AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

25 Jan. 2021

AIG Interview
Presentation:
Motor
Third-Party
Liability Claims
Analysis and
Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset Data Visualization

Model & Prediction

Poisson GLI

Model & Prediction

Poisson GLM
Poisson Lasso & Rid
Gradient Boosting

Final Validation

0 & A

Link to complete code and analysis: https://yipeichan.github.io/claims.html

AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset

Data Visualization

Model &

Poisson GLM

Model &

Prediction
Poisson GLM
Poisson Lasso & I

Poisson Lasso & Ridge Gradient Boosting Model

Final Validation

Project Concept

▶ Problem to solve :

How can we predict the number of claims a policyholder would file, given his age, his car brand, and so on?

- My approach to solve the problem :
 - 1. Explore the structure and properties of the dataset
 - 2. Choose the proper models to answer the question
- Methodology : After exploring the data with visualizations,
 - 1. Generalized Poisson Linear Model
 - 2. Poisson Lasso Regression, Poisson Ridge Regression
 - 3. Gradient Boosting Model
- Goals achieved by this project :
 - Explored relationships between the risk factors and ranked the influences of risk factors on claim numbers
 - 2. Investigated the efficacy of using modern machine learning algorithms to do P&C ratemaking
 - 3. Make your hiring decision easier!

AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset

Data Visualization

Model & Prediction

Model &

Poisson GLM
Poisson Lasso & Ridge
Gradient Boosting
Model

Final Validation



Data Exploration- The Dataset

- CASdatasets Package : Proposed by Christophe Dutang ¹ on OpenML
- ▶ Used in this study is freMTPL2freq dataset :
 - 1. Risk features were collected from motor third-part liability policies in France
 - 2. 678,013 samples, 12 explanatory variables

Variable Name	Description	Key
IDpol	Policy ID	(link with the claims dataset)
ClaimNb	Number of claims during the exposure period	
Exposure	Period of exposure (in years)	
VehPower	Power of the car	
VehAge	Vehicle age (in years)	
DrivAge	Driver age (in years)	
BonusMalus	Bonus/malus, between 50 and 350	<100: bonus; >100: malus in France
VehBrand	Car brand	Unknown categories
VehGas	Car gas	Diesel or regular
Area	Density value of the city where the car driver lives in	"A" for rural to "F" for urban centre
Density	Density of inhabitants of the city where the car driver lives in	Number of inhabitants per square-kilometer
Region	Policy region in France	

AIG Interview
Presentation:
Motor
Third-Party
Liability Claims
Analysis and
Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset Data Visualization

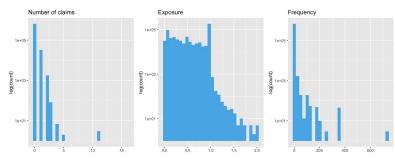
Model & Prediction

Model &

Poisson GLM
Poisson Lasso & Ridg
Gradient Boosting
Model

Final Validation

- ► Among the 678,013 policies, there were 34,060 filed claims, i.e. 5.02% notified claims.
- Potential Problems :
 - 1. Mean should equal to Variance in Poisson distribution ⇒ Use Negative binomial if Overdispersed
 - 2. More 0s than are expected in Poisson regression? ⇒ Incorporate the logit model for predicting excess 0s
 - 3. Varied exposure periods (observations not comparable)
 - ⇒ Add offset of Exposure term to the model



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration Data Visualization

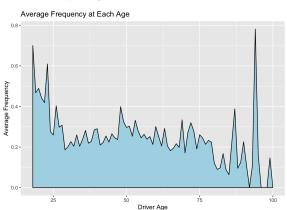
Model & Prediction

Model & Prediction

Gradient Boosting

Final Validation

- Exposure : the duration of the insurance coverage
- Claim frequency: claim count per unit of exposure
- Did driver age influence frequency?
 - 1. The highest mean frequency happens at age 94
 - 2. Drivers between age 18 to 23 tends to have higher mean frequency



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

Data Visualization

Prediction
Poisson GLM

Model &

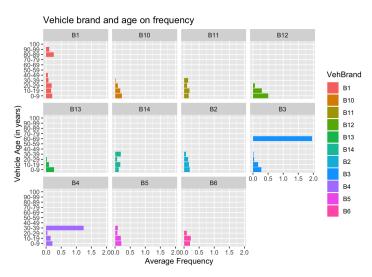
Prediction

Poisson Lasso & Ridge Gradient Boosting Model

Final Validation



Did vehicle brand and age influence frequency?



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration
The Dataset
Data Visualization

Model & Prediction

Poisson GL

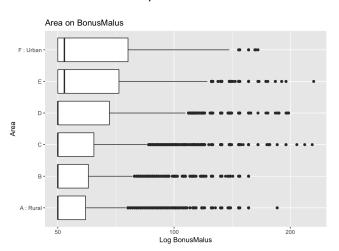
Model & Prediction

Poisson GLM
Poisson Lasso & Ridge
Gradient Boosting
Model

Final Validation

O & A

▶ What is the relationship between area and bonus-malus?



