HW2-3

Maedeh Karkhaneh Yousefi

February 25, 2022

I reckon a linear model, especially linear regression is the best model to fit the data. That's because the target is continuous. I used MinMaxScaler and OrdinalEncoder as a preprocessing procedure.

Part I
Regression

Linear Regression
The linear regression score is: 0.7296646657433239 (73 percent)

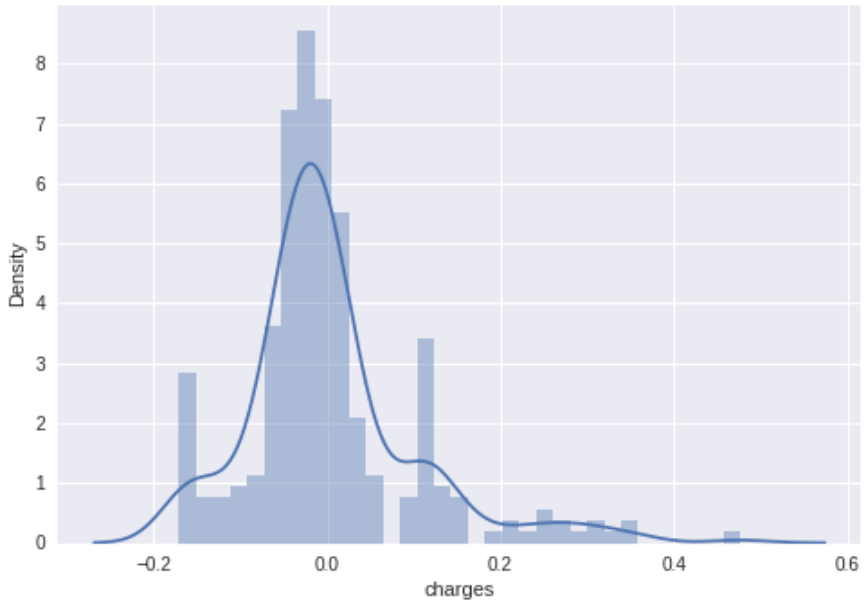

Figure 1: The distplot of prediction-ytest for charges. One can see 2 other peeks, which may show 2 more classes.

Linear Regression (Polynomial Features included)
The score is : 0.8199104895969911 (82 percent)

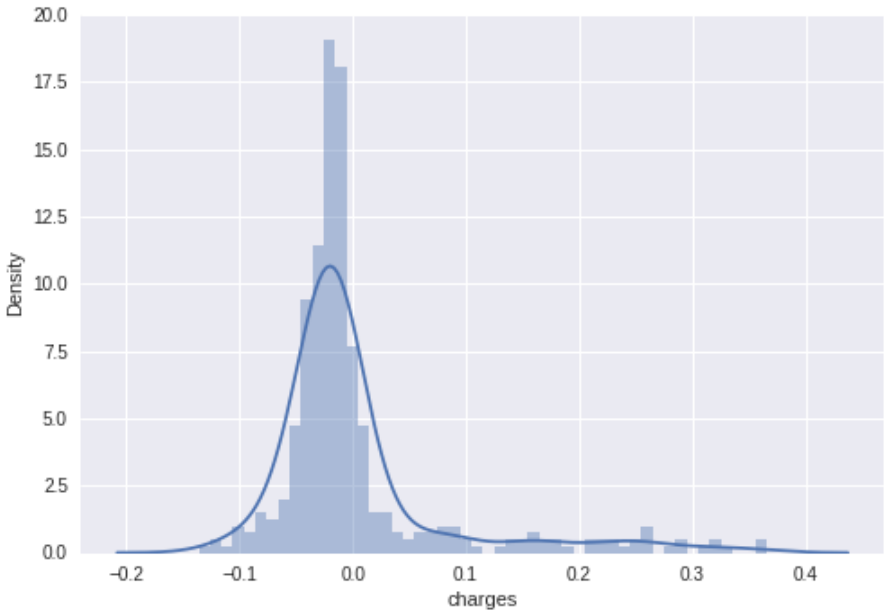

Figure 2: The distplot of prediction-ytest for charges. You can see that those other peaks are gone and most of the error is distributed over the range of -0.1 to 0.1.

