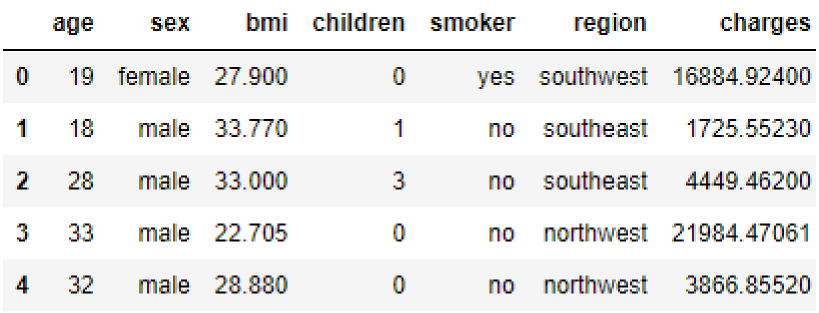
Insurance Charges Data Set – Supervised Machine Learning: Regression

Dataset Kaggle Link (https://www.kaggle.com/mirichoi0218/insurance).

The data consists of health insurance customers' information and the amount of charges they would pay for the insurance. A snippet of the first 5 rows is shown below:



	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

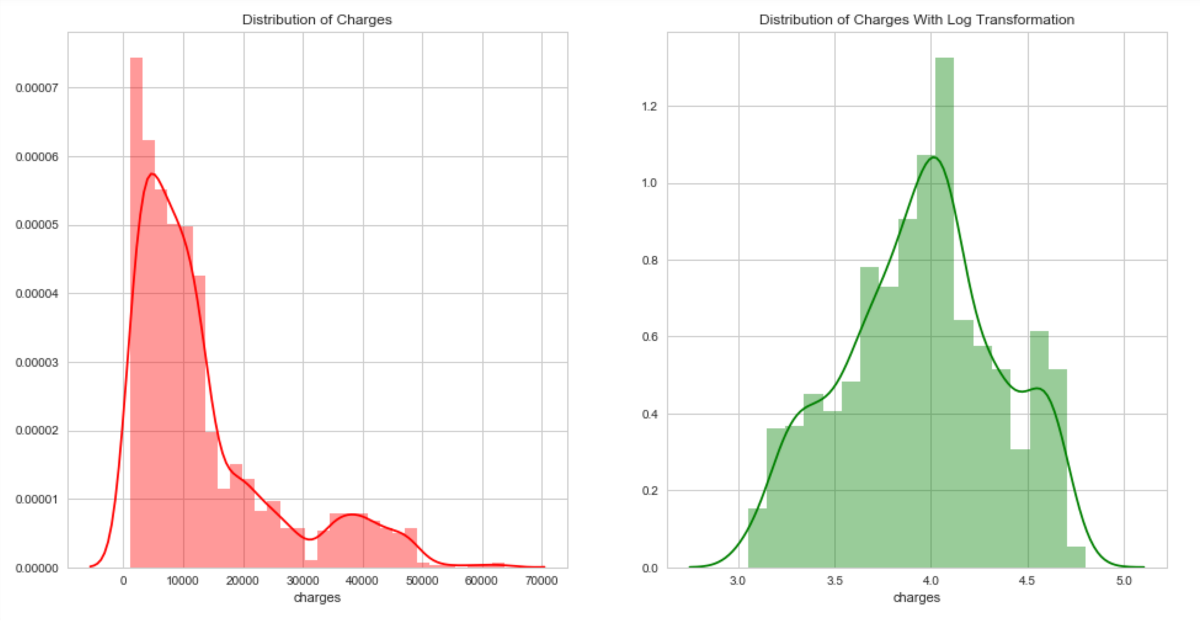
The data taken from Kaggle was clean and no values were missing. It consists of **1338 rows,** i.e., customers. A description of the features is given below:

- Age: Age of customer.
- Sex: Insurance customer's gender: Female or Male.
- bmi: Body mass index.
- Children: Number of children covered by health insurance / Number of dependents.
- Smoker: If Insurance customer smokes: Yes or No.
- Region: The customer's residential area: Northeast, Southeast, Southwest, or Northwest.
- Charges: cost of the health insurance

