Online Shoppers Purchasing Intention Prediction

Midterm Report for DATA1030, Fall 2021

Keying Gong

Github Repository: https://github.com/keying-go/DATA1030_Project

1. Introduction

Online shopping has changed the traditional way of shopping and became the main distribution channel. Thus, businesses need to understand customers' behavior patterns to improve the shopping experience, customize promotions, and ultimately increase sales revenue.

Just like in brick-and-mortar stores, salespersons may speculate customers' purchasing intentions based on their past experience and interactions with the customers, we can predict customers' purchasing intentions based on their behaviors on shopping websites as well.

Specifically, knowing the main features that influence the final decision of purchasing is helpful to detect customers with a large potential to buy and conduct more efficient marketing.

The dataset contains 12,330 shopping sessions from a bookstore site, each belonging to a different customer in 1 year to avoid influence from campaigns, holidays, and user types.

For each session, 10 continuous variables and 8 categorical variables are recorded. The goal is to build a classification model using the 17 variables to predict whether the session ends up with a purchase or not.

The 8 categorical variables include a "Special Day", which indicates the closeness to a special day (e.g. Valentine's day) in which the order is more likely to be finalized. Special Day has 6 unique values: 0.0, 0.2, 0.4, 0.6, 0.8, 0.1. For example, for Valentine's Day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8 (1 Sakar).

"Traffic Type" represents the source that had referred the customer to the website (e.g. banner, SMS).

"Operating System", "Browser", "Region" and "Traffic Type" are categorical variables represented by integers. However, the original dataset did not state the label associated with each number, probably due to sensitive data. This paper will state conclusions involving these variables with type numbers. With information about the original label, we can easily apply the conclusions.

Other data types are well-documented in the original dataset.

Using this dataset, Sakar and his team have concluded that the clickstream data during the visit convey important information for online purchasing intention prediction (Sakar).

Baati and the team achieved the highest accuracy of 86.78% and an F1 score of 0.60 using a random forest classifier, compared to other classifiers (Baati).

2. Exploratory Data Analysis

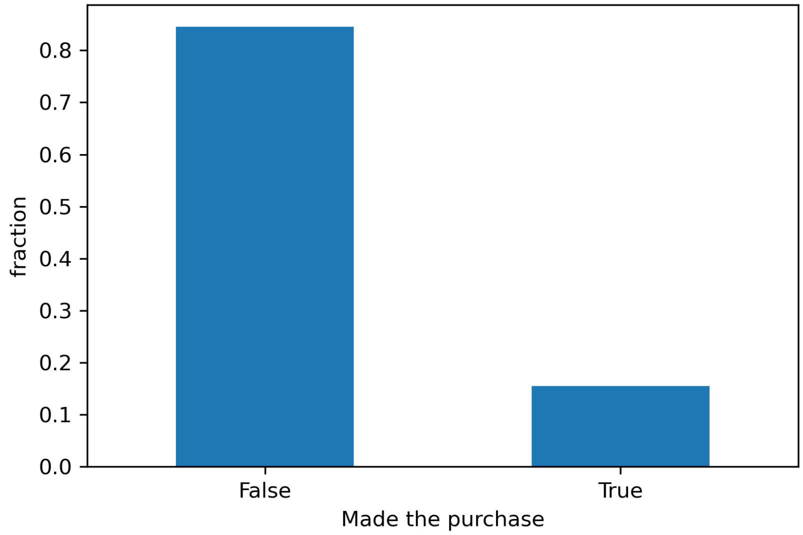

Figure 1 Of the 12,330 total sessions in the dataset, 84.5% (10,422) sessions did not generate revenues, and the remaining 15.5% of sessions ended up with a purchase. The target variable is imbalanced, with most of the sessions being False. Thus, we need to stratify Also, precision and f1 score should also be considered in addition to the accuracy metric.

