Supervised Machine Learning: Regression

Dataset Introductions

The dataset used here is Medical charges dataset. This dataset was obtained from kaggle.This dataset was used for Regression problems

The dataset contains the following columns

Features	Descriptions
Age	The age of the customer
Sex	Sex of the customer
Children	How many children does the customer have
BMI	BMI of the customer
Region	Region of the customer
Charges	Medical expense of the customer(amount to predict in the model)

Dataset Observation length = 1338

Feature deletion

Before we start with the actual analysis we write a function to find features having unique value as much as observation length.We get 0 features with all unique values

Main objective of the project

This project focuses on prediction then uses interpretability to understand feature importance get more insights

Exploratory data analysis

Numerical column analysis

Charges

- Continuous variable determining the cost of insurance for a customer
- This is the target variable

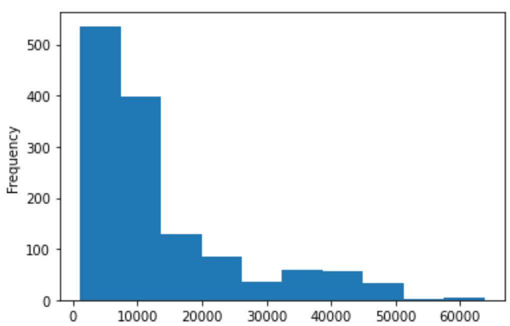
Null values

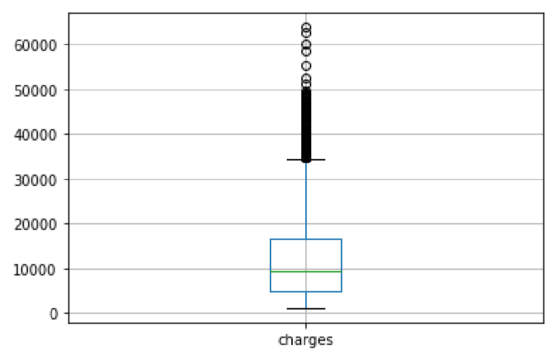
```
data['charges'].isnull().sum()
```

```
0
```

We shall first check for any null values by using the isnull function. The output was that this feature had 0 null values

Distribution





This does not follow a normal distribution. This means that most customers are getting the same charges where only some customers are getting higher charges and around 35000 afterwards we are able to notice some high outliers. So we will analyze why is it so

col_0	count
smoker	
no	3
yes	130

First we check if a customer is a smoker or non smoker and the charge is above 35000. We can see majority of customer who paid above 35000 was a smoker

We can also check for bmi scale among the outliers customers Majority of them were not normal So we can conclude if you are a smoker or not normal on bmi scale you will have a higher medicalc charges

