Anish Arora

100803162

Medical Insurance Cost Prediction

Statement of Work - V1

# CONTENTS

[CONTENTS 1](#_Toc55581187)

[ABSTRACT 2](#_Toc55581188)

[INTRODUCTION 2](#_Toc55581189)

[SCOPE OF WORK 2](#_Toc55581190)

[DELIVERABLES 2](#_Toc55581191)

[MILESTONES 2](#_Toc55581192)

[DATASET INFORMATION 3](#_Toc55581193)

[DATA ASSUMPTIONS 3](#_Toc55581194)

[DATA LIMITATIONS 3](#_Toc55581195)

[TESTING PROCESS 3](#_Toc55581196)

[ACCEPTANCE 4](#_Toc55581197)

# 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 | 23-Nov-2020 |
| Modelling | 23-Nov-2020 |
| Prototyping | 23-Nov-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 there were encoded in numbers using encoding technique.

# 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: 6 Nov 2020  
Signature: Anish Arora