# Investigation of Medical Cost variation through machine learning techniques

**Sijia Li**
Boston University
Boston, MA
scartt@bu.edu

**Zhenglei Jiang**
Boston University
Boston, MA
zljiang@bu.edu

**Zhangde Song**
Boston University
Boston, MA
sozhangd@bu.edu

**Qingtian Huang**
Boston University
Boston, MA
qthuang@bu.edu

## Abstract

Medical cost Varies greatly from person to person and the underlying reason is often obscure. This research intends to gain insight of factors that contributes to the variation of medical cost by modeling patient's medical data with machine learning techniques. To make better sense of the data, multiple regressions and clustering techniques were employed with each modeling either some or all patient's features with the medical cost data. Multiple models indicate a direct positive correlation between smoking and higher medical cost in addition to other findings.

## 1 Introduction

Various factors account for the variation of medical cost in the population and in this project aims to make sense of the factors that contributes to this variation. A data set containing patient's age, sex, bmi, number of subsidiaries and their smoking habit was chosen and regression techniques

## 2 Data processing

### 2.1 Machine learning with scikit-learn:

Scikit-learn (Sklearn) is a widely-used library tfor machine learning in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib. In this experiment we used it for data visualization and models.

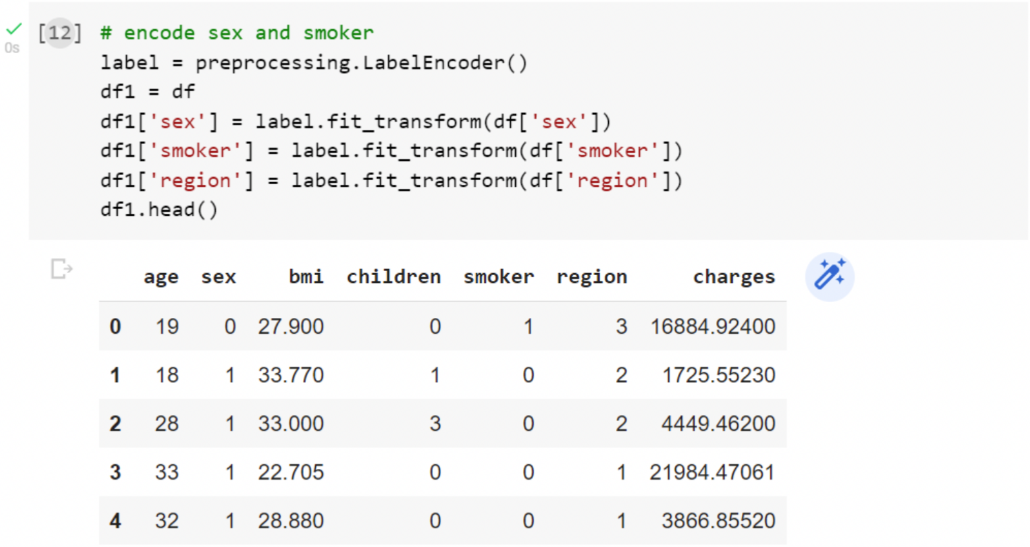
### 2.2 Data Cleaning and Generating Training/Testing Sets:

Before doing experiment, we cleaned the data so that they can be processed and to avoid process failures. In this step, all formatted data were manually inspected and all null values were removed. The processed data is then split into two parts with 80 percent being used as training data and the remaining 20 percent as testing data.

### 2.3 Data Encoding:

To transform data to a machine recognizable set, all data first passed through the encoding step. Data such as ”Male” and ”Female” are not numerical and therefore cannot be recognized by the machine learning models, and therefore we applied encoding to these values. Numerical data such as age or medical cost were kept in numerical form whereas smoking status, sex and region were encoded with respective numerical values.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).



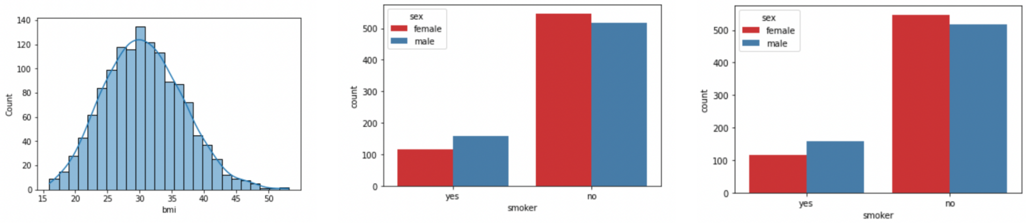
```
[12] # encode sex and smoker
     label = preprocessing.LabelEncoder()
     df1 = df
     df1['sex'] = label.fit_transform(df['sex'])
     df1['smoker'] = label.fit_transform(df['smoker'])
     df1['region'] = label.fit_transform(df['region'])
     df1.head()
```

```
   age  sex     bmi  children  smoker  region      charges
0   19    0  27.900         0       1       3  16884.92400
1   18    1  33.770         1       0       2   1725.55230
2   28    1  33.000         3       0       2   4449.46200
3   33    1  22.705         0       0       1  21984.47061
4   32    1  28.880         0       0       1   3866.85520
```

