products that were visited by the user.

**ProductRelated\_Duration:** This refers to the duration of time that was spent on pages belonging to the specific ProductRelated\_Duration category.

Graphical user interface, application, table

Description automatically generated

**BounceRates:** This pertains to the ratio of unique visitors who access the website via the specific page and depart without initiating any further interactions or engagements with the platform.

**ExitRates:** This denotes the ratio of the total number of website pageviews that terminate on the particular page.

**PageValues:** The mean page value that is determined by calculating the average of the target page value and/or the eCommerce transaction completion value.

Table

Description automatically generated with low confidence

**SpecialDay:** This value indicates the proximity of the browsing date to significant occasions or holidays (such as Christmas or Mother's Day), when the likelihood of completing a transaction is higher.

**Month:** This contains the month in which the pageview was recorded, presented as a string.

**OperatingSystems:** This denotes a numerical value that represents the specific operating system employed by the user during the page viewing session

**Browser:** This denotes integer value that represents the web browser utilized by the user to access the page.

Chart, line chart

Description automatically generated

**Region:**  This refers to the integer value that indicates the geographical region where the user is situated.

**TrafficType:**  This denotes the integer value that represents the classification of traffic to which the user belongs.

**VisitorType:** A string that represents whether a visitor is a “New Visitor”, “Returning Visitor”, or “Other”.

**Weekend:** A boolean defining the timing of the session is on a weekend or not.

**Revenue:** This refers to a boolean value that indicates whether or not the user has finished the purchase.

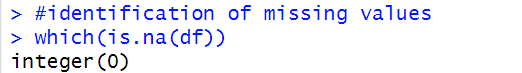
### Insights and Recommendations

* Most purchasing customers are returning customers
  + Collect contact information for targeted marketing like abandoned card email, promotion on item of interest etc.
* Special days are not driving sales
  + Expand Special Days tracking to include holidays
  + Track the closeness to each holiday to improve on probing and offer triggers
* Traffic Type 2 driving significant purchasing customers
  + Identify the traffic type.
  + Understand what works/collect more info
* Similarly identify OS and Browser 2 and collect more information on what works.

## Data Pre-processing

Data Cleaning:

1. **Handling missing values**

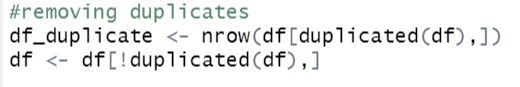


It is evident that there is no missing data present in the dataset. However, we noticed the presence of redundant data.

1. **Handling duplicate data**

To remove duplicate data, the following operation is performed.

