Anish Arora

100803162

Medical Insurance Cost Prediction

Statement of Work – V2

# CONTENTS

[CONTENTS 1](#_Toc57753338)

[ABSTRACT 2](#_Toc57753339)

[INTRODUCTION 2](#_Toc57753340)

[SCOPE OF WORK 2](#_Toc57753341)

[DELIVERABLES 2](#_Toc57753342)

[MILESTONES 2](#_Toc57753343)

[DATASET INFORMATION 3](#_Toc57753344)

[DATA ASSUMPTIONS 3](#_Toc57753345)

[DATA LIMITATIONS 3](#_Toc57753346)

[EXPLORATORY DATA ANALYSIS 3](#_Toc57753347)

[PRE-PROCESSING THE DATA 7](#_Toc57753348)

[STATISTICAL ANALYSIS 8](#_Toc57753349)

[FEATURE ENGINEERING & ITS BENEFITS 9](#_Toc57753350)

[TESTING PROCESS 9](#_Toc57753351)

[ACCEPTANCE 10](#_Toc57753352)

# ABSTRACT

Having a Health Insurance is beneficial and necessary for several reasons. If one has a medical insurance, they are more likely to get the care they require and live a healthier and longer life compared to people not having medical insurance. People without a medical insurance receive less timely care, have worse health outcomes and a lack of insurance is a financial burden for them and their families.

Stroke and heart disease can be expensive, they are in fact a leading cause of medical bankruptcy. Having a high quality medical insurance can help one get access to the care they need and manage their health without large medical bills.

# INTRODUCTION

As we can see through the dataset that the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values, therefore we are going to use regression analysis for this problem. It involves determining the best fit line that passes through all the data points in such a way that distance of the line from each data point is minimized. There are different regression techniques like linear regression, ridge regression, lasso regression etc. that we are going to use and then compare the results to find the best out of them.

# SCOPE OF WORK

The objective of this project is to understand and analyze the factors that influence the price one pays for their health insurance and predict the costs of insurance for individuals. By using different machine learning regression models and comparing the results, we are going to achieve the best model which is more accurate and effective.

# DELIVERABLES

At the end, there will be a python file and a document outlining the entire project that can be used to define variables that influence the price of health insurance and predicting the costs of insurance for individuals. It helps to identify the factors that are responsible for the difference in insurance prices for various individuals.

# MILESTONES

|  |  |
| --- | --- |
| **Milestone** | **Estimated Delivery Date** |
| Statement of Work | 06-Nov-2020 |
| Data Acquisition and Understanding | 1-Dec-2020 |
| Modelling | 18-Dec-2020 |
| Prototyping | 18-Dec-2020 |
| Deployment | 18-Dec-2020 |

# DATASET INFORMATION

The data used for this analysis is obtained from the medical cost personal dataset that consisted of 7 variables and 1,339 records. These records describe the various factors related to the beneficiary – age, sex, bmi, children, smoker, region, charges.

Data Source: Miri Choi (2018, Feb 20). “Medical Cost Personal Datasets”, from <https://www.kaggle.com/mirichoi0218/insurance>

|  |  |
| --- | --- |
| Variables Information | |
| *Variable Name* | ***Description*** |
| *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 – yes/no* |
| *region* | *the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.* |
| *charges* | *Individual medical costs billed by health insurance* |

# DATA ASSUMPTIONS

As the dataset is clean and every variable is important in our analysis so there are no assumptions required.

