Data Analysis Project

DA-6833

Rinu Thomas

2024-07-16

# 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

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

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

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

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

[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 |

Observation: There are several outliers with BMI values above 50, indicating that there are individuals with unusually high BMI compared to the rest of the dataset.@fig-outliers-charges shows the outliers in charges

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

Observation: There are numerous outliers with insurance charges above 40,000, indicating that there are individuals with unusually high insurance charges compared to the rest of the dataset.

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

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

Observation: The histogram shows a right-skewed distribution, indicating that most individuals have lower insurance charges, with fewer individuals having very high charges.

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

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

Observation: There is a clear and substantial difference in insurance charges between smokers and non-smokers. Smokers have much higher charges on average. Charges for smokers are more variable than for non-smokers, with a wider range of values. High charges are more common among smokers and are not considered outliers, whereas for non-smokers, such high charges are less common and considered outliers.

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

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

Observation: The Southeast region has the highest median insurance charges, followed by the Northeast, Northwest, and Southwest. The Southwest region has the lowest median charges. All regions have numerous outliers with very high charges, particularly the Southeast and Northeast.

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

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

Observation: There is a clear upward trend in insurance charges as age increases. This suggests that insurance charges tend to increase with age. The relationship appears to be somewhat linear, with higher age associated with higher charges, but there is a noticeable spread in the data.

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

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

Observation: There is a noticeable upward trend in insurance charges as BMI increases. This suggests that individuals with higher BMI tend to have higher insurance charges. The relationship appears to be somewhat nonlinear, with higher charges associated with higher BMI values, but there is considerable spread in the data.

[Figure 8](#fig-children) shows bar graph of average insurance charges by number of children

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

Observation: There is a general trend where insurance charges increase with the number of children up to three children, after which there is a slight decrease for four children, and a significant decrease for five children. The drop in average charges for households with five children could be due to various reasons such as discounts for larger families, different insurance plans, or data anomalies.

## 4.2 Basic statistical analysis

[Table 2](#tbl-resulttable1) shows a simple model with one predictor using linear regression analysis.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Simple Models with One Predictor.   | predictor | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | --- | | age | (Intercept) | 3182.7705 | 650.44959 | 4.8931854 | 0.0000010 | | age | age | 257.7010 | 15.64837 | 16.4682290 | 0.0000000 | | bmi | (Intercept) | 1097.5183 | 1155.49870 | 0.9498222 | 0.3422855 | | bmi | bmi | 396.1993 | 36.90861 | 10.7346011 | 0.0000000 | | children | (Intercept) | 12529.1303 | 310.97832 | 40.2894006 | 0.0000000 | | children | children | 664.6262 | 189.64091 | 3.5046560 | 0.0004644 | | sex | (Intercept) | 12486.8320 | 328.19612 | 38.0468610 | 0.0000000 | | sex | sexmale | 1527.0407 | 460.82644 | 3.3137004 | 0.0009325 | | smoker | (Intercept) | 8417.8744 | 158.98688 | 52.9469743 | 0.0000000 | | smoker | smokeryes | 23805.2654 | 352.46691 | 67.5390070 | 0.0000000 | | region | (Intercept) | 13475.8747 | 472.25777 | 28.5349983 | 0.0000000 | | region | regionnorthwest | -1012.7454 | 666.36289 | -1.5198107 | 0.1286728 | | region | regionsoutheast | 1272.9030 | 643.90193 | 1.9768585 | 0.0481567 | | region | regionsouthwest | -1311.6783 | 661.49615 | -1.9828964 | 0.0474777 | |

Observation:

* Age, BMI, number of children, sex, smoking status, and living in the Southeast and Southwest regions are significant predictors of insurance charges.
* Smoking status has the largest effect on insurance charges, significantly increasing them.
* The intercept for BMI and living in the Northwest region are not significant predictors of insurance charges in these models.

## 4.3 Full analysis

[Table 3](#tbl-resulttable2) shows linear regression analysis charges is analyzed across all predictors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Linear Regression Analysis.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | -1.184037e+04 | 757.390243 | -15.6331202 | 0.0000000 | | age | 2.526119e+02 | 9.072504 | 27.8436841 | 0.0000000 | | bmi | 3.364993e+02 | 21.859169 | 15.3939664 | 0.0000000 | | children | 4.550461e+02 | 105.022395 | 4.3328479 | 0.0000154 | | sexmale | 8.252665e-01 | 255.209406 | 0.0032337 | 0.9974202 | | smokeryes | 2.415304e+04 | 317.486710 | 76.0757369 | 0.0000000 | | regionnorthwest | -2.197996e+02 | 368.749286 | -0.5960679 | 0.5511909 | | regionsoutheast | -1.020635e+03 | 369.922657 | -2.7590496 | 0.0058447 | | regionsouthwest | -7.674076e+02 | 366.387191 | -2.0945263 | 0.0363267 | |

Observation: After the data was split into training and testing, the linear regression analysis was conducted.

* Age, BMI, number of children, smoking status, and living in the Southeast or Southwest regions are significant predictors of insurance charges.
* Smoking status still has the largest effect on insurance charges, significantly increasing them.
* Sex and living in the Northwest region are not significant predictors of insurance charges in this model.

[Figure 9](#fig-linear) helps visualize which features have a significant effect on charges.

|  |
| --- |
| Figure 9: Linear Regression Coefficients. |

[Table 4](#tbl-resulttable3) helps to capture complex interaction between features.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Random Forest Analysis.   |  | Feature | Importance | | --- | --- | --- | | age | age | 41276698254 | | sex | sex | 2333500565 | | bmi | bmi | 45657041834 | | children | children | 6364617046 | | smoker | smoker | 205835757547 | | region | region | 6128386015 | |

Observation:

* Smoking status, BMI, and age are the most important predictors of insurance charges, followed by the number of children and region. Sex has the least importance.
* The importance rankings from the random forest analysis are consistent with the findings from the linear regression models, where smoking status, age, and BMI were significant predictors.

