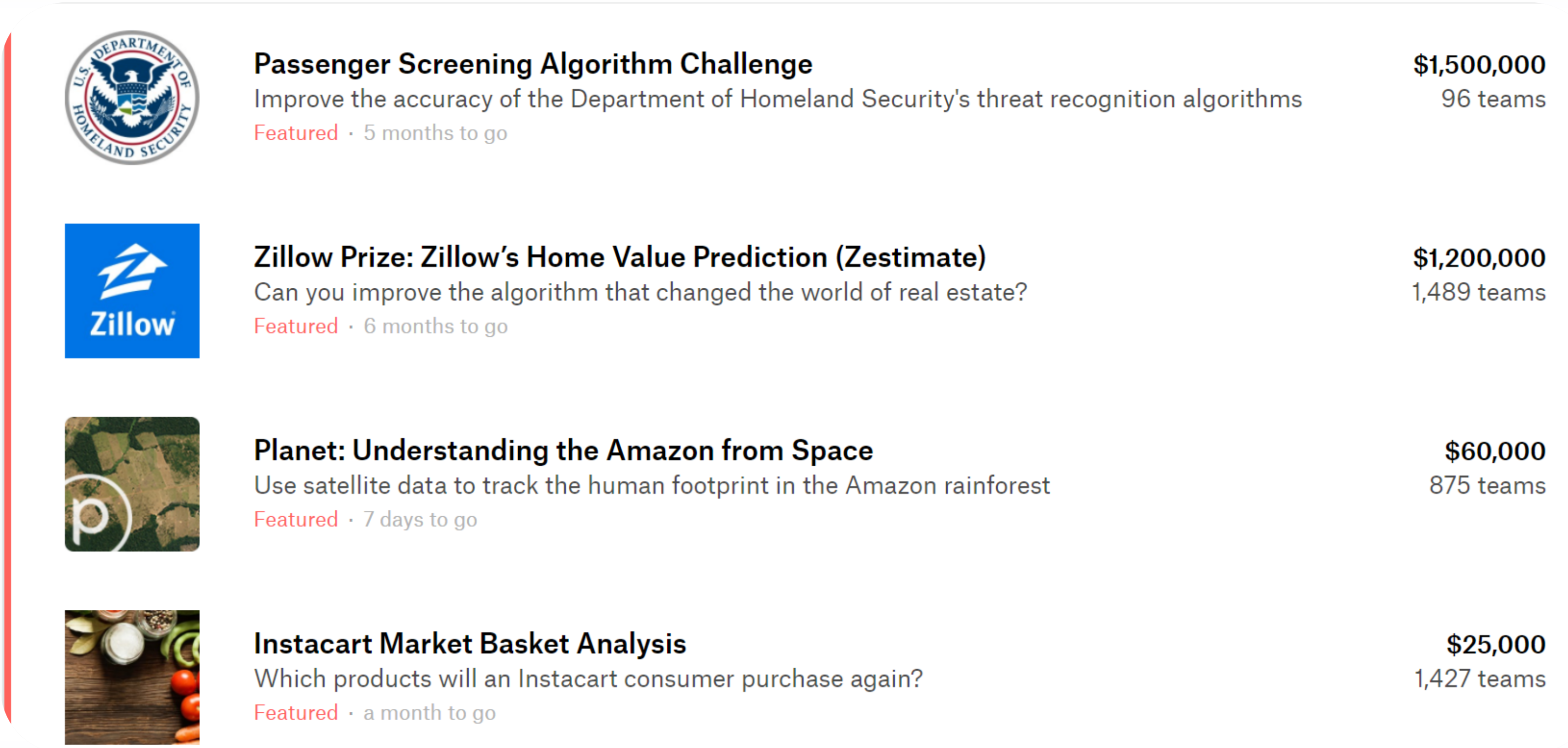
Building intuition about the data

Video overview

1. Getting domain knowledge
2. Checking if the data is intuitive
3. Understanding how the data was generated

Get domain knowledge



Passenger Screening Algorithm Challenge
Improve the accuracy of the Department of Homeland Security's threat recognition algorithms
Featured · 5 months to go
$1,500,000
96 teams

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 6 months to go
$1,200,000
1,489 teams

