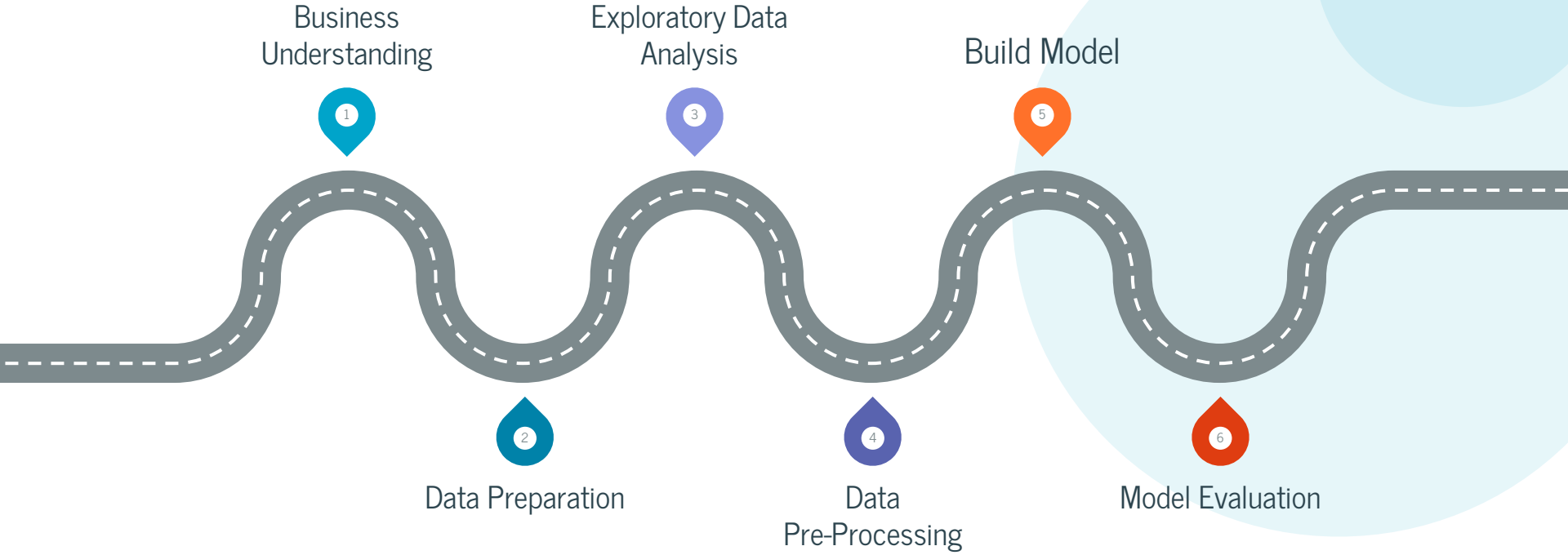
Personal Medical Cost Prediction Analysis

Using Regression Machine Learning

Ilham Taufiqurrahman U.

ROADMAP


Business Understanding
1
Data Preparation
2
Exploratory Data Analysis
3
Data Pre-Processing
4
Build Model
5
Model Evaluation
6

1 Business Understanding

Business Understanding

- Problem:

As a health insurance company, competing with competitors regarding **price and quality** is the main thing. For that, it is necessary to **adjust the price** from year to year. then from the total of all existing bills plus the company's operational costs as well as profits, then divided by the number of customers. then we will get an adjustment price and can make an appeal with competing companies.

- Goals:

Predict the **total bill of insurance each customers** based on behavior.

- Objective:

Build a **regression machine learning** that can predict total bill of insurance customers, so company can adjust the price for each customers.

- Success Criteria:

Predict the total bill of insurance customers with the benchmark : **80% of R^2 and RMSE < 0.4**

2 Data Preparation

Data Overview

No	Column	Description
1	Age	Age of primary beneficiary
2	Sex	Insurance contractor gender: female, male
3	BMI	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height
4	Children	Number of children covered by health insurance
5	Smoker	Smoking : yes, no
6	Region	The beneficiary's residential area in the US
7	Charges	Individual medical costs billed by health insurance

Data Condition

```
df.duplicated().sum()
```

```
1
```

```
df[df.duplicated(keep=False)]
```

	age	sex	bmi	children	smoker	region	charges
195	19	male	30.59	0	no	northwest	1639.5631
581	19	male	30.59	0	no	northwest	1639.5631

```
df = df.drop([581])
```

- There are 2 identical rows, for index [195] and [581]. It's necessary to drop one of them.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

- There are 7 columns and 1338 rows and there are no missing values from the data.

