Springboard – DSC

Capstone Project II

Predicting individual annual cost of medical insurance by ML method

Final Report

Author: Wong Lok Hang Ronald
Mentor: Wayne Ang

Content

1. Introduction

Medical cost is one of the largest burdens of Americans, statistics from the Centers for Medicare and Medicaid Services (CMS) indicate that health care spending per person surpassed $10,000 in 2016 and then march steadily higher to $14,944 in 2023. In view of the rising medical cost, learning the factors affecting the medical cost becomes more and more important. For the consumers, they can understand the factors causing increase of health insurance. If the health insurance is no longer a black box to them, they can have better planning and management on their own medical insurance. For start-up companies, they can understand the behavior of their customers and set up their target customers easily.

1.1 Objective

The ultimate goal of this project is to adopt ML method to build up a medical cost prediction model to achieve:

1) Aiding start-up companies to investigate the behaviors of customers and learn the factors affecting the medical cost. Hence, they can set up a better marketing strategy, search for more potential customers and set up their target customers.

2) Helping consumers to understand their current healthcare plan, for instance, are they spending above or below the average compared with people who have similar background. (This can be one of the services provided by the start-up companies)

3) Helping consumers to have a better planning on their medical insurance such as planning no. of children, changing smoking behavior and achieving possible bmi. (This can be one of the services provided by the start-up companies)

2. Dataset

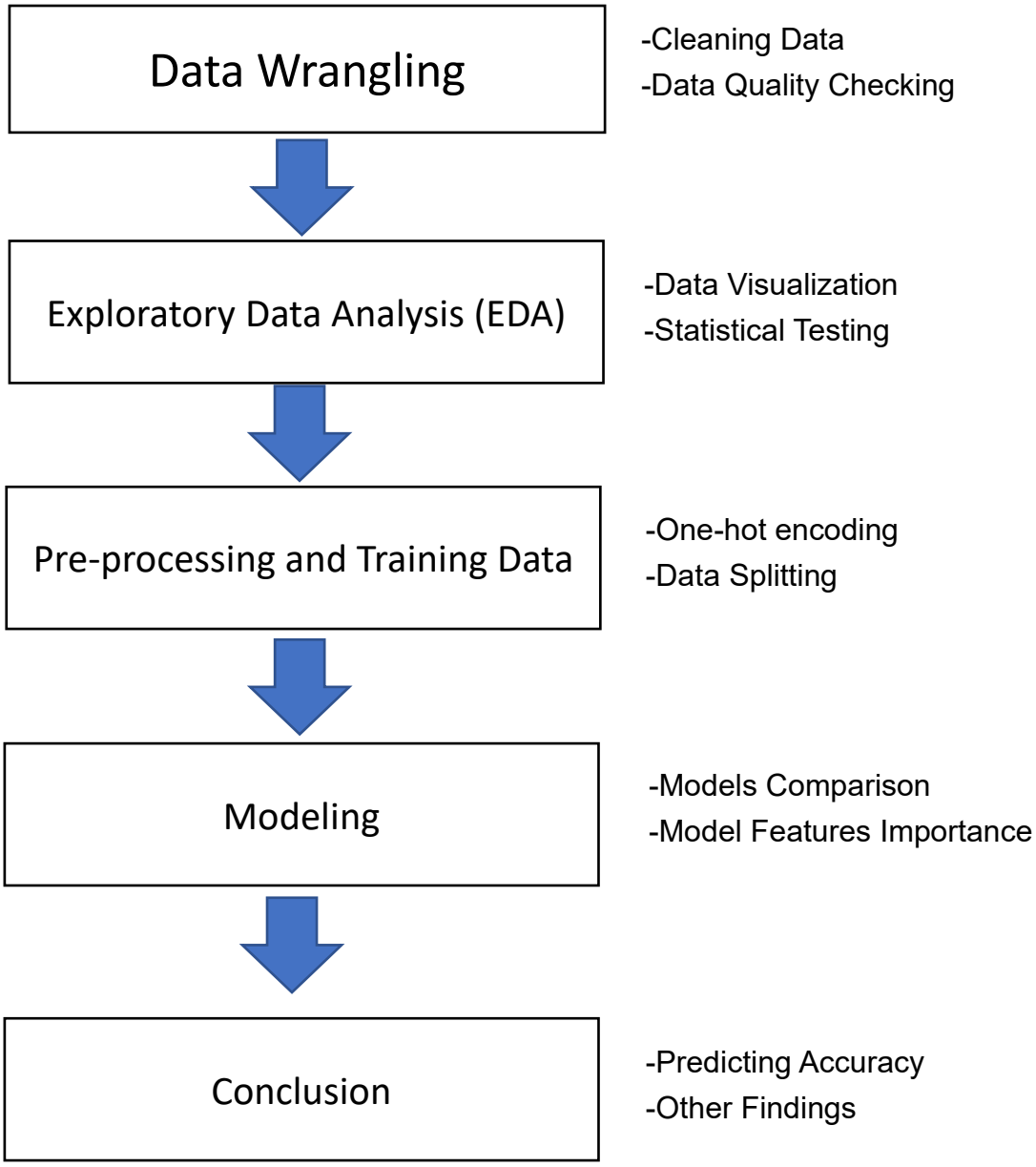
The dataset adopted in this project is sourced from website Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

The adopted dataset contains information of 1.3k persons with their personal and information and individual medical costs billed by health insurance. The dataset contains 7 columns and they are explained as follow:

- age: Age of primary beneficiary
- sex: Insurance contractor gender (female, male)
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking behavior (no : non-smoker, yes : smoker)
- region: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

3. Project Outline

This project followed the following workflow to establish a medical insurance cost prediction model and understanding the customers' behavior:



4. Data Wrangling

The main objective of data wrangling is cleaning the data such as dealing with missing value and screening unreasonable data. The initial condition of the data is as follow:

```
Data columns (total 7 columns):
 #   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
```

Figure 4.1 – Initial data condition

The data seems to be clean and does not contain any missing values. Then the next step is checking any repeating data. By checking, 2 identical data rows have been spotted. This is probably due to repeat enter of the same survey and the report will omit one of the two identical rows. The number of rows becomes 1337 instead of 1338.