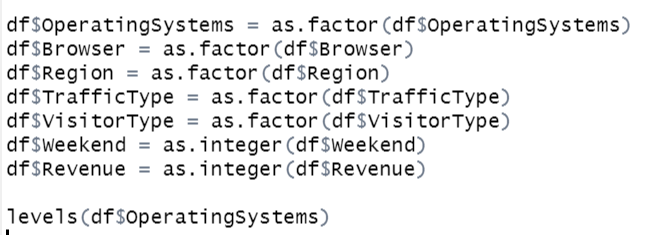






1. **Transforming categorical attributes into factor types**

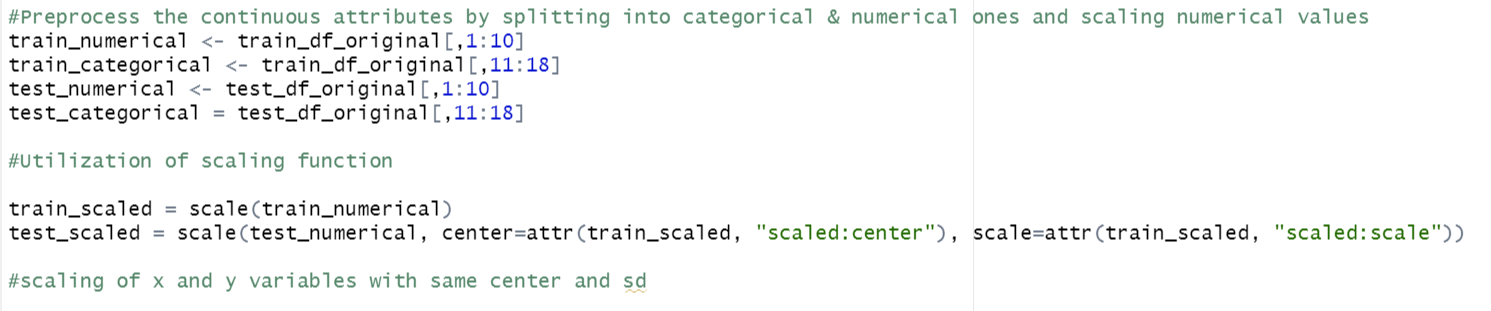
We further transform all categorical attributes into factor type and split the data into training and testing data.



1. **Utilization of scaling function in the testing and training data:**

Pre-process the continuous attributes by splitting into categorical & numerical ones and scaling numerical values.

Scaling is a technique used for comparing data that isn’t measured in the same way. The normalizing of a dataset using the mean value and standard deviation is known as scaling. This is also known as data standardization.



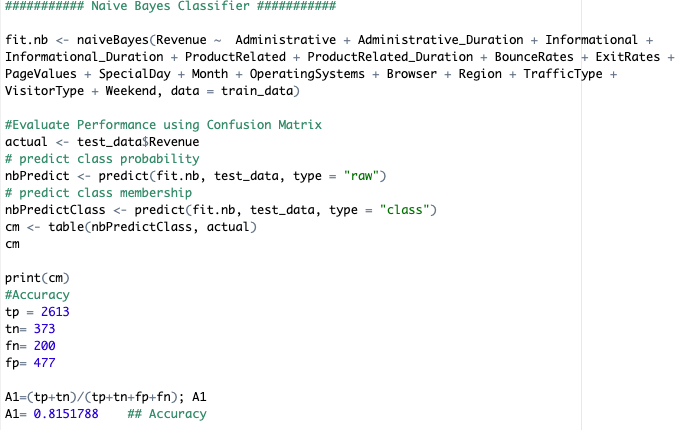
## Classification Models executed by team

### Naive Bayes Classification

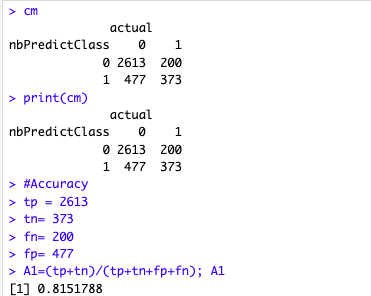
Naive Bayes is a probabilistic classification technique based on Bayes' theorem, which states that the probability of a hypothesis (or class) given the evidence (or data) is proportional to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis. In other words, it calculates the probability of each class given the input features and then selects the class with the highest probability as the output.

A naive Bayes classifier considers each of these features to contribute independently to the probability regardless of any possible correlations.

Code:



Output:



Upon Classification, the Naive Bayes Model returns an Accuracy of **81.51%**

### K-NN Classification

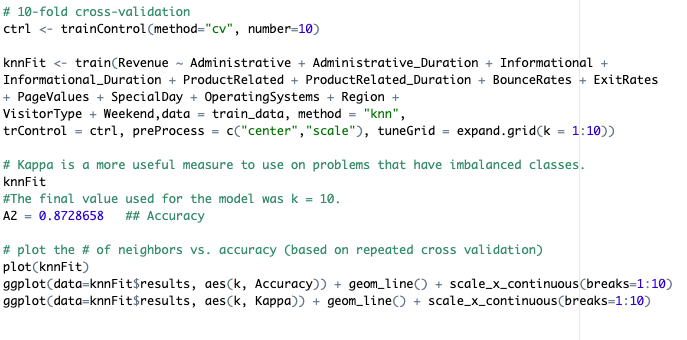
The k-nearest neighbours algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

Regression problems use a similar concept as classification problem, but in this case, the average the k nearest neighbours is taken to make a prediction about a classification. The main distinction here is that classification is used for discrete values (Revenue- factor (0/1)), whereas regression is used with continuous ones. However, before a classification can be made, the distance must be defined. Euclidean distance is most commonly used calculation criterion.