Data Condition

From that table we get:

- Mean of age and median of age not really different, so its possible that age has normal distribution
- It's happen to BMI and Children, that their mean and median was so close.
- Different from other columns, mean of charges is higher than its median. The possibility of the distribution is right skewed or positively skewed,
- From categorical columns, we know that male and female not so much different. But smokers (no) is dominating the data that there are 1063 people. And the Southeast region is the most region in this data.

```
df.describe()
```

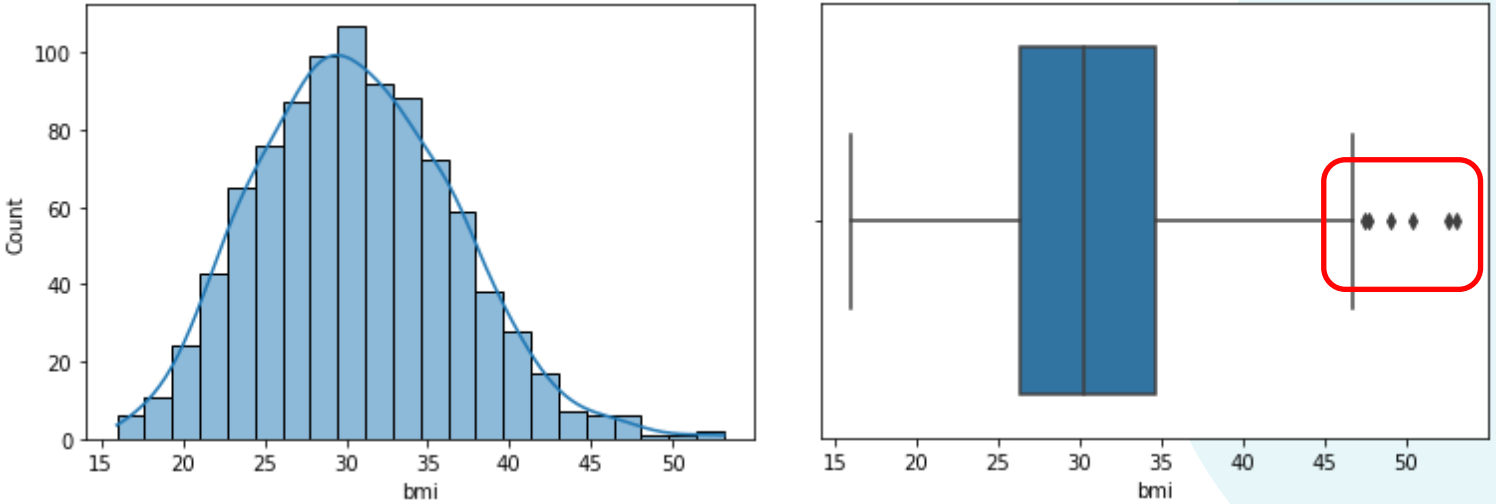
	age	bmi	children	charges
count	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487
std	14.044333	6.100468	1.205571	12110.359656
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.290000	0.000000	4746.344000
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	53.130000	5.000000	63770.428010

```
df.describe(include='object')
```

	sex	smoker	region
count	1337	1337	1337
unique	2	2	4
top	male	no	southeast
freq	675	1063	364

