Capstone Project

Medical Insurance Charge Prediction Model
Using Supervised Regression Model

K A P Siriwardena
28-04-2022

1.0 Introduction

People are always confused about their medical insurance and don't know the cost of insurance at different ages and conditions like sex, bmi, region, no of children etc.

A Supervised Machine Learning Model was developed in this project to make predictions of the insurance cost they will have to pay.

2.0 Data Set

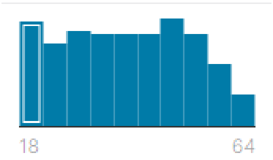
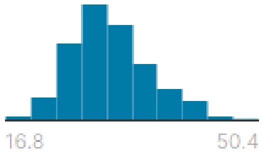
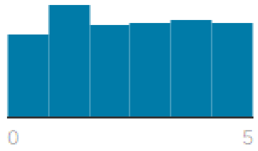
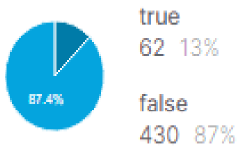
Focused data set is consisting of 3630 data points with 6 different attributes. Below table describes the columns of the data set.

Column Name	Description
age	Age of user
sex	Gender of user, Male/Female
bmi	Body Mass Index of user
smoker	If the user is smoker or not
region	Region where user lives
children	Number of children user have
charges	Actual Charge, user has to pay

3.0 Exploratory Data analysis

3.1 Data distribution

Below table summarizes the focused group for this experiment. It is clear that 58% of the group is male and 87% are nonsmokers. Age lies between 18yrs to 64 yrs.

age	bmi	children
		
sex	**smoker**	**region**
Male – 58% Female – 42%		Northeast – 30% Southeast – 28% Other -43%

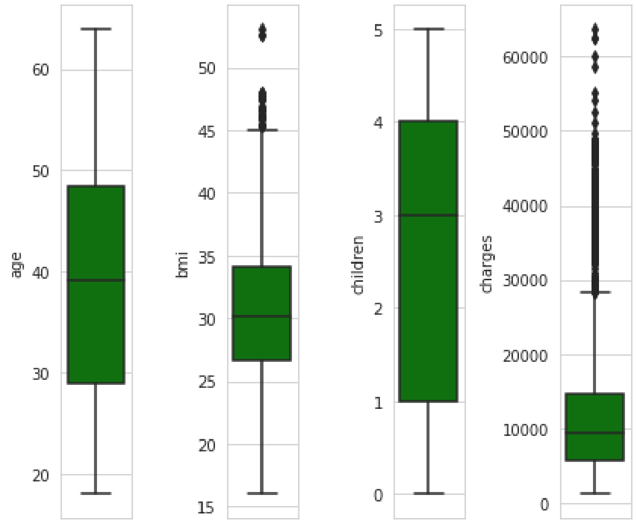
3.2 Missing data

Further analysis indicates that no missing values in the dataset

```
#   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   age       3630 non-null   float64
 1   sex       3630 non-null   object
 2   bmi       3630 non-null   float64
 3   smoker    3630 non-null   object
 4   region    3630 non-null   object
 5   children  3630 non-null   int64
 6   charges   3630 non-null   float64
```

