Predicting Insurance Premiums

Maddie Bauer
DSC 680 - Summer 2021
https://madelinebauer.github.io/PredictingInsurancePremiums/

Which Domain?

Health insurance is market that will never die out as most people have (or should have) some level of health insurance. This has especially been expressed this year due to the worldwide COVID-19 pandemic. Illnesses, accidents, medications, surgeries, etc. all can add up in your yearly budget, but by having health insurance an individual is not liable for paying an entire medical bill. One of the many important tasks within the health insurance industry includes determining policy premiums. With predictive modeling, insurance companies can determine an accurate and fair policy premium based on a customer's behaviors and attributes. By analyzing customer habits and behaviors over time, insurers are able to anticipate future behaviors and provide the right insurance product(s) and policy premiums. This project will aim to predict future premiums and verify the results by using regression models.

References

[1] https://www.healthcare.gov/how-plans-set-your-premiums/ - what can/cannot affect premiums

[2] https://statisticsbyjim.com/regression/predictions-regression/ - making predictions with regression analysis

[3] https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70 - ways to evaluate regression models

[4] https://favtutor.com/blogs/types-of-regression - types of regression models

[5] https://medium.com/analytics-vidhya/predicting-medical-insurance-costs-machine-learning-e1e4e7c4e8ed - predicting costs with machine learning

[6] https://www.lotuslabs.ai/accurate-insurance-claims-prediction-with-deep-learning/ - Deep Learning

[7] https://www.formotiv.com/predictive-analytics-in-insurance/ - predictive analytics in the insurance industry

[8] https://www.duckcreek.com/blog/predictive-analyitics-reshaping-insurance-industry/ - predictive analytics in the insurance industry

[9] https://online.maryville.edu/blog/predictive-analytics-in-insurance/ - what is predictive analytics?

[10] https://www.alchemer.com/resources/blog/regression-analysis/ - what is regression analysis?

Which Data?

What is the dataset you'll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

The dataset being used is a dummy dataset as real insurance data is not available to the average citizen due to privacy laws. However, I am confident that the dataset I am using is comparative to real-world data because according to healthcare.gov, premiums are set by looking at only your age, location, tobacco use, individual vs. family enrollment and plan category [1]. My dataset consists of age, gender, BMI, number of children, tobacco use, region in the United States and yearly charges. I also collected data from my close family and friends who did not mind sharing it with me so that I could further my learning. This adds more data points to the initial dataset which can help make better predictions. The only inconsistency was that a few had a hard time remembering their BMI from their most recent doctor's visit. So, if they couldn't remember their exact BMI, they estimated.

Research Questions? Benefits? Why analyze these data?

Like I mentioned above, health insurance isn't going anywhere any time soon. Changes may be made in the years to come, but health coverage is a must in today's world.

1. Can we predict charges for new individuals?

2. Which variables in the dataset truly affect the premium?

3. Is tobacco use a main factor in determining the premium amount?

4. Does the size of your family affect the premium amount?

5. What attributes can individuals change (tobacco use, regions, etc.) to lower their premiums?

What Method?

I plan to start with importing and merging the two datasets. I will then briefly view summary statistics and determine if there are any outliers or missing values that need to be taken care of. I will then explore the data through many visualizations to get a better understanding of what I have available to me. I will then split the data into training and testing sets so that I can run regression models. At this point I am planning to use simple linear regression, support vector machine, and random forest models. I may end up altering these or adding more models in once I get going. I will then compare all of models and determine which one performed the best. I can then use the model to predict premiums for new individuals.

Potential Issues?

I feel fairly confident with the plan I've outlined above. I have considered completing this project solely with R as Python has been my go-to throughout this program. I think it might be beneficial to have a project with R in my portfolio to showcase my skills. However, I don't want to be slowed down by having to continuously look things up in textbooks or online to remember how to code in R.

Concluding Remarks

While the healthcare industry is always evolving, I think it is safe to say it will always be around. Individuals cannot afford to pay all of their medical expenses up front, especially when accidents or surprising diagnoses occur. Thoughtful use of predictive analytics has only recently allowed insurance companies to improve their premium pricing accuracy, create customized plans for individuals and their families and help build stronger relationships with their customers. By analyzing customer behaviors and attributes overtime, we can accurately predict future premiums by using regression analysis.