# Final Project ML Regression Treatment Charges Prediction

IBM Machine Learning Professional Certificate

Course 02: Supervised Machine Learning: Regression

IBM

# Contents

- Dataset Description
- Main objectives of the analysis.
- Applying various regression models.
- Machine learning analysis and findings.
- Models flaws and advanced steps.

# Data Description Section

# Introduction

Today we will explore and work on a dataset dedicated to the cost of treatment of different patients. The cost of treatment depends on many factors: diagnosis, type of clinic, city of residence, age and so on. We have no data on the diagnosis of patients. But we have other information that can help us to make a conclusion about the health of patients and practice regression analysis to create a predictive model capable of predicting the charges of insurance depending on the patient features. In any case, I wish you to be healthy! Let's look at our data.

# Dataset Description 01

| | 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 |

## 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

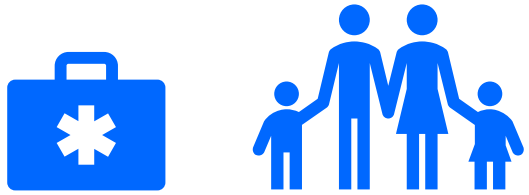
# Dataset Description 02

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

**Mean:**
**Age : 39**
**Bmi : 30.6**
**Children: 1**
**Charges : 13270$**

**Min:**
**Age : 18**
**Bmi : 15.96**
**Children: 0**
**Charges : 1121.87$**

**Max:**
**Age : 64**
**Bmi : 53.13**
**Children: 5**
**Charges : 63770.43$**



# Dataset Description 03

```
data.isnull().sum()
```

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

Great, there is **no missing values** within our features !

# Data Analysis Section

# Main Objective of the analysis:

In this section I am showing the correlation between features to find the most influence feature on our target which is insurance charges.

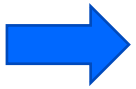
Furthermore, I am studying the normality of the features through techniques such as square root, Log Transformation, Box cox Transformation

After that I am building different regression models based on advanced techniques such as GridSearch , ML pipelines, and Hyperparameters tuning to get the best predictive model in terms of accuracy and to sho what are the flaws of each model.

# Data Analysis & Cleaning 01

- Converting categorical features into numerical features

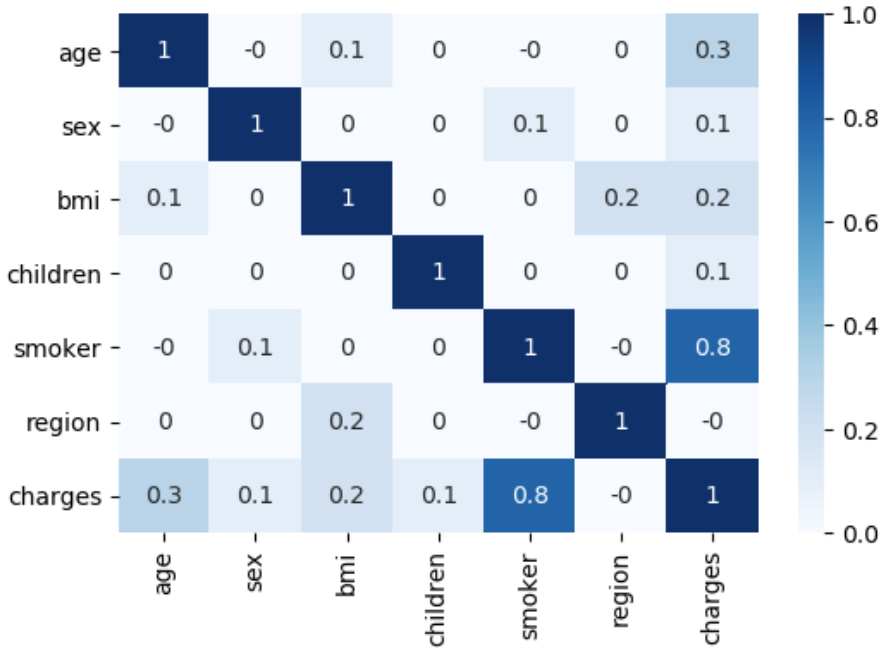
| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 0 | 19 | female | 27.900 | 0 | yes | southwest | 16884.92400 |
| 1 | 18 | male | 33.770 | 1 | no | southeast | 1725.55230 |
| 2 | 28 | male | 33.000 | 3 | no | southeast | 4449.46200 |
| 3 | 33 | male | 22.705 | 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 |

