Generalized Linear Models

-- General Session Modeling and Big Data

June 28, 2021

Hao Li, ACAS MAAA FRM

ISO / Verisk Analytics

SERVE | ADD VALUE | INNOVATE

HAO LI, ACAS, MAAA, FRM

Lead Data Scientist, Insurance Analytics
ISO / Verisk Analytics

BACKGROUND

Hao Li is a Lead Data Scientist with Verisk Analytics, based in Buffalo Grove, IL, leading a team of data scientists focusing on predictive modeling in personal line pricing. Hao has over 9 years experience working in the banking and the insurance industry with a focus on predictive modeling and actuarial pricing.

PROFESSIONAL DESIGNATIONS AND ACTIVITIES

Hao Li is an Associate of the Casualty Actuarial Society (ACAS), a Member of the American Academy of Actuaries (MAAA) and a Financial Risk Manager of Global Association of Risk Professionals (FRM).

EDUCATION

- Master of Probability and Statistics, Auburn University, USA
- Master of Finance, Auburn University, USA
- BSc in Management, Shanghai University of Engineering Science, China

SELECTED EXPERIENCES

RAPA Symbol V2.0

- Developed the Other-Than-Collision coverage models for Risk Analyzer Personal Auto Symbol

VINhistory Score

- Led the development of by-coverage VINhistory score for Personal Auto to further improve rating efficiency by leveraging history of vehicles from prior and current owners

RAHO Environmental V2.1

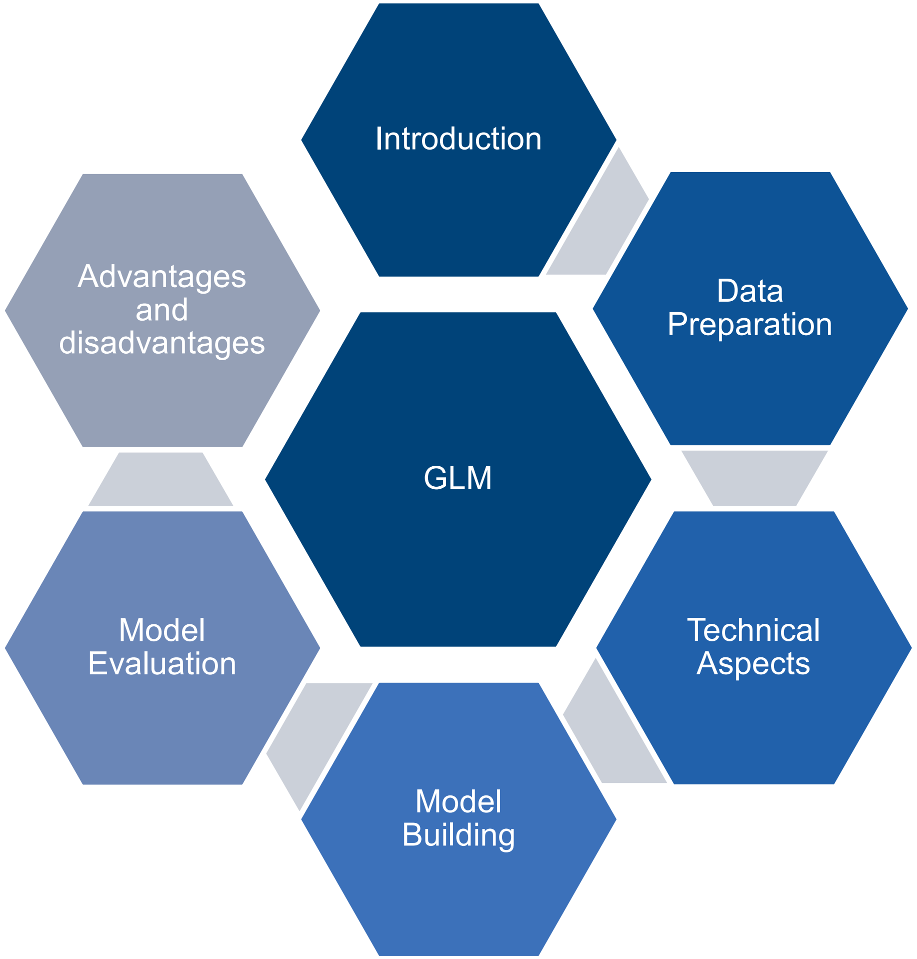
- Currently leading the effort to develop by-peril loss cost models for Home Owners insurance using environmental information – weather, elevation, road features, census, business points, distance to coast, and etc., in presence of standard rating variables.
- Leading the effort to refresh/rebuild pipeline for major 3rd-party data to support a suite of products offered by Verisk

INTERESTS AND EXPERTISE

- Analytics | Data Science | Actuarial
- Underwriting
- Risk Segmentation | Risk Classification
- Econometrics | Risk | Finance
- Product Research | Product Development

Contents

- Introduction
- Data Preparation
- Technical Aspects of the GLM
- Model Building
- Model Evaluation
- Advantages and disadvantages

