Prediction Model for Computing Insurance Premiums

Business Understanding

Understanding the factors that influence insurance premiums is important for insurance companies because it can help them to set premiums that are fair and appropriate for their customers. This is especially important for the underwriting process, where insurance companies assess the risk of an insurance event occurring for each policyholder and determine the appropriate premium based on that risk. By analyzing a dataset containing information on policyholder characteristics such as age, BMI, sex, number of children, smoker status, and region, insurance companies can gain insights into the relationships between these characteristics and premiums, and use this information to set premiums aligned with the level of risk. For instance, companies may find that premiums are generally higher for policyholders who are older, have higher BMIs, are male, are smokers, or have more children, all other variables being kept constant.

This dataset can also be used to predict premiums for new customers, using statistical techniques such as multiple linear regression or classification. By accurately predicting premiums, insurance companies can improve their risk management and financial stability, and provide better value to their customers.

Research Question

The main objective of this project is to build a model that will predict the insurance premiums that can be used by insurance companies to manage risk and ensure financial stability.

Objectives

- To identify the most significant features in determining insurance premiums
- To build a linear regression model that can accurately predict insurance premiums based on input features from the Kaggle insurance premium prediction dataset.
- To assess the performance of the predictive model and identify potential areas for improvement.

Data Understanding

Data Source

The dataset used for this project was obtained from Kaggle.

Data Description

Our data was in csv format and contained data grouped in columns of dependent and independent variables. The insurance.csv dataset contained 7 columns and 1338 observations (rows). The 7 columns contained 4 numerical features (age, bmi, children and expenses) and 3 categorical features (region, sex and smoker).

Data Preparation

Loading the data

At the beginning of the process, the necessary libraries were imported and then the insurance.csv dataset was loaded onto the jupyter notebook using pandas.

Reading and checking the data

The data was read and then checked for anomalies, outliers, missing values and duplicates. This was to determine the next course of action that would ensure the data would be set for use. During this process, it was established that even though the data did not have missing values, it had duplicates and outliers.