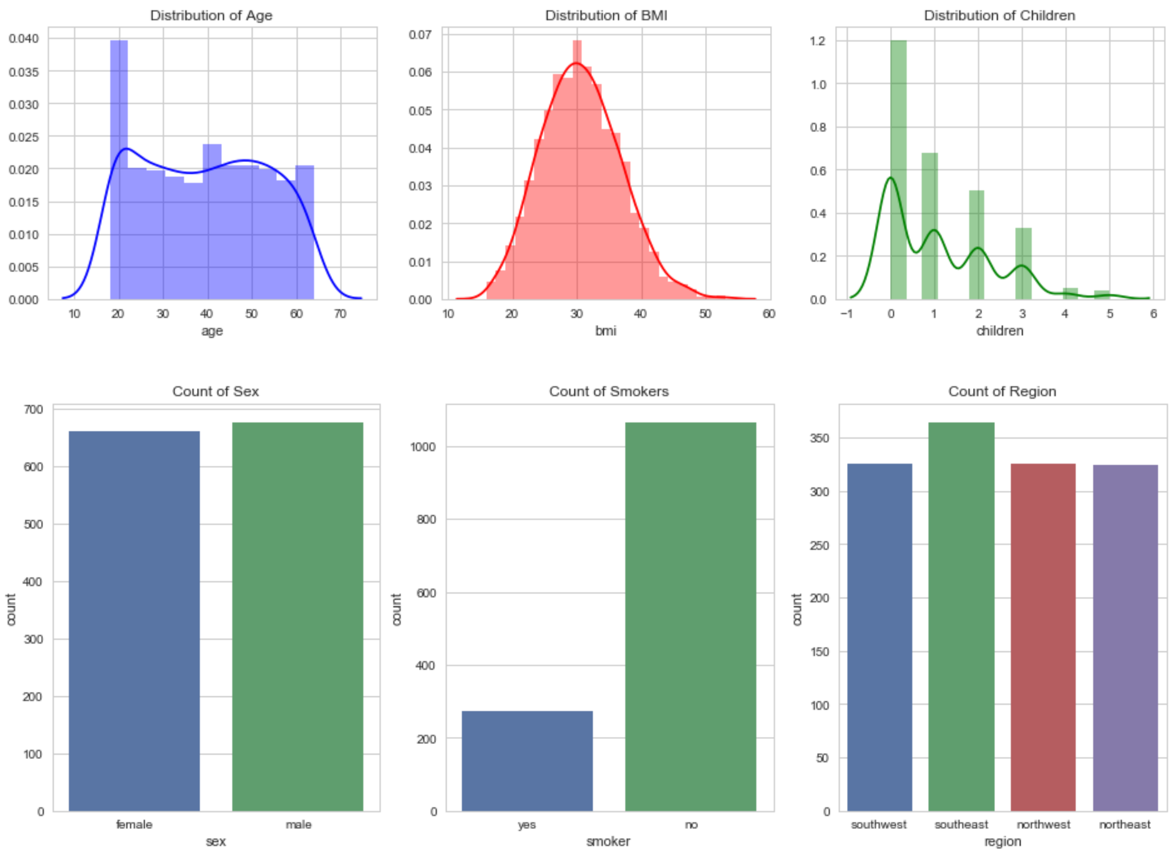
The aim of the regression model will be to **predict** the charges based on customer information, NOT interpretation.

Distribution of Data

The charges is the target variable and it was found to be not normally distributed so a log transformation was applied.



Just to visualise how the rest of the data is distributed, the features were also plotted.

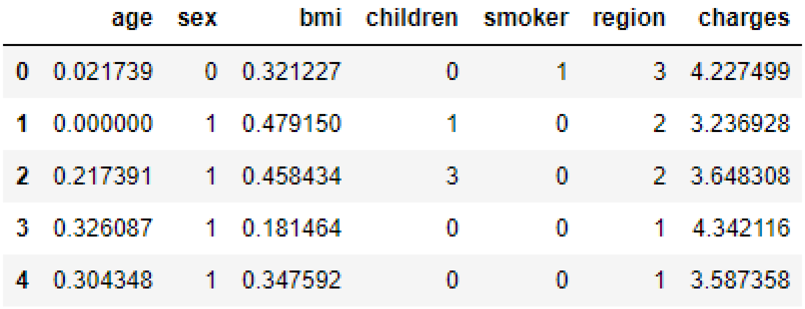


Feature Engineering

Categorical features (sex, smoker, region) were converted to numerical using Label Encoding:

- Sex: Male = 1, female = 0
- Smoker: Yes = 1, No =0
- Region: Southwest = 3, Southeast = 2, Northwest = 1, Northeast = 0 (another way of doing this could have been with one-hot encoding)

Age and BMI features were scaled down using MinMax scaling. After these two steps and taking log of the target, the data set looked as follows,



	age	sex	bmi	children	smoker	region	charges
0	0.021739	0	0.321227	0	1	3	4.227499
1	0.000000	1	0.479150	1	0	2	3.236928
2	0.217391	1	0.458434	3	0	2	3.648308
3	0.326087	1	0.181464	0	0	1	4.342116
4	0.304348	1	0.347592	0	0	1	3.587358

Regression Models

A total of **6 models** were explored after splitting the data into **training and testing** sets (70:30).

