**Predictive and Decision-Making Analysis Report On Health Insurance**

University Canada West

BUSI 650 (Section- 20)- Business Analytics

Prepared by: Aditya Arte

Student ID: 2305505

Date: February 25, 2024

**Table of Contents**

[**Introduction 3**](#_bo54zp529gp2)

[**Methodology 5**](#_aqt31tw3ri5v)

[**Results and Discussion 6**](#_fquhuiv69gyb)

[Descriptive Analysis 6](#_enjt0zdr5a29)

[Prediction Model 8](#_ggdlpee3encw)

[**Conclusion 11**](#_80z4n53nihzq)

[**References 12**](#_qtsk7wdrcq6h)

# **Introduction**

Business Analytics (BA) is a methodology that uses data to derive valuable business insights. Business Analytics employs analytical processes such as descriptive and statistical analysis to decipher the data and gain clarity regarding future trends, which can be helpful in decision-making. Business Analytics has been implemented across various fields such as healthcare, medicine, stock market, forecasting, and others for accurately making critical business decisions. Recently, the focus has been shifting from Business Analytics to Prediction Analysis as corporations need to have a firm knowledge of their field of operations. Prediction Analysis focuses primarily on statistical analysis, wherein the data is analyzed using statistical combinations and machine learning to predict the likelihood of a particular event (Moslehpour, 2021).

Health insurance is one field requiring Business Analytics to help them issue insurance to potential customers. The business needs to know how much insurance a specific individual needs and how much premium it should collect.

Health Insurance companies collect customer data when they issue insurance; this data can predict how much insurance the next customer will need and how much premium he should pay. Predicting such things requires a good prediction model to predict such events or numbers accurately. The author has tried to create a predictive model for an insurance company using the dataset provided. He has tried to process the data using descriptive and predictive analysis and attempted to predict the amount of premium that should be collected from the customers.

In the forthcoming part of this paper, the author explains his work while explaining more about prediction and Decision-making.

**Background**

Health insurance companies provide insurance to their customers; this entails that ordinary people approach such companies and pay them a certain amount every month, and if and when a medical emergency arises, the insurance company covers the costs of the treatment. Medical costs have skyrocketed today, and no average individual can pay for a medical emergency without support. Health Insurance provides this financial support in exchange for a monthly premium.

As important as health insurance is to a person, the same way knowing the risk of suffering a loss is to an insurer. Health Insurance companies can not simply offer people insurance at any premium rate. The premium rate should be appropriate with respect to the insured value and the health condition of the insured person. A comparatively healthy person will not require a hefty amount of insurance as their chances of falling ill are lower than an unfit person; hence, the premium rate will also be low. Predictive Analysis plays a crucial role in determining the amount of premium charged as it processes customers' health-related data and predicts how much premium that person should be charged.

# **Methodology**

First, the author checked for any null values in the dataset. No null or inaccurate data was found in the dataset, so he moved to the descriptive analysis part. The author calculated the minimum, maximum, and average values of the given data in the descriptive analysis. The author supported his descriptive analysis with tables and charts. After completing the descriptive analysis, the author worked towards making a predictive model. He started with regression of the data and progressed towards making the most accurate piecewise regression model. In the further part of the paper, the author elaborates on the steps implemented in doing the things mentioned earlier.

