JSC370 Final Project

Xinpeng Shan

2022/4/17

Introduction

As medical treatments become more costly during the global pandemic, it has imposed heavy medical burdens on people who requires medical services. According to The National Health Expenditure Accounts, US healthcare spending had grew by 9.7% in 2020 and the medical cost has a 19.7% share in the nation's GDP. To give people information on the crucial factors that may increase there health insurance charges in the US and give people idea on the total insurance charges they need to pay, the dataset used in this research is regarding the medical costs billed by health insurance companies in the US. To solve the problem of accurately predicting the insurance charges for individuals who live in the United States, the aim of the project is to build a machine learning model that can best predict the individual insurance charges, provide that the relavent informations of that individual is given.

Research Question

Since the aim of this research is to predict individual insurance charges in the US, two main research questions are formulated below. By the end of the research, the three research questions will be answered and supported by evidences.

1. Which factor is the most crucial one regarding on the insurance charges paid by individuals?
2. Do females pay more than males on insurance charges?
3. How to predict insurance charges by personal informations accurately?

Methods

Data Source

The dataset is an open-source dataset acquired from Kaggle, it is a sample of medical insurance information from the whole population in the US (The year that the data is aquired is not provided). There are 1338 observations in the dataset with variables like age, gender, bmi, etc. for each individual observations. The summary statistics of all the variables in this dataset summarized in result section.

Data wrangling and data exploration

The following steps are used on data wrangling and data exploration:

- Checking the variable types and if there are missing values in the dataset.
- Visualizing the distribution of each variable to check if there are distinctive shapes in those variables.
- Creating interactive visualizations to showcase the relationship between different factors and the insurance costs billed by insurance companies.

Machine Learning models

Six machine learning models are builded in this research including linear regression, regression tree, bagging, random forest, boosting and XGBoost, in order to compare the performance of each model to find the best

fitted one on predicting the insurance charges.

Results

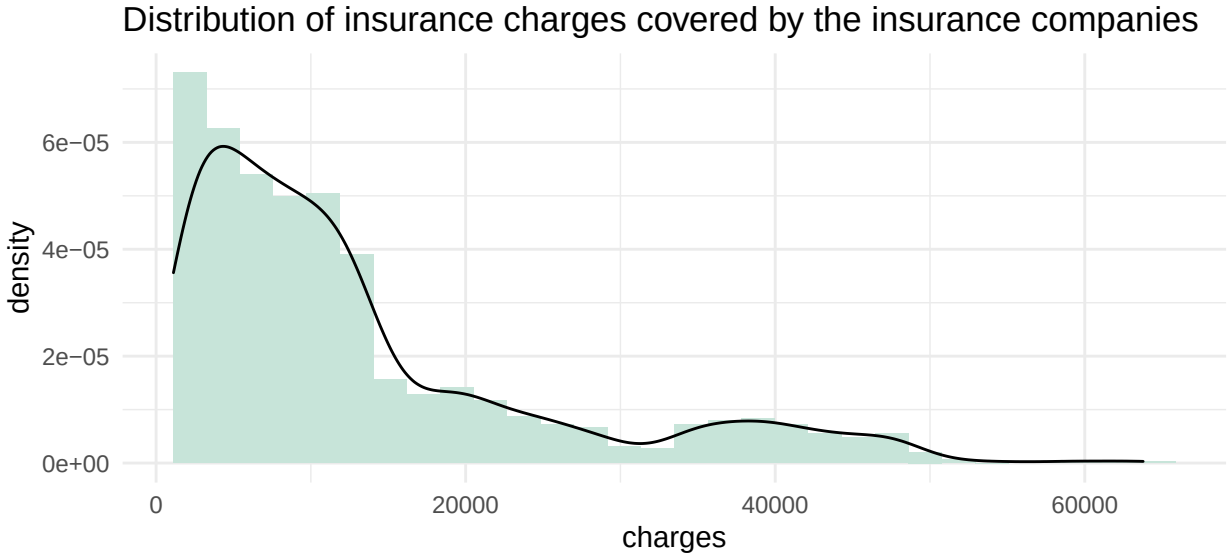
Summary Statistics and Discription of each Variable