Planet: Understanding the Amazon from Space
Use satellite data to track the human footprint in the Amazon rainforest
Featured · 7 days to go
$60,000
875 teams

Instacart Market Basket Analysis
Which products will an Instacart consumer purchase again?
Featured · a month to go
$25,000
1,427 teams

Get domain knowledge, example

Task: Predict advertiser’s cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	…
78db044136	s	0.28	s_2	3	0	…
68a0110c33	s	1	s_2	1	13	…
2r39fw11w3	p	1.2	p_1	3	419	…

Check if the data is intuitive

…	*Age*	…
…	21	…
…	45	…
…	336	…
…	19	…
…	…	…

Check if the data is intuitive

…	*Age*	…
…	21	…
…	45	…
…	336	…
…	19	…
…	…	…

- Is *336* a typo?

Check if the data is intuitive

…	***Age***	…
…	**21**	…
…	**45**	…
…	**336**	…
…	**19**	…
…	…	…

- Is *336* a typo?
- Or we misinterpret the feature and age 336 is normal

Check if the data is intuitive

Task: Predict advertiser's cost

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	…
78db044136	s	0.28	s_2	***3***	***0***	…
68a0110c33	s	1	s_2	1	13	…
2r39fw11w3	p	1.2	p_1	3	419	…

Google Adwards 데이터임 - - > 관련 Knowledge획득하여 컬럼의미 이해

Check if the data is intuitive

Task: Predict advertiser's cost

the task is to predict what will be the costs for the advertiser (how much he will pay to Google, column Cost) when the parameters (e.g. keywords) are changed

Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	*is_incorrect*
78db044136	s	0.28	s_2	***3***	***0***	***True***
68a0110c33	s	1	s_2	1	13	***False***
2r39fw11w3	p	1.2	p_1	3	419	***False***

The ads are grouped in groups, there are features AdGroupId AdGroupName describing them.
A campaign corresponds to some specific parameters that an advertiser sets. Similarly, there are ID and name features CampaignId, CampaignName.
And finally there is some information about keywords: KeywordId and KeywordText. Slot is 1 when ad is shown on top of the page, and 2 when on the side. Device is a categorical variable and can be either

Understand how the data was generated

Understand how the data was generated

It is crucial to understand the generation process to set up a proper validation scheme

What was the algorithm for sampling objects from the database?
i.e, host sample the objects random or over - sample the
particular class(for data be more class balanced)

model의 validation scheme 정하는데 중요한 정보임 - - >
train, test set이 서로 다른 algorithm으로 생성되었다면 test
set의 일부를 validation set으로 사용할 수 없음