AIG Interview
Presentation:
Motor
Third-Party
Liability Claims
Analysis and
Prediction

Yi-Pei Chan

Project Concept

Data Exploration
The Dataset
Data Visualization

Model & Prediction

Model & Prediction

Poisson GLM
Poisson Lasso & Rid
Gradient Boosting
Model

Final Validation

Model and Prediction - Poisson GLM

Before training the models, randomly select 30% of the data and set aside as testing set to find the best fitting model

► Poisson GLM Model

```
glm(formula = ClaimNb ~ VehPower + VehAge + DrivAge + BonusMalus +
   VehBrand + VehGas + Density + Region + Area, family = "poisson",
   data = data[(data$data == "train"), ], offset = log(Exposure))
```

- Statistically significant variables (Signif. level 1%):
 VehAge, DrivAge, BonusMalus, VehPower, Density, etc.
- Overdispersion Test
 Overdispersion test

```
data: poissonglm

z = 3.9191, p-value = 4.444e-05
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0.243314
```

- 1. Small p-value :The test confirms the overdispersion
- 2. The alpha value very close to zero: Overdispersion may not be a serious concern here

AIG Interview
Presentation:
Motor
Third-Party
Liability Claims
Analysis and
Prediction

Yi-Pei Chan

Project Concept

Data Exploration
The Dataset
Data Visualization

Model & Prediction

oisson GLIVI

Model & Prediction

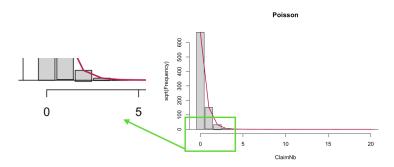
Poisson GLM
Poisson Lasso & Ridge
Gradient Boosting
Model

Final Validation



Model and Prediction - Poisson GLM

Hanging rootogram : Only 2 count is a little under predicted



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset

Data Visualization

Model & Prediction

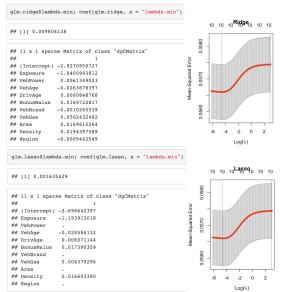
Model &

Poisson GLM
Poisson Lasso & Ridg
Gradient Boosting

Final Validation

O & A

Model & Prediction - Poisson Lasso & Ridge Regression



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration
The Dataset
Data Visualization

Model & Prediction

Poisson GLIV

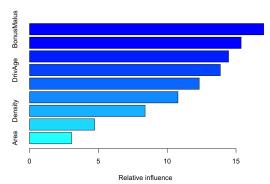
Model & Prediction

Poisson Lasso & Ridge Gradient Boosting Model

Final Validation



Model & Prediction - Gradient Boosting Model



var rel.inf
BonusMalus 17.014808
Region 15.372979
VehAge 14.459134
DrivAge 13.862481
VehBrand 12.328304
VehPower 10.782009
Density 8.396521
VehGas 4.728894
Area 3.054871

AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset Data Visualization

Model & Prediction

Model &

Poisson GLM
Poisson Lasso & F

Final Validation

O & A

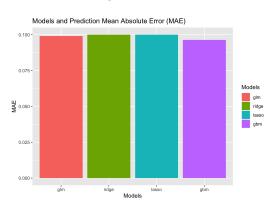
Model



Final Validation

Use the test set to find the best fitting model

- The claim number prediction MAE for test set with
 - 1. Poisson GLM: 0.09905573
 - 2. Poisson Ridge GLM: 0.09988506
 - 3. Poisson Lasso GLM: 0.09996999
 - 4. Gradient Boosting Model: 0.09630762



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset

Data Visualization

Model & Prediction

Model &

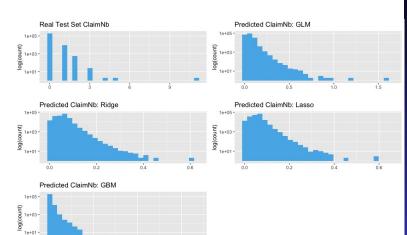
Poisson GLM Poisson Lasso & R

Final Validation



Final Validation

Evaluation of the Predicted Number of Claims in the Test Set



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset

Data Visualization

Model & Prediction

Model &

Poisson GLM Poisson Lasso & Ric

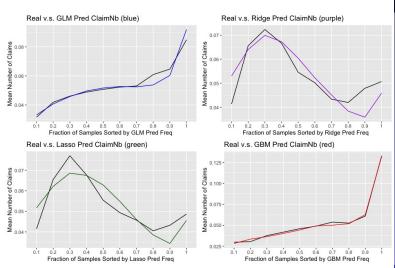
Final Validation

00. 1



Final Validation

Evaluation of the Predicted Number of Claims in the Test Set



AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset Data Visualization

Model & Prediction

Poisson GLN

Model & Prediction

Poisson GLM
Poisson Lasso & Ridge
Gradient Boosting
Model

Final Validation





Link to complete code:

https://yipeichan.github.io/claims.html

AIG Interview Presentation: Motor Third-Party Liability Claims Analysis and Prediction

Yi-Pei Chan

Project Concept

Data Exploration

The Dataset Data Visualization

Model & Prediction

Poisson GLN

Model & Prediction

Poisson GLM
Poisson Lasso & Ridg
Gradient Boosting

Final Validation