IBM PROFESSIONAL CERTIFICATE: Supervised Learning - Regression

Ibrahim Mohamed

July 2022

MAIN OBJECTIVE

- This analysis' primary goal is to predict the Insurance Charges using a linear regression and other regularisation regressions.
- This investigation tries train-test-split and cross-validation to get an idea of how these two strategies can influence model selection in different ways.
- Show the correlation between the features and the target predicted value and the most feature with impact on it.

ABOUT THE DATA

- The data set used in this analysis is a dedicated to the cost of treatment of different patients.
- The cost of patient treatment depends on many factors like the diagnoses, city, age, type of medical facility.
- This data doesn’t include data about the patient diagnoses but we have general info about his health.
- This data set has 1338 records and 7 variables.

DATA EXPLORATION

- Features:

- **age:** age of customer | patient
- **sex:** male-female
- **bmi:** body mass index
- **children:** number of children
- **smoker:** smoking or not smoking
- **region:** residential area
- **charges:** treatment charges

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

DATA EXPLORATION

- Description:

Mean

age : 39
bmi : 30.6
children: 1
charges : 13270$

Std

age : 14
bmi : 6
children: 1
charges : 12110$

Max

age : 64
bmi : 53.13
children: 5
charges : 63770.43$

Min

age : 18
bmi : 15.96
children: 0
charges : 1121.87$

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

DATA EXPLORATION

■ Data Types & Null Values

- Our Data Types are the following
- Our data doesn't have any missing values

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

	data
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

EXPLORATORY DATA ANALYSIS

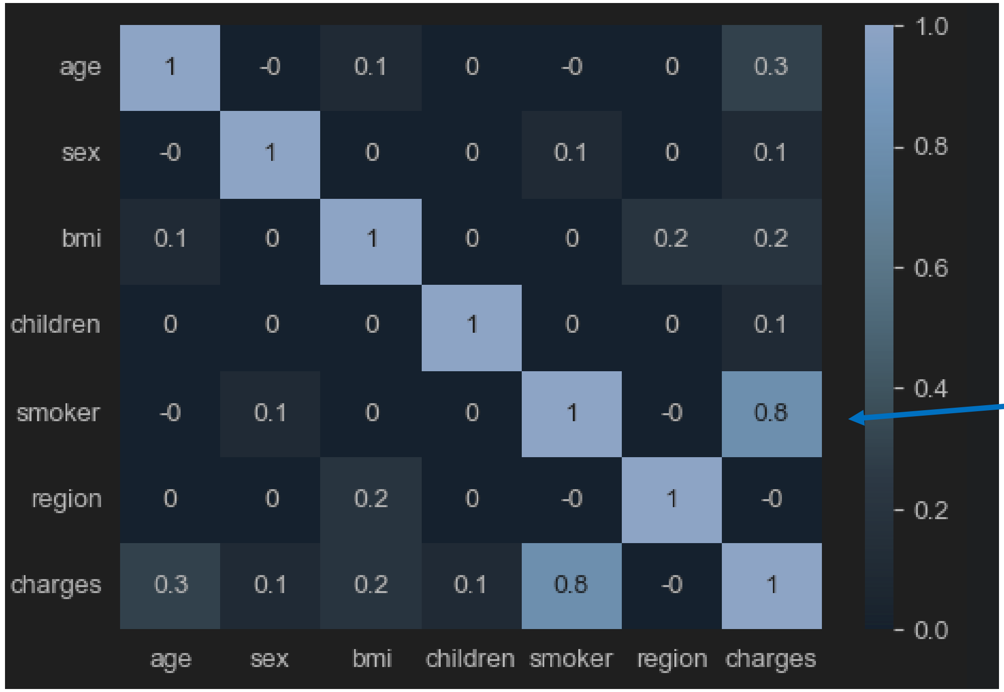
- Converting categorical features into numerical features

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

	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
5	31	0	25.740	0	0	2	3756.62160
6	46	0	33.440	1	0	2	8240.58960
7	37	0	27.740	3	0	1	7281.50560
8	37	1	29.830	2	0	0	6406.41070
9	60	0	25.840	0	0	1	28923.13692