# DATA LIMITATIONS

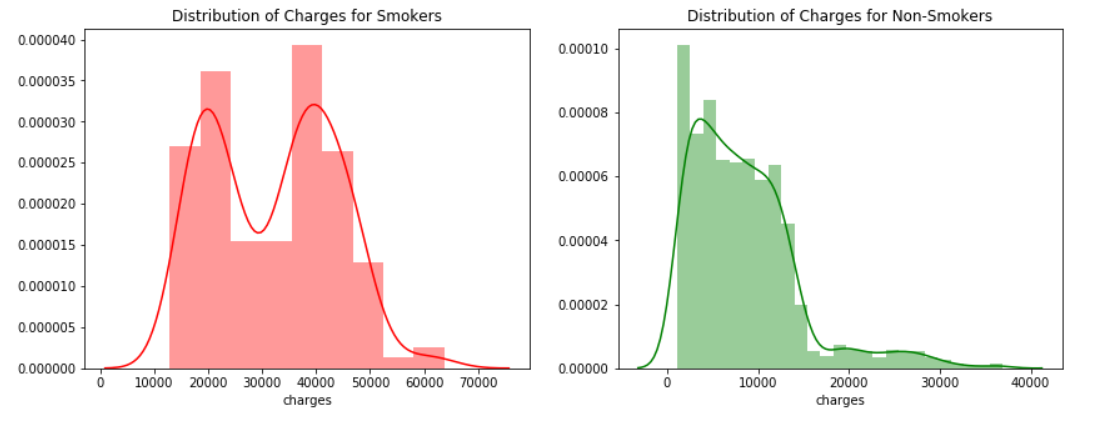
The dataset is simulated based on demographic statistics from the US Census Bureau according to the book Machine Learning with R by Brett Lantz.

As we can see in the dataset, there are so many categorical variables like sex, smoker, region therefore they were encoded in numbers using encoding technique.

# EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

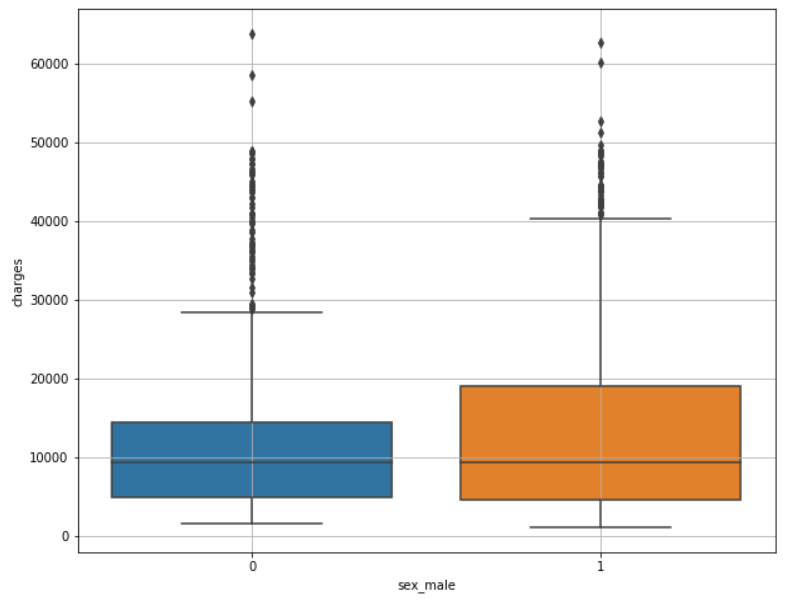
1. **Charges for smokers & charges for non-smokers:**



Following can be concluded from the graphs above:

* Distribution of Charges for Smokers shows that smokers face charges higher than the non-smokers and the charges for most smokers are between 15000 to 50000.
* Distribution of Charges for Non-Smokers shows that the target variable is skewed to the right i.e. maximum number of non-smokers are charged little for their insurance. The graph also shows that there are outliers in the data.

