Medical expenditure prediction

Portfolio project

Rehan Elahi

Contents

1 Problem description

In this project, we will predict annual medical expenditures for health insurance buyers. We are given the following features:

- **age**: Age of primary beneficiary.
- **sex**: Gender of the beneficiary.
- **BMI**: Body mass index.
- **children**: Number of children covered by health insurance.
- **Smoker**: Smoking status.
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges (the dependent variable)**: Individual medical costs billed by the health insurance company.

For context, health insurance is a contract between a company and a consumer. The company agrees to pay all or some of the insured person's healthcare costs in return for payment of a monthly premium. Predicting medical expenditures correctly would help the company charge appropriate premiums.

2 Methodology

I will first use exploratory data analysis (EDA) techniques to learn more about the dataset. Then I will move on to the modelling part, where I will use a linear and non-linear regression modelling technique and find out which one suits this problem the best. Lastly, I will conclude with a section on the performance of each model and their limitations. All code will be available in a GitHub repository.

3 Exploratory Data Analysis (EDA)

3.1 Getting a summary of the dataset

```
      age            sex                 bmi           children       smoker             region
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    charges
 Min.   : 1122
 1st Qu.: 4740
 Median : 9382
 Mean   :13270
 3rd Qu.:16640
 Max.   :63770
```

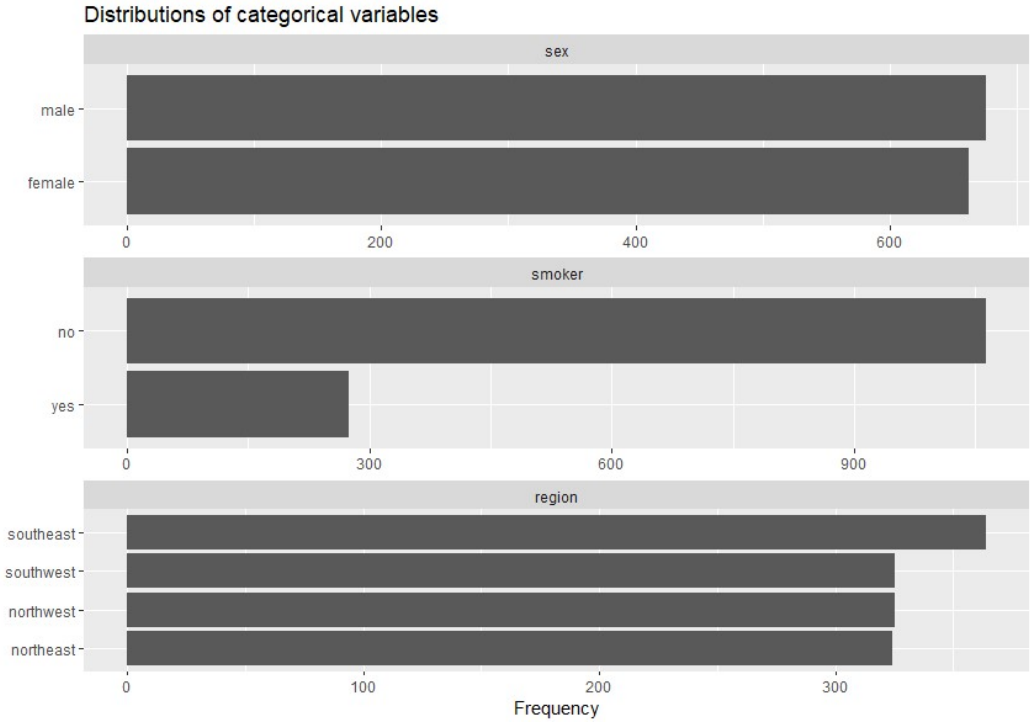
We can see that we have three independent numerical variables in age, children and bmi and three independent categorical variables in sex, smoker and region.