col_0	count
Bmi_range	
Normal	1
Overweight	4
Obese	57
OverObese	71

Age

- A continuous variable
- Indicating the age of the customer

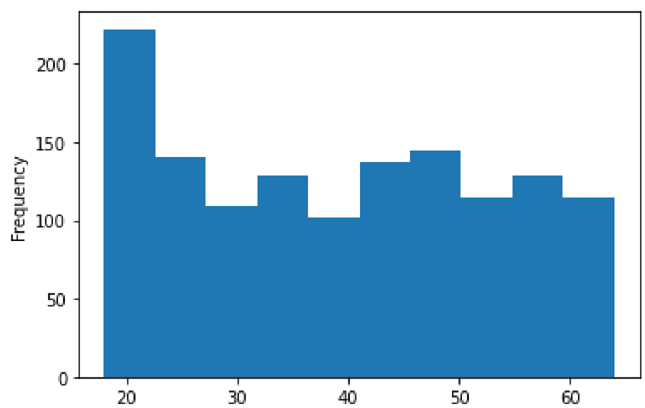
Null values

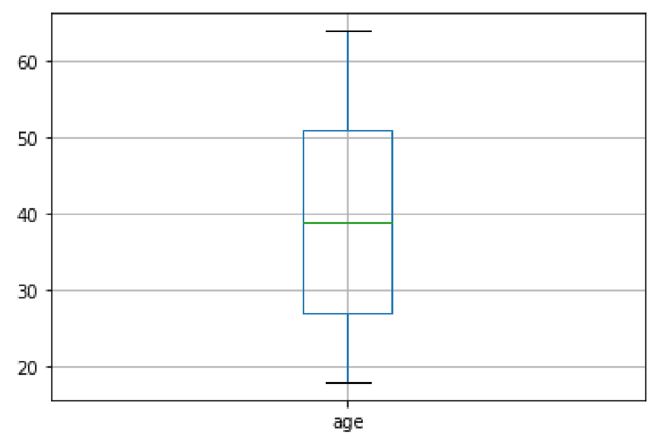
We shall first check for any null values by using the isnull function.
The output was that this feature had 0 null values

```
data['age'].isnull().sum()
```

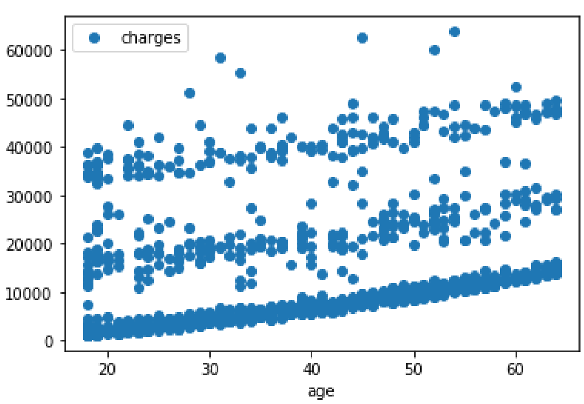
```
0
```

Distribution





We can see that the majority of the customers were of age 20. The boxplot shows no outliers



We can infer minimum charge per age group increases as age increases

	0	1
0	0-20	8713.482413
1	20-40	10686.686643
2	40-60	15888.757668
3	60-80	21063.163398

We can see the average charge per age group
This also make it clear that the average cost also increases as age group increases

BMI

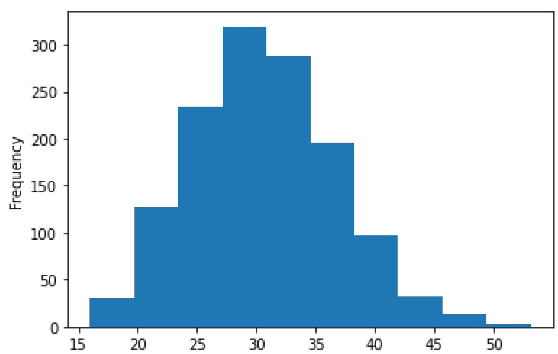
- Continuous variable telling the body mass index of a customer

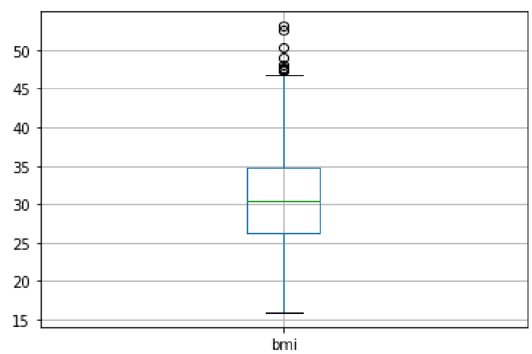
Null values

We shall first check for any null values by using the isnull function.
The output was that this feature had 0 null values

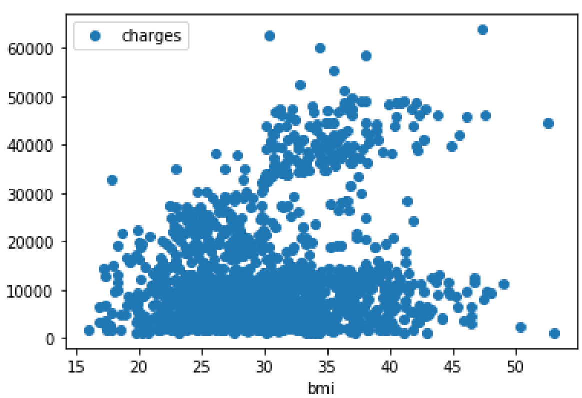
```
data['bmi'].isnull().sum()

0
```

Distribution





We can see that bmi range is somewhat distributed in an normal distribution scale .There are also some outliers.These outliers are important because this indicates that customers has been overly obese



