Noah Caulfield

Dr. Edgar Eduardo Ceh Varela

CS 460

3/10/2024

HW3: Linear Regression

*NOTE: Screenshots are only of where an output were expected within the. ipynb file. Not every step has an output.

1.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

2.

```
df_subset    #print
```

	age	sex	bmi	children	smoker	region	charges	sex_idx	smoker_idx	region_idx
0	19	female	27.9	0	yes	southwest	16884.924	0	1	3
1	18	male	33.77	1	no	southeast	1725.5523	1	0	2
2	28	male	33.0	3	no	southeast	4449.462	1	0	2
3	33	male	22.705	0	no	northwest	21984.47061	1	0	1
4	32	male	28.88	0	no	northwest	3866.8552	1	0	1
...	...	...	...	...	...	...	...	...	...	...
1333	50	male	30.97	3	no	northwest	10600.5483	1	0	1
1334	18	female	31.92	0	no	northeast	2205.9808	0	0	0
1335	18	female	36.85	0	no	southeast	1629.8335	0	0	2
1336	21	female	25.8	0	no	southwest	2007.945	0	0	3
1337	61	female	29.07	0	yes	northwest	29141.3603	0	1	1

3.

```
print(f"y Test shape: {y_test.shape}")
```

```
X Train shape: (1070, 6)
y Train shape: (1070,)
X Test shape: (268, 6)
y Test shape: (268,)
```

Create a simple Linear Regression n

4.

```
Mean Squared Error (MSE): 33635210.431178406
R^2 Score: 0.7833463107364539
```

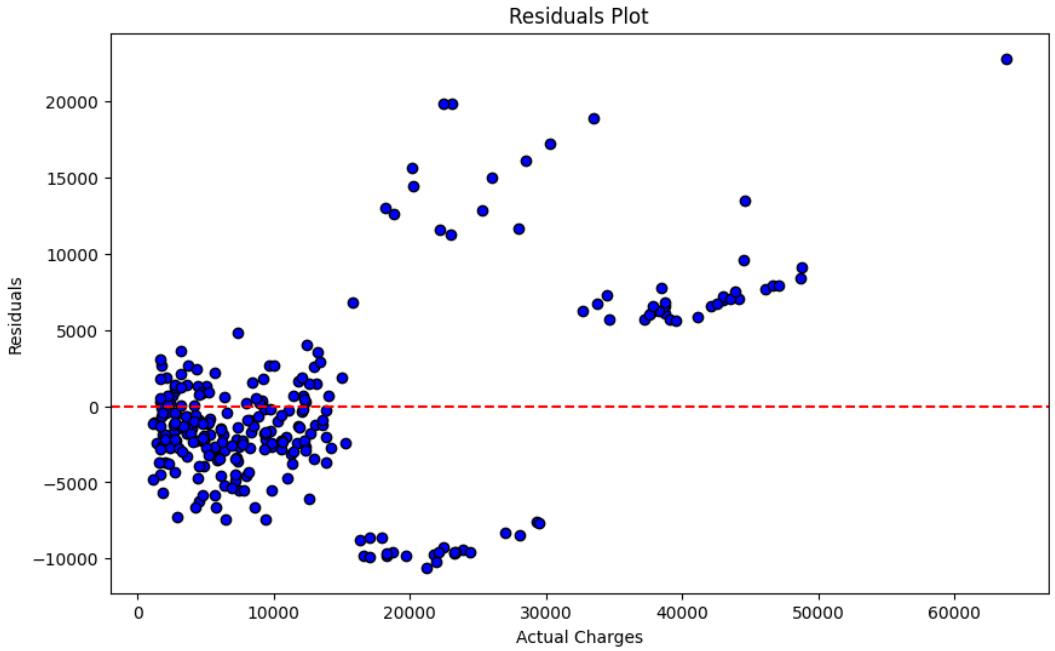
5.

```
Prediction 1: 8924.41
Prediction 2: 7116.30
Prediction 3: 36909.01
Prediction 4: 9507.87
Prediction 5: 27013.35
```

6.

```
Mean Absolute Error (MAE): 4186.5088983664355
Mean Squared Error (MSE): 33635210.431178406
Root Mean Squared Error (RMSE): 5799.587091438356
R² Score: 0.7833463107364539
```

7.