Figure 1: Screenshot of Data Encoding Code

## 2.4 Data Visualization:

Before any experiments, the cleaned data set was first vitalized to find some pattern that could be further exploited within later experiments. In this step, a histogram graph of age count, bmi count, charge count, were individually plotted, followed along with the male female smoker ratio, regional gender comparison, and regional smoker comparison histograms.



(a) BMI distribution histogram

(b) Smoker Count Histogram

(c) Region Count Histogram

Figure 2: Data Visualization Histograms

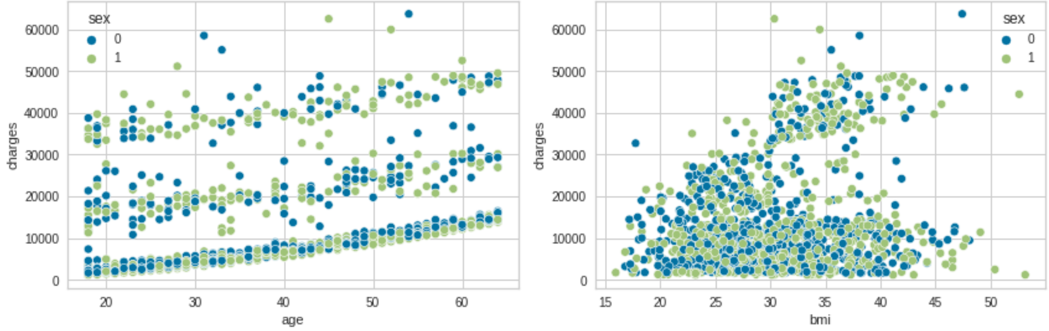
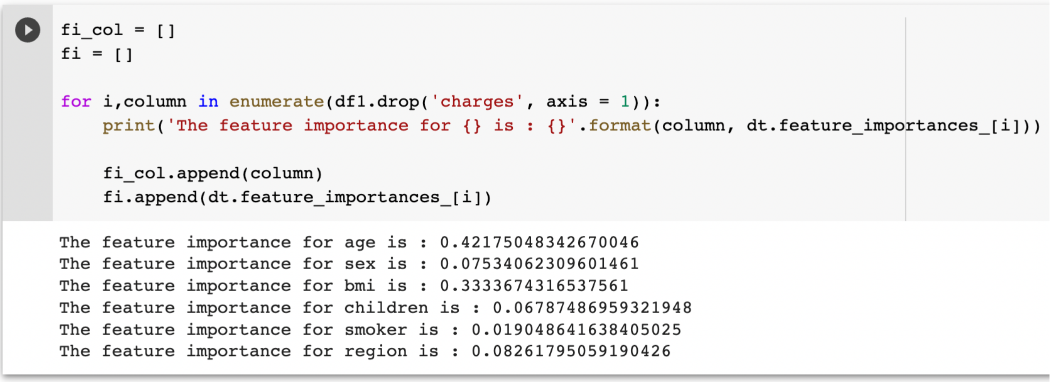


Figure 3: Categorical relations Age VS. BMI VS Smoker VS Charge

In this step, it is clear from the basic histogram that among the data set: BMI value of the sample follows a Gaussian distribution, the sex distribution is relatively even and the region feature is very even amongst the data. There seems to be some distinct pattern regarding smoker feature that could be exploit later.

### 2.5 Importance:

Data set undertook an importance evaluation before the experiment. This step calculates the importance of each feature of the patient for this data based on its a decision tree method which rank feature correlation to the medical cost and allows us to know which features could be potentially more valuable to exploit.



```
fi_col = []
fi = []

for i,column in enumerate(df1.drop('charges', axis = 1)):
    print('The feature importance for {} is : {}'.format(column, dt.feature_importances_[i]))

    fi_col.append(column)
    fi.append(dt.feature_importances_[i])
```

```
The feature importance for age is : 0.42175048342670046
The feature importance for sex is : 0.07534062309601461
The feature importance for bmi is : 0.3333674316537561
The feature importance for children is : 0.06787486959321948
The feature importance for smoker is : 0.019048641638405025
The feature importance for region is : 0.08261795059190426
```

Figure 4: Screenshot of Decision Tree Importance Ranking

With the importance data, age, BMI, and smoking status appear to be in high correlation with medical cost so these few features are focused in the later experiments.

