415-hw3

May 12, 2023

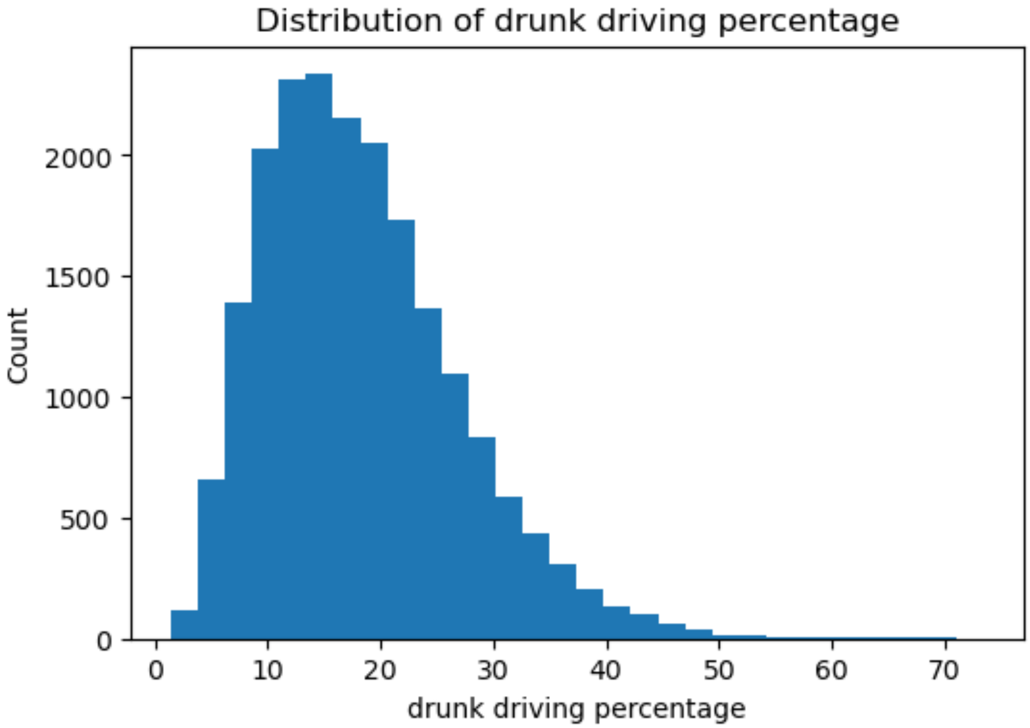
1 415-HW3

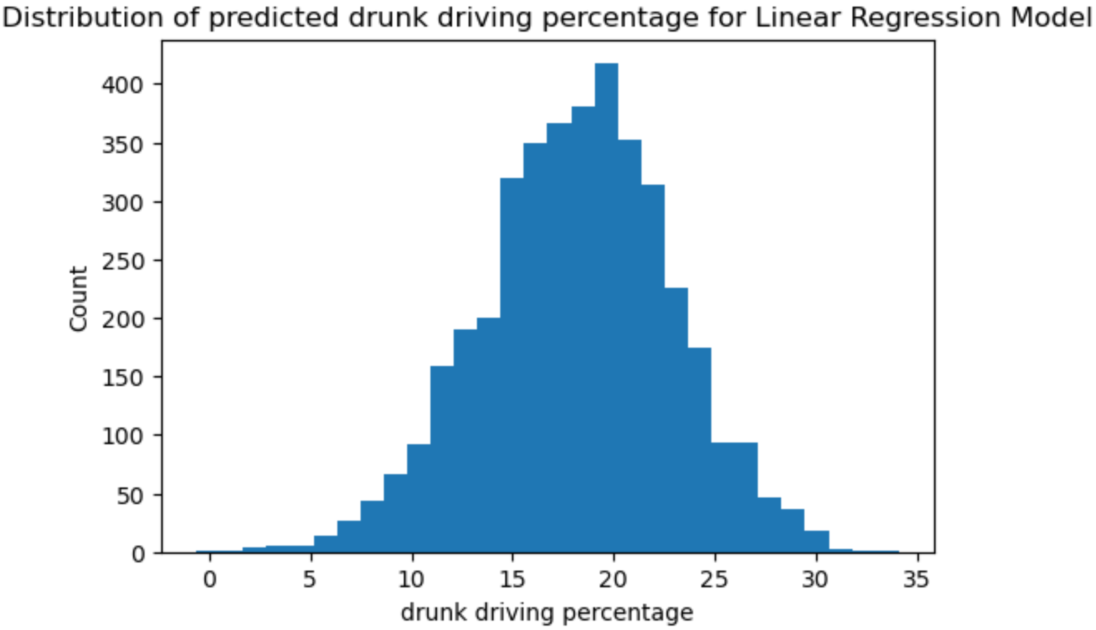
1.1 Shuo Han

1.2 Linear Models

1.2.1 1 Train a linear model

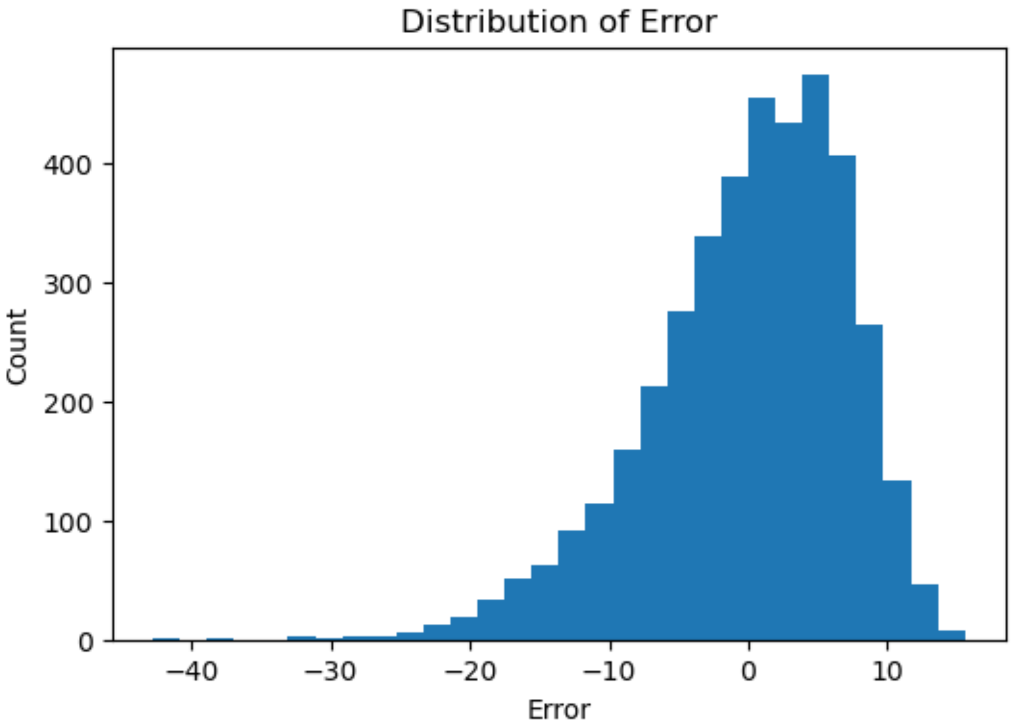
```
MSE: 51.1845162483201
```





By using test/train split, I have trained a linear model that takes all of the columns in the dataset and tries to predict the percentage of drunk driving accidents here. The histogram of the drunk driving percentage in the dataset is right skewed, so the majority of the clustered census tracts has lower drunk driving percentage, and there are relatively fewer higher drunk driving percentages. Also, there are some extreme values in the dataset that are significantly higher than the majority of the data, which affect the distribution. The histogram of the predictions made by the linear model for the test set is almost symmetrically and normally distributed, so the model's predictions are centered around the true values, and the model is making predictions with relatively low bias, but we still need further analysis.