Now, create a polynomial model

8.

```
Array ( [[4.600000e+01 0.000000e+00 0.000000e+00 1.000000e+00 1.995000e+01
  2.000000e+00 2.116000e+03 0.000000e+00 0.000000e+00 4.600000e+01
  9.177000e+02 9.200000e+01 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 1.000000e+00 1.995000e+01 2.000000e+00 3.980025e+02
  3.990000e+01 4.000000e+00]
 [4.700000e+01 0.000000e+00 0.000000e+00 0.000000e+00 2.432000e+01
  0.000000e+00 2.209000e+03 0.000000e+00 0.000000e+00 0.000000e+00
  1.143040e+03 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 5.914624e+02
  0.000000e+00 0.000000e+00]
 [5.200000e+01 0.000000e+00 0.000000e+00 2.000000e+00 2.486000e+01
  0.000000e+00 2.704000e+03 0.000000e+00 0.000000e+00 1.040000e+02
  1.292720e+03 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
  0.000000e+00 4.000000e+00 4.972000e+01 0.000000e+00 6.180196e+02
  0.000000e+00 0.000000e+00]] )
```

9.

```
LogisticRegression
LogisticRegression(max_iter=1000)
```

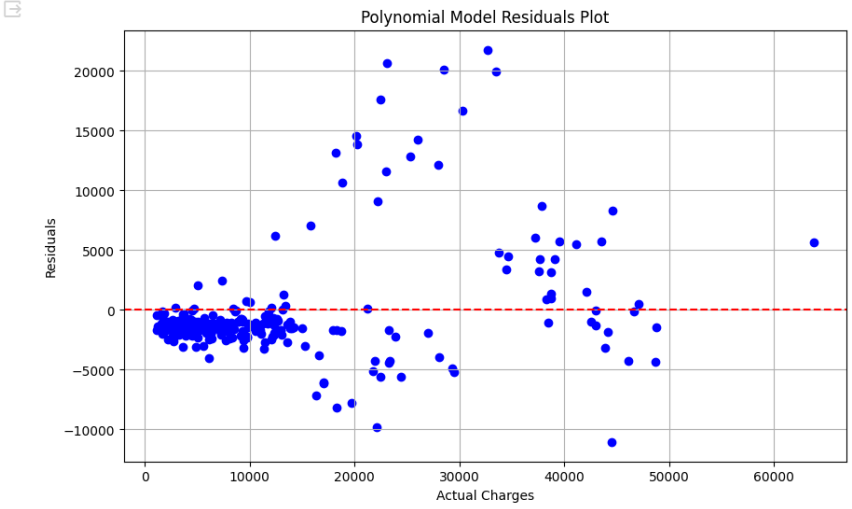
10.

```
First 5 class predictions: [0 0 0 0 0]
First 5 class probabilities:
 [[0.74046388 0.25953612]
 [0.70331278 0.29668722]
 [0.69990682 0.30009318]
 [0.70664373 0.29335627]
 [0.73773522 0.26226478]]
```

11.

```
Accuracy: 0.45
Precision: 0.6
Recall: 0.25
F1 Score: 0.35294117647058826
AUC-ROC: 0.6354166666666666
```

12.



Conclusions:

It appears that the polynomial regression model offers better accuracy compared to a simple linear regression model by capturing the non-linear relationships between variables. However, this trade-off may also include overfitting of the training data, leading to poor generalization on unseen data. In the polynomial model, the residuals are skewed more on the x-axis along the 0 marking on the y-axis.

To further improve these results, first off, I frankly need much more experience with these processes to adequately feel as though I can make improvements. However, more tuning of the model (Possibly using other model types) could be implemented to compare their performances. Also, random search could be used to fine-tune the hyperparameters systematically. Additional Regularization could also be used, along with more cross-validation.