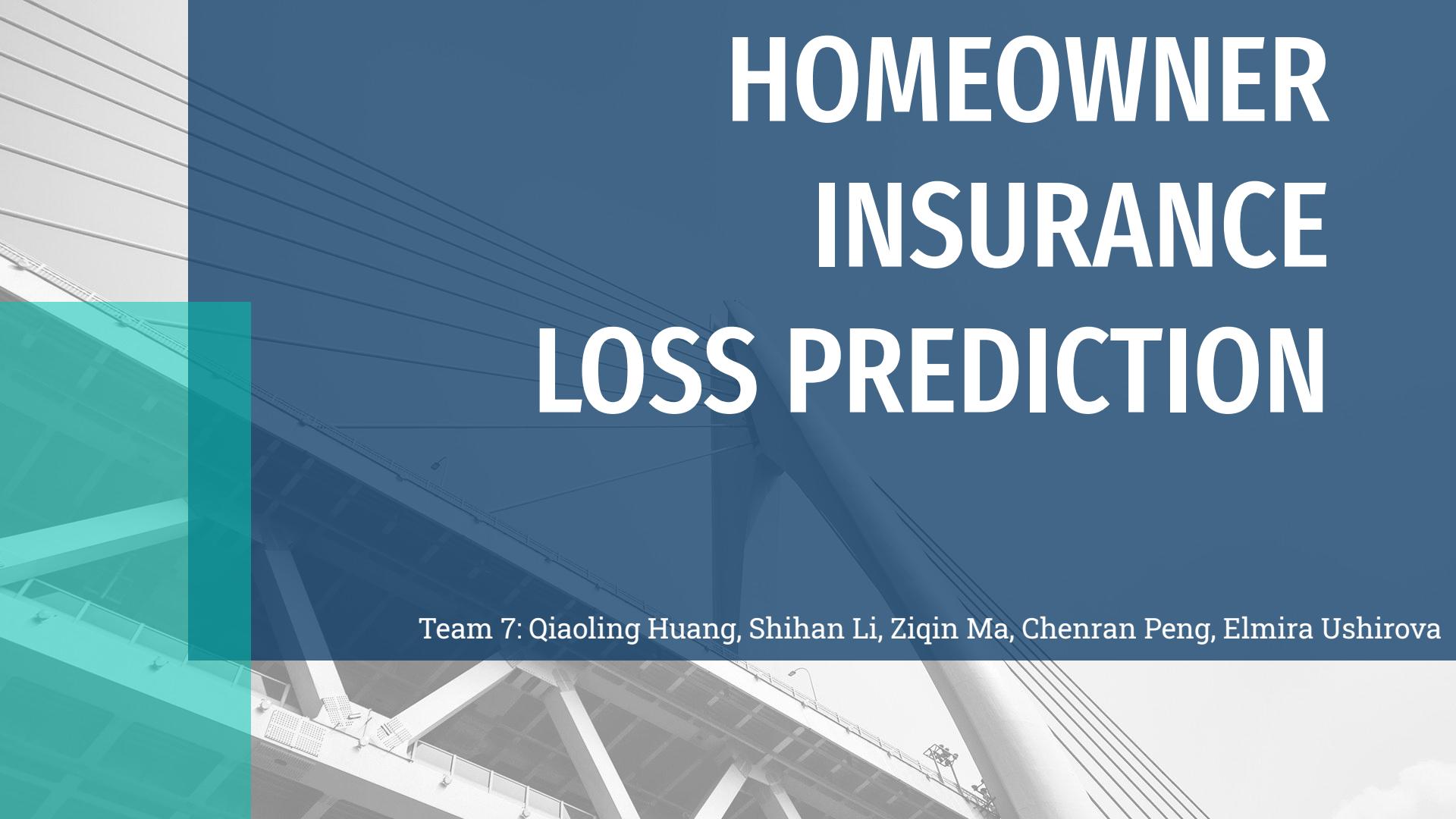


# HOMEOWNER INSURANCE LOSS PREDICTION



Team 7: Qiaoling Huang, Shihan Li, Ziqin Ma, Chenran Peng, Elmira Ushirova

# AGENDA

**01**

**Business Problem**  
Our Business Problem

**02**

**Datasets and Source**  
Datasets and sources used in  
this project

**04**

**Results**

**05**

**Criticism of the Results and Future  
Work**

**03**

**Methodology**

**06**

**Conclusion  
and  
Acknowledgment**

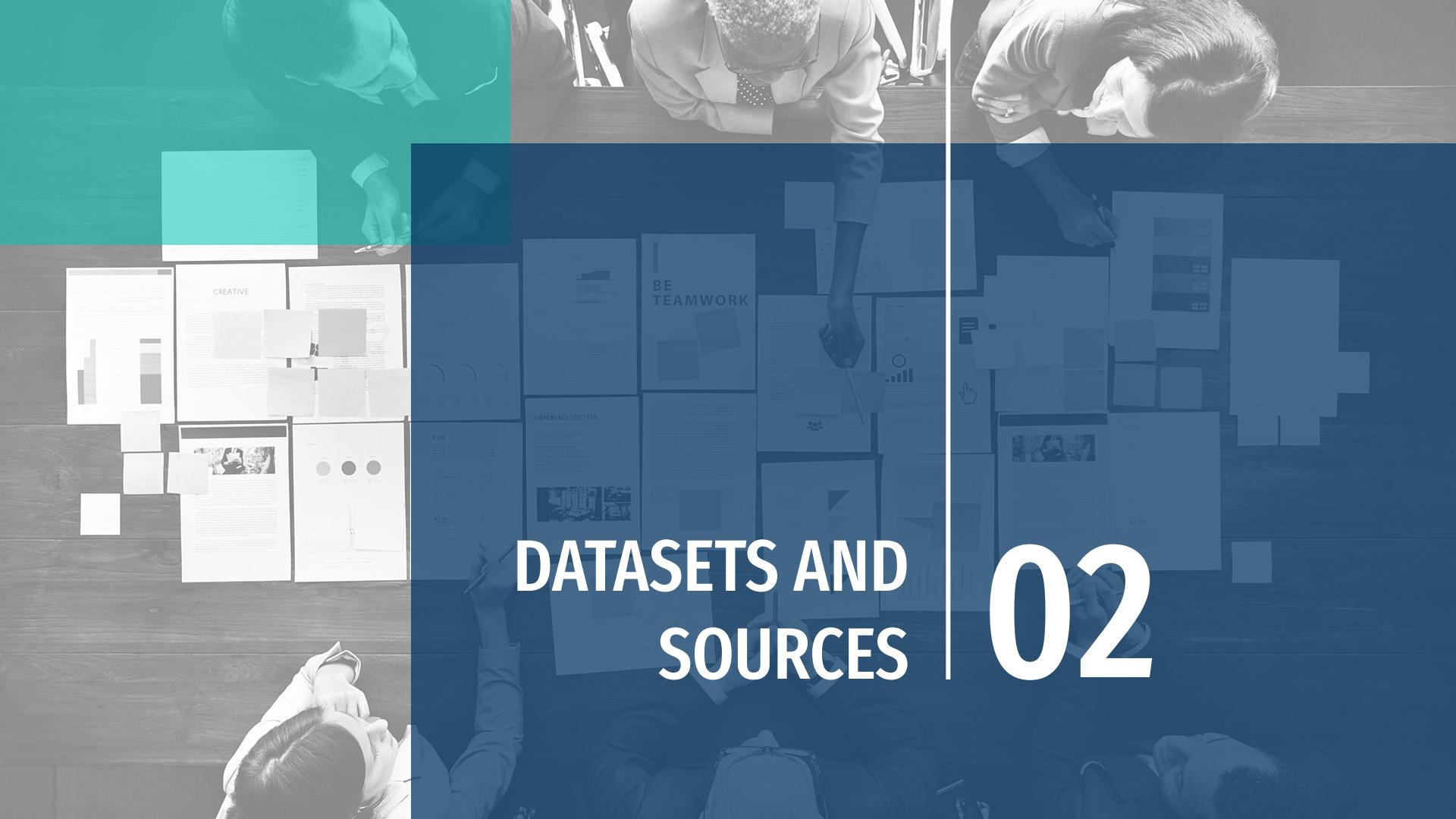


# BUSINESS PROBLEM

01

# Business Problem

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 AND SOURCES | 02

# Datasets: Company Data

**Data provider:** Fast-growing home insurance company in the New England area.

## **Data:**

- Anonymized zip-code level
- 5 years (2013 - 2018)
- Actual loss
- Old model loss prediction
- Earned Exposure
- Population
- Area
- Number of properties

# Datasets: Crime from the FBI



**FBI crime data** includes numbers of different crime categories by State and city.

20 tables (1999 - 2018):