Table 1: Informations on variables in the dataset

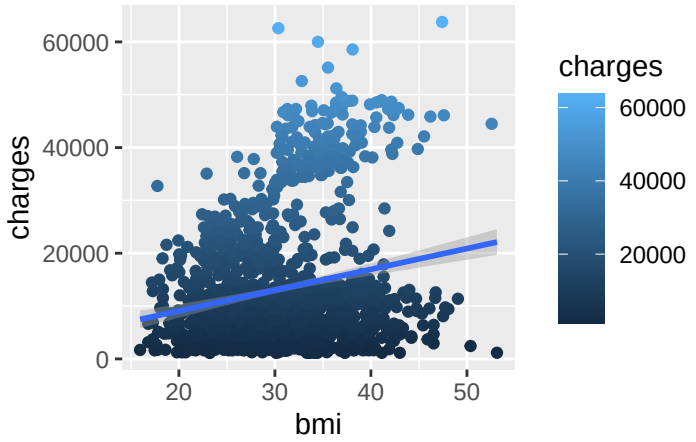
Variable	Description	type	missing values
age	age of beneficiary	numarical	0
sex	gender of insurance contractor	categorical (male, female)	0
bmi	body mass index of beneficiary	numarical	0
children	number of children covered by the insurance	categorical (0, 1, 2, 3, 4, 5)	0
smoker	if the beneficiary smokes	categorical (yes, no)	0
region	the beneficiary's residances in US	categorical (northeast, northwest, southeast, southwest)	0
charges	the medical cost covered by the company	numarical	0

Variable Distributions

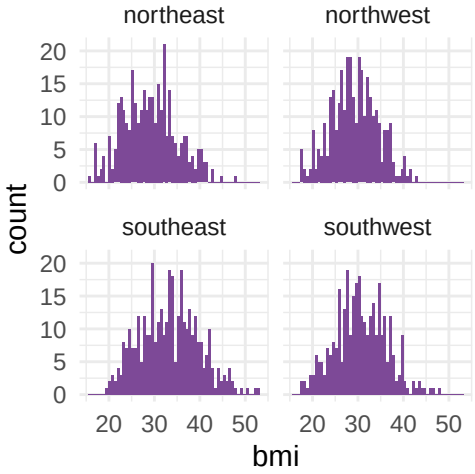
To do further analysis using the dataset, we need to first understand the distribution of the variable we want to predict, which is **charges**. From Figure 1, we can see that the distribution of charges is extremely right-skewed, indicating that a large percantage of observations in the dataset have insurance charges close to 0.



The scatterplot below showcase the distribution of bmi, we can notice that as bmi increases, the value of insurance charges increases. From the histograms of bmi by region, bmi has a normal distribution shape for all four regions in the US.

