Insurance Premium - Supervised Learning - Regression

Kushagra Singhal

2021-08-26

1 Objective

The objective of this report is to present the prediction of basic insurance premiums given some attributes of the insured person. We will also touch upon interpretation of the model in terms of different predictors used.

2 Dataset and Attributes

The dataset contains information about basic insurance premiums charged to 1338 individuals. The set of features included are age, sex, body-to-mass index, number of children, smoker flag, and location region. Table 1 represent a sample of the data. Such dataset can be used to predict insurance premiums for new customers.

The information about the data columns is also provided in table 2. As can be seen, there are total 6 features and 1 target column called charges. The data types are a mix of integer, floating and object. There are a total of 1338 samples in the dataset.

Some of the descriptive statistics of the data are also shown in table 3. We can observe some statistics like the mean age is around 39 years with the range being 18 to 64 years. This makes sense as the insurance buyers are generally adults and insurance providers do not want to provide cover to very aged people. We also observe that the dataset is more or less even in terms of sex of the individuals. In terms of smoking behavior, the dataset is skewed towards non smokers.

Table 1: Sample Data

	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

Table 2: Information about columns in the dataset

	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

Table 3: Descriptive statistics for the dataset

statistic	age	sex	bmi	children	smoker	region	charges
count	1338	1338	1338	1338	1338	1338	1338
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.2	NaN	30.66	1.09	NaN	NaN	13270.42
std	14.05	NaN	6.1	1.20	NaN	NaN	12110.01
min	18.0	NaN	15.96	0.0	NaN	NaN	1121.87
25%	27.0	NaN	26.3	0.0	NaN	NaN	4740.28
50%	39.0	NaN	30.4	1.0	NaN	NaN	9382.03
75%	51.0	NaN	34.69	2.0	NaN	NaN	16639.91
max	64.0	NaN	53.13	5.0	NaN	NaN	63770.42

3 Data Exploration Summary

Although we looked at some descriptive statistics of the dataset in section 2, it will be more insightful to explore the data further. In the following subsections, we look at the dataset in more details.

3.1 Distributions

Figure 1 shows the distributions for numerical columns of the dataset. As can be observed, BMI is is almost normally distributed. The age feature is more or less uniformly distributed with some skew towards lowest ages. Number of children and charges seem to be exponentially distributed. No outliers seem to exist for these features.

Figure 2 shows the bar plots for the categorical features. As can be observed, male to female ratio is almost 1. There are more non-smokers than smokers. Also, the spread of individuals across regions is pretty uniform.