**Violent crime:** Murder and nonnegligent manslaughter, Rape, Robbery, Aggravated assault

**Property crime:** Burglary, Larceny- theft, Motor vehicle theft, Arson.

Table 8

Offenses Known to Law Enforcement

by State by City, 2018

State	City	Population	Violent crime	Murder and nonnegligent manslaughter		Rape <sup>1</sup>	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson <sup>2</sup>
				Murder	nonnegligent manslaughter								
ALABAMA	Abbeville	2,551	18	0	2	0	16	49	14	33	2		
	Adamsville	4,323	19	0	1	4	14	289	42	230	17		
	Alabaster	33,501	92	0	2	10	80	579	56	497	26		
	Albertville	21,428	24	0	6	10	8	802	194	492	116		
	Alexander City	14,548	314	2	5	15	292	610	92	484	34		
	Aliceville	2,315	9	0	0	1	8	24	7	16	1		
	Andalusia	8,753	82	1	6	6	69	467	76	368	23		
	Anniston	21,592	646	7	43	68	528	1,696	413	1,131	152		
	Ardmore	1,428	5	0	0	1	4	35	15	18	2		
	Ashford	2,145	4	0	0	0	4	32	2	24	6		
	Ashland	1,918	52	0	0	1	51	57	16	35	6		
	Athens	26,177	13	3	0	8	2	768	110	630	28		
	Auburn	65,585	240	4	16	29	191	1,396	115	1,206	75		
	Bay Minette	9,304	73	0	3	3	67	250	35	196	19		
	Bayou La Batre	2,500	36	0	4	5	27	305	70	203	32		