3.2 Checking for invalid and unique values

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
age	0	0.0000000	0	0	0	0	integer	47
sex	0	0.0000000	0	0	0	0	character	2
bmi	0	0.0000000	0	0	0	0	numeric	548
children	574	0.4289985	0	0	0	0	integer	6
smoker	0	0.0000000	0	0	0	0	character	2
region	0	0.0000000	0	0	0	0	character	4
charges	0	0.0000000	0	0	0	0	numeric	1,337

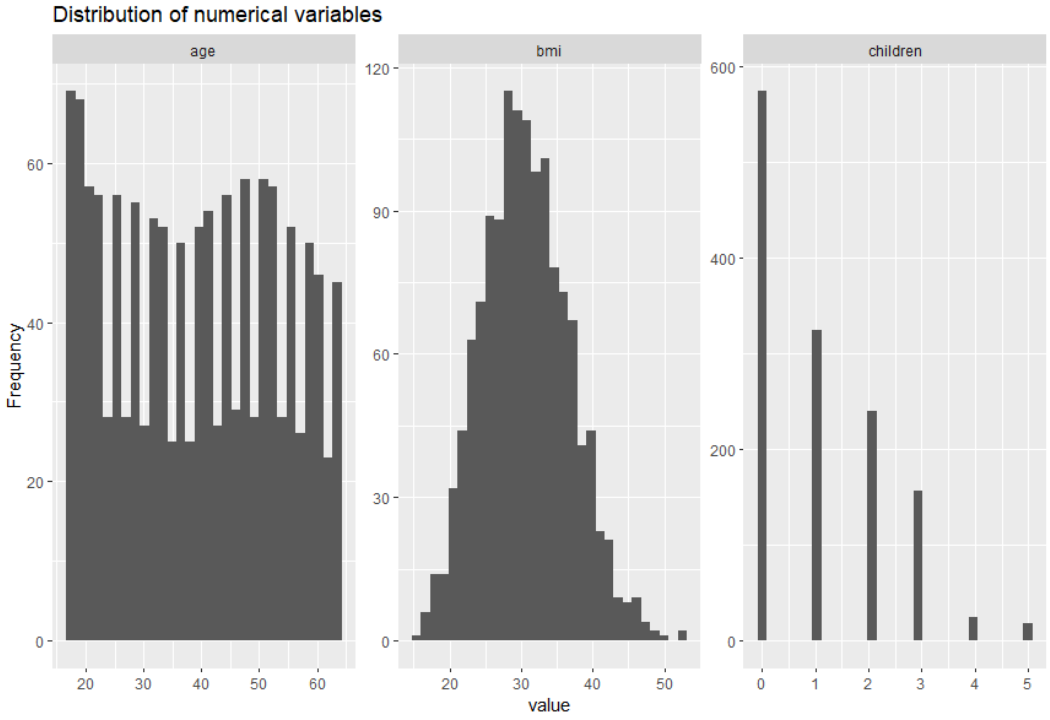
There are no invalid (NA, Inf) values for any of our independent variables. So, we do not need to replace any values by imputation methods. There are zeroes in the feature children as they represent a genuine category. They are not to be mistaken as an invalid value.

3.3 Plotting distributions for categorical features



The categorical features, sex and region have pretty much equal number of values for each category whereas the feature, smoker has greater number of observations for the category 'no' than 'yes'.

3.4 Plotting distributions for numerical features



Out of the three, we have two continuous numerical variables, in age and BMI and one discrete numerical variable, in BMI. We can see that BMI's distribution looks like a normal distribution, but we would have to confirm it using the Shapiro-Wilk Test, which we will do next.