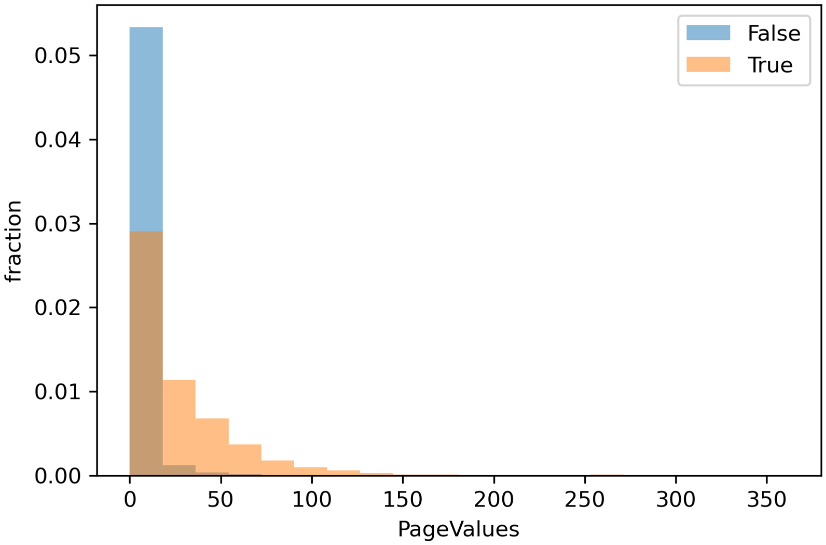

Figure 2 This histogram shows the distribution of the average value for the web page the user visited before completing the transaction. Both distributions are not normally distributed, and most of them are equal to 0. However, sessions that generate revenues have page values larger than 0, and the distribution is skewed to the left.

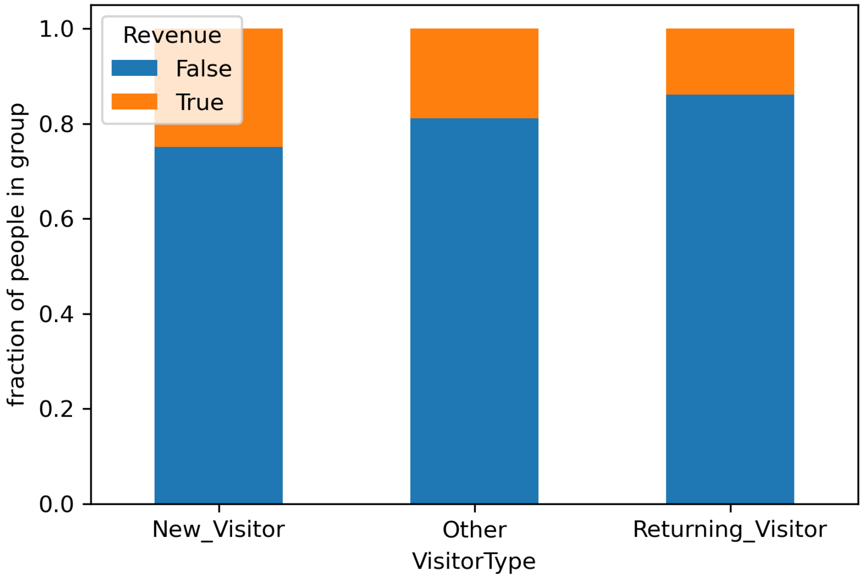

Figure 3 This figure displays the stacked barplot of visitor types with respect to whether there is revenue generated or not. Compared to returning visitors, new visitors have larger fractions of people who made the final purchase. Of the 12330 sessions, there are only 85 “Other” visitors, probably due to missing data.

3. Method

3.1 Data Splitting and Preprocessing

Since the target variable is imbalanced, stratified splits are used to split the data. First, use a stratified train-test split to allocate 20% of the observations to the testing group. For the remaining 80% observations, use 5-fold cross-validation to split them into training and validation groups. Since the data are collected from different customers in 1 year, the dataset is independent and identically distributed(iid), and has no group features and time series. The goal for this classification is to predict a new visitor’s purchasing intention (not in the dataset) in the future, and the data have no group structure on users, so it should be applicable.

In each instance of cross-validation, use a pipeline to preprocess the data. Within the pipeline, use StandardScalar to preprocess the continuous variables, since all of them follow a tailed distribution. Because all of the categorical variables cannot be ranked or ordered, use OneHot Encoder to preprocess all the categorical variables. After preprocessing, the data have 79 features.

3.2 Model Selection

After splitting and preprocessing the data in the pipeline, 5 Machine Learning algorithms are trained for comparison: Logistic regression with L2 regularization, a Random Forest Classifier, a Support Vector Machine Classifier, a K Nearest Neighbors Classifier, and an XGboost Classifier. All the algorithms’ hyperparameters were tuned using grid search cross-validation to loop through all the combinations of the parameters listed in the table.

