Predicting Online Shoppers *Purchasing* Intention

Prepared by
Md Rahat Mahmud

1. Task

We used the binary classification data-set based on online shoppers purchasing intention. This problem seemed to be significantly challenging with both categorical and numerical features. It gave us hands on tasks like normalizing/standardizing the data, multidimensional analysis, encoding and so on. We analyzed the customers behavior by focusing on different types of pages visited by them in specific sessions and the total time spent in each of these categories. Our goal was to identify customers shopping pattern and predict their future intention by basically targeting the **revenue** feature. We also investigated whether other features might be effective to the decision. Details about the other features and the overall data-set was presented at the **Dataset and Metric** section.

2. Approach

We used Python code basically for the coding part on the **Jupyter notebook**.

Libraries will be used are **sklearn, pandas, numpy** and so on. For visualization, we will use **Matplotlib library** and **Tableau software** for showcasing the overall result.

Normalizing and predictive accuracy are important part of the analysis. The classification problem dataset is challenging enough to be used different algorithms and compare their accuracies.

I am planning to apply the algorithms and tools I am learning in my class; INFS-762. They are:

- KNN Classifier (K- Nearest Neighbors)
- Decision Tree Classifier
- Logistic Regression
- Artificial Neural Network (Multilayered Perceptron Classifier)
- SVM (Support Vector Machine Classifier)

Description of the tools:

1. K-NN Model:

We analyzed and tried to predict future behavior based on the following Euclidian formula:
For k=1, here the nearest neighbors from the two different points x and y. d is the distance between x and y

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Basically, the python code was implemented for at the Jupyter Notebook.

2. Decision Tree Model:

We applied Entropy and Information Gain for finding the leaf node in order to reach decisions and Forecast.

Entropy = $-\sum_i p_i \log_2 p_i$

Information Gain = $-\sum_{v \in A} \frac{T_v}{T} Entropy(T_v)$

(A = variable, T = Target, and v = unique value in variable (A))

We are also planning to apply **Random Forrest** and other techniques which can create multiple decision tree model for better predictions.

3. Artificial Neural Network:

We used this deep learning model. It uses techniques to weigh respective variables in the data set like **Back Propagation** and **Feed Forward** into multiple hidden layers for the prediction. We used **sigmoid** function for the final prediction stage.

4. SVM (Support Vector Machine Classifier):

SVM tried to find a hyperplane with the maximum margin among the support vectors.
We fitted the SVM model at the function to get a result

3. Dataset and Metric

Dataset: Online Shoppers Purchasing Intention

Features of the Dataset:

Number of Instances: 12330
Target attribute: We will use the 'Revenue' attribute as the target attribute.
Number of Attributes: 18
Associated Tasks: Classification, Clustering
Dataset Characteristics: Multivariate
Last Updated: 8/31/2018

Challenges:

- The dataset demanded hands on tasks like normalizing/standardizing the data, multidimensional analysis, and encoding.
- The text variables will need to be weighted into integer values. (including the binary values)

Limitations:

- This dataset had no missing value. So, we did not have the chance to work on handling the missing values.
- Small size: The dataset is pretty small; 12330 instances. It might affect the overall analysis.

4. Results

After analyzing the models, we have seen that for the KNN model, the test set accuracy is 0.85. For the logistic Regression model, the test set accuracy is 0.87. For the Decision Tree model, the test set accuracy is 0.88. For the NNET model, the test set accuracy is 0.87, and for the SVM model, the test set accuracy is 0.83. So, we see that the NNET model is the best model for the dataset as it offers the most accurate test data set.