Project Final Report: InSURE

Sukriti Khanal, Jaya Kola, Poojitha, Krishnavamsi Sanisetty

1. Project Introduction

1.1 What is health insurance?

Health insurance is a type of insurance that covers medical expenses related to hospitalization costs, the cost of medicines, or doctor consultation fees. The cost of treating the illness can cause severe financial strain on the accumulated savings. Hence, finding an insurance plan compatible with your financial capacity is a must. With continuous rise in medical costs, there comes cases when you might have to go so far as to compromise on your children's education quality or default on your home loan payments if any sudden illness occurs.[1] Hence, the key elements in health insurance charges are advance planning on the insurance and to make sure you are able to understand which factors are being considered in its calculations and how to optimize the charges. By considering various factors, our project focuses on the computation of medical insurance charges.

1.2 Project goal

The dataset being used in our project is a medical insurance dataset gathered from Kaggle. The goal of the project is ***to identify the factors that might be related with predicting the medical insurance charges and assign weightage to these factors to obtain an equation***. The variables in question are age, BMI (Body Mass Index), children, gender, and location. In other words, ***a statistical model is performed using these parameters to predict the insurance cost incurred***. The goal is not limited to predicting these charges but goes as far as ***to compute the reliability of the model and tries to strengthen it*** by taking significant data into account.

1.3 Project Significance

Health insurance provides financial protection in case of serious health problems or accidents.[2] The significance of the project is to establish a better health care system i.e., predicting insurance charges is one of the most appropriate methods.

2. Data Analysis

2.1 About the Dataset

The raw data comprises of medical insurance data. It has 1338 rows and 6 columns with mixed data types containing numerical values: discrete variables, continuous variables, and categorical variables. Because the categorical values were present in the dataset, they are converted to the dummy variable for the analysis. And the final dataset comprises of nine (independent variables) and 1 (dependent variable). The features can be broadly classified into the following categories:

1. Medical Information- This includes age and BMI.
2. Family details- number of children a person has.
3. Personal information- whether the person is a smoker or a non-smoker.
4. Cardinal directions- whether the person is residing in the northeast, northwest, southeast, southwest

region.

2.2 Preliminary Data Processing

After conducting a preliminary investigation into each feature, we found that all the parameters are useful for our data analysis. For the categorical variables such as gender, we have created two dummy variables (female and male) and for the location, we have created four dummy variables (northeast, northwest, southeast and southwest). BMI refers to the Body Mass Index which is derived from the mass and height of the human body. In addition, the data comprises of charges, age, location, smokers, and children. With these, there are a total of 9 independent variables and 1 dependent variable.

The features present in our database are:

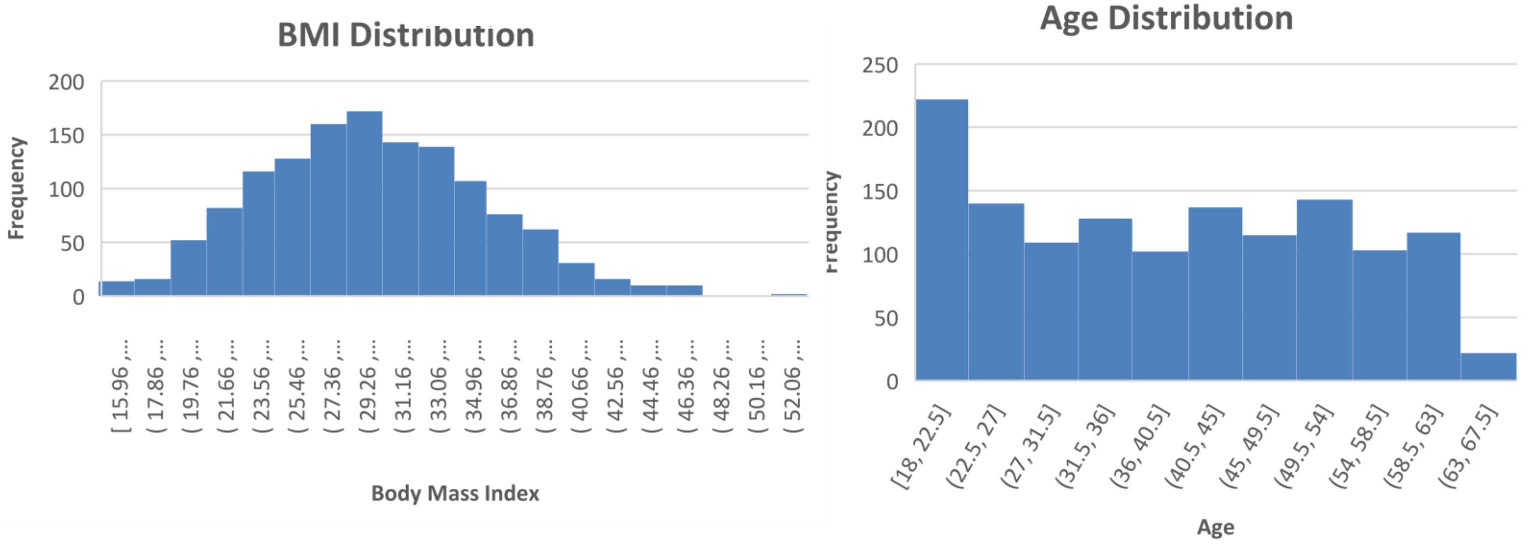
Variables	Status
Age	Independent variable
BMI	Independent variable
Children	Independent variable
Gender	Independent variable
Location	Independent variable
Charges	Target Variable

Variables	Status
Age	Independent variable
BMI	Independent variable
Children	Independent variable
Male	Independent variable
Female	Independent variable
Northeast	Independent variable
Northwest	Independent variable
Southeast	Independent variable
Southwest	Independent variable
Charges	Dependent variable

2.3 Exploratory Data Analysis and Further Data Processing

This section is comprised of visualizing each variable and going in depth to see the attributes separately (univariate) as well as comparing it with other variables (multivariate) analysis to find a relationship with each other. In Figure 1, it can be observed that data in the BMI column follows a normal distribution whereas, age column is slightly skewed, similarly insurance charges is positively skewed in our dataset. Conversely, Figure 2, shows the relationship of two attributes (gender and insurance) and (location and insurance). With the box plot, we can witness a few outliers in the dataset as well, however, since these outliers constitute of a significant dataset and are better captured in relation to the other attributes in the dataset, we believe that keeping them would preserve the essence of the dataset.