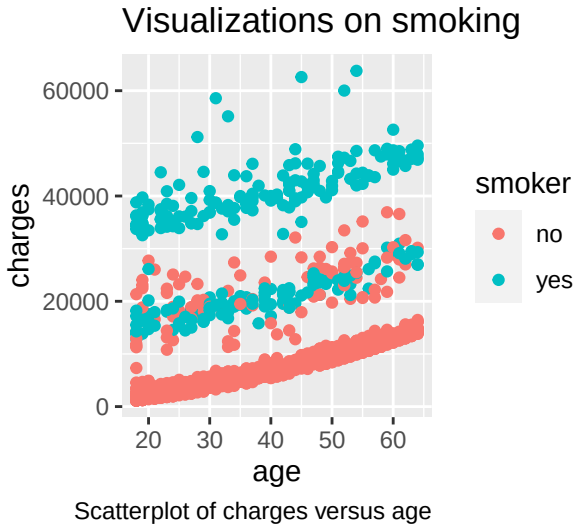


Charges versus bmi colored by charges

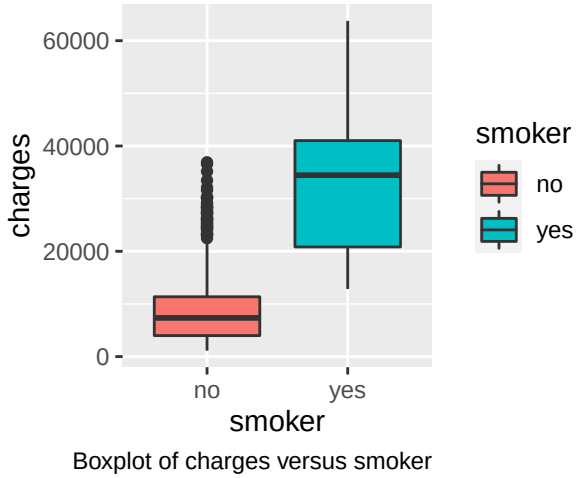


Histograms of bmi by region

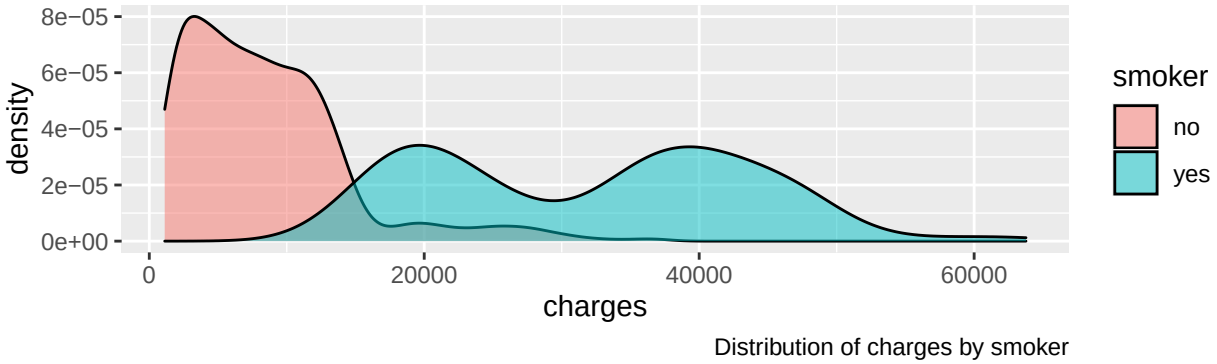
The plots below are regarding whether smoking has an influence on insurance charges. From Figure 3, we can notice that smokers have much higher insurance charges compare to individuals who do not smoke. The high insurance charges also indicates the worsening of health situation that people are in. From the distribution of charges by smoker, we can see that the distribution of charges for non-smokers is right-skewed, indication the lower cost they spend on medical insurance, while the distribution of charges for smokers have a bimodal distribution, with the two modes both have charges larger than the mode for the distribution of non-smokers.



Scatterplot of charges versus age



Boxplot of charges versus smoker



Distribution of charges by smoker

Interactive visualizations

Smoking and Insurance Charges

The first interactive visualization is a barplot with age as x-axis, charges as y-axis and colored by smoker, we can easily notice that the insurance charges is higher across all ages for people who are smokers.

Region and Insurance Charges

To see if people at different age have similar insurance charges in different regions of US, the scatterplot and heatmap on the mean insurance charges for people in different region and age are created. We can notice that in southeast and northeast region, people who are around 20 years old have large insurance charges covered compare with people at the same age in southwest and northwest regions. Also, people with higher age tend to have larger insurance charges.

Number of Children and Insurance Charges

To compare the insurance charges paid by people who have different numbers of children, I created the scatterplot and barplot below. In the scatterplot, the size of the bubble is proportional to the mean charges paid by people at that specific age with that number of children.

For people at a younger age and with less than 4 children, the insurance charges do not differ by a great amount. However, for people over 50 years old, the amount of insurance charges are higher for people who have 1-3 children. From the data we have, people who have 4 children have a higher amount of insurance charges at a younger age compared with people who have other numbers of children, but it may be caused by the lack of data collected for people who have 4 or 5 children.

Machine Learning Models

Before building the models, the dataset is split into training and testing datasets. The training dataset has 75% of data and the testing dataset has 25% of data. Here are the number of observations in the each dataset:

Table 2: Number of rows in the two dataset

dataset	Number of Rows
train	993
test	345

Linear Regression

Firstly, I built a linear regression model with all variables as predictor and charges as responses using the training dataset. From Table 3, we can see that age, bmi, smoker and children have p-values that are extremely small, which indicates that it is very likely that those variables explains the variation in the response.

Table 3: p-values of each predictor in the linear regression model built using training dataset

	Pr(>|t|)
age	< 2e-16
sexmale	0.48599
bmi	< 2e-16
children	0.000577
smokeryes	<2e-16
regionnorthwest	0.29430
regionsoutheast	0.33319

	Pr(>|t|)
regionsouthwest	0.03229

Then, I built the two models below to decide which one to use as the final model.

- model 1: Prodictor: age, bmi, smoker, children
- model 2: Prodictor: age, bmi, smoker

The information of the two models are summarised in the table below.