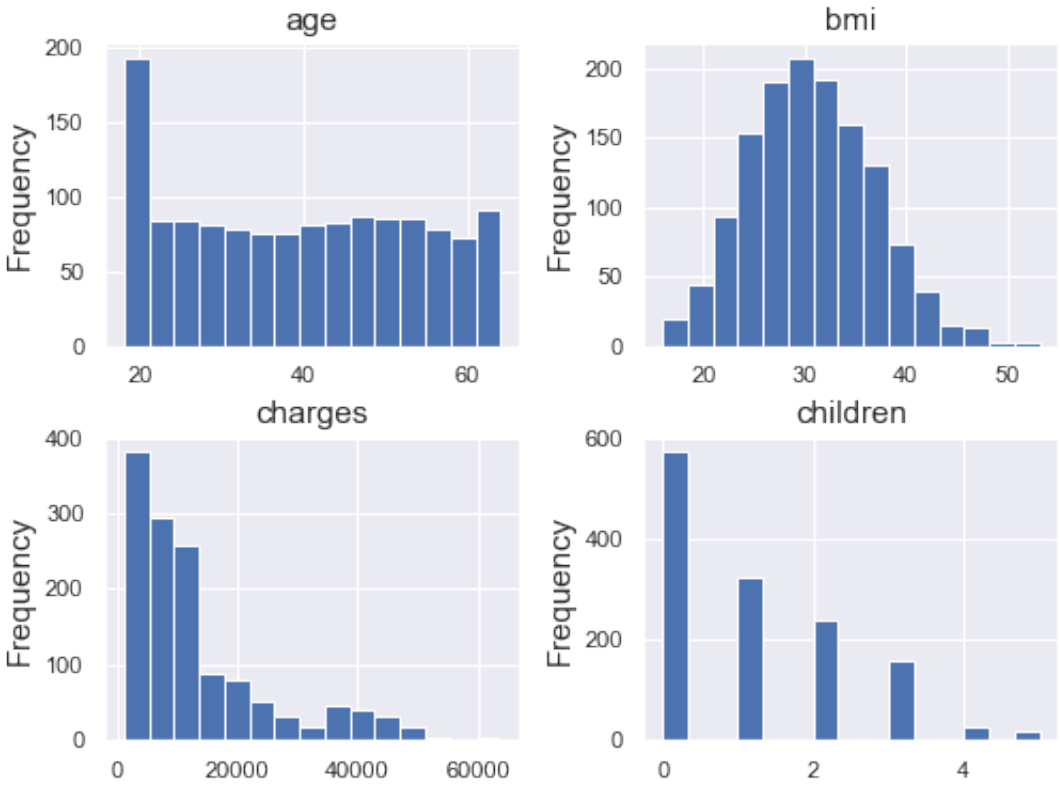

Figure 1: Distributions for the numerical columns in the dataset

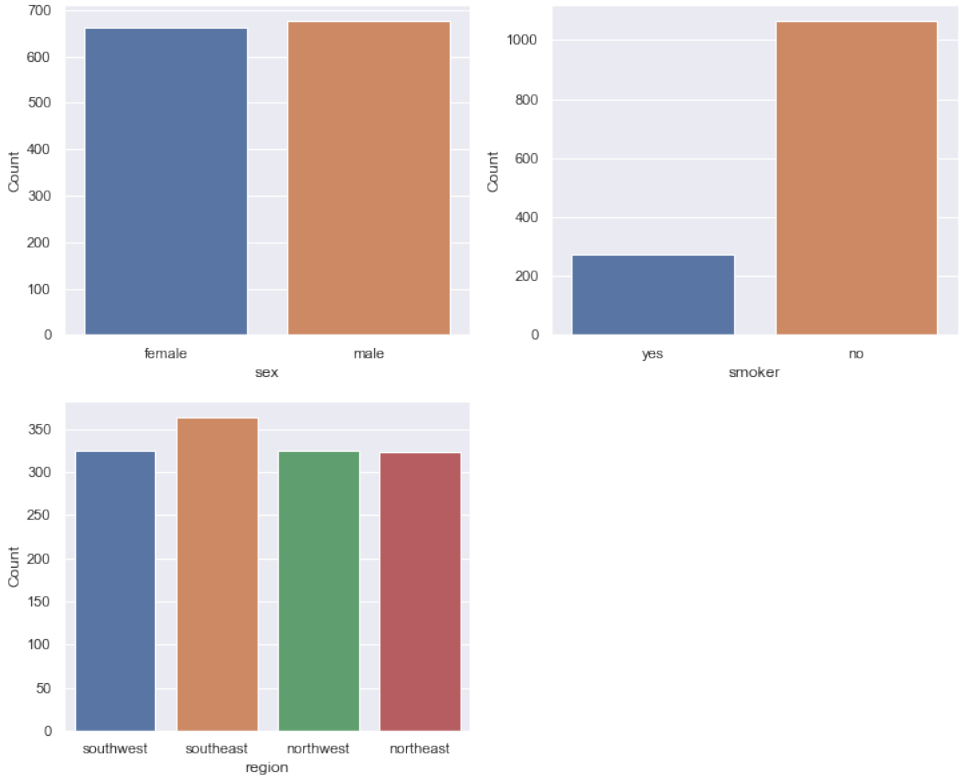

Figure 2: Bar plots for the categorical columns in the dataset

3.2 Premium Relationship with Features

Next, we are interested in exploring the relationship of charged premiums with the features in the dataset. For this, we plot the correlations of each pair of columns. The plots are shown in figure 3. As can be observed from the charts in the last row, the premium charges are seem to be correlated with age, bmi but not very strongly with number of children.