	Birmingham	210,564	4,025	88	180	941	2,816	13,295	2,555	8,988	1,752	145	
	Boaz	9,678	74	0	3	3	68	364	68	259	37		

Source: <https://ucr.fbi.gov/crime-in-the-u-s>

# Datasets: Lightning

\$451 million in property damage per year (NFPA).

Google Cloud Platform public dataset.

Daily information about the number of strikes and location (around US territory).

20 tables (1999 - 2018):

- Day
- Number of strikes
- Geo point



Source:

<https://console.cloud.google.com/marketplace/details/noaa-public/lightning?filter=solution-type:dataset&filter=category:climate&id=d18e2712-bc50-471a-bf22-a2de3d9489d9&authuser=1>

Row	day	number_of_strikes	center_point
54801	2019-12-31	6	POINT(-75.5 32.4)
54802	2019-12-31	6	POINT(-75.4 32.4)
54803	2019-12-31	6	POINT(-74.9 32.4)
54804	2019-12-31	6	POINT(-74.9 32.9)
54805	2019-12-31	6	POINT(-75.9 32.3)
54806	2019-12-31	6	POINT(-75 32.4)
54807	2019-12-31	7	POINT(-76.6 31.7)
54808	2019-12-31	7	POINT(-75.8 31.7)
54809	2019-12-31	7	POINT(-75.5 32)
54810	2019-12-31	7	POINT(-74.6 32)
54811	2019-12-31	8	POINT(-72.3 34)

