

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

...	<i>Age</i>	...
...	21	...
...	45	...
...	336	...
...	19	...
...

Check if the data is intuitive

...	<i>Age</i>	...
...	21	...
...	45	...
...	336	...
...	19	...
...

- Is 336 a typo?

Check if the data is intuitive

...	<i>Age</i>	...
...	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	<i>is_incorrect</i>
78db044136	s	0.28	s_2	3	0	<i>True</i>
68a0110c33	s	1	s_2	1	13	<i>False</i>
2r39fw11w3	p	1.2	p_1	3	419	<i>False</i>

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**

Check if the data is intuitive

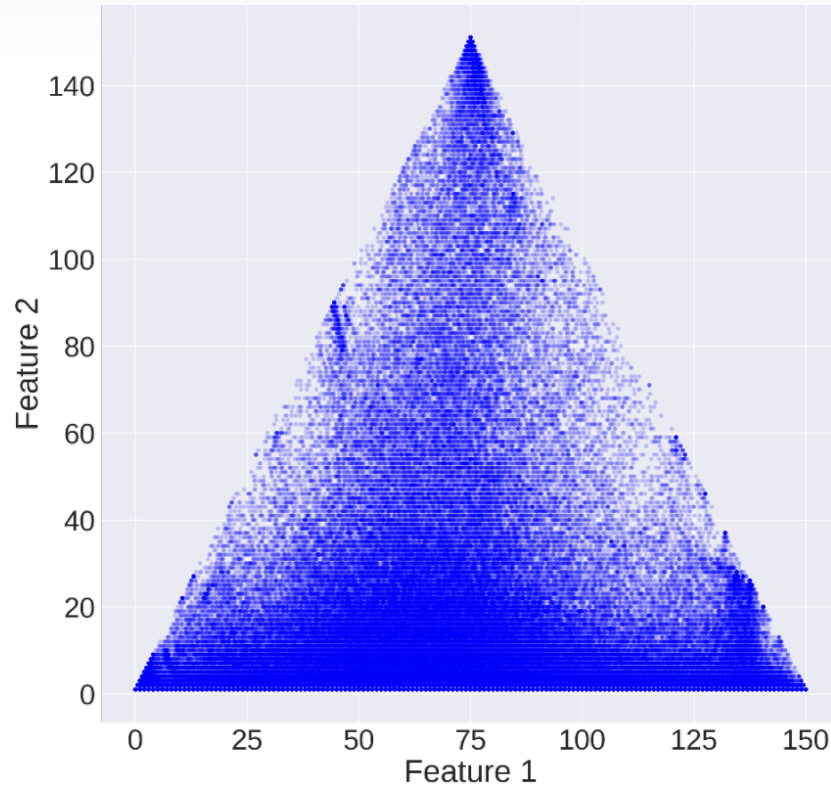
Task: Predict advertiser's cost

Data:

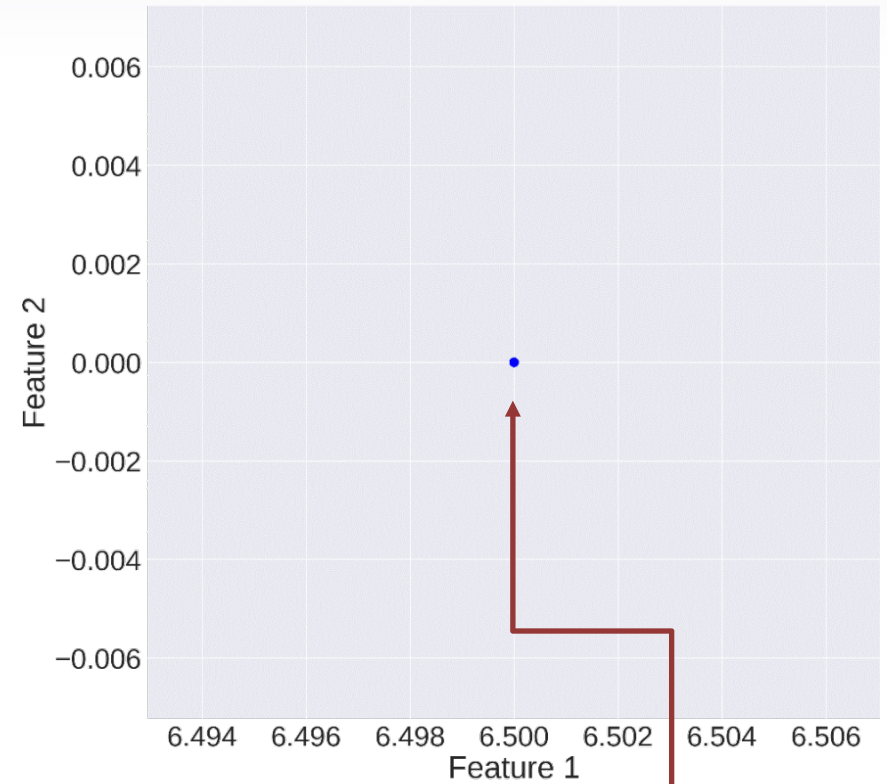
AdGroupId	AdNetwork Type2	MaxCpc	Slot	Clicks	Impressions	<i>is_incorrect</i>
78db044136	s	0.28	s_2	3	0	<i>True</i>
68a0110c33	s	1	s_2	1	13	<i>False</i>
2r39fw11w3	p	1.2	p_1	3	419	<i>False</i>

Understand how the data was generated

Train