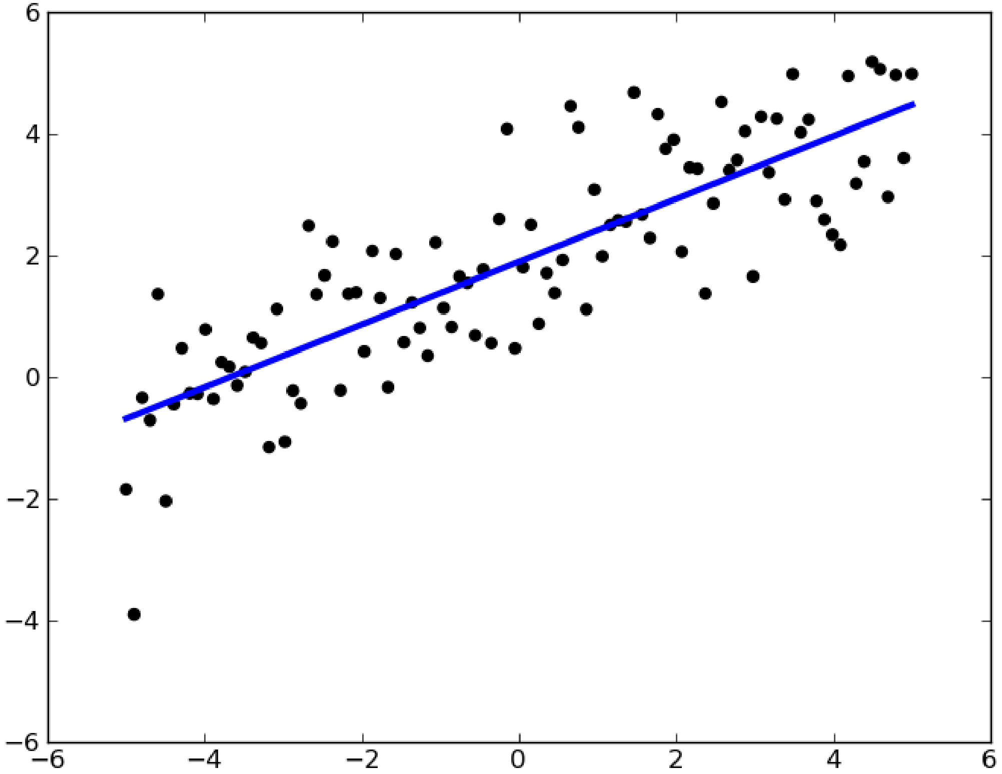


Introduction

- Generalized Linear Model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

$$\mu = E[Y]$$



* Sources: scikit-learn generalized linear model, ordinary least squares

❑ What exactly the problem are we solving?

- ❖ Lines of business?
- ❖ Rate-making or reserving or underwriting or claim analytics?

Data Preparation

- Data scope
- Target
- Predictors
  - Insurance data: policy/insured characteristics
  - Non-insurance data
- Treatment of missing values and outliers

❑ What's the data at hand in general?

❑ What is target?

  ❖ Depending on the business problem, whether the target is chosen properly?

❑ What are the predictors?

  ❖ Description of the predictors

  ❖ Any rationality certain predictors need to be considered?

❑ Are there any missing values or outliers existing?

  ❖ If yes, what was the treatment?