# Exploratory Data Analysis

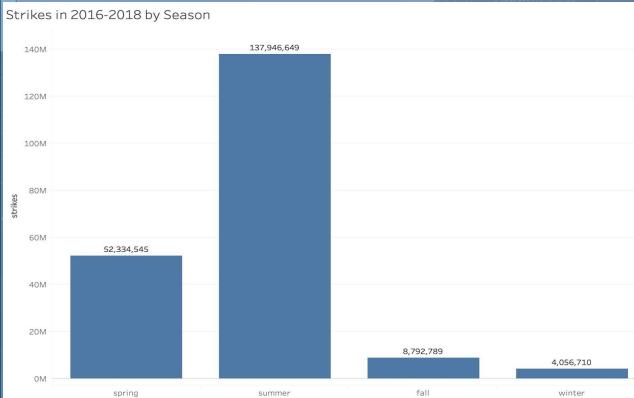
## Company Data



## Crime Data



## Lightning Data



The background of the slide features a collage of three images. On the left, a man in a white shirt is seen from behind, looking down at a table. In the center, a woman wearing glasses and a dark jacket is looking down at something. On the right, a desk is shown with a laptop, a white mug, and a desk lamp. A large vertical blue bar runs down the center of the slide, containing the text.

03

## METHODOLOGY

# NAs imputation

## Crime

We replaced NAs in the original data by computing the mean of each State and multiply by the population ratio of the missing city.

For those population that we don't have, we assume that we replaced with mean population with other cities without any crime record.

## Lightning

We matched geocodes of strikes with a unique zip code based on distance. If lightning data is missing under certain zip code, we assumed that 0 strikes occurred in that area.

# Create New Variables

1 year	3 years	5 years	10 years	20 years
--------	---------	---------	----------	----------

Sum of Original

State Level

Nation Level

Violent Crime/Total Crime  
Property Crime/Total Crime

Each Type of Violent Crime/Violent Crime  
Each Type of Property Crime/Property Crime

Total: 205 New Variables

Crime

1 year	3 years	5 years	10 years	20 years
--------	---------	---------	----------	----------

Group by Each Season

Sum of Original

State Level

Nation Level

Strikes and Events/Area

Strikes and Events/Population

Strikes and Events/Property

Total: 315 New Variables

Lightning

# Sample Design

## In Time and Out of time

2016 to 2018 as our in-time sample

2014 and 2015 as our out of time sample

## CAP

99.3 percentile on losses as our Y

We divided losses into 4 levels: 0, low, medium, high.

## Stratify

By year, State and loss level.

70% Train  
30% Test

Considering our data shape, we did downsampling, randomly dropped 50% of zero loss from the in-time sample to balance zero and non-zero losses. We ended up with 5 samples: Train\_reduced, Test\_reduced, Train\_full, Test\_full, and Validation.

# Feature Selection

We used wrapper method (backward elimination) for the feature selection, we ended up with 33 variables in total.

After the feature selection, the result showed that crime data have more predicting power than lightning.

# Selected Variables

Crime (Total 23 variables)			
<b>Robbery at State level for 1 year</b>	Larceny-theft /Property crime at 20 years level	Murder/Property crime at 10 years level	Motor vehicle theft at State level 10 years
Motor vehicle theft/Property crime 3 years level	Aggravated assault /Total Violent crime 1 year level	Motor vehicle theft /Property crime at 20 years level	Larceny-theft /Property crime at 10 years level
Property crime at State level 10 years	Robbery at State level 5 years	Motor vehicle theft at State level 3 years	Burglary at State level 5 years
Property crime /Total crime 10 years level	Burglary at State level 10 years	Property crime at State level 5 years	Burglary/Property crime 3 years level
Aggravated assault /Total Violent crime 5 year level	Property crime at nation level 10 years	Burglary at State level 20 years	Larceny-theft/Property crime 3 years level
Larceny-theft at State level 5 years	Violent crime/Total crime 10 years level	Larceny-theft at State level 20 years	

