IAI5101: Foundations of Machine Learning for Engineers & Scientists

Winter 2023

Assignment 1

Submission Deadline: 17th Feb 2023 on Brightspace

Submission: Submit your Python Notebook. Only 1-2 in a group allowed.

Bay clinic is a medical centre in Brazil that operates with a unique mission of blending research and education with clinical and hospital care. The medical center has a huge head force of 25,000 employees, and as a result of the combined effort of those employees, the medical center has been able to handle approximately 3 million visits so far. In recent times, the hospital was incurring losses despite having the finest doctors available and not lacking scheduled appointments. To investigate the reason for the anomaly, a sample data dump of appointments ***medicalcentre.csv*** is hereby presented. The collected data provides information on the patient's age, gender, appointment date, various diseases, etc. To cut costs, predict if a patient will show up on the appointment day or not (i.e., predict the appointment status) by completing the following:

A. Data Wrangling & Feature Engineering (45 marks):

1. Prepare the data for downstream processes, e.g., deal with missing values, duplicates
2. Determine the frequency of distinct values in each feature set
3. Initialize a function to plot relevant features within the dataset to visualize for outliers
4. Count the frequency of negative Age feature observations, and remove them
5. The values within AwaitingTime are negative, transform them into positive values
6. ML algorithm requires the variables to be coded into its equivalent integer codes. Encode the string categorical values into an integer code
7. Separate the date features into date components
8. ML algorithms work best when the input data are scaled to a narrow range around zero. Rescale the age feature with a normalizing (e.g., *min_max normalization*) or standardization (e.g., *z_score standardization*) function.
9. Conduct variability comparison between features using a correlation matrix & drop correlated features

B. Model Development (20 marks):

1. Develop a Naïve Bayes classifier to predict the outcome of the test data using Python. The performance of the classifier should be evaluated by partitioning the dataset into a train dataset (70%) and test dataset (30%). Use the train dataset to build the Naïve Bayes and the test dataset to evaluate how well the model generalizes to future results.

C. Model Evaluation & Comparison (35 marks):

1. Write a Function to detect the model's Accuracy by applying the trained model on a testing dataset to find the predicted labels of Status. Was there overfitting? **(10 marks)**
2. Tune the model using GridSearchCV **(5 marks)**
3. Using the same data set partitioning method, evaluate the performance of a SVM and Decision tree classifier on the dataset. Compare the results of the Naïve Bayes classifier with SVM and Decision tree model according to the following criteria: Accuracy, Sensitivity, Specificity & F1 score. Identify the model that performed best and worst according to each criterion. **(10 marks)**
4. Carry out a ROC analysis to compare the performance of the Naïve Bayes, SVM model with the Decision Tree model. Plot the ROC graph of the models. **(10 marks)**