Algorithms	Parameters
Logistic regression with L2 regularization	C: 0.01, 0.1, 1, 10, 100, 1000, 10000
Random Forest	max_depth: 3, 10, 30, 50, 79 max_features: 0.25, 0.5,0.75,1.0
Support Vector Machine	gamma:0.01, 0.1, 1, 10 C: 0.01, 0.1, 1, 10
KNN	n_neighbors: 1, 10, 20, 30, 40, 50, 60, 70, 100 weights: 'uniform', 'distance'
XGBoost	max_depth: 1, 10, 30, 50

Figure 4 Parameters used for tuning

Since the dataset is imbalanced, I choose the f1 score as the evaluation metric to select the optimal parameters. Run the algorithm 5 times with 5 different random states to measure the uncertainties due to splitting. For the indeterministic algorithms - random forest and XGBoost, the pipeline also specifies the random states in each run.

After tuning the parameters, the mean test f1scores and standard deviation of the test scores are calculated with the best parameters to compare the algorithms' performance. The baseline f1 score by predicting all data as the majority class - no revenue is 0, since the true positive would be 0. All the algorithms outperform the baseline f1 score. Random Forest has the highest f1 score, with a mean of 0.6620 and a standard deviation of 0.0045. Thus, the Random Forest classifier is chosen as the final model.

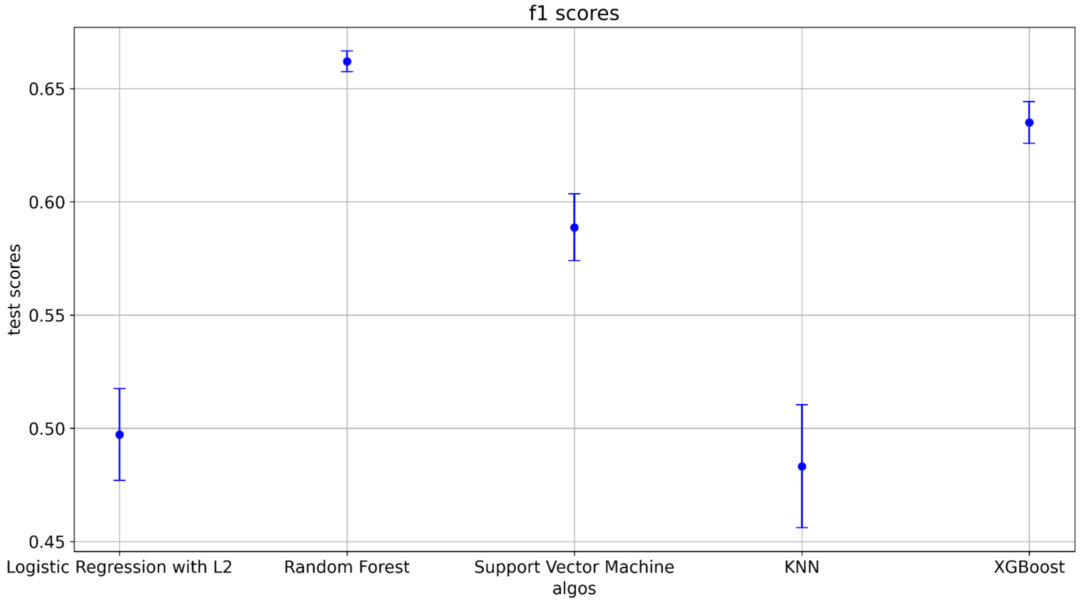

Figure 5 F1 score of each model

3.3 Final Model Formation

Reviewing the results from GridSearchCV, we pick the best hyper-parameters: max_depth = 10 and max_features = 0.5 for our choice of the model - Random Forest. Then, we trained the random forest model with 25 different random states. We allocate 80% of the data to the training set and 20% to the testing set. Record the model, f1 score, accuracy score, baseline accuracy, and test sets for each random state.

4. Results

4.1 Evaluation of Model

For the 25 random states, we have an average baseline accuracy of 0.8439 with a standard deviation of 0.0067. The trained models have an average accuracy of 0.9035 with a standard deviation of 0.0049. The trained models have an accuracy that is 6% percent higher than the baseline, which is 8.9 standard deviations above the baseline.