3 Exploratory Data Analysis

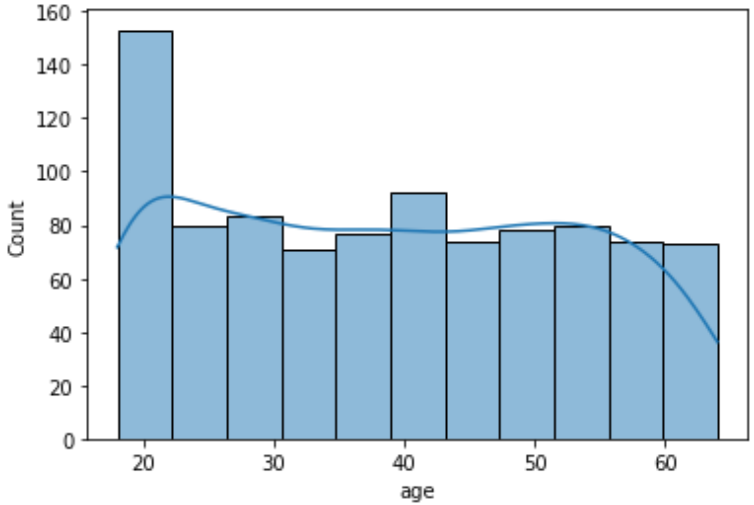
Exploratory Data Analysis

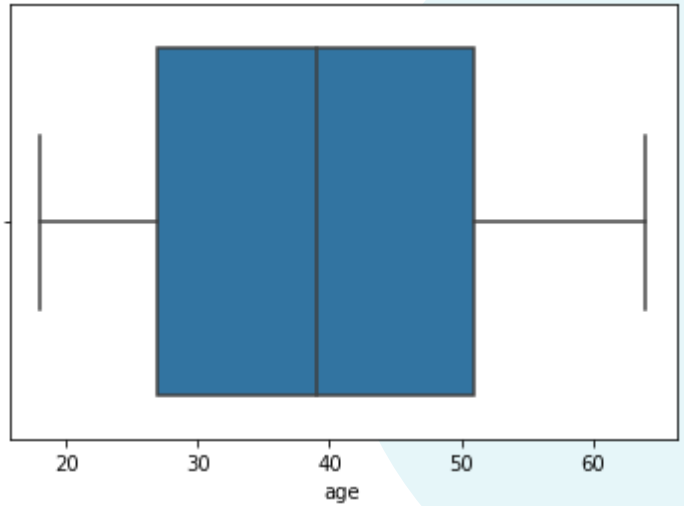


From BMI graphic we get:

- BMI distribution is close to normal distribution, but there are some outliers after Q3

Exploratory Data Analysis

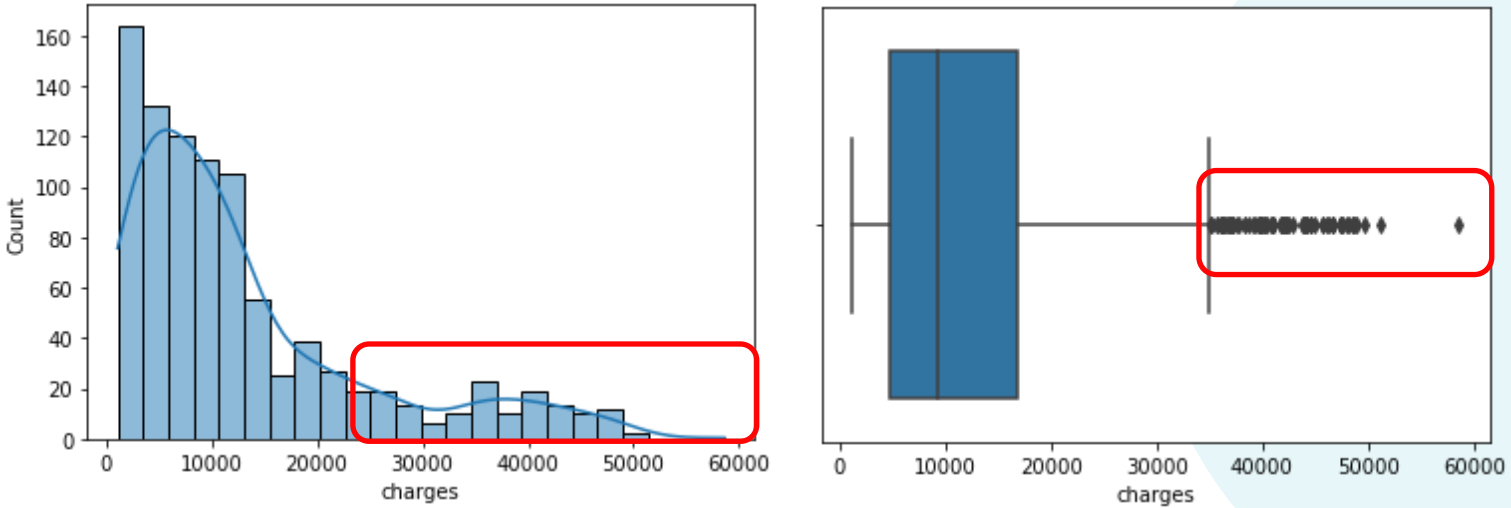




From Age graphic we get:

- Age is almost like uniform distributed but there are more customer at 18-20 years old and its not an outliers.