# Selected Variables

## Lightning (Total 10 variables)

<b>Total strikes for 5 years</b>	Event number in fall for 3 years	Strikes in winter for 5 years	Event number in winter for 5 years
Total strikes for 10 years	Total events in 10 years	To total strikes in fall for 10 years	Strikes in winter for 10 years
Event number at nation level for 10 years	Event number in summer for 3 years		

# Model Design

Used Train\_reduced  
without oldmodel

Used Train\_reduced  
with oldmodel

GLM, Tree-based(Random Forest, XGBoost), Neural Network

Check performance of models on 4 sample datasets:  
Test\_reduced, Train\_full, Test\_full, Validation



# 04 | RESULTS

# Results - without oldmodel

Selected features do have predictive power on losses.

Model	Train_reduced AUC	Test_reduced AUC	Train_full AUC	Test_full AUC	Validation AUC
Benchmark	0.5	0.5	0.5	0.5	0.5
Simple GLM	0.631	0.516	0.6	0.606	0.480
Random Forest	0.834	0.481	0.719	0.759	0.523
XGBoost	0.777	0.605	0.668	0.672	0.554
Neural Network	0.551	0.526	0.528	0.579	0.54



# Important Variables – without oldmodel

<b>Motor vehicle theft/Property crime 3 years level</b>	<b>56%</b>
<b>Larceny-theft/Property crime 3 years level</b>	<b>14%</b>
<b>Larceny-theft at State level 5 years</b>	<b>7%</b>
<b>Strikes in winter for 5 years</b>	<b>6%</b>
<b>Robbery at State level for 1 year</b>	<b>4%</b>
<b>Total strikes for 5 years</b>	<b>3%</b>

# Results - with oldmodel

Selected variables have additional predictive power on the oldmodel.

Model	Train_reduced AUC	Test_reduced AUC	Train_full AUC	Test_full AUC	Validation AUC
Benchmark (oldmodel)	0.572	0.61	0.598	0.554	0.641
Simple GLM	0.649	0.619	0.643	0.651	0.600
Random Forest	0.931	0.538	0.808	0.802	0.592
XGBoost	0.735	0.607	0.699	0.679	0.608
Neural Network	0.607	0.552	0.582	0.621	0.547