We can see that as bmi increases the cost also increases in a upward trend

col_0	count
Bmi_range	
Underweight	41
Normal	206
Overweight	386
Obese	389
OverObese	316

This shows the overall distribution of customers we can see that a lot of people are obses in the dataset

Children

- Discrete variable indicating number of children a customer has

Null values

We shall first check for any null values by using the isnull function.
The output was that this feature had 0 null values

```
data['children'].isnull().sum()
```

```
0
```

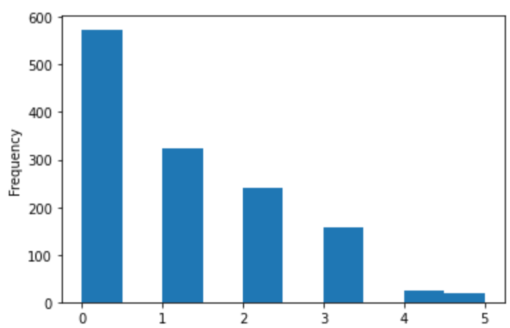
Unique values

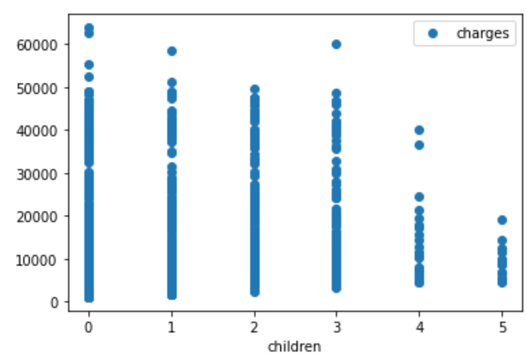
5 unique respectively
1, 2 ,3 ,4 ,5 ,0

```
data['children'].unique()
```

```
array([0, 1, 3, 2, 5, 4])
```

Distributions






```
col_0       count
children
0             574
1             324
2             240
3             157
4              25
5              18
```

```
{'0': 12365.975601635882,
 '1': 12731.171831635793,
 '2': 15073.563733958328,
 '3': 15355.31836681528,
 '4': 13850.656311199999,
 '5': 8786.035247222222}
```

Majority of them are with no children and paid charges in variety scale but customers with 5 kids payed less average medical charges compared to others

Categorical Column Analysis

Sex

- Indicates the sex of the customer
- Categorical variable

Null values

We shall first check for any null values by using the isnull function.

```
data['sex'].isnull().sum()
```

```
0
```

The output was that this feature had 0 null values

Unique values

Has two 2 unique values respectively

- Male
- Female

```
data['sex'].unique()
```

```
array(['female', 'male'], dtype=object)
```

Distribution

```
        0             1
0  female  12569.578844
1    male  13956.751178
```

col_0	count
sex	
female	662
male	676

Both count of male and female is almost identical but the mean of male is more than that of a female

Smoker

- Indicating whether customer smokes or not
- A categorical variable

Null values

We shall first check for any null values by using the isnull function.

The output was that this feature had 0 null values

```
data['smoker'].isnull().sum()
```

```
0
```

Unique values

There are 2 unique respectively
- Yes
- No

```
data['smoker'].unique()
```

```
array(['yes', 'no'], dtype=object)
```

Distribution

	0	1
0	Smoker	32050.231832
1	NotSmoker	8434.268298

col_0	count
smoker	
no	1064
yes	274

Majority of people dont smoke also average medical charges for a smoker is 4 times than that of a smoker

Region

- Denotes the area of residence of the customer
- Categorical variable

Null values

We shall first check for any null values by using the isnull function.

