Justin Ho, vunet-id: hojc, justin.c.ho@vanderbilt.edu

Vanderbilt Data Science Coding Challenge Fall 2024

Background:

Since the rise of e-commerce, online shopping has become a widely popular form of entertainment, on par with social media, offering consumers the convenience of browsing and purchasing from virtually anywhere. Whether from a smartphone, laptop, or tablet, shoppers can effortlessly explore countless online stores at the click of a button, whether at work, in transit, or during leisure time. Many people enjoy browsing through their favorite websites in search of appealing products, though a large percentage of these sessions often end without any purchases being made.

In contrast to in-person shopping, online businesses can leverage detailed user data gathered from website interactions and browsing behavior to predict whether a customer is likely to make a purchase during their session. Metrics such as bounce rate, exit rate, and the number of product page views, combined with outside factors like proximity to major holidays, can be analyzed to forecast the likelihood of a purchase and revenue for the ecommerce store. These insights offer businesses valuable tools to optimize website design, layout, and interactivity, enhancing the overall shopping experience for the user and increasing the chances of the ecommerce store earning revenue from completed transactions.

Problem Statement:

Is a machine learning model able to be built in order to predict whether a user will make a purchase during an online shopping session based on their browsing behavior and external factors such as time of year. The goal of the model is to identify key features from a dataset that can be used to estimate the likelihood of a transaction taking place during the online shopping session in order to create an accurate and precise model with high recall.

Hypothesis:

A machine learning model can be built to solve the binary classification problem of predicting the likelihood of a purchase taking place based on data gathered from ecommerce websites and users' browsers.

Methods:

To train and test a machine learning model that can solve the aforementioned problem, a dataset from UC Irvine's Machine Learning Repository titled Online Shoppers Purchasing Intention Dataset was investigated. The dataset consists of several variables that detail the type of browsing done by a user during their online session including what types of pages they visited and for how long, how often they bounced between pages and exited a page without returning, the closeness in time the user visited a website before completing a transaction, the proximity of time towards a special holiday, whether or not a user is a new or returning visitor to the site, and the time of year the site is being visited. The variables of these values are all used by the machine learning model to predict the revenue variable which is a binary classification of whether or not a transaction occurred (1 = successful transaction).

First, the dataset was loaded in from a csv file and used to create a dataframe through the pandas python library. The data was then cleaned by dropping several undesired variables such as page visitation statistics for non-product related sites, information on the computer and browser being used, region, month, and traffic type. Since the dataset was over 10,000 entries, a random sample of around 5,000 data entries was taken to be assigned as the data frame being examined. The data for the visitor type variable was then converted into numeric value by converting the string 'Returning_Visitor' to the value 1 and the string 'New_Visitor' to 0. This allowed for data standardization and ease of fitting the data to the machine learning model. Finally, the data frame was standardized using the StandardScaler from Scikit's machine learning library.

A random forest model was used because of the complex relationship between the variables used to determine whether a transaction occurs. The cleaned data frame consists of nine variables that are used to determine the output of the binary classification of the revenue field. The complex relationship between the variables can be examined using a random forest model because the random forest is made up of an ensemble of decision trees, each of which examine a random set of features to predict the transaction outcome. The random forest model is better than logistic regression for this problem because it adds in an extra layer of randomness due to the random features examined by each tree, and it allows for feature importance to be easily determined as not every variable will be examined by each tree. Together, the decision trees use majority vote to determine the classification of the outcome.

To create the model, the data was split into 80% training and 20% testing data, and subsequently split into data versus results by creating a target data frame containing just the data of the revenue variable. The training data was then fit to the random forest classifier.

Results and Discussion:

The accuracy, precision, and recall of the model were then evaluated by predicting the results of the testing data and comparing them to the actual testing outcome. The accuracy of the model was 0.9027355623100304, precision was 0.5611510791366906, and recall was 0.6902654867256637. The accuracy of the model was very high indicating that the random forest model had a high percentage of correct precisions.

Furthermore, the hyperparameters of the random forest, number of estimators and max tree depth, were explored to fine tune the model. The hyperparameters that gave the best random forest model were found to be 270 estimators and a max depth of 5. The improved model was then evaluated again and gave the metrics, accuracy: 0.9078014184397163, precision: 0.7068965517241379, and recall: 0.5899280575539568. The accuracy and precision of the improved random forest are slightly higher than the initial model, however, this came at the sacrifice of a slightly worse recall score.

The evaluation metrics indicate that the random forest model was successful in predicting whether or not the data gathered from a user's browser and websites visited could be used to determine if a purchase would occur. However, upon examining the confusion matrix as seen in the notebook, The model is much better at predicting when a purchase will not be made rather than when a purchase is going to be made, While the model predicts more purchases made than false purchases, it is still not nearly as accurate as it is when predicting non-purchase visits. This could be because of some hidden layers that are not being evaluated within the data variables, thus, this problem may require further exploration using more advanced models such as a neural network.

Resources:

6.3. Preprocessing Data. scikit. (n.d.-a). https://scikit-learn.org/stable/modules/preprocessing.html

Ibm. (2024, October 2). *What is Random Forest?*. IBM. https://www.ibm.com/topics/random-forest

Karabiber, F. (n.d.). *Binary classification*. Learn Data Science - Tutorials, Books, Courses, and More. https://www.learndatasci.com/glossary/binary-classification/

Normalize. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html#sklearn.preprocessing.normalize

Online shoppers purchasing intention dataset. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Pandas.dataframe#. pandas.DataFrame - pandas 2.2.3 documentation. (n.d.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Shafi, A. (2024, October 1). *Random Forest classification with Scikit-Learn*. DataCamp. https://www.datacamp.com/tutorial/random-forests-classifier-python