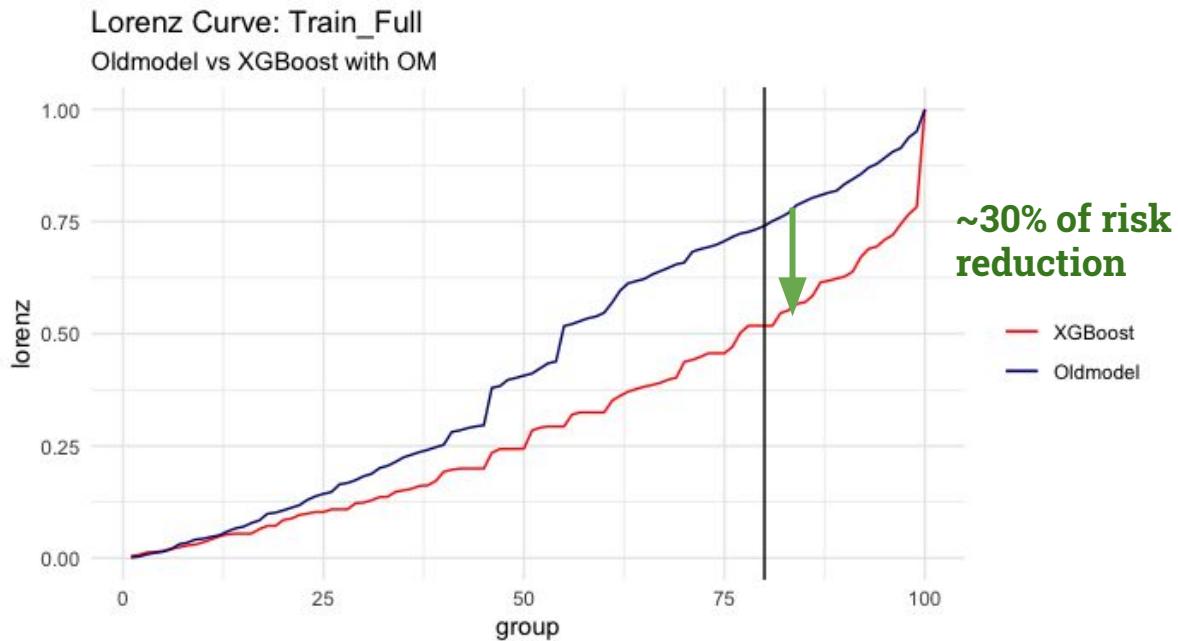


# Important Variables - with oldmodel

Oldmodel	81%
Number of strikes for recent 5 years	14%
Violent crime /Total crime ratio for recent 10 years	1%

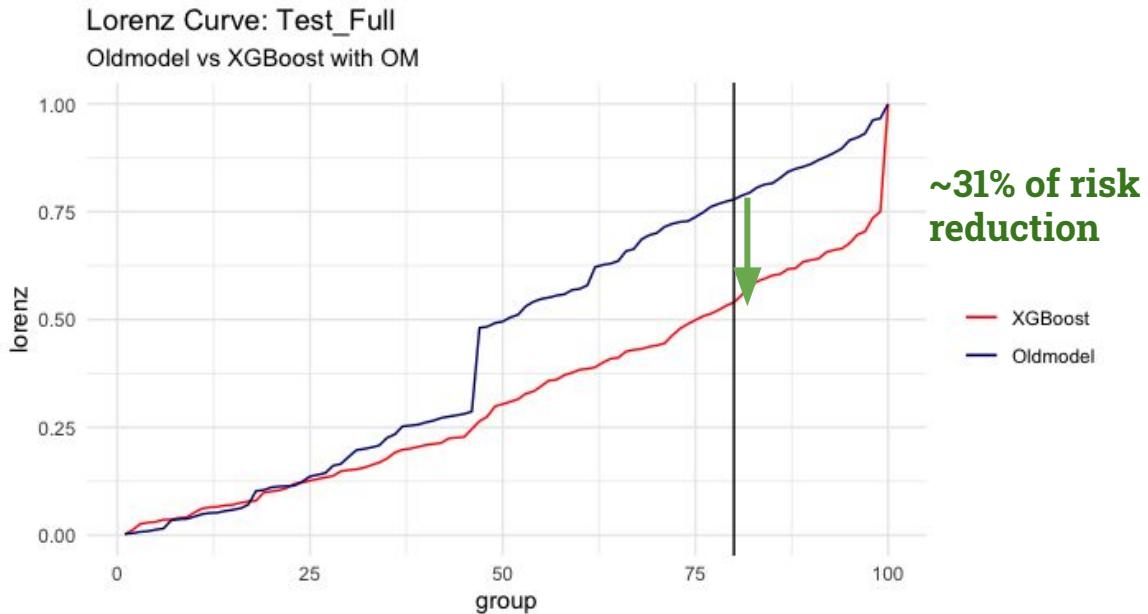


# Lorenz Curve Comparison - Train





# Lorenz Curve Comparison - Test





## CRITICISM OF THE RESULTS AND FUTURE WORK



05

# Criticism of the results and future work

The company data was a sample from a bigger dataset. Adding the same variables to that bigger dataset may result in different AUC scores

Different methods of feature selection might bring different sets of variables; Re-train the model with only important ones can improve the performance.

Adjusting some parameters (i.e. distribution) might give better results.

More features can be introduced to the model: earthquake, solar radiation, gas leak, etc.





# CONCLUSION AND ACKNOWLEDGMENT

# 06

# Conclusion

Some new features have predictive power by themselves, as well as with existing model predictions.



In terms of business application, our new model can reduce the risk of loss payment by 30%.

**Thank you!**