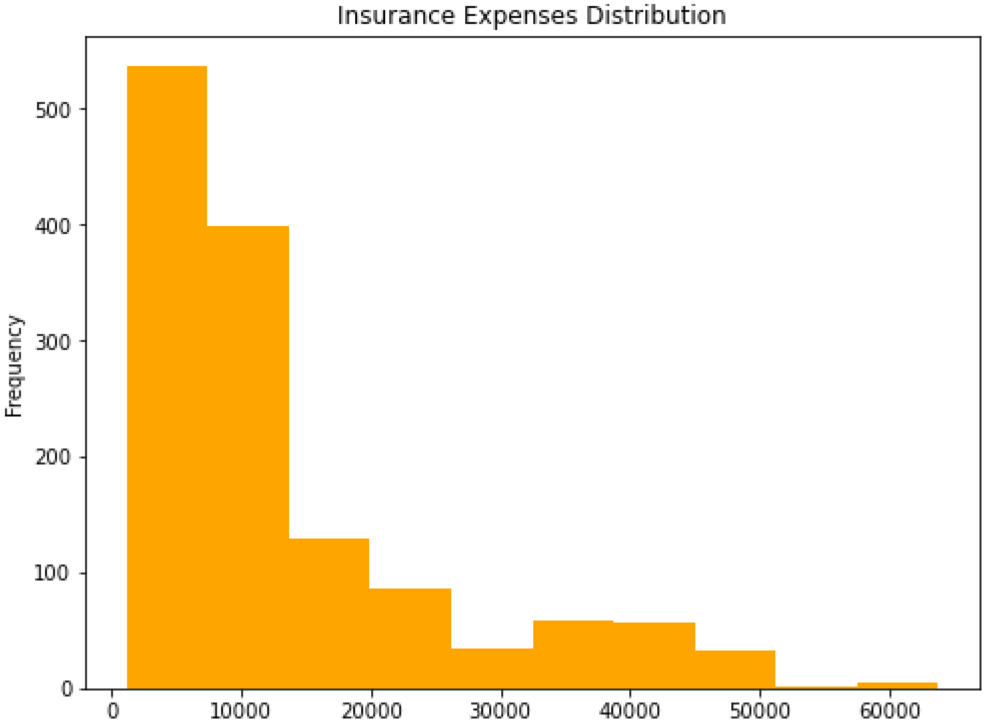
Cleaning the data

The data was then cleaned and pre-processed to ensure that it was in a usable form. This was done by removing duplicates. The outliers were kept because they were essential for training the model. Outliers are only removed after performing model diagnostics and it has been established that they are irrelevant for the study, otherwise they are kept.

External data source validation

The data set was measured against a reliable external data source (The Kaiser Family Foundation) to ensure that it was in line with what it should be and checked for any additional issues with the data set.

KFF published a report that established annual premiums for employer-sponsored family health coverage reaching USSD 22,463. The typical annual deductible health coverage for employees would be USD 1,644 which was not very far from that of the insurance dataset as expressed in the graph below:



Exploratory Data Analysis

The data sets were analyzed and trends found by using statistics and visualizations to aid in comprehending the data set. There were several questions that were answered in this step by comparing the predictor variables with the target variable which was expense (premium) using data visualization tools. The questions answered and variable relationships established include:

- At what age is smoking prevalent? (Age vs smoker)
- What regions contribute the highest insurance premiums? (Region vs premium)
- What regions contribute the highest number of smokers? (Region vs smoker)
- Depending on age, what premium would a smoker be charged? (Age vs smoker vs premium)

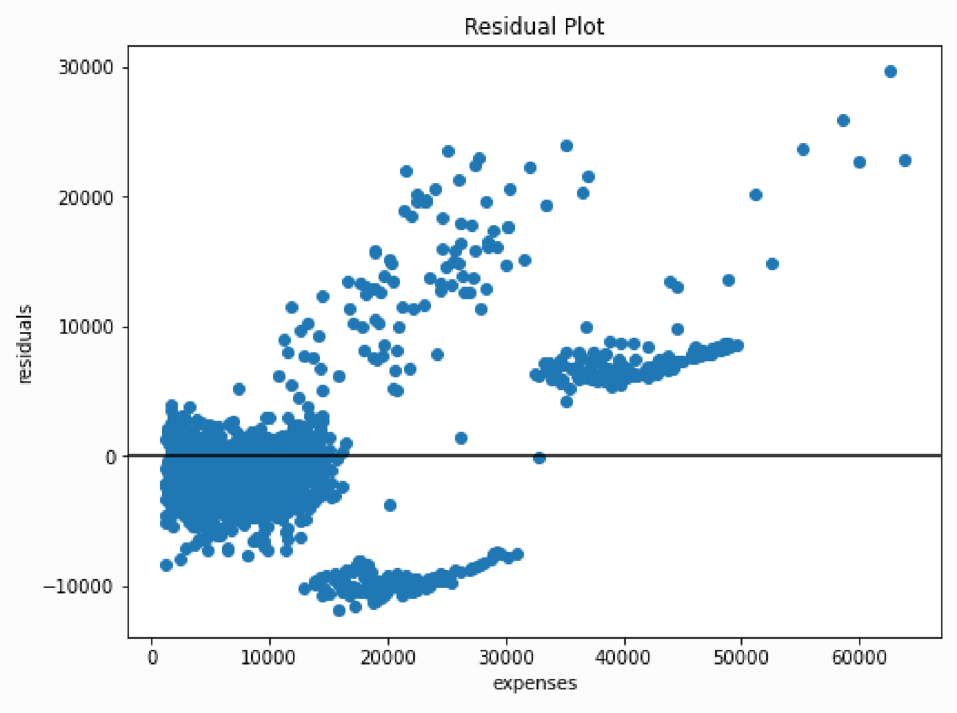
- How would bmi affect the rate of premium that a smoker would be charged? (BMI vs premium vs smoker)
- How does sex affect the rate of premium? (sex vs premium)
- How does being a smoker affect the rate of premium? (smoker vs premium)
- How does the number of children affect the rate of premium? (children vs premium)