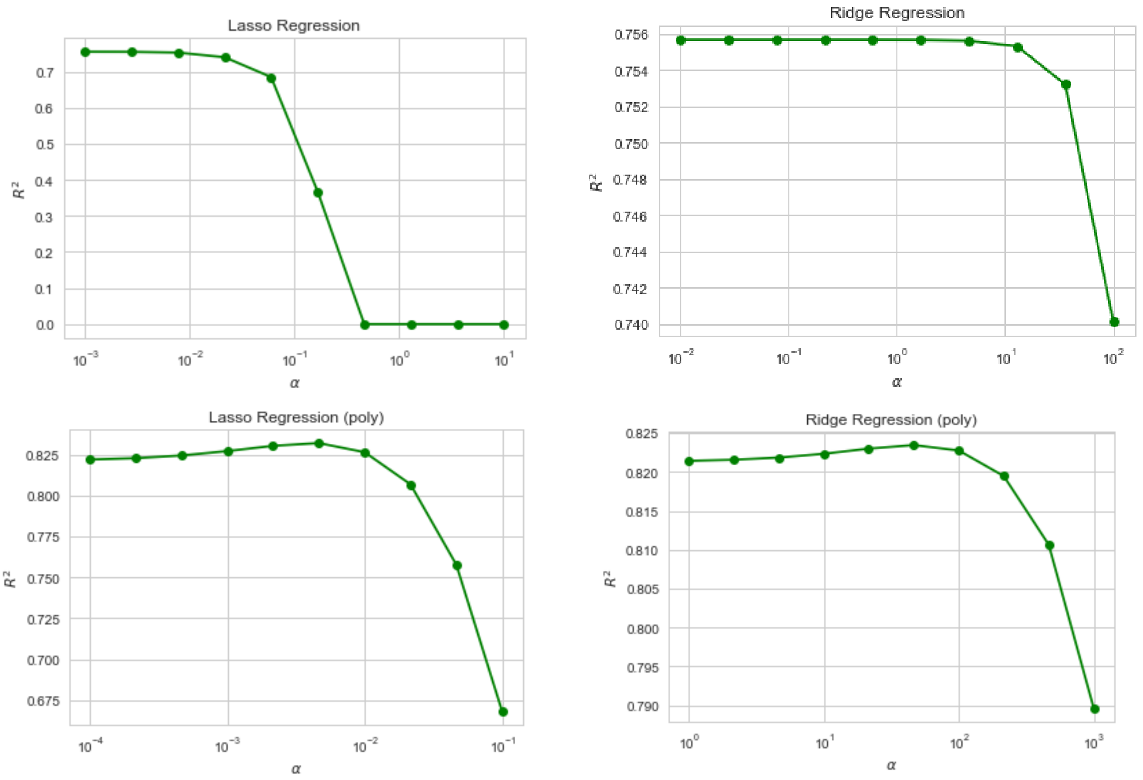
3-Fold cross validation sets were also made.

The models were as follows:

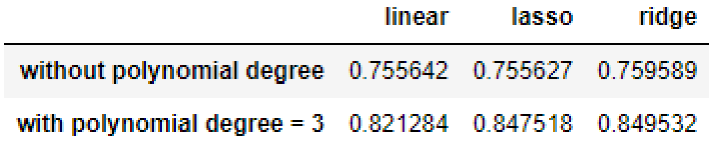
- Linear Regression
- Lasso Regression [alpha = 0.001]
- Ridge Regression [alpha = 0.1]

- Polynomial Regression with degree 3
- Polynomial Lasso Regression with degree 3 [alpha = 0.003]
- Polynomial Ridge Regression with degree 3 [alpha = 60]

The alpha values were chosen on the basis of the results of regularization shown below.

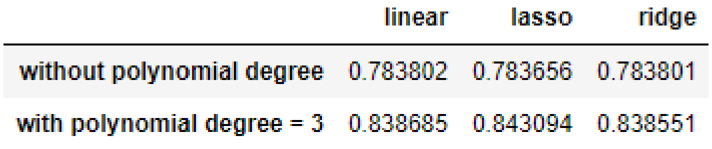


The accuracy of the models on the training set is given in the table below,



	linear	lasso	ridge
without polynomial degree	0.755642	0.755627	0.759589
with polynomial degree = 3	0.821284	0.847518	0.849532

As seen, without polynomial feature, linear, lasso and ridge all three are comparable. With polynomial feature, regularization shows better accuracy. The testing accuracy was found as,



	linear	lasso	ridge
without polynomial degree	0.783802	0.783656	0.783801
with polynomial degree = 3	0.838685	0.843094	0.838551

The testing accuracy is also consistent with the training accuracy and indicates the same results with the best model being **polynomial lasso regression**. The models without polynomial features were relatively underfitting.

Conclusion

In conclusion, the model with polynomial degree 3 and alpha 0.003 for Lasso regularization turned out to be the best of the chosen models. Different models can also be explored, and different prediction algorithms can also be explored to find better models. Furthermore, interpretation of results can also be done.