Springboard Capstone Project 1: Analysis of Online Shopping Dataset

Michael Thabane

Kaggle Dataset Repository

March 18, 2020

Introduction

Business Problem

- Rapid increase in online shopping leads to need of finding ways to know customer's intention
- Use data from web sessions to answer this question
- Supervised classification data science problem which can be solved using ML classification algorithms

Approach

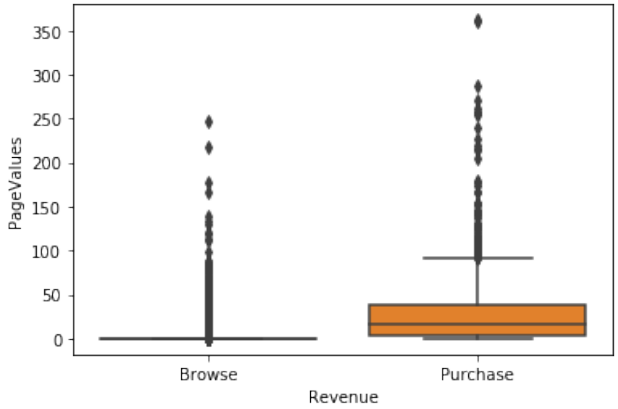
Data Acquisition and Wrangling

- Data was taken from the Kaggle Dataset Repository
- Values of some features are derived from the URL information of the pages visited by the user
- Some features represent the metrics measured by "Google Analytics" for each page in the e-commerce site
- No wrangling was needed and the data was then split into two data sets based on Revenue

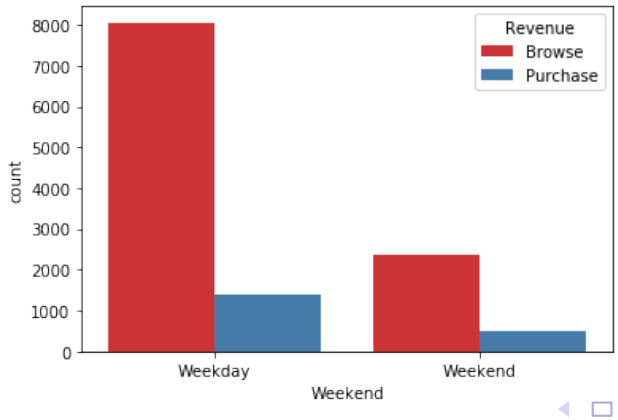
Approach

Descriptive Analysis

- Browsing Customers visit less pages on average than Purchasing Customers illustrated in the boxplot below



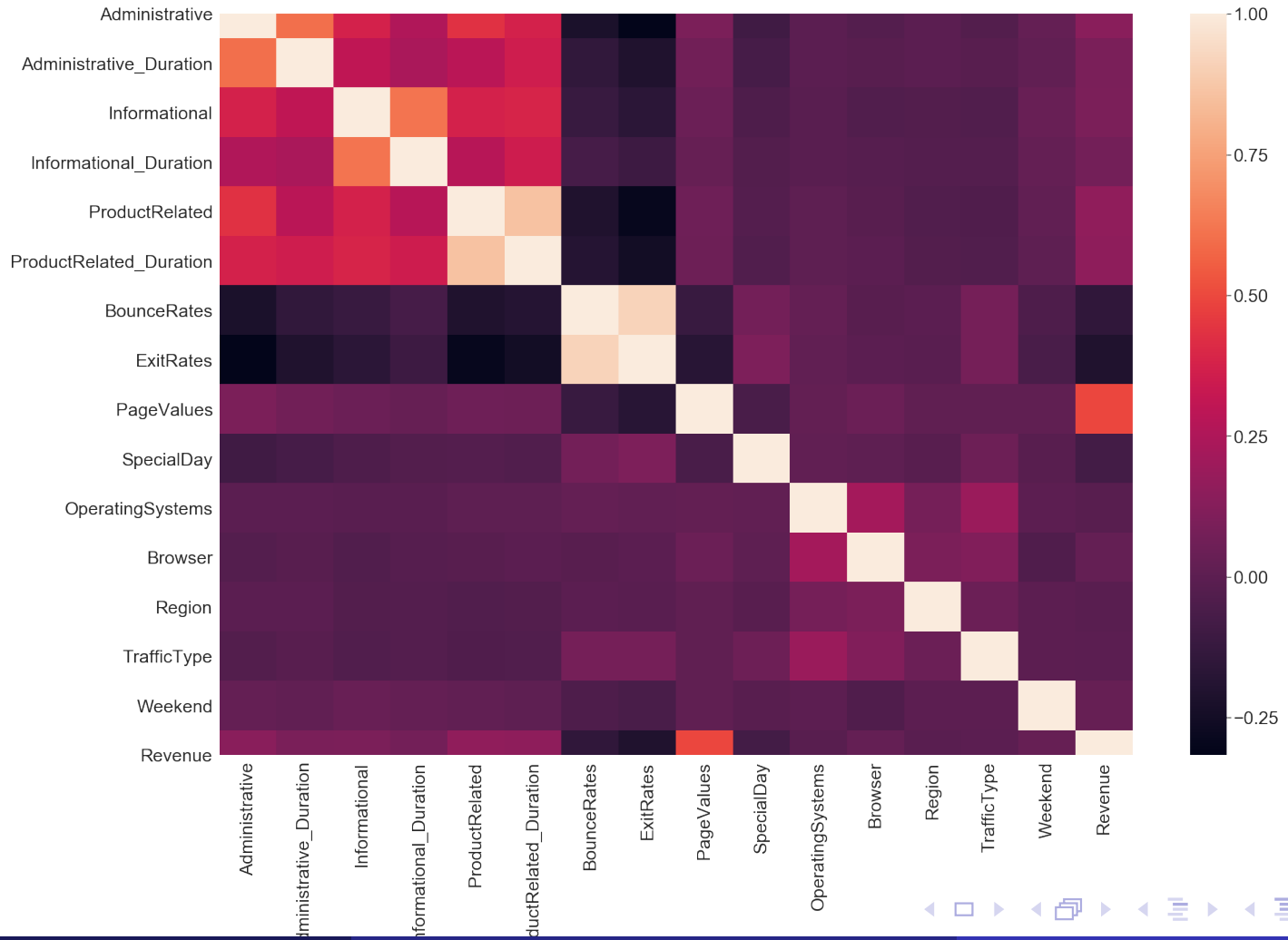
- Ratio between browsing and and purchasing customers is lower on the weekend vs. week days shown in the histogram below



