Introduction

The purpose of this project was to identify the factors that influence both accidents and insurance rates, both on a state-wide level and on a more specific basis.

Crash Analysis

The crash analysis focused on fatal accidents. By far, motorcycles caused the most fatal accidents per 1000 miles traveled. Younger drivers were more likely to be involved in fatal accidents. Our analysis showed that snow and sleet were the most dangerous weather conditions. Rural roads tended to have a higher rate of fatal accidents. The distinction between urban and rural roads was moderately correlated (0.4) with accident rate. The age of drivers within an area did not affect crash rates (each had a correlation of less than 0.1).

Other factors that could have impacted the crash analysis are the make and model of cars, analyzing all crashes not just fatal ones (these might be more prevalent in urban areas where cars are not traveling at large speeds), quality of drivers education in each state, as well as state/city ordinances that affect drivers.

Statewide Demographics Analysis

For the analysis of statewide demographics and how they relate to car insurance premiums, we utilized data from the United States Census bureau on population, including breakdowns by age and gender, distribution of area and population between Urban and Rural, and household income. We also examined insurance premiums by state, from the Insurance Information Institute, and the crash data discussed above.

Overall, the results of this analysis step were largely inconclusive. There were few strong correlations present between the demographic factors and the insurance rates. The strongest correlation was with the Urban and Rural percentages, which was approximately 0.6 for both area and population. Additionally, population density had a moderate correlation of 0.48 for Rural areas and 0.41 for Urban areas. There was a surprising moderate negative correlation (-0.44) between insurance expenditures and accident rates. None of the population demographics had a meaningful correlation with insurance premiums.

Insurance Claims Machine Learning Models

To conclude our investigation we applied our knowledge to attempt and fit our data to a regression model as well as a classification model. For the regression analysis, we attempted to predict the cost of an insurance claim based on a number of predictors, such as Income and account age. Despite our efforts, the best results we were able to achieve with this model was an accuracy approaching 0.3. This indicates that the data within our dataset was not effective in predicting the claim cost, so we would need additional information to further our model.

Our Classification model attempted to predict the result of a claim; whether it was accepted by the insurance company or not. This model proved to be much more accurate than our regression model, with our best results coming from a Random Forest Classifier producing an impressive 98% successful prediction rate. Across all the models we tested, including Naive-Bayes, Decision Trees, SVM, and KNN, all models were able to produce at least a 80% accuracy rate.

Conclusion

Our analysis largely shows a need for more data. Correlations and machine learning show that we are likely missing data on many of the most important factors involved in determining insurance premiums and claim amounts.