Georges ESSOH

DSC 680: Milestone 3: White Paper Draft

Date :04/10/2022

**Accurately predicting medical costs based on personal data**

**Business Problem**

Health insurance companies make profits by collecting more money than needed to cover

their beneficiaries' medical treatment. Medical costs are challenging to forecast and

estimate because of the variety of factors influencing them. This project aims to

accurately anticipate or predict insurance prices based on factors like people's data, such

as age, BMI, whether they smoke, etc. Figure out the main factors that influence

insurance costs is also an objective of this project. Such estimations could be used to

develop computer models that adjust the price of annual premiums based on predicted

treatment costs.

**Background/History**

National Healthcare Expenses (NHE) in the United States increased 4.3% in 2016 to the US $ 3.3 trillion (US $ 10,348 per person). This represents 17.9% of the US gross domestic product (GDP) [1]. When trying to control and reduce these unsustainable increases in health care costs, healthcare institutions must be able to predict an individual's expected future costs, which can benefit a variety of stakeholders. There is sex. For health insurers and increasingly healthcare systems, accurate predictions of expected expenses can help in business planning in various ways.

**Data Explanation**

The dataset we chose was obtained on Kaggle(<https://www.kaggle.com/mirichoi0218/insurance>) to address this problem. The dataset contains seven (07) features and about 1338 instances or rows, and here is the definition of each feature.

1- **Age**: age of the primary beneficiary

2- **Sex:** insurance contractor gender, female, male

3- **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/m²) using the

ratio of height to weight, ideally 18.5 to 24.9

4- **Children:** number of children covered by health insurance, number of dependents

5- **Smoker:** smoking or not

6- **Region:** the beneficiary's residential area in the US, northeast, southeast, southwest,

northwest.

7- **Charges:** individual medical costs billed by health insurance. Charges is our target feature.

**Methods**

We have identified three main steps necessary before generating our final predictive model, namely Exploration Data Analysis, Feature / variable selection and finally the consolidation of the two previous steps.

• **Step I:** Here we will be conducting an exploratory data analysis (EDA) on our dataset. The dataset contains seven (07) different variables that could contribute to our predictive model. At this level, gaining insights into the data is necessary. At this level, it is crucial to identify and deal with missing values and outliers. Visualization will be handy insofar as it will allow us to understand how each variable is distributed. The type of distribution predetermines the type of model. Since most models assume that the data distribution is normal if otherwise, the distributions should be normalized. At this level, it would also be necessary to proceed to the encoding of certain categorical variables for the sake of the next step, which is the selection of variables, because most algorithms take continuous values as input. Renaming the columns for the sake of simplicity is another thing that will be done.

• **Step II:** At this stage, the focus will be on feature/variable selection from the 15 variables within the dataset after encoding. Several methods such as filtering by variance, Feature selection by correlation, and Feature Selection Using a Wrapper will be tested. The advantage of some of these methods is that they generate top feature data frames. The overall feature score was then determined that provided the final feature rankings.

• **Step III:** At this stage, the results of Step I and Step II are combined. The final features that will be used in our initial predictive models build. The models will then be run using cross validation, the summary of results generated, and discussions will follow.

**Analysis**

* **Exploratory Data Analysis:**

The first graph of our appendix (appendix A) presents pair plots (scatter plots) of all numerical variables of our data set and histograms showing the data distribution. The idea is to know the data distribution for each of these variables. So, we can see that "charges" our target variable is right-skewed. At the same time, none of the scatterplots show any clear correlation between variables or very little correlation with the target variable "charges.

The weak correlation between the variables and between the variables and the target variable 'charges' is confirmed by the heat map below.

Chart, treemap chart

Description automatically generated

**Bivariate analysis**

Nevertheless, age and BMI seem to be positively correlated with the target variable 'charges’ (Appendix B).Our Exploratory Data Analysis reveals some interesting facts that medical costs seem to increase with the number of children, as shown in the chart (Appendix C). Medical costs for men seem higher than those for women (Appendix D). Finally, the exploratory analysis of the data reveals that smoking status significantly increases medical costs regardless of the region of the United States (Appendix E).

To complete our analysis, let's look at the distribution of our charges. As the diagram below indicates, the load distribution is right-skewed, implying mean > median > mode. Here the distribution tells that most people have medical expenses near 13K dollars/year, and then the number of people having higher medical costs decreases with increased medical costs.