The f1 baseline score is 0 since it failed to tell the true positive. The trained models give an average f1 score of 0.6572 with a standard deviation of 0.0155. Based on the f1 scores and accuracy scores, the trained Random Forest model has a good prediction performance.

4.2 Interpretation of Findings

Random Forest is a rather complex model. To interpret our results, we calculate the global feature importances and local feature importances.

Global feature importances are calculated based on both the mean decrease in impurity and the permutation feature importances (with and without encoding categorical data).

First, global feature importances for each of the 79 preprocessed features are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. The top 10 important features are shown below.

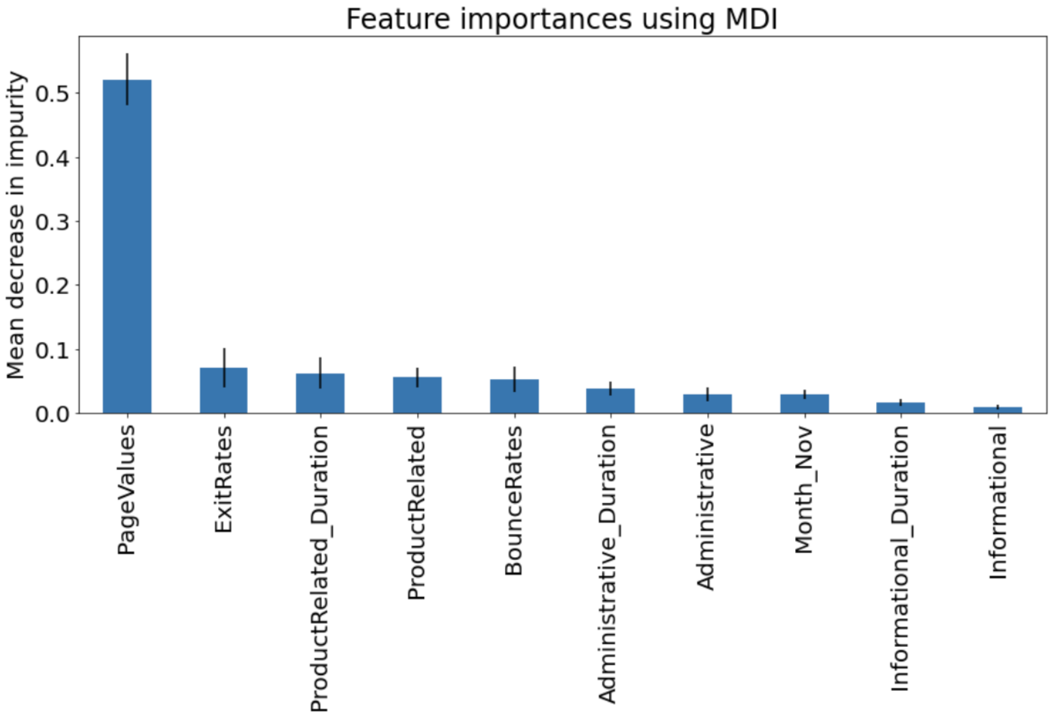

Figure 6 Global feature importances using the mean decrease in impurity

Next, we calculate the feature importance using permutations - calculate importance based on the decrease in accuracy with each feature left out.