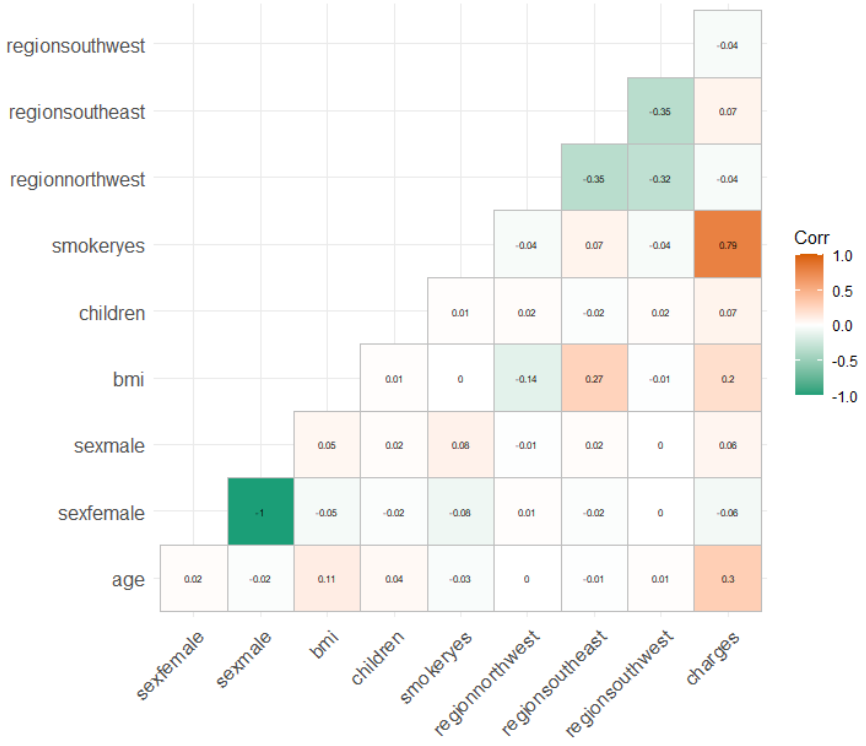
3.5 Confirmation of normality for BMI

```
        Shapiro-Wilk normality test

data:  EDAdata$bmi
W = 0.99389, p-value = 2.605e-05
```

For the distribution to be normal, the p-value must be greater than 0.05. It is not in this case and hence we would need to perform a transformation when we use this predictor for the linear regression model. For other models, we do not need to.

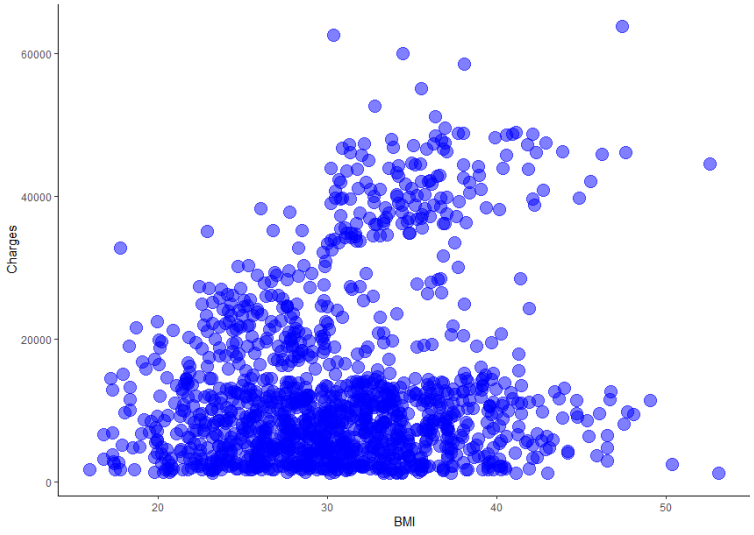
3.6 Correlation Matrix



As can be seen from the correlation matrix, smoker, age and BMI are positively correlated to charges. The variables sex, children and region have very little correlation with charges.

