Building intuition about the data

If we take a look at the computations hosted in the Kaggle, well, you'll notice, they are rather diverse. Sometimes, we need to detect threats on three dimensional body scans, or predict real estate price, or classify satellite images.

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

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

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

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

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	…
…	…	…

- 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	…

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*

Understand how the data was generated

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

Check if the data is intuitive

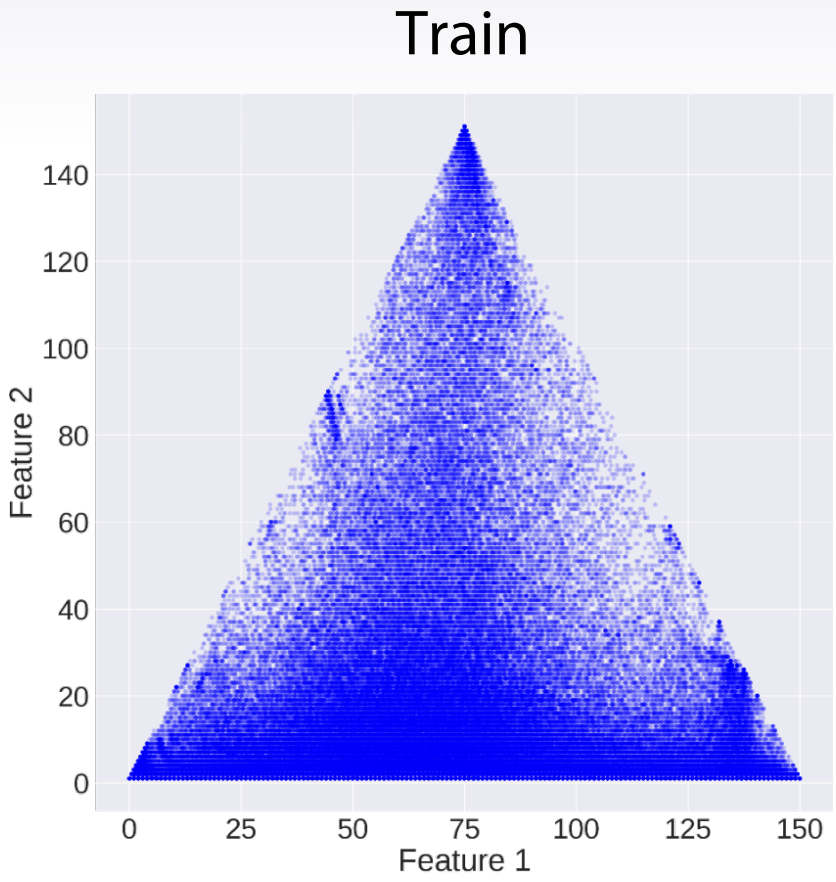
Task: Predict advertiser's cost

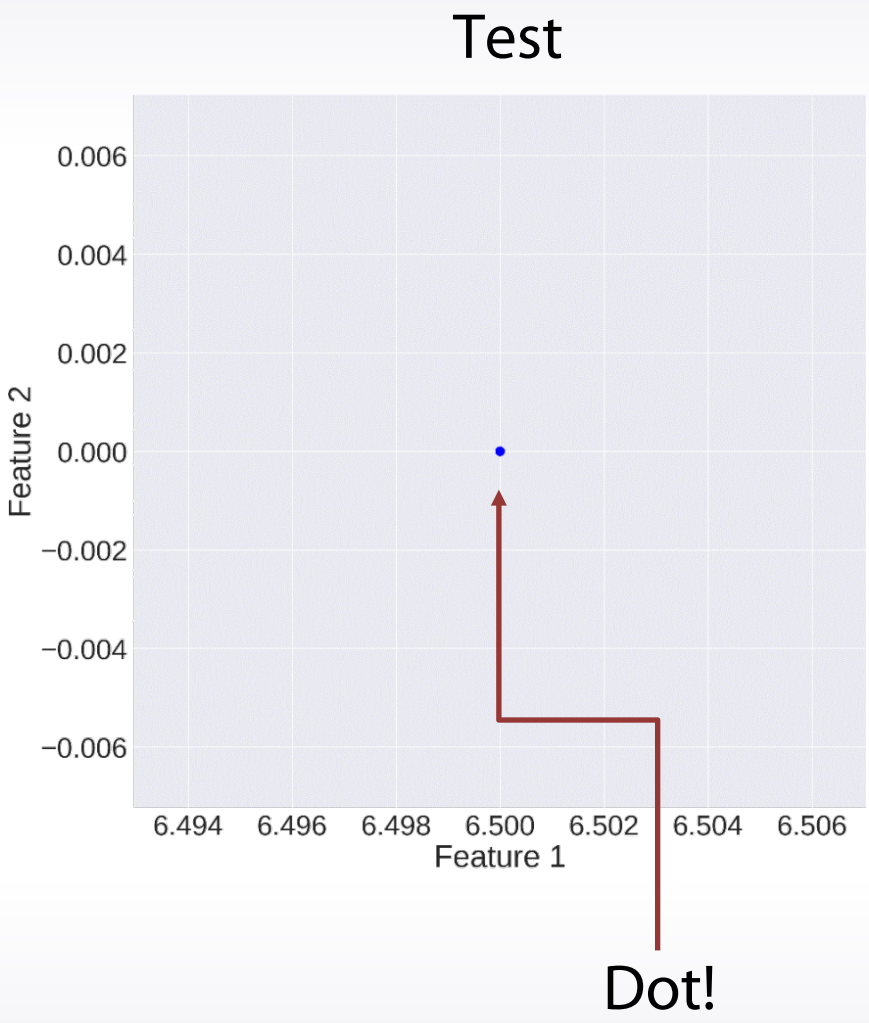
Data:

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*

For example, in our case, we could create a new feature, is_incorrect, and mark all the rows that have errors. Probably, our models will find this feature helpful.

