Lab 2 - Predictive Modeling, Stat 415, Spring 2023

Wing Yan Yau

12th May 2023

Starting with the census tracts data set, we first dropped the first 2 indexing columns, and split them into training and testing sets with the latter having size of one-third of the whole data set. The purpose of this report is to examine different ways in predicting the drunk driving rate of different states, using the given information in the data sets.

1 Linear Models

We first train a linear model using the training data. The mean square error of the validation is around 51.8. The distribution of the prediction error (predictive value - real value) is slightly left-skewed, indicating that the model is more likely to underestimate the drunk driving rate than to overestimate. Most of the distribution mass is centered around 0, showing a good hint of a working model.

We then try to improve the model by regularization, but the mean square error attains minimum around the choice of coefficient for the regularization being close to 0 i.e. regularization is not particularly helpful in this case. Quantitatively, the regulated linear models still have a mean square error around 51.8.

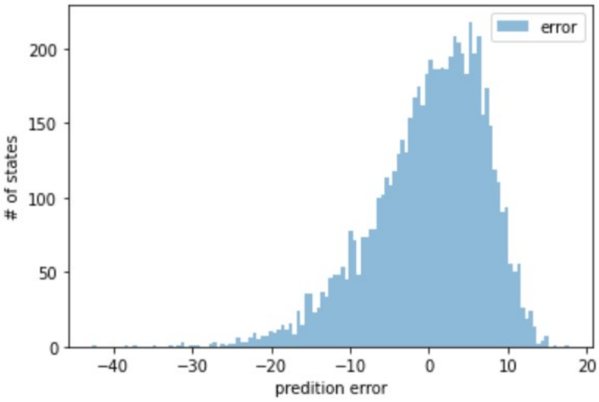

Figure 1: Histogram(distribution) of error of prediction made by linear model

We also look into the distribution of the drunk driving percentage in the data set as well as that of those made by the linear models. It is clear that the models tend to give a narrower range of output then the real values.

2 Random Forest

We then proceed to investigate in the performance of random forest on this data set. Different number of trees are tested, the resulting mean square errors seem to converge to around 48 as number of trees increases. Hence, random forest seems to slightly outperform linear model in this task.

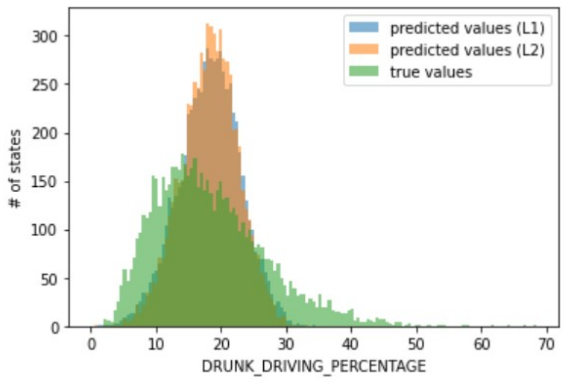

Figure 2: Histogram(distribution) of prediction outputs made by linear models, and of true values

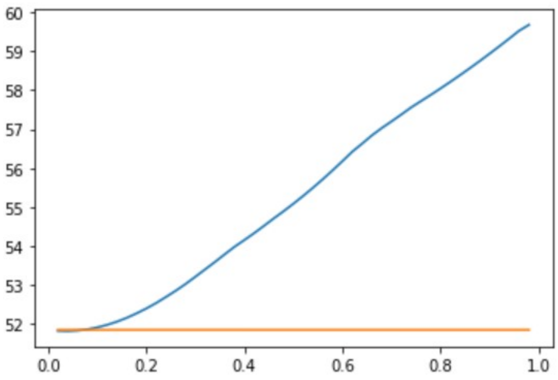

Figure 3: Line plot on how different coefficients for the regularization change the mse (y-axis) with blue: L2, orange:L1

n estimators	mse
10	104.1
20	52.9
30	50.4
40	49.3
50	48.8
60	48.6

Table 1: Table to recommendation coefficient for the 3 metrics

Like linear models as we have looked into in the previous section, the distribution of the predicted values by random forest model is more concentrated than that of the actual ones. One thing worth noticing is that, the right-skewness of the actual distribution is resembled in the predicted values distribution very clearly in this model.

In order to improve the performance of the random forest models, we attempt the tune the parameters of them, including number of trees, maximum tree depth, maximum leaf nodes and number of features to consider when looking for the best split. We get the best mean sqaure error as 48.3. The optimal setting output by performing grid search says setting higher limit of maximum leaf nodes and maximum depths give better performance. Combining with the former investigation on number of trees, we can conclude in this case, overfitting caused by too much tree nodes or too deep a tree is unlikely and 50 trees in a forest is rather ideal.