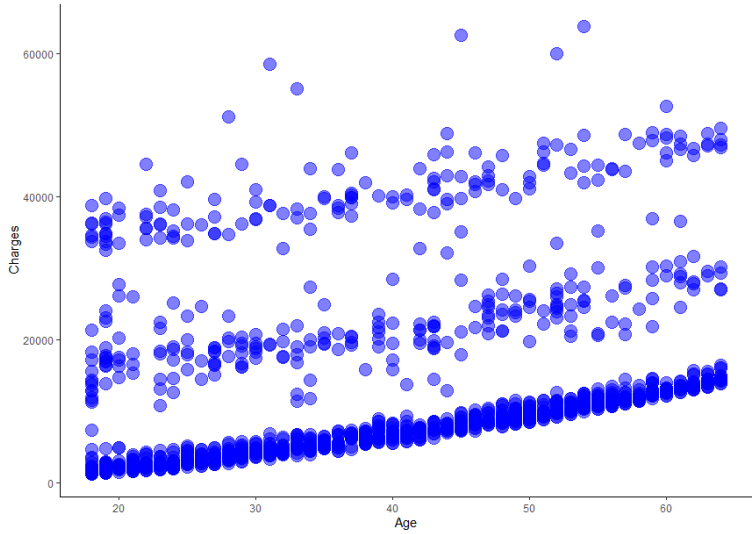
3.7 Inspecting relationships between independent continuous numerical variables and our dependent variable

3.7.1 Charges vs. BMI



The variable BMI is positively correlated with charges. As the BMI increases, so do the charges. This also makes sense, as the more obese a person is, the greater the chance of health problems. The relationship, however, seems non-linear.

3.7.2 Charges vs. Age



The variable age is positively correlated with charges. As the age increases, so do the medical charges.

3.8 Outlier Check

3.8.1 Age

```
OK: No outliers detected.
- Based on the following method and threshold: iqr (1.7).
- For variable: EDAdata$age
```

We have no outliers for age.

3.8.2 BMI

```
4 outliers detected: cases 117, 848, 1048, 1318.
- Based on the following method and threshold: iqr (1.7).
- For variable: EDAdata$bmi.

--------------------------------------------------------------------------------
Outliers per variable (iqr):

$`EDAdata$bmi`
      Row Distance_IQR
117   117     1.300457
848   848     1.392922
1048 1048     1.547029
1318 1318     1.585556
```

For BMI, we find 4 outliers. It is better to drop them before we use models which are sensitive to outliers.

4 Modelling

4.1 Multiple Linear Regression

Now that the EDA is finished, we can begin to model the relationship between our dependent and independent variables. The first technique we are going to use is multiple linear regression. Before we can progress to modelling, we will have to carry out data-preprocessing to make the data ready to be fed into the model.

4.1.1 Encoding categorical variables:

The categorical variables smoker, sex and region were encoded. Refer to the code to see the values.

4.1.2 Transforming BMI & Age to fit normal distribution

Since linear regression requires numerical variables to be normally distributed, we try a range of transformations to do so and select the best ones.

```
Best Normalizing transformation with 1338 Observations
 Estimated Normality Statistics (Pearson P / df, lower => more normal):
 - arcsinh(x): 1.1583
 - Box-Cox: 1.0303
 - Center+scale: 1.1676
 - Exp(x): 140.9823
 - Lambert's W (type s): 1.0509
 - Log_b(x+a): 1.156
 - orderNorm (ORQ): 1.1909
 - sqrt(x + a): 1.0441
 - Yeo-Johnson: 1.0299
Estimation method: Out-of-sample via CV with 10 folds and 5 repeats

Based off these, bestNormalize chose:
Standardized Yeo-Johnson Transformation with 1338 nonmissing obs.:
 Estimated statistics:
 - lambda = 0.4431716
 - mean (before standardization) = 8.129085
 - sd (before standardization) = 0.8927178
> shapiro.test(newdatalr$bmi)

        Shapiro-Wilk normality test

data:  newdatalr$bmi
W = 0.99857, p-value = 0.3406
```

For BMI, we see that Standardized Yeo-Johnson transformation was the best one and as confirmed by the Shapiro-Wilk test (p value > 0.05), we have successfully managed to transform BMI to fit a normal distribution.

```
Best Normalizing transformation with 1338 Observations
 Estimated Normality Statistics (Pearson P / df, lower => more normal):
 - arcsinh(x): 4.0296
 - Box-Cox: 2.9796
 - Center+scale: 3.7771
 - Exp(x): 130.4232
 - Lambert's W (type s): 3.2383
 - Log_b(x+a): 4.0341
 - orderNorm (ORQ): 1.3107
 - sqrt(x + a): 2.8459
 - Yeo-Johnson: 2.9808
Estimation method: Out-of-sample via CV with 10 folds and 5 repeats

Based off these, bestNormalize chose:
orderNorm Transformation with 1338 nonmissing obs and ties
 - 47 unique values
 - Original quantiles:
  0%  25%  50%  75% 100%
  18   27   39   51   64
```