# Data Analysis & Cleaning 02

- **Studying the correlations between features using Heat Map!**




```
charges     1.000000
smoker      0.787251
age         0.299008
bmi         0.198341
children    0.067998
sex         0.057292
region     -0.006208
Name: charges, dtype: float64
```

We can notice that the strongest correlation is between “smoker” feature and our target “charges”. Where the feature does not affect our target at all is “region” which will be dropped from our dataset

# Determining Normality 01

Making our target variable normally distributed often will lead to better results
If our target is not normally distributed, we can apply a transformation to it and then fit our regression to predict the transformed values.

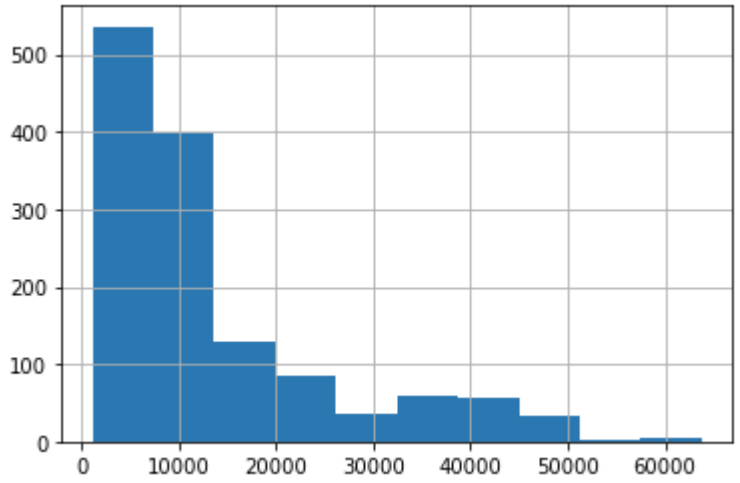
How can we tell if our target is normally distributed? There are two ways:

1- checking the visual distribution of the data.

2- calculating the P-value.

