2020 Traveler’s Competition

Yue Du, Broderick Prows
URI

Methods

We started by exploring the distributions of the claim cost and features. We saw that only 6% of the policies even reported a claim, and few reported more than a single claim. This meant the data was 0 inflated and would not fit well in a traditional linear model.

We did some research on how to handle 0 inflated data and tried to use quantile regression to estimate the top 6% of the data, but the performance was still poor.

We thought that we could use a classification algorithm to identify the policies that were likely to have a claim, and then predict claim cost for those, but this too was not a very good model.

We fit a Tweedie GLM model, which has been widely used in the insurance industry because of its ability of modeling frequency and severity of insurance claims, but this still did not return a high gini index.

We found a paper by Yi Yang, Wei Qian, Hui Zou (2018) that proposed a gradient boosted method for tweedie models, a popular model in actuarial problems.

Did not understand GINI index until late

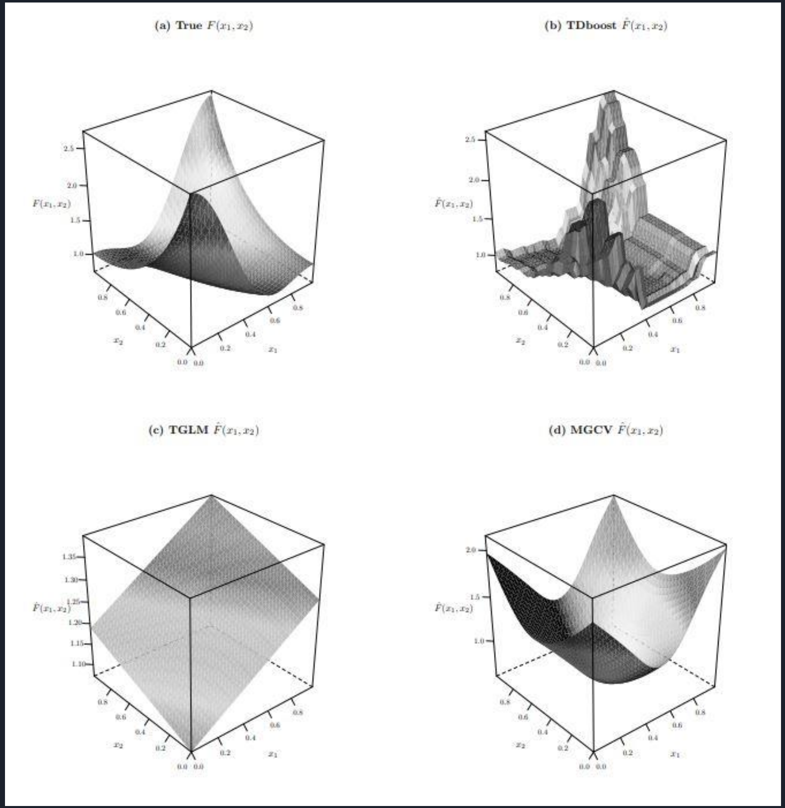
TDboost

Advantages

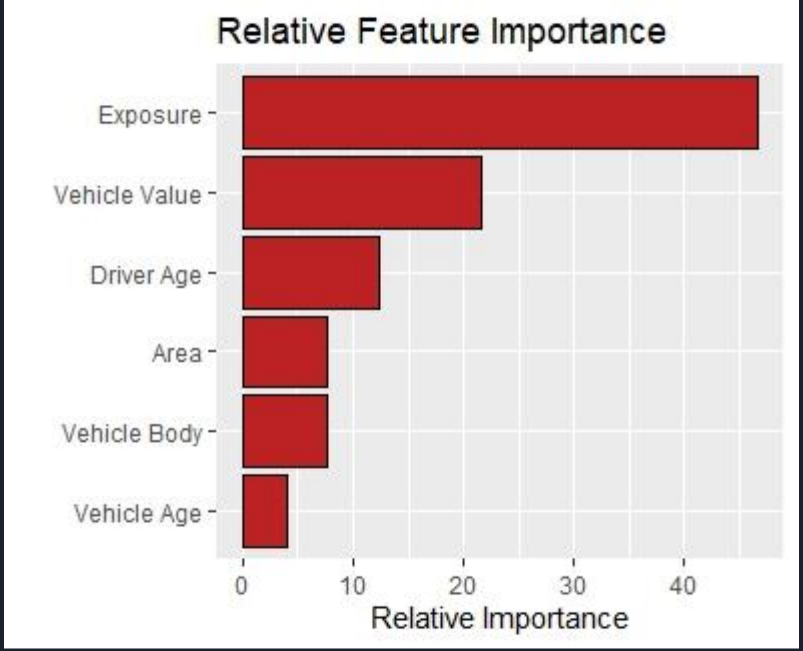
- No Assumption of Linear Relationship
- No need to do variable selection or specify interactions
- Not a black box relative to other nonparametric methods