For transforming age to fit a normal distribution, we use the orderNorm Transformation technique. But lets see whether the Shapiro-Wilk test confirms the normality:

```
> shapiro.test(newdata$age)

        Shapiro-Wilk normality test

data:  newdata$age
W = 0.9447, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  output_age$data_Transformed
W = 0.98939, p-value = 2.879e-08
```

The first Shapiro-Wilk Test screenshot shows the p-value before age was transformed and it was obvious, as seen in the EDA, that it was not normally distributed. The second screenshot tells us that even after transformation, it fails to fit a normal distribution, but the p-value improves by almost double. So, we would go ahead with the transformed values of age for modelling.

4.1.3 Outlier removal

The outliers for BMI, identified in the EDA section, were removed.

4.1.4 Train/Test split

Before proceeding to the modelling, we need to divide the dataset into a test and training set. We use a ratio of 0.8 to do so.

4.1.5 Multiple linear regression models

We build three regression models, based on selecting different features. Our first model 'regressor1' selects all independent variables to predict values for the dependent variables. Shown below, is the summary of the model:

```
Call:
lm(formula = charges ~ age + sex + smoker + bmi + children +
    region, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-11135.9  -2927.4   -948.6   1303.3  30104.2

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11874.07    1088.62 -10.907   <2e-16 ***
age                263.41      13.17  20.007   <2e-16 ***
sexmale            -66.46     368.60  -0.180   0.8569
smokeryes        23644.60     455.76  51.879   <2e-16 ***
bmi                326.51      31.49  10.368   <2e-16 ***
children           373.35     153.42   2.433   0.0151 *
regionnorthwest   -138.36     531.47  -0.260   0.7947
regionsoutheast   -981.53     528.83  -1.856   0.0637 .
regionsouthwest  -1058.57     534.69  -1.980   0.0480 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6002 on 1058 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.7528
F-statistic: 406.9 on 8 and 1058 DF,  p-value: < 2.2e-16
```

We see that the variable, sex, is statistically insignificant and hence, in the next model, 'regressor2' we will drop it.

```
Call:
lm(formula = charges ~ age + smoker + bmi + children + region,
    data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-11168.6  -2918.1   -962.5   1307.4  30075.2

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11901.69    1077.30 -11.048   <2e-16 ***
age                263.40      13.16  20.016   <2e-16 ***
smokeryes        23639.48     454.67  51.993   <2e-16 ***
bmi                326.32      31.46  10.373   <2e-16 ***
children           373.19     153.35   2.434   0.0151 *
regionnorthwest   -137.11     531.18  -0.258   0.7964
regionsoutheast   -980.42     528.56  -1.855   0.0639 .
regionsouthwest  -1055.81     534.23  -1.976   0.0484 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5999 on 1059 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.7531
F-statistic: 465.4 on 7 and 1059 DF,  p-value: < 2.2e-16
```

We see that after dropping sex, the R-squared value remains the same, but the adjusted R-squared value increases from 0.7528 to 0.7531, which is an indication that the decision to drop sex was right. Let's see if dropping region can further improve the r-squared and adjusted R-squared values:

```
call:
lm(formula = charges ~ age + smoker + bmi + children, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-11844.7  -2982.5   -942.2   1303.8  29673.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11917.50    1038.78 -11.473   <2e-16 ***
age            263.88      13.17  20.033   <2e-16 ***
smokeryes    23610.98     454.03  52.003   <2e-16 ***
bmi            308.00      30.19  10.201   <2e-16 ***
children       376.76     153.46   2.455   0.0142 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6009 on 1062 degrees of freedom
Multiple R-squared:  0.7532,    Adjusted R-squared:  0.7523
F-statistic: 810.3 on 4 and 1062 DF,  p-value: < 2.2e-16
```

Although, we see that now all features are statistically significant, there is a decrease in both r-squared and adjusted r-squared values for this model, 'regressor3'. In the next section, we find out the Root-Mean Squared Errors (RMSE) and the Normalized Root-Mean Squared Errors (NRMSE) for all these models, which will help us select the best one, amongst the multiple linear regression models.