[Figure 10](#fig-randomforest) helps visualize how each feature have a significant effect on charges.

|  |
| --- |
| Figure 10: Random Forest Feature Importance. |

# 5. Discussion

## 5.1 Summary and Interpretation

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

### 5.1.1 What was done:

* **Data Loading and Preparation**:
  + Loaded the insurance dataset from a CSV file.
  + Performed exploratory data analysis to understand the distribution and relationships within the data.
  + Conducted data cleaning processes, such as handling missing values and transforming variables.
* **Exploratory Data Analysis (EDA)**:
  + Visualized the distribution of insurance charges using histograms and box plots.
  + Examined the relationships between insurance charges and various predictors, including age, BMI, number of children, smoking status, and region.
* **Statistical Modeling**:
  + Built simple linear regression models to explore the impact of individual predictors on insurance charges.
  + Developed a multiple linear regression model to understand the combined effect of multiple predictors on insurance charges.
  + Conducted a random forest analysis to determine the importance of different features in predicting insurance charges.
* **Visualization**:
  + Created various plots, including scatter plots, bar charts, and box plots, to visualize the relationships between predictors and insurance charges.

### 5.1.2 What was found:

* **Key Predictors**:
  + **Smoking Status**: The most influential predictor, with smokers incurring significantly higher insurance charges compared to non-smokers.
  + **BMI**: A strong positive relationship with insurance charges, indicating higher charges for individuals with higher BMI.
  + **Age**: A positive relationship with insurance charges, with charges increasing as age increases.
  + **Number of Children**: A significant positive impact on insurance charges, although the effect was smaller compared to smoking status, BMI, and age.
  + **Region**: Regional differences were observed, with the Southeast region associated with higher charges and the Southwest region associated with lower charges.
  + **Sex**: The impact of sex on insurance charges was negligible.
* **Feature Importance**:
  + The random forest analysis confirmed that smoking status, BMI, and age were the most important features in predicting insurance charges.
  + The number of children and region also contributed to the predictions, while sex had the least importance.

### 5.1.3 What it means:

* **Policy and Pricing**:
  + The findings highlight the importance of considering multiple factors when setting insurance premiums. Smoking status, BMI, and age are critical factors that significantly impact insurance costs.
  + Insurance companies can use these insights to adjust premiums more accurately based on the risk profile of individuals.
* **Health Interventions**:
  + The strong association between smoking and higher insurance charges underscores the need for targeted health interventions to reduce smoking rates, which could, in turn, lower insurance costs.
  + Addressing obesity through health programs could also help reduce insurance charges associated with higher BMI.
* **Regional Considerations**:
  + Regional differences in insurance charges suggest that localized strategies might be necessary to address specific regional health challenges and cost structures.
* **Further Research**:
  + Future studies could explore the interactions between these key predictors to understand their combined effects on insurance charges.
  + Additional data could be collected to validate these findings and explore other potential predictors.

## 5.2 Strengths and Limitations

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

### 5.2.1 Strengths:

1. **Comprehensive Exploratory Data Analysis (EDA)**:
   * Detailed EDA helped in understanding the distribution and relationships within the data.
   * Visualizations such as histograms, box plots, scatter plots, and bar charts provided clear insights into the data and highlighted important patterns.
2. **Use of Multiple Statistical Methods**:
   * Employed both linear regression and random forest models to analyze the data.
   * Linear regression provided insights into the individual and combined effects of predictors.
   * Random forest analysis helped in identifying the importance of different features.
3. **Handling of Categorical Variables**:
   * Effectively transformed and factored categorical variables to ensure accurate modeling.
   * Considered regional differences and the impact of smoking status, which are critical factors in insurance pricing.
4. **Feature Importance Analysis**:
   * The random forest feature importance analysis provided a clear ranking of predictors, confirming the results of the linear regression models.
5. **Clear and Informative Visualizations**:
   * Created visualizations that were easy to interpret and clearly illustrated the relationships between variables and insurance charges.
6. **Thorough Interpretation**:
   * Provided detailed interpretations of the results, explaining the significance and impact of each predictor on insurance charges.

### 5.2.2 Limitations:

* **Assumptions of Linear Regression**:
  + Linear regression models assume linear relationships between predictors and the outcome, which may not fully capture more complex, non-linear relationships.
  + Assumes independence of observations, which might not hold true in real-world data.
* **Potential Overfitting with Random Forest**:
  + Random forest models can be prone to overfitting, especially with a small dataset or highly correlated features.
  + Overfitting might lead to overly optimistic feature importance scores that do not generalize well to new data.
* **Outliers and Data Distribution**:
  + The dataset contains outliers, particularly in insurance charges, which can impact the results of linear regression models.
  + The skewed distribution of insurance charges might affect the assumptions of normality in linear regression.
* **Limited Dataset**:
  + The analysis was conducted on a specific insurance dataset, which may limit the generalizability of the findings to other datasets or populations.
  + Potential biases in the dataset could affect the results and their applicability.

## 5.3 Conclusions

*What are the main take-home messages?*

The analysis highlights the significant factors affecting insurance charges and provides valuable insights for insurance pricing and policy formulation. By focusing on key predictors such as smoking status, BMI, age, number of children, and region, and considering their combined effects, insurance companies can develop more accurate and equitable pricing strategies. Additionally, public health interventions targeting smoking and obesity can help reduce overall insurance costs. Continued research and validation are essential to further refine these insights and ensure their applicability across diverse populations and datasets.

*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>