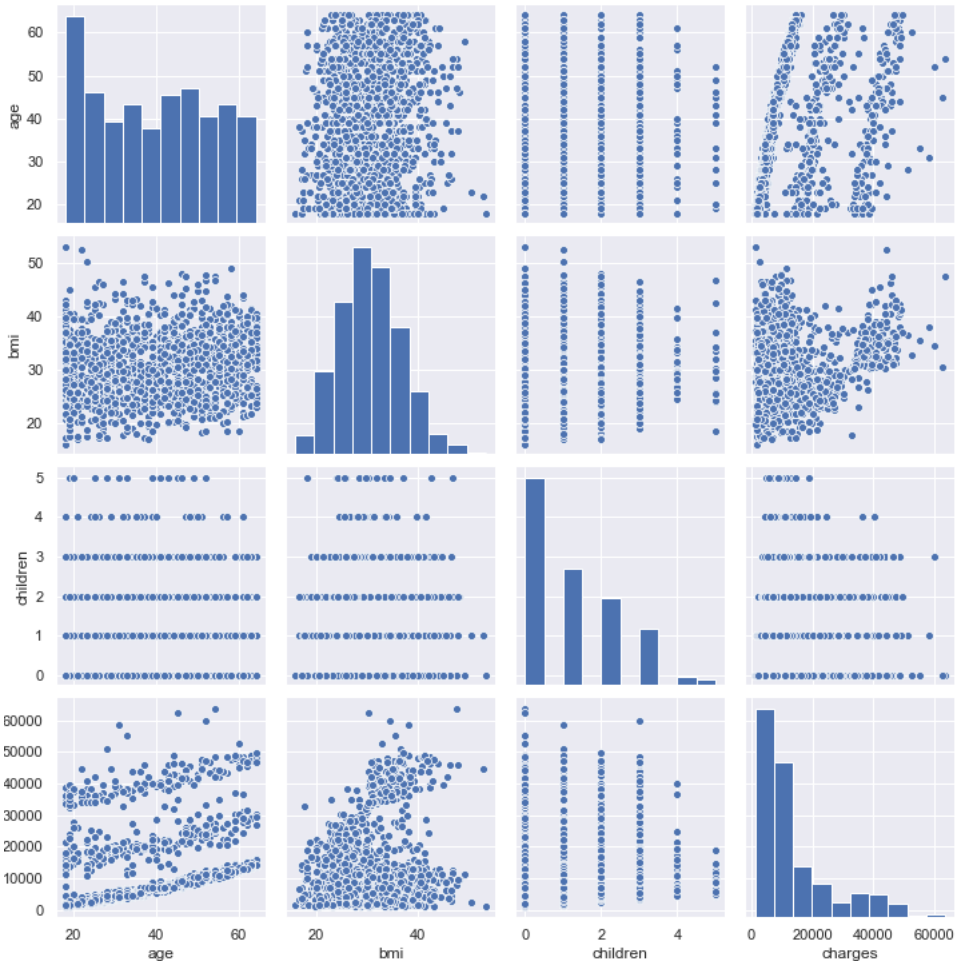

Figure 3: Pair plots for the numerical columns in the dataset

The box plots for different categorical features are shown in figure 4. It can be observed that the premiums charged are clearly higher for smokers. There is no clear distinctions based on sex but we see a larger variance for males. In terms of regions, southeast region has the highest variance whereas the median value is slightly higher for north east region.