4.1.6 Error metrics

```
        Metric        Regressor1        Regressor2        Regressor3
1         RMSE   6282.1508601887  6282.63540371758  6278.52753443021
2 NRMSE_minmax 0.109369744447122 0.109378180156977 0.109306663788749
```

We see that The RMSE for Regressor3 is the best at $6278.52. However, the RMSEs for Regressor2 and Regressor3 aren't much different. Regressor2 gave us the best combination of R-Squared and adjusted R-squared values, thereby explaining the most variance and because the RMSEs aren't much different, we would choose Regressor2 as the best multiple linear regression model.

4.2 Support Vector Regression

This is the second regression technique that we are going to use. We will be choosing the radial basis kernel function for the modelling. All the steps for the data-preprocessing bit are the same except that we would not need to remove outliers for BMI as the support vector regression algorithm is robust to outliers. Also, we would not try to fit age and BMI to a normal distribution because there's no need to do so. We will use simple standardization using the scale() function in R, to feature scale these variables.

4.2.1 Default model

For the default support vector regression model, I chose all the independent variables to form a relationship with the dependant variable. The values of the hyper-parameters cost, gamma and epsilon were what they were set to as default, 1, 0.1 and 0.1 respectively. It gave me the following values of RMSE and R-squared:

RMSE	Rsquared	MAE
5196.052855	0.810397	3128.190417

The values of RMSE and R-squared are clearly better than the previous modelling technique. This means that not only has the goodness of fit improved, but so has the prediction quality.

4.2.2 Dropping features

We will try to drop features as we did for Regressor2, the best multiple linear regression model, to see if it decreases the RMSE. The feature that we will drop is sex, which was found to have very little correlation with the dependent variable.

RMSE	Rsquared	MAE
5169.4789304	0.8123495	3002.3360458

We see that this decreases the RMSE by approximately $27. Also the goodness of fit improves by a little bit. Dropping this feature would hence be a good option. The feature, region, was dropped too, after this, but the RMSE worsened (screenshot shown below) and I decided that the best model would be the one that has all independent features except for sex.

RMSE	Rsquared	MAE
5228.3666330	0.8086692	2978.5011164

4.2.3 Tuning hyperparameters

We will now try to optimize the hyperparameters cost, gamma, epsilon and sigma to see if we can improve the RMSE for the model we have chosen. Also, this time around, we perform a 5-fold cross validation, instead of a train/test split. The train/test split method was helpful in identifying the features that should be in the model.

```
     RMSE  rsquared    C gamma sigma epsilon
4746.229 0.8146375 10.0   0.1   0.1    0.10
```

Tuning hyperparameters did work and the RMSE decreased by about $423 or 8.18%, than what it was when there were default hyperparameter values.

The results say that choosing the model with all independent variables except for sex, optimized hyperparamters (shown above) and a 5-fold cross validation would produce the best possible predictions for the charges, as far as support vector regression is concerned.

5 Conclusion

For our problem, we tried two regression techniques, a linear one, in multiple linear regression and a non-linear one, in support vector regression. We can conclude that the support vector regression outperformed the multiple linear regression model when it came to both, error metrics and goodness of fit. There was a difference of $1532.2 in the RMSE, which is huge, and 0.06 in the R-squared metric, which as I have described earlier, is how well the regression technique fits the data, between the best models of both techniques. This is due to the fact that the problem at hand is better described by a non-linear model, as some variables, that we inspected, such as BMI, did not show a linear relationship with respect to charges. Also, what further improved the performance of the support vector regression model was the fact the we were able to tune hyper-parameters. As far as the features that gave us the best models for both techniques, involved all except for sex. So, for people working on this problem, I would suggest dropping it as there is an improvement in the error metrics and goodness of fit, when we do so.