Final Project ML
Treatment Charges Prediction

IBM Machine Learning Professional Certificate – Supervised Machine Learning: Regression

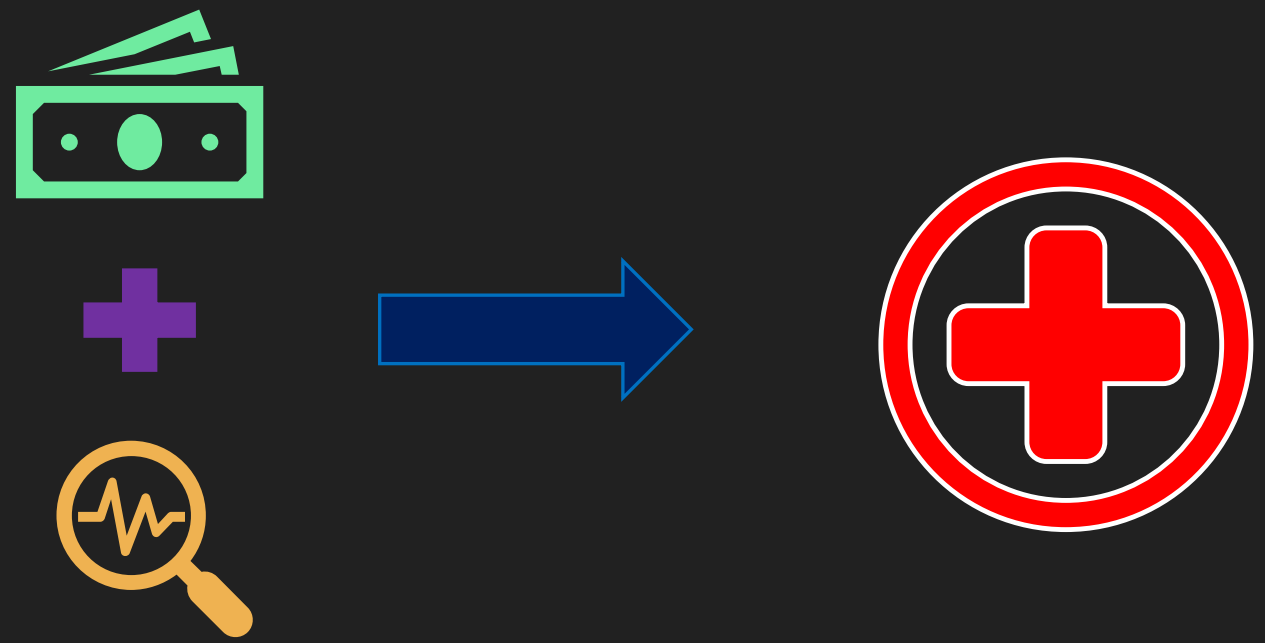
Bruno Facco Almeida

Summary

- Introduction
- Data Description
- Data Normalization
- Normality Analisys
- Applying Regression Models
- Results from the models

Introduction

- This dataset provide information about treatment costs, and using linear regression and data analisys we will provide a model to predict the insurance costs based on the information provide by the data.