In the next step, the report checks for any unreasonable values by simple data visualization:

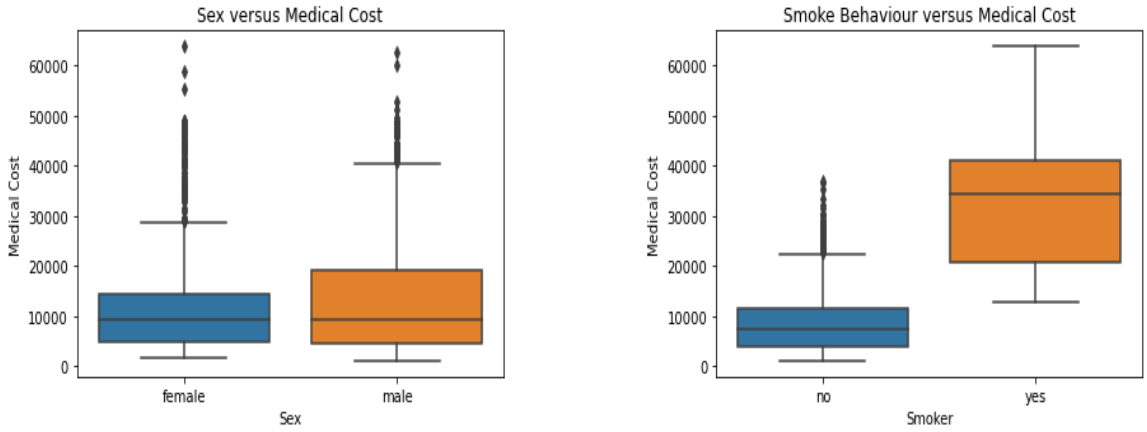


Figure 4.2 – Preliminary visualization for data wrangling (1)

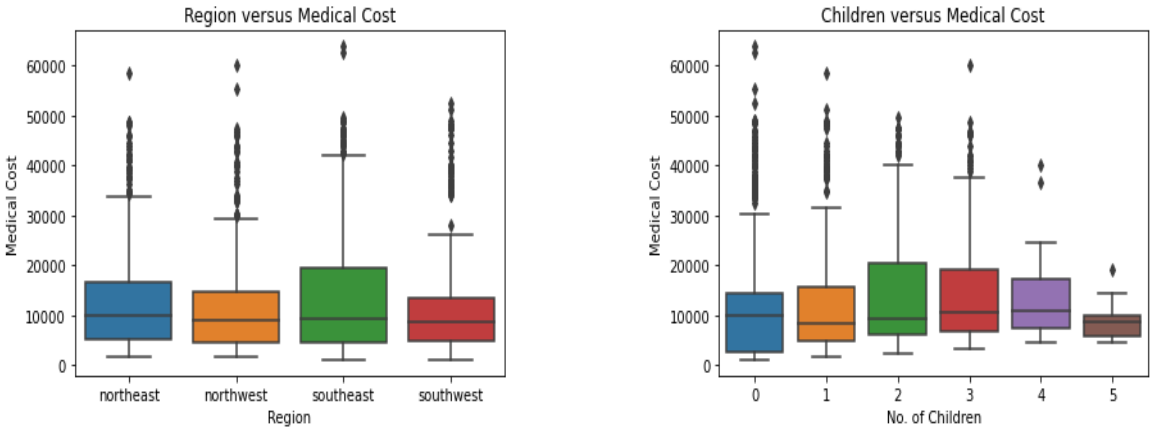


Figure 4.3 – Preliminary visualization for data wrangling (2)

From figure 4.2 & 4.3, there values of region, no. of children, smoker and sex are appropriate. For medical cost, some outliers have been noticed. As the outliers are still lie on the reasonable range, they will not be modified at this stage.

5. Exploratory Data Analysis (EDA)

The objective of EDA is to explore any underlying relationship and pattern within the data. This can provide more insight on selection of model and understand the inter-relationship between different features.

5.1 Heatmap

The correlation between all numeric features are summarized in a heatmap:

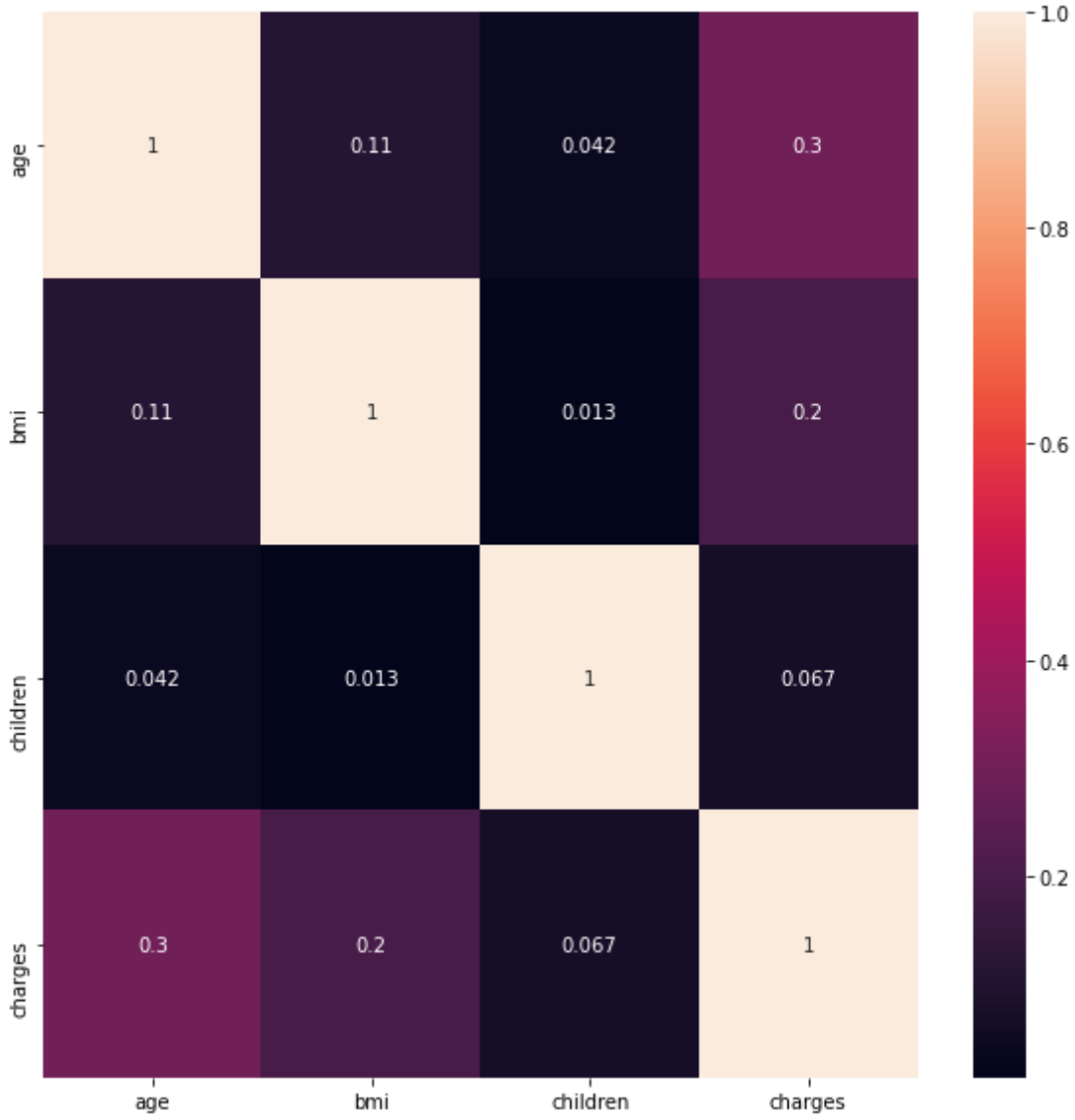


Figure 5.1 – Heatmap showing correlation of all numeric features

From the heatmap, no strong correlation between features can be found.