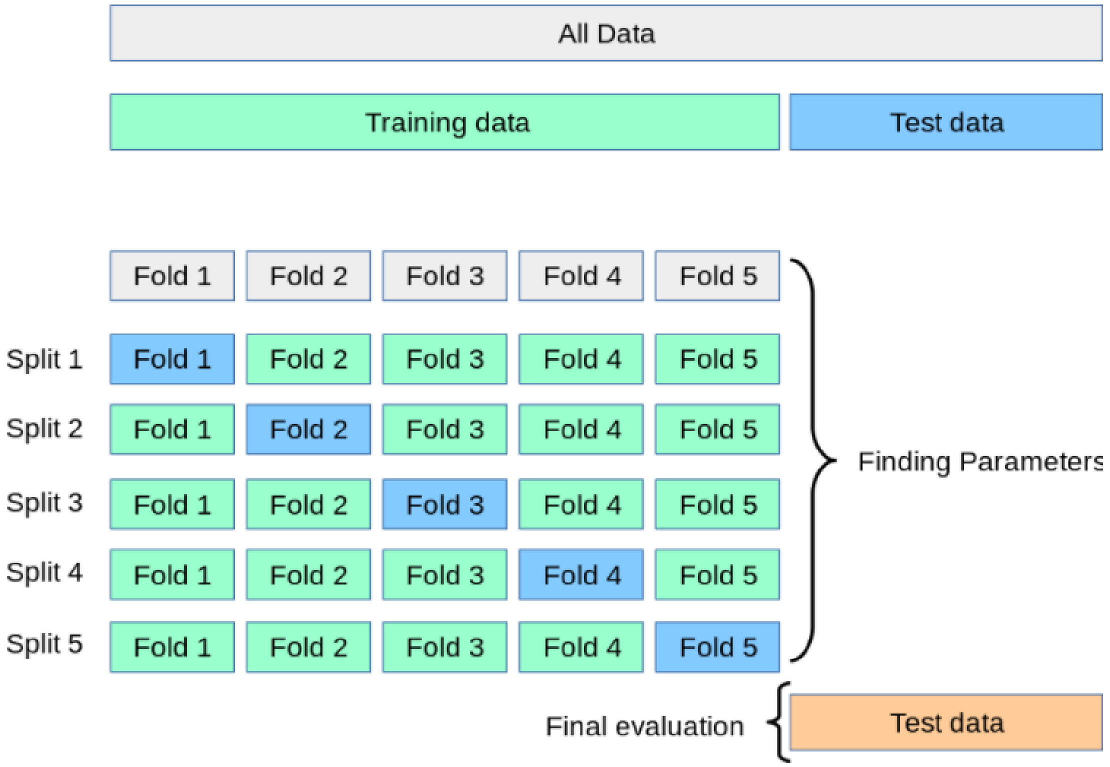
Technical Aspects of GLM

- Distribution
  - Frequency: Negative binomial (a more general case of Poisson)
  - Severity: Gamma
  - Pure premium: Tweedie
- Link function
- Weight
- Offset
  - Some components of the rating plan held constant while analysts are updating the signals from others

❑ What distribution should be used to work with the target?
❑ What are the proper link functions?
❑ Is there a weight needed?
❑ Under what situation, an offset should be considered?

Model Building

- Data split
  - Train / Validation / Test
  - Cross-validation



All Data

Training data | Test data

Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 1: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 2: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 3: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 4: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 5: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Finding Parameters

Final evaluation: Test data

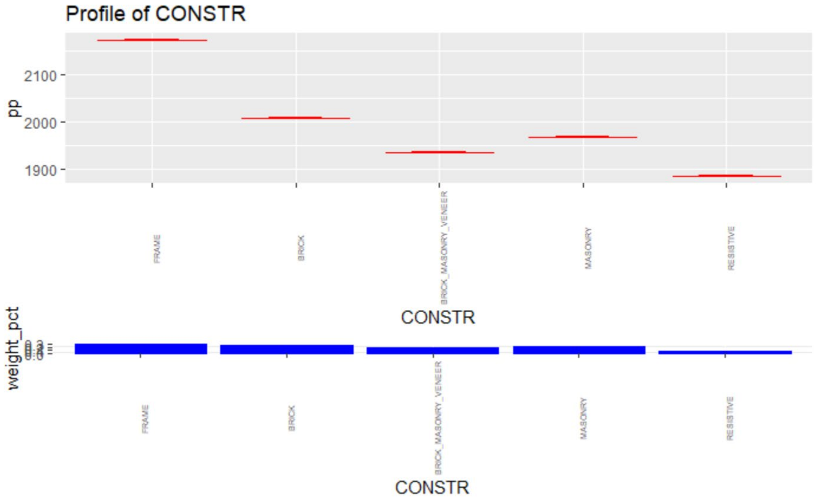
❑ How is the data split handled?

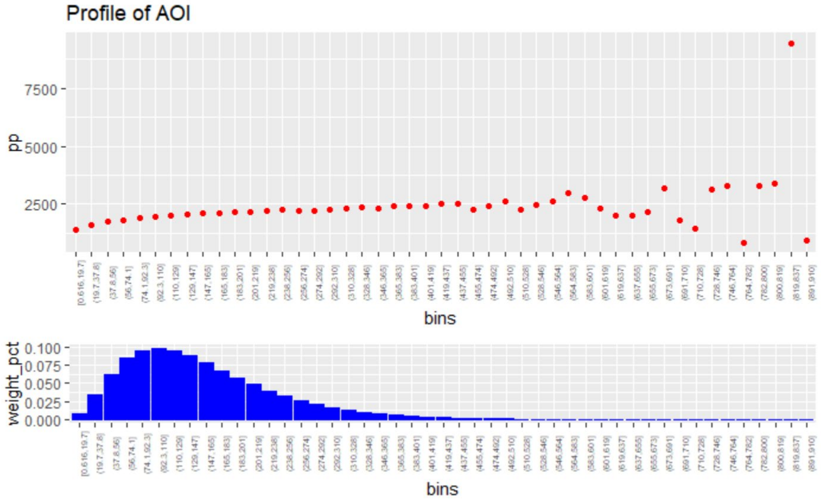
- ❖ What’s the portion of train, validation and test?
- ❖ Is cross validation used? What’s the fold?