# **Results and Discussion**

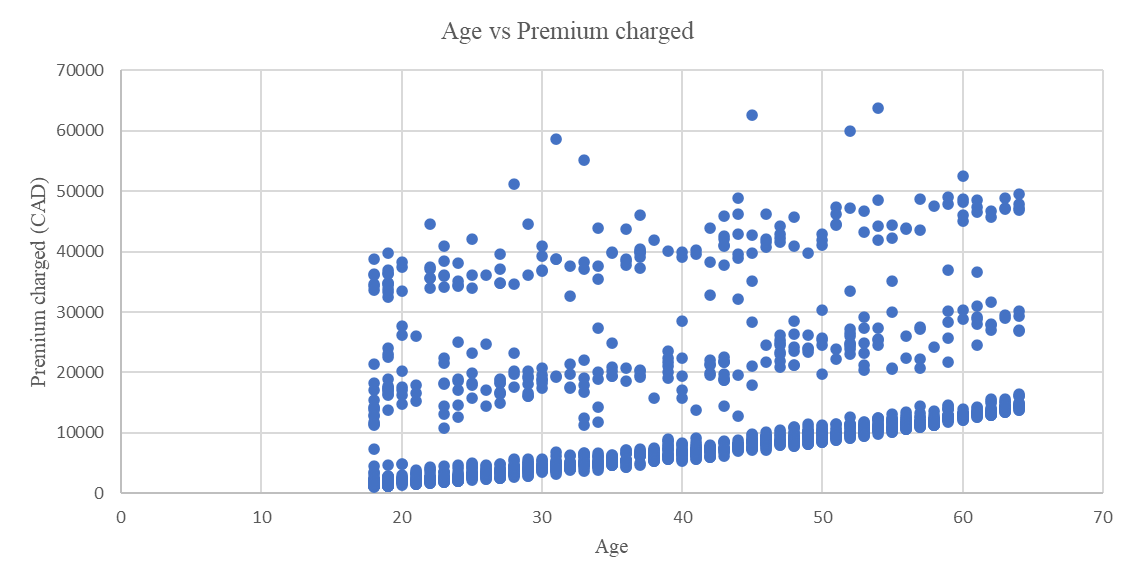
## **Descriptive Analysis**

The author calculated the average age of customers was 39.2 years, while the youngest customers were 18 and the oldest were 64. This indicates that people from different age groups, from early working to retirement, avail health insurance facilities. The average BMI of customers was 30.66, with the lowest BMI of customers at 15.96 and the highest BMI of customers at 53.13. BMI, or Body Mass Index, measures an individual's physical fitness. The lower the BMI, the more fit the person is. On average, customers have a single child. The data on customers' children will be helpful if the insurance also offers coverage for children of customers. The average premium charged by the insurance company was $13270.42, the lowest premium was $1121.87, and the highest was $63770.43. The standard deviation of the premium charged is $12105.48 (Considering the overall data, this standard deviation is relatively less). The level of covariance of premium charged with age is $50836.78, which gives us an idea of how much the charge increases with age. The correlation of premium charged with age is 0.3. The covariance of the premium charged with BMI is $14636.36, and the correlation of the premium charged with BMI is 0.2. The covariance of the premium charged with the number of children is $991.93, and the correlation of the premium charged with the number of children is 0.07. From the previous data, we can observe that the premium rate is highly affected by the age of customers.

To get an easier understanding of this observation of the effect of age on the premium rate, the author has created a scatter plot of the same (see Figure 1).

**Figure 1**

Scatter Plot



*Note.* From Aditya\_Arte\_Predictive\_Analysis Excel file

To provide more insight into the data, the author also created a table and a graph that show the average premium collected from people who belong to different demographics, Table 1 and Figure 2, respectively.

