Competitive Project: Midterm Report

Dudley Irish

CS6350

October 25, 2024

I. Introduction

In this project we are participating in the Kaggle competition, ML2024F, Income level prediction for mortgage.

II. Experimental Method

The procedure is as specified by the Kaggle competition terms. There are two data files with training and test data. The training data is labeled and test data has an identifying number for each case. The procedure is to train your model on the training data then generate a submission using the test data. Details are available on the Kaggle website.

The data was prepared for use by converting all category fields to number ranges and replacing all continuous fields with binary thresholds. The thresholds were determined by taking the mode of the training data.

The models were constructed and run using a library Copyright (C) 2013 Quan Wang ¡wangq10@rpi.edu¿, Signal Analysis and Machine Perception Laboratory, Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. This library is written in C++ and can be called from either MATLAB or Octave. More information is available at Decision Tree and Decision Forest

III. Analysis

At this point in the project we have made two submissions:

(a) The model was a decision tree with 23 levels. The resulting score was 0.73315.

(b) The model was a decision forest with each tree having 23 levels and the forest having 10 trees. The resulting score was 0.73957.

So far the performance has been disappointing. There are a number of meta-parameters that could be adjusted to improve the performance but these initial numbers suggest that we will not use decision trees for our final submission.

IV. Conclusion

The plan for the project was to perform initial experiments with each machine learning technique through out the semester and the concentrate on the approach that seems to work best for this data. Our results from these first two submissions suggest that a decision tree/forest is not the best fit for this data.