5.2 Pair Plot

The pair plot has been adopted to further visualize the relationship between different features as follow:

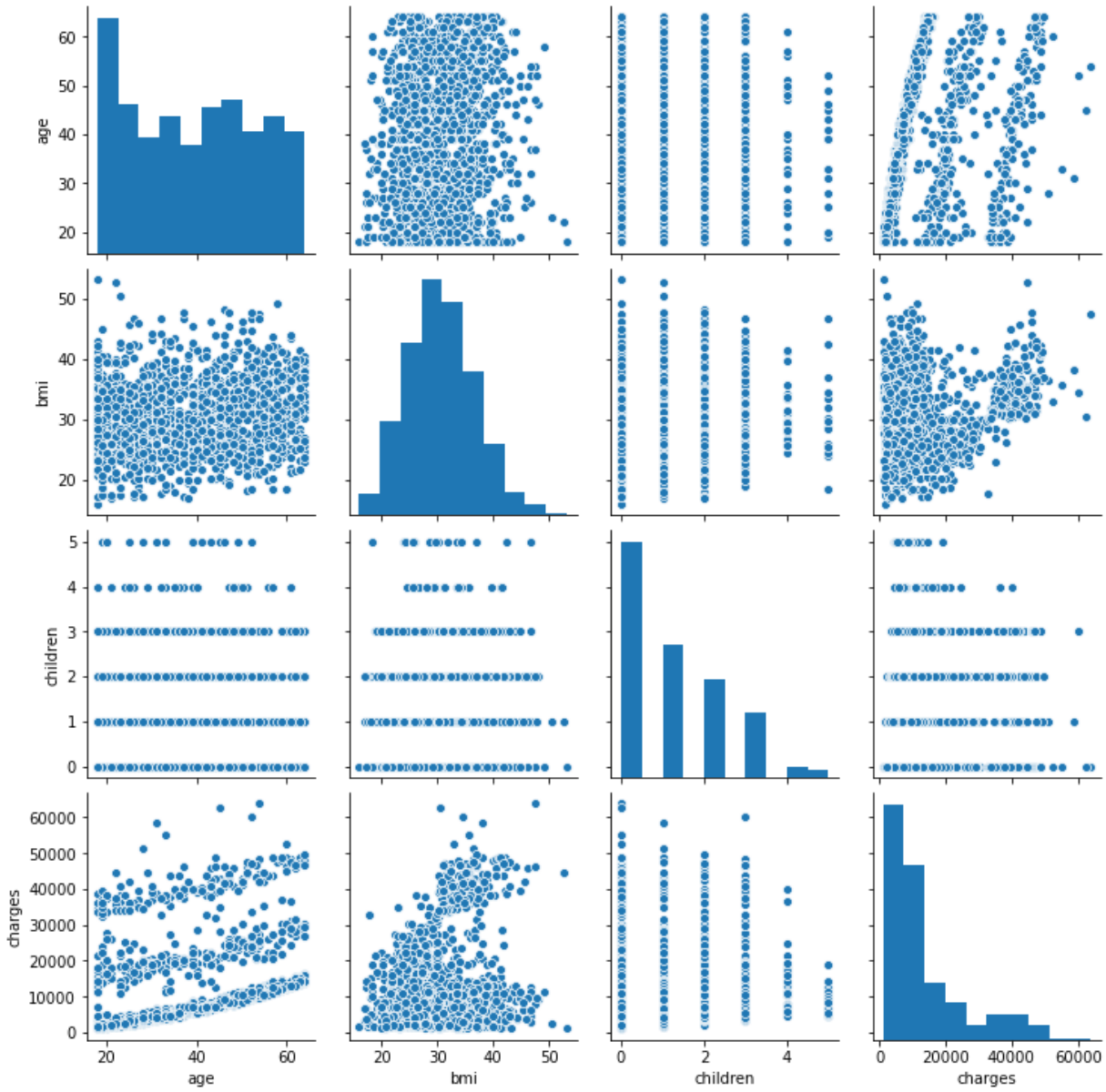


Figure 5.2 – Pair plot showing different features versu charges

From the pair plot, the following observations have been spoted:

1) Age: There are 3 discrete line we can draw from the scatter plot. This indicate clusters may exist

2) Sex: No obvious different in insurance fee between male and female

3) bmi: There is no obvious trend but we can observe that most people which have 30k+ insurance free have bmi 30 or higher.

4) children: The people which have 4 and 5 children tend to have a lower insurance fee compared with people who have 1-3 children

5) smoker: Smoker have higher insurance fee

6) region: There is no obvious different between people in different region

5.3 Plotting features against dependent variable

As one of the main objectives of the project is predicting the medical cost (dependent variable). Visualization including plotting features against medical cost has been carried to explore the relationship between features and the dependent variable.