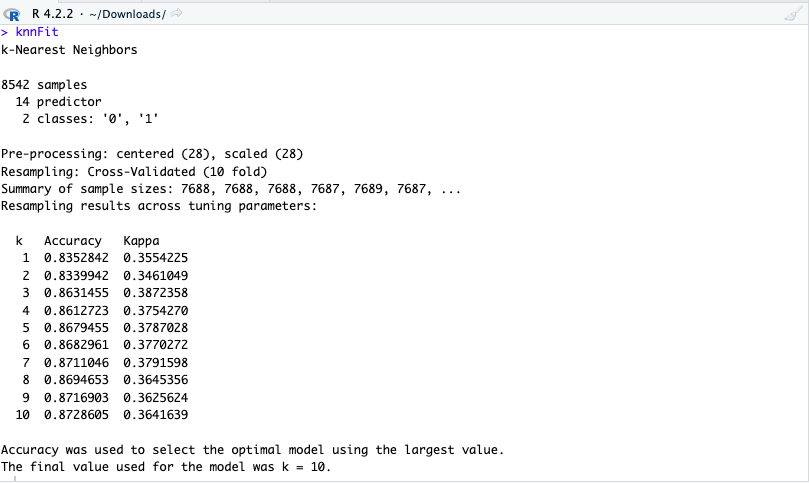
Euclidean distance: This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

When we try k-NN, by default it achieves accuracy of 87.28%, we further try visualizing model accuracy of different k, which shows that k=10 is the one that achieves the highest accuracy. Its result is approximately identical to NBC, but it is still reasonable since there are continuous and categorical attributes and k-nn is an expert on cluster, not on classification.

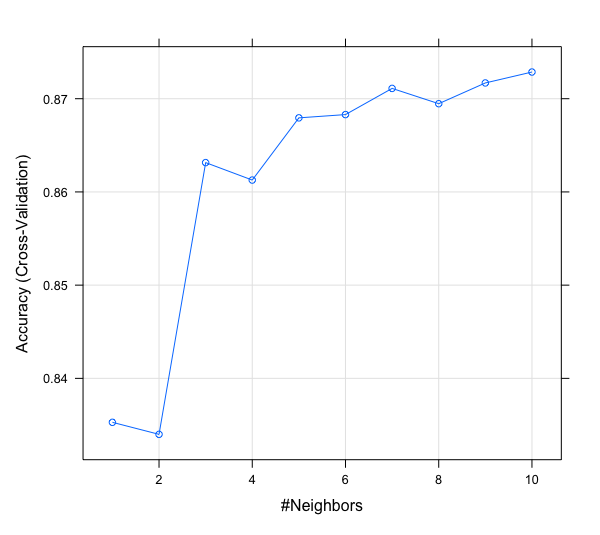
Code:



Output:



Plotting K Vs Accuracy:



### Decision Tree

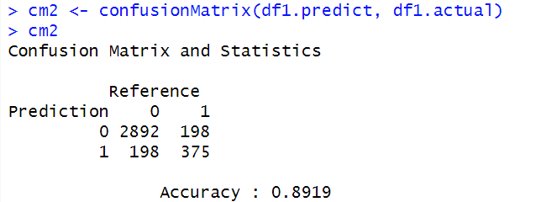
Decision tree is an efficient non-parametric method that can be used for both classification and regression. A decision tree has two main components: internal decision nodes and terminal leaves. Each internal node in the tree implements a test function using one or more features and each branch descending from that node is labelled with the corresponding discrete outcome.

During testing, when a new instance is given, the test pointed out by the root node is applied to the instance and according to the output of the decision node the next internal node that will be visited is determined. This process is then repeated for the subtree rooted at the new node until a leaf node is encountered which is the output of the constructed tree for the given test instance.