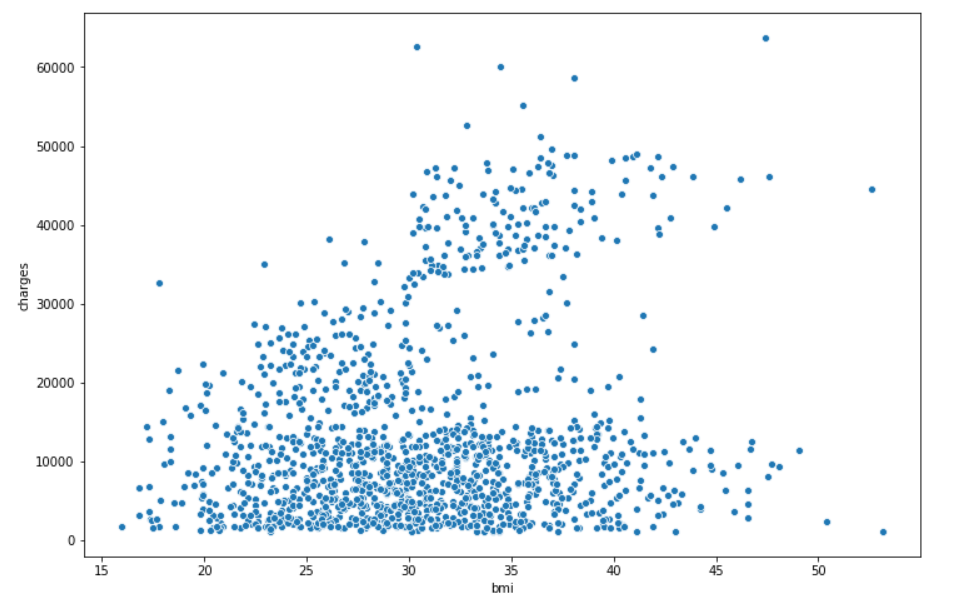
1. **Charges for males and Females:**



Following can be concluded from the graph above:

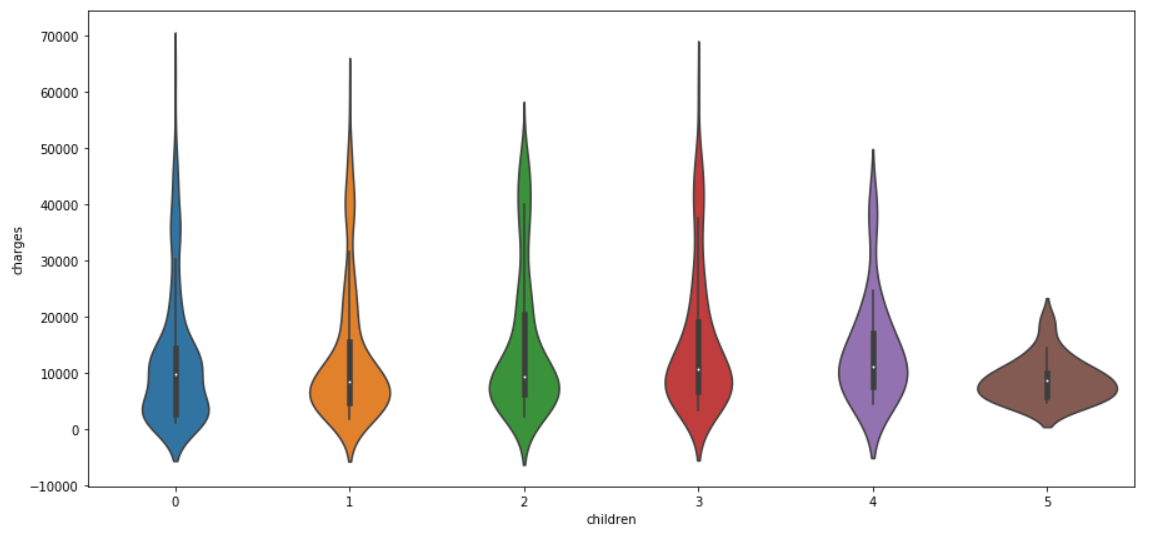
* There is not much difference between the charges for men and women.
* Presence of a lot of outliers is also observed specially in the Female part.

1. **Affect of BMI on charges:**

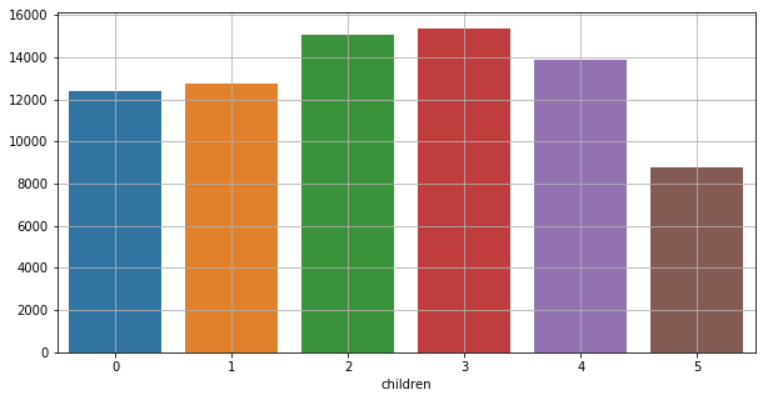


BMI above 30 is considered obese and we can see that as the BMI increases above 30 the charge rates shoot up.