Figure 1: Univariate Analysis of different variables



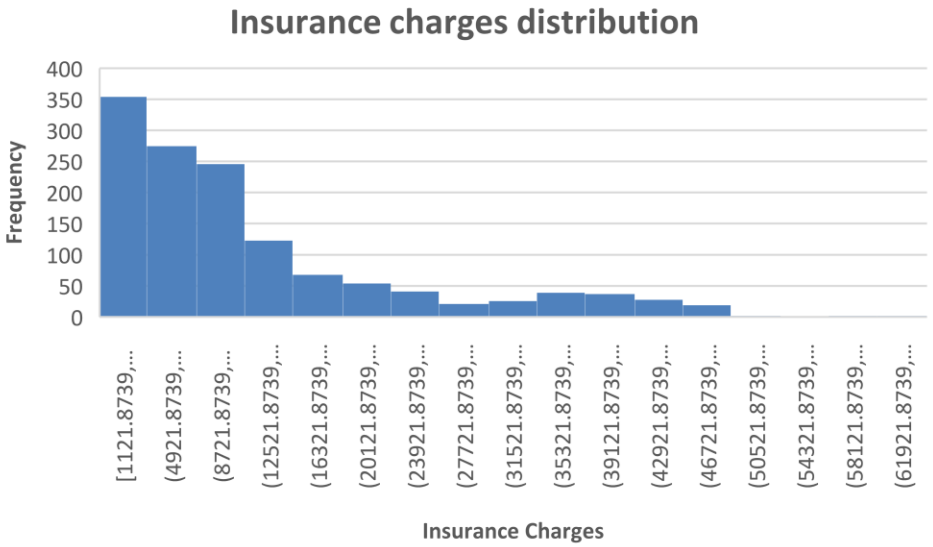
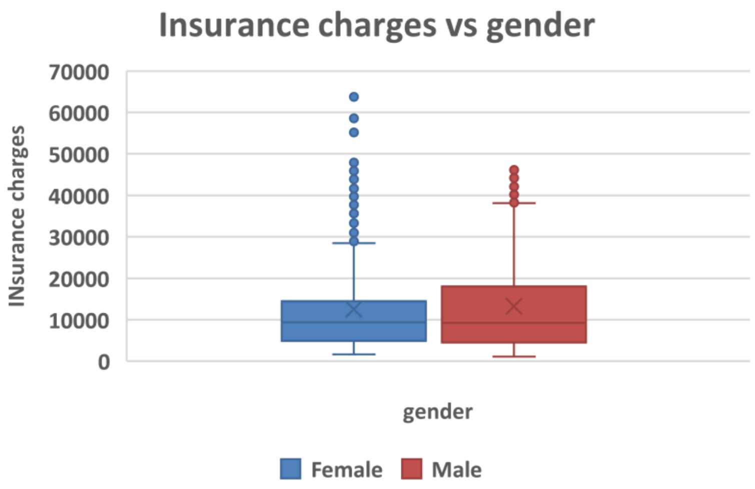
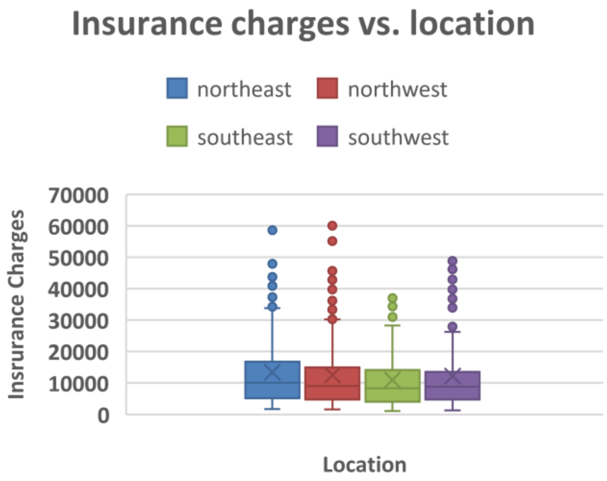


Figure 1 : Bivariate analysis of different variables





3. Methods

In this section, the sampling methods, the models applied and the metric we have chosen to evaluate the performance of our models would be discussed.

3.1 Models

3.1.1 Correlation

The data is explored using correlations and scatter plots to visualize the relation between various parameters/attributes and how these attributes will help to predict the insurance charges. The correlation analysis will help us understand what would happen to the insurance charges when there is an increase/decrease in the parameters.

3.1.2 Multiple Linear Regression

Since, the correlation can provide an indication of whether an attribute is significantly correlated with the target variable and does not consider other variables present in the dataset, we have resorted to fit the dataset into multi-linear regression model. The linear regression model considers all the attributes in question and computes the weightage that each of these independent variables have on the dependent variable. Hence, we have computed linear regression to provide a linear equation that can be used to compute the insurance charges. The data is processed using MS- Excel. Multi-Linear Regression evaluates the relationship between a dependent variable and two or more independent variables. (Gulden & Nese,2013)

The equation to obtain the best fit line is as follows:

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \cdots + \beta n * xn$$

Where, y = Dependent variable, (x1, x2….) = Independent variables, β0 = constant term, βi = coefficients relating to y and x1.

It is assumed that the error term has a mean value of 0. The accuracy of the estimation model was judged by employing statistical measures such as multiple linear regression between observed and predicted insurance charges.

