Advanced Machine Learning – Assignment 2

Individual Assignment

Read the Tahoe Healthcare Systems case carefully and then answer the following questions. (The data for the case can be found in the Tahoe Healthcare Data.csv file.) All of your answers below should be computed taking the dataset as representative of what will happen in a given year if nothing is done to reduce the readmissions rate.

(i) Suppose we consider the status quo and don't implement CareTracker. How much will this cost Tahoe due to the loss in Medicare reimbursements? You can assume that (as stated in the case) the loss would be \$8,000 per re-admitted patient. (Hint: Your answer should be \$7,984,000 but you should explain where this comes from!)

(ii) Suppose CareTracker was implemented for all AMI patients. What would be the net change in cost (over the status quo) from doing this? Should Tahoe implement CareTracker for all AMI patients?

(iii) Despite this seemingly bad news it may be possible to apply CareTracker to the subset of the AMI patient population who are most at risk (of readmission) and save money in the process. To investigate this, Tahoe want to build a classification model to predict those AMI patients who will be admitted and those who won't. Before doing so, however, they want to estimate an upper bound on the possible savings they could make using such a classification model. Specifically, suppose Tahoe had perfect foresight regarding what patients will need to be readmitted. What savings could they make in this event?

(iv) One simple idea for a classification algorithm is to classify based on the severity.score variable since (and you can check this) the mean value of *severity.score* is considerably higher for readmitted patients (*readmit30* $= 1$) than it is for patients who were not readmitted (*readmit30* $= 0$). Consider then the simple classifier which predicts that a patient will be readmitted if the patient's severity score, S, satisfies $S > S^*$, for some fixed threshold S^*.

Write a piece of code that computes the cost savings (over the status quo) using the whole dataset (*Tahoe_Healthcare_Data.csv*) for values of S^* increasing from 25 to 100 in increments of 1. Create a graph of these estimated cost savings as a function of S^*. What is the best value (in terms of cost savings) for the threshold S^*?

(v) Classification algorithms based on *severity.score* alone are very crude, however, and Tahoe now expects that they can do better using other more sophisticated algorithms. In particular, they want to use logistic regression to fit a model using the entire dataset to estimate the probability of readmission. Write a piece of code to fit a logistic regression model. The goal is to predict *readmit30* as a function of the other covariates. Train your model on the train data (*train.csv*) and report the accuracy and the confusion matrix of your model on the test dataset (*test.csv*). Note that we don't employ any regularisation for the logistic regression model, so there is no need for hyper-parameter tuning using the validation data.