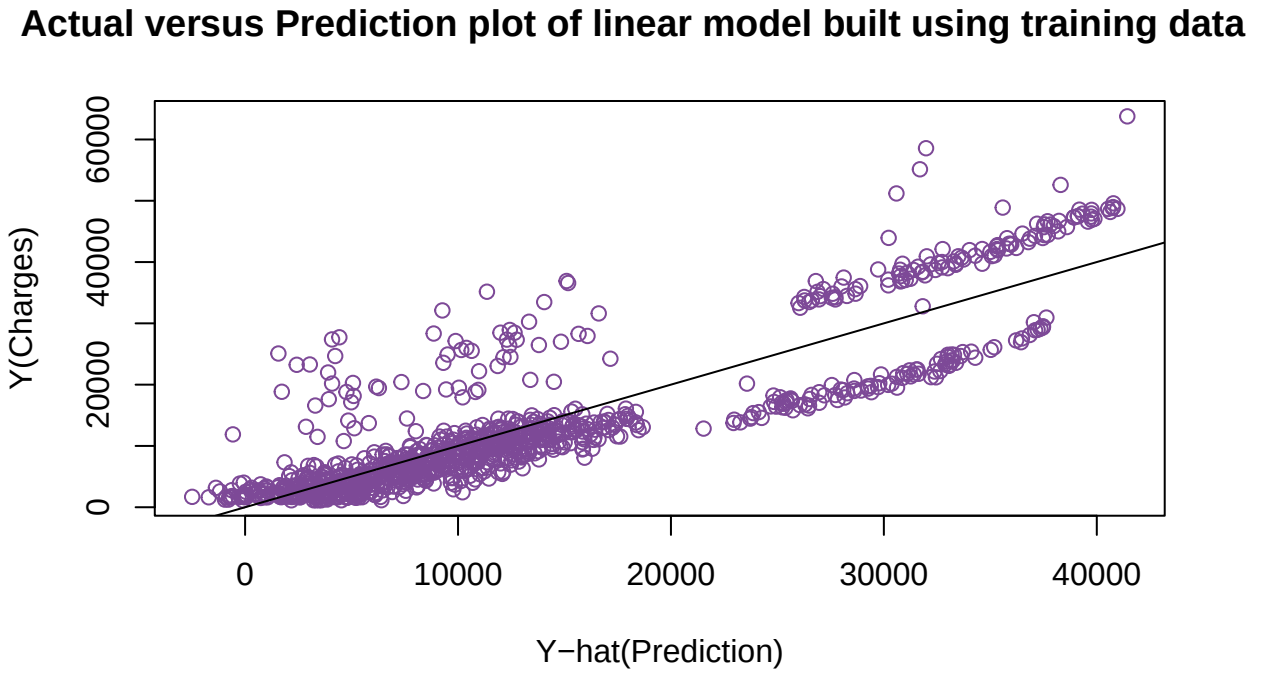
Table 4: Adjusted R squared of the two models

model	Number of predictors	Adjusted R squared
model 1	4	0.7657707
model 2	3	0.7640206

From the table above, we can notice that although model 2 has a slightly (0.0015) lower Adj R squared compared with model 1, it has 1 less predictor, which makes it a more feasible linear model because it can use minimal information to create a reasonable prediction. Therefore, we choose model 2 to be the final model.

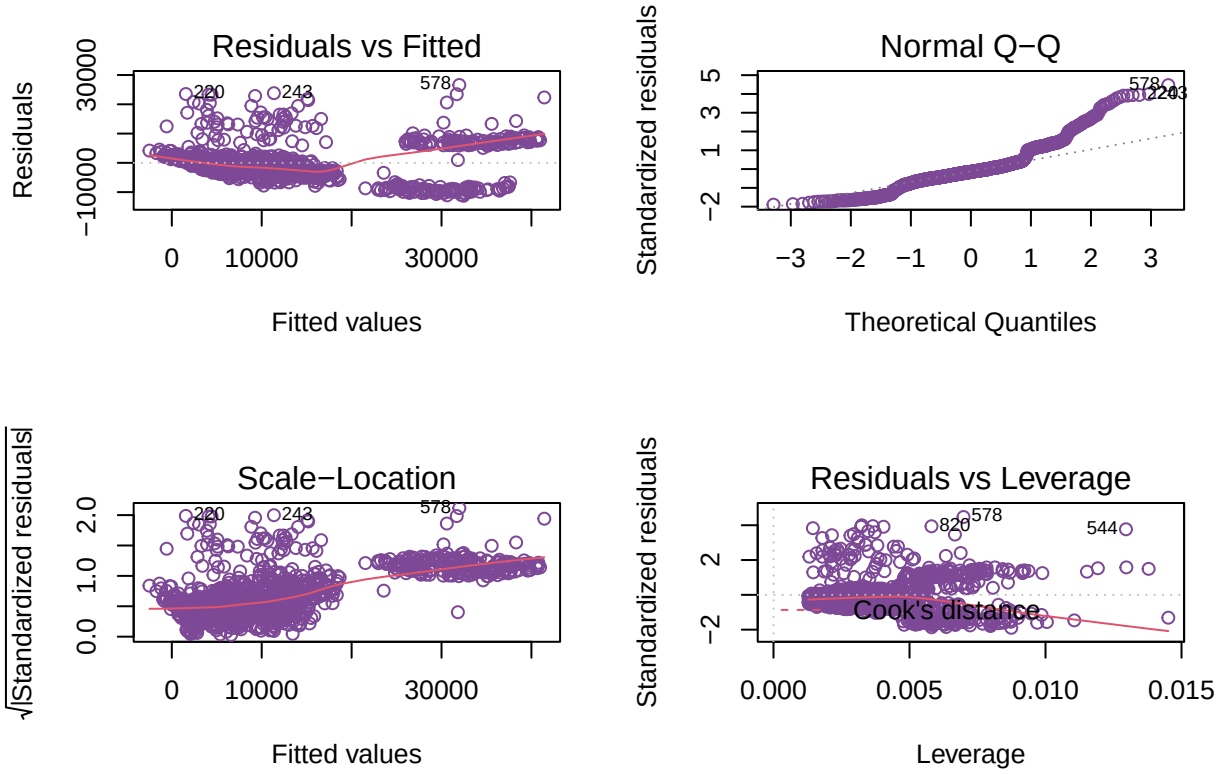
- Analyse performance of the final linear regression model

