Predicting Insurance Premium Charges

Mudassir & Kaushik

Introduction

Insurance companies primarily make profit off collecting premium that exceeds insured losses and related expenses.

Insurers, whether life or nonlife must assess the risk and likelihood of claims and the value of the claims.

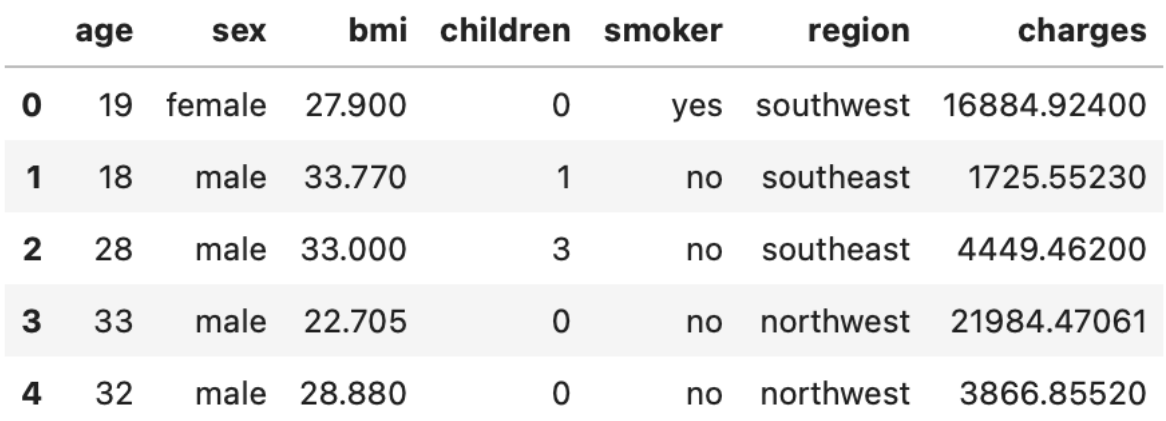
If an insurer is predominant in acquiring a certain life risk disease, the calculated premium is going to way above that of an average person.

Thus, the *problem statement* scope of this project is to perceive the impact of the premium a person pays with factors including one’s age, gender, location, predisposed smoking habits etc.

Data Collection

The data has 1338 records and seven fields being *age, sex, bmi, no. of children, smoker, region*, and the premium charges.

Below is a preview of the dataset chosen



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Pre-Processing of the data:

There can be three types of data's

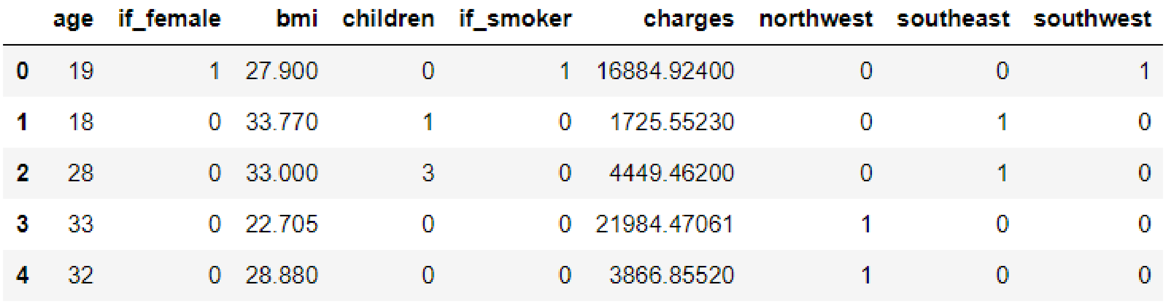
- Structured Data (e.g.: tables)
- Semi-Structured Data (e.g.: emails)
- Unstructured Data (e.g.: videos)

Fortunately, the dataset we are using is structured and doesn't have any null values.

However, we did have to preprocess the categorical data column's such as the sex, smoking habit, and region into a binary format so as to ease up processing and the model we are using is able to understand and extract valuable information.

Also, based on testing the correlation between different variables against the premium charges, we can rule out *no. of children* from our dataset.

Here's the preview of the dataset after it has been pre-processed



	age	if_female	bmi	children	if_smoker	charges	northwest	southeast	southwest
0	19	1	27.900	0	1	16884.92400	0	0	1
1	18	0	33.770	1	0	1725.55230	0	1	0
2	28	0	33.000	3	0	4449.46200	0	1	0
3	33	0	22.705	0	0	21984.47061	1	0	0
4	32	0	28.880	0	0	3866.85520	1	0	0

Training the Dataset

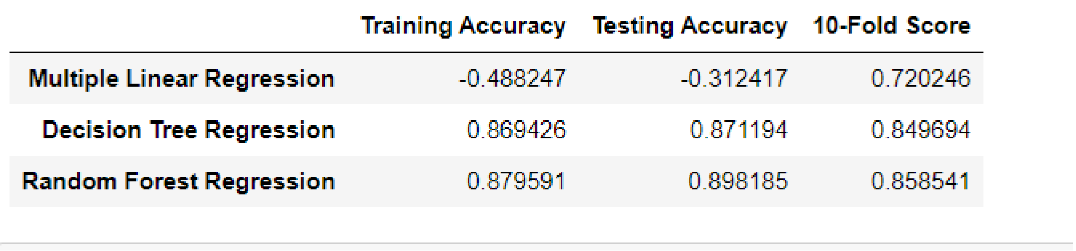
We created two variables “X” and “y” with “X” containing all columns but “charges” and the “y” variable containing only the “charges” column. We are trying to predict the “charges” column also being the variable “y” based on the values or inputs of the other columns or the “X” variable. We divided the dataset in the ratio of 75-25 i.e. 75% of the data is used as training set and 25% as testing set.

Methodologies used:

- Multiple Linear Regression
- Decision Trees

- Random Forest Regression

The R-Squared was used to determine the best methodology and the following results were obtained.



	Training Accuracy	Testing Accuracy	10-Fold Score
Multiple Linear Regression	-0.488247	-0.312417	0.720246
Decision Tree Regression	0.869426	0.871194	0.849694
Random Forest Regression	0.879591	0.898185	0.858541

We decided to go ahead with Random Forest Regression model as it has the highest accuracy when compared to the other models.

Random Forest Regression is a supervised learning algorithm that merges predictions from numerous machine learning algorithms to get a better prediction a sole model.

It works by making many decision trees while training and giving the average of classes as predictions of all trees.

It is usually more accurate than the other models and works well on different kinds of problems. However, overfitting occurs easily and is often difficult to understand the reason behind the predictions and decision made by the model.

Conclusion:

The results we ended up were more along the initial prediction.

People with higher *BMI’s, Habit of Smoking & Age* influence the premium cost of insurance.

A habit of *Smoking* among insurers having a higher premium is plausible since smokers pose a higher risk of contracting a wide array of diseases & health complications thus making the premium cost justifiable in comparison to non-smokers.

In the case of higher *BMI’S*, it is used as a metric of measuring obesity among the insurers. Even though people with lower BMI’s pay the same for food/exercise decisions, externalities associated with it namely the mortality rate pushes the premium price of the insurance.

Lastly, *Age* being a factor that is quite justifiable and non-avoidable in any case since though people are likely to contract diseases in any age, the likeliness & seriousness of the contracted disease is often higher and life threatening, and the rate of recovery from said disease being much smaller in the case of older people makes this premium cost reasonable.