The output was that this feature had 0 null values

```
data['region'].isnull().sum()
```

```
0
```

Unique values

```
data['region'].unique()
```

```
array(['southwest', 'southeast', 'northwest', 'northeast'], dtype=object)
```

Four unique values

- Southwest
- Southeast
- Northwest
- Northeast

Distributions

Southeast had the highest customers and also highest average among other categories

col_0	count
region	
northeast	324
northwest	325
southeast	364
southwest	325

	0	1
0	northeast	13406.384516
1	northwest	12417.575374
2	southeast	14735.411438
3	southwest	12346.937377

Feature Encoding

Feature encoding was applied on Categorical columns to produce a total of 12 columns from 7 columns

	age	bmi	children	charges	sex_female	sex_male	region_northeast	region_northwest	region_southeast	region_southwest	smoker_no	smoker_yes
0	19	27.900	0	16884.92400	1	0	0	0	0	1	0	1
1	18	33.770	1	1725.55230	0	1	0	0	1	0	1	0
2	28	33.000	3	4449.46200	0	1	0	0	1	0	1	0
3	33	22.705	0	21984.47061	0	1	0	1	0	0	1	0
4	32	28.880	0	3866.85520	0	1	0	1	0	0	1	0

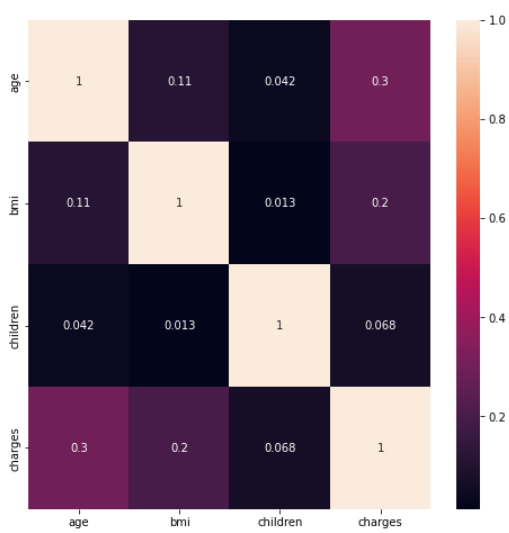
Feature Transformations

Checking the transformation effect on target feature(charges)

R2 score with box cox transformation = 0.47990628299448956
R2 score without box cox transformation = 0.7164012706034976

We will not apply box cox transformation on target variable

Correlation and interaction terms



There is not a strong positive or negative correlation between any features so we will not add any interaction terms

Vanilla linear regression

Scaling

Among three scalers standard Scaler had the least loss and highest r2 score

```
standarsScaling 37704801.58174754
minmaxScaling 37827235.04227932
robustScaling 37827235.042279325
standarsScaling 0.7164012706034976
minmaxScaling 0.7154803806270009
robustScaling 0.7154803806270008
```

Cross validation

Using Kfold cross validation with 3 splits the score we obtained was 0.7458022110624182

Polynomial features

```
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold, cross_val_predict
from sklearn.pipeline import Pipeline
estimator = Pipeline([("Polynomial", PolynomialFeatures()),
                      ("StandardScaler",StandardScaler()),
                      ("Linear Regression", LinearRegression())])
params = {"Polynomial__degree":np.arange(1,10,1)}
grid = GridSearchCV(estimator, params, cv = kf)
grid.fit(X, y)
print(grid.best_params_, grid.best_score_)
```

```
{'Polynomial__degree': 2} 0.8332961643091904
```

Using GridSearchCV we were able to find the best degree and best score we achieved using Kfold with 3 splits which were Polynomial degree = 2 ,Polynomial score = 0.8332961643091904

The final score was 0.8340521783293577

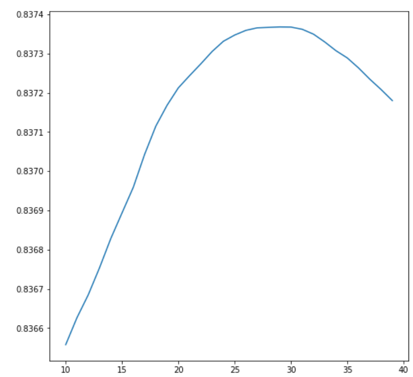
Lasso regression

```
standarsScaling 37825811.747555375
minmaxScaling 37825820.62618677
robustScaling 37823874.262637585
standarsScaling 0.7154910860161909
minmaxScaling 0.7154910192350787
robustScaling 0.7155056589150753
```

Scaling

The robust scaling showed the least error and highest r2 score

So we will applying the robust scale for lasso regression



Hyper parameter tuning

We can see that the highest r2 score was achieved around 30

Using hypermater as 30 and kfold cross validation with 3 splits and robust scaling we got the highest r2 score of 0.837367277739997