Data Description

- Dataset provide by Kaggle
- Medical Cost Personal Dataset
- Columns:
  - Age
  - Sex – Male, Female
  - Bmi – Body Mass Index
  - Children – Number
  - Smoker - Smoke or not
  - Region – Residential Area
  - Charges – Treatment Costs

	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

Data Description

- No Null Values
- Data Description

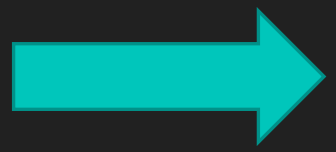
	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.505232	30.663397	1.094918	0.204783	1.515695	13270.422265
std	14.049960	0.500160	6.098187	1.205493	0.403694	1.104885	12110.011237
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1121.873900
25%	27.000000	0.000000	26.296250	0.000000	0.000000	1.000000	4740.287150
50%	39.000000	1.000000	30.400000	1.000000	0.000000	2.000000	9382.033000
75%	51.000000	1.000000	34.693750	2.000000	0.000000	2.000000	16639.912515
max	64.000000	1.000000	53.130000	5.000000	1.000000	3.000000	63770.428010

```
df.isnull().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

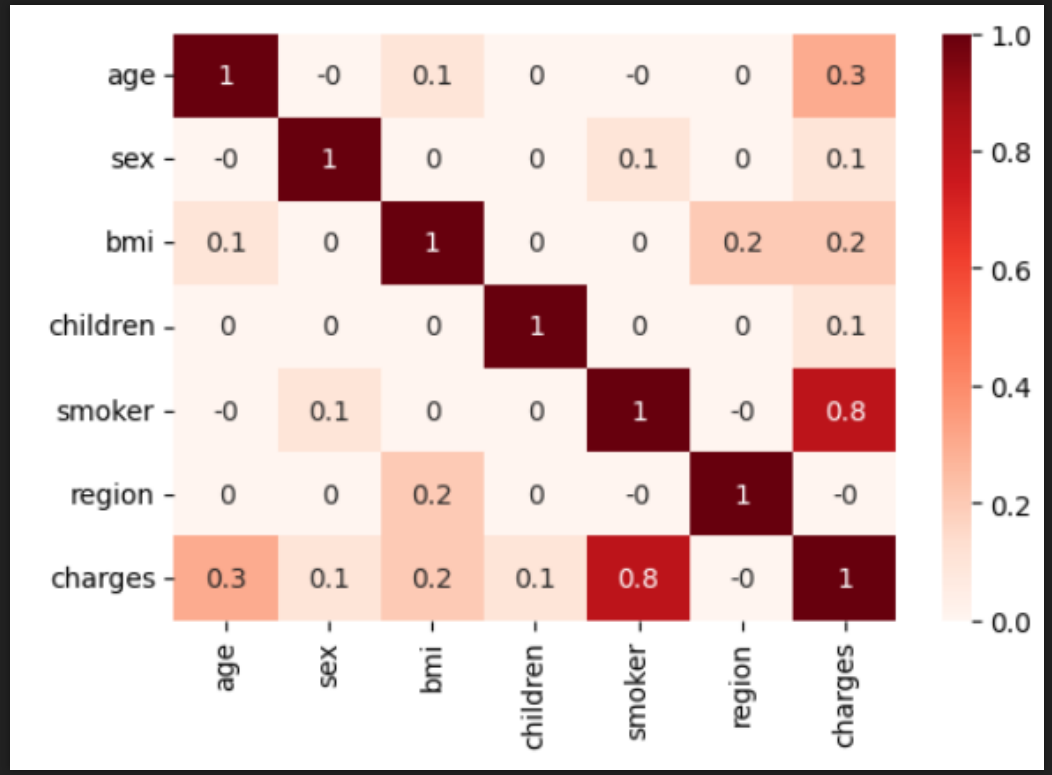
Data Normalization

	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



	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

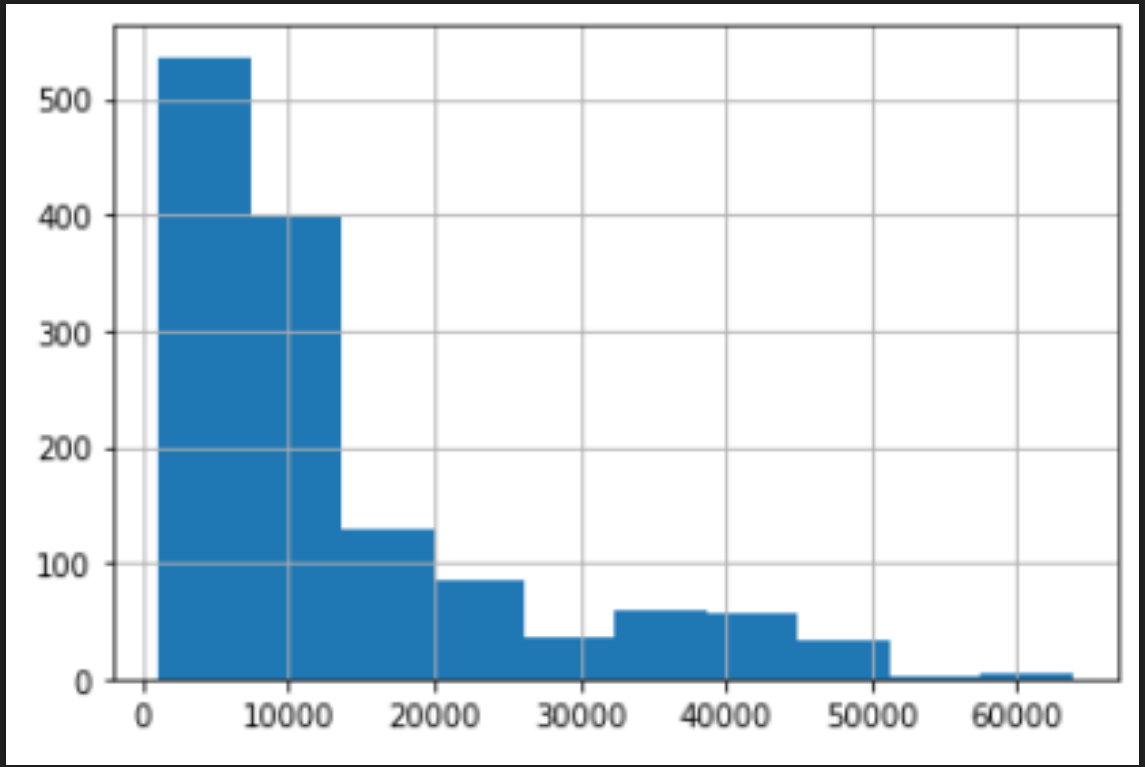
Normalization Analisys



charges	1.000000
smoker	0.787251
age	0.299008
bmi	0.198341
children	0.067998
sex	0.057292
region	-0.006208

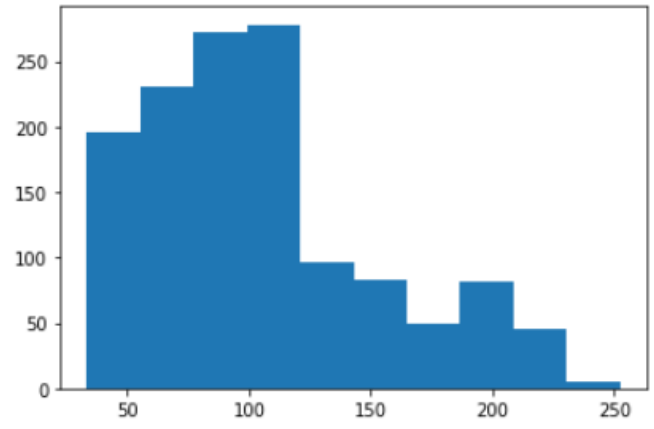
Normality Analisys

Statistic: 336.88