3.2 Significant Association

Since the dataset is not normally distributed, we were not able to make use of the z- test to identify the significant difference in the insurance charges based on the attributes. Hence, we used chi-squared test, which is an analysis of independence that is conducted to test the influence of dependent variables (McHugh,2013). It is a statistical computation that compares the observed distribution of data with the expected distribution. The level of significance was set at 5%. The formula for calculating a Chi-Square is:

$$\sum_{\boldsymbol{i-j}} \chi 2 = \frac{(\boldsymbol{O} - \boldsymbol{E})^2}{\boldsymbol{E}}$$

O Observed (the actual count of cases in each cell of the table); **E** Expected value (calculated);$i-j$ is the correct notation to represent all the cells, from the first cell (*i*) to the last cell (*j*); χ^2 is the cell Chi-square value

3.3 Performance Metrics

In order to predict the model performance, we first divided our dataset into two parts training and testing dataset. 80% of the data is used as the training set which is randomly taken from the dataset to ensure there is no partiality and the rest 20% of the unseen data is the testing set which is fitted into the linear regression model obtained from the training set. The predicted values in the testing dataset are further compared with the actual values to test the reliability of the model. We used absolute error (AE), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). All these metrices focus on the error made in predicting the insurance charges.

4. Results

4.1 Correlation Analysis

The correlation analysis of the health insurance dataset shows that there are six attributes which have positive correlation with the insurance charges whereas there are three attributes which have negative correlation. However, we have found that only the smoker attribute has a significant positive correlation of 0.79 and that this attribute would be an important factor while calculating insurance charges.

Particulars	*age*	*bmi*	*children*	*Male*	*Female*	*Smoker*	*northeast*	*northwest*	*southwest*	*southeast*
charges	0.30	0.20	0.07	0.06	(0.06)	0.79	0.01	(0.04)	(0.04)	0.07

4.2 Linear Regression Analysis

The linear regression model computed an R squared value of 0.75 showing that 75% of the variation in the dataset can be explained using the linear regression model. The linear regression analysis of the health insurance dataset indicates that, among the nine attributes, the weightage for smoker is the highest one, such that one's smoking status would impact a lot more in calculating the insurance charges as compared to other attributes. It can be assumed that, perhaps, smokers are more prone to diseases and need to frequent the hospital regularly, increasing their medical expenses. Enrolling these individuals in an insurance plan might mean that the insurance company might have increased expenditures, and that could be one of the reasons why they are charged with higher insurance charges. Other factors like number of offspring, location as well as gender too have an influence on the insurance charge accrued by an individual. These attributes show a high significance level in calculating the target variable.

Particulars	*Coefficients*	*Standard Error*	*P-value*
Intercept	(12,422.82)	1,000.68	1.46355E-33
age	256.86	11.90	7.78322E-89
bmi	339.19	28.60	6.49819E-31
children	475.50	137.80	0.000576968
Male	-	-	#NUM!
Female	131.31	332.95	#NUM!
Smoker	23,848.53	413.15	-
northeast	352.96	476.28	0.46
northwest	-	-	#NUM!
southwest	(607.09)	477.20	#NUM!
southeast	(682.06)	478.96	0.15

4.3 Significant Association

4.3.1 Significant Association between gender and insurance charges

A chi- squared test is performed on the insurance charges and gender. And the insurance charges are divided into five categories (lowest, low medium, medium, medium high, highest) and gender (male and female). Using the chi-squared test, we calculated the p value of 0.05, which shows that there is a significant association between gender and insurance charges. On other words, the gender of a person does have an impact on the insurance charges that is paid by an individual.

