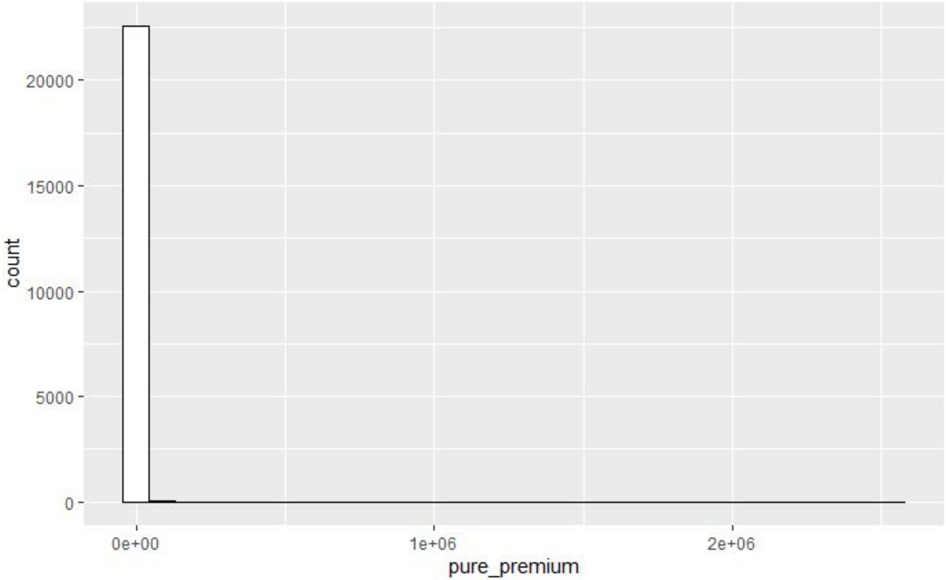
InNova Auto Insurance Company Modeling

Group 3:
Anubha Agrawal
Jacob Gursky
Zhaoliang Zhou

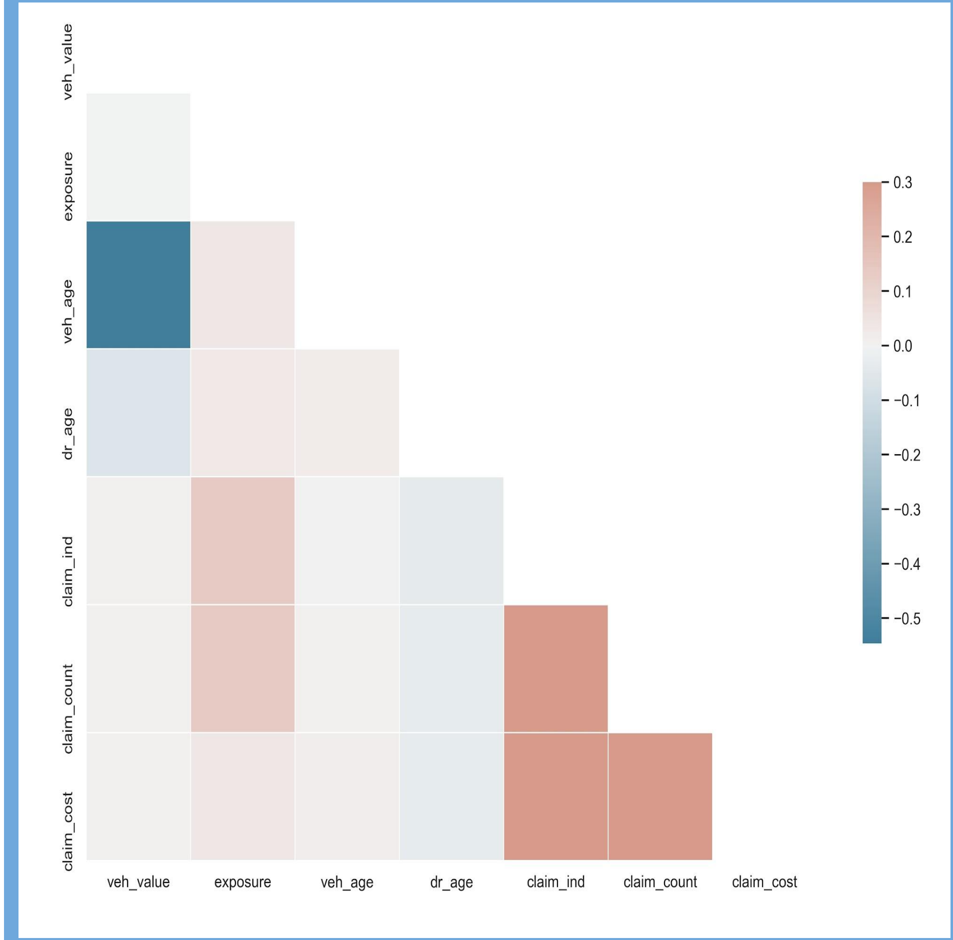
Exploratory Data Analysis