P value: 7.019 e-74

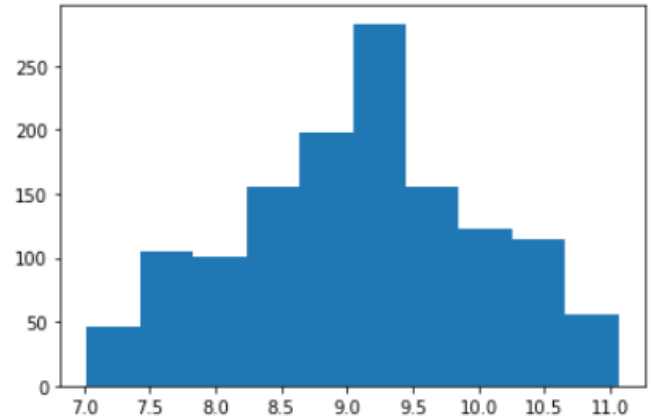


Normality Analisys

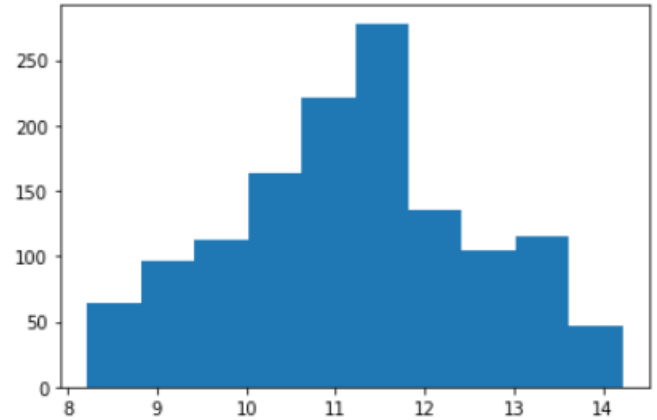
NormaltestResult(statistic=112.4605295472106, pvalue=3.7975744156203163e-25)



NormaltestResult(statistic=52.71670509113935, pvalue=3.5703676381337117e-12)



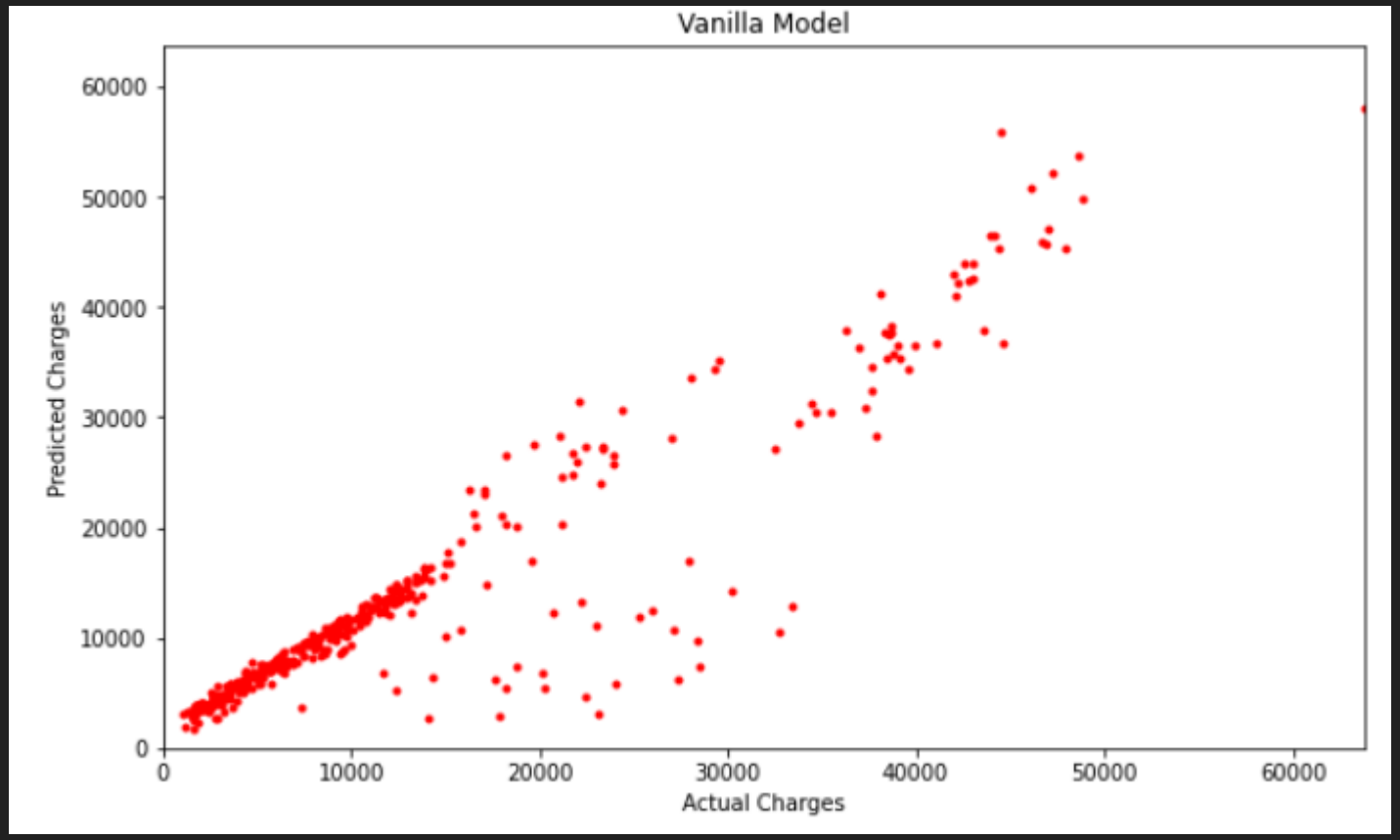
NormaltestResult(statistic=54.4181017156977, pvalue=1.5249631686757666e-12)



Box Cox Transformation has the best suit with the data