# Determining Normality 02

## Normality Visualization



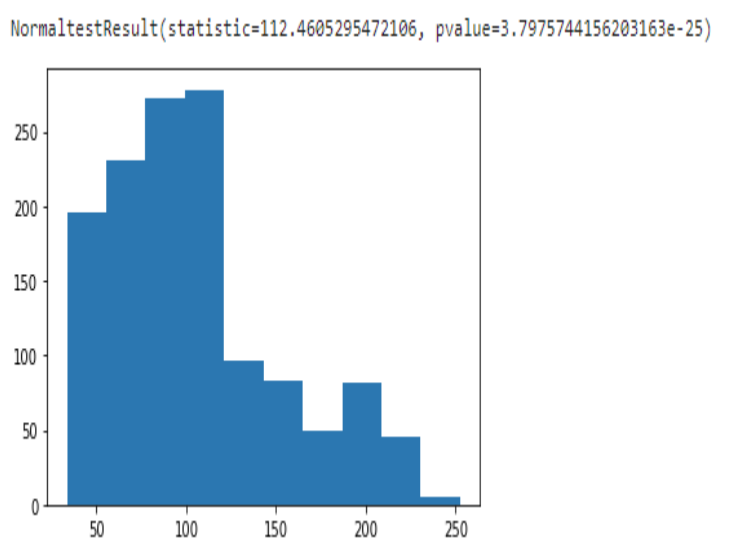

## Normal test Result

statistic = 336.8851220567733

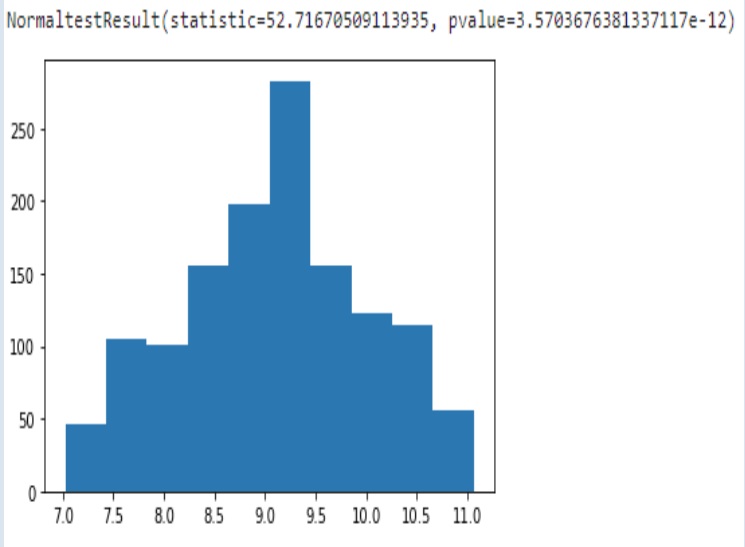
p-value = 7.019807901276e-74

# Determining Normality 03

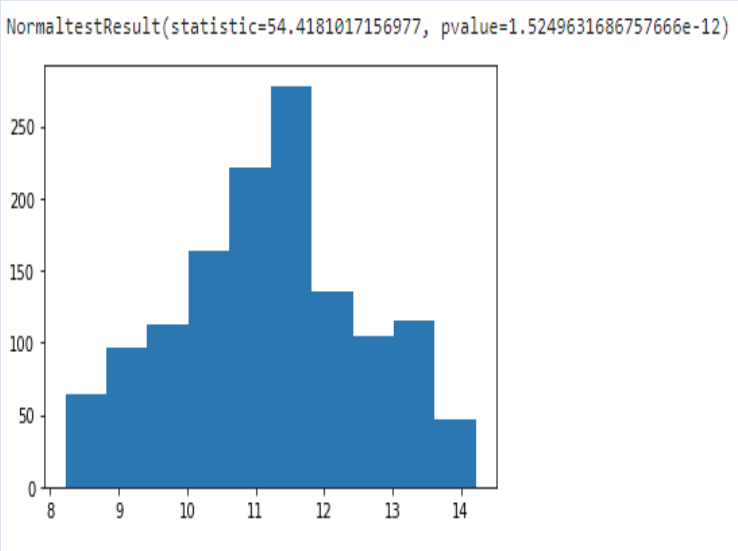
## Square root




## Log Transformation




## Box cox Transformation




As shown in the table on the right there is no big difference between log & Box Cox transformations so for the sake of simplicity, we can go with Log transformation! 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 & Findings

# Machine Learning Analysis & Findings

In the following analysis will compare between 4 different regression models Vanilla, Lasso, Ridge, and ElasticNet in terms to their accuracy in predicting the charges of treatment for patients. Where I am going to use the following techniques to help me in developing robust models:

Standard scaling, Polynomial effects, Regularization regression, cross-validation method, Grid Search, metric measurements such RMS and R2 Score.