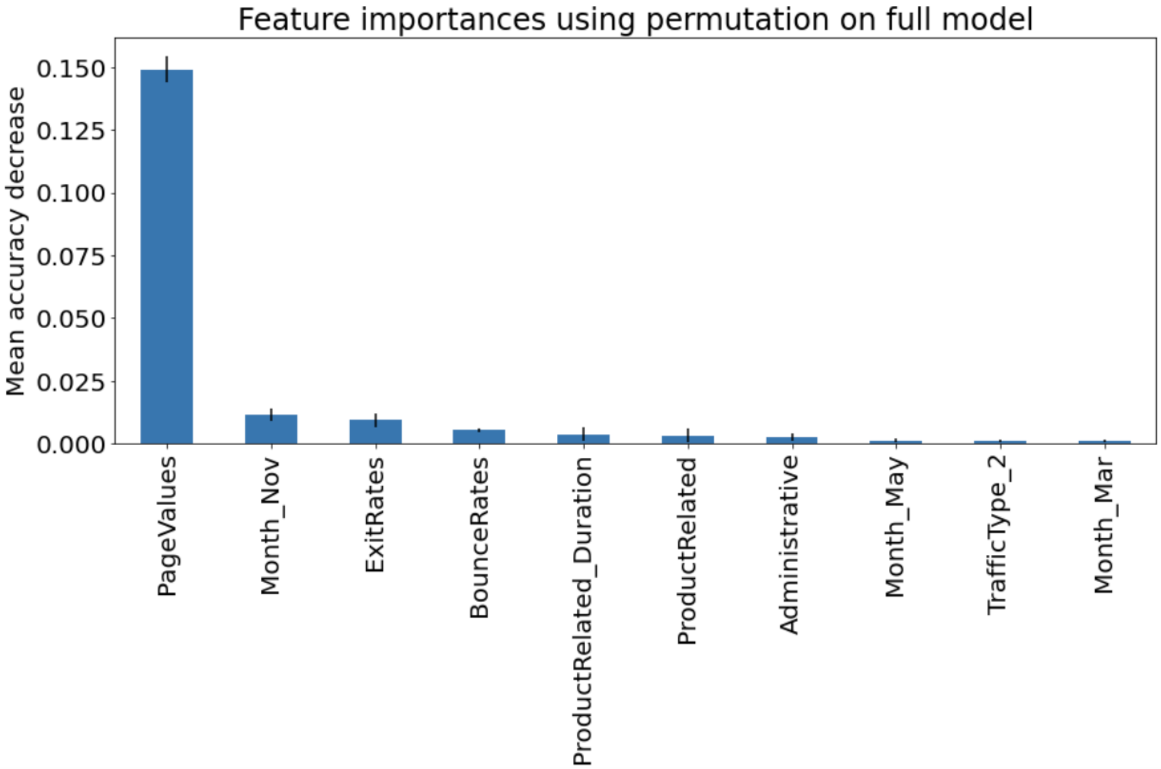

Figure 7 Global feature importance using permutations (preprocessed data)

We can see that in both methods, “PageValues” is the most predictive feature, and the top 10 features have a lot in common.

We also performed permutations on the test set using the unprocessed features by shuffling the values for each variable iteratively. The model has the lowest with shuffled values in "PageValues". We arrive at the same conclusion that "PageValues" is the most predictive feature. "Month" as a combined categorical feature that contains values ranging from Jan to Dec, also has high predictive power.