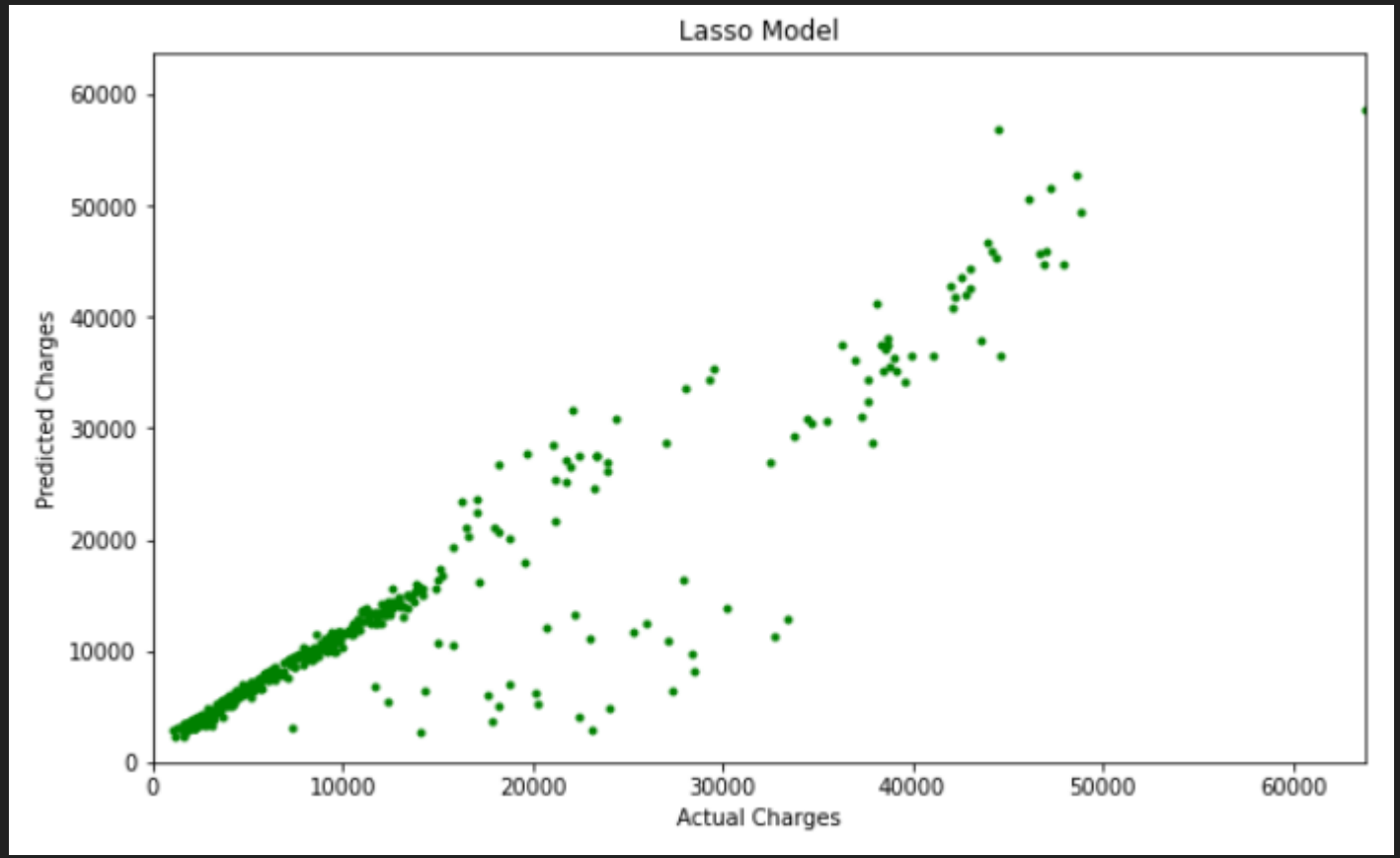
Applying Regression Models

- Vanilla Regression Model:
  - RMSE: 4496.56
  - R2: 0.8621



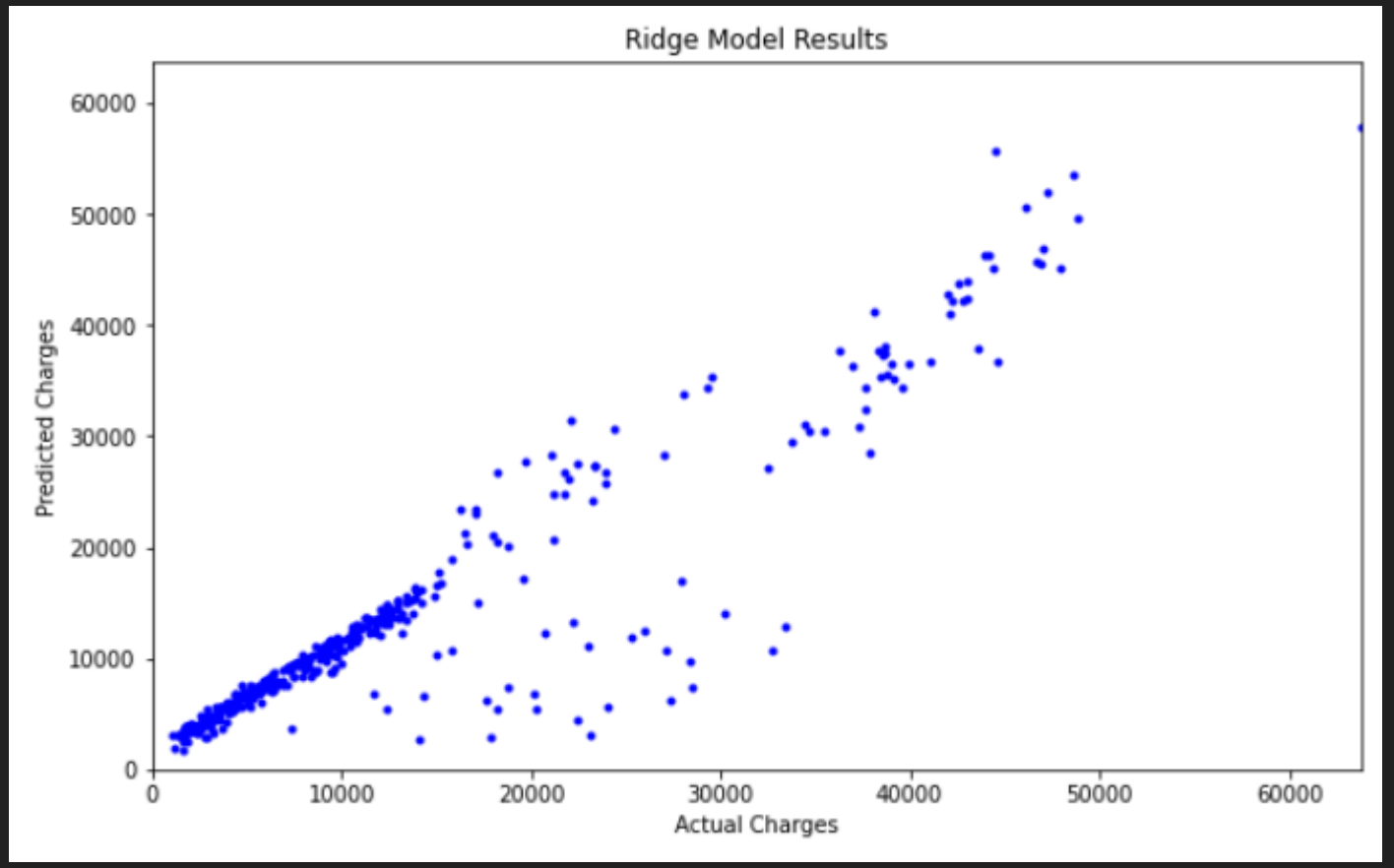
Applying Regression Models

- Lasso Regression Model:
  - RMSE: 4496.57
  - R2: 0.8621



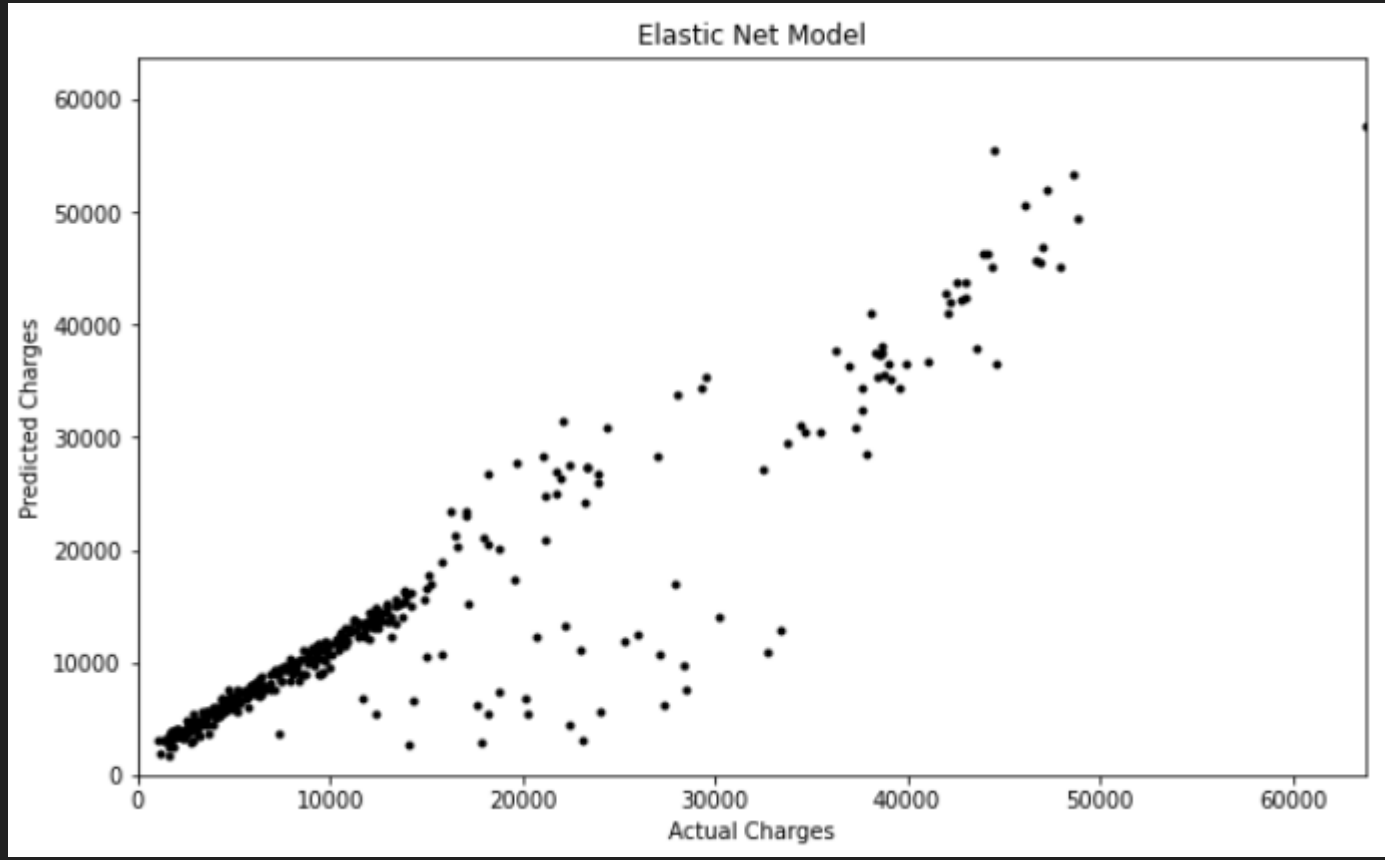
Applying Regression Models

- Ridge Regression Model:
  - RMSE: 4494.68
  - R2: 0.8622



Applying Regression Models

- Elastic Net Regression Model:
  - RMSE: 4494.41
  - R2: 0.8622



Results from the models

- Based on the results from the different regression models, each regression model had a similar result, don't having a big difference applying each one.
- Considering this, we can use anyone of the models to determine the insurance costs.
- Other point is that the model is better to predic low values of cost compared with high costs.

	RMSE	R2
Linear	4496.560111	0.862103
Lasso	4496.577652	0.862102
Ridge	4494.682980	0.862218
ElasticNet	4494.417701	0.862234