SVR with Linear Kernel

The SVR (kernel = linear) is: 0.7110971727822234 (71 percent)

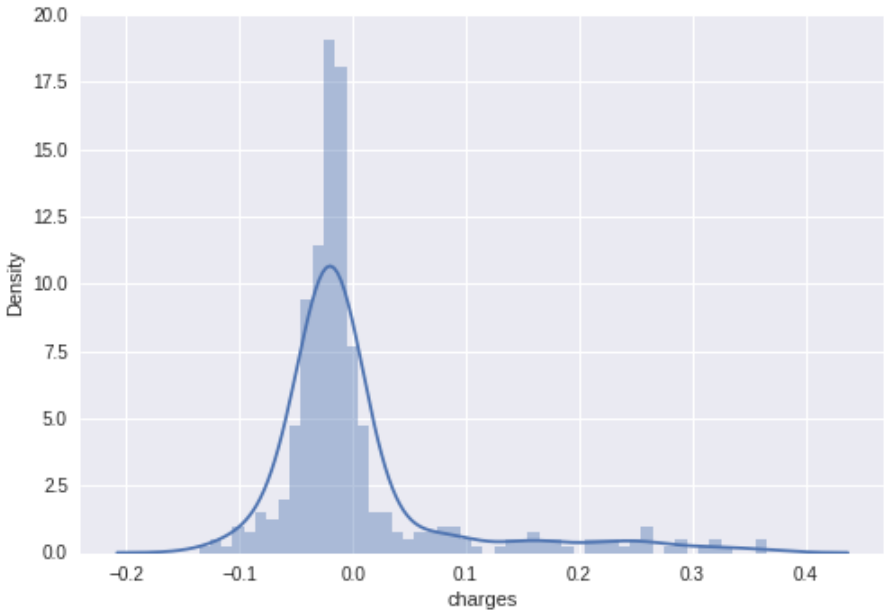

Figure 3: The distplot of prediction-ytest for charges. The results show a distribution of error between the range of -1 to -1, which was less for linear regression.

SVR with Poly Kernel

The SVR (kernel = poly) score is: 0.6842361729120802 (68 percent)

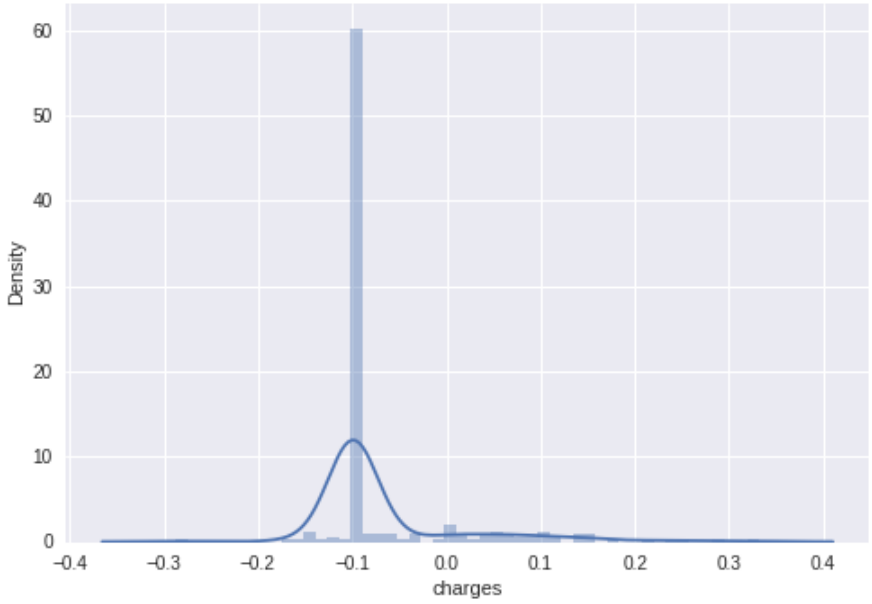

Figure 4: The distplot of prediction-ytest for charges. There is an obvious peak at -0.1, which shows that most of the data have an error of 0.1.