From the Actual versus Prediction Plot, we can see that the points are scattered around the fitted line and there are very few points that appear to be far away from the line vertically, which indicates there are few outliers in the dataset. Since there is no point that is horizontally further away from the mean, there appears to be no significant leverage that pulls the regression line by a great amount.



In the Residual vs. Fitted plot, there is a curved shape, which indicates a slight violation of the linearity assumption. From the Normal Q-Q plot, we can notice that there is some deviation at the right end of the Normal Q-Q plot, which means there is a slight violation of the normality assumption. There is an upward trend on the Scale-Location plot, which is caused by the unequal variance. And finally, in the Residual vs. Leverage plot, There are some amount of points that lie below Cook's distance, which demonstrates the

influential points in the training dataset.

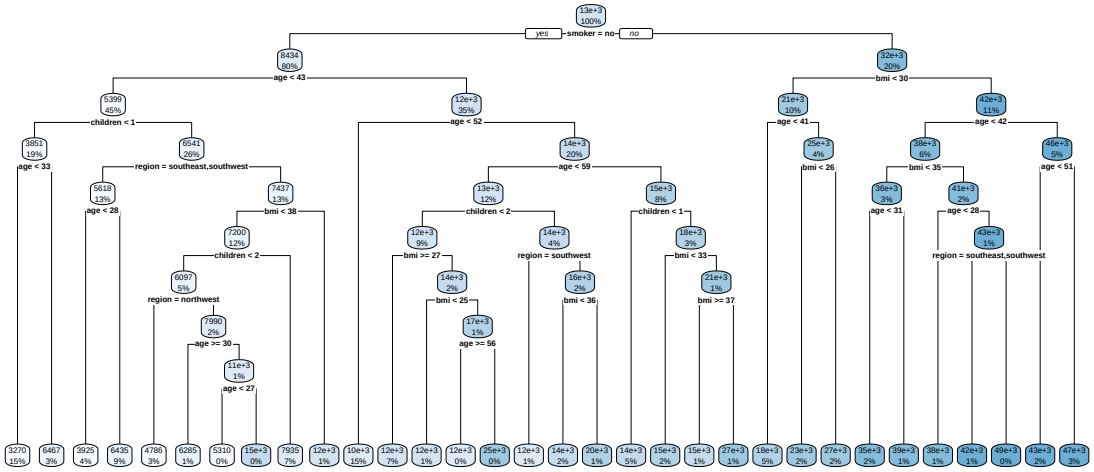


- Validate the model using the test dataset

To check the performance of the final model, I built another model using the same predictors but using the test dataset. It achieves an adj R squared of 0.7048884, which means that our model explains around 70% of the variation in total charge in future data. There is no big difference in the violation of assumptions. The estimated coefficients do not differ by a great amount from the training dataset. Therefore, I can conclude that the model is validated.