1.2.2 2 histogram of the linear model errors



The histogram of the linear model errors is left skewed, so we can tell the majority of the errors are positive, meaning that the predicted values are lower than the actual values. Since the error is a skewed distribution, so the model may be systematically underestimating the target variable. However, based on the histogram of the predicted values above, we can tell that the model is making predictions that are centered around the true values, but is having a harder time predicting values that deviate from the mean.

1.2.3 3 Tune the linear model with L1 and L2 regularization

```
MSE (Lasso): 51.15978190188532

MSE (Ridge): 51.18451512341784
```

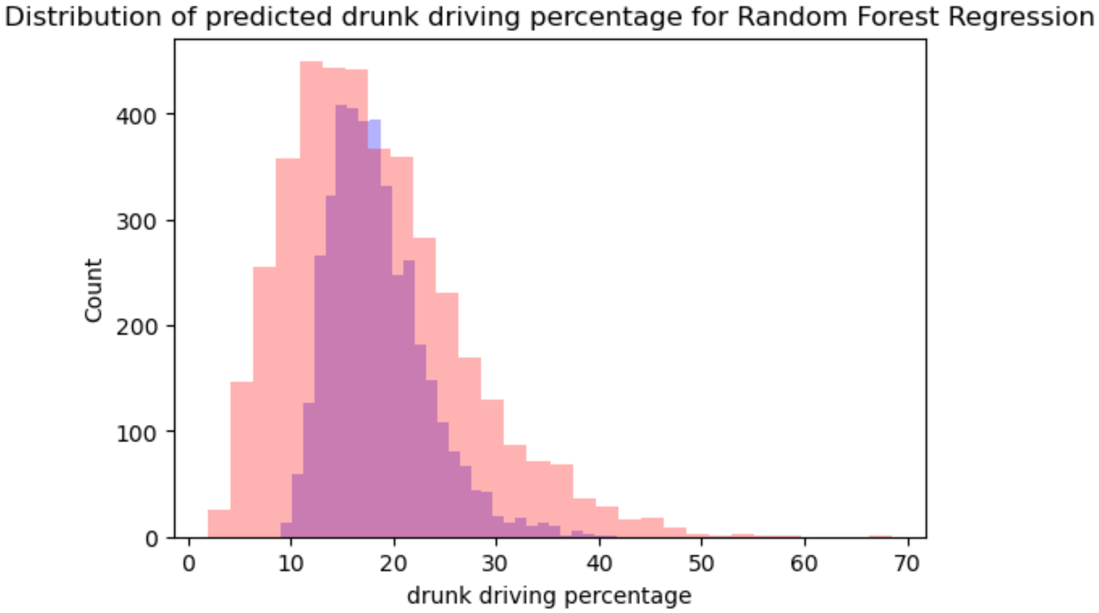
The MSE for the orginal linear model is 51.1845162483201, so the MSE improves some with L1 Penalty. The MSE improves only a little after L2 Penalty, almost no improvement.

1.3 Random Forrest

1.3.1 5 Train a random forrest model on the same dataset. Report the MSE.

```
MSE (Random Forest): 46.755260427392216
```

1.3.2 6 histogram of the forrest outputs vs the true data distribution



The histogram of the forrest outputs is right skewed, and the true data distribution is also right skewed, so they are of really close distribution as shown above and the forest model is able to capture the general shape and trend of the true data distribution.

1.3.3 7 Tune your random forrest

```
MSE (Random Forest): 46.410704587457786
```

I have used GridSearchCV() to tune my random forest here, and my RMSE has improved only a little after tuning. There are some parameters tested here: n_estimators that specifies the number of trees in the forest, max_depth that limits the maximum depth of each decision tree in the forest, min_samples_split that determines the minimum number of samples required to split an internal node, and max_features that determines the maximum number of features that can be used to split each internal node. However, the model does not improve a lot in the best one, so it seems there is no parameters really important.

