Claim Cost Prediction for Auto Insurance Data

Harsh Dubey, Ankit Kumar

University of Massachusetts Amherst
BE REVOLUTIONARY™

BE REVOLUTIONARY™

Objective

To predict claim cost for each policy ID using a set of 20 predictors

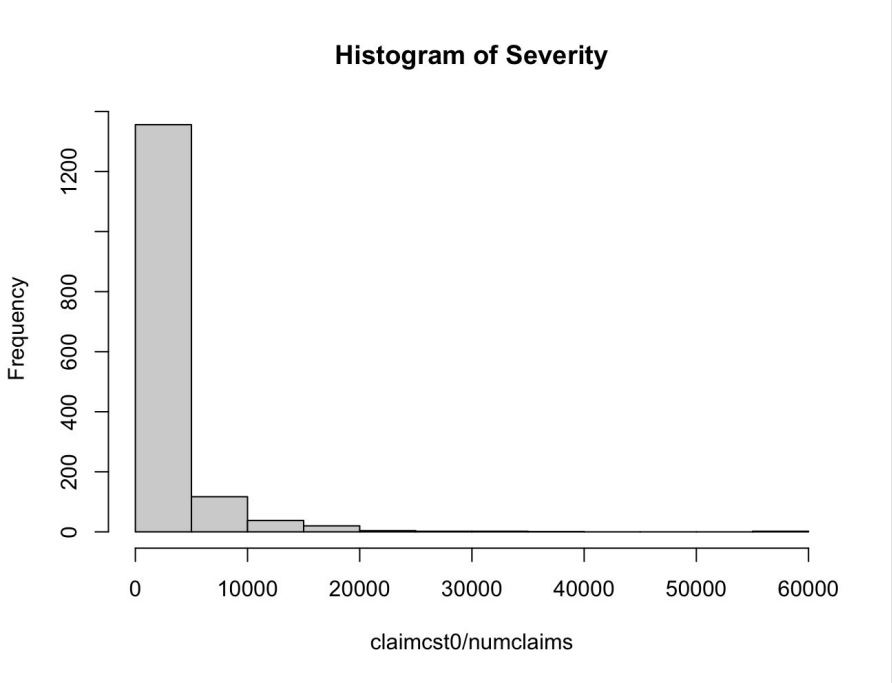
University of Massachusetts Amherst

What methods did we consider?

- Initial exploration of the claim cost variable with respect to the other features showed us that about only 6.8% of the policies have a claim. So, we clearly had a 0 inflated target variable
- We thought of two modeling strategies - **Single Stage** and **Two Stage**
- **Single Stage Model**
  - Linear Regression
  - LightGBM
  - Neural Networks
  - Random Forest
  - Tweedie Model
- **Two Stage Model**
  - Random Forest - Light GBM
  - Poisson - Light GBM
  - Poisson - Inverse Gaussian **(Chosen Model)**

What method did we choose in the end and why?

- The distribution of claim cost (claimcst0) is heavily concentrated around 0 and rightly skewed.
- Generalized Linear Models (GLMs) were preferred instead of Ordinary Least Squares to address the normality violations.
- Concerns with single model for claimcst0 prediction
- Opted for alternative approach (De Jong, 2008)
- **claimcst0 = Frequency ∗ Severity = numclaims ∗** $\frac{\textbf{claimcst0}}{\textbf{numclaims}}$
- **Adopted Two-Stage Model:**
  - Predicts **Frequency** of claims (numclaims)
  - Predicts average **Severity** of a claim (claimcst0/numclaims)



What method did we choose in the end and why?

- Predict **‘Frequency’**:
  - Defined as number of claims for each policyholder.
  - Count data for each policyholder.
  - Options: Poisson Regression, Negative Binomial Regression.
  - Training data: Mean = 0.073, Variance = 0.078.
  - Chosen approach: Poisson Regression.
- Predict **'Severity'**:
  - Defined as claim cost divided by the number of claims.
  - Claim cost distribution (claimcst0): Strongly right-skewed.
  - Options: Gamma Regression, Inverse Gaussian Regression.
  - Gamma Regression had dispersed predictions; Inverse Gaussian Regression chosen due to better results, avoiding extreme predictions on the right tail.

Model	GINI Index
Training	0.23827
Kaggle (Public)	0.19932
Kaggle (Private)	0.20366

Variable Selection

- **Stage 1: Predicting numclaims**
  - We had 13 categorical variables in the dataset!
  - Not all categorical variables had sufficient number of observations in each category
  - We ran chi-sq tests to check how many of those categorical variables are related
  - We also ran a one-sample t-test to check for significant variables

Chi-Sq test variables

vehicle body
vehicle age
gender
area
age category

One sample t-test variables

exposure
area
age category
term length

Final Select

vehicle age
gender
area
age category
term length
exposure
vehicle value

- **We used a stepwise regression technique with all the variables of the Final Select column**

Variable Selection

- **Stage 2: Predicting severity**
  - We followed the same process for predicting severity as we did for predicting numclaims
  - The final model consisted of Chi-sq test variables and one sample t-test variables

Chi-sq test variables
vehicle body
vehicle age
gender
area
age category

One sample t-test variables
exposure
gender
area
age category
term length
driving history score

Final Select
vehicle value
gender
area
age category
term length
driving history score

- **We used a stepwise regression technique with all the variables of the Final Select column**

What variables help explain pure premium?

Final Models after stepwise regression:

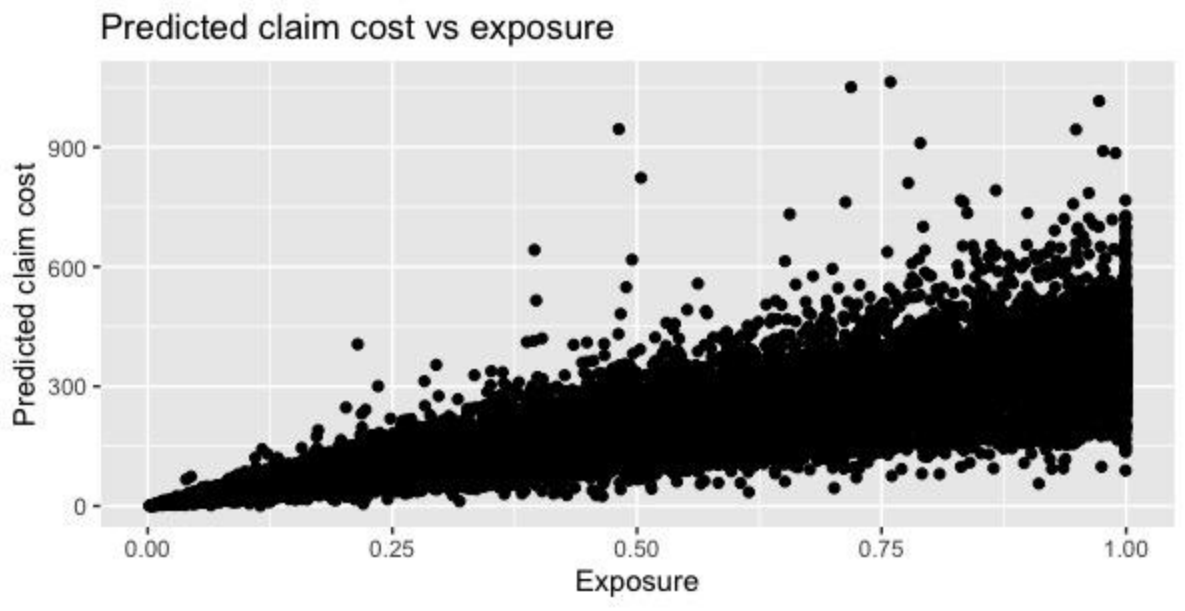
Frequency Model

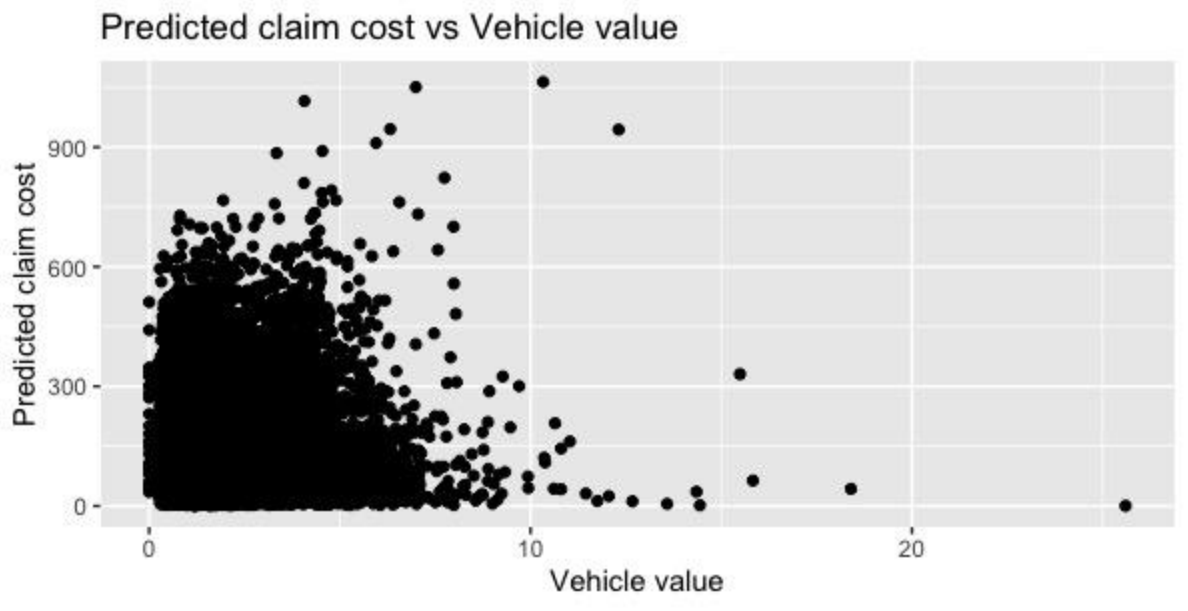
numclaims ~ exposure + age category + area + vehicle value + vehicle age
+ vehicle value:vehicle age + area:vehicle value

Severity Model

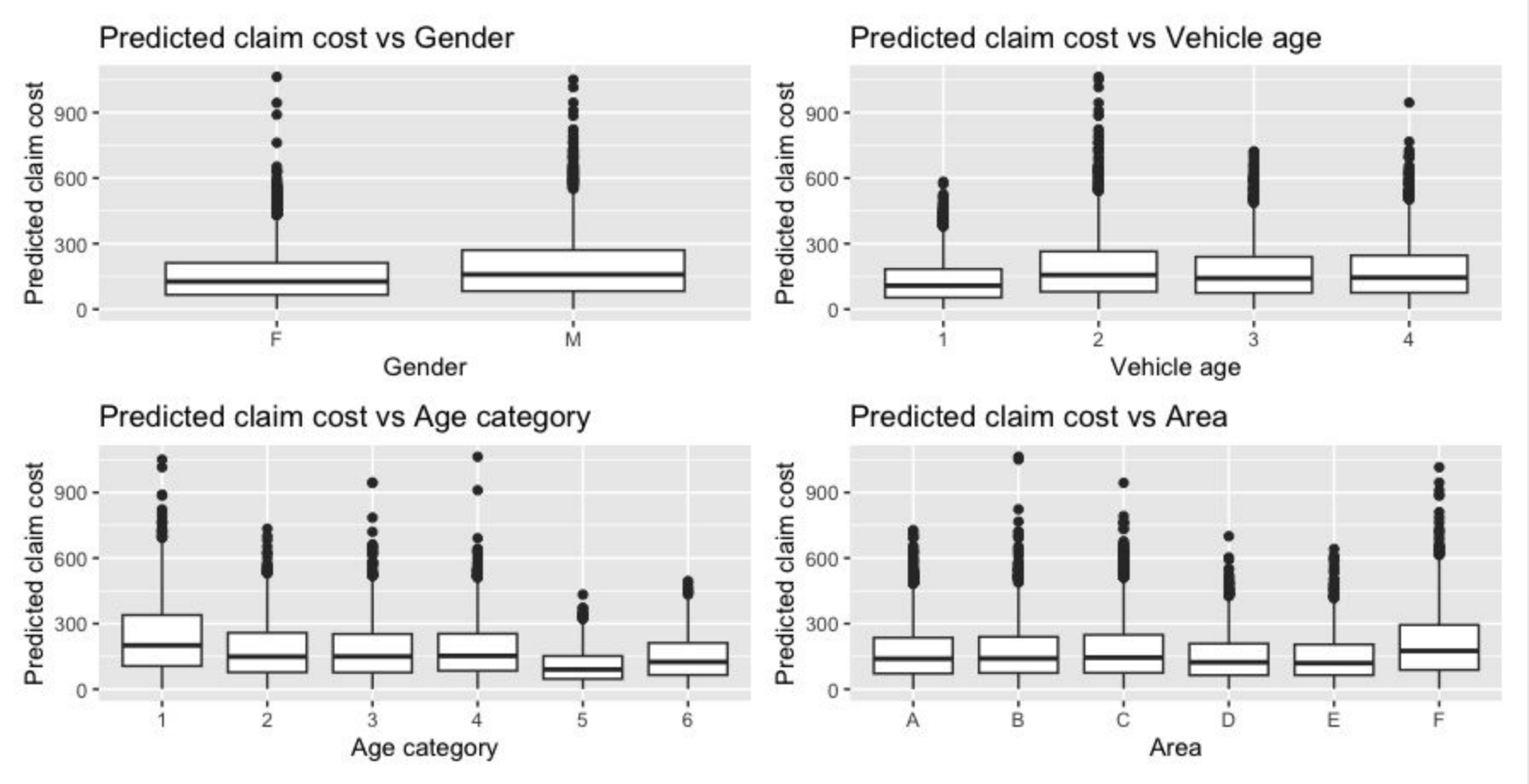
(claimcst0/numclaims) ~ gender + vehicle age + age category

What variables help explain pure premium?



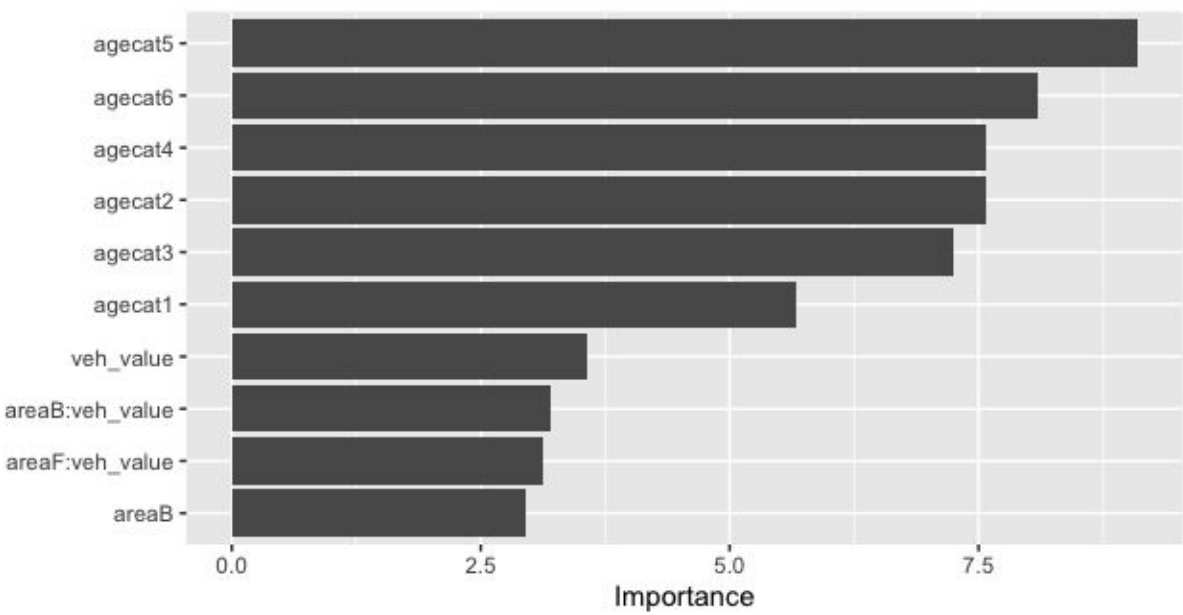


What variables help explain pure premium?

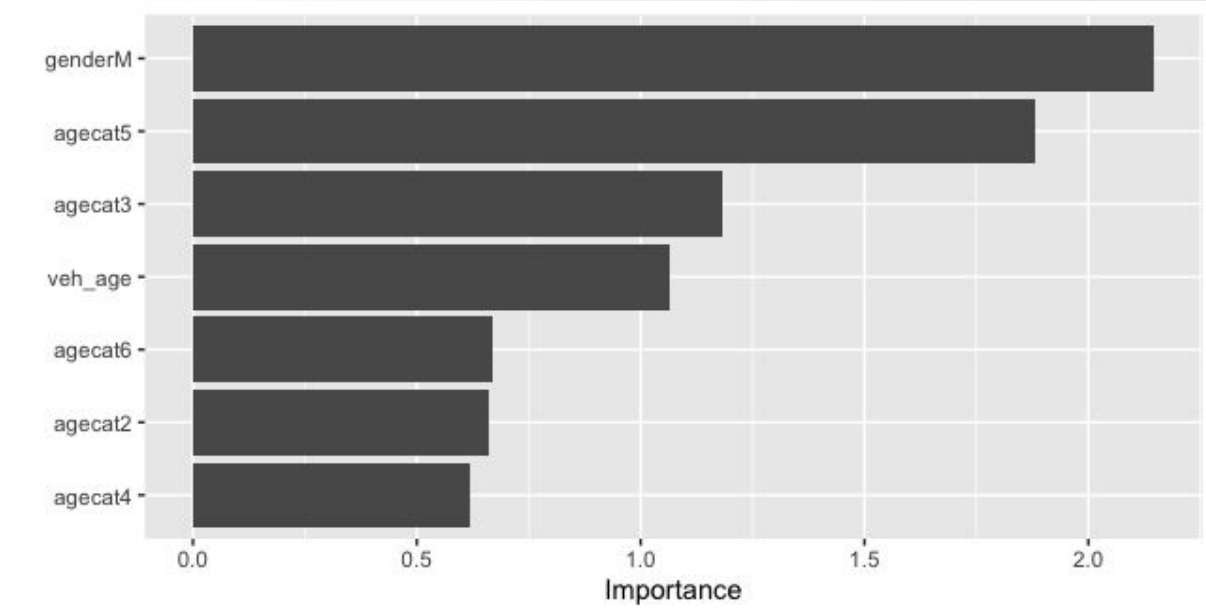


Relative importance of Variables

Frequency Model



Severity Model



References

1. De Jong, Piet, and Gillian Z. Heller. Generalized linear models for insurance data. Cambridge University Press, 2008.
2. Ye, Chenglong, et al. "Combining predictions of auto insurance claims." Econometrics 10.2 (2022): 19.
3. Noll, Alexander, Robert Salzmann, and Mario V. Wuthrich. "Case study: French motor third-party liability claims." Available at SSRN 3164764 (2020)


Thank You!
Questions?
University of Massachusetts Amherst
BE REVOLUTIONARY™

Additional Questions?

- **What other variables not in the data set do you think might be useful?**
  - Income
  - Accident/Maintenance history
- **Any concerns about the resulting model?**
- **What questions do you have about the data?**