Response variables

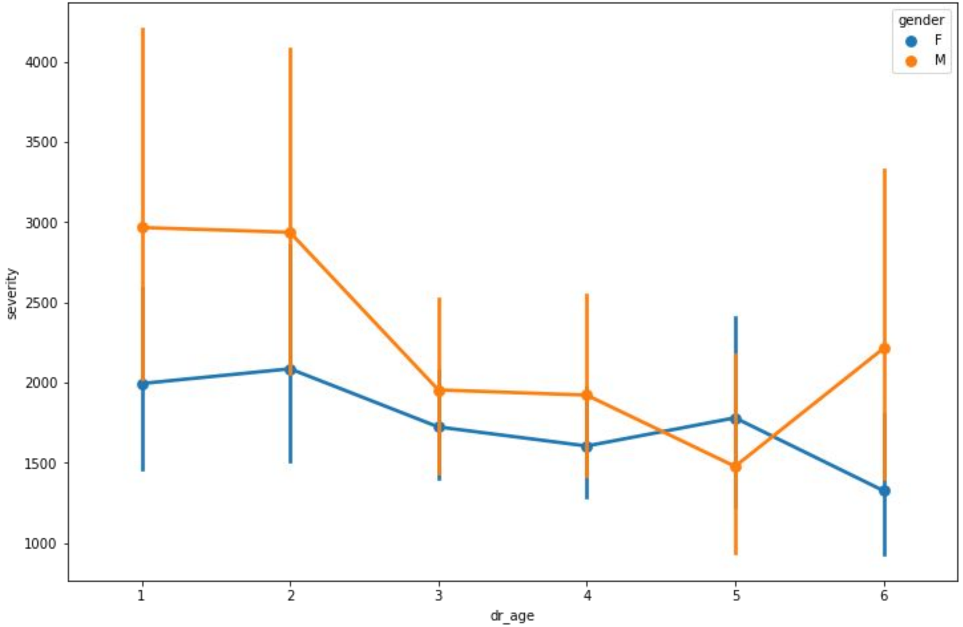
1. **Pure premium** = claim cost / exposure
2. **Severity (average cost)** = claim cost / claim count
3. **Frequency** = claim count / exposure