Ridge Regression

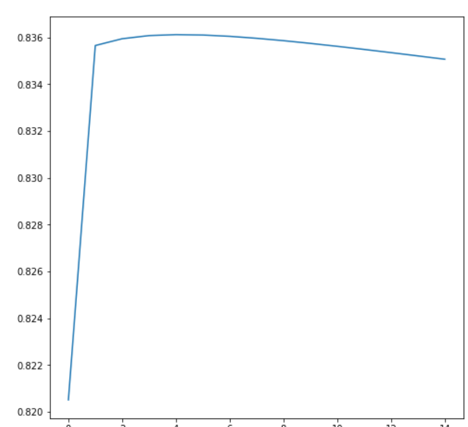
```
standarsScaling 37822065.32520359
minmaxScaling 37817732.090661205
robustScaling 37798433.721370555
standarsScaling 0.7155192649370227
minmaxScaling 0.7155518575978734
robustScaling 0.715697011339059
```

Scaling

The robust scaling showed the least error and highest r2 score

So we will applying the robust scale for Ridge regression

Hyperparameter tuning



We can see that the highest r2 score was achieved around 5

Using hypermater as 5 and kfold cross validation with 3 splits and robust scaling we got the highest r2 score of 0.8360975785580597

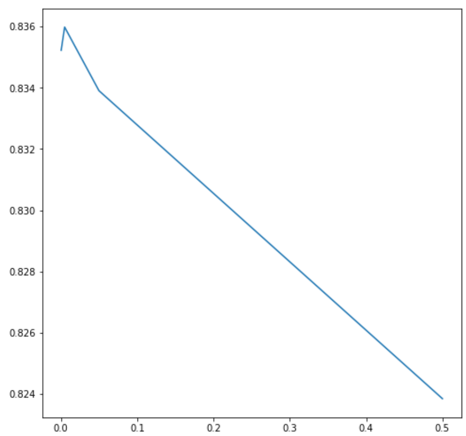
ElasticNet

Scaling

Applying scaling showed worst performance
So we won't be applying any scaling

```
standarsScaling 41318464.931991614
minmaxScaling 76458458.29431425
robustScaling 69552534.80681138
standarsScaling 0.689220903870258
minmaxScaling 0.42491351991679127
robustScaling 0.47685680152030385
```

Hyperparameter Tuning



The highest R2 score score was achieved at hypermater 0.005
Using hyperparameter of 0.005 with Kfold cross validation with 3 splits and no scaling the highest r32 score achieved was

Model Selection

Model	Scaling	Hyperparameter	R2_score
Linear Regression(vanilla)	Standard scaler	-	0.8340521783293577
Lasso Regression	Robust scaler	30	0.837367277739997
Ridge regression	Robust scaler	5	0.8360975785580597
ElasticNet regression	No scaler	0.005	0.8354538560738195

Thus Selecting lasso regression for higher predictability

Interpretability and Feature importance

	0	1
54	sex_male region_southwest	-125.802421
38	children region_southeast	-105.639566
45	sex_female region_northwest	-48.637454
23	bmi^2	-34.174945
22	age smoker_yes	-30.378901
...	...	...
48	sex_female smoker_no	770.804855
19	age region_southeast	1078.269615
32	bmi smoker_yes	1303.083589
12	age^2	5232.353776
75	smoker_no^2	14133.839845

Here

- The model is able some of the actual relationship in the data accurately example: bmi smoker_yes we already know that if both cause to higher medical charges and age^2 Where medical charge increases as age increases

- But the model also learnt the trend completely opposite of what we expected Example smoker_no^2 and age smoker_yes We know that if a customer is not a smoker the medical charge should be less but here it adds the most value for the target.This happened because of the data has more smoker than non smokers

Conclusion

The final model was able to achieve the r2 score of 0.837367277739997.But for interpretability model learnt some trends and misunderstood some trends due to lack of data

Next Steps

-Request more data to increase interpretability
-Delete the feature misunderstood by the model to improve prediction
-As the data was bias towards the smoker category .Need to balance out this categories