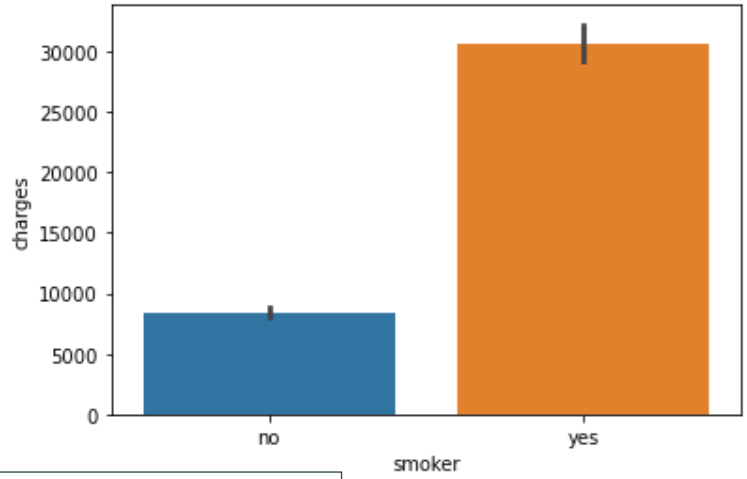
Exploratory Data Analysis

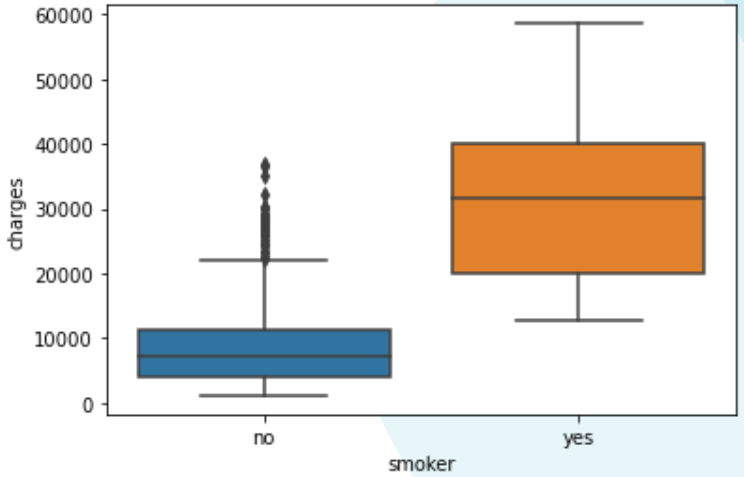


From Charges graphic we get:

- Charges clearly seen as right skewed/positively skewed distribution and a lots of outliers after Q3.