* Sources: scikit-learn cross-validation: evaluating estimator performance

Model Building

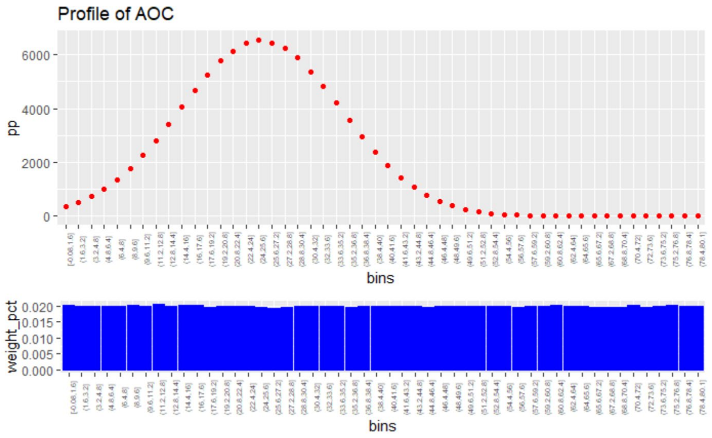
- Explanatory Data Analysis

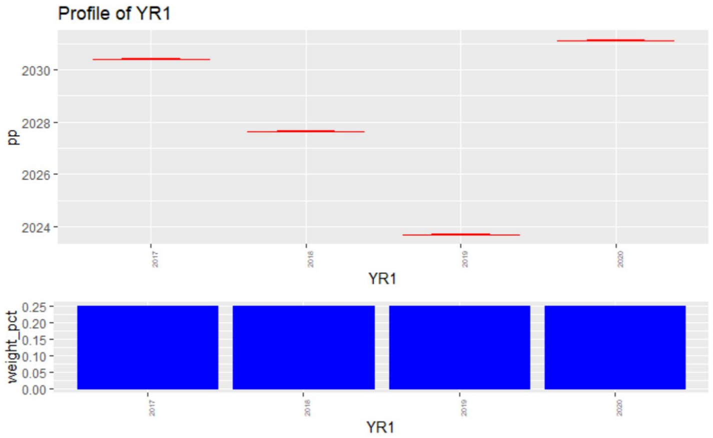




❑ Any insight gained form the explanatory data analysis?

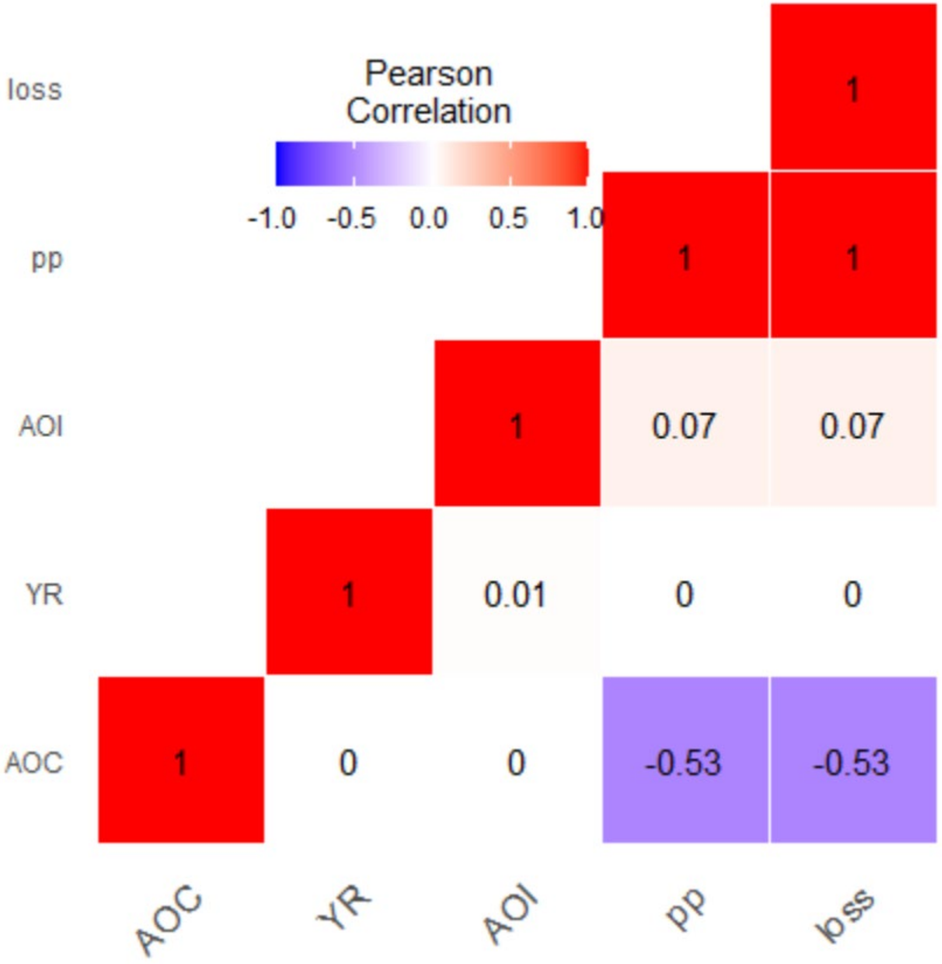
- ❖ Any visual relationship between the target and features?
- ❖ What level is used as base line for categorical variables?
- ❖ Any further binning can be done on individual categorical variables?
- ❖ What potential transformation can be used for continuous variables?