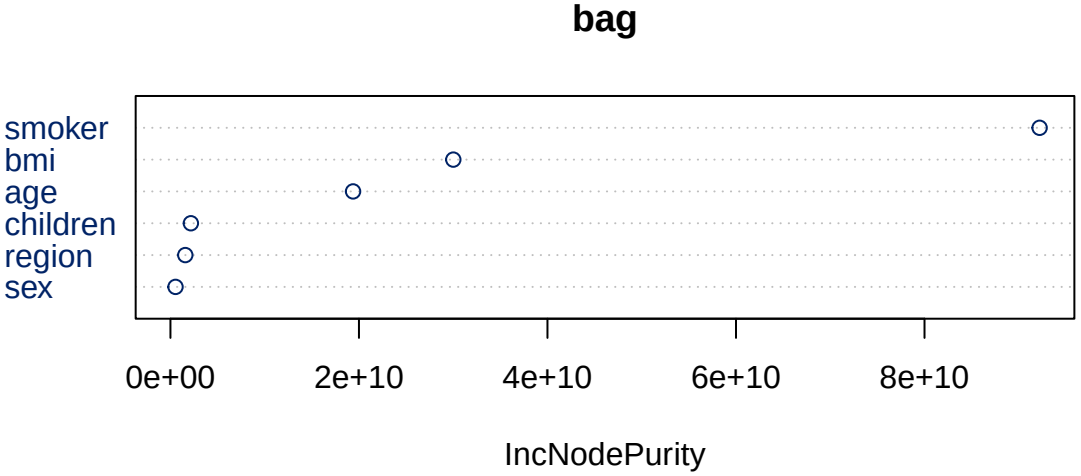
Regression Tree

In this section, we built a regression tree to predict charges. To fit a regression tree with reasonable number of splits, we need to determine the optimal complexity parameter (cp) that have the minimal xerror to prune the tree. This process reduce the complexity of regression tree and prevent building a model that is overfited to the training dataset. From the CP table, the minimal xerror is 0.15573 with 7 splits, and 0.00196 as the optimal cp. The pruned tree is shown below.



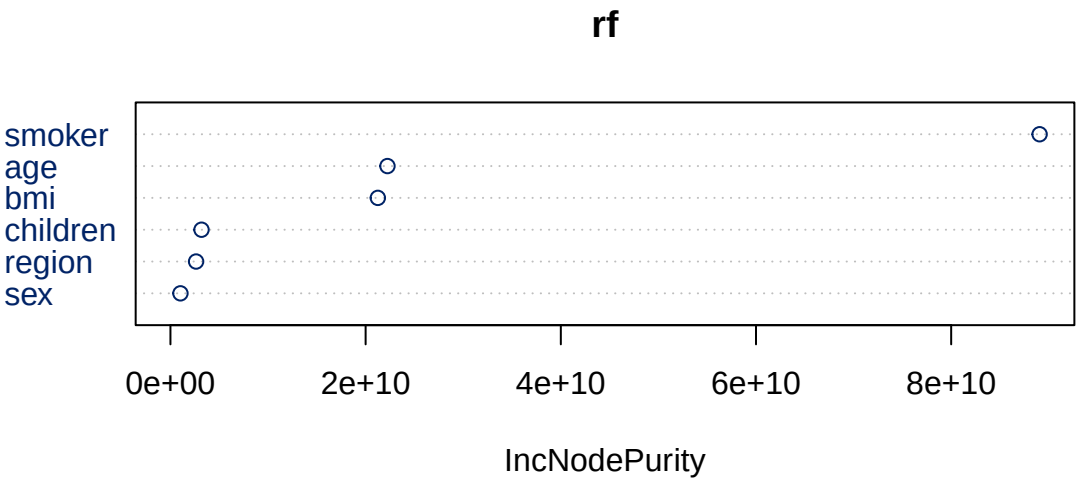
Bagging

A bagging model is built to predict charges. From the variable importance plot below, **smoker** is the most important feature on predicting insurance charges, followed by bmi and age.