## 3 Experiment

The ultimate goal of this experiment is to gain insight into the difference of medical cost from person to person, so we set out to approach this problem gradually with multiple machine learning models. This approach requires aggrandized work while it should help us minimize under-fitting, over fitting problems and other general entry level machine learning problems that could occur due to the limitation of machine learning algorithms. In this specific experiment, linear regression, polynomial regression, random forest regression and logistic regression were each employed to train certain feature-medical cost models with a clustering technique which helps to provide further insight.

### 3.1 General Evaluation:

Each regression model in this experiment employs different algorithms, but follows a similar evaluation scheme: **MSE**, **MAE** and $\mathbf{R^2}$

**MSE**: Mean Squared Error is a statistical estimator that measures the average squared difference of the predicted data from actual value. This measurement directly reflects the fidelity of train model to raw data, and the equation is shown as follow:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

**MAE**: Mean Absolute Error measures the error between paired trained model prediction and actual test data. With yi being prediction and xi being the true value, MAE also reflects the accuracy of the trained model.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$\mathbf{R^2}$ coefficient represents the percent of variation and in our models this value ranges from 0-1 indicating the performance of the regression model. The norm for most machine learning models is to have a $\mathbf{R^2}$ value of smaller than 0.7 to be considered a usable model and the greater the value usually indicates better performance.

$$R^2 = 1 - \frac{u}{v}$$

### 3.2 Simple Linear Regression: Age VS. BMI VS. Charge:

From data visualization and importance indicator, age, bmi and smoker are highly correlated with output variable charges, sex and region do not have significant impact on output variable charges.

We chose Age and BMI as the first one-variable simple linear regression. The resulting model made a $\mathbf{R^2}$ value of 0.04 and 0.11, both significantly lower than the 0.7 threshold signifying a completely inaccurate model that significantly under-fit our data. This is expected as linear regression is limited to 1 variable and is not great to account for data that don't form a good linear pattern.

$$y = ax + b$$

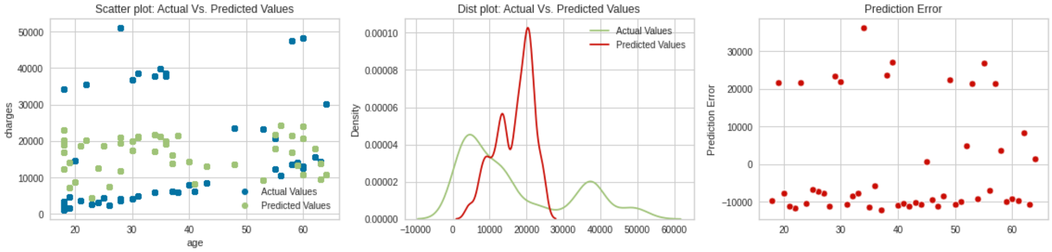


Figure 5: Simple Linear Regression: Age

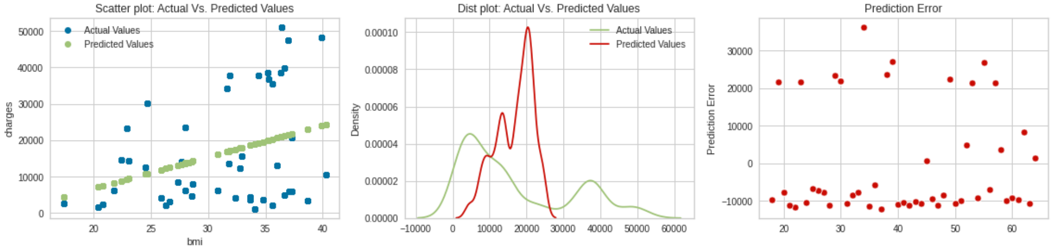


Figure 6: Simple Linear Regression: BMI

### 3.3 Multiple Linear Regression: Age VS. BMI VS. Smoker VS. Charge:

In this section we can see two or more variables in relation with charge. With age and bmi 0.15 value is better than both previous tests yet still significantly lower than the accepted 0.7 value. This low value reveals that the medical cost is not strongly linearly related to either age, BMI or age and BMI combined value.

$$y = a_1x_1 + a_2x_2 + a_3x_3b$$

As visualization shows, ”smoker” stands out in raw data scatter pilots to form some general correlation to higher cost. Hence, an additional multiple regression containing “Age”, “BMI” and “Smoker” value was trained. $\mathbf{R^2}$ value is 0.76 which is better than any previous regression and falls into an acceptable range ¡ 0.7. This indicates a relatively strong linear relationship between the three combined factors.