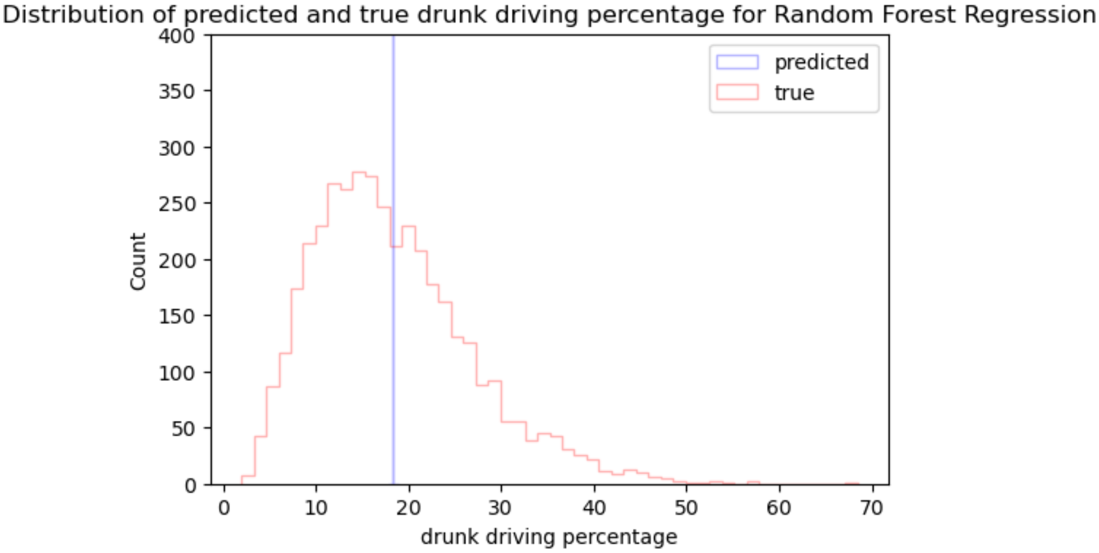
1.4 Neural Networks

1.4.1 8 train a 3 layer neural network

```
Mean Squared Error: 72.5758
```

1.4.2 9 distribution of outputs of the neural network

```
[9]: (0.0, 400.0)
```

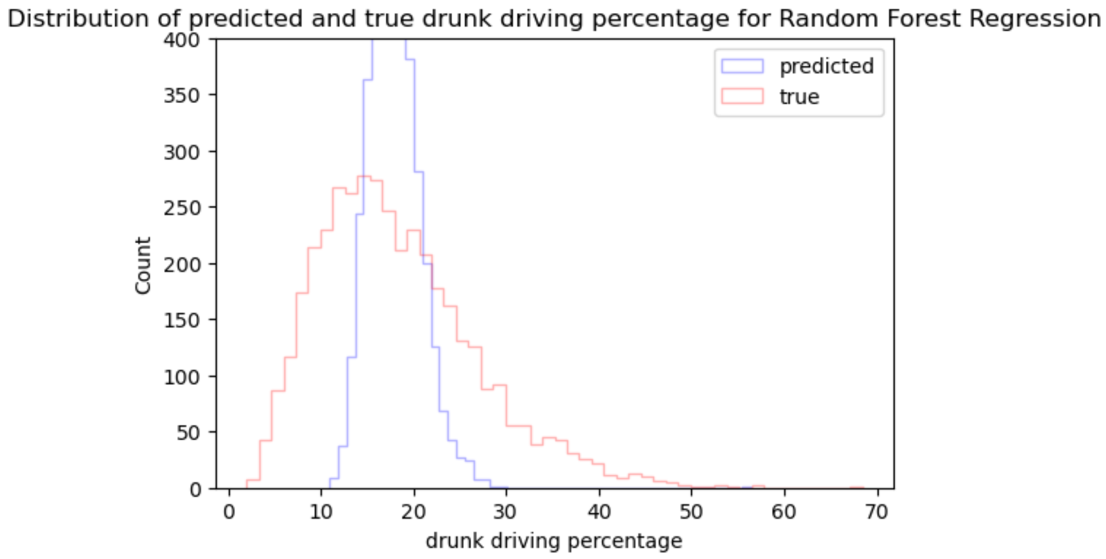


All the predicted values are around the mean of the true vlues in the test set, so the neural network does converge to the mean output. So I will further retrain the neural network in 10, and we can see the converge-to-mean problem is solved.

1.4.3 10 Tune the net by adjusting the optimizer, the number of layers in the net, and the activation functions

```
Mean Squared Error: 87.7312
```

```
[11]: (0.0, 400.0)
```



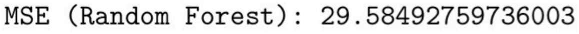
Here, we changed the optimizer to SGD with weight decay, added another hidden layer with ReLU activation, and added a dropout and regularization in my neural net. Right here we get a larger MSE compared to the original one, but the neural network does not converge to the mean output right now, so these help to improve.

1.5 Transfer Learning

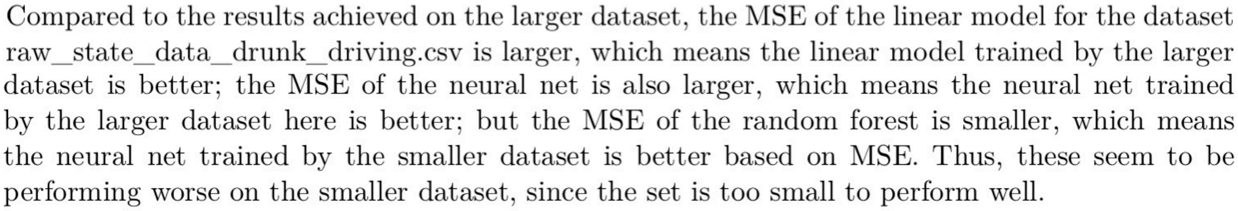
1.5.1 11

```
MSE (linear model): 219.36567142046178
```

```
Mean Squared Error for neural net: 104.2809
```



```
MSE (Random Forest): 29.58492759736003
```



Compared to the results achieved on the larger dataset, the MSE of the linear model for the dataset raw_state_data_drunk_driving.csv is larger, which means the linear model trained by the larger dataset is better; the MSE of the neural net is also larger, which means the neural net trained by the larger dataset here is better; but the MSE of the random forest is smaller, which means the neural net trained by the smaller dataset is better based on MSE. Thus, these seem to be performing worse on the smaller dataset, since the set is too small to perform well.

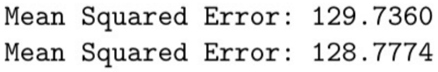
1.5.2 12

```
MSE on census-tracts-dataset.csv: 51.1845162483201

MSE on raw_state_data_drunk_driving.csv: 6694.327542791737
```

The MSE of the linear model trained on the census-tracts-dataset.csv data is 51.1845162483201 from question 1, and the MSE of the trained linear model to make predictions on the data from raw_state_data_drunk_driving.csv is 6694.327542791737, which is significantly higher, so it suggests that the trained linear model did not transfer well.

