Online Shoppers Purchasing Intention

Gargee Singh – 501068464

Supervisor – Dr. Tamer Abdou

29-May-2024

Abstract

Shoppers go online first in over 60% of shopping occasions (ThinkwithGoogle 2018, n.d.) and this constantly flowing stream of clicks holds immensely insightful data about how and at exactly what point during their web session, do customers decide to make the payment or exit the website or not complete the transaction. Naturally, this has led to companies wanting to harness the power of all this data, learn from it and make their offerings increasingly more lucrative. The combined power of e-commerce, logistics and social media has always intrigued me, and I chose Clickstream analytics for this project using the Online Shoppers Purchasing Intention Data Set (UC Irvine Machine Learning Repository, n.d.).

Using this sequential clickstream dataset, my project's primary focus will be to answer the question whether machine learning can be used to predict a customer's intention to make the purchase or not followed by the second question which is a customer's likelihood to leave the website without buying a product more commonly known as website abandonment rate. I will try to answer these questions and also try to establish whether or not there is a significant relationship between different pages a customer visits on the website. The combined information from the model should give us answers that can help predict a website's purchase conversion as well as abandonment ratio.

The data set I will be using has ten numerical and eight categorical attributes and although there are no null values, it is imbalanced. As initial pre-processing steps I will be working on methods (e.g., Over Sampling, Under Sampling or resampling with different ratios) to balance the data and check for correlation between numerical attributes which would need to be normalized before being used in the model. Once finalized I will proceed to feature selection methods to use only the most relevant features for the models. Once finalized with a balanced ratio between class labels

and relevant features the data set will be fed to the different models using the standard train to test ratios.

I propose to start with a classification model using the attribute ‘Revenue’ as the class Label to establish a decision tree of customer’s making the purchase or not. For comparison I also plan to run a random forest model as it is a stronger modeling technique and will help me limit overfitting as well as errors due to any biases in the data which I might overlook in the decision tree. For modeling I plan to use aggregate page views data along with session and user information present in the data. I also plan to explore Support Vector Machine learning models as well as Logistic Regression on my data set so as to have a few good comparisons to make.

In the end I will be using a common matrix to measure the scalability and performance of each model before recommendations about the best model that predicts whether the shopper will make the purchase or is merely scrolling through the shop in their current session.

Link to the Data Set, I downloaded the CSV file from this link –

https://archive.ics.uci.edu/dataset/468