Medical Insurance Cost Forecast

Truc Tran

Introduction

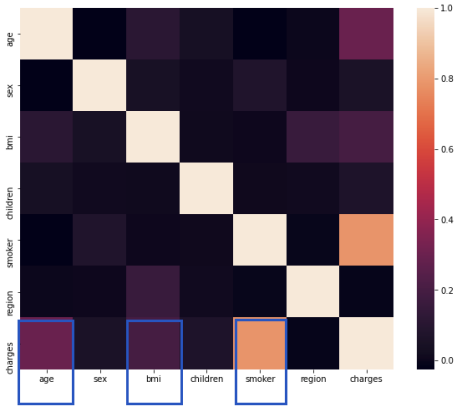
Medical expenses are challenging to estimate because of the complexity of individual medical conditions, accidents, technology, and geography. As a result, American health insurance firms struggle to determine how much to charge for insurance coverage. They must precisely predict medical bills for their clients to make a profit. Health insurance firms chose to fund and invest in forecasting models to estimate individual medical spending bills. This project aims to create an accurate medical insurance cost forecast system for future clients using data obtained from other people around the United States.

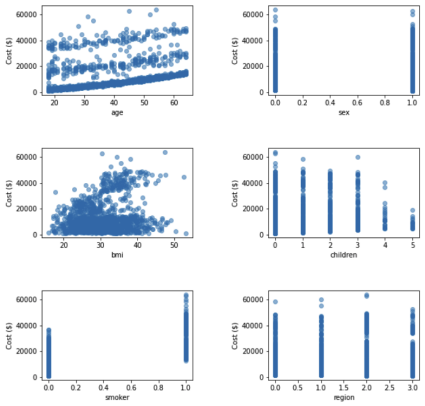
Data Wrangling

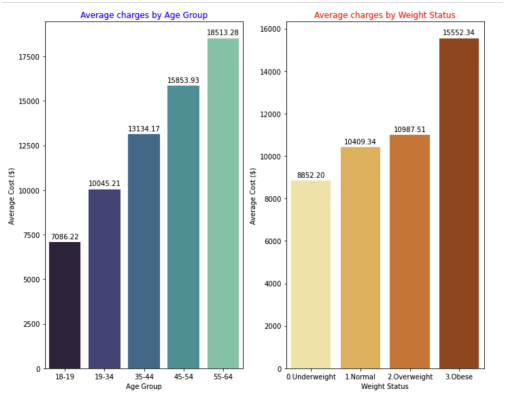
There are 1338 rows and 7 columns in the dataset. The seven columns are age, gender, children (the number of dependents), smoker, region, and charges. This collection has only four areas: southeast, southwest, northeast, and northwest. The data includes a person's personal information and their annual medical bills if they have health insurance. There is no missing information in any of the columns. I add two new features,age_group, and weight_statu, depending on a person's age and BMI, to easily illustrate the distribution of various groups.

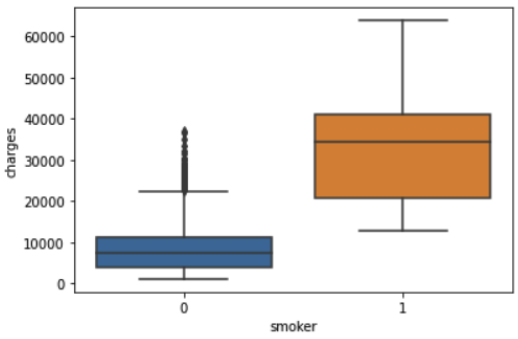
Exploratory Data Analysis

The categorical features are sex, smoker, region, weight_status, and age_group. The rest are numerical features: age, bmi, children, and charges—numerical grouping characteristics for each region to compare the difference in charges and its distribution among the region. I applied the average charges for each region, weight_status, and age_group. The distribution indicates that age and weight are correlated to the medical insurance cost, especially on the smoker feature. Most people who do not smoke have medical expenses, starting from $ 0 to $30,000, while smokers have charges starting at least $12,000 to more than $60,000 annually.
I created a heatmap and scattered plots to see the relationship and correlation amongst the features.









Based on the graphs, heatmap, and scatterplot, there are positive correlations between the medical insurance charges with smoker, bmi, and age.

Preprocessing and Training

During the preprocessing stage, I convert all categorical columns to numerical values, subsequently one-hot encoded. The target features "charges" will be stored in the y data frame, and the rest in the X data frame. In all data frames, I divided the dataset into an 80/20 train/test split. Three models were used: linear regression, random forest, and lasso regression. Before training the dataset for each model, I scale the data to ensure consistent measurement. I utilize a cross-validation process to identify the best k for the linear regression model, which divides the training data into k folds and trains the model on k-1 folds. The best k is 7; the top six most influential features are smoker_yes, age, weight_status_3.Obese, bmi, age_group_19-34, age_group_55-64, and smoker_no. With a random forest model pipeline, I impute the missing value with median, then apply feature scaling and split the train set into 5-fold cross-validation. The best parameters for the random forest model are n_estimator = 100, simpleimputer_strategy = mean, and use StandardScare() function. In this model, the four important features are smoker_no, smoker_yes, bmi, and age. The best parameters for the lasso regression model are lasso_normalize: True, lasso_selection: random, pca_n_components: 16.

The R2 score, mean absolute error, and root mean square error are the three metrics used by each model to evaluate performance using their ideal parameters. The findings of the three models are displayed in the table below.

	R2	Mean Absolute Error	Root Mean Squared Error
Linear Regression	0.799771	4008.447484	5536.117762
Random Forest	0.860833	2736.647824	4615.413312
Lasso Regression	0.796857	4021.451883	5576.260168

According to the results table, the random forest outperforms all three criteria, with an average smallest error of about $2736.64.

Modeling

According to the random forest model, the average individual charge is $13,590.31, while the actual cost is $13,100.04. This dataset's average absolute error is $2736.64. The four most important features are smoker no, smoker yes, bmi, and age. Health insurance companies must consider three primary factors that influence their clients' medical costs: smoking status, age, and BMI, particularly for those who smoke, are elderly and are obese. Companies can review and collect the client's critical information, as suggested by the model above, and then use the random forest model to forecast their client's medical bills and calculate the ultimate health insurance fee.