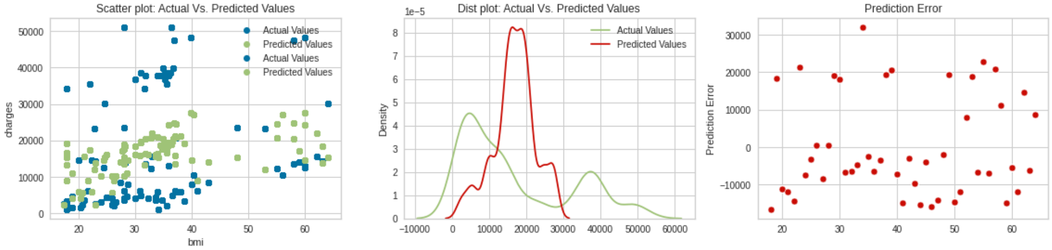


Figure 7: Multiple Linear Regression: AGE, BMI

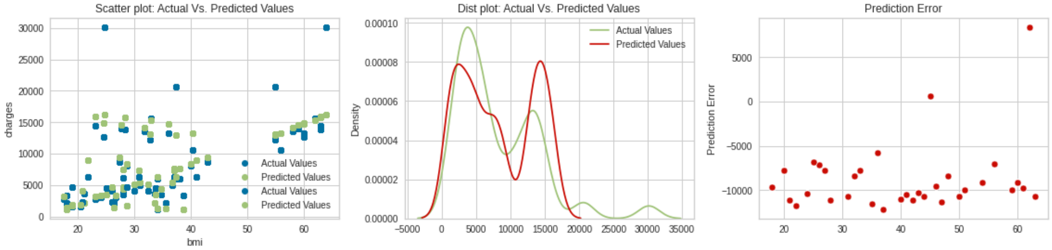


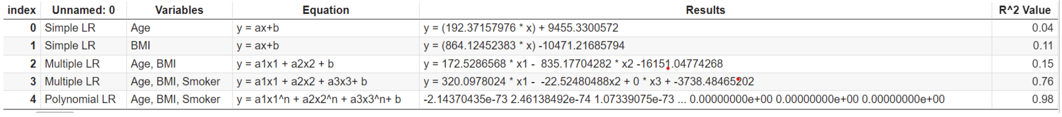
Figure 8: Multiple Linear Regression: AGE, BMI, Smoker

### 3.4 Polynomial Regression: Age VS. BMI VS. Smoker VS. Charge:

To further verify this correlation, a polynomial linear regression was trained using these data. Polynomial regression is a special case of linear regression with multiple variables. Though polynomial linear regression could generate over-fitting issues, it should provide a more accurate plot and, along with above models, providing a understanding of the data correlation. The equation and result is as shown below:

$$y = a_1 x_1^i + a_2 x_2^j + a_2 x_3^k + b$$

The $\mathbf{R^2}$ result of 0.98 is significantly better than simple and muptiple linear regression. Due to better fitting pattern of the polynomial regression model. Though this model could suffer from the over-fitting problem, along with the previous models, a clear correlation between “smoker” status with medical cost is observed, and this feature will be further investigated later.



| index | Unnamed: 0 | Variables | Equation | Results | R^2 Value |
|---|---|---|---|---|---|
| 0 | Simple LR | Age | y = ax+b | y = (192.37157976 * x) + 9455.3300572 | 0.04 |
| 1 | Simple LR | BMI | y = ax+b | y = (864.12452383 * x) -10471.21685794 | 0.11 |
| 2 | Multiple LR | Age, BMI | y = a1x1 + a2x2 + b | y = 172.5286568 * x1 - 835.17704282 * x2 -16151.04774268 | 0.15 |
| 3 | Multiple LR | Age, BMI, Smoker | y = a1x1 + a2x2 + a3x3+ b | y = 320.0978024 * x1 - -22.52480488x2 + 0 * x3 + -3738.48465202 | 0.76 |
| 4 | Polynomial LR | Age, BMI, Smoker | y = a1x1^n + a2x2^n + a3x3^n+ b | -2.14370435e-73 2.46138492e-74 1.07339075e-73 ... 0.00000000e+00 0.00000000e+00 0.00000000e+00 | 0.98 |

Figure 9: Linear regression summary

### 3.5 Random Forest Regression: Hyperparameter Tuning:

Considering the over fitting problem of polynomial regression, decision tree models are introduced to mitigate the limitation of above linear regression techniques. In this experiment, a random forest model, one algorithm that combines ensemble learning methods and tree framework to randomly draw decision trees and take average of result output data as leaf nodes to produce stronger predictions, was introduced. With random forest model’s decision tree patterns and its more robust