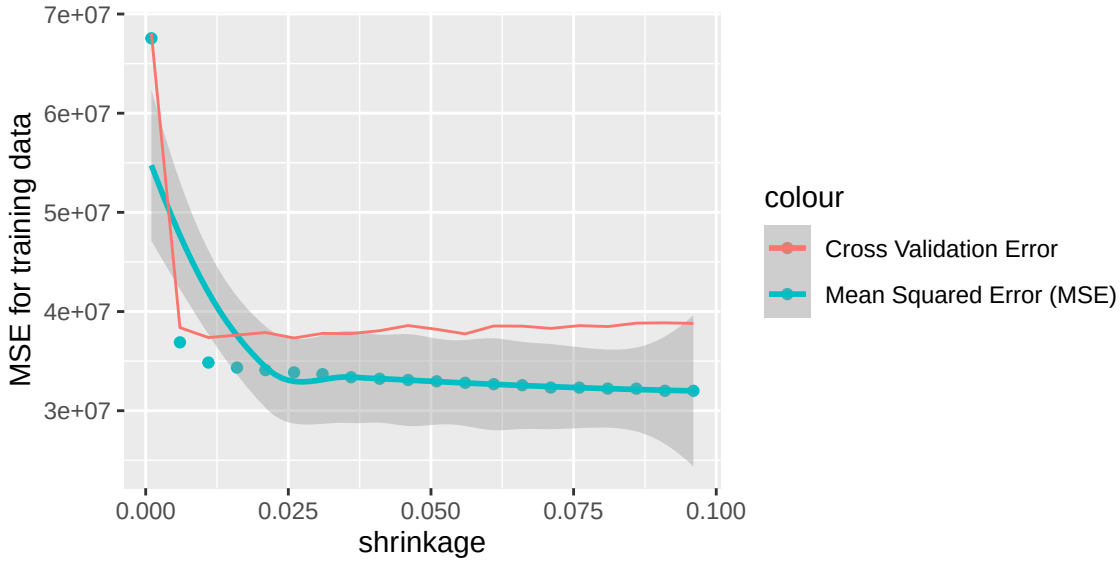
Random Forest

A random forest model is built to predict charges. The variable importance plot is shown below. Similar to the Bagging model,**smoker** is the most important feature on predicting insurance charges, followed by bmi and age.



Boosting

In this section, boosting model with 1000 trees is built using the training dataset. To pick the best value for the shrinkage parameter, a sequence of boosting models is built with shrinkage parameter values from a range of 0.001 to 0.1 with a step size of 0.005. The training MSE and cross-validation error are calculated for each model and plotted in the graph below. We can notice that the MSE for training data is the smallest when the shrinkage parameter is around 0.025 and the cross-validation error is flattened at around 0.01. Thus, 0.025 is chosen for the shrinkage parameter to build the final boosting model. In this way, the model will not only have a good performance on the training dataset but also can be used to predict unseen future data.



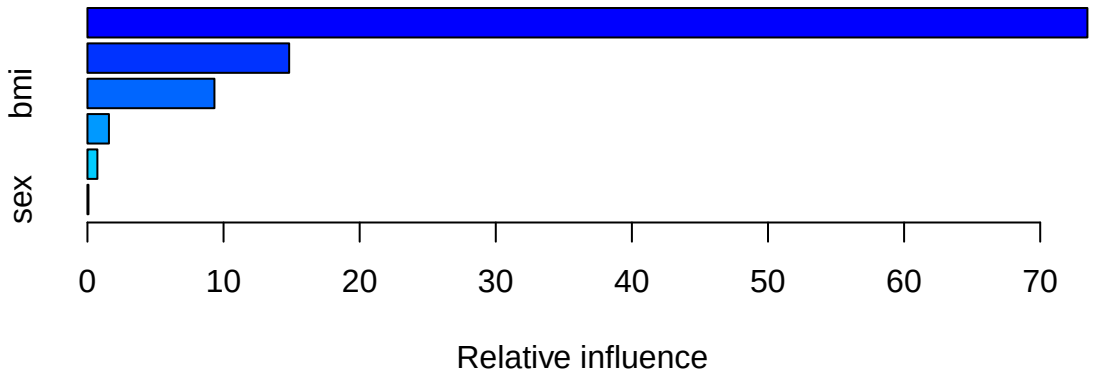


Table 5: Relative influence of each variable in Boosting Model

	var	rel.inf
smoker	smoker	73.4671862
age	age	14.8196906
bmi	bmi	9.3296705
children	children	1.5797069
region	region	0.7302521
sex	sex	0.0734936

From the variable importance plot and the table witch generalized the variable importance values for each variable. The level of importance for each variable have a similar order compare with the previous models. We can notice that **smoker** have a relative influence of 73.2600934 which is five times higher than the relative influence of age. On contrast, **sex** is the variable with least relative influence with a value of 0.0451726

XGBoost

Finally, we created Extreme Gradient Boosting model to predict the salary and set up a grid search on max_depth, nrounds and eta. We trained the XGBoost model using the tuning grid and plot the variable importance plot below.