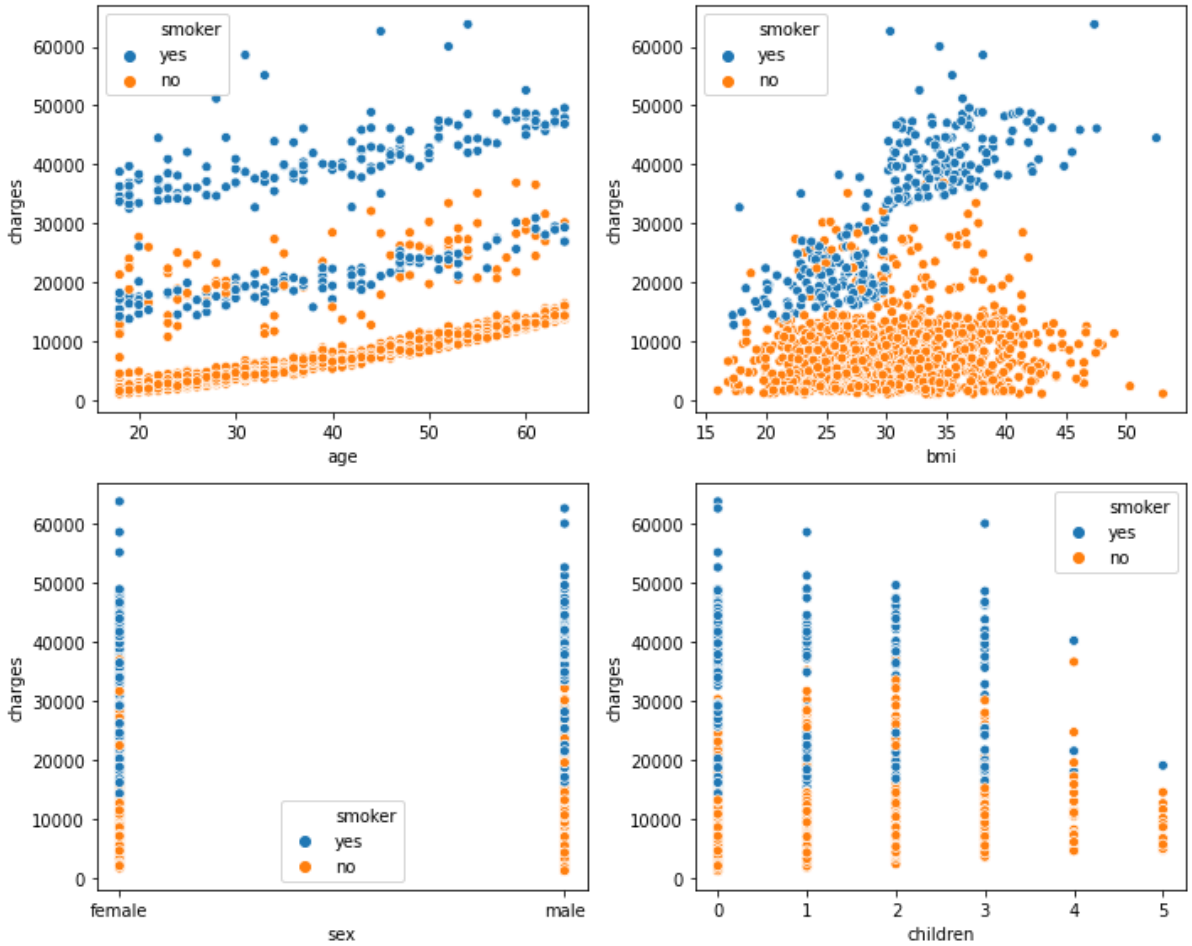


Figure 5.3 – Cluster of smoking behavior

When visualizing features versus charges, one caterical variable are held in order to locate any possible clusters. The only cluster can be found in this

study is smoking behavior. From Figure 5.3, despite of different situations (varying in gender, family size, bmi and age), the non-smokers tends to have a lower medical cost. This provides a preliminary thought that smoking behavior is highly correlated with the charges.

5.4 Statistical Test

In order to investigate the relationship between different features, 3 statistical test including t-distribution, one-way ANOVA and chi-square test have been adopted and serve in different purpose.

t-distribution

The t-distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis. The Student's t-distribution also arises in the Bayesian analysis of data from a normal family.

In this project, t-test is adopted to compare the significant of binary categorical features e.g. sex and smoking behavior. The null hypothesis is that the average medical cost for customer is the same regardless their gender and smoking behavior. The alternative hypothesis is that these means are different. The test to use here is the two-sample t-test. We are assuming the two groups have equal variance. The result is as follow:

Feature	t-statistic	P-value
Smoker	-46.64	1.4e-282
Sex	-2.12	0.03

For both gender and smoking behavior, the p-value < 0.05. Hence, for both cases we can reject the null hypothesis and accept the alternative hypothesis that these means are different (in 95% confident interval). In other word, gender and smoking behavior may cause different in medical cost.

One-way ANOVA

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of two or more samples (using the F distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way".

The ANOVA tests the null hypothesis, which states that samples in all groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. These estimates rely on various assumptions (see below). The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values.

In this project, One-way ANOVA is adopted to compare the significant of categorical features which contains more than two different levels e.g. region and family size. The null hypothesis is: there is no difference in the population means of the different levels of factor A (the only factor). The alternative hypothesis is the means are not the same.

Feature	f-statistic	P-value
region	2.92	0.03
children (Family size)	3.27	0.006

As the p-value of both region and family size are less than 0.05. Hence, for both cases we can reject the null hypothesis and accept the alternative hypothesis that these means are different (in 95% confident interval). In other word, region and family size may cause different in medical cost.

Chi-square test

Chi-square test is adopted to check for inter-feature correlation. The Chi-Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.

In this project, Chi-square test has been adopted to evaluate the relationship between all categorical independent variables and dependent variable. The result is as follow:

1) smoking behavior probably correlates with sex and age
2) Sex probably correlates with bmi and smoking behavior
3) family size probably correlates with age
4) Region probably correlates with bmi and large family size (5 children)

6. Preprocessing and Training Data Development

6.1 Preprocessing

The preprocessing of data consists of one-hot encoding and standardization of the data.

One-hot encoding

A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.

Categorical data can be either nominal or ordinal. Ordinal data has a ranked order for its values and can therefore be converted to numerical data through ordinal encoding. An example of ordinal data would be the ratings on a test ranging from A to F, which could be ranked using numbers from 6 to 1. Since there is no quantitative relationship between nominal variables' individual values, using ordinal encoding can potentially create a fictional ordinal relationship in the data. Therefore, one-hot encoding is often applied to nominal variables, in order to improve the performance of the algorithm.

For each unique value in the original categorical column, a new column is created in this method. These dummy variables are then filled up with zeros and ones (1 meaning TRUE, 0 meaning FALSE).

In the project dataset, the one hot encoding has been adopted to facilitate the modelling part in next stage as follow:

1) Sex: transformed from male and female to 1 and 0 respectively
2) Smoker: transformed from smoker and non-smoker to 1 and 0 respectively
3) Children: transformed from one variable to five variables which indicates Number of children covered by health insurance / Number of dependents (1: if no. of children of customer match with the variable, 0: if not match)
4) Region: transformed from one variable to four variables (1: if the region march with customer’s location, 0: if not match)

Standardization

In statistics, the standard score is the number of standard deviations by which the value of a raw score (i.e., an observed value or data point) is above or below the mean value of what is being observed or measured. Raw scores above the mean have positive standard scores, while those below the mean have negative standard scores.

It is calculated by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This process of converting a raw score into a standard score is called standardizing or normalizing (however, "normalizing" can refer to many types of ratios; see normalization for more).

In this project, standardization is used to transform the numerical independent variables (bmi, age) and the aim of standardization is to convert all numerical independent variables to similar scaler. This is a common technique adopted in the ML. Usually the models cannot distinguish the normal range and physical meaning of features, the models treats the numerical features equally despite the scaler or number of figure of them are in great different. If the features are standardized, they will become dimensionless and described as no. of standard derivation from mean. Hence, the models will not be affected by the units of features.

6.2 Training Data Development

The last step before establishing the model is splitting the data into train set and test set. The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets.

In this report, the dataset has been split to 70% of training data and 30% of testing data.

7. Modelling

As the targeting dependent variable is a numerical variable, ordinary least squares (OLS) and random forest regressor have been adopted and compared in this section.

7.1 Ordinary least squares (OLS)

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

Preliminary OLS Model

For the initial model, all features have been adopted and after the train data has been fit to the model, a summary of the model can be drawn as below:

Dep. Variable:	charges	**R-squared:**	0.729
Model:	OLS	**Adj. R-squared:**	0.726
Method:	Least Squares	**F-statistic:**	207.1
Date:	Sat, 13 Feb 2021	**Prob (F-statistic):**	7.66e-252
Time:	10:34:24	**Log-Likelihood:**	-9472.6
No. Observations:	935	**AIC:**	1.897e+04
Df Residuals:	922	**BIC:**	1.903e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6462.7392	305.143	21.179	0.000	5863.884	7061.594
age	3686.3005	202.075	18.242	0.000	3289.720	4082.881
bmi	1772.7240	212.844	8.329	0.000	1355.008	2190.440
sex_1	70.4981	402.658	0.175	0.861	-719.735	860.731
smoker_1	2.268e+04	508.044	44.648	0.000	2.17e+04	2.37e+04
children_0_1	42.0528	416.949	0.101	0.920	-776.227	860.333
children_1_1	263.7529	474.098	0.556	0.578	-666.684	1194.190
children_2_1	1563.5188	523.927	2.984	0.003	535.290	2591.747
children_3_1	873.4268	600.356	1.455	0.146	-304.796	2051.650
children_4_1	2677.4248	1172.073	2.284	0.023	377.184	4977.666
children_5_1	1042.5631	1334.109	0.781	0.435	-1575.679	3660.805
SW_1	1426.1022	358.018	3.983	0.000	723.478	2128.726
SE_1	1242.4228	364.633	3.407	0.001	526.816	1958.030
NW_1	2113.1498	364.828	5.792	0.000	1397.160	2829.139
NE_1	1681.0644	364.648	4.610	0.000	965.427	2396.702

Omnibus:	230.569	**Durbin-Watson:**	1.987
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	523.089
Skew:	1.337	**Prob(JB):**	2.59e-114
Kurtosis:	5.506	**Cond. No.**	1.57e+16

From the summary, there are 5 features with P>|t| less than 0.05 which indicates their coefficient are not significantly different with 0 in 95% confident interval.

Then the above model is used for predicting the test data and the result is shown below:

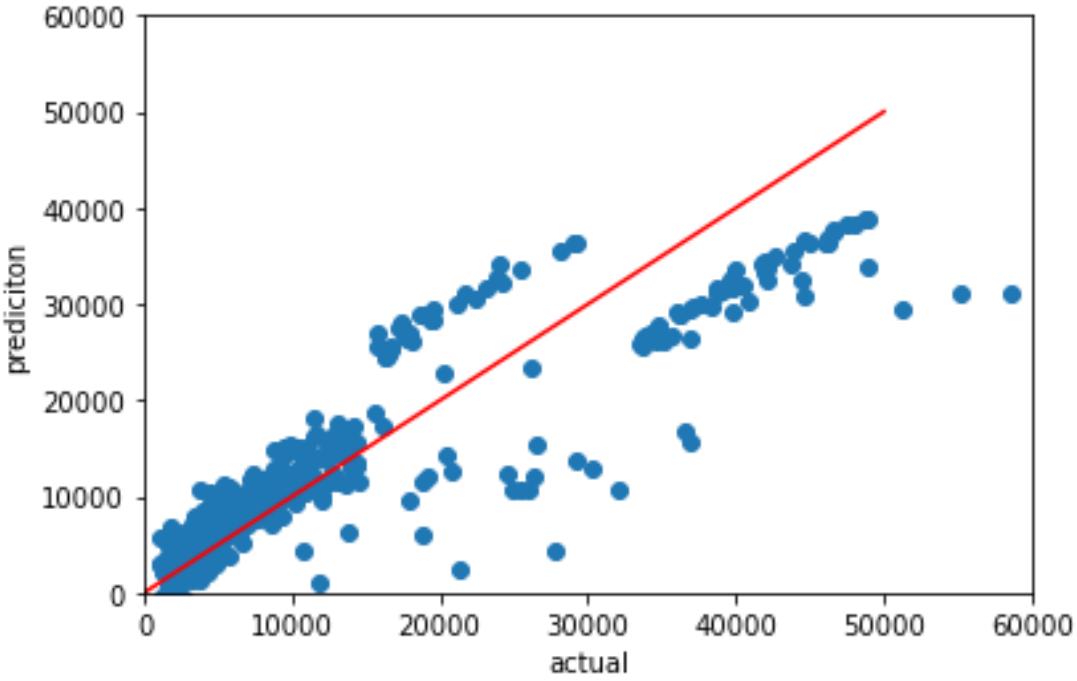


Figure 7.1 – Actual charges against prediction charges (initial OLS)

The R^2 score of the initial model (compare actual and prediction) is 0.786 and the root mean square score (RMSE) is 6034. These metrics will be compared with different models to find the model with the best performance.

Feature Selection - Recursive Feature Elimination and Cross-Validation Selection (RFECV)

To increase the accuracy and simplicity of OLS model, RFECV has been adopted to eliminate irrelevant feature. Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. By recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid.