Diagram

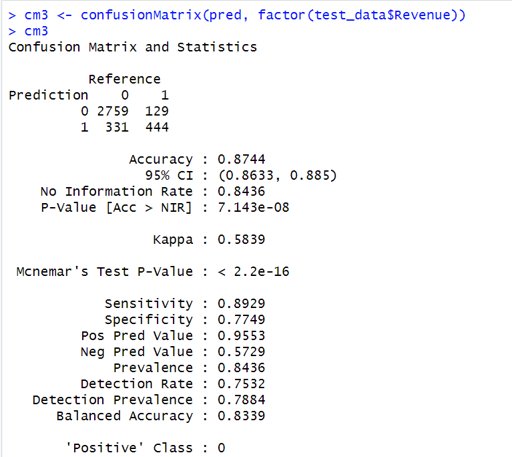
Description automatically generated

Accuracy: **89.19%**



### Linear SVM and Radial SVM

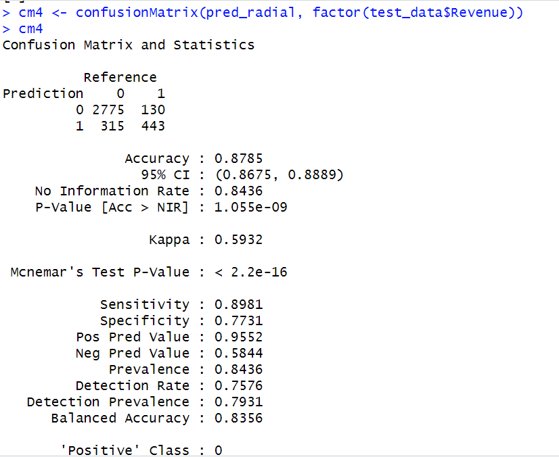
#### Linear SVM:



Hence, we tried Linear Support Vector Machine by default, which achieves an accuracy of 87.44%

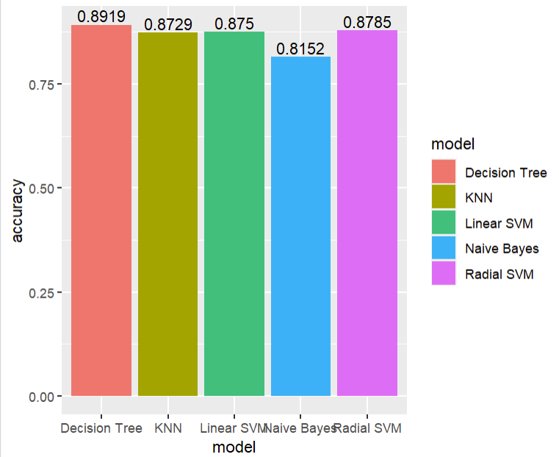
#### Radial SVM:

Evaluating Metrics:



Further, Radial-basis-function SVM is applied, by default it achieves accuracy of 87.85% which is approximately the same as that of Linear SVM.

## Comparison



Therefore, with evidence from the bar chart, we can affirm that the Decision Tree model performs the best among all the algorithms we have implemented.

## Insights

In this project, we explore methods to predict purchasing decision of consumers based on the 17 attributes. This real-world dataset that inspires us to analyse the behaviour of customers on E-commerce website.

We see that sometimes inference from our daily life experience could be inaccurate. For instance, we would believe “SpecialDay” has strong influence on the decision, since, it is in weekend we have more free time to browse the online shopping mart. From Decision tree, results show that PageValues, ProductRelated and ExitRates are among the most important predictors.

Based on our findings, we can offer recommendations to the website owner to adjust their advertising budget. One possible strategy would be to allocate more funds towards months when customers are more likely to complete their transactions, which could lead to increased return rates.

Also, they could re-design the informational pages to best cater to the customers’ demands for their longer stay, which further increases the probability of successful transactions.