USA health insurance cost analysis

Summary

Analysis of USA health insurance data to predict the cost of premiums.

Outline

- Business Problem
- Data and methods
- Results
- Conclusions

Business Problem

- This information could be used by potential customers researching how much their premiums would be.
- This could also be used by health insurance companies as well to compare their premiums or how they have weighted an individuals variable.

Data and methods

- The data used in this analysis is from Kaggle.
- This is a collection of data on various factors that can influence medical costs and premiums for health insurance in the United States. The dataset includes information on 10 variables, including age, gender, body mass index (BMI), number of children, smoking status, region, income, education, occupation, and type of insurance plan.
- I used multi linear regression on the raw data, dummies for the categorical variables and log transformed the continuous variables to ascertain the most accurate predictive model

Results

- Using the coefficients I gathered the top 5 variables which increase the cost of health insurance premiums and the top 5 which lower them.

TOP 5 FEATURES FOR COST INCREASE

	coefficients
Smoker_yes	5008.534372
Coverage_level	2500.48536
BMI	1599.932202
Family_medical_history	1402.349659
Medical_history	1401.419856

TOP 5 FEATURES FOR LOWER COST

	coefficients
Unemployed	-1487.756091
Student	-992.579746
South-west Region	-798.248727
North-west Region	-703.109754
South-east Region	-499.342339

Results

	RAW MODEL	DUMMIES MODEL	LOG TRANSFORMED MODEL
TRAIN MSE	4305117.22861077	2255545.56821866	2261219.7409103
TEST MSE	4302655.68724418	2259460.9798744	2264208.15957426
ADJ. R-SQUARED	0.779601002417424	0.884528944300211	0.884238459053467

- The Mean squared error and R-squared improved significantly between the raw and dummied data.
- It had a tiny affect between the dummied data and log transformed data.
- The train MSE and test MSE are very close in the log transformed data. The model fits well.

Conclusions

- This model can measure an individuals health information and predict their health insurance premium with 88% accuracy.
- Current health, habits and family health history play a significant role in the cost of premiums

MOVING FORWARD

- With more time, I would have looked at the outliers and used GLS for the heteroscedasticity contained within the data to perhaps render a more accurate predictive model.