Approach

Correlation

- Based on the plot below the page value rate is the most strongly correlated variable to the session being a purchase



Approach

Hypothesis Testing

- Test four hypotheses described below:
  1. A Special Day is independent of a user making a Purchase
  2. A Returning User is independent of the user making a purchase
  3. Weekends are independent of a user making a Purchase
  4. The Informational Duration means for the two groups of data (Purchase, Browse) are equal.
- There was significant evidence to reject all four hypotheses

Approach

Baseline Modeling

- Baseline models used were:
  1. K-nearest Neighbours Classifier
  2. Logistic Regression Classifier
- The training and test split was 80-20
- Using a 5-fold cross-validation the accuracy of models are as follows:
  1. KNeighbors Accuracy = 0.86 (+/- 0.03)
  2. Logistic Regression Accuracy = 0.88 (+/- 0.01)

Model	User Decision	Precision	Recall	F1 Score	Support	Comments
K-Nearest Neighbors	Browse	0.86	0.98	0.91	2060	Optimized parameters: n = 6
	Purchase	0.62	0.18	0.29	406	
Logistic Regression	Browse	0.88	0.98	0.92	2060	Optimized parameters: Penalty = l1
	Purchase	0.73	0.3	0.43	406	C = 0.00021544346900318823

Approach

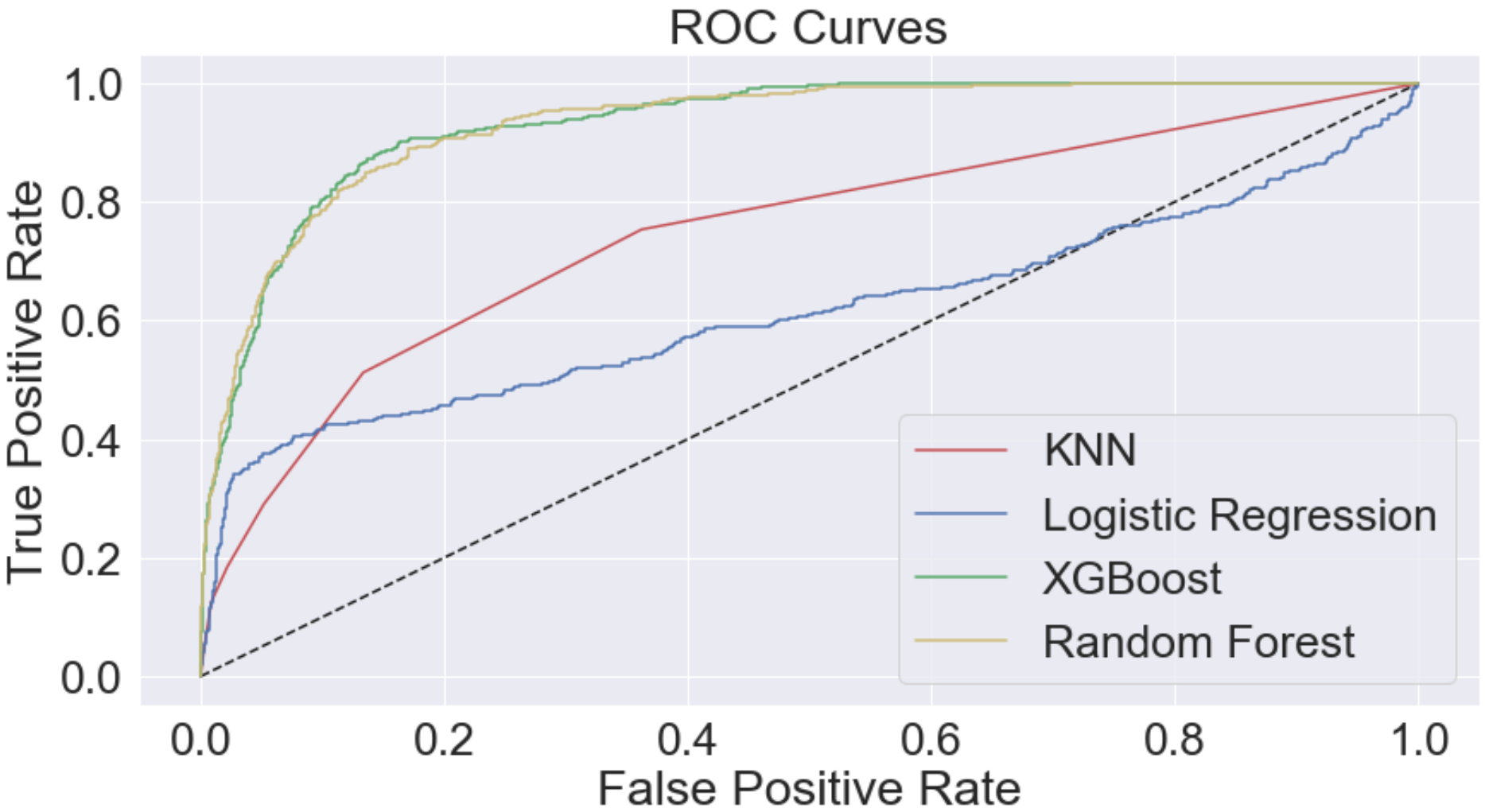
Extended Modeling

- Baseline models used were:
  1. XGBoost Classifier
  2. Random Forest Classifier
- Using a 5-fold cross-validation the accuracy of models are as follows:
  1. XGBoost Accuracy = 0.90 (+/- 0.03)
  2. Forrest Accuracy = 0.89 (+/- 0.03).

Model	User Decision	Precision	Recall	F1 Score	Support	Comments
