TOPIC: Predicting the Health Insurance Premium in the United States

GROUP #19

1. Prashant Dhungana-30080130
2. Prateek Kaushik-30229287
3. Mariya Mathews-30192182
4. Nisha Pillai-30158934

INTRODUCTION

In Canada the health services are provided at no cost because it is covered and funded by the government. The system is called “Universal Health Care System” and was passed by legislation in 1984 under the Canada Health Act. However, the same does not apply for our neighboring country, the United States. The health services in the United States are mostly private and unlike Canada the people in the States have to pay for their health service fees. Due to the privatization of the healthcare system in the United States there is a competition within the industry to provide better services which makes it expensive for the people to afford. Therefore, health insurance companies play a huge role as they cover the cost of health services in exchange for an insurance premium. By purchasing health insurance people protect themselves from any future uncertainties and a huge debt. In order for the insurance companies to cover the cost of health services they must analyze how much premium to charge for each person. Every person has different day to day habits and can vary in their lifestyle choices which impacts how the insurance companies charge their premiums.

The objective for this project is to develop the best model in predicting insurance premiums based on several factors using multiple linear regression. In order to predict the insurance premiums, we first must determine which predictor variables are significant and develop a model with predictors that explains a high variability in terms of the premiums charged. This topic is very important as people residing in Canada can use this model to predict the insurance premiums they will be charged according to their personal information if they ever plan on moving to the United States.

The motivation for our study is to find the factor that most influences insurance price and help people plan accordingly to reduce their financial burden of paying high premiums. We will be using multiple linear regression to address this query.

METHODOLOGY

DATASET

In order to answer our research topic we have acquired a dataset that has 7 variables and 1338 observations. The dataset contains both numerical and categorical values. The dataset is clean and no further cleaning is required. The variables are explained below according to the source of the dataset:

Response Variable:

Charges: Insured individual medical annual costs billed by the health insurance in USD($). (Quantitative)

Predictor Variables:

1. Age: Age of the primary Insured. (Quantitative)
2. Sex: Insured gender “Male or Female”. (Qualitative)
3. BMI: Body Mass Index. (Quantitative)
4. Children: Insured number of dependents. (Quantitative)
5. Smoker: Insured smoking habits “Yes or No”. (Qualitative)
6. Region: Insured region of residence “Southwest”, “Southeast”, “Northwest”, “Northeast”. (Qualitative)

The dataset was acquired through the Kaggle website and is an open-sourced data licensed under “Open Data Commons”. [https://www.kaggle.com/datasets/mirichoi0218/insurance/]

WORKLOAD DISTRIBUTION

STEPS	METHODOLOGY	TEAM MEMBER
Step 1	Model building & Model selection including Stepwise, Forward, Backwards, All possible regressions selection, T-test & F-test	Prashant Dhungana
Step 2	Model building continued with Interaction model and Quadratic (higher) order model	Prateek Kaushik
Step 3	Checking the Regression Assumptions including Linearity, Independence, Equal variance, Normality, Multicollinearity, and identifying if any Outliers	Nisha Pillai
Step 4	Box-cox transformations in case of non-normality, Model evaluation by reiterating and refining process and Predicting the insurance premium	Mariya Mathews
Step 5	Preparing Final report	All team members with an equitable contribution

End of Project Checkpoint