University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 - Fall 2021

Assignment 1: Online Shoppers Purchasing Intention

TOTAL MARKS 100

For this assignment, please use the Online Shoppers Purchasing Intention dataset from the UCI Machine Learning Repository. (The direct link is located at https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#.)

Note that the dataset was used in this publication: Sakar, C.O., Polat, S.O., Katircioglu, M. *et al.*, Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computation and Applications,* **31,** 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0. You are invited to consult this paper to obtain a deeper understanding of the domain and the descriptions of the features. You will also notice that the algorithms used in the original paper are different from the ones used in this assignment.

Our aim in this assignment is twofold: First, to construct models against this data and, second, to contrast the results we obtain against those reported in Section 4.1 of the reference paper.

Instruction:

1. This is an individual assignment. Submit your assignment using BrightSpace, before the due date.
2. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that, if you choose to use GitHub, the date and time of last change to your repository should be **before** the assignment deadline.
3. All students will be required to interactively demonstrate their assignments, as will be scheduled by the corrector Farnaz Sadeghi (fsade079@uottawa.ca), using the Zoom platform.
4. All students are **required to turn their cameras on during these demonstrations, to enable us to verify your identity**. Please have your student card and/or a photo id (e.g., a driver's licence) ready for inspection.
5. Use Scikit-Learn to complete the assignment.

Topic: Supervised learning – Binary classification

The aim of this learning task is to predict whether a client will complete or abandon their purchase, i.e., the class label is **Revenue (True, False).** Note this dataset is imbalanced, and actually reflects the real-world, where most shoppers end up not purchasing anything.

Import the data into your machine learning environment. Next, construct models using the following four (4) types of algorithms: a single decision tree, a random forest (RF) learner, a support vector machine (SVM), and a k-nearest neighbor (k-NN) classifier. Similar to Section 4.1 of the reference paper, you should use the holdout method of evaluation, namely use 70% of the data for training, and 30% for testing. It is advisable to repeat this process at least ten (10) times, and to report the average values.

Please submit the following.

a. Show the data after feature transformation and feature selection. [**12 marks**]
   Feature transformation and feature selection are pre-processing steps followed before doing the actual machine learning.
   For this dataset, you should follow the same procedure as detailed in the reference paper to transform the data.
   For feature selection, use one (1) correlation-based feature selection process.
b. Show the code you wrote to construct the four (4) models. **[16 marks]**
c. Show the four (4) confusion matrices corresponding to the models and calculate the f1-scores. **[8 marks]**
d. Show the ROC Curves to contrast the four (4) models. **[5 marks]**

e. Rebalance the data by oversampling the minority class. Next, repeat steps (b) to (d) for the balanced data. [**12 marks**]
f. Rebalance the data by undersampling the majority class. Next, repeat steps (b) to (d) for the balanced data. [**12 marks**]

Submit a **report on BrightSpace**, detailing the following.

g. Discuss the results you obtained and the lessons you learned when analysing this data. Your summary should include a **decision** as to which one of the four (4) machine learning algorithms you would use as well as a **motivation** for your choice. In addition, your discussion should reflect on the results before and after **balancing** the classes. **[25 marks]**
h. Contrast the results you obtained during this assignment with those of the reference paper by Sakar et. al., notably the results as reported in Section 4.1 (Tables 3 to 8). Be sure to discuss and to motivate any differences in methodologies, and results, and to highlight similarities. **[10 marks]**