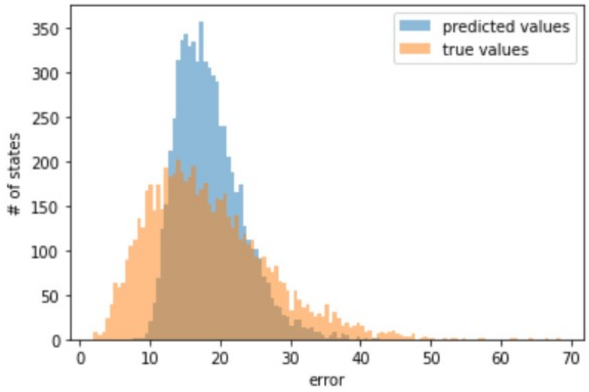

Figure 4: Histogram(distribution) of prediction outputs made by random forest, and of true values

3 Neural Networks

A 3 layer neural network with Sigmoid activation is trained on this data set. The mean square error is . A very undesirable situation ariased from the prediction, that is, all of the output converges to the mean of the output of the training data set. Retraining the network from scratch does not show any sign of hope, the problem is still there, with the output mean at 19.077, for all 3 times of retraining. We try to soothe it away by tuning the neural network by adjusting the optimizer, the number of layers in the net, and the activation functions in use. Both optimizer and layers number adjustments fail to show visible improvements, while the distribution of the predictions finally get wide enough of being like a spectrum instead of a line when plotted as histogram, after we use ReLU as the activation function. The mean square error is around 72.1 after such tuning.

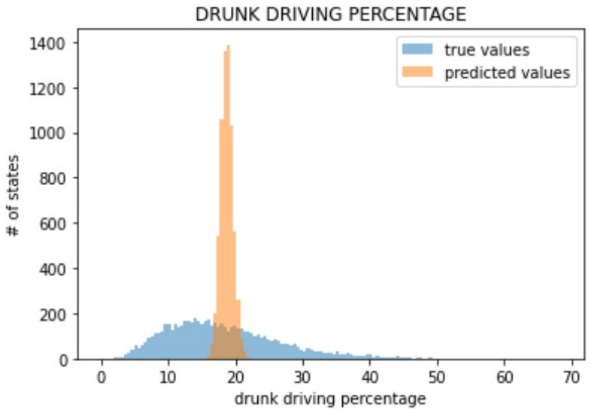

Figure 5: Histogram(distribution) of prediction outputs made by 3 layer sigmoid-activated neural net, and of true values

4 Transfer Learning

Turning our attention to the state data drunk driving data set in this section, we first try to make prediction using a linear model, a random forest and a neural net by training them on this data set. The mean square errors on the validation set of the data set are, 52.1, 48.9 and 142.

Then, we take a linear model and train it on the data from census tracts data set and use it to make predictions on the data from state data drunk driving data set, the mean square error is 4501 with one of the prediction being 264% which is non-sense. It seems like the model does not transfer.

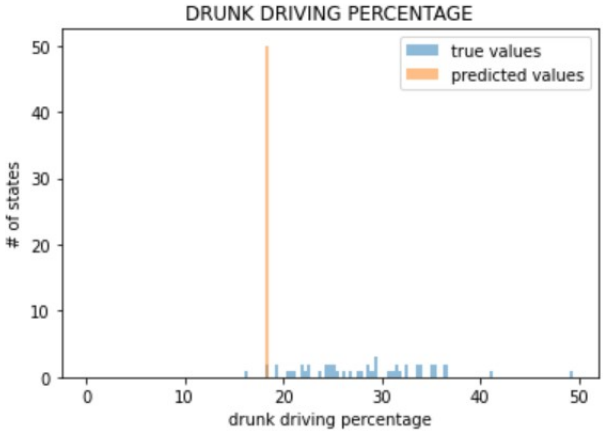

Figure 6: Histogram(distribution) of prediction outputs made by linear models, and of true values

We then try to achieve transfer via further training by fine tuning. The resulting mean square error drop to around 142, which is way better than the original 4501 but it is very similar to that of the net trained on only state data drunk driving data set.

5 Visualization

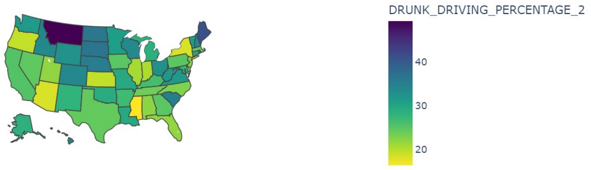

Figure 7: Choropleth map of the US States colored by the actual percentage of drunk driving accidents in that state

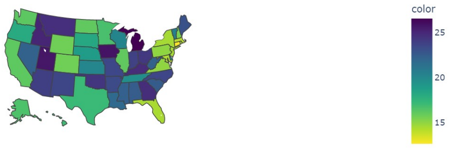

Figure 8: Choropleth map of the US States colored by the percentage of drunk driving accidents in that state, predicted by the above random forest model

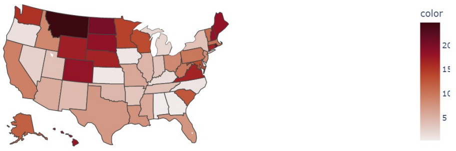

Figure 9: Choropleth map of the US States colored by the absolute value of error by the above random forest model

By observing the maps, we can see it is the most difficult to predict well the drunk driving rate for states with high drunk driving rate, for example Montana and Maine. A possible reason for this is the narrow range of output of the predictive model, and they are better in predicting value nearby the mean. While we can see from the first map of actual data, most of the states have drunk driving below 30%, hence those with value higher than that could be viewed as hard for models, proposed in this lab, to predict correctly.