1. **Number children and charges:**

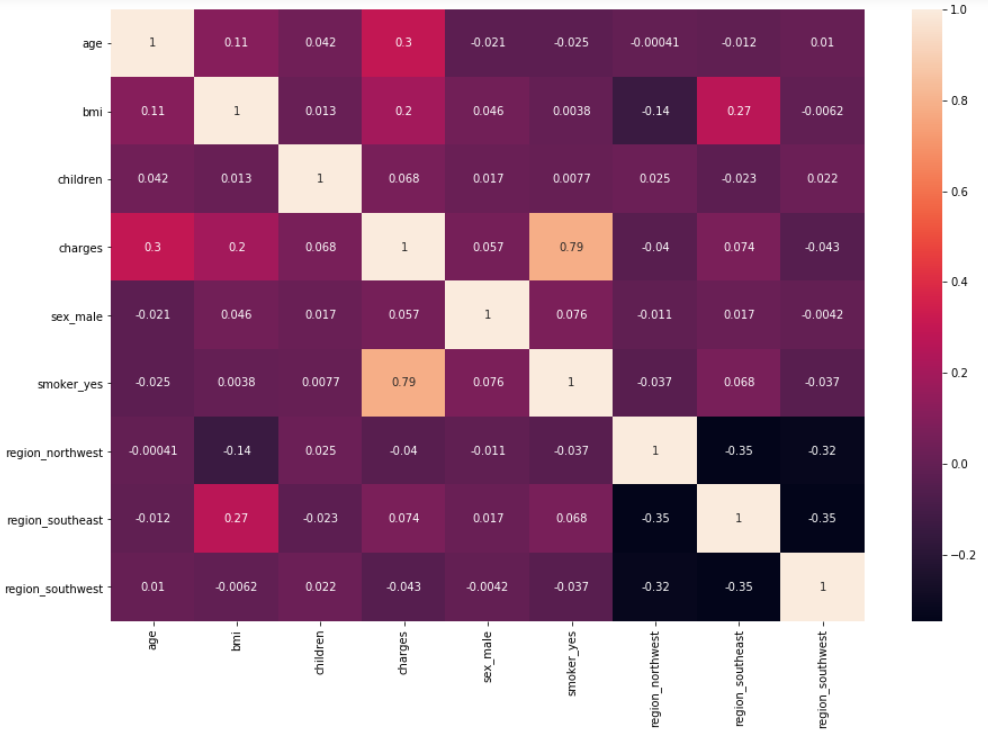


The violin plot shows that people with 5 children have the lowest charges, but it is difficult to say anything about the highest charges. Hence, a bar chart is plotted to get more insights.



The bar chart clearly shows that people with 3 children have the highest charges. It also validates the result of the violin plot saying that people with 5 children have the lowest charges.

1. **Correlation Matrix:**



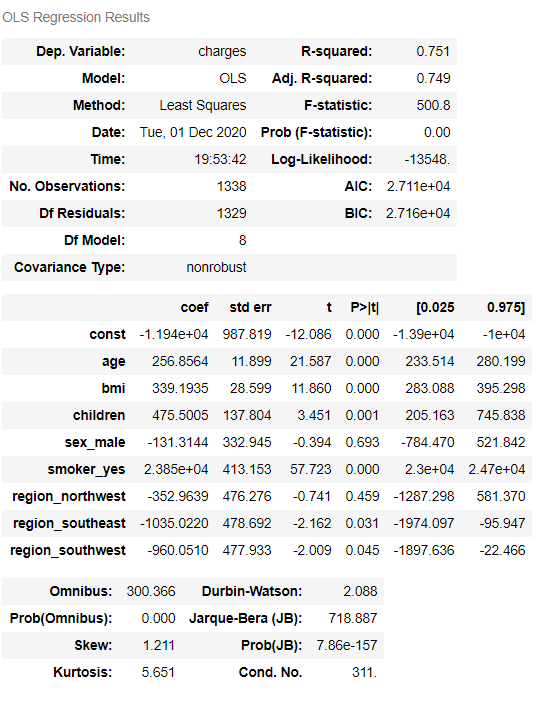
The highest correlation with Charges is with Smokers and the lowest correlation is with different regions.

