Pre-Employment Report 1

Wyatt Jones

January 23, 2019

For this project the two models that I chose are a Random Forest classifier and an Ensemble of Multi Layer Perceptron Neural Networks. In order to process the data I imputed missing numerical data with the mean of that feature and for categorical data I created a new class for missing values. I also explored imputation with the median and most common class however, the results did not differ significantly. In the future, if the missing values are not missing randomly I would explore the impact of Multiple Imputation by Chained Equations (MICE) on model performance. Once the data was cleaned I rescaled the features so that they had zero mean and unit variance and then did a dimensionality reduction with PCA. In my experience, training machine learning models in my research has shown me that this preprocessing often leads to significant improvements in model performance.
I explored several alternative models before finally deciding to use a Random Forest classifier and an Ensemble of Multi Layer Perceptron Neural Networks. The Random Forest classifier is able to achieve near perfect performance on the training set as is shown in (Figure **??**). In order to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model I used cross validation with 5 folds. I used a validation curve and learning curve to examine the degree of overfitting and to determine the max number of features to use. I found that while the Random Forest Classifier performs very well on the training data and the 5-fold cross validation has a mean of 0.93702 with std 0.00188 which shows some evidence of over fitting.

	precision	recall	f1-score	support
0	1.0000	1.0000	1.0000	31912
1	1.0000	1.0000	1.0000	8088
micro avg	1.0000	1.0000	1.0000	40000
macro avg	1.0000	1.0000	1.0000	40000
weighted avg	1.0000	1.0000	1.0000	40000

Figure 1: Random Forest Classification Report

My second method I used was a Ensemble of Multi Layer Perceptron (MLP) Neural Network estimators. I first used a single MLP and found that I had good results and so to reduce variance I used a bagging of MLP networks with a higher regulation penalty. I found that the Ensemble of MLP networks has very good performance on the training data which can be seen in (Figure **??**) but not as good as the Random Forest classifier's results in

(Figure **??**). The MLP ensemble does significantly outperform the Random Forest classifier based on the 5-fold cross validation which for the Ensemble MLP has a mean of 0.99042 and std of 0.00077.

	precision	recall	f1-score	support
0	0.9975	0.9996	0.9985	31912
1	0.9986	0.9902	0.9944	8088
micro avg	0.9977	0.9977	0.9977	40000
macro avg	0.9980	0.9949	0.9965	40000
weighted avg	0.9977	0.9977	0.9977	40000

Figure 2: MLP Ensemble Classification Report

The Random Forest classifier was significantly faster to train as compared to the Ensemble MLP with training time taking 20 seconds and 13 minutes respectively. If decision makers are interested in a more interpretable model I would favor the Random Forest classifier since it is easy to modify the max depth of the decision trees which makes interpretation easy to understand. If there are large amounts of data that will be added continuously in the future I would favor the Ensemble MLP due to the ability to easily do online updates and since the MLP could be easily written in Tensorflow to fully make use of all GPU and CPU resources available. My significant experience working with Tensorflow also shown me the ease with which neural network models can be modified and added to as this model moved towards a production environment. Therefore, the model that I recommend is the Ensemble MLP due to the lower likelihood of overfitting, ease of performing online learning, ability to fully use all available computational resources, and the large degree of modularity available.