EXPLORATORY DATA ANALYSIS

- Correlation between features
  - From the heat map we notice that there is a strong correlation between the smoking and the chargers and a very weak correlation between the region and the charges



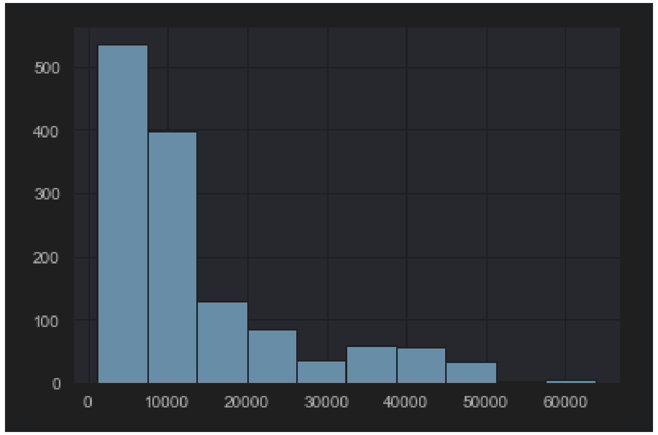
	charges
charges	1.000000
smoker	0.787251
age	0.299008
bmi	0.198341
children	0.067998
sex	0.057292
region	-0.006208

EXPLORATORY DATA ANALYSIS

Data Normality

- Implementing a model to a normally disturbed target value leads to better results, so we will make a test to our value to make sure it is normally distributed. These tests are visualizing the data and the p-value calculations

Visual



P-value calculations

```
NormaltestResult(statistic=336.8851220567733, pvalue=7.019807901276197e-74)
```

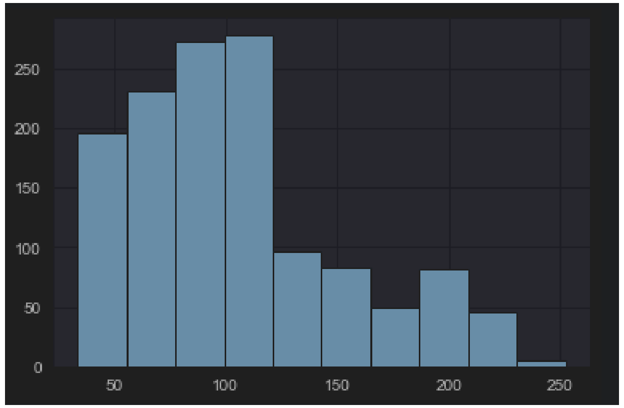
- statistic = 336.8851220567733
- p-value = 7.019807901276e-74

EXPLORATORY DATA ANALYSIS

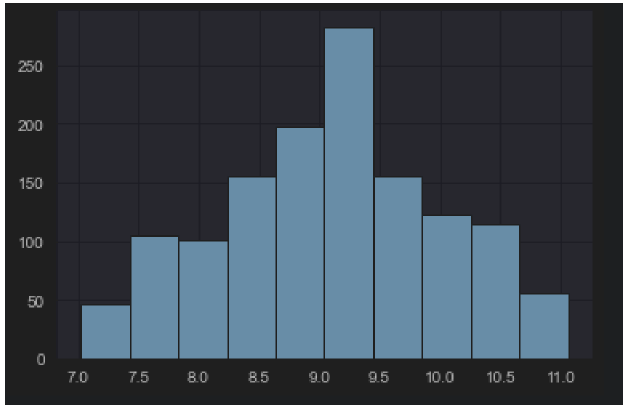
- Data Normality

  - From the previous, we conclude that the target value is not normally distributed, so we have to apply a transformation

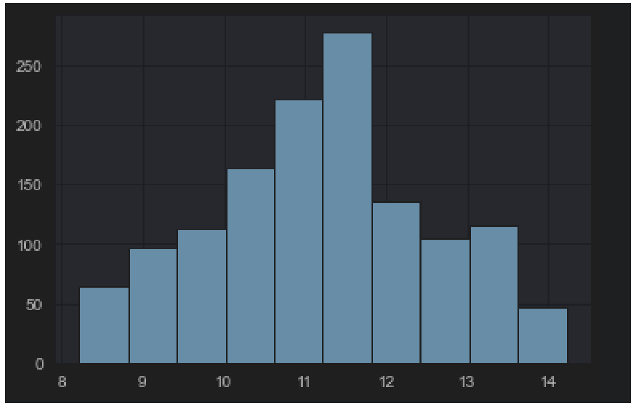
Square root transformation



Log Transformation



Box Cox Transformation



EXPLORATORY DATA ANALYSIS

- Data Normality

  - In order to keep things simple, we can use the log transformation as there isn't much of a difference between it and the Box Cox transformation, as indicated in the table on the right. To make our target distribution more normalized!