Model Building

- Correlation/Association

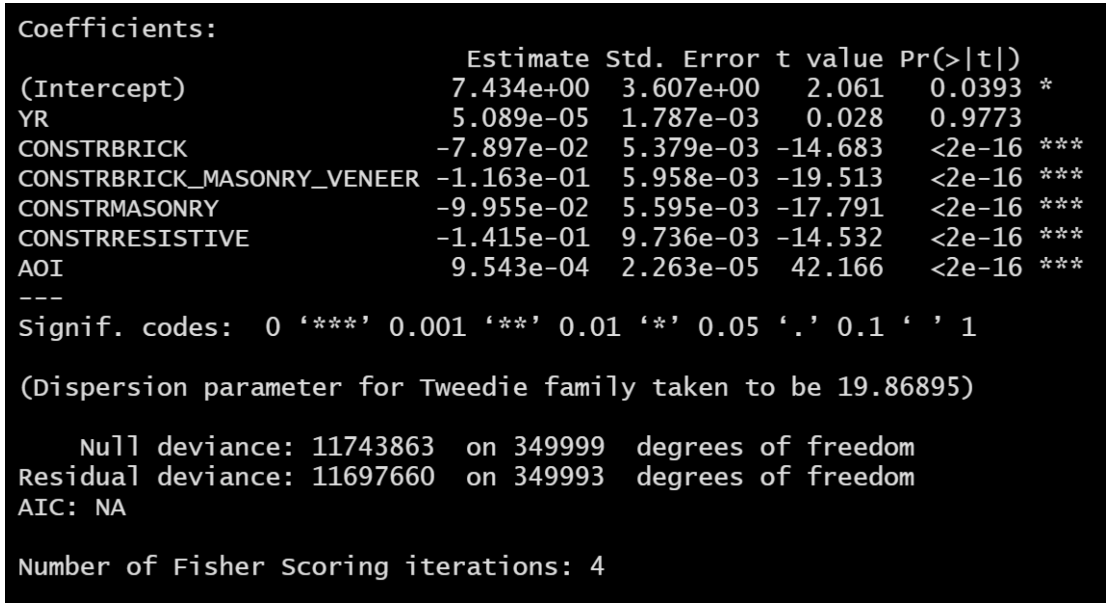


❑ Is correlation or association evaluated against groups of variables?

- ❖ Can we identify highly correlated variables?
- ❖ Is there multicollinearity among features?

Model Evaluation

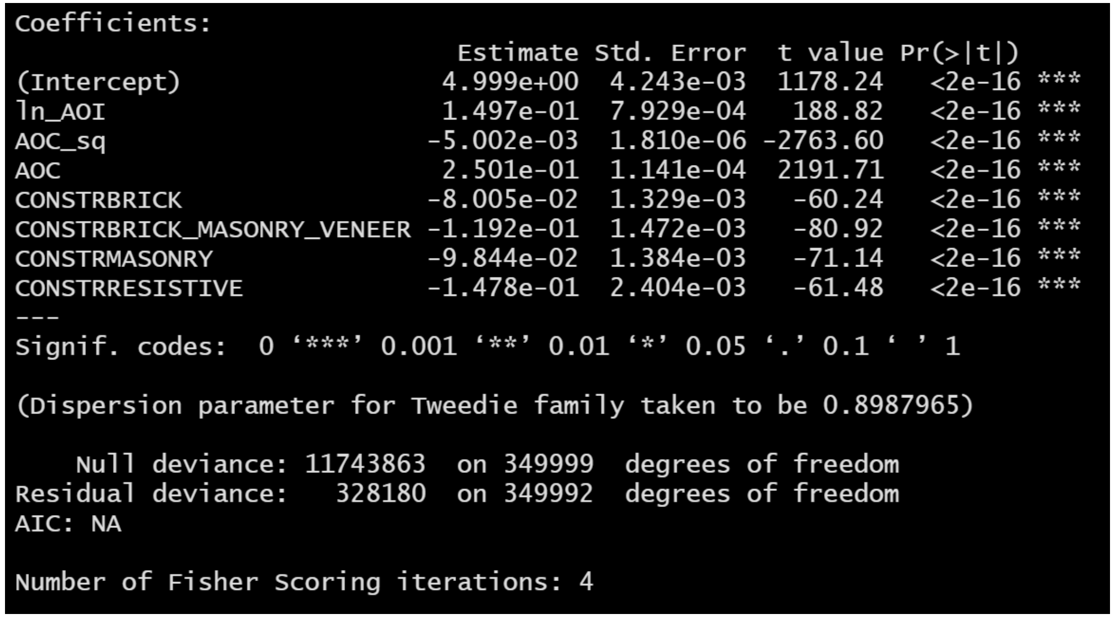
- Coefficient table



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 7.434e+00  3.607e+00   2.061   0.0393 *  
YR                          5.089e-05  1.787e-03   0.028   0.9773    
CONSTRBRICK                -7.897e-02  5.379e-03 -14.683   <2e-16 ***
CONSTRBRICK_MASONRY_VENEER -1.163e-01  5.958e-03 -19.513   <2e-16 ***
CONSTRMASONRY              -9.955e-02  5.595e-03 -17.791   <2e-16 ***
CONSTRRESISTIVE            -1.415e-01  9.736e-03 -14.532   <2e-16 ***
AOI                         9.543e-04  2.263e-05  42.166   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 19.86895)

    Null deviance: 11743863  on 349999  degrees of freedom
Residual deviance: 11697660  on 349993  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

❑ What variables are included in the model?

- ❖ What’s the magnitude and direction of coefficient?
- ❖ Are they reasonable?
- ❖ Are all the coefficients statistically significant?

