Jefferson Veliz

Professor Avinash Jairam

Data Mining for Business Analytics

July 5th, 2023

Lab 2- Medical Insurance

The aim of this project is to develop a linear regression tool to predict medical insurance prices based on various factors such as age, number of children, BMI, location, smoker status, and gender. The motivation behind this project is to create a model that can accurately estimate insurance prices, which can be valuable for individuals and insurance companies alike.

To start the project, I loaded a dataset from Kaggle containing the relevant columns mentioned earlier. I then performed data exploration to understand the dataset better. The dataset did not have any missing values (NAs), ensuring a reliable foundation for analysis. The dataset consisted of 676 males and 662 females, and it covered regions from the southeast, southwest, northwest, and northeast. I also investigated the number of smokers in the dataset and found that most individuals were non-smokers, providing an important characteristic to consider during the analysis.To gain further insights into the data, I utilized various visualizations. First, I created a correlation plot that revealed a slight correlation between age and insurance charges. This correlation is an important factor to consider when building the predictive model. Next, I generated a scatter plot with a linear regression line between BMI and insurance charges. The resulting linear regression showed a slight positive correlation, indicating that higher BMI values were associated with higher insurance charges. Additionally, I used a pie chart to visualize the distribution of smokers and non-smokers in the dataset. This visualization revealed that a significant portion of the individuals were non-smokers. Furthermore, I created a bar plot to analyze the distribution of individuals within each region. The bar plot indicated that most of the dataset consisted of individuals from the southeast region. To prepare the data for regression modeling, I encoded categorical variables using the "get\_dummies" method for the region and sex columns. This step enabled the utilization of these variables in the linear regression model. I proceeded to split the data into training and testing sets and developed a linear regression model using the training data. The model's intercept term was calculated to be -12250.255454599479, representing the baseline insurance price when all feature values are zero. To assess the model's performance, I calculated several metrics. The R-squared score, which measures how well the model fits the data, was found to be 0.7642348545269875, indicating that the model explains approximately 76.42% of the variance in insurance prices. The mean squared error (MSE), a measure of the average squared difference between predicted and actual values, was calculated as 38108732.489800245. A lower MSE indicates a better fit of the model to the data. The root mean squared error (RMSE), which measures the average deviation between predicted and actual values, was determined to be 6173.227072593413, indicating an average deviation of approximately 6173.23 units. Additionally, the mean absolute error (MAE), measuring the average absolute difference between predicted and actual values, was computed as 4292.580187720698, indicating an average absolute deviation of approximately 4292.58 units. In conclusion, the linear regression model demonstrated promising predictive power in estimating medical insurance prices based on the provided features. The R-squared score and other evaluation metrics suggest that the model has significant predictive capabilities. However, further analysis and testing may be required to assess the model's robustness and generalizability to ensure its suitability.