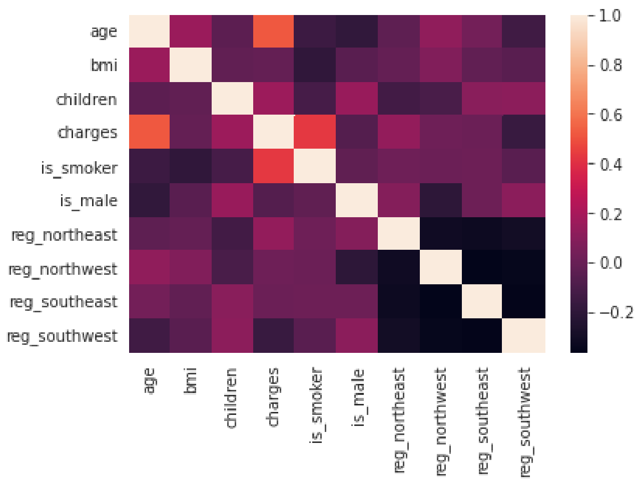
3.3 Outlier analysis

Outliers play vital role towards the accuracy (r2-score) of any machine learning algorithm. Therefore identifying those outliers and filtering out is a must before going for the model building process. Outliers recorded in BMI and CHARGES columns. Data set further narrowed down to 3006 points after filtering out those outliers.



3.4 Correlation analysis

Correlation analysis carried out to examine the correlation of different attributes with the final charges. Accordingly, region southeast is not considered in this model building process due to its low correlation to the output.



Feature	Correlation
reg_southwest	-0.163218
is_male	-0.067385
bmi	-0.011718
reg_southeast	0.011019
reg_northwest	0.019978
reg_northeast	0.139680
children	0.163299
is_smoker	0.432600
age	0.520700

4.0 Feature Engineering

4.1 One hot encoding

Four different "regions" identified in the data set. Encoding those categorical variables to numerical variables has been done with one hot encoding.

Feature encoded	Encoded column names
Region	reg_northeast
	reg_northwest
	reg_southeast
	reg_southwest

4.2 Binary encoding

Sex and Smoker columns converted in to binary for the model building process.

Feature encoded	Encoded column names
Sex	is_male
Smoker	is_smoker

5.0 Methodology (Solution approach , tools)

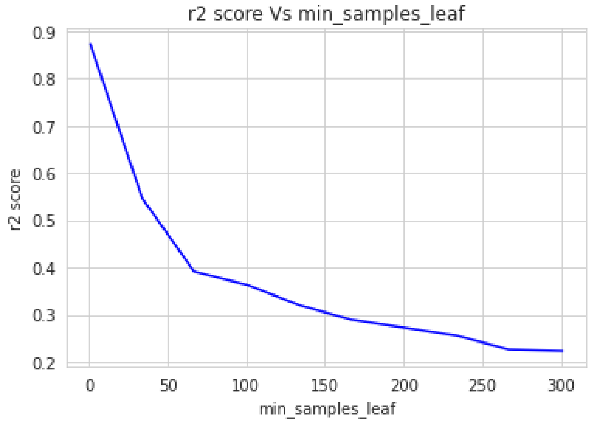
Four regression models considered in this approach. Those are,

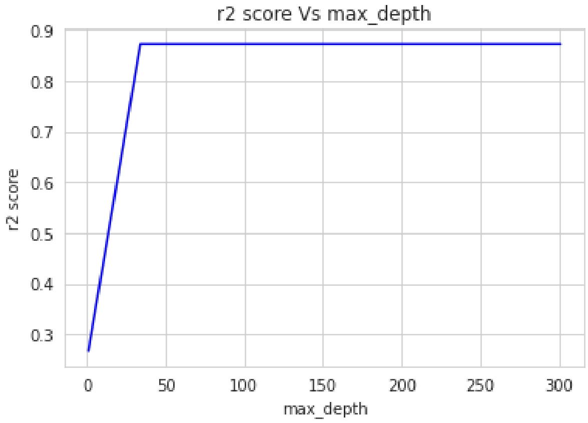
1. Linear Regression
2. XGBoost Regressor
3. RandomForest Regressor
4. Decision tree Regressor

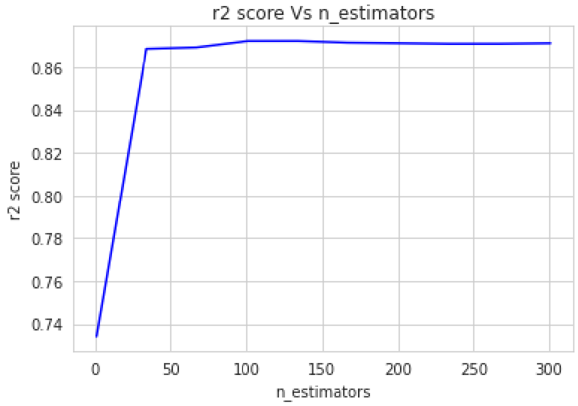
Among those, RandomForest Regressor scored the best results for the r2 score and mean squared error. Comparison of the different models and the final evaluation is available in the results section.