	Transormation	P-value
0	Square-Root	3.797574e-25
1	Log	3.570368e-12
2	Box Cox	1.524963e-12

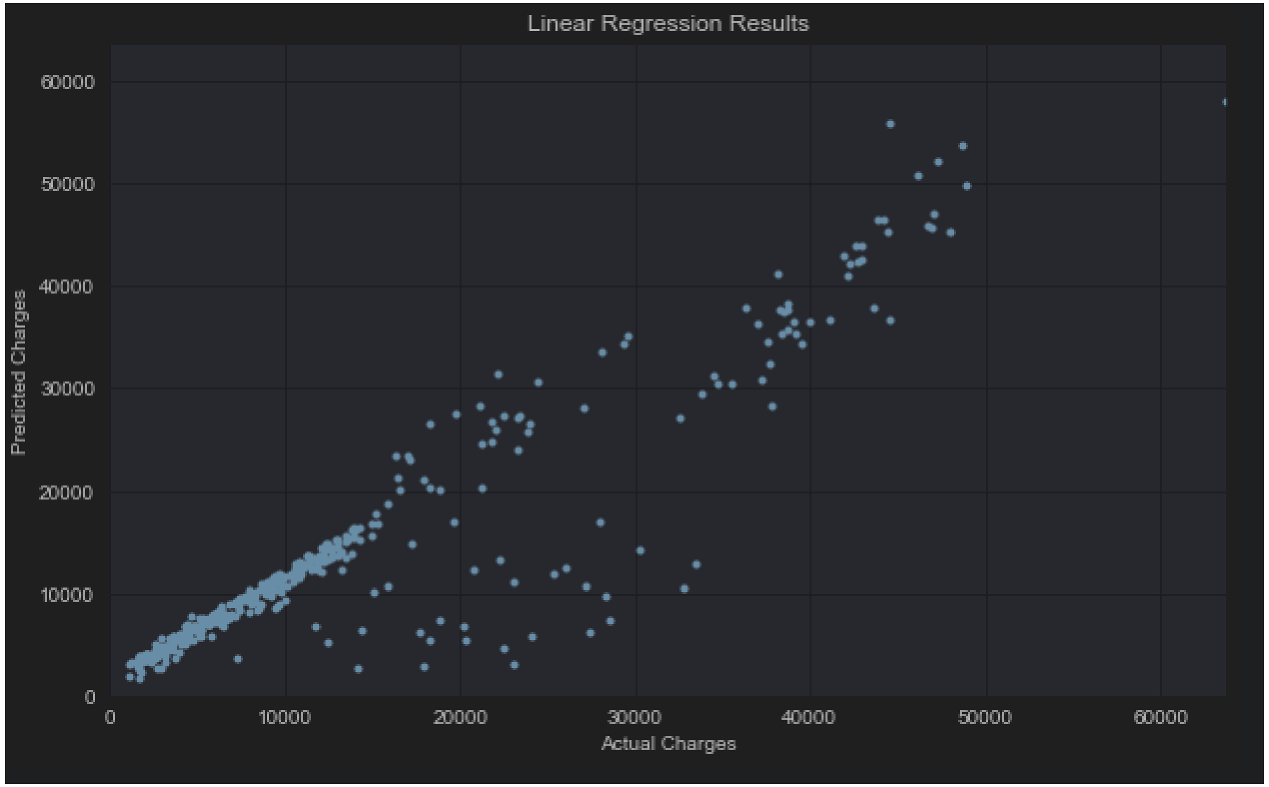
MACHINE LEARNING ANALYSIS

- **Linear Regression Model**

Model Features and Parameters:

- Model = LinearRegression()
- Polynomial Features degree = 2
- Standard Scalar

RMS_score	R2_Score
4496.560110896	0.862102995



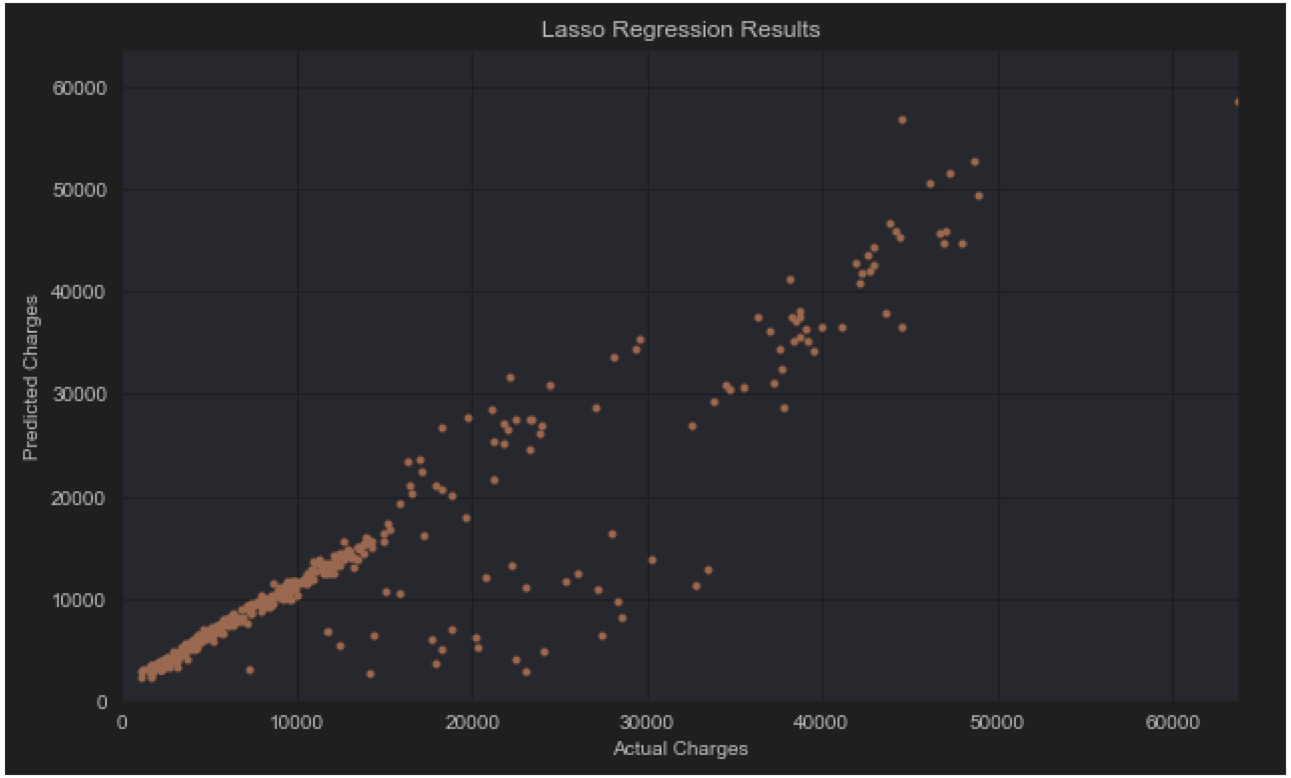
MACHINE LEARNING ANALYSIS

Lasso Regression Model

Model Features and Parameters:

- Model = Lasso()
- Polynomial Features degree = 2
- Standard Scalar
- Alpha = 13.7454
- max_iter = 10000