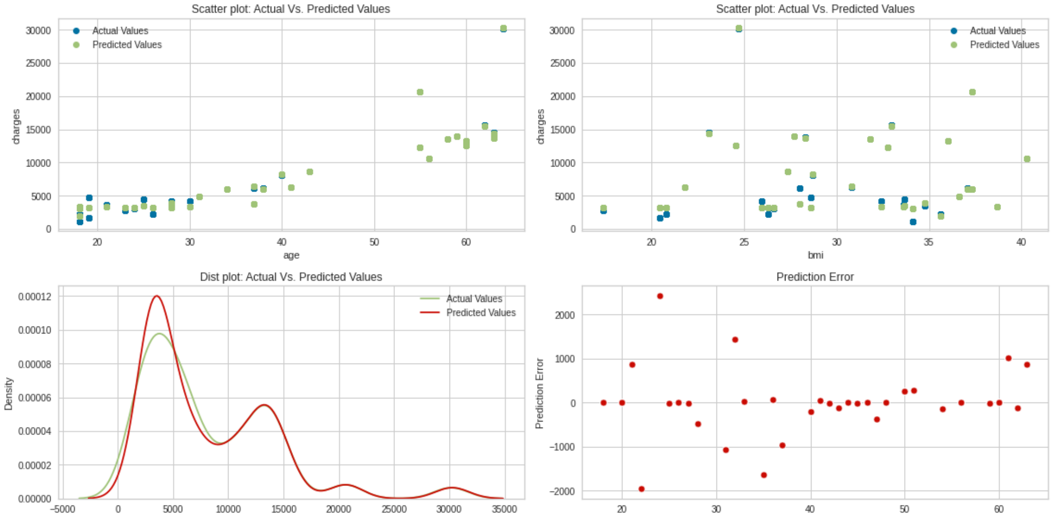


Figure 10: Polynomial Linear Regression Equations and Bias

prediction features, it is expected to produce a different model and might have a better performance in predicting with regard to the over fitting problem. Therefore, here we will be using RandomForestRegressor from sklearn to do the experiment.

To train a better model, we tried to fine-tune the hyperparameters and tested the performance of each. By experimenting and searching, we decided the hyperparameters needed further modification are: n estimators (number of esimators used), max leaf nodes (maximum number of leaf nodes in the decision tree), min samples split (minimum number of samples are splited into), and max features (maximum number of features used).

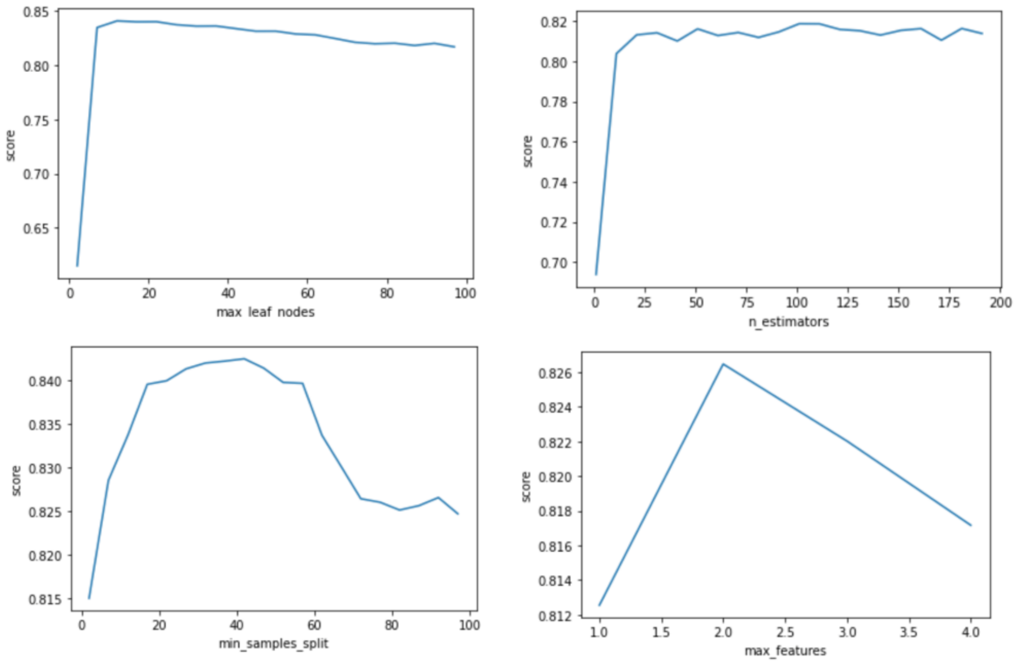


Figure 11: RFRegressor Hyperparameter Tuning

This figure shows the R2 Score change with respect to hyperparameter change within a chosen range. Here we decide to train the model using the following parameter:

n estimators = 100 (default value)

max leaf nodes = 10

min samples split = 35

max features = 2.0

### 3.6 Random Forest Regression: Result Visualization:

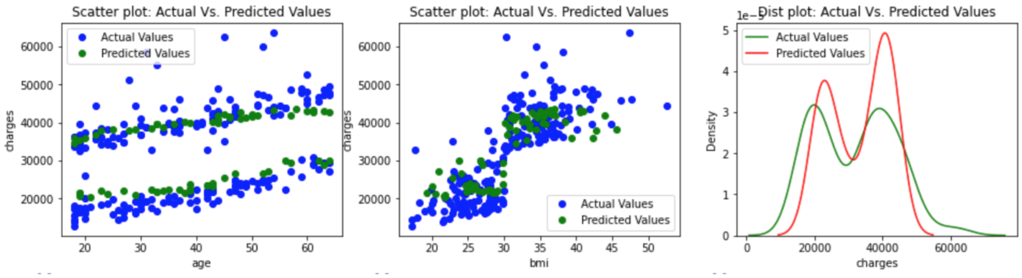


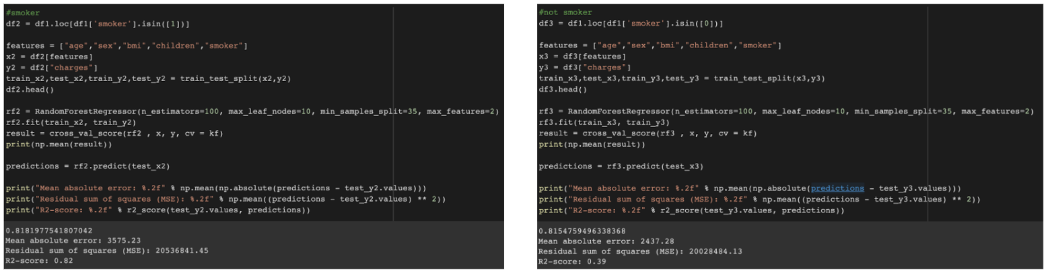
Figure 12: Random Forest Regression: Age  BMI  Prediction

By observing the prediction results of the model with parameters mentioned in the previous section, we found that the prediction with x-axis of age is generally linear, while with respect to BMI is clustering. The prediction in charges shows that there is certain error with the model in prediction performance, but generally can do a moderately accurate prediction. Evaluating the dot plots of predictions, the outliers of the data are not cleaned, which may influence the result.

### 3.7 Random Forest Regression: Models for Smoker  Non-Smoker:

From what we observed from the data visualization, there is major difference between the distributions of smokers and non-smokers. Therefore, although we believe that the Random Forest Regressor can be a potential model to do prediction for our data, we think we may train different models for smokers and non-smokers, so that we can get a more accurate result in prediction. Therefore, we split the data set into smokers and non-smokers and train the model again.

In the models trained for smokers and non-smokers respectively, the R2 Scores have a major difference with that of the model for the data set as a whole. However, this does not mean that this model would fail in achieving a good prediction performance with smokers since it may suffer from underfitting or overfitting problem. Therefore, we did 5-fold cross validation to both models and experiment again, and this gave us a stable R2 Score around 0.82.



```
#smoker
df2 = df1.loc[df1['smoker'].isin([1])]

features = ["age","sex","bmi","children","smoker"]
x2 = df2[features]
y2 = df2["charges"]
train_x2,test_x2,train_y2,test_y2 = train_test_split(x2,y2)
df2.head()

rf2 = RandomForestRegressor(n_estimators=100, max_leaf_nodes=10, min_samples_split=35, max_features=2)
rf2.fit(train_x2, train_y2)
result = cross_val_score(rf2 , x, y, cv = kf)
print(np.mean(result))

predictions = rf2.predict(test_x2)

print("Mean absolute error: %.2f" % np.mean(np.absolute(predictions - test_y2.values)))
print("Residual sum of squares (MSE): %.2f" % np.mean((predictions - test_y2.values) ** 2))
print("R2-score: %.2f" % r2_score(test_y2.values, predictions))
```

```
0.8181977541807042
Mean absolute error: 3575.23
Residual sum of squares (MSE): 20536841.45
R2-score: 0.82
```

```
#not smoker
df3 = df1.loc[df1['smoker'].isin([0])]

features = ["age","sex","bmi","children","smoker"]
x3 = df3[features]
y3 = df3["charges"]
train_x3,test_x3,train_y3,test_y3 = train_test_split(x3,y3)
df3.head()

rf3 = RandomForestRegressor(n_estimators=100, max_leaf_nodes=10, min_samples_split=35, max_features=2)
rf3.fit(train_x3, train_y3)
result = cross_val_score(rf3 , x, y, cv = kf)
print(np.mean(result))

predictions = rf3.predict(test_x3)

print("Mean absolute error: %.2f" % np.mean(np.absolute(predictions - test_y3.values)))
print("Residual sum of squares (MSE): %.2f" % np.mean((predictions - test_y3.values) ** 2))
print("R2-score: %.2f" % r2_score(test_y3.values, predictions))
```

```
0.8154759496338368
Mean absolute error: 2437.28
Residual sum of squares (MSE): 20028484.13
R2-score: 0.39
```

Figure 13: R2 Score and 5 Fold Cross Validation

For random forest regression, although the R2 Score it can reach is not as high as polynomial regression can, but it is less biased and suffers less from overfitting.

### 3.8 Cluster K-Means, Gaussian Mixture:

Clustering technique is an unsupervised machine learning method which categorize data into groups while avoiding the inaccuracy from human assumptions.

Before model the clustering, we tried normalized the feature ranging from 0 to 1. Because especially for K-means Modeling, not normalizing data may leads to inaccurate modeling.But the result turns out to be very messy.

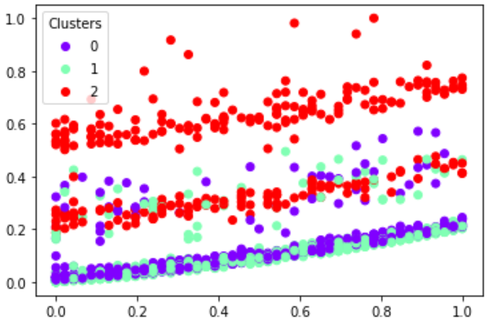


Figure 14: Normalized Data Modelling with KNN(3)

The reason is listed as follow. 1. The predominate feature is ”charges”, which originally ranging from 10000 to 60000 and takes 10 time more importance than any other feature. 2. There are also 2 binary features: ”sex” and ”smoker”, which take value 0 or 1. These 2 feature have little contribution on distinguishing smoker or non-smoker.

I believe there are better parameter to normalize the data. However, I didn't normalized the data here. Luckily, the most importance feature also takes the larger scale. I tried both K-Means model and Gaussian Mixture Model. There is the graph showing the result.

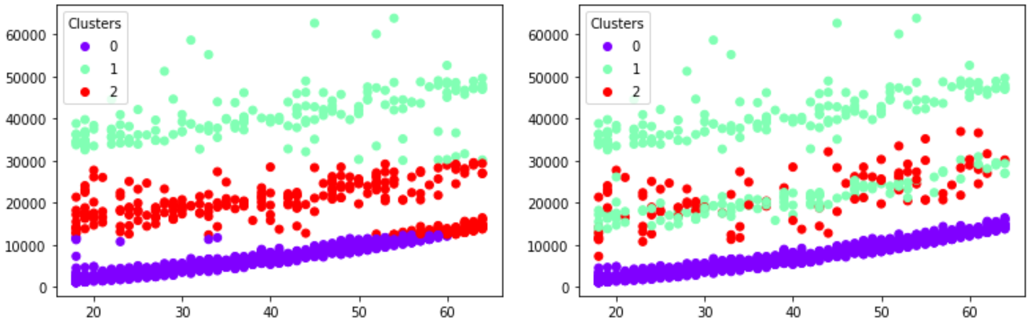


Figure 15: KNN Model and GMM Model with 3 Cluster

The three centers of KNN model: [0: [36, 6296] 1: [40, 40761] 2: [45, 18505]] The three centers of GMM Model: [0: [36, 6296] 1: [40, 40761] 2: [45, 18505]] Surprisingly, as I compare the above GMM model with the Charges VS Age graph with Smoker label on. I found great similarity. I decide to make a classifier based on GMM model. Therefore, I tried GMM Model with the same input, only changing the cluster = 2. Based on the graph, I works well. And this motivate me to make the logistic classifier in the below.

### 3.9 Smoker Logistic Regression:

As established in section 3.2, a direct positive correlation between a patient's "smoke" status and his or her medical cost was observed in the linear regression models. To further exploit this trait, our team decided to reverse the process and try to identify whether a patient is a smoker based on his or her medical cost.

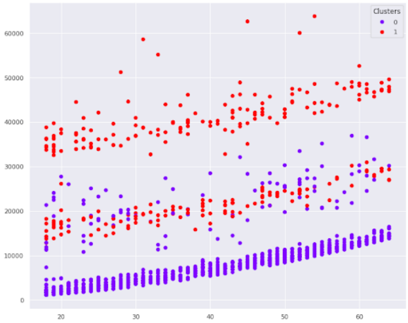


Figure 16: Charges VS Age graph with Smoker label

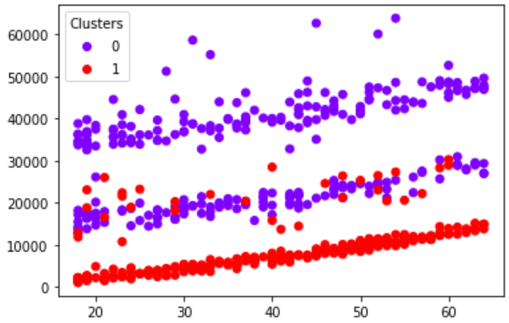


Figure 17: GMM with 2 Clusters

Logistic Regression, which traditionally is being used to estimate the probability of an event occurring, is the ideal chosen model. The basic model regresses the data to a logistic distribution given the equation:

One problem of this experiment is the data is unbalanced with the number of smoker and non-smoker. Therefore, before the experiment, we balanced the data frame by taking all smokers' data point and sample the same number of non-smoker data point to avoid unbalance.

In this experiment, patient's medical cost was directly trained by this model to predict weather a person is smoker and the trained graph is shown:

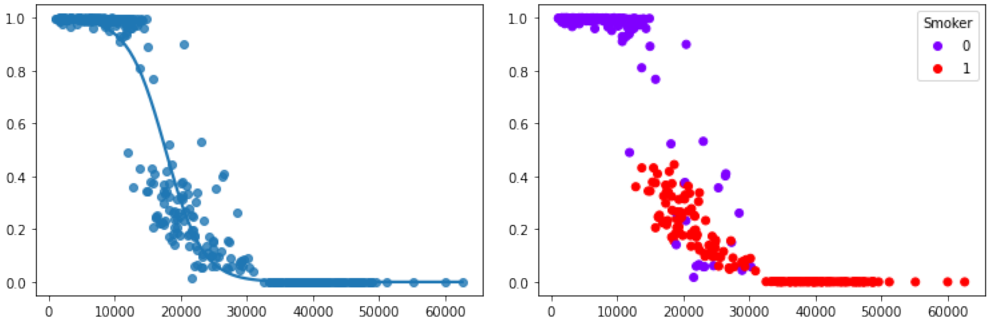


Figure 18: Charge

The parameter is trained to be

[-0.0655244 , 0.0863435 , -0.19792367, -0.08792067, 0.36826806, 0.00040367].

This achieves accuracy score($R^2$) with 0.95 with both training set and test set.

Since charge is the most influential feature overall, the graph above shown the accuracy of the sigmoid prediction of the logistic regression. On the right side, the graph is labeled with smoker and non-smoker by different colors.On the left, there is the sigmoid line ploted on the graph with parameter of 0.00040367.

## 4 Conclusion

Overall seven machine learning models were trained in this experiment to correlate patient medical cost to their feature sets in this experiment with two of the model resulting in $\mathbf{R^2}$ value of below 0.7 signifying numerous weak correlation between our tested feature sets and the variance of medical cost, revealing why medical cost raw data could be quite obscure to naked eyes.

Amongst the five acceptable regression models, all model es indicates a strong positive correlation between smoking feature and higher medical cost. This discover was further confirmed by a logistic regression model resulting in over 95 percent accuracy on predicting weather a patient is a smoker given his or her medical data.

In terms of predicting medial cost, the linear and random forest regression managed to obtain a $\mathbf{R^2}$ value of 0.98 and 0.8122 respectively indicating a relative accurate prediction model. Though the polynomial regression could sufferer from over fitting problem, the two models should provide a very good machine learning algorithmic estimation of medical cost for this experiment.

## 5 Limitation

Algorithm wise, each regression models are not tailor designed to fit the best need of this data set, hence all above conclusions inevitably suffer from certain degree of over and under fitting problem. To further improve the reliability of the conclusion, more robust machine learning algorithm to better fit the data set could be devised.

Data wise, for a grand inquiry about personal medical cost, the data-set was a relatively small containing only 1300+ data entries. Hence, to train a more robust model, a larger data set could be helpful to a more accurate training result.

Finally, this experiment was undertaken as a final project for introductory to machine learning course and all team members were first time learner of this topic. Though all work are to the best of the best knowledge of the researchers, there could be unavoidable incorrect assumptions, inaccurate modeling, imperfect training methodologies either due to unfamiliarity to the topic or programming packages used.

Weisberg [2005] **?**

## References

S. Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.