**Table 1**

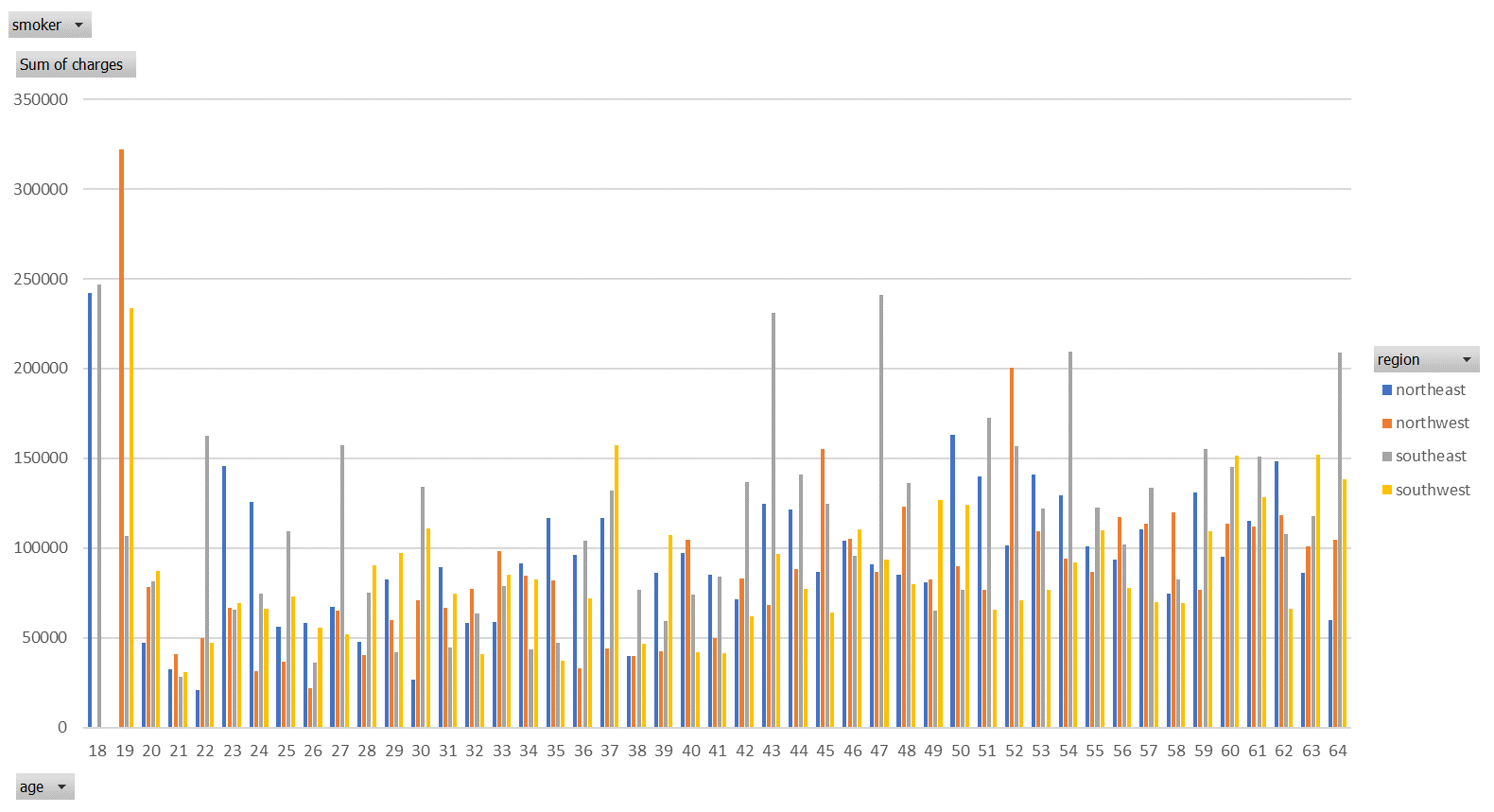
Average Premium Charged

| **Average of charges** | **Column Labels** |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Row Labels** | **northeast** | **northwest** | **southeast** | **southwest** | **Grand Total** |
| **no** | **9165.531672** | **8556.463715** | **8032.216309** | **8019.284513** | **8434.268298** |
| female | 9640.426984 | 8786.998679 | 8440.205552 | 8234.09126 | 8762.2973 |
| male | 8664.042222 | 8320.689321 | 7609.003587 | 7778.905534 | 8087.204731 |
| **yes** | **29673.53647** | **30192.00318** | **34844.99682** | **32269.06349** | **32050.23183** |
| female | 28032.0464 | 29670.82495 | 33034.82072 | 31687.98843 | 30678.99628 |
| male | 30926.25258 | 30713.18142 | 36029.83937 | 32598.86285 | 33042.00598 |
| **Grand Total** | **13406.38452** | **12417.57537** | **14735.41144** | **12346.93738** | **13270.42227** |