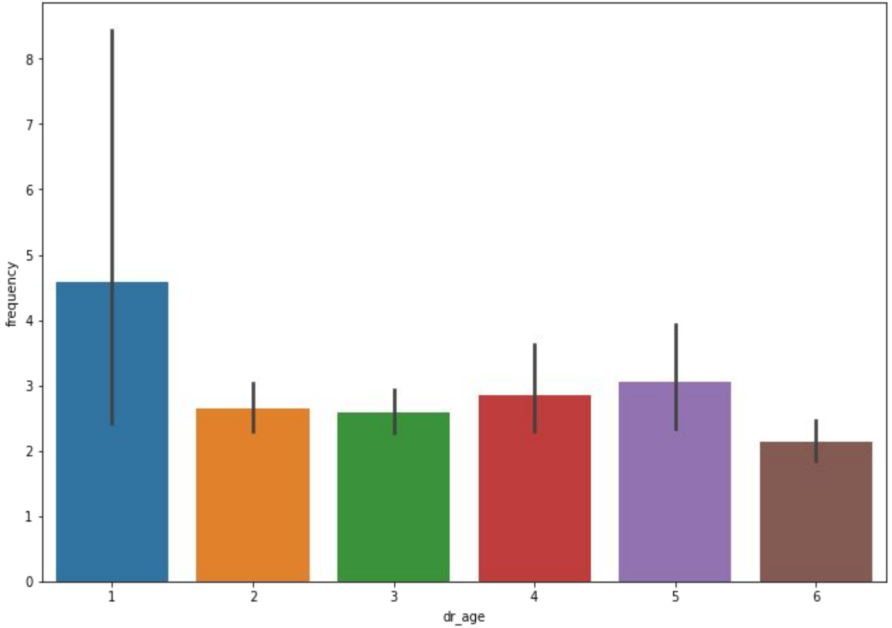


Correlation matrix

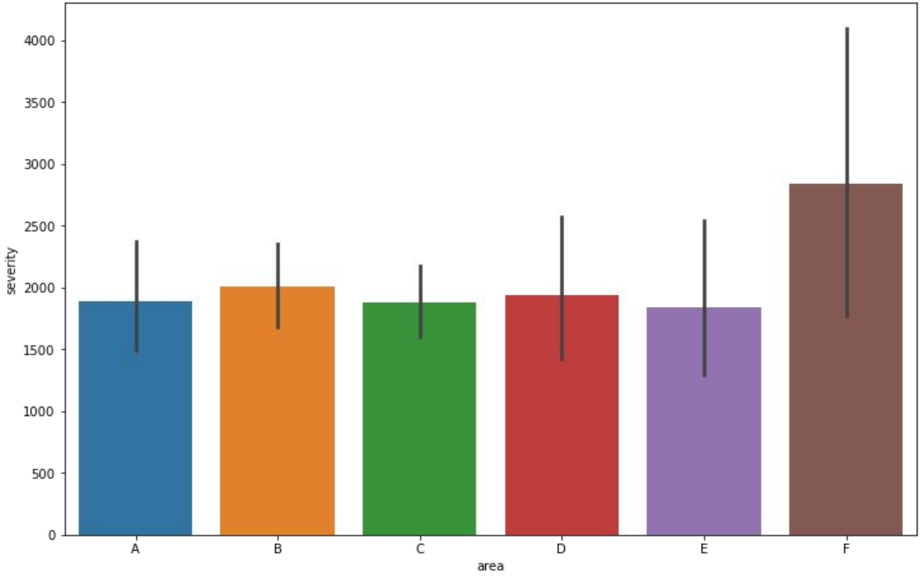


Interesting Visualizations

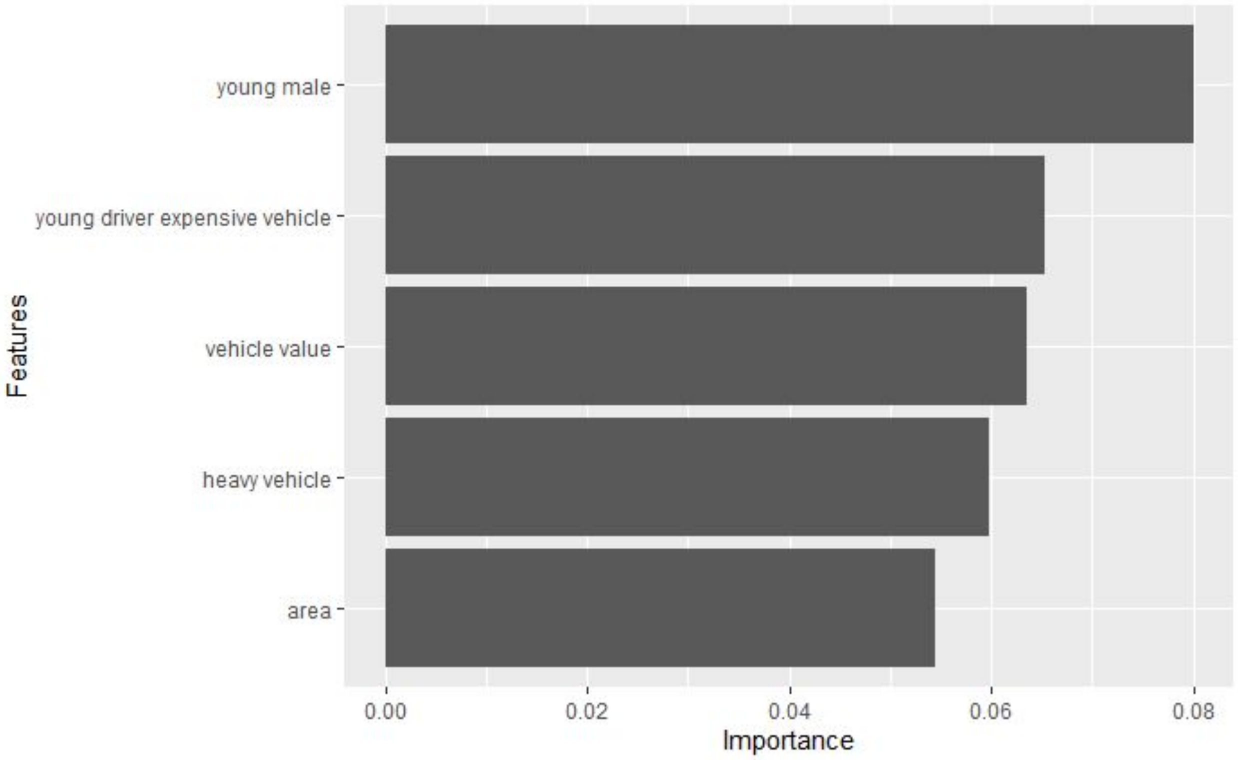




Interesting Visualizations



Important variables explaining severity



Creating a Risk-Prediction Model

What Makes a Good Ratemaking Model?

A good ratemaking model should:

- <u>Identify and segment risk</u>
- <u>Be informative</u>
- <u>Be understandable</u>

“Complex” models (boosting, neural nets, ensembles, stacking) struggle with:

- Interpretability
- Regulatory constraints
- Stability of model over time

What we started with: <u>Complexity</u>

- No transparency to model results
- Long computation time
- Very sensitive, hard to tune

What we ended with: <u>Practicality</u>

Simple linear models + domain knowledge

Satisfies all three criteria for a good risk model

Our Approach

Two components:

1. Base ratemaking model
2. Relativities to boost risk segmentation

Base Ratemaking Model Ensemble:

- **Indicator Model** (Logistic)
  - Will the insured file a claim at all?
- **Frequency Model** (Poisson)
  - How many claims will there be, if any?
- **Severity Model** (Gamma)
  - How much will an insured's claim cost?

Why not have a single pure premium model?

