Home Insurance Loss Prediction

Cohort A Team 7

Qiaoling Huang, Shihan Li, Ziqin Ma, Chenran Peng, Elmira Ushirova

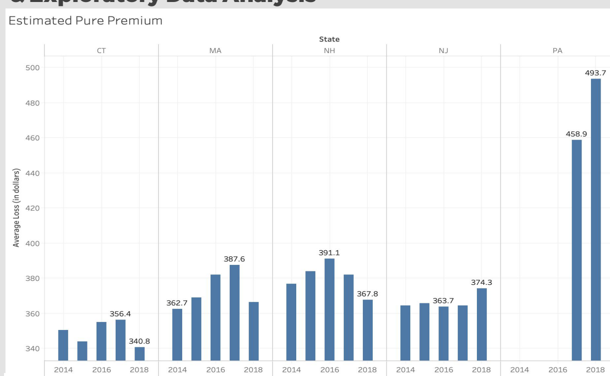
Introduction and Problem Statement

Recommended and required in some regions, house insurance covers losses and damages to an individual's house and assets in the house. We would like to explore different factors that are associated with risks of homeowner insurance and come up with new efficient variables and a precise model that can predict house damage losses at zip code level in the USA.

Datasets sources & Exploratory Data Analysis

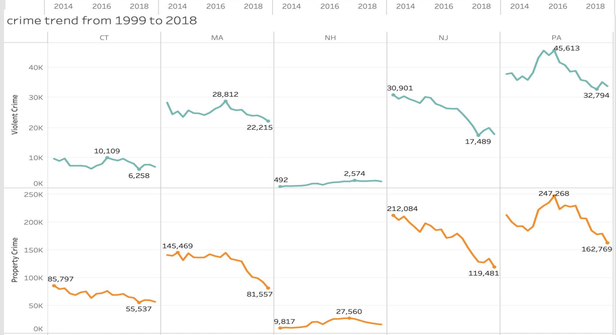
Insurance losses data

- New England
- 2013 – 2018
- Actual Loss
- Predicted loss
- Earned Exposure
- Population
- Number of properties for each zip code



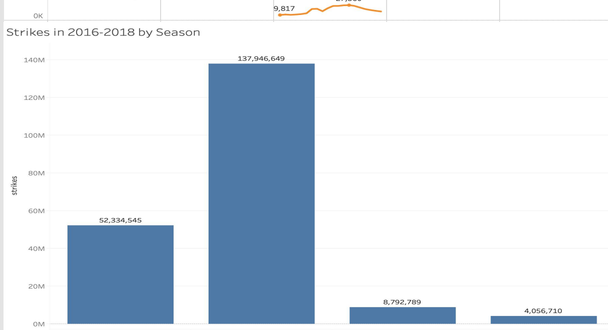
Crime data

- 1999 – 2018
- Different crime categories
- Violent crimes
- Property crimes



Lightning data

- 1987 - 2019
- Number of strikes(daily)
- Geo-points



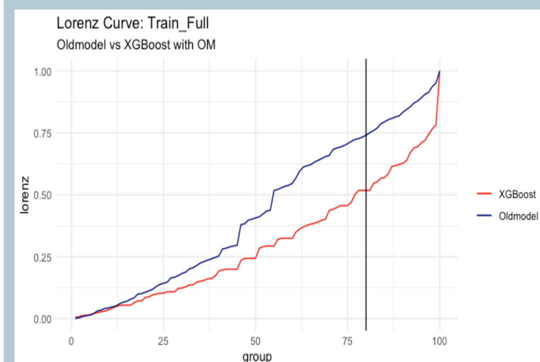
Methodology

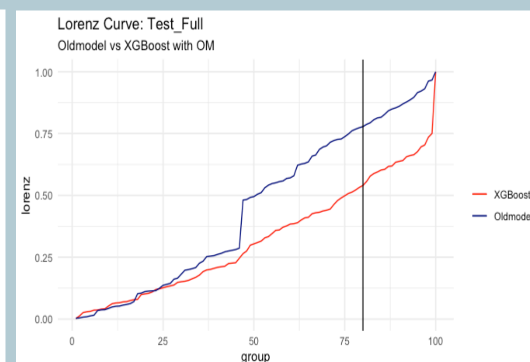
NA Imputation → Feature Engineering → Sample Design → Feature Selection

Modeling

4 basic models	Train & Fit	Old model inclusion
• GLM, Fandom Forest, XGBoost, Neural Network	• Train on 1 dataset (Train_reduced) • Fit on other 4 datasets	• Train models with oldmodel prediction or not

Model	Distribution	Oldmodel	AUC: NCAT/ee				
			TRAIN_Reduced	TEST_Reduced	Train_Full	Test_Full	Validation
Oldmodel	default	Y	0.572	0.61	0.598	0.554	0.641
XGBoost	default	Y	0.735	0.607	0.699	0.679	**0.608**
Oldmodel	default	N	0.5	0.5	0.5	0.5	0.5
XGBoost	default	N	0.777	**0.605**	0.668	0.672	**0.554**





Results

After running 4 models, XGBoost preforms the best in both cases.

Without Oldmodel:
- Our variables do have predictive power
- AUC scores greater than random

With Oldmodel:
- Our variables with the OM can improve the prediction
- In Train_Full, our model can reduce the risk of payment by 30% at the 80% of the company market
- In Test_Full, our model can reduce the risk of payment by 31% at the same 80% threshold

Criticism of the Results and Future Work

- The company data itself was a sample from a bigger dataset. The addition of the same variables to a bigger dataset may result in different AUC scores.
- Adjusting some parameters (i.e. distribution) might give better results.
- Different methods of feature selection bring different sets of variables.
- More features can be introduced to the model: earthquake, solar radiation, gas leak, etc.