RMS_score	R2_Score
4496.577651935	0.862101919



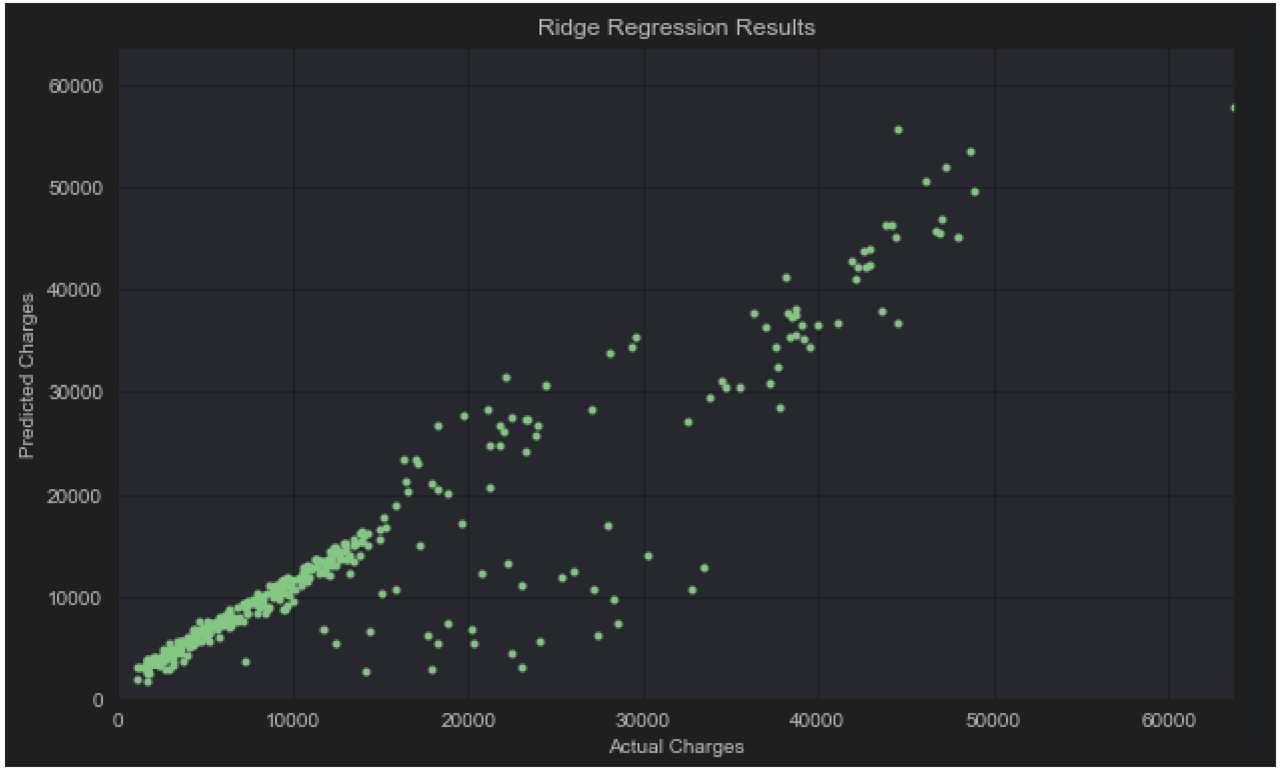
MACHINE LEARNING ANALYSIS

- **Ridge Regression Model**

Model Features and Parameters:

- Model = Ridge()
- Polynomial Features degree = 2
- Standard Scalar
- Alpha = 0.55974
- max_iter = 10000