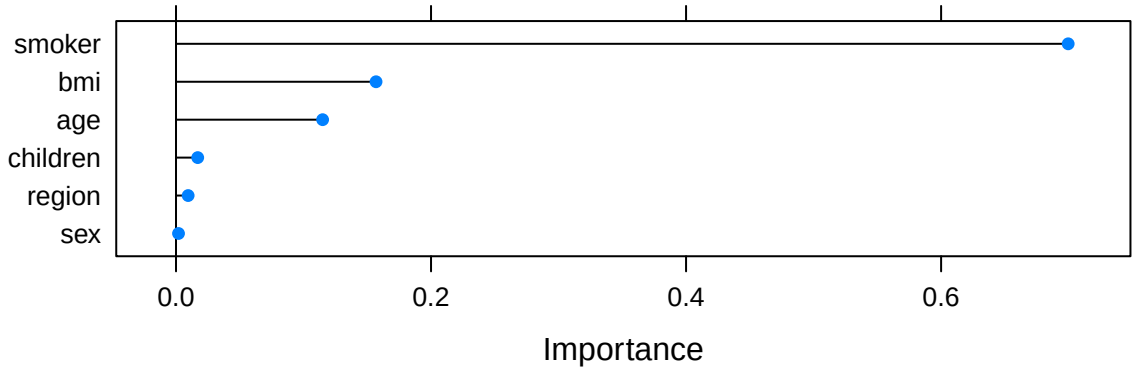


Figure 1: Variable Importance Plot of XGBoost Model

Conclusion and Summary

Key Findings

- From the visualizations and the predictors in the linear regression model, we can notice whether the beneficiary smokes affect the amount of insurance bills that the insurance company pays by a large amount (more than 20000 higher on insurance bill for person who smokes), while variable bmi and age also have a positive association with charges (the coefficients are 289.05 and 205.65 respectively in test dataset).
- The final linear regression model has 3 predictors including age, bmi, smoker. The linear model explains around 76% of the variation of the response variable (charges) in the training dataset and around 70% of the variation in the testing dataset
- From the variable importance plots of Bagging, Random Forest, Boosting and XGBoost, all of them suggest that **smoker** is the most important factor that will affect the insurance charges, while **sex**, **children** and **region** are factors that do not have large influences on insurance charges.

Summary tables

Linear Regression

The table below summarized the coefficients of the linear regression model built using the training and testing dataset.

Table 6: Estimated coefficients of the two linear models built using training and testing dataset respectively

Dataset used to build the linear model	Coefficient of age	Coefficient of bmi	Coefficient of smokeryes
train	277.91	328.52	23581.99
test	205.65	289.05	24421.41

Mechine learning models

The table below summarized the Test RMSE for each mechine learning model. From the result, we can see that Extreme Gradient Boosting Model have the smallest RMSE, with a value of 4162.544, followed by Regression Tree (4836.064) and Random Forest (5091.062). Since , The test RMSE for all 5 models are not terriblly large consider that **charges** takes value from 0 up to 60000. Since smaller in test RMSE gives insights on the performence of each model in future datas, we will choose XGBoost to be the final model on predicting charges.

Table 7: Test RMSE for each mechine learning model

Model	test RMSE
Regression Tree	4763.592
Bagging	5303.973
Random Forest	5112.717
Boosting	5945.542
Extreme Gradient Boosting (XGBoost)	4294.575

Since the higher insurance bills that the insurance company pays, the larger the bill from the hospital, we can see the drawbacks of smoking from the statistical results. To reduce the medical burden on individuals, we could start by quitting unhealthy habits like smoking and increasing the amount of time we spend on physical exercise to lower the bmi.

Conclusion

To answer the research question, whether people smokes is the most crucial factor regarding on the insurance charges paid by individuals. It is both supported by the coefficient of **smokeryes** from table ? and the variable importance plots of all machine learning models built using the training dataset.

For the second research question, from the p-value table, "sexmale" has a p-value of 0.48599, which suggests that sex do not explain the variation on insurance charges. Also, from the variable importance plots, it is easy to see that sex is the factor that do not have much influence on charges. So we can conclude that female do not pay more than males on insurance charges.

Finally, for the third research question, we decide on using the Extreme Gradient Boosting model on predicting the insurance charges since it have the smallest RMSE on the testing dataset.

Strengths and limitations

One of the greatest strength of this research is that we aimed to build models that are flexable, simple enough to be used and are not over-fitted to the training dataset. In addition, we meticulously checked all model assumptions when building the linear model, to make sure that the model are used to its intended purposes.

However, there still exist some limitations to the research. Since the year that the dataset is aquired is not specified, it limits the ability of model to have an accurate prediction on datas acquired in other years, as the insurance charges may differ accross years. Since the dataset only have 1338 observations, there is a trade-off between having enough datas in the training dataset to build a more accurate machine learning model and to measure the performance of model on testing dataset.

Appendix

Data source: https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv

Github repo: https://github.com/xinpeng13/JSC370/tree/main/midterm-report

Heathcare spending research: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical