Project Report
on
Insurance Cost Prediction

596: Regression and Time Series Analysis

By,
Atharva Adbe - aa2159
Anchala Krishnan - fa439
Rutu Desai - rmd228
Shubham Kokane - ssk203

Under the Guidance of
Prof. Ying Hung

Rutgers University-New Brunswick

Aim of the Project:

- ➔ To predict the insurance cost within different regions of the US given specific attributes.
- ➔ To find a correlation between the attributes and insurance cost charges.

Dataset Source:

- ➔ https://www.kaggle.com/mirichoi0218/insurance

About the Dataset:

- ➔ Age: age of the primary beneficiary
- ➔ Sex: insurance contractor gender, female, male
- ➔ Bmi: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, an objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- ➔ Children: Number of children covered by health insurance / Number of dependents
- ➔ Smoker: Smoking
- ➔ Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- ➔ Charges: Individual medical costs billed by health insurance

Questions from Dataset:

- ➔ What will be the Insurance cost of an individual based on different attributes in the dataset?
- ➔ What are the most significant features while predicting the Insurance Costs of Individuals?
- ➔ How bad is smoking for the person’s health and how it’ll affect the Insurance Cost?
- ➔ Which region in the USA has the Highest Insurance Cost Premium?
- ➔ Can the number of Children/dependents increase the Insurance Cost?
- ➔ Does the insurance cost differ based on an individual’s gender or not?
- ➔ How do different age groups have a distribution of Insurance Costs?
- ➔ How BMI can have an effect on Insurance Costs?

Model/Techniques Used:

- ➔ We pre-processed the dataset and divided the dataset into training and testing data with the ratio of 80:20 % respectively.
- ➔ We used Forward Selection and Backward Elimination Approaches to find the significant variables for the model.
- ➔ We applied Regression Models on the significant variables got from the above 2 approaches including Multiple Linear Regression, Random Forest Regression, and Decision Tree Regression to compare different models based on their accuracy, AIC value, and no. of dependent variables so as to choose the best among all models.

Results:

- ➔ This was the results from different approaches and regression models:

Approach	**Significant Variables**	**Multiple Linear Regression**	**Random Forest Regression**	**Decision Tree Regression**
Full Model	Age, sex, BMI, region, children, smoker	79%	**87%**	68%
Forward Selection	Age, BMI, Smoker, Region	79%	**85%**	66%
Backward Elimination	Age, Smoker, Region	**77%**	69%	64%

Table-1 Approaches and the Accuracies for different models applied

Approach	**AIC**	**Regression Model**	**Number of Independent variables**	**Accuracy**
Full Model	27112.45	Random Forest Regression	6	87.17%
Forward Selection	27120.7	Random Forest Regression	4	85.02%
Backward Elimination	27255.2	Multiple Linear Regression	3	77.25%

Table-2 Best Approaches and their results

Conclusion:

- ➔ In conclusion, we started out with three different models for testing purposes; random forest modeling, decision tree, and multiple linear regression. Within each model, we performed forward selection and backward elimination in addition to testing with all the variables. By performing the ANOVA test on the top 3 models, we conclude that random forest modeling with all the variables fits our dataset the best since it has the lowest AIC score with the highest accuracy rate.

Future Scope:

- ➔ Add more attributes including hours of sleep, hours of exercise, diet restrictions, etc.
- ➔ Add more rows to the dataset.
- ➔ Use hybrid variable selection method
- ➔ Apply different types of models and compare for precision
- ➔ Add Confidence Intervals to the predictions.