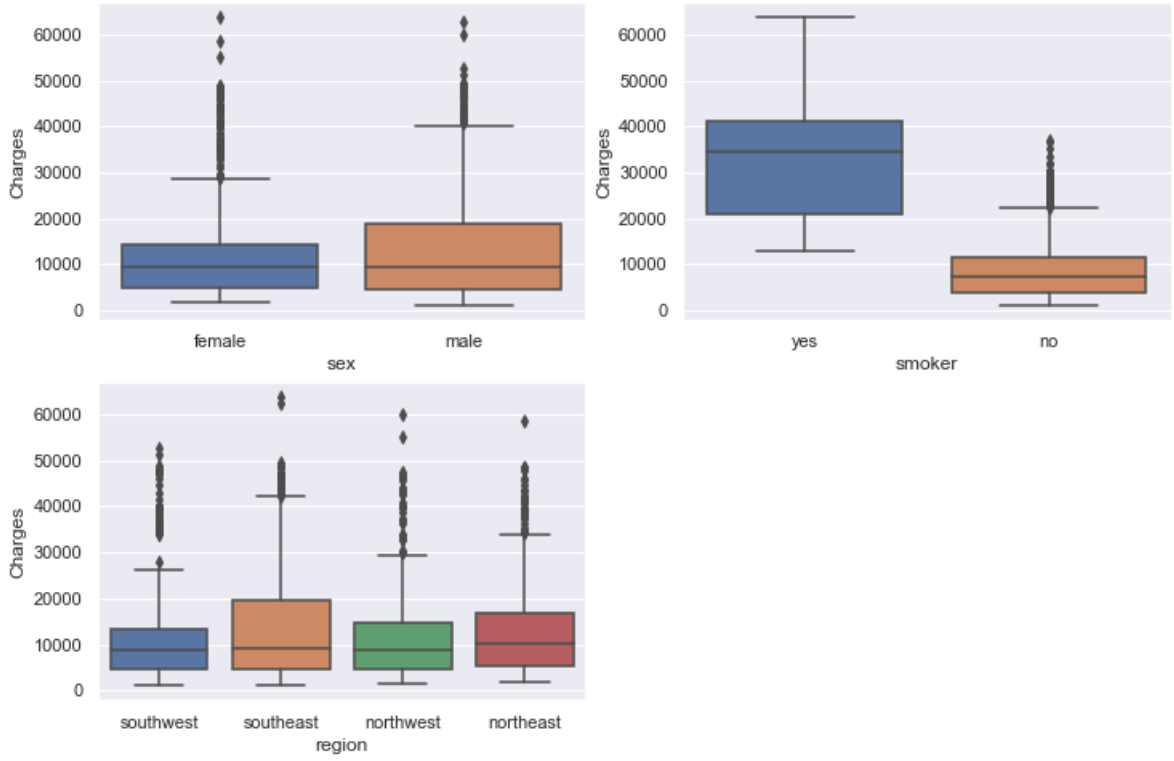

Figure 4: Box plots for the categorical columns in the dataset

3.3 Data Cleaning and Feature Engineering

As already established, there are no missing values for any of the features or the target. As such, cleaning is no needed. However, the dataset contains some categorical features which we will need to convert to numerical ones to be able to use those for learning models. All the three categorical features- sex, smoker, and region - are non-ordinal. Hence, we will simply use the one-hot encoding to convert them to numerical features. The sample data after applying one hot encoding is shown in table 4.

Looking at the data, we do not see use of any scaling that is necessary. We will scale the date when we use regularization techniques.

Table 4: Sample dataset after applying one hot encoding

	age	bmi	children	charges	sex male	smoker yes	region northwest	region southeast	region southwest
0	19	27.9	0	16884.9	0	1	0	0	1
1	18	33.8	1	1725.6	1	0	0	1	0
2	28	33.0	3	4449.5	1	0	0	1	0
3	33	22.7	0	21984.5	1	0	1	0	0
4	32	28.9	0	3866.8	1	0	1	0	0

3.4 Key Findings

Below are some of the findings from the data until now:

- There are three numerical and three categorical features to work with
- The categorical features were non-ordinal and were converted to numerical features using one hot encoding
- Pair plots seem to suggest that age and BMI are correlated with premium charges whereas number of children is uncorrelated
- Smokers seem to pay higher premiums compared to non-smokers

3.5 Hypothesis Testing

Three possible hypothesis are mentioned below:

- Number of children does not affect the insurance premiums
- Average premiums for smokers are higher than non-smokers
- Average premiums for males are equal to females

3.5.1 Is there a gender bias?

Let the true mean of the premiums for males and females be μ_m and μ_f respectively. Then, we perform the following hypothesis testing:

$$H_0 : \mu_m = \mu_f \tag{1}$$
$$H_1 : \mu_m \neq \mu_f \tag{2}$$

This is an important hypothesis to test because the sex of an individual should not be a major factor for deciding the insurance premiums. Such discriminating practices may not be seen in the best light in today's world.