XGBoost	Browse	0.92	0.96	0.94	2060	Optimized parameters: Learning rate = 0.05, Max Depth = 3, Min Child Weight = 1, gamma = 0.1 , Colsample by tree = 0.6, n-estimators = 300
	Purchase	0.73	0.56	0.63	406	
Random Forest	Browse	0.92	0.96	0.94	2060	Optimized parameters: n-estimators = 100, Max Depth = 20
	Purchase	0.75	0.56	0.64	406	Min Sample Split = 12, Min Sample Leaf = 3

Approach

Model Comparisons



Conclusions and Future Work

- Hypothesis tests shows that the "Weekend" and "Special Days" makes a statistically significant difference on the outcome of a session ending with a purchase
- Hypothesis tests also show that there is a statistically significant difference between outcome groups for the variable "Informational Duration"
- From the model prediction analysis we can conclude that the XGBoost and Random Forest models are the best performing when predicting if a user session will include a purchase
- From the Feature Influence Analysis we know that "Page Value" has a positive impact on the outcome of a session resulting in a purchase
- A new or improved product recommendation system can be implemented in order to engage users so the average session will have more page values and lower exit rates.

Recommendation for the Client

- My recommendation to the client would be to use an XGBoost model when looking to predict purchases by session as this model performs the best
- Based on the report we know that the "Page Values" per session has a positive impact on the session concluding with a purchase
- Recommend implementing a product recommendation system to encourage users to explore more products on the site and increase the "Page Values" per session

Thanks