Check if the data is intuitive

Task: Predict advertiser's cost

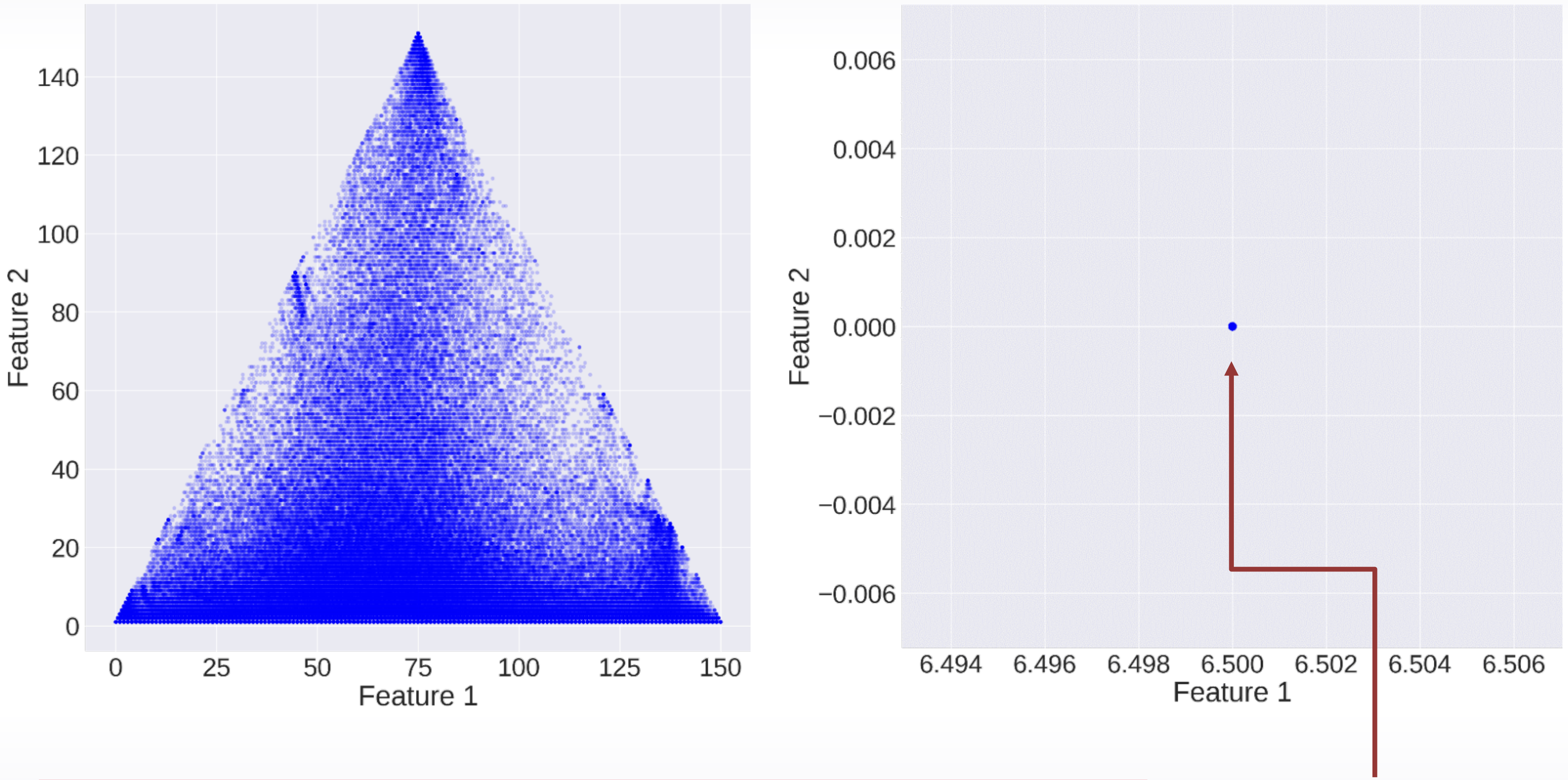
Data:

AdGroupId	AdNetwork Type2	MaxCpc	Slot	*Clicks*	*Impressions*	*is_incorrect*
78db044136	s	0.28	s_2	***3***	***0***	***True***
68a0110c33	s	1	s_2	1	13	***False***
2r39fw11w3	p	1.2	p_1	3	419	***False***

domain knowledge에 의하면 Clicks < Impressions 이면 정상(Correct)

Understand how the data was generated

EDA_video2.ipynb에 상세내용 있음


Train

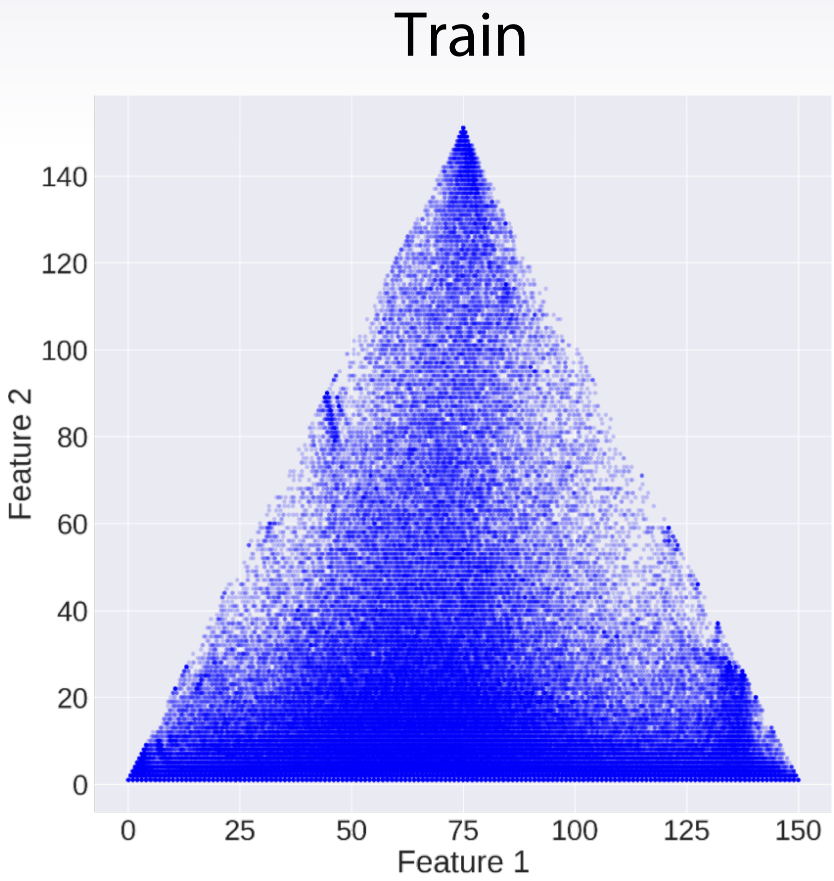
Test

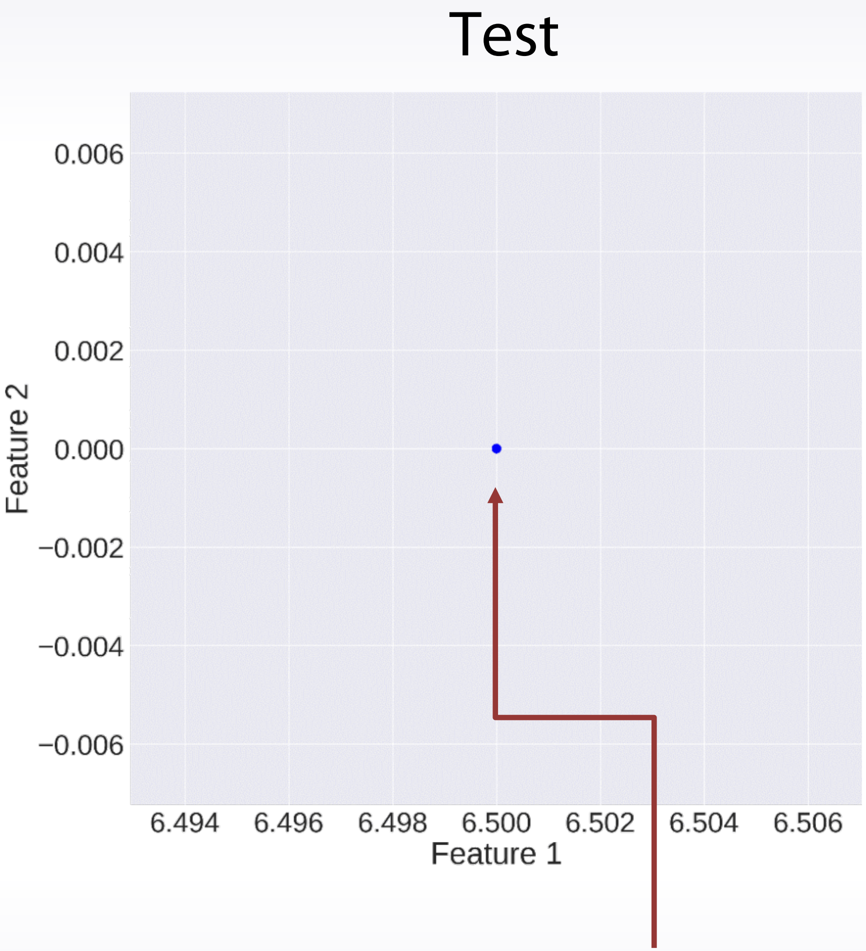
Dot!

train, test set이 서로 다른 algorithm으로 생성되었을 가능성이 높다

adjusted train set to match test set - - > validation score improved

Understand how the data was generated


Train


Test

Dot!

#days in *train* > #days in *test*
#rows in *train* < #rows in *test*

Conclusion

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem
- **Check if the data is intuitive**
  - And agrees with domain knowledge

Conclusion

- **Get domain knowledge**
  - It helps to deeper understand the problem
- **Check if the data is intuitive**
  - And agrees with domain knowledge
- **Understand how the data was generated**
  - As it is crucial to set up a proper validation