To test this hypothesis, we first use the independent two sample two-tailed t-test which is already implemented in python packages. We perform this test with equal variance assumption as true and false respectively. The significance level chosen is 5%. The results are shown in table 5. As observed, the p-values are below the significance level of 5% and hence we fail to reject the null hypothesis of no bias. However, it should e noted that if the significance level was 3% or lower, we would have not rejected the null hypothesis.

Table 5: T-test results for the hypothesis test for gender bias

Equal Variance	t-statistic	p-value	Decision
True	2.0975	0.0361	Reject
False	2.100	0.0358	Reject

One point of discussion may be the choice of significance level (or confidence interval). As such kind of bias can may have significant ramifications on the insurance company, such as discriminatory practices charges, we should make sure that we only reject the null hypothesis when it is very highly likely that a bias is present. Hence, a confidence interval of 99% may be more appropriate.

To summarize, the dataset contains attributes of individuals along with their insurance premiums. We did not find any outliers or missing values. As such, the dataset was of good quality. However, some of the attributes like sex and number of children may be of limited predictive value. Some more features like medical history, drinker flag, employment type, travel history, adventure sports interest etc. could be of more use as well.

4 Regression Models

In this section, we will look at three regression models. First, is the multiple linear regression model which serves as a baseline model. Second, we look at a polynomial regression model where we include the higher order terms and capture feature interactions. Last, we look at the regularized regression using L_1 penalty, i.e., LASSO regression. The LASSO regression will help

trim the feature selection from the polynomial model and help in better interpret-ability.

4.1 Baseline - Linear Regression

The most straightforward model is a multiple linear regression model. This was implemented using sklearn in python. None of the features has been dropped from the dataset. We have not used the intercept here.

4.2 Polynomial Regression

The next model explored is the polynomial regression. The higher degree and interaction terms are considered only for the numerical features. We then proceed with the linear regression including these additional features.

4.3 LASSO Regression

We then perform the LASSO regression without scaling the features as created in the last regression. The aim is to see if LASSO makes any feature selection and improves results.

4.4 LASSO Regression with scaling

Next we include the scaling of features in the LASSO regression. The scaling of features helps in understanding the importance of features.

5 Summary of Results

This section presents a summary of results obtained after training and testing the different models. The train test split is 70:30 and the splits are same across models. The best hyper-parameters were selected based on 5-fold cross validation and were same for the LASSO regressions.

Table 6 and 7 show the RMSE and coefficients of determination for different methods. It can be observed that the performances are very similar. There is only a marginal increase in performance when polynomial features are included or when LASSO is used. One of the reasons is that the features like age and number of children may not influence the premium charges a

Table 6: Comparison of RMSE for different regression methods

Method	Training RMSE	Test RMSE
LinearRegression	6479.454154	6099.792116
PolynmialLinearRegression	6066.250150	5826.564691
LassoPolynmialLinearRegression	6086.317515	5822.059212
LassoScalePolynmialLinearRegression	6102.406022	5790.377776

Table 7: Comparison of R^2 for different regression methods

Method	Training R^2	Test R^2
LinearRegression	0.844608	0.863852
PolynomialLinearRegression	0.865307	0.876621
LassoPolynomialLinearRegression	0.864344	0.876825
LassoScalePolynomialLinearRegression	0.863569	0.878254

lot. As similar story is posted in figure 5 where the scatter plots are very similar.