*Note.* From Aditya\_Arte\_Predictive\_Analysis Excel file

**Figure 2**

Average Premium Charged



*Note.* From Aditya\_Arte\_Predictive\_Analysis Excel file

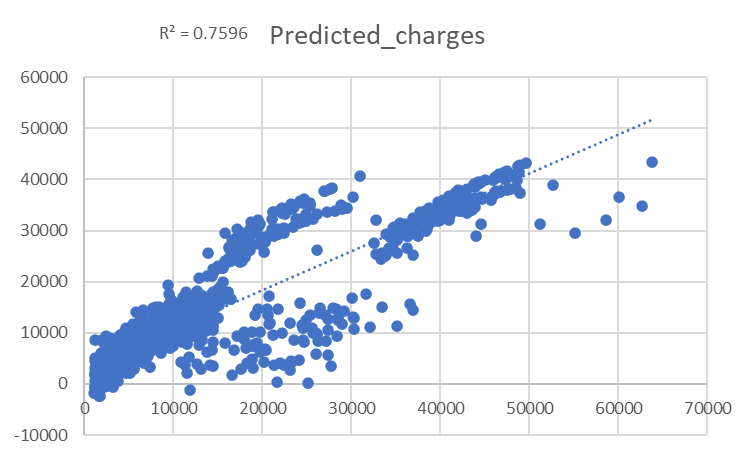
## **Prediction Model**

The author made a prediction model by performing regression on the entire dataset. Regression analysis is a statistical method of analysis that accounts for the relationship between multiple parameters (Schneider, 2010). The author used Excel for his whole work, and as Excel does not accept categorical values while performing regression, the author had to resolve this issue. The author converted the categorical data into numerical data with the help of dummy variables. After the conversion of data, this data was used for regression. The R square value found here was 0.751, meaning the simple linear regression model had 75.1% accuracy.

The author grouped the data region-wise to improve the model's accuracy further. The author felt that by grouping the data region-wise and performing piecewise linear regression on it, the model's accuracy would improve as the trend of the premium rate will be different as per the regions. The R square value achieved after piecewise regression grouping region was 0.7596. This implied that the accuracy of this model was 76%. The author used a scatter plot of the available data and predicted data of premium charges to get this value (see Figure 3).

**Figure 3**

Scatter Plot

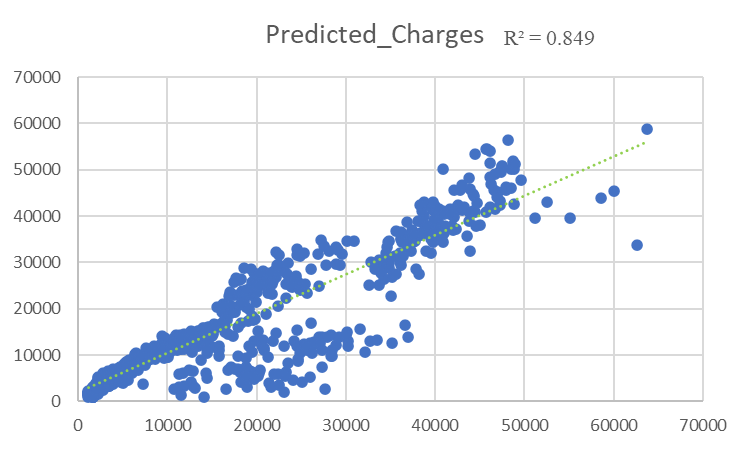


*Note.* From Aditya\_Arte\_Predictive\_Analysis Excel file

After careful thought, the author realized that he could achieve greater accuracy by grouping the data according to distinct combinations of the categorical data. The categorical data included parameters such as gender, smoking habit in yes or no format, and regions. The author developed a piecewise regression model using groups of distinct combinations of the categorical parameters. The author achieved the R square value of 0.849 or an accuracy level of 85%. The scatter plot in Figure 4 below shows the distribution of data.

**Figure 4.**

Scatter Plot of the final model



*Note.* From Aditya\_Arte\_Predictive\_Analysis Excel file

This means that with this prediction model, the premium charges can be accurately predicted for the next customers.

# **Conclusion**

The author created a prediction model that predicted the premium charged with 85% accuracy. The predicted premium charge was found to be different from the available data. This means that the insurance company can now adjust the premium charges and minimize their supposed losses in the future. The author feels that his prediction model will most justify the premium rates as his prediction model considers each factor affecting the premium rate.

Although the author claims to have created a highly accurate prediction model, he also feels that with more parameters and data, he will be able to make an even more precise prediction model with increased accuracy. This high level of prediction model can only be achieved after the insurance company improves its data retrieval process to mine more effective or valuable data.

# **References**

Moslehpour, M. (2021, December 31). *(PDF) Predictive Analytics in Business Analytics: Decision Tree*. ResearchGate. Retrieved February 25, 2024, from https://www.researchgate.net/publication/357447580\_Predictive\_Analytics\_in\_Business\_Analytics\_Decision\_Tree

Schneider, A. (2010, November). *Linear Regression Analysis*. ncbi. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/