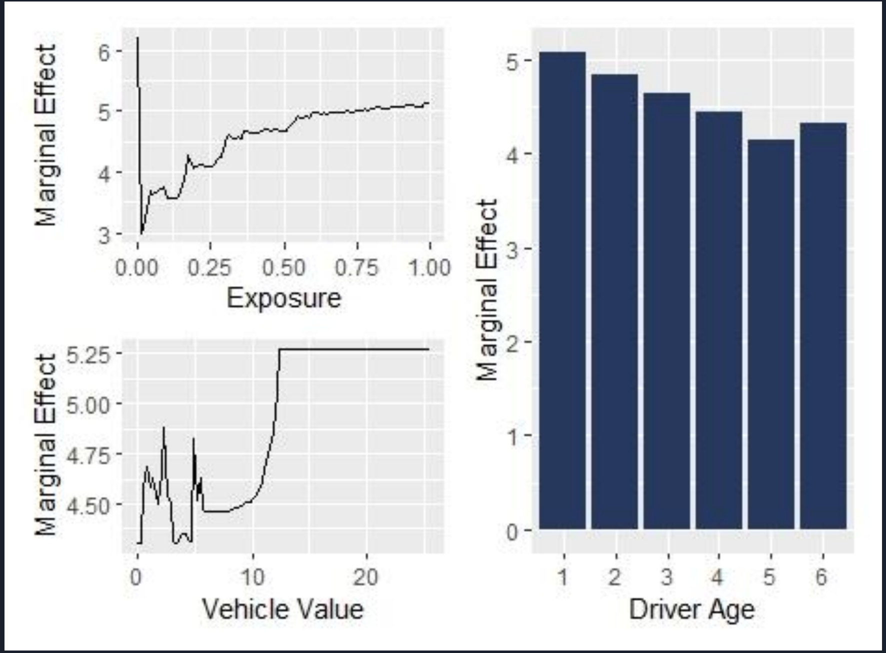
Performance

- GINI Train = .47
- GINI Test = .55
- Kaggle Results = .1125



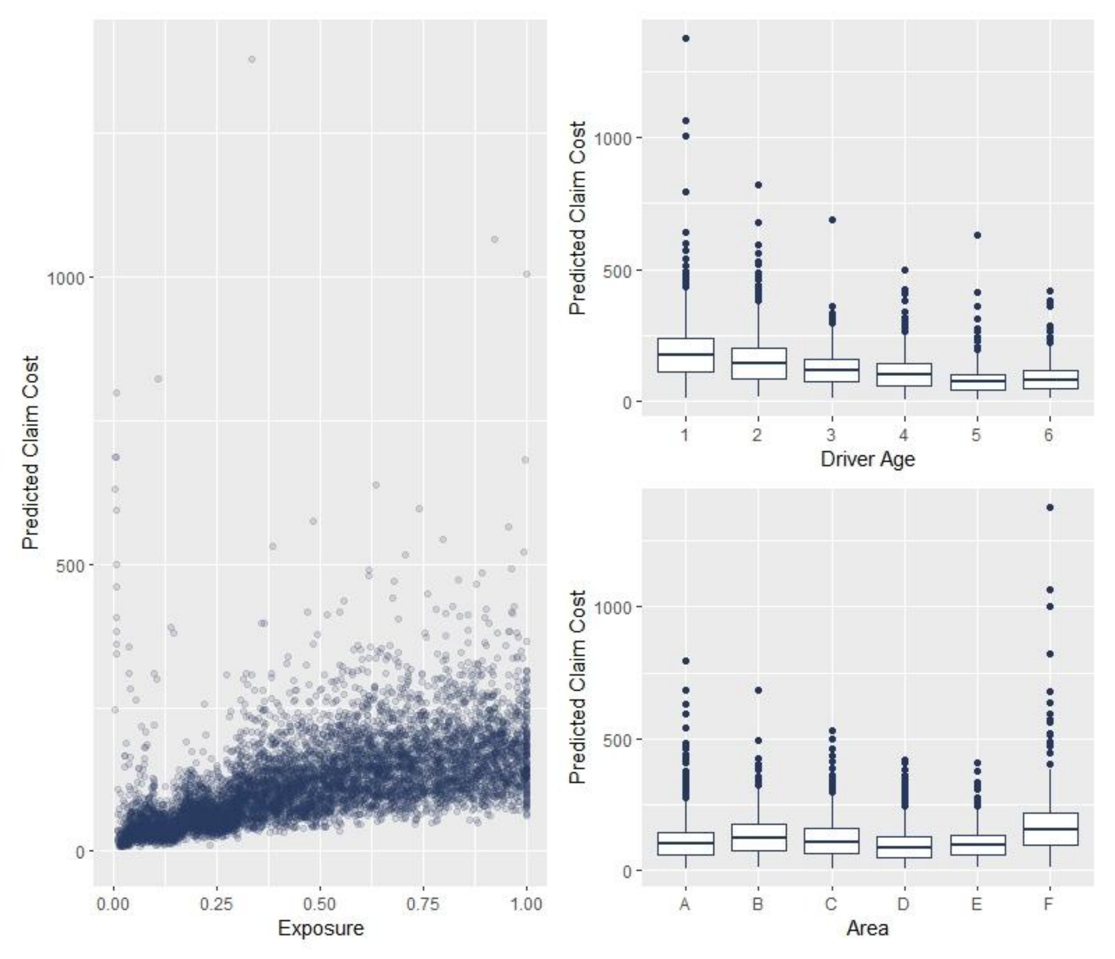
Business Findings - Feature Importance





How do features affect output

- Exposure seems to capture some of the variation
- Increased variance in the lower Driver Age predictions
- Increased variance in Area F



Data we wish we had

Previous driving history

- # of tickets
- # of accidents

Vehicle

- Ownership Status
- Maintenance history

Personal Info

- Commute Time to work
- Avg. Yearly Mileage

Possibly useful, but dangerous/impossible

- Reaction time
- Vision and Hearing
- Driving with children

While this data might be useful, it would lead to a model that is discriminatory!

Questions

- What is Exposure?
- Why was vehicle value standardized?