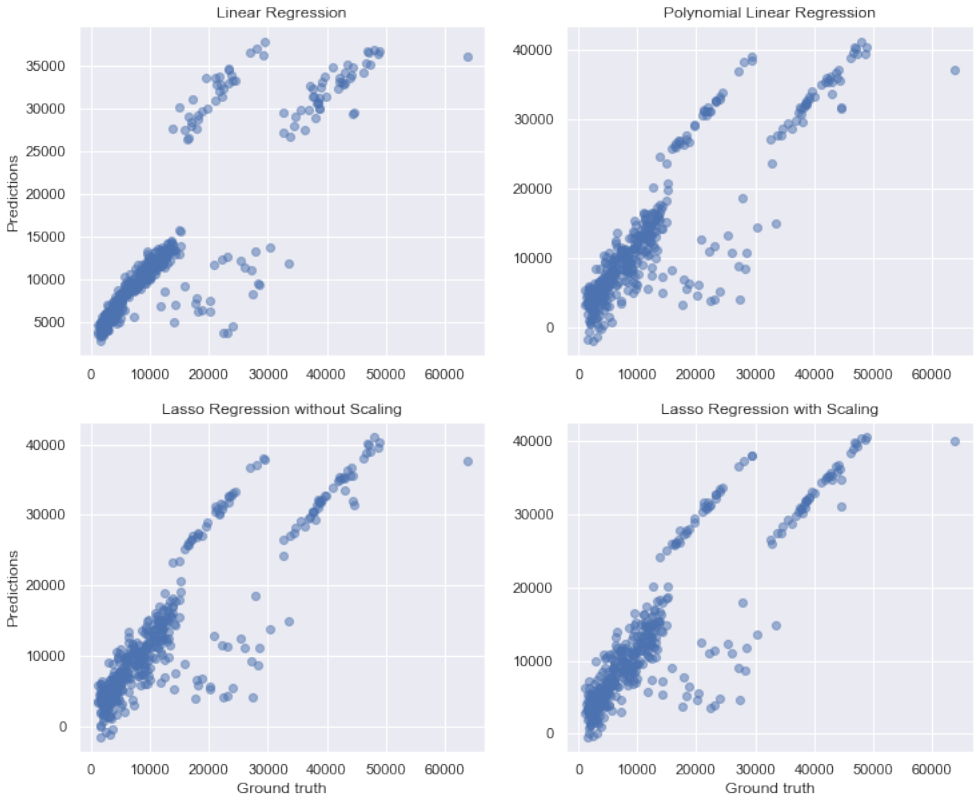

Figure 5: Predictions vs Ground Truth scatter plots

Figures 6 to 9 depict the importance of features in different regressions. It can be noticed that in all cases the smoking attribute leads to a large insurance premium. The BMI attribute is also important as per all regressions but the linear regression. One of the striking differences between the scaled LASSO and other models is the region dependence. One would not expect the premiums to vary much across regions. In the scaled LASSO, the negative dependence on children squared is a bit unintuitive though. However, it seems that the data as seen in figure 3 points to the higher charges for lesser number of children.

