Data Analysis Project

DA-6833

Rinu Thomas

2024-06-12

# 1. Summary/Abstract

The primary goal of this research project is to develop a predictive model to estimate medical insurance costs for individuals based on their health-related and personal attributes. This model will help in understanding the key factors influencing insurance charges and assist insurance companies in pricing policies more accurately.

# 2. Introduction

## 2.1 General Background Information

### 2.1.1 Importance of Predictive Modeling in Insurance

Predictive modeling plays a vital role in the insurance industry. Accurate predictions of insurance costs enable companies to:

* Assess risk more effectively.
* Set premiums that reflect the actual risk associated with individual policyholders.
* Enhance customer satisfaction by offering personalized and fair pricing.

On the consumer side, individuals can benefit from understanding the factors that influence their insurance costs, allowing them to make informed choices about their health and financial planning.

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

The dataset used for this analysis is sourced from Kaggle. The dataset comprises 2.7K rows and 7 columns and contains the following features:

* **Age**: Age of the individual.
* **Sex**: Gender of the individual (male/female).
* **BMI (Body Mass Index)**: A measure of body fat based on height and weight.
* **Children**: Number of children/dependents covered by the insurance plan.
* **Smoker**: Smoking status of the individual (yes/no).
* **Region**: Geographical region in the US where the individual resides (northeast, southeast, southwest, northwest).
* **Charges**: Medical insurance costs billed by the insurance company (target variable).

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

The primary question we aim to answer using the “Medical Insurance Cost Prediction” dataset is:

**“What are the key factors influencing medical insurance costs, and how accurately can we predict these costs based on individuals’ demographic and health-related attributes?”**

### 2.3.1 Outcomes of Interest

The main outcome of interest in this analysis is:

* **Medical Insurance Costs (charges)**: The amount billed by the insurance company for an individual’s medical insurance.

### 2.3.2 Specific Predictors

The specific predictors (independent variables) we will focus on include:

* **Age**: How age influences insurance costs.
* **Sex**: The impact of gender on insurance costs.
* **BMI (Body Mass Index)**: The relationship between BMI and insurance costs.
* **Children**: How the number of dependents affects insurance costs.
* **Smoker**: The effect of smoking status on insurance costs.
* **Region**: Regional differences in insurance costs within the US.

### 2.3.3 Relations/Patterns of Interest

I am looking for several key relationships and patterns in the data, including:

1. **Age and Insurance Costs**:Investigating whether older individuals tend to have higher medical insurance costs.
2. **Gender Differences**:Analyzing whether there is a significant difference in insurance costs between males and females.
3. **BMI and Insurance Costs**:Exploring the relationship between an individual’s BMI and their medical insurance costs.Identifying any nonlinear patterns where very high or very low BMI values might correlate with higher insurance costs.
4. **Impact of Smoking**:Assessing how smoking status (yes/no) affects insurance costs.Quantifying the cost difference between smokers and non-smokers.
5. **Number of Children**: Examining whether having more dependents (children) increases insurance costs.
6. **Regional Variations**: Identifying differences in insurance costs across different regions in the US (northeast, southeast, southwest, northwest).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

The methods used to clean, process and analyze the are as follows:

* **Data Cleaning and Preprocessing:**
  + Load the data
  + Convert categorical variable
  + Identify missing values
  + Find Outliers
* **Exploratory Data Analysis:**
  + Summarize the data
  + Visualize the data
* **Model building and Analysis Approach:**
  + Train-Test split the data
  + Linear Regression Model
  + Random Forest Model
  + Model Evaluation

## 3.1 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

The dataset used in this analysis is sourced from Kaggle. The dataset was downloaded directly from Kaggle website. It can be accessed at the following URL: [Medical Insurance Cost Prediction](https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction).

The methods used to clean, process and analyze the are as follows:

* **Data Cleaning and Preprocessing:**
  + Load the data (can be found in processingcode2.R)
  + Convert categorical variable (can be found in processingcode2.R)
  + Identify missing values (can be found in processingcode2.R)
  + Find Outlier (can be found in eda2.R)
* **Exploratory Data Analysis:**
  + Summarize the data (can be found in eda2.R)
  + Visualize the data (can be found in eda2.R)
* **Model building and Analysis Approach: *(Have not started yet)***
  + Train-Test split the data
  + Linear Regression Model
  + Random Forest Model
  + Model Evaluation

## 3.2 Statistical analysis

The data will be built and analyzed as follows:

* **Model building and Analysis Approach: *(Have not started yet)***
  + Train-Test split the data
  + Linear Regression Model
  + Random Forest Model
  + Model Evaluation

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   |  | age | sex | bmi | children | smoker | region | charges | | --- | --- | --- | --- | --- | --- | --- | --- | |  | Min. :18.00 | female:1366 | Min. :15.96 | Min. :0.000 | no :2208 | northeast:658 | Min. : 1122 | |  | 1st Qu.:26.00 | male :1406 | 1st Qu.:26.22 | 1st Qu.:0.000 | yes: 564 | northwest:664 | 1st Qu.: 4688 | |  | Median :39.00 | NA | Median :30.45 | Median :1.000 | NA | southeast:766 | Median : 9333 | |  | Mean :39.11 | NA | Mean :30.70 | Mean :1.102 | NA | southwest:684 | Mean :13261 | |  | 3rd Qu.:51.00 | NA | 3rd Qu.:34.77 | 3rd Qu.:2.000 | NA | NA | 3rd Qu.:16578 | |  | Max. :64.00 | NA | Max. :53.13 | Max. :5.000 | NA | NA | Max. :63770 | |

[Figure 1](#fig-outliers-bmi) shows the outliers in BMI

|  |
| --- |
| Figure 1: BMI Outliers |

[Figure 2](#fig-outliers-charges) shows the outliers in charges

|  |
| --- |
| Figure 2: Charges Outliers |

[Figure 3](#fig-charges) shows histogram of insurance charges

|  |
| --- |
| Figure 3: Distribution of Insurance charges |

[Figure 4](#fig-smoking) shows box plot of insurance charges by smoking status

|  |
| --- |
| Figure 4: Insurance Charges by smoking status. |

[Figure 5](#fig-region) shows box plot of insurance charges by region

|  |
| --- |
| Figure 5: Insurance Charges by Region. |

[Figure 6](#fig-age) shows scatter plot of insurance charges by age

|  |
| --- |
| Figure 6: Insurance Charges by Age. |

[Figure 7](#fig-bmi) shows scatter plot of insurance by BMI

|  |
| --- |
| Figure 7: Insurance Charges by BMI. |

**?@fig-children** shows bar graph of average insurance charges by number of children

|  |
| --- |
| Figure 8: Average Insurance Charges by Number of Children. |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>