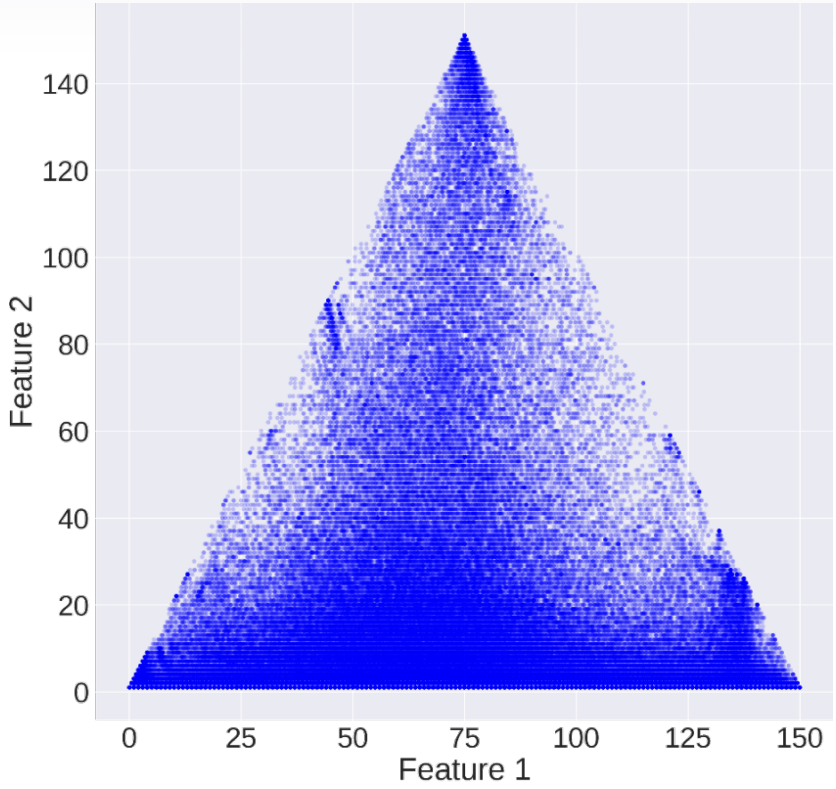
Understand how the data was generated



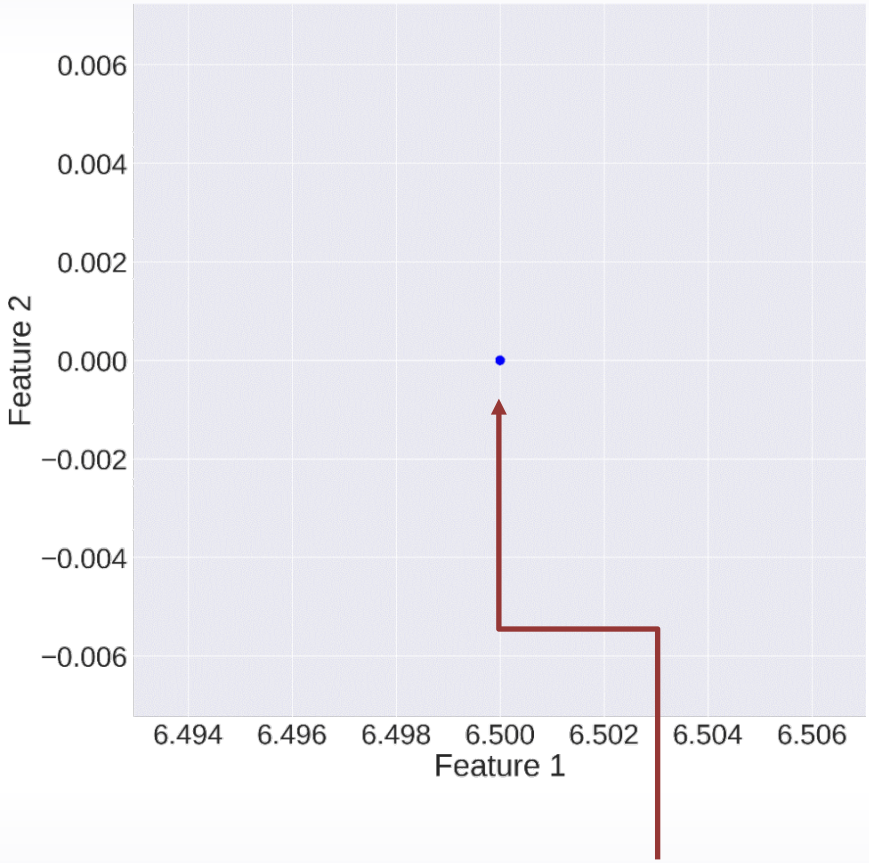


Understand how the data was generated

Train



Test



#days in *train* > #days in *test*

#rows in *train* < #rows in *test*

Conclusion

in this video, we've discussed several important exploratory steps.

- **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