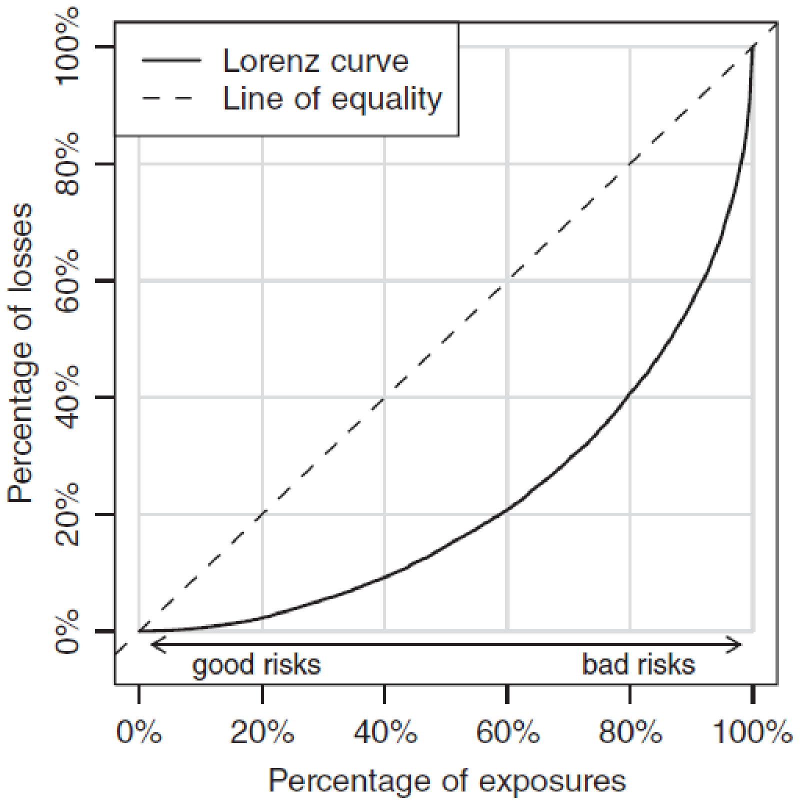


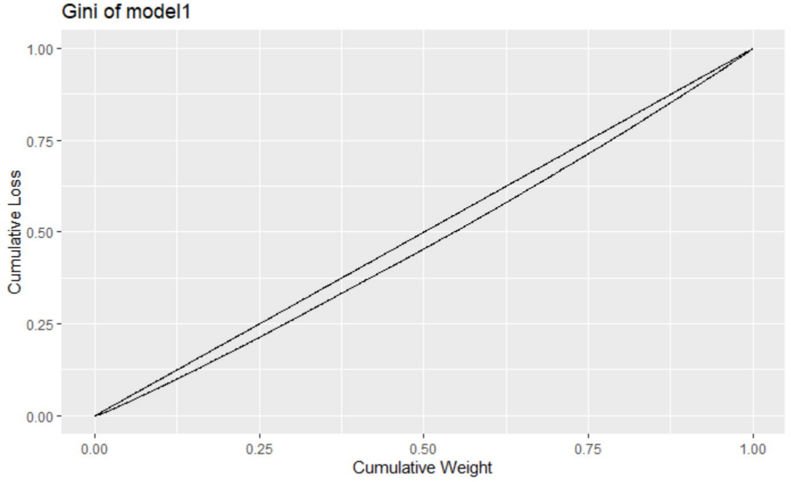
```
Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 4.999e+00  4.243e-03  1178.24   <2e-16 ***
ln_AOI                      1.497e-01  7.929e-04   188.82   <2e-16 ***
AOC_sq                     -5.002e-03  1.810e-06 -2763.60   <2e-16 ***
AOC                         2.501e-01  1.141e-04  2191.71   <2e-16 ***
CONSTRBRICK                -8.005e-02  1.329e-03   -60.24   <2e-16 ***
CONSTRBRICK_MASONRY_VENEER -1.192e-01  1.472e-03   -80.92   <2e-16 ***
CONSTRMASONRY              -9.844e-02  1.384e-03   -71.14   <2e-16 ***
CONSTRRESISTIVE            -1.478e-01  2.404e-03   -61.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 0.8987965)

    Null deviance: 11743863  on 349999  degrees of freedom
Residual deviance:   328180  on 349992  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

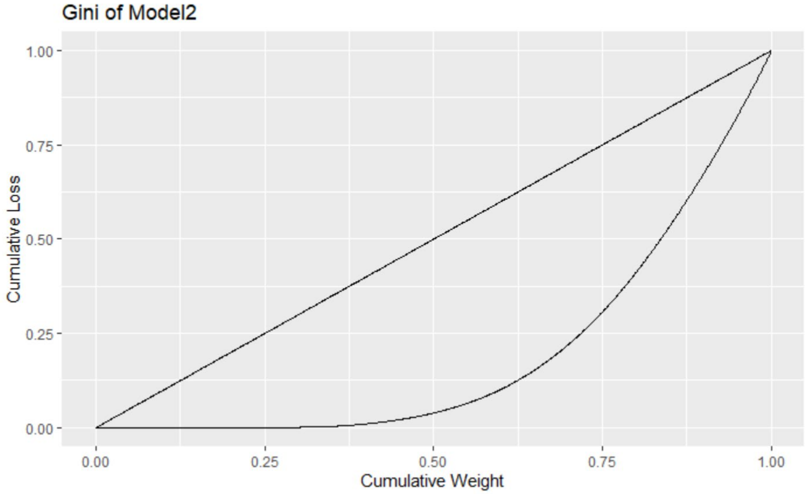
Model Evaluation

- Gini
  - Index = 2 * area between equality and Lorenz curve

❑ Does model performance improve between different models?



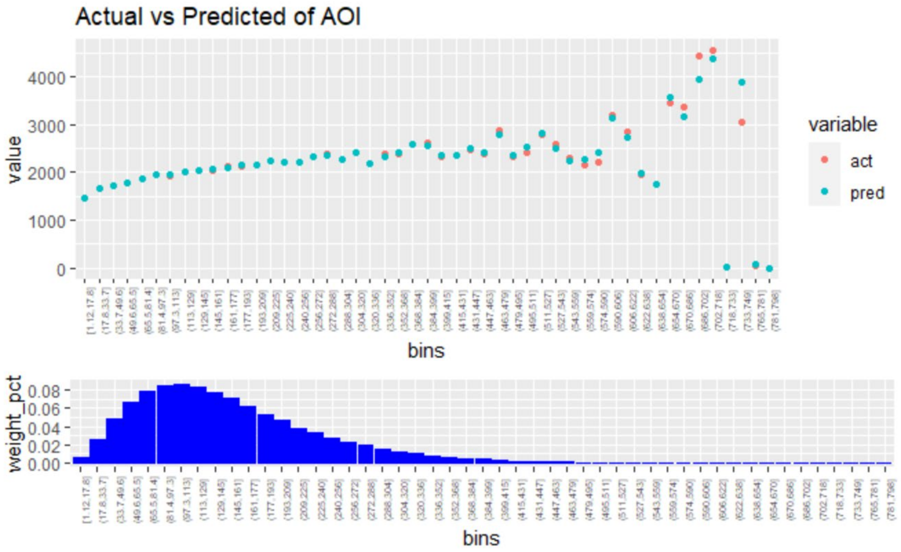


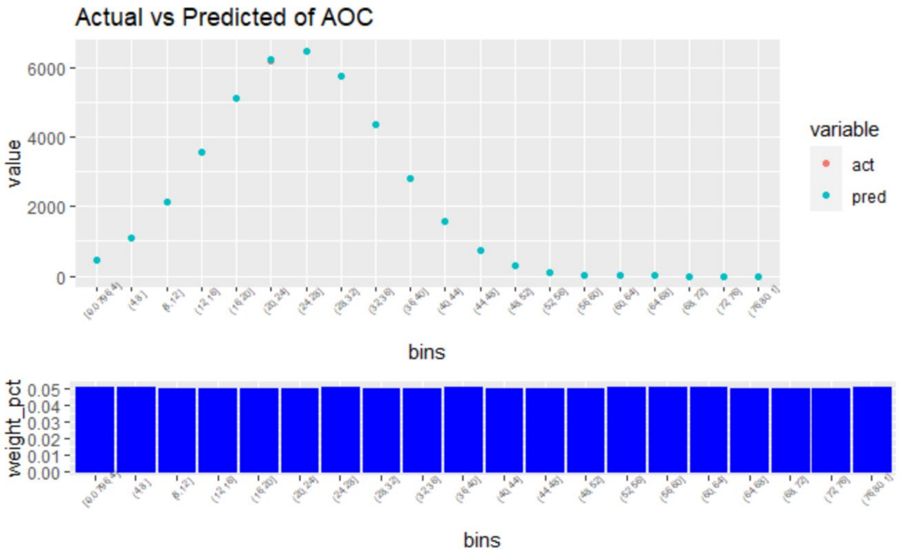