1.5.3 13 achieve some transfer via training

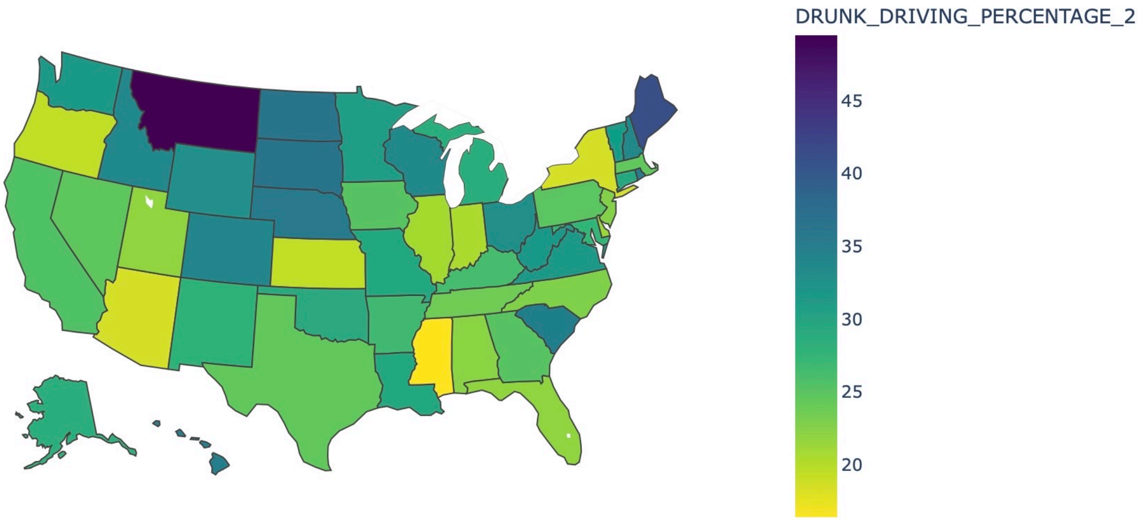


```
Mean Squared Error: 129.7360
Mean Squared Error: 128.7774
```

After fine-tuning, the MSE of neural network here get even worse in MSE when compared to the result from question 11, we can see that the MSE is signigicantly larger, so this does suggest that transfer learning does not help here.

1.6 Visualization

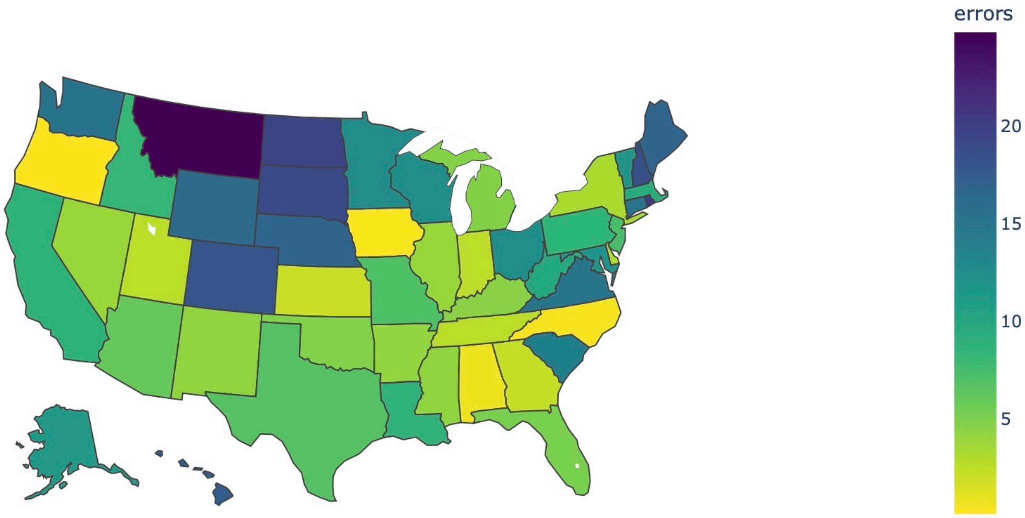
1.6.1 14



1.6.2 15

```
State with maximum error: MT
State with minimum error: OR
```

Using the random forest model for data census-tracts-dataset.csv after fine tuning, I have predicted the percentage of drunk driving accidents and the errors for these predictions. Based on the prediction error, we can see that, MT, Montana, is the easiest to predict with the maximum error; OR, Oregon, is the hardest to predict with the minimum error.


errors
20
15
10
5