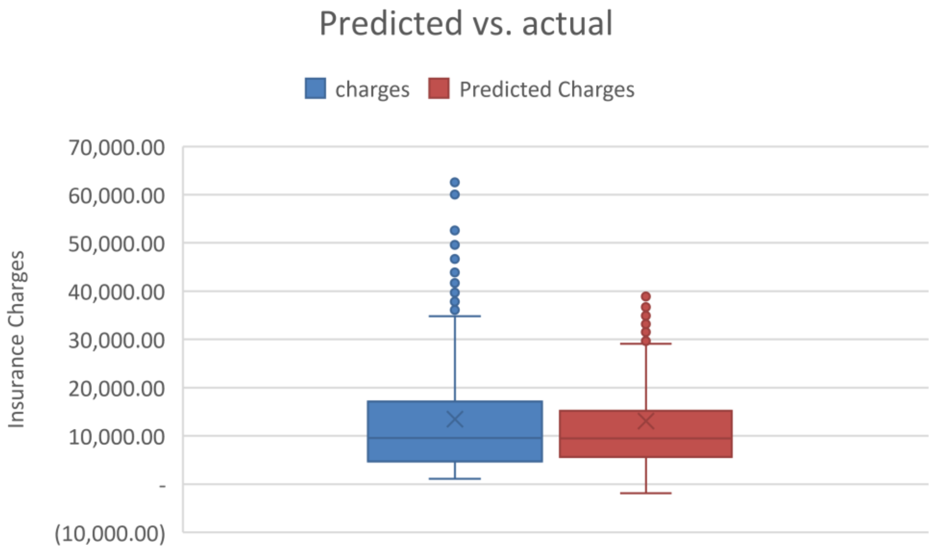
4.3.2 Significant Association between location and insurance charges

The insurance charges are divided into five categories and the location is divided into four (northeast, northwest, southeast, southwest). While computing the p value using the chi squared test between (southeast vs. southwest) showed a value of 0.0038, (northwest vs. northeast) showed 0.22, (southwest vs. northwest) showed 0.099 and (southeast vs. northeast) showed 0.00977. Thus, there is a significant association between two of these variations but not on the other two when computing the insurance charges. Hence, we can conclude that the location where a person is residing might have impact on the insurance charges when comparing southeast and southwest as well as southeast and northeast.

4.4 Prediction and performance metrics

The prediction of the insurance charges using the regression model equation from the train data in the test data showed a mean absolute percentage error of 0.42. Although there are a few disparities in the dataset's reliability when dividing the dataset, a significant difference is not observed.

Particulars	Mean Absolute Error	Mean Squared Error	Mean Absolute Percentage Error	RMSE
Whole worksheet	4,170.89	36,501,893.01	0.42	6,041.68
Training data	4,151.56	36,372,447.12	0.42	6,030.96
Testing data	4,225.23	37,175,951.41	0.42	6,097.21



5. Conclusion

In this group project, we have taken the dataset about health insurance and there are certain independent attributes that are used to calculate the insurance charges paid by an individual. Upon visually analyzing the data, we have found that although there is just one significant correlation between the independent variable and the dependent variable, there are other values which needs to be taken into consideration while evaluating the charges paid for insurance as signaled by linear regression. A linear regression model is performed, which results in an R squared value of 0.75, indicating that the model can accurately predict the insurance charges using the attributes given at least with 75% accuracy. This shows that it is a good model to predict the insurance charges. Furthermore, there is a significant association of insurance charges with gender and with the location in certain regions but not at others. On the other hand, the prediction metrics assessed mean absolute percentage error at 42% showcasing a medium reliability on the model used.

6. Future Work

Although the model had an R squared value of 0.75, there are certain limitations to the dataset, which is acting as the constraint. If these factors were to be addressed in the dataset, we could have expected a higher R squared value and a more accurate prediction. The dataset is not symmetrical and hence it might be more biased towards one group as compared to the other. In the future, we can make sure to obtain equal data from people in different age groups and as well as people from both genders.

In addition to this, there are many other variables which might affect the insurance charges of a person, for instance, past medical history, occupation, type of plan chosen, consumption habits, however, the dataset only considers five major attributes, and hence we are not able to predict the insurance charges with high accuracy level. With the inclusion of these attributes, we might be able to obtain a better prediction result, but a 100% accuracy is highly unlikely.

References

McHugh ML. The chi-square test of independence. Biochem Med (Zagreb).2013; 23(2):143-9. doi: 10.11613/bm.2013.018. PMID: 23894860; PMCID: PMC3900058.

Gulden Kaya Uyanik, Nese Guler, A Study on Multiple Linear Regression Analysis, Procedia - Social and Behavioral Sciences, Volume 106, 2013, Pages 234-240, ISSN 1877-0428. https://doi.org/10.1016/j.sbspro.2013.12.027.

[1] https://www.iciciprulife.com/health-insurance/what-is-health-insurance.html.

[2] https://www.healthcare.gov/blog.