RMS_score	R2_Score
4494.682979659	0.862218104



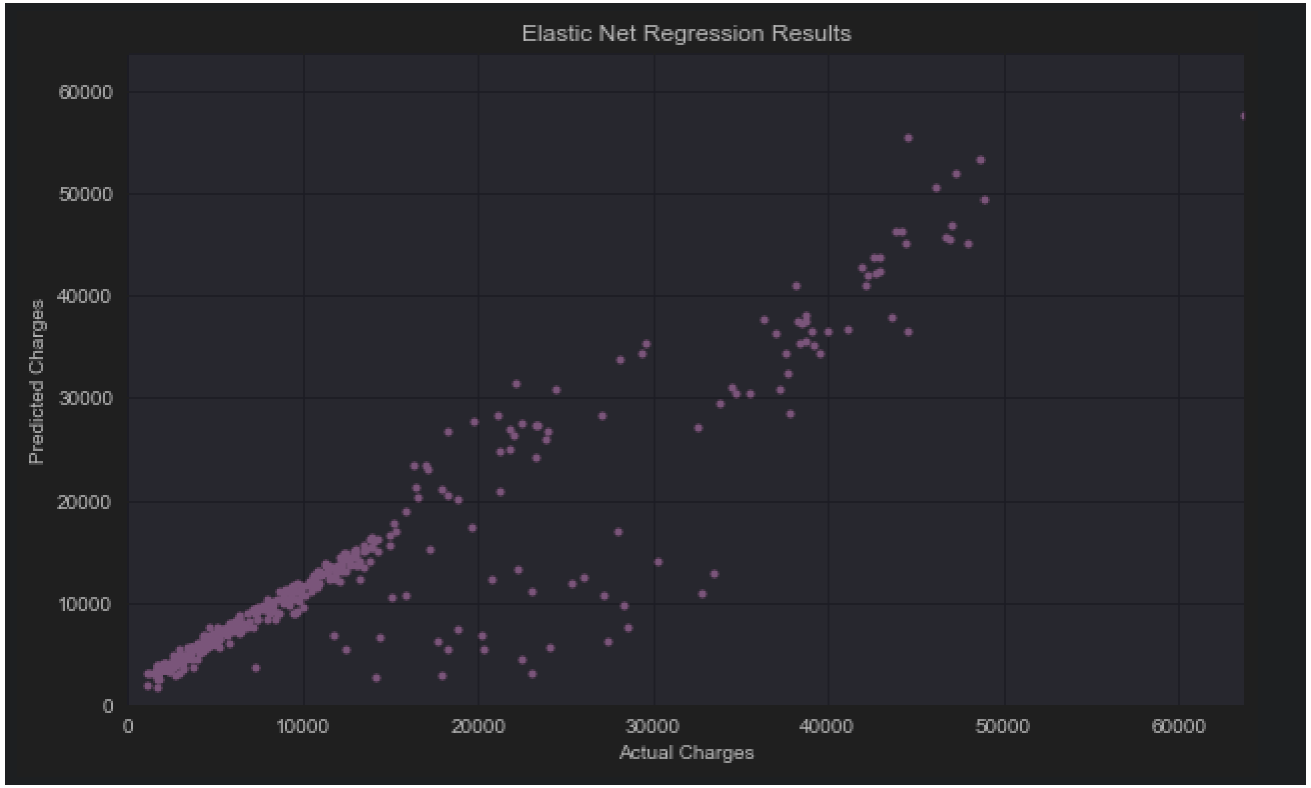
MACHINE LEARNING ANALYSIS

- **ElasticNet Regression Model**

  Model Features and Parameters:

  - Model = ElasticNet()
  - Polynomial Features degree = 2
  - Standard Scalar
  - Alpha = 0.008111
  - L1 ratio = 0.9
  - max_iter = 10000

RMS_score	R2_Score
4494.417700642	0.862218104



MACHINE LEARNING ANALYSIS

- ## Models Comparison

The used models have closely similar results and they are close to each other. It won't be a problem to choose anyone to get the job done. But according to the highest result it would be the ElasticNet model.

	RMSE	R2	RMSE-SGD	R2-SGD
Linear	4496.560111	0.862103	4531.504262	0.859951
Lasso	4496.577652	0.862102	4570.227510	0.857548
Ridge	4494.682980	0.862218	4512.691171	0.861112
ElasticNet	4494.417701	0.862234	4528.496874	0.860137

MACHINE LEARNING ANALYSIS

- Regularization

When a regularization is added to our models it conducted a worst results, so we will be using the models without regularization and the ElasticNet Model will be our selected one for this dataset.

	RMSE	R2	RMSE-SGD	R2-SGD
Linear	4496.560111	0.862103	4531.504262	0.859951
Lasso	4496.577652	0.862102	4570.227510	0.857548
Ridge	4494.682980	0.862218	4512.691171	0.861112
ElasticNet	4494.417701	0.862234	4528.496874	0.860137

ANALYSIS NEXT STEPS

- **Models Flaws and Strength and further suggestions**

From a simplicity point of view, linear regression provided high prediction results and the simplest and fastest model from a parameter point of view, but other models Lasso, Ridge and ElasticNet gave higher results.

However, more results have been obtained since then. We used a complex and slow grid search technique to find the best parameters, but it took a long time. Therefore, there is a trade-off in the end. The performance of these models is high for large datasets, but the training process can be time consuming. If you decide to use a linear model, the accuracy is relatively sacrificed, but the training process will be much faster