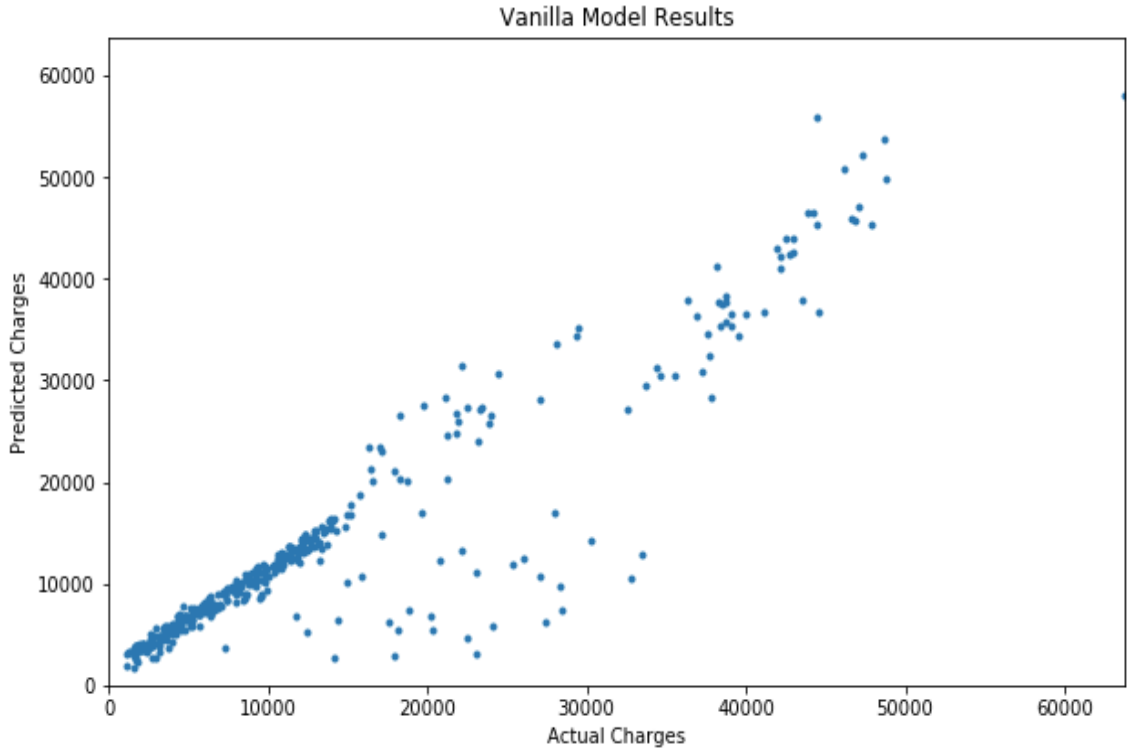
# Machine Learning Analysis 01

## Vanilla Regression Model:

Model Features and Parameters:

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

| RMS_score | R2_Score |
|---|---|
| 4496.560110896 | 0.862102995 |



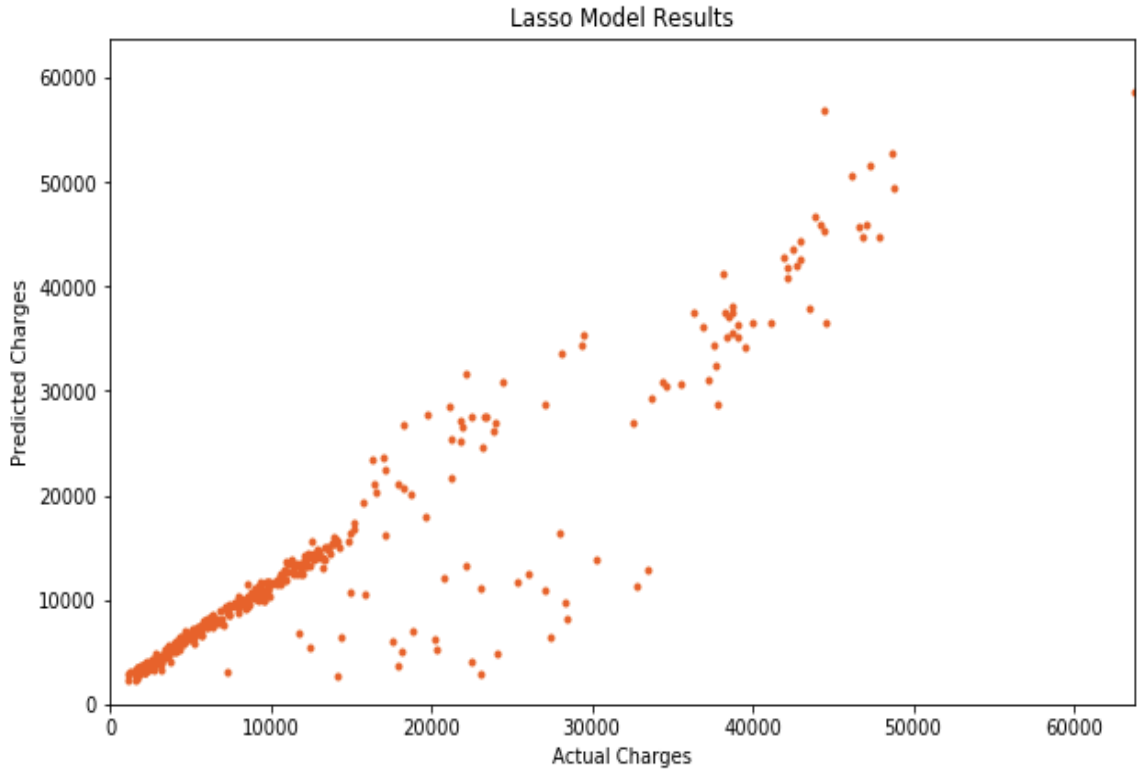
# Machine Learning Analysis 02

## 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 |



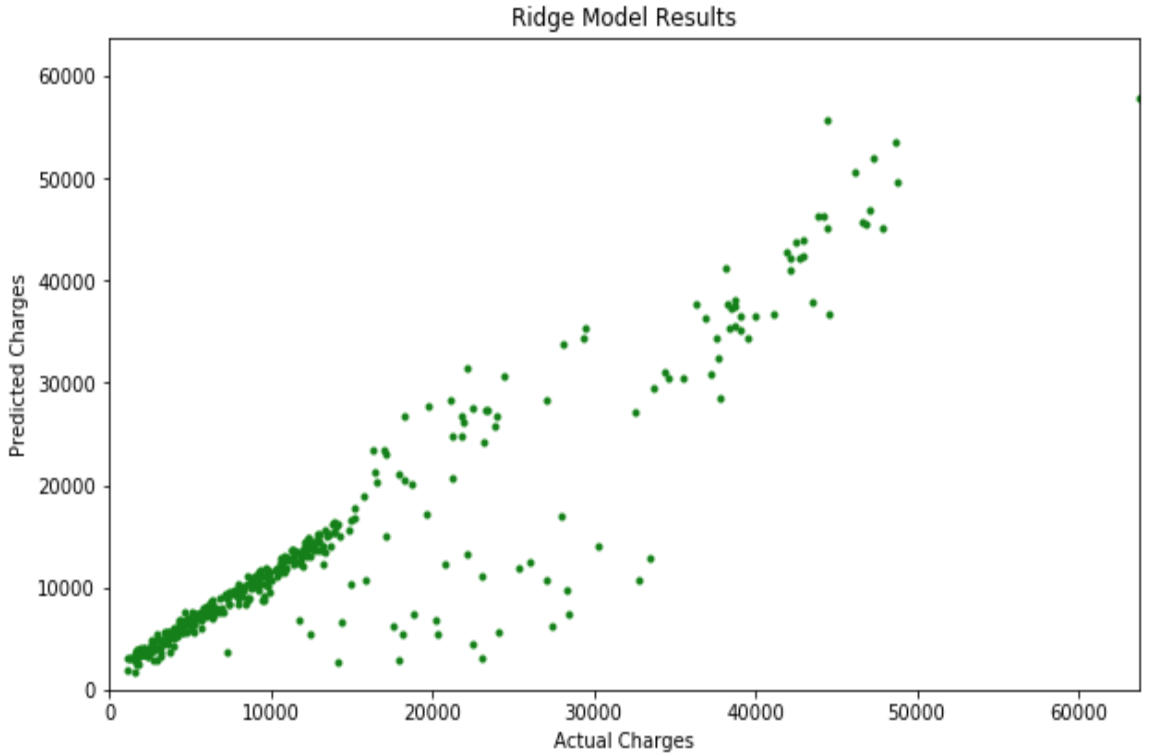
# Machine Learning Analysis 03

## 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 |



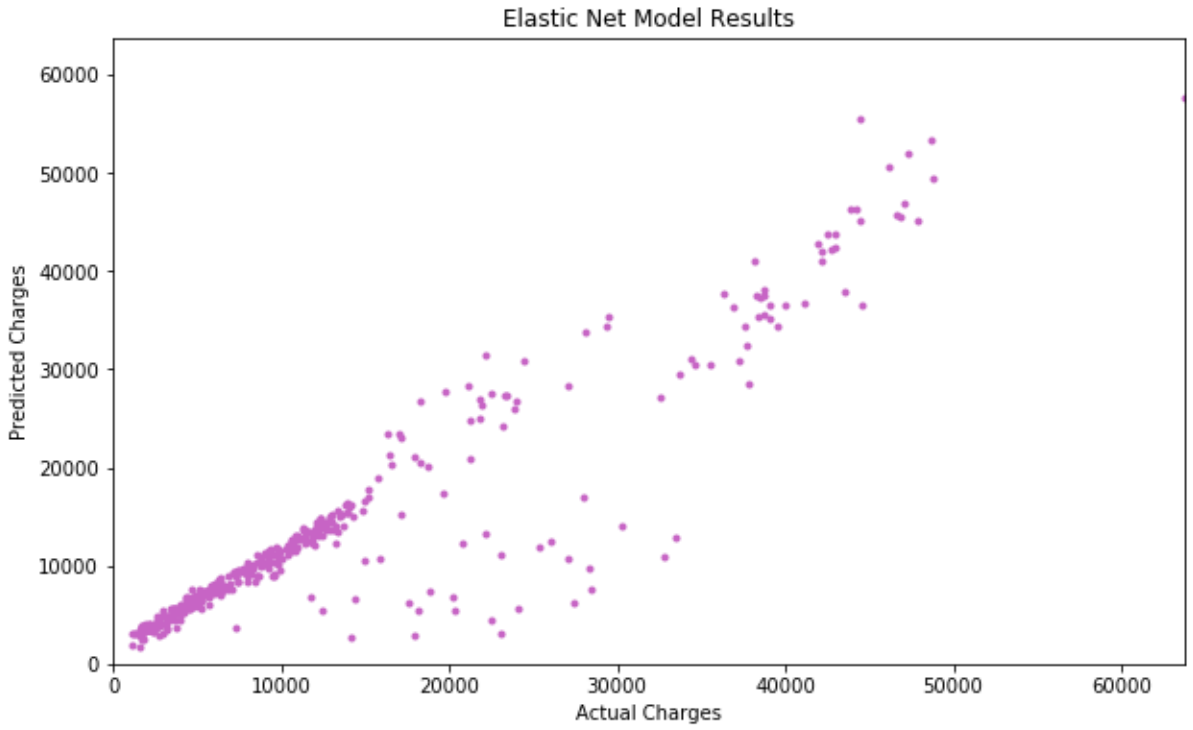
# Machine Learning Analysis 04

## 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 05

## Models Comparison

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

As shown in the data frame all the models provide very good prediction results and these results are so close to each other, But at the end we must choose one model for our dataset and this depends on the highest result.

Below I ordered the models descending:

1- ElasticNet
2- Ridge
3- Vanilla Linear
4- Lasso

# Machine Learning Analysis 06

## Adding regularization terms:

Let’s add regularization terms to our models and check how this will affect our results!

| | RMSE | R2 | RMSE-SGD | R2-SGD |
|---|---|---|---|---|
| **Linear** | 4496.560111 | 0.862103 | 4540.863842 | 0.859372 |
| **Lasso** | 4496.577652 | 0.862102 | 4533.386800 | 0.859835 |
| **Ridge** | 4494.682980 | 0.862218 | 4527.088149 | 0.860224 |
| **ElasticNet** | 4494.417701 | 0.862234 | 4525.848464 | 0.860301 |

As shown above we ended up with worst results ☹ so we can be satisfied with old models and Choose Elastic Net as highest accuracy model in terms of prediction the charges of treatment.

# Models flaws and strengths and advanced steps

# Machine Learning Analysis 07

## Models Flaws and Strength and further suggestions:

In terms of simplicity, we can say vanilla linear regression provided high predictive results and the simplest and fastest Model in terms of parameters but if we look to other models Lasso, Ridge and ElasticNet they provided higher results but in they were more complex and slower since when we used grid search technique to search about best fitting parameters, they took longer time so at the end it is a tradeoff if we have bigger dataset then the performance will be higher with these models, but the training process will take a longer time where if we choose vanilla model will relatively sacrifice by some accuracy but the training process will be much faster.

# Thank you

## IBM Machine Learning Professional Certificate

Supervised Machine Learning: Regression