Random Forest Regressor
The score is : 0.8262341327080196 (83 percent)

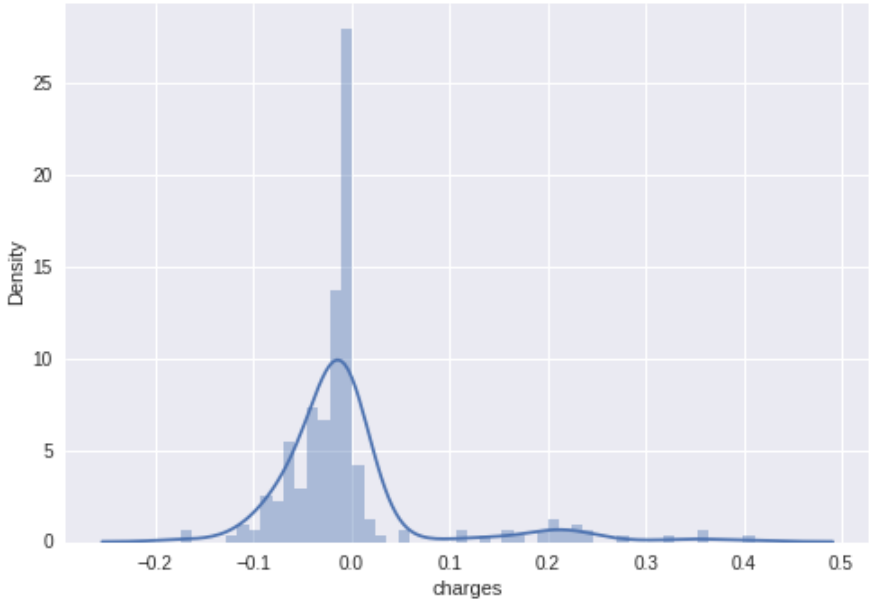

Figure 5: The distplot of prediction-ytest for charges. The errors are biased toward the left part of the plot.

PCA I used PCA for finding the two most important components in the models.

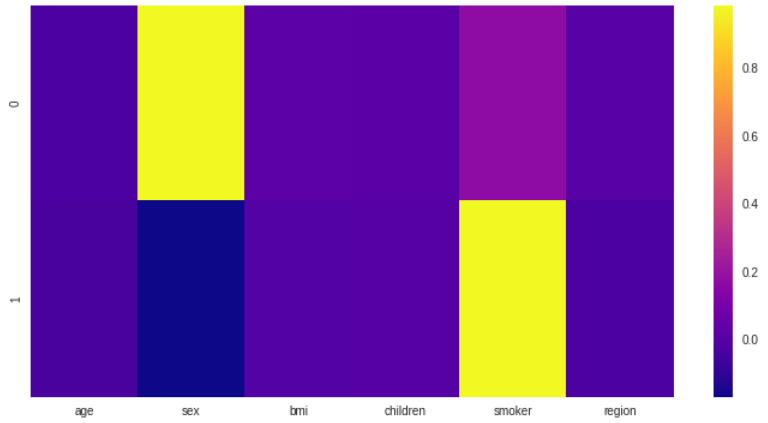

Figure 6: From the plot one can see that smoke and sex have the most effect on those two components.

KNN

I could only use KNeighborsClassifier after classifying 'charge' into 5 classes, each representing a range of 1000 values.

	precision	recall	f1-score	support
0.0	0.86	0.96	0.91	221
1.0	0.72	0.74	0.73	103
2.0	0.71	0.34	0.46	35
3.0	0.67	0.42	0.52	19
4.0	0.80	0.67	0.73	24
accuracy			0.81	402
macro avg	0.75	0.63	0.67	402
weighted avg	0.80	0.81	0.79	402

The confusion matrix is:

```
[[212   9   0   0   0]
 [ 26  76   0   1   0]
 [  9  12  12   1   1]
 [  0   8   0   8   3]
 [  0   1   5   2  16]]
```

Part II
Clustering

Figure 7: Charge-bmi lmplot with hue=smoker.

I used Kmeans for clustering, which gave me completeness score = 0.9999999999999996 and homogeneity score = 0.7772742676688331. The results are amazing actually!