5.1 Model Training

Hyper parameter tuning for the RandomForest Regressor is done to get the best hyper paraments for the model training process. Below graphs show the behavior of those hyper parameters with the r2_score.





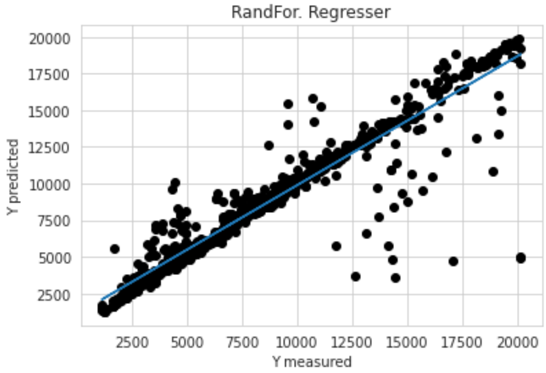


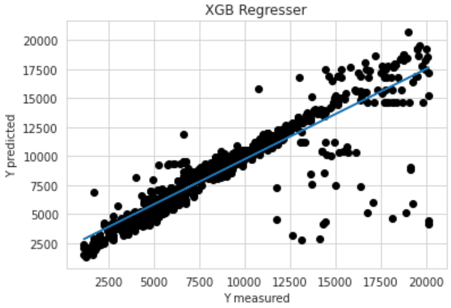
6.0 Results

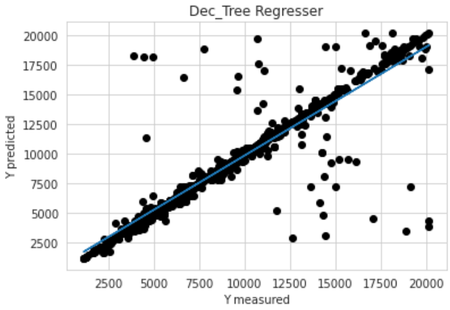
Below table summarizes the results of the four models considered.

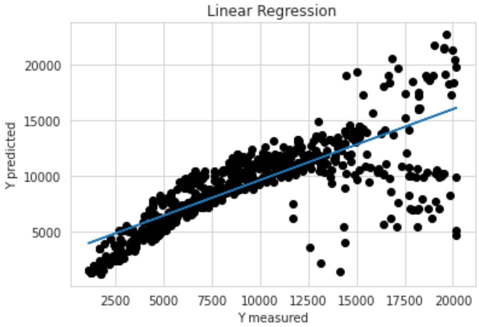
Model Name	Mean Square Error	Root Mean Square Error	R2 Score
Linear Regression	7,124,544	2669.18	0.627561
XGB Regressor	3,932,259	1982.99	0.794439
Random Forest R.	2,284,328	1511.39	0.880586
Decision Tree R.	3,429,477	1851.88	0.820723

6.1 Cross validation









7.0 Conclusion

With the above results it can be clearly seen that the RandomForest Regressor has the best r2_score and the minimum mean squared error.

8.0 Discussion

Goal of this experiment was to build a model that can be used to effectively predict the charges of the medical insurance of the people at their different ages and conditions.

Among the tested models, RandomForest Regressor performed well with an r2_score of 0.88 and mean square error of 2,284,328.

Having more data would have been taken this experiment towards more good results ensuring a better practical usage.