Exploratory Data Analysis

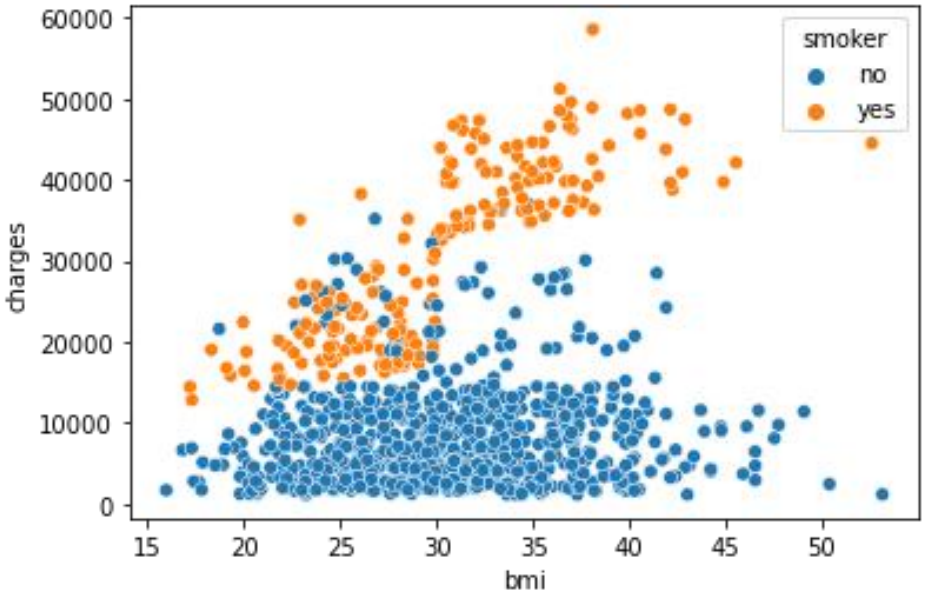




	charges	
	len	mean
smoker		
no	741	8407.907285
yes	194	30708.903177

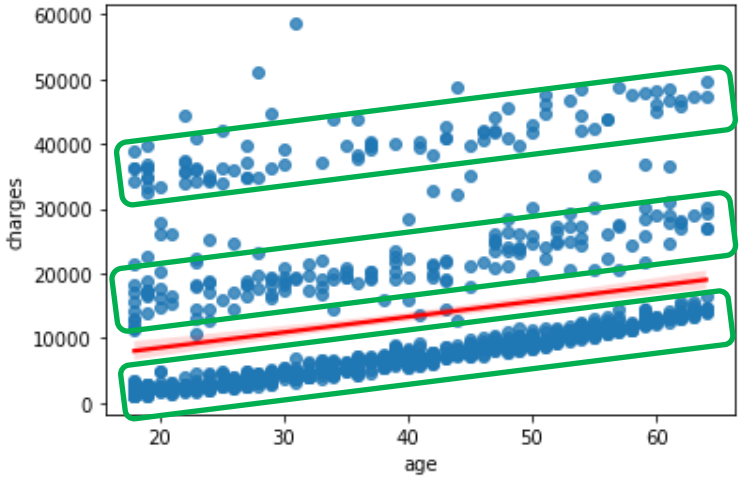
From the data, there are significant differences between who smoking and non-smoking groups viewed by their insurance charges. Smokers group has insurance cost **3.5 times** more than non-smokers group

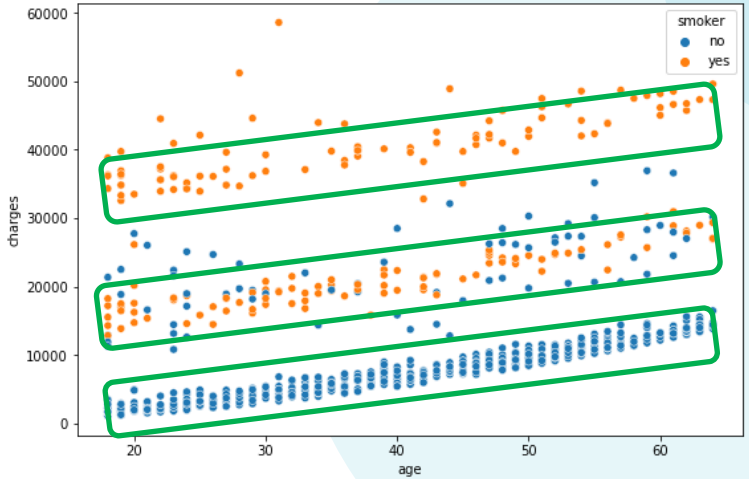
Exploratory Data Analysis



There are **strong correlation** for smoker from their **BMI and their charges**. It seems that the charges will go up when their BMI goes up. But its doesn't happen for non-smoker group because the distribution is evenly distributed on low charges

Exploratory Data Analysis





- From the scatter plot of age-charges, seems there are **3 segmentation of customer** that the increase of age, their charges is so increase too.
- And if divided into their smoke behavior, 3 segmentation have each characteristics. Bottom segmentation is for **non smokers**, middle segmentation is **mix** of non smokers and smokers, and top segmentation is just for **smokers.**

4 Data Pre-Processing

Data Pre-Processing

- Label Encoding
- One-hot Encoding

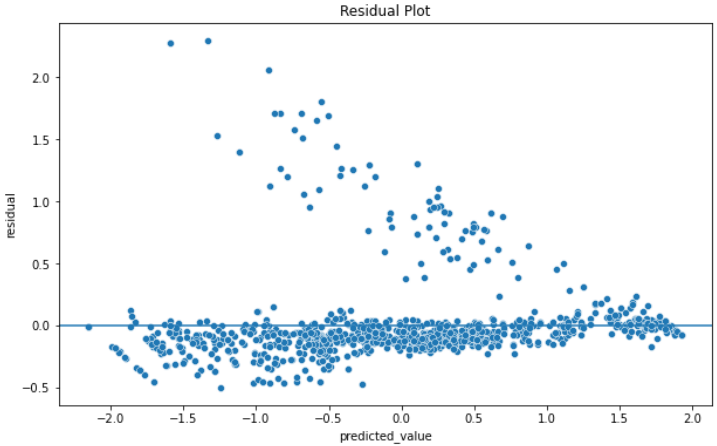
Drop column that have low correlation with target