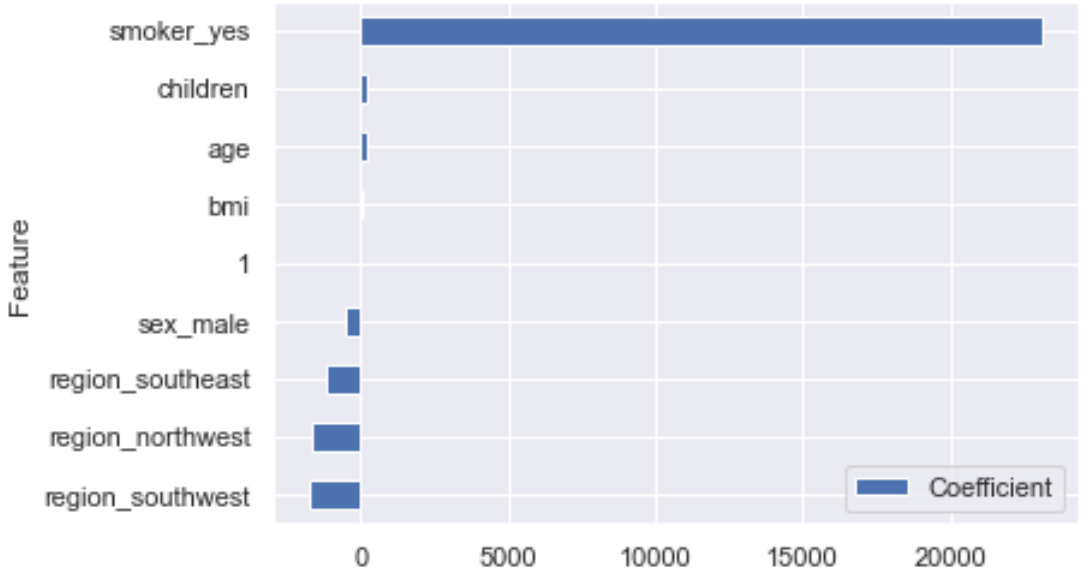

Figure 6: Feature Importance - Linear Regression

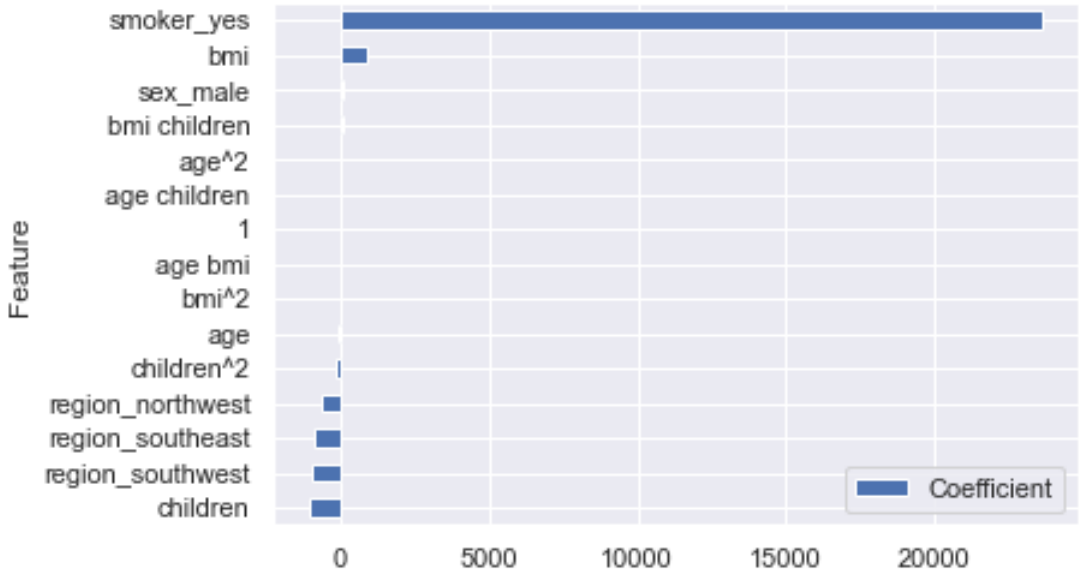

Figure 7: Feature Importance - Polynomial Regression

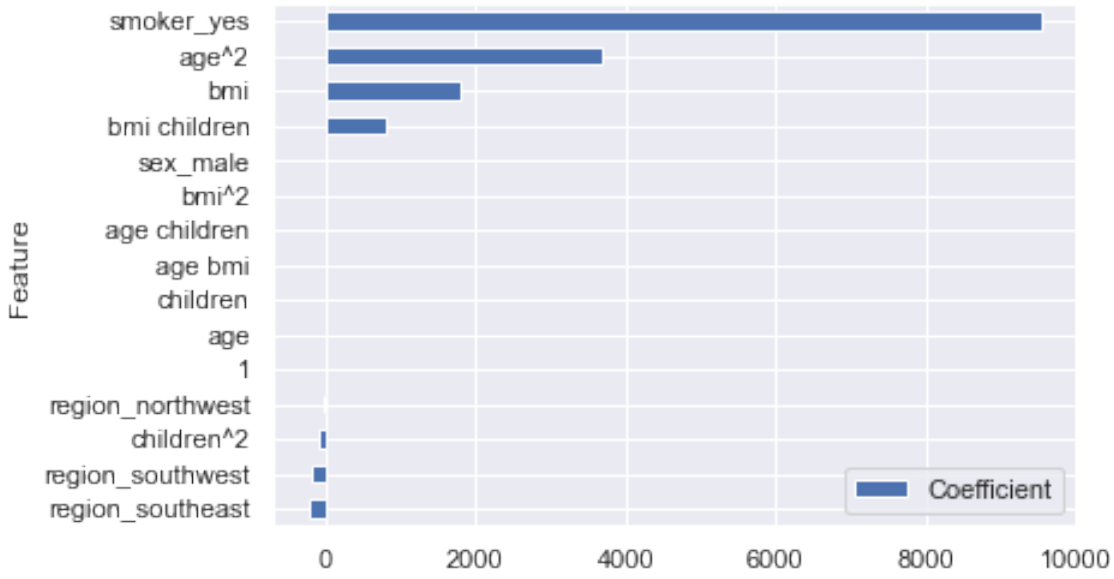

Figure 8: Feature Importance - LASSO Regression

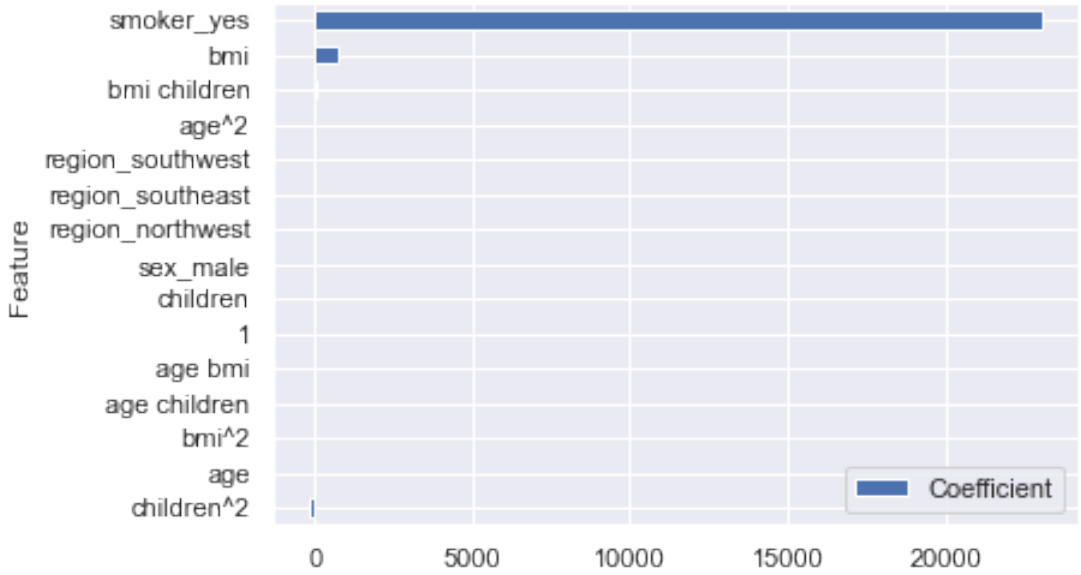

Figure 9: Feature Importance - Scaled LASSO Regression

6 Model Choice

There is not much to gain in terms of model performance as indicated by the RMSE and R-squared metrics. However, The scaled LASSO is the best choice among the alternatives. It is marginally better in terms of performance when compared to other models. Further, the interpretation of coefficients also makes more sense. For example, smokers are charged way more than non smokers, people with high BMI are also charged a bit more.

7 Key Findings and Insights

Overall, it is observed that the linear regression using LASSO does a good job explaining the variance (close to 87%). Further, the error is also limited to under 5800 units on the test set. The major drivers of the insurance premiums are smoking behavior and BMI which makes sense as well. It was also observed that adding polynomial features was not very beneficial in terms of performance but interpretation is a bit better with LASSO and polynomial features.

8 Next Steps

Some of the next steps could be:

- Higher degree polynomials: it would exciting to see how LASSO performs with polynomials of degree three and higher.
- Stepwise Regression: Another model that can be explored is stepwise regression. Such a model will also help select more important features and get rid of redundant features.
- Ridge Regression: One can also explore how the ridge regression performs in comparison to the LASSO.
- Some more transformations on the data can be explored like reducing skewness, or taking logarithms. In practice, such steps are generally useful and can lead to better performance.