Modeling

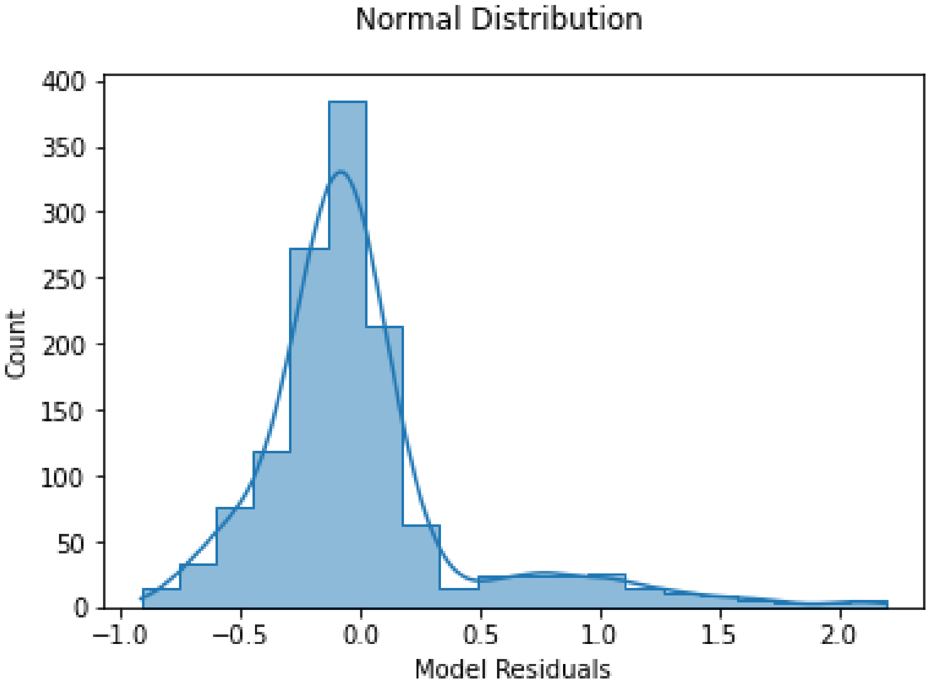
Data modeling commenced by checking the correlation between the predictor and target variables. This resulted in weak correlations for all predictor variables and a further analysis needed to be and was conducted.

Since the data set consisted of both categorical or numerical variables, binary encoding was used to convert the data into a form that can be used by the Ordinary Least Squares (OLS) model.

Using the Least Squares Method (LSM) method, in the first model, it was noted that it was statistically significant with a probability of F-statistic below 0.05 and the model was able to explain 74.8% of the variation in the insurance premium. Additionally, the Mean Absolute Percentage Error (MAPE) of the model was 42.26% meaning that for any prediction made by the model, the prediction will have an expected error of 42.26%. Lastly, the linearity and heteroskedasticity assumption was tested and there existed a non-linear relationship between the predictors and the response variable. The plot shown below, also shows that there was some evidence of heteroskedasticity as the residual displays a funnel-shape.



Due to the high MAPE, non-linear relationship and heteroskedasticity log transformations were carried out to transform the data to approximately conform to normality. After conducting log transformations on the x and y variables, it was noted that the model improved. This was seen where the adjusted R-squared increased to 0.764 from 0.748 which meant that the model explained 76% of the variation in the insurance premiums (USD). Moreover, the MAPE of the model was 3.15% meaning that for any prediction made by the model, the prediction would have an expected error of 3.15%.After conducting the log transformation, it was noted that the residuals were nearly normally distributed as shown below.