# PRE-PROCESSING THE DATA

The categorical features were converted into dummy features as the first step towards pre-processing.

The next step would be towards defining the features as independent and dependent in the form of X and y respectively, and finally scale the independent features X.

# STATISTICAL ANALYSIS



Summary from the Statistical Analysis:

1. The p-value for the sex region is 0.692 and hence it is irrelevant. We have also deduced the same from the visualization that the charges were not biased to any gender.
2. The charges are very much dependent on smoking and hence the p-value for that region is 0.00.
3. The p-value for the children is also 0 hence it can be said confidently that the charges vary with the number of children a person has. Same is true for age and BMI.
4. Coming to the region column as the p-value is large it can be concluded that the variation in charges is not dependent on the region.
5. All the visualization conclusions align with the statistical summary, but it should be kept in mind that the visualizations may be confusing at times and may not always tell what really is going on. Hence, statistical analysis should always be done to confirm the findings.

# FEATURE ENGINEERING & ITS BENEFITS

Feature engineering refers to a process of selecting and transforming variables when creating a predictive model using machine learning or statistical modeling (such as deep learning, decision trees, or regression). The process involves a combination of data analysis, applying rules of thumb, and judgement. It is sometimes referred to as pre-processing. The data used to create a predictive model consists of an outcome variable, which contains data that needs to be predicted, and a series of predictor variables that contain data believed to be predictive of the outcome variable. A "feature" in the context of predictive modeling is just another name for a predictor variable. Feature engineering is the general term for creating and manipulating predictors so that a good predictive model can be created.

Most of the time the data is in the form of a table and each column in the table is called its feature. These features of the raw dataset may not produce the best results from the algorithm. Modifying, combining, and deleting these features results in a set that can produce better training for the algorithm. Without feature engineering, the accuracy of the machine learning algorithm reduces significantly. An algorithm that is fed the raw data is unaware of the importance of features. It makes predictions in the dark. Feature engineering is the guiding light in this scenario. The complexity of the algorithm reduces when there are relevant features. The results will be accurate even when the algorithm used is not fit for the situation. Simpler models are often easier to understand, code, and maintain. Feature engineering is thus a guide to your algorithms.

# TESTING PROCESS

As we are using regression models therefore, we are going to use R², RMSE and CV Score to test the accuracy and the effectiveness of the different models and find out the best one.

{\displaystyle R^2 = \frac {\text{Variance explained by the model}}{\text{Total variance}}}

R-squared is always between 0 and 100%

* 0% represents a model that does not explain any of the variation in the [response](https://statisticsbyjim.com/glossary/response-variables/) variable around its [mean](https://statisticsbyjim.com/glossary/mean/). The mean of the dependent variable predicts the dependent variable as well as the regression model.
* 100% represents a model that explains all of the variation in the response variable around its mean.

Usually, larger the R2, the better the regression model fits your observations.

[rmse](https://www.statisticshowto.com/wp-content/uploads/2016/10/rmse.png)

**Where**:

* f = forecasts (expected values or unknown results),
* o = observed values (known results).

**Cross Validation**

Cross-validation is a useful tool when the size of the data set is limited. In a perfect world, our data sets would be large enough that we could set aside a sizable portion of the data set to validate (i.e., examine the resulting prediction error) the model we run on the majority of the data set. Unfortunately, this type of data is not always available, especially in social science research.

To combat the issue of limited data, while still being able to assess the fit of the model, we use *cross-validation*. Essentially, cross-validation iteratively splits the data set into two portions: a test and a training set. The prediction errors from each of the test sets are then averaged to determine the expected prediction error for the whole model.

# ACCEPTANCE

I Anish Arora, a student at Durham College, consent to and accept the terms stated in this Statement of Work by initialing and signing each page below.

Date: 1st Dec 2020  
Signature: Anish Arora