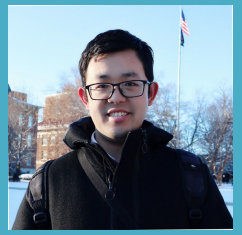
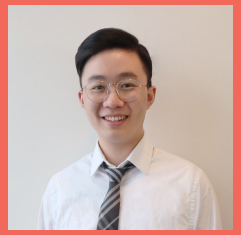
Credit No-hit Prediction

By Team 2
Yulong Gong, Muyan Xie, Yangyang Zhou, Yichi Zhang

About Us



Yulong Gong



Muyan Xie

Yichi Zhang

Yangyang Zhou

Business Problem

Credit status is one of the most important predictors in business setting, developing a successful credit-status predicting model could help our clients:

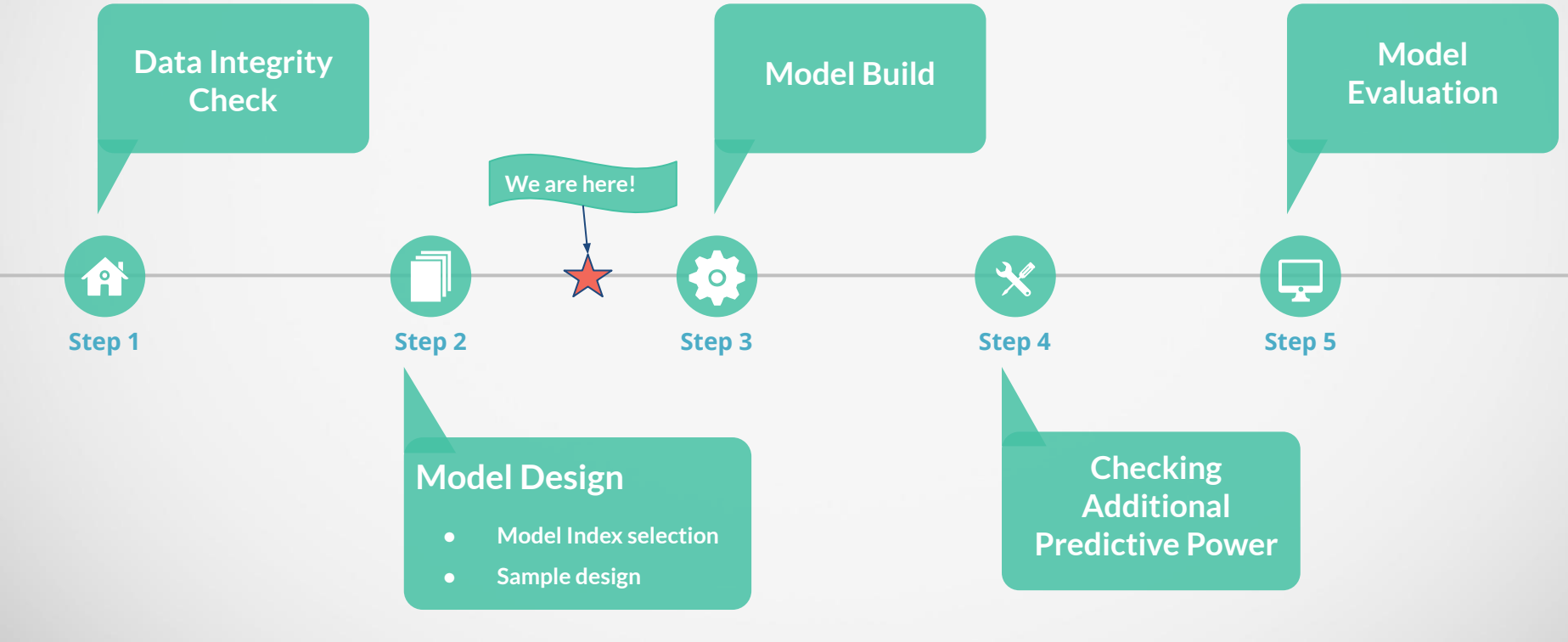
Predict consumer credit hit or no-hit for marketing promotional purpose.

- Hit: There is credit record, represented as 0 in our dataset.
- No_hit: Don't have credit record, represented as 1 in our dataset.

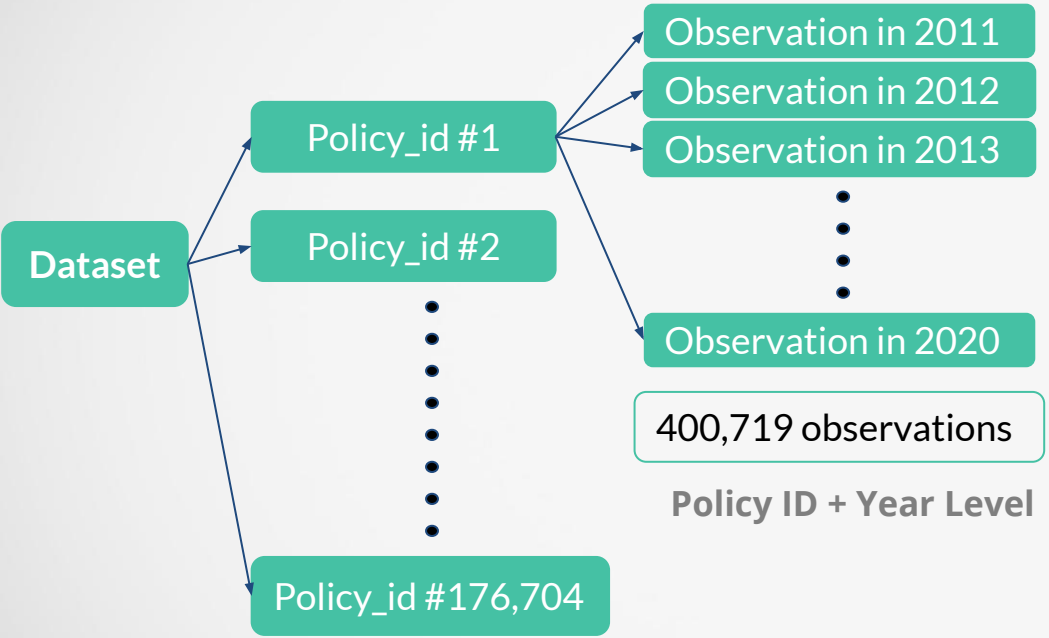
Reduce company's potential loss.

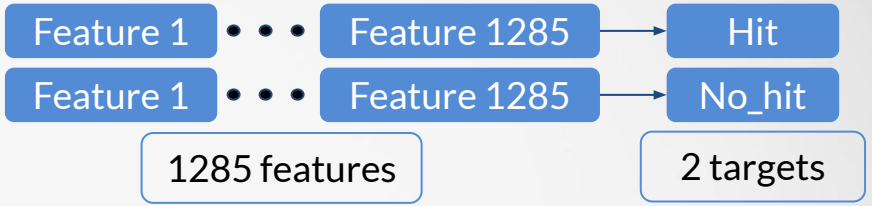
- Strategically target the consumer with credit records.

Overview of the Process



Data Overview





Hit: there is credit records, represented as 0.

No_hit: there is no credit records, represented as 1.

Data Integrity Checking

Policy ID Level

176,704 unique policy_id

Hit or No-hit Transformation

The total number of Policy_ids for which transformation records exist is **14170** (8% of unique ids)

Previous years recorded	Last recorded year	Hit/No hit	Number	Final Results
Hit or No-hit or No record	2020	0	5400	0
		1	144	1
	2019	0	6641	0
		1	554	1
	2018	0	721	0
		1	79	1
	2017	0	502	0
		1	50	1
	2016	0	166	0
		1	35	1
	2015	0	123	0
		1	20	1
	2014	0	58	0
		1	14	1
	2013	0	7	0
		1	2	1

9.0%

No-hit = 1

15893
In total IDs

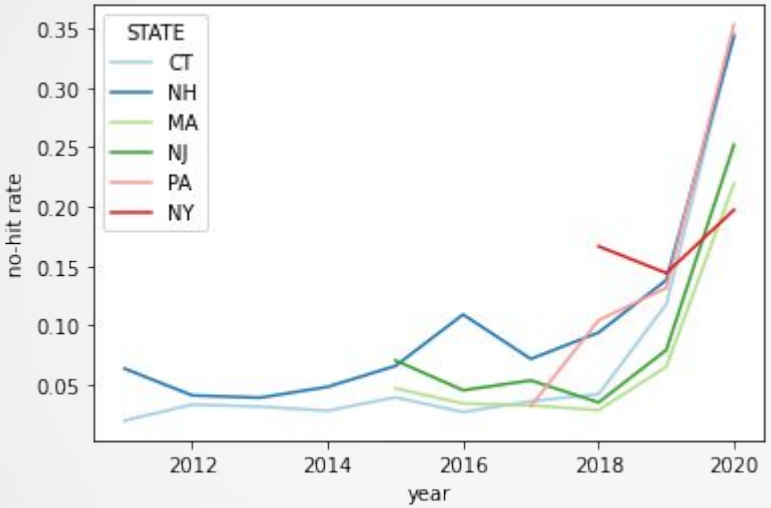
91.0%

Hit = 0

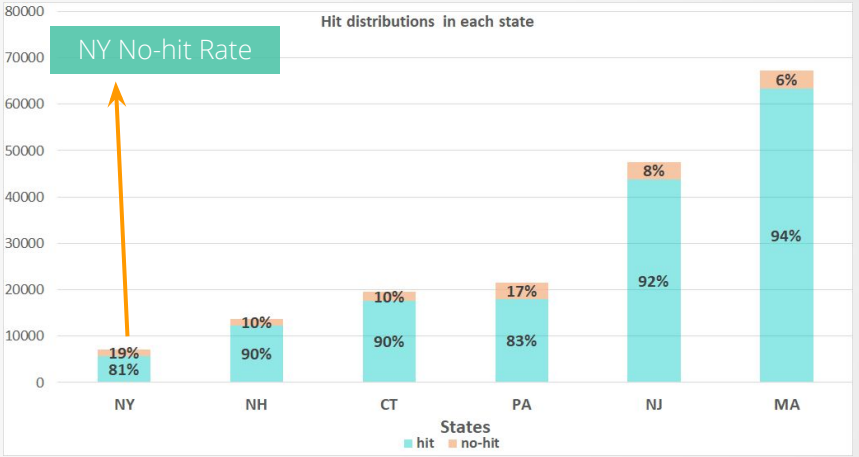
160811
In total IDs

Policy ID Level

No-hit Rate By Year



No-hit Distribution By State



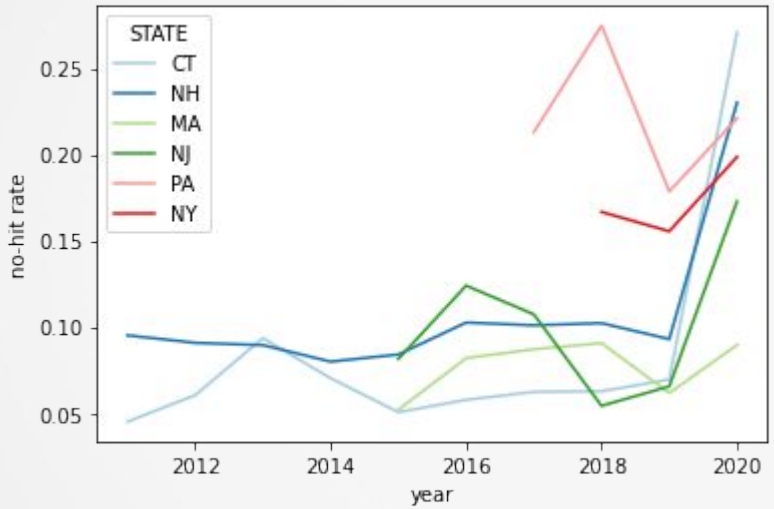
- Not all states have records since 2011
- 2020 has abnormal no-hit rate

- NY has least records but highest no-hit rate
- MA and NJ have low no-hit rate

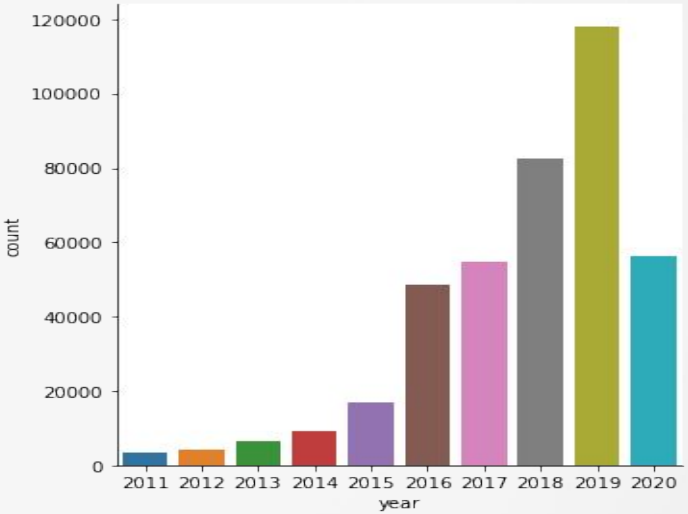
Policy ID + Year Level

400,719 observations

No-hit Rate By Year



Policy ID Distribution



- 2020 no-hit ratio increases may due to the decreasing policy amounts
- **PA** and **NY** remain a relatively high no-hit rate over years. All states have significant increments from 2019 to 2020.
- The number of Policy ID drops rapidly in 2020

Model Design

Model Design

in-sample train

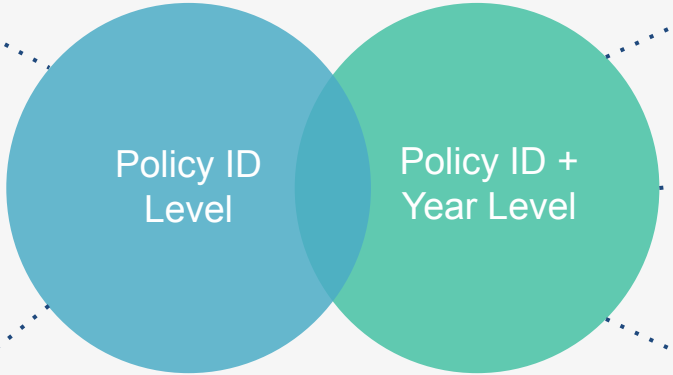
contains 80% of policies in 2016-2019

in-sample validation

contains 20% of policies in 2016-2019

out-of-sample test

contains policies in 2011-2015 and 2020



in-sample train

contains 80% of policies in 2016-2019

in-sample validation

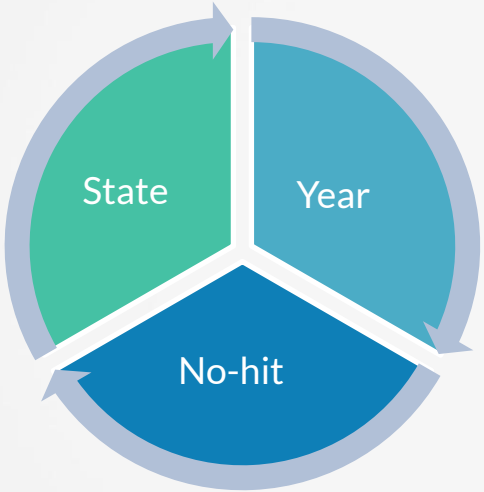
contains 20% of policies in 2016-2019

out-of-sample test

contains policies in 2011-2015 and 2020

91.0% Hit **9.0% No-hit**

Policy ID Level



Stratification

Policy ID Level design results:

- Total rows: 176,704 rows
- In sample : 131,913 rows
  - train(80%) : 105,530 rows
  - validation(20%): 26,383 rows
- Out of sample: 44,791 rows

Policy ID Level

	DISTRIBUTION			NO HIT RATIO		
STATE	IN-Train and IN-Valid	OUT 11-15,20	OUT YEAR LEVEL	IN-Train and IN-Valid	OUT 11-15,20	OUT YEAR LEVEL
MA	42%	27%	14%	4%	15%	14%
NJ	31%	15%	7%	6%	23%	23%
PA	13%	9%	4%	12%	35%	35%
CT	8%	21%	34%	8%	14%	8%
NH	5%	16%	36%	11%	9%	8%
NY	1%	13%	6%	14%	30%	20%

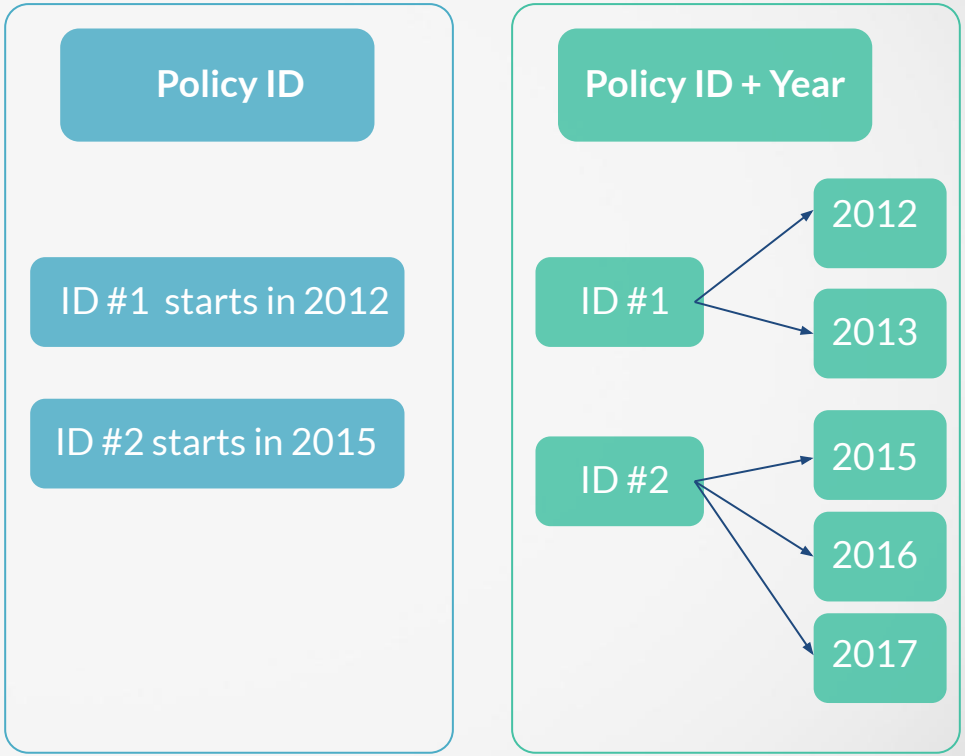
Policy ID Level design results:

- Total rows: 176,704 rows
- In sample : 131,913 rows
  - $train_{(80\%)}$: 105,530 rows
  - $validation_{(20\%)}$: 26,383 rows
- Out of sample: 44,791 rows

Policy ID + Year Level

Policy_id + year level design results:

- Total rows: 400,719 rows
- In sample : 302,471 rows
  - train(80%): 241,802 rows
  - validation(20%): 60,669 rows
- Out of sample: 98,248 rows



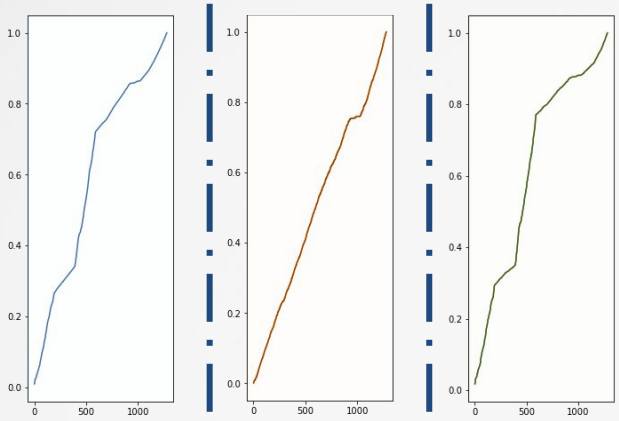
Policy ID + Year Level

Policy_id + year level design results:

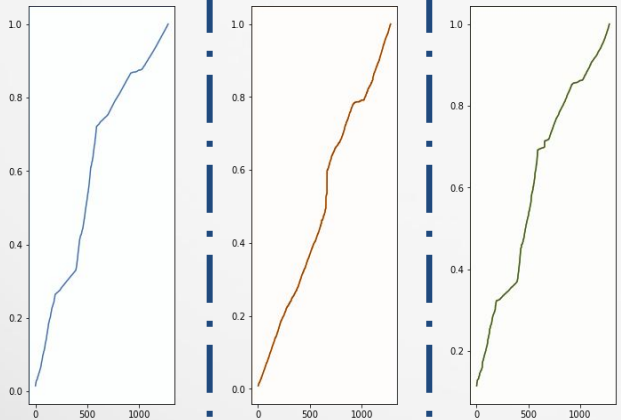
- Total rows: 400,719 rows
- In sample : 302,471 rows
  - train(80%): 241,802 rows
  - test(20%): 60,669 rows
- Out of sample: 98,248 rows

	DISTRIBUTION			NO HIT RATIO		
STATE	IN-Train and IN-Valid	OUT 11-15,20	OUT-ID LEVEL	IN-Train and IN-Valid	OUT 11-15,20	OUT-ID LEVEL
MA	53%	14%	27%	7%	14%	15%
NJ	25%	7%	15%	7%	2%	23%
PA	9%	4%	9%	19%	35%	35%
CT	8%	34%	21%	9%	8%	14%
NH	5%	36%	16%	15%	8%	9%
NY	1%	6%	13%	17%	20%	30%

Feature Selection Results



Policy ID level
(1285 features)



Policy ID + Year level
(1285 features)

Column 1: Random Forest

Column 2: XGBoost

Column 3: Decision Trees

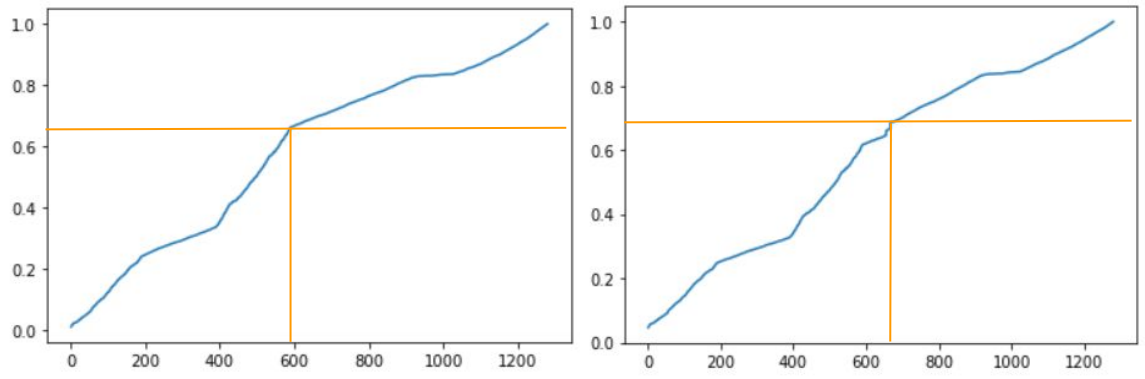
Aggregate Feature Selection Results

Policy ID level

- 176,704 Rows
- 657 Features

Policy ID + Year level

- 400,719 Rows
- 657 Features



How to select?

Average importance score of three tree-based models.

How much portion?

Elbow principle to maintain around 70% variance.

How many?

Finally picked highest 650 features plus 7 identity objects.

Plans for second part of capstone

Model Build

- logistic regression
- tree-based models

Checking Additional Predictive Power

Check the variable predictive power again to enhance model performance.

Model Evaluation

- Gini

Used in Finance industry to predict credits.

- AUC

Evaluation metrics for checking model performance.

- K-S test

Evaluate the model performance based on data distribution.

Q&A

Thanks