The adjusted R-squared in the second model was better; however, upon the introduction of interaction terms in the third model, the adjusted R-squared increased to 0.821. This meant that the model explained 82.1% of the variation in the insurance premiums (USD). Additionally, the MAPE of the model was 2.32% which meant that for any prediction made by the model, the prediction would have an expected error of 2.32%. The following insights were deduced from the third model regarding the predictor variables:

From the data above the following can be noted:

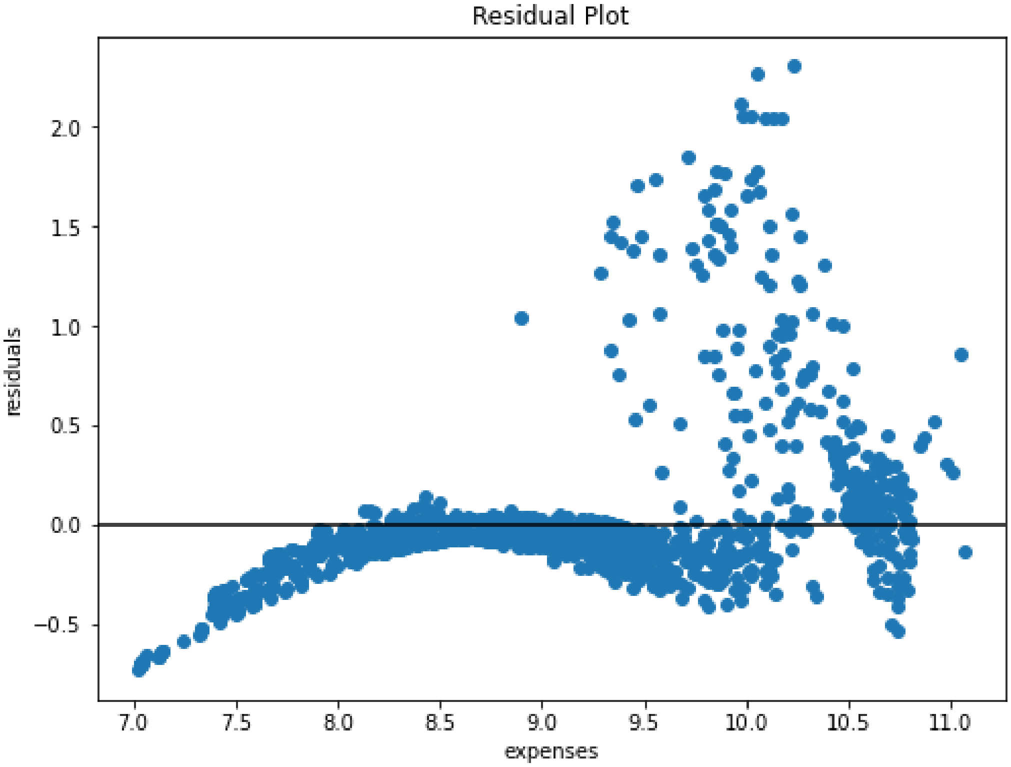
- For each increase of 1 year in age, there would be an associated increase of 4.16% in expenses.
- For each increase of 1 unit in BMI, there would be an associated decrease of 0.1% in expenses.
- For 1 increment in the number of children , there would be an associated increase of 10.63% in expenses.

- For each male, there would be an additional deductible expense of USD 1.088 $(np.exp(0.0847))$ compared to a female.
- For each smoker, there would be an expected deductible expense of USD 3.607 $(np.exp(1.2829))$ compared to a non-smoker.

In this case, the model developed was :

$$y = 0.0416 * age - 0.001 * bmi + 0.1063 * children + 0.0847 * sex + 1.2829 * smoker - 0.0286 * region1 - 0.0351 * region2 - 0.0333 * age \times smoker + 0.05077 * bmi \times smoker + 7.0528$$

Lastly, the heteroskedasticity was relatively less prevalent compared to our previous models as seen below.



Conclusion

Determination of insurance premiums can be a complex and time-consuming process. This predictive model would be useful to insurance companies to ensure that they can accurately estimate insurance premiums based on a variety of input features and reduce the risk of adverse selection.