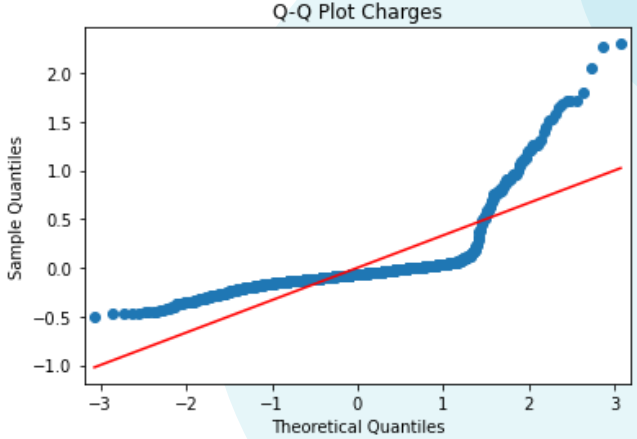
Handling outliers using **capping method**

- Fit and Transform data train
- Transform data test

5 Build Model and Evaluation

Gradient Boosting Regressor



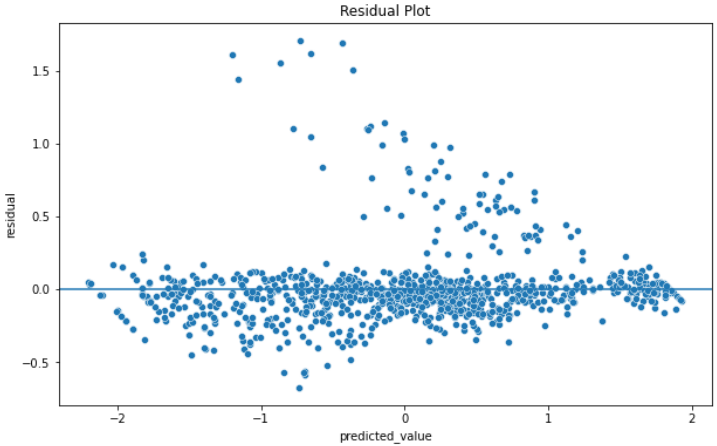


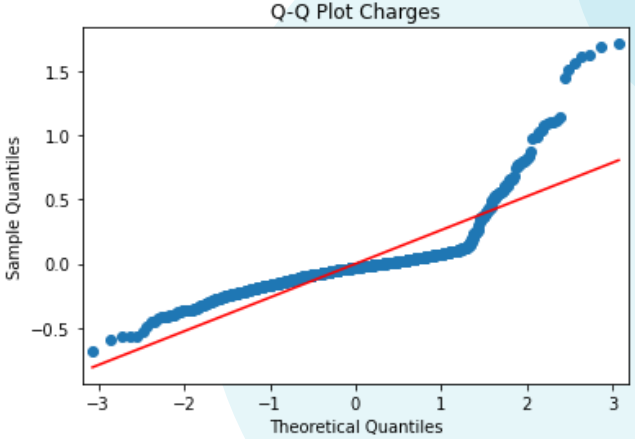

```
eval_regression(gbr)
```

```
rmse (train): 0.3333800789864192
rmse (test): 0.4077955104563977
r2 (train): 0.888857722935009
r2 (test): 0.8444537199067195
```

From the evaluation results, it can be seen that the residual distribution is **not normally distributed**, indicated by the residuals not in line with the red fit line. However, the Gradient Boosting Regressor performa has a fairly good RMSE test and R^2 test

LightGBM Regressor






```
eval_regression(lgbm)
```

```
rmse (train): 0.2619162423055413
rmse (test): 0.43146991501434967
r2 (train): 0.931399882016545
r2 (test): 0.8258691229958506
```

From the results of the LightGBM Regressor the residual is **slightly better** than the Gradient Boosting Regressor, but the model has a slightly decreased performance for the RMSE test and R^2 test.

Business Recommedation

- Make a **stop smoking campaign** to customers so their health is much better marked by cost of their health services which will decrease if they do not smoke.

"It is health that is real wealth, and not pieces of gold and silver"

-Mahatma Gandhi-

Thank you!

</do_you_have_any_question?>

</reach_me_at :>

Ilham Taufiqurrahman

ilhamtaufiqur@gmail.com