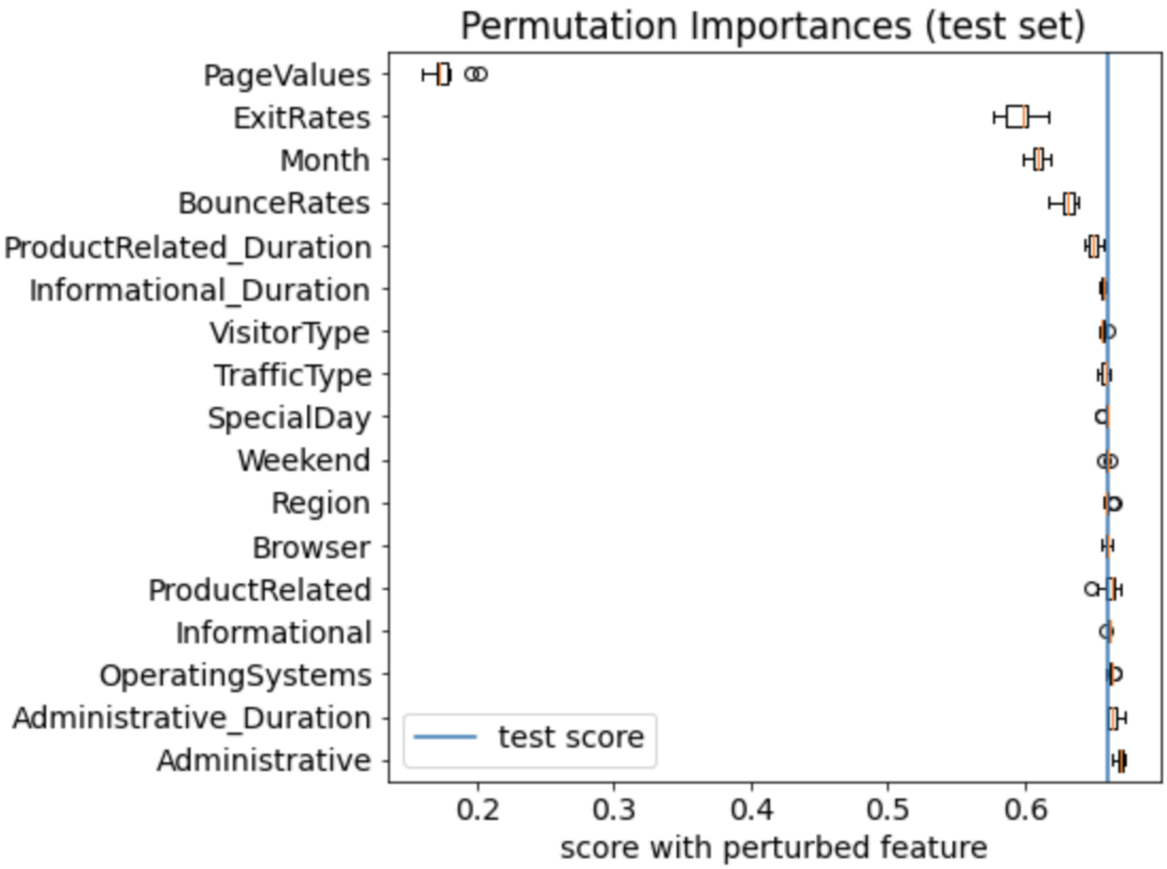

Figure 8 Global feature importances using permutations (unprocessed data)

Local feature importances help us to interpret the features' importance for each data point. We calculate the SHAP values for local feature importance. The SHAP force plot shows how each feature affects the prediction result for each session.

For example, for the 10th session, the “PageValues” has the highest power to increase the likelihood of purchase and “ProductRelated_Duration” has the largest power to decrease the likelihood of purchase.

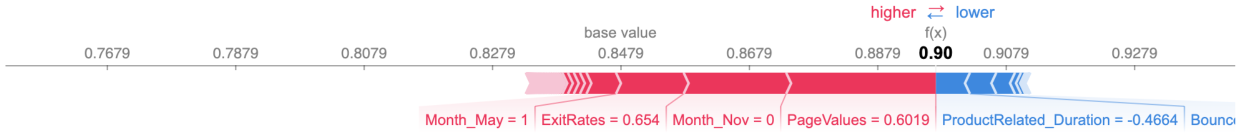

Figure 9 SHAP force plot for 10th data

For the 1st session, the value we get from Page Values has the largest power to decrease the likelihood of purchase and the time spent on Informational pages has the largest effect on increasing the likelihood of purchase.

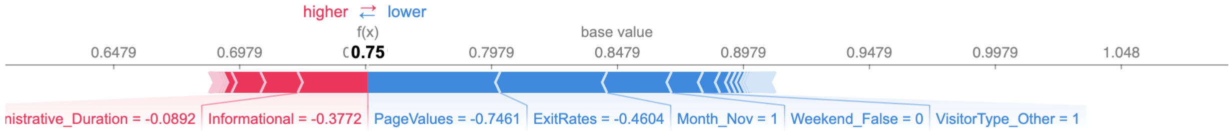

Figure 10 SHAP force plot for 1st data

5. Outlook

From the interpretation of the Random Forest model, "PageValues" is the most important factor in predicting whether the customer will make the final purchase. Thus, we can target the customers based on the page values of their browsing session, and conduct marketing to those who have strong intentions to purchase. We also see that the month "November" has strong predictive power. This is probably due to the holiday season and BlackFriday campaigns.

With better computation capacity, we may further improve the performance of the XGBoost model by having more values for tuning the hyper-parameters. XGBoost Classifier might give a better result as a gradient boosted decision tree model.

References

Data Source: UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data Set, archive.ics.uci.edu/ml/datasets/Online Shoppers Purchasing Intention Dataset.

Baati, Karim, and Mouad Mohsil. "Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest." Edited by Ilias Maglogiannis et al., *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I*, U.S. National Library of Medicine, 6 May 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7256375/#idm140554427882944aff-info.

Sakar, C. Okan, et al. "Real-Time Prediction of Online Shoppers' Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks." *Neural Computing and Applications*, Springer London, 9 May 2018, link.springer.com/article/10.1007/s00521-018-3523-0.