* Sources: scikit-learn cross-validation: evaluating estimator performance

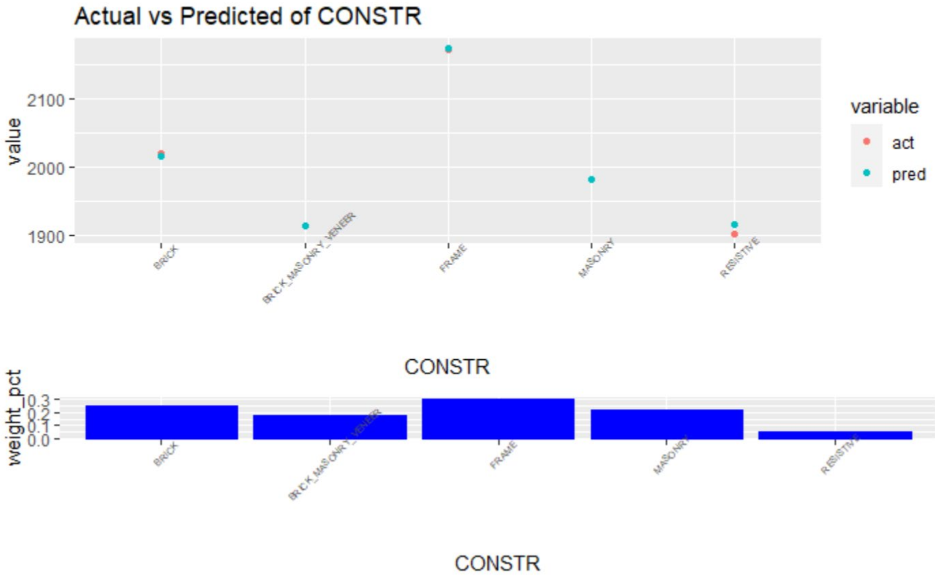
Model Evaluation

- Actual vs Predicted



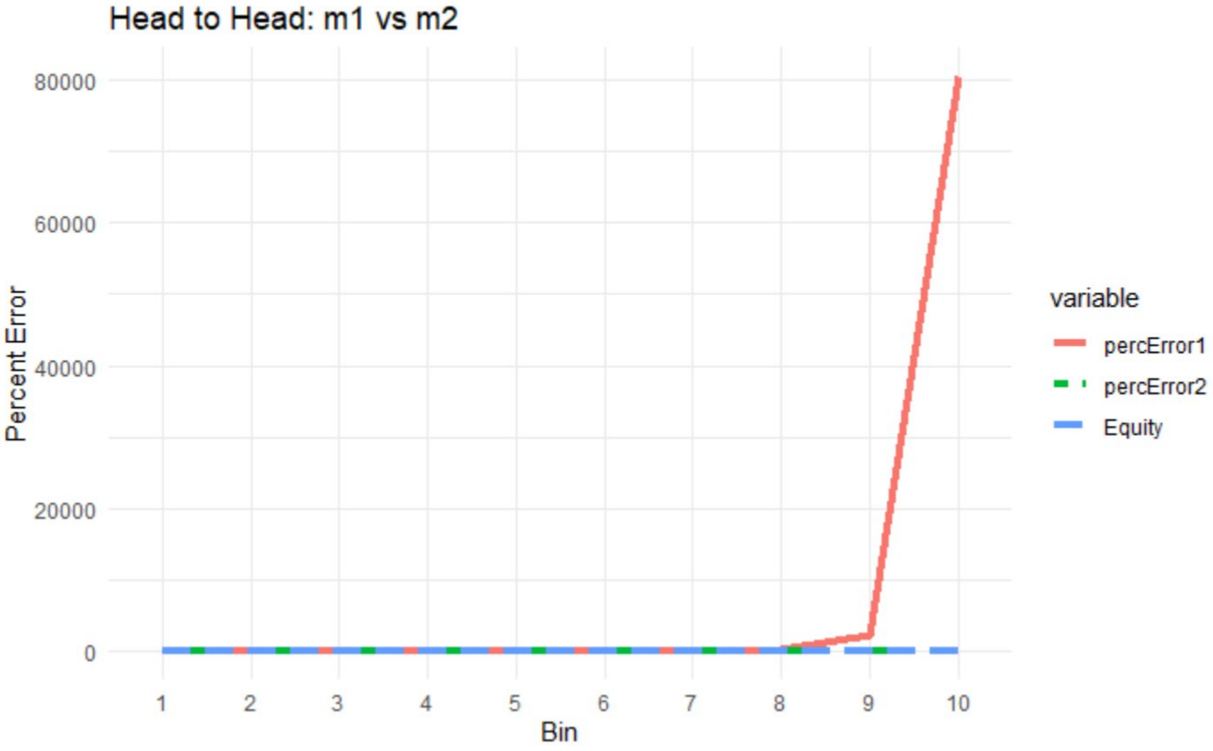


❑ Is prediction by individual variable close to actual value?



Model Evaluation

- Head-to-head (double lift chart)



❑ Which model produces a prediction close to the actual between two competing candidates?

   ❖ This is similar to double lift chart

   ❖ A single error metric can be derived to show which model is overall better than the other one

Model Evaluation

- Nested model comparison

$$F = \frac{D_S - D_B}{\#\ of\ added\ parameters\ * \ \widehat{\emptyset}_S}$$

$$where\ D = 2 \times (ll_{saturated} - ll_{model})$$

- Penalized measure of fit

$$AIC = -2 \times ll + 2p$$

$$BIC = -2 \times ll + p \cdot log(n)$$

- Residual based analysis
  - Response residual
  - Working residual
  - Pearson residual
  - Deviance residual

❑ When to use F test for comparing two models?

❖ When F is larger than the critical value, we conclude that there is significant difference between big and small model

❑ When can AIC and BIC be useful?

❑ Do we see residuals showing random pattern, constant variance and normally distributed?

❖ Only useful for continuous distribution

Advantage and Disadvantage of GLM

- Advantage:
  - Help to understand associative relationship between features and target
  - When project requires a strong interpretability from the models

- Disadvantage:
  - Prediction accuracy due to constraint of “linear” framework
  - Unstable result when handling features with multicollinearity and thin data
  - Requires significant iteration and modeler’s intervention to improve model

References

- Goldburd, Mark, Anand Khare, Dan Tevet, and Dmitriy Guller. 2020. Generalized Linear Models for Insurance Rating, 2nd Ed. Arlington VA, Casualty Actuarial Society
- de Jone, Piet and Gillian Z. Heller. 2008. Generalized Linear Models for Insurance Data. New York: Cambridge University Press
- Faraway, Julian 2005. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Boca Raton FL: Chapman & Hall/CRC
- Haste, Trevor, Robert Tibshirani, and J.H. Friedman 2017, 2nd Ed. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer

Verisk™

SERVE | ADD VALUE | INNOVATE