Chart, histogram

Description automatically generated

* **Data preprocessing:**

The dataset has three categorical columns: Gender, Smoker, and Region. Gender consists of men and women, and data analysis showed that men had higher medical costs than women. For this reason, we have created a new column for men and a new column for women. The smoker column consists of smokers and non-smokers, and smokers have been observed to have higher medical costs than non-smokers, so smokers are in the new column and non-smokers are new. I coded it in a column. The Region column has four segments: Southeast, Southwest, Northeast, and Northwest. The data analysis found that the southeastern region was the most spent, followed by the northeastern, northwestern, and southwestern regions, so we coded each of these regions as a new column, as shown in the graph below**.**

**Table

Description automatically generated with medium confidence**

* **Feature selection:**

Several methods such as filtering by variance, Feature selection by correlation, and Feature Selection Using a Wrapper were tested. At the end of the feature selection process, these features are retained for model buildings***: age, BMI, smoker, and non-smoker.***

* **Validation process:**

Before modelling, it is important to standardize the results validation process to ensure that it is the same for each of the models tested and that the same data must be used in each model. For this, we have chosen cross-validation. The goal of cross-validation is to test the Model's ability to predict new data that was not used in estimating it. It has a single parameter called 'k,' which indicates the number of groups that the data would be split into. Here we train the Model on k-1 datasets and test on the kth dataset, and this process repeats till the k value we set, and in our case, we selected the k value as 10

* **Model Building:**

The 'charges' value to be predicted being continuous, we are in a classic case of regression. We have chosen three powerful models: Linear Regression, KNeighbors regression, and XGboost regression. The results are compiled in the table below.

|  |  |  |
| --- | --- | --- |
| Models | R squared | RSME |
| Linear regression | 0.74 | 6121.29 |
| KNeighbors regression | 0.84 | 4809.72 |
| XGboost regression | 0.86 | 4553.69 |

**Conclusion**

The analysis we have just completed suggests that several factors can affect the medical costs of a population, from apparent factors like smoking and age status to less apparent factors like region of residence. That men have higher medical cost than women in average. Smokers have higher medical cost than non-smoker and as the number of children increases so the medical costs. Despite the great results, there are several ways this model can be improved for future use because nothing guarantees that it will perform if moved to a production environment. Scalability testing would be required, so see if the model can accurately predict medical expenses from new data. Automatic processes can be put into place to blend the data without manual work. This includes data collection as well, by using methods such as web scraping to automatically acquire the data, and automatic correlation testing to determine the best variables from new data as well.

***XGboost regression*** is the model with the best results after tuning and the one that will be moved to production**.**

**Assumptions**

Part of the assumptions made is to ensure the integrity of the information collected. We understand the sensitivity of collecting such data, but we must ensure the quality of the data collection process.

**Limitations**

One of the main limitations we face is the size of our dataset. Indeed, the more data we have, the better it is for our models. It is important to considerably increase the size of our databases by collecting as much data as possible.

**Challenges/Issues**

This is a regression problem, and one obvious issue is how to deal with wildly different results from using different approaches for features selection. I will need to develop a plan to deal with such irregularities that might lead me away from the focus of this project. Another issue would be the results that are based on gender and smoking status, which are two categorical features, and adequately encoding the categorical features is key to success.

**Recommendations**

Despite the great results, there are several ways this model can be improved for future use because nothing guarantees that it will perform if moved to a production environment. Scalability testing would be required, so see if the model can accurately predict medical expenses from new data. Automatic processes can be put into place to blend the data without manual work. This includes data collection as well, by using methods such as web scraping to automatically acquire the data, and automatic correlation testing to determine the best variables from new data as well.

**Implementation Plan**

The best and fastest implementation strategy is to implement an application that will generate web pages for our associates and brokers. This application will allow direct entry of personal data and create an estimate of the person's medical expenses and health insurance premium. This will considerably speed up the quotation process and reduce the margin of error. The app could b implemented via the Amazon AWS platform, and this would reduce our operating expenses if we set up everything in-house.

**Ethical considerations**

It is always very sensitive to work with personal data of a medical type due to obvious confidentiality reasons, and such data should be treated with caution. It should be noted that the goal is to limit human intervention during the process of determining a client's premium for 1) celerity and 2) to avoid bias.

**Appendix**

Appendix A

**A picture containing chart

Description automatically generated**

Appendix B

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Appendix C

Chart, box and whisker chart

Description automatically generated

Appendix D

Chart, box and whisker chart

Description automatically generated

Appendix E

Chart, box and whisker chart

Description automatically generated