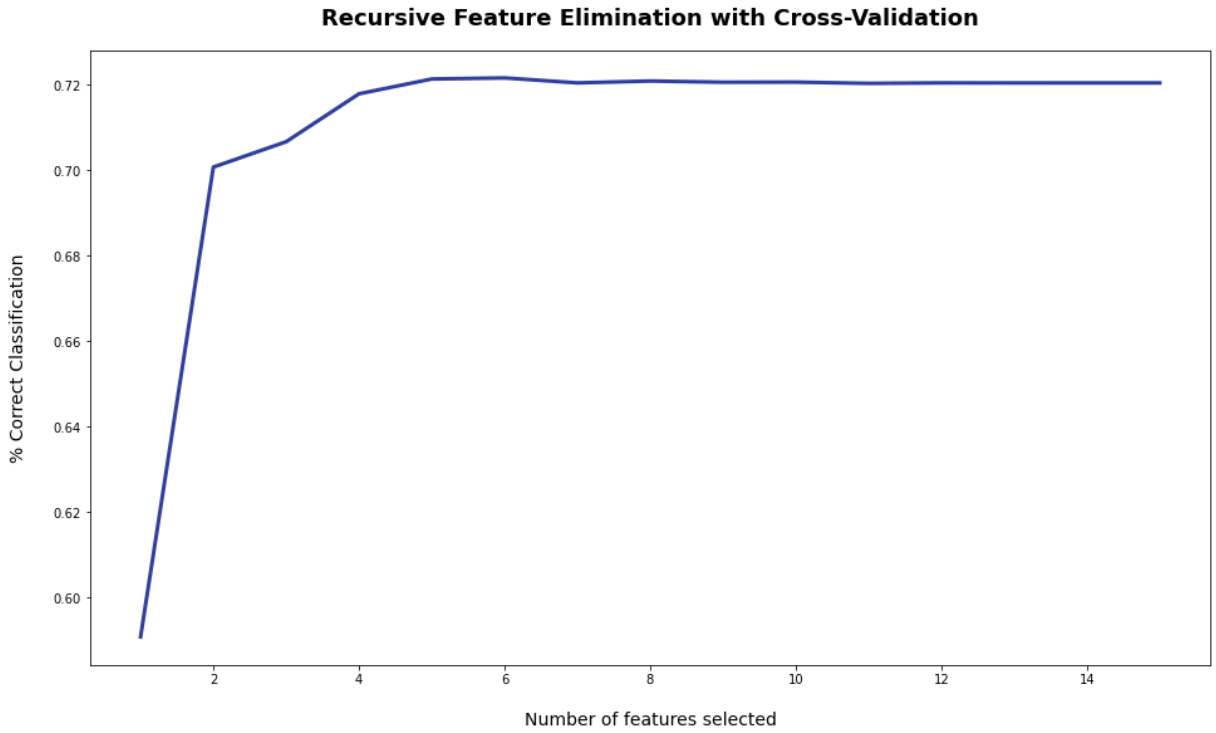


Figure 7.2 – Result of RFECV for OLS model

From figure 7.2, REFCV method suggested the optimal number of features is 6. Then, the another OLS model has been established with only importing 6 most important feature this time, the result is as follow:

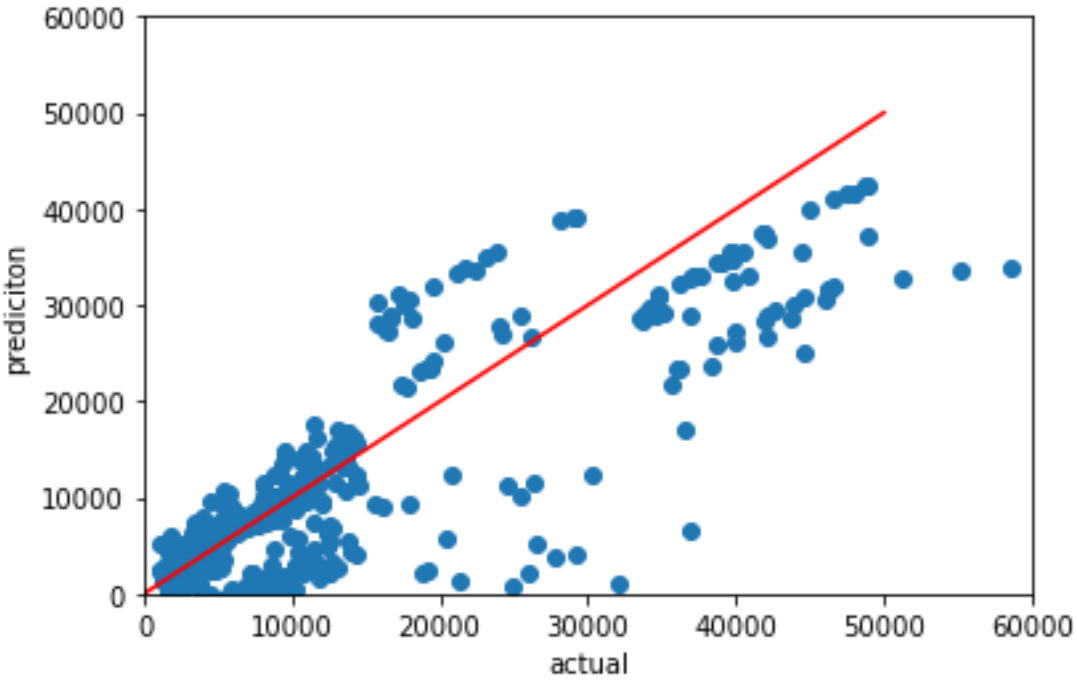


Figure 7.3 – Actual charges against prediction charges (OLS with 6 most important features)

The R^2 score of the model (compare actual and prediction) is 0.643 and the root mean square score (RMSE) is 7793. Although the model becomes much simpler, the prediction power decreases.

Re-construct OLS by only dropping the 5 features which P>|t| is less than 0.05

As the RFECV method does not come up a better model. Another approach is adopted to fine tune the model parameters in this section. This approach uses the OLS summary report of the initial OLS model as reference. In the new model, the five features which P>|t| is less than 0.05 has been eliminated as P>|t| less than 0.05 indicates the independent variable is not correlated with the dependent variable. The result of the new model are shown below:

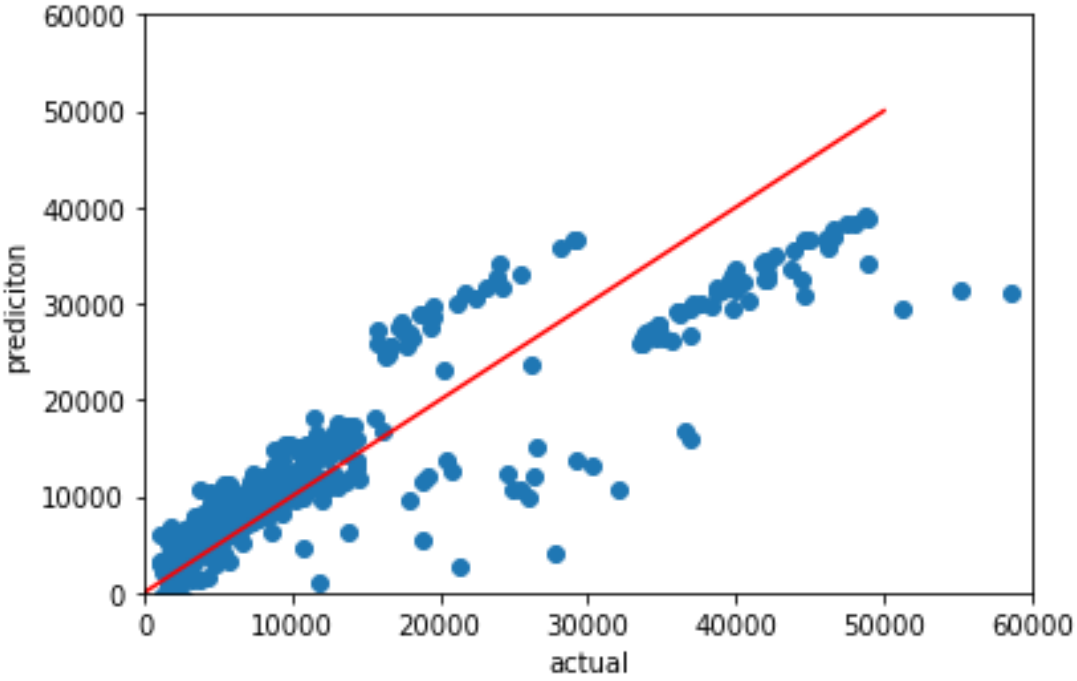


Figure 7.4 – Actual charges against prediction charges (OLS dropping 5 features which P>|t| is less than 0.05)

The R^2 score of the model (compare actual and prediction) is 0.785 and the root mean square score (RMSE) is 6042. Among the three OLS models, this models has the greatest prediction power and second in simplicity.

OLS Model Selection

The performance of the 3 OLS models are summarized as follow:

Model	R2 Score	RMSE
All Features	0.786	6033
Best 6 Features	0.643	7793
Drop 5 Features	0.785	6042

As the last model has the highest R2 score and lowest RMSE score, it has been selected as the best OLS model.

Cross Validation

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

To prevent over-fitting, cross validation has been performed in the best OLS model and the result is shown below:

Cross-validation No.	R2 Score
1	0.749
2	0.752
3	0.738
4	0.660
5	0.714

As the R2 score of the 5 cross-validation are not fluctuating, the chance of

over-fitting can be regarded as low.

7.2 Random Forest Regressor

Beside OLS, this project also establishes random forest regressor (RFR) model to compare the model performance with OLS. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

Hyperparameter

When adopting RFR model, there are two hypermeters (no. of decision tree in the forest and the metrics of evaluating the model) need to be adjusted. In this report, GridSearchCV method has been adopted to fine-tune the hyperparameter. The model performance of 0 to 500 decision tree has been evaluated by both maximum absolute error (mae) and maximum square error (mse).

The result shows that the model perform best with 133 decision trees using mse as the metric.

Initial RFR model

Same as OLS model, the initial model utilizes all the features in the dataset and the performance of the initial RFR model is shown below:

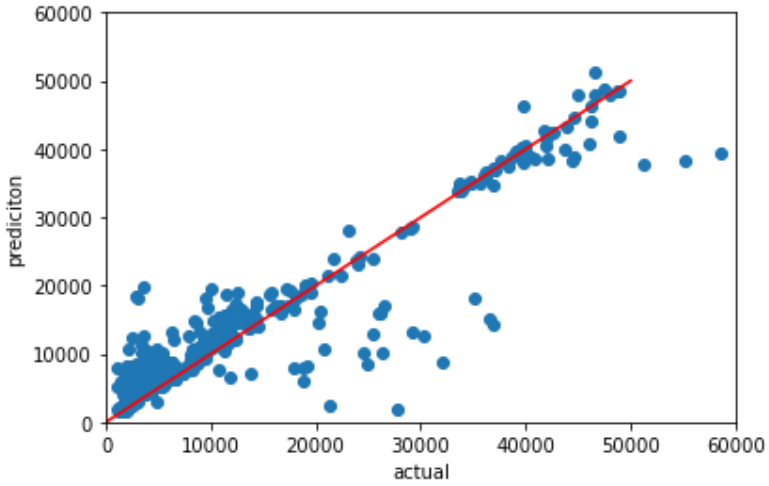


Figure 7.5 - Actual charges against prediction charges (Initial RFR model)

The R^2 score of the model (compare actual and prediction) is 0.862 and the root mean square score (RMSE) is 4840. These results will then be compared with different RFR models.

Feature Selection - Recursive Feature Elimination and Cross-Validation Selection (RFECV)

The same RFECV methods are then adopted to find optimal no. of features for the RFR model. By going through the same steps in Section 7.1, the report finds 6 best features are the optimal no. of features and results are as follow:

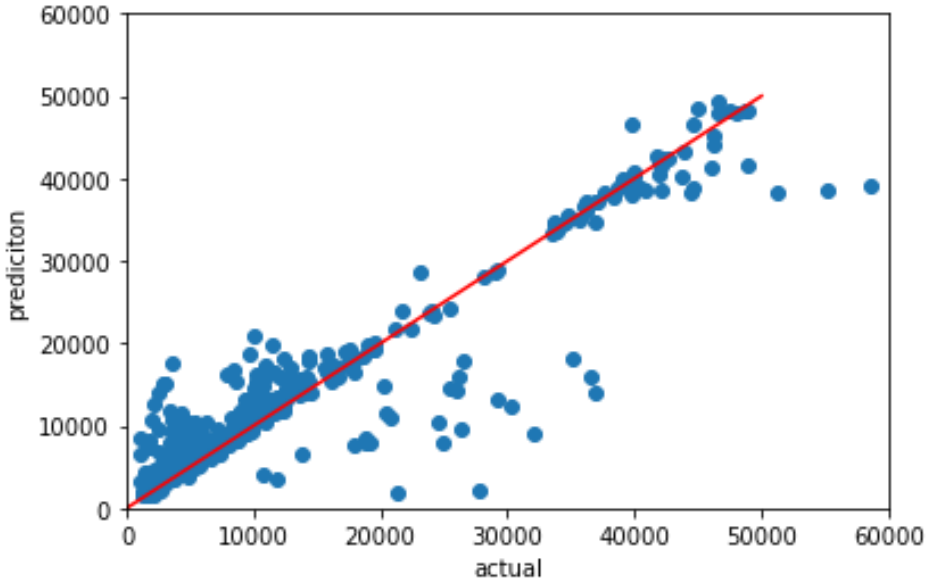


Figure 7.5 - Actual charges against prediction charges (RFR model with 6 most important features)

The R^2 score of the model (compare actual and prediction) is 0.862 and the root mean square score (RMSE) is 4843. These results will then be compared with different RFR models.

RFR Model Selection

The performance of the 2 RFR models are summarized as follow:

Model	R2 Score	RMSE
All Features	0.862	4840
Best 6 Features	0.862	4843

Although the two models have very similar predicting power, the second model (best 6 features) is a much simpler model than the first model. Therefore, the RFR models with best 6 features are regarded as a better RFR model.

Cross Validation

To prevent over-fitting, cross validation has been performed in the best OLS model and the result is shown below:

Cross-validation No.	R2 Score
1	0.825
2	0.865
3	0.820
4	0.770
5	0.817

As the R2 score of the 5 cross-validation are not fluctuating, the chance of over-fitting can be regarded as low.

7.3 Summary of modelling and features importance

The best RFR and OLS model are compared in the below table:

Model	R2 Score	RMSE
Best OLS	0.785	6042
Best RFR	0.862	4844

As the best RFR is simper and has greater prediction power than the best OLS model. The Random Forest Regressor with the best 6 features are adopted as the final selected model in this project.

From the RFECV methods, the feature importance of the RFR model is shown below:

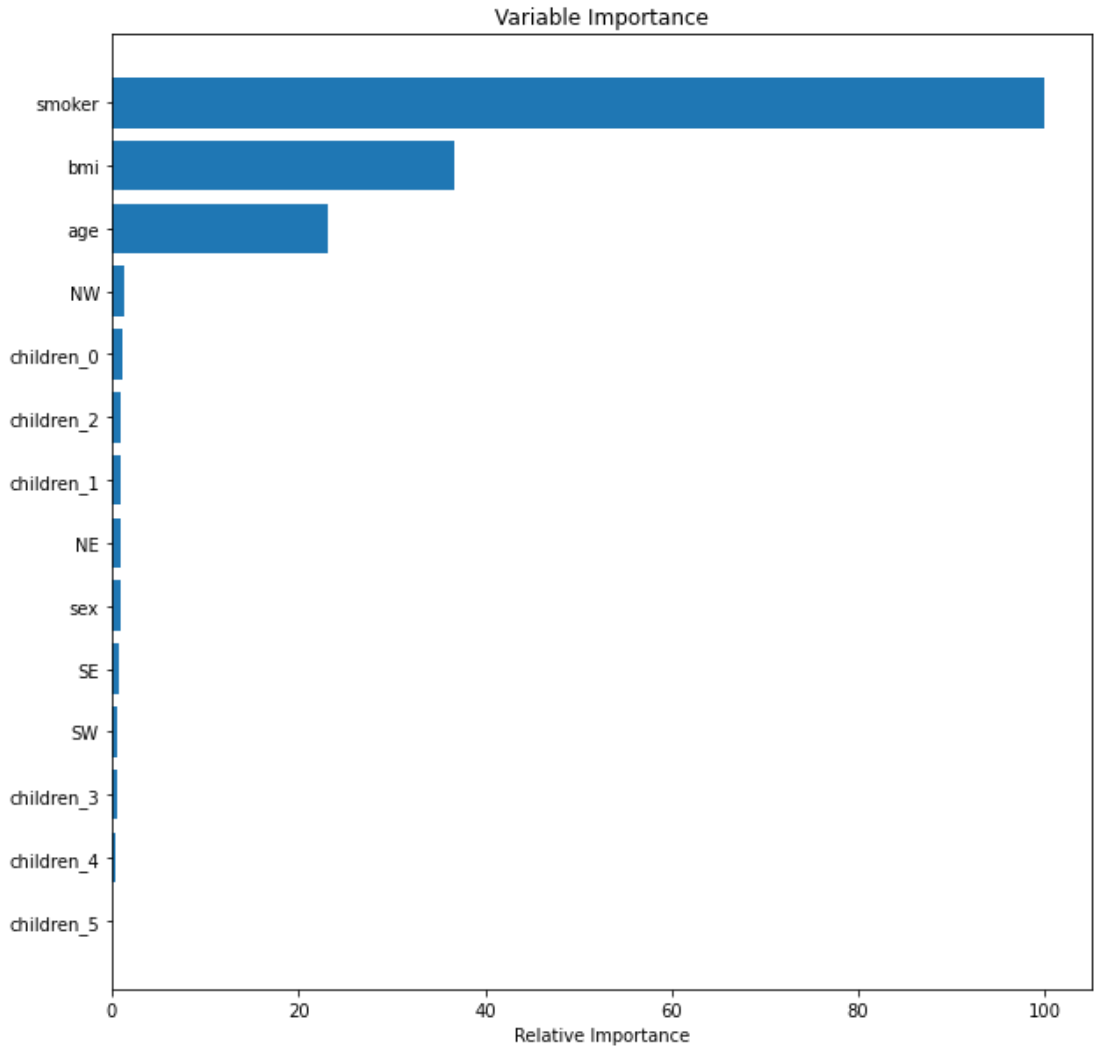


Figure 7.6 – Feature importance from RFECV

8. Result and discussion

The main findings and insights of the report are as follow:

1) This report establishes a medical cost prediction models (with r2 score 0.862) which can help to predict medical cost of potential customers. The start-up companies can help their customers planning and evaluating their medical insurance.

2) The six most important features are smoker, bmi, age, living in northwest region and having zero or two children. This finding can help companies understanding the market condition and provide directions for start-up companies to set-up their target customers.

3) Some features such as gender, no. of children greater than 2, region other than northwest are non-essential factors.

4) RFECV is not helpful to the OLS model but can effectively eliminate less important features in the RFR model. This may be due to the underlying principle of RFECV and RFR are more similar (using subset of features and keep adding / subtracting no. of features in parallel trials)

9. Future Studies

Due to constraints of time and resources, the project has both limitations and rooms for improvement. The directions of future studies can be as followed:

1) Acquire more data. The quantity of the gathered customers information is low (1337 customers). More sampling can be carried out randomly to obtain more data. This can increase the predicting power of the ML methods and further reduce the probability of overfitting.

2) Comprise more customer information in the dataset. The current result indicates that the medical cost is highly related to the health and behavior of the customers. Information such as medical history and drug can be incorporated in the project. Hence, more features can be analysis.

3) More ML methods such as Neural Networks and K-Nearest Neighbors can be established for comparison. This report compared the performance OLS and RFR models and it is possible to find a better model if more ML methods have been reviewed.

10. Conclusion

In conclusion, the project explored the important factors affecting individual medical insurance cost and established a model for predicting the medical insurance cost based on six features. With the models, start-up companies can understand relationship between features and medical cost, help their customers to plan and manage their medical insurance and find potential customers.