With 3 separate base models, we can:

- Identify response-specific relationships
- Easily diagnose segmentation issues
- Better performance, same interpretability

Calculating our Base Predicted Loss:

Base Loss = Ind. x Sev. x Freq. x Exposure

Improving Risk Identification with Relativities

Problem: Linear models find expected cost, but tend to underestimate large risks

Historical claims data omits very risky customers due to underwriting

Solution: Use known risk factors

Base predictions multiplied by relativities (risk multipliers)

Relativities allows us to:

- Account for selection bias in claims data
- Better classify risk (score gain: **0.17** to **0.22)**

We calculated relativities using loss ratio and count ratio, based on risk segmentation score

Some relativities used in our model:

>1 : Higher risk, <1: Lower risk

- Driver Age 1: **1.794**
- Driver Age 5: **0.569**
- Male: **1.559**
- Area F: **2.88**
- Area D: **0.745**

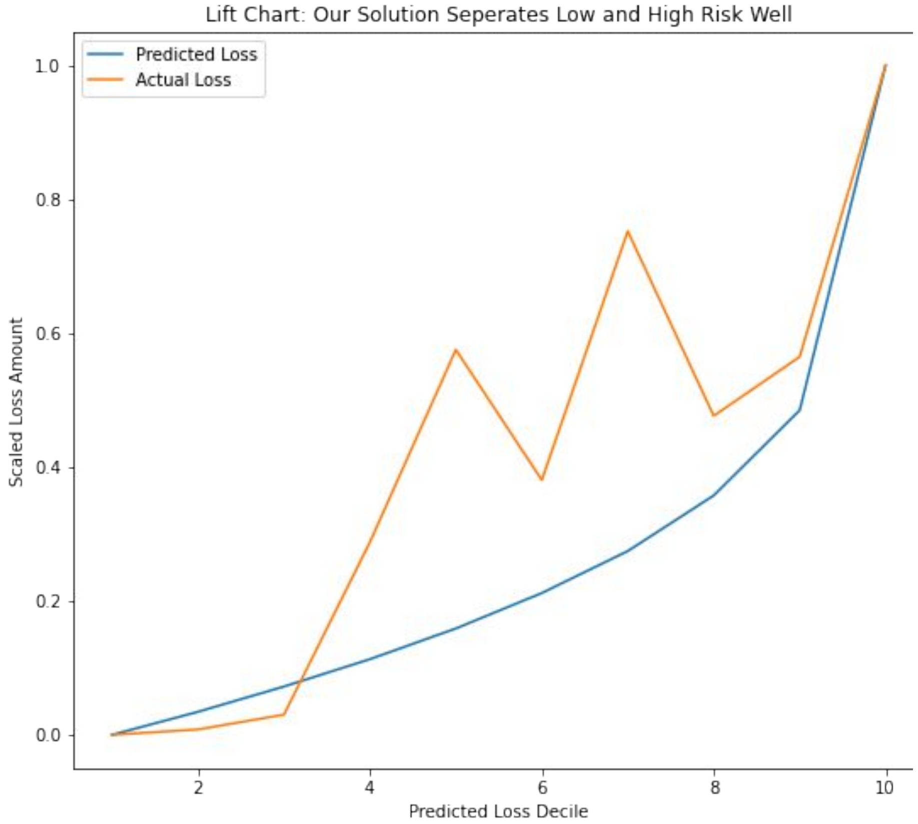
Validating our Model

“Lift” chart can be used to show risk differentiation in a model

Loss scaled to show relative risk between policies (0 = least risky, 1= most risky)

Large difference between 1st and 10th actual loss quantiles show good risk segmentation

Some nonmonotonicity that could be improved



Where can we go from here?

Future Investigation

1. Incorporating credibility into relativity calculation
2. Use mixed models to better capture “area” effect
3. Add more models tuned for different loss amounts
4. Change linear model loss to optimize risk segmentation
5. Examine stability of solution under changing conditions

Questions about Data

- **Veh_age and Dr_age**

  Veh_age takes values 1 (youngest) to 4 (oldest) and Dr_age takes value from 1 (young) to 6 (old).

  How were the veh_age and dr_age variables binned?

- **Area**

  Area takes values A to F. Is the Area categorized based on zip-code or areas given a risk rating or in some other way?

Potentially useful variables

- Education level
- Job type
- Marriage status
- Area type ex- urban, rural
- Number of children
- If driver is a single parent

Conclusion
and
Q&A