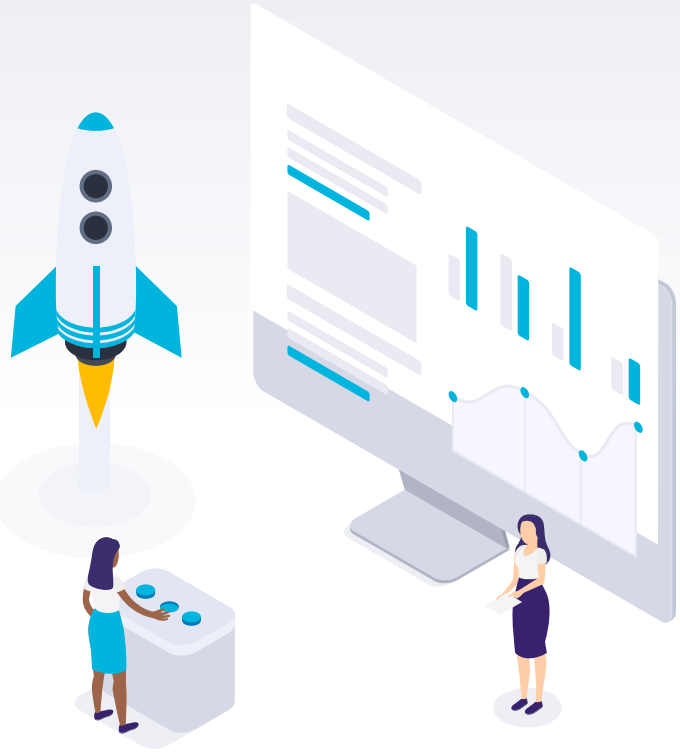
Analysis of Online Shoppers' Purchase Intention

By Sahil Shah

Contents

Motivation

- In the era of online shopping, buyers and sellers want to understand the behavior of online customers.
- Businesses will want to improve their revenue by focusing on customers and marketing their products.
- Thus, *analysis of the online shoppers' purchase intention* has become an emerging field in data mining.



Problem Definition

- Analyzing clickstream data for successful online businesses to extract the clicks and behavior through web page requests.

Dataset Description

- The dataset is taken from UCI Machine Learning dataset.
- URL : https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
- Total of 12,330 records where each row corresponds to the session data of the particular users.

Dataset Attributes

Quantitative Data

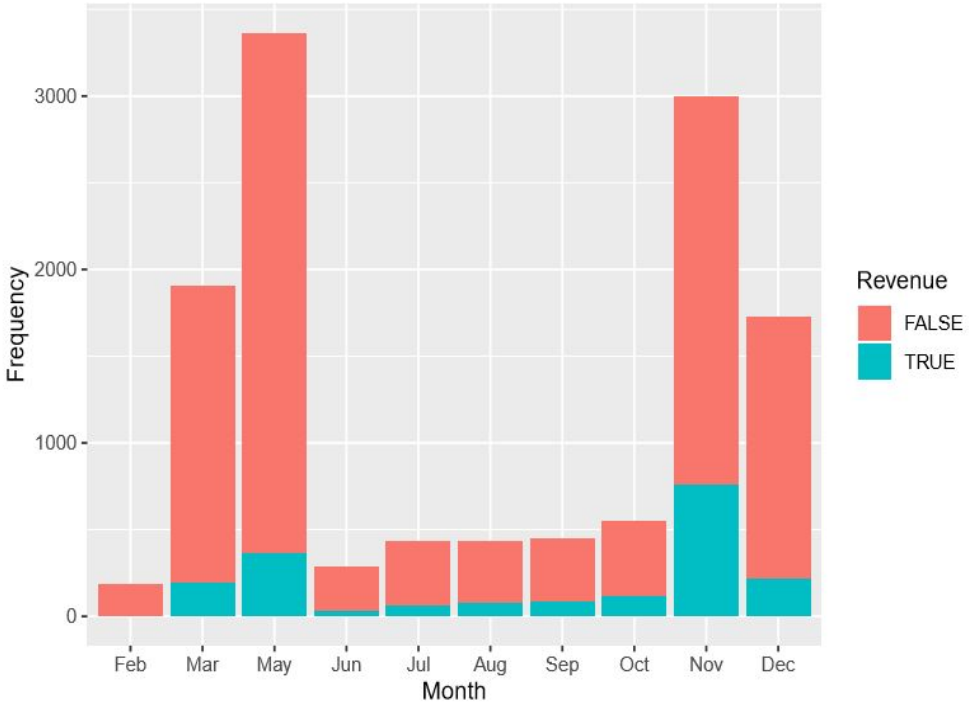
Feature Name	Description
Administrative	Number of pages visited by the visitor about account management
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages
Product related	Number of pages visited by visitor about product related pages
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce rates	Average bounce rate value of the pages visited by the visitor
Exit rate	Average exit rate value of the pages visited by the visitor
Page value	Average page value of the pages visited by the visitor
Special day	Closeness of the site visiting time to a special day

Dataset Attributes

Qualitative Data

Name	Description
OperatingSystems	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
VisitorType	Visitor type as New Visitor, Returning Visitor, and Other
Weekend	Boolean value indicating whether the date of the visit is weekend
Month	Month value of the visit date
Revenue	Class label indicating whether the visit has been finalized with a transaction

Data Analysis



- The graphs represents the data, with respect to the corresponding months of the year.
- It also describes the frequency of the customers that helped the organisation to generate the revenue.

Key Issues Addressed

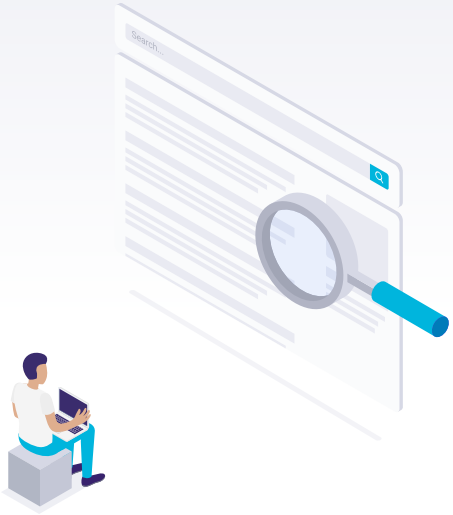
- Since e-commerce is getting more extensive, sellers need to understand what factors come into play for shoppers.
- Enable websites to better target ads or other factors which may lead to an increase in sales.
- Finding out problems in the existing market as well as knowing the capabilities of a brand getting released in the new market.

Alternative Ways to Resolve Issues

- In our work, we used, KNN, K-Means, Decision Trees and Logistic Regression to resolve our issues.
- This can also be done using Gradient Boosting, Random Forest and Adaboosting.

Related Work & Limitation

- There were limitations in our dataset which did not allow us to better predict the purchasing intent with Clustering models.
- To use Clustering models, we would need more variables and observations such as socio-economic or demographic information
- Since the dataset was qualitative, had to perform additional pre-processing to convert to quantitative data for different algorithms.



Algorithms

- Candidate Algorithms :
  - Logistic Regression (Regression)
  - Decision Tree (Classification)
  - K-Means (Clustering)
- Core Algorithms:
  - K-Nearest Neighbors (Classification)

Our Approach

- Building various classification, Regression and Clustering Algorithms and comparing them. Which includes Logistic Regression, Decision Tree, K-Means Clustering and K-NN Classification.
- Values of different evaluation metrics like Accuracy, Recall/Sensitivity, and Specificity are calculated to compare the performance of each of the algorithms.
- Use these classification models to predict the shopper's intentions.

Validation

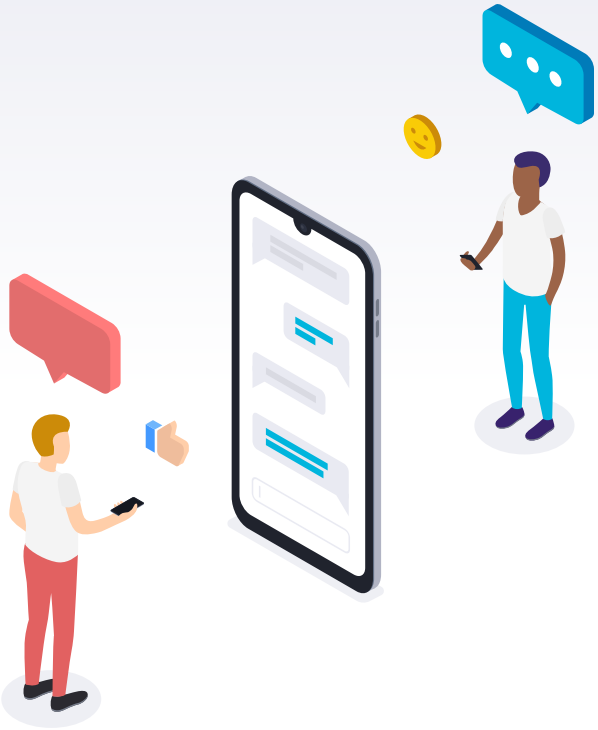
- For each algorithm we splitted the data 75:25, i.e. ¾ as the Training and ¼ as Testing.
- The training data and test data for each algorithm is same.
- The training of the model is done on the training data.
- Also preprocessing of the data, with respect to conversion of factors is done as requirement of the algorithm.

Results

Algorithms	Accuracy(%)	Sensitivity/Recall (%)	Specificity(%)
Logistic Regression	88.97	98	39.62
Decision Tree	89.65	95.13	59.75
K-Means Clustering	45.06	47.03	59.14
K-Nearest Neighbors	87.90	88.51	78.26

Conclusion

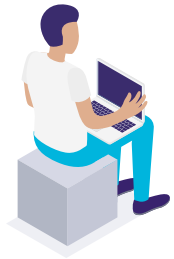
- Identified Decision Tree as a suitable model to predict the purchase intention of the shopper visiting the web-pages more accurately.
- The comparison of the recall in Logistic and Decision Tree are high, which represents True revenue.



Future Work

- Since we have got decent accuracy for this diversified dataset where data points are very sparse, the accuracy can be increased by using deep-learning.
- Also, by incorporating different Ensemble Methods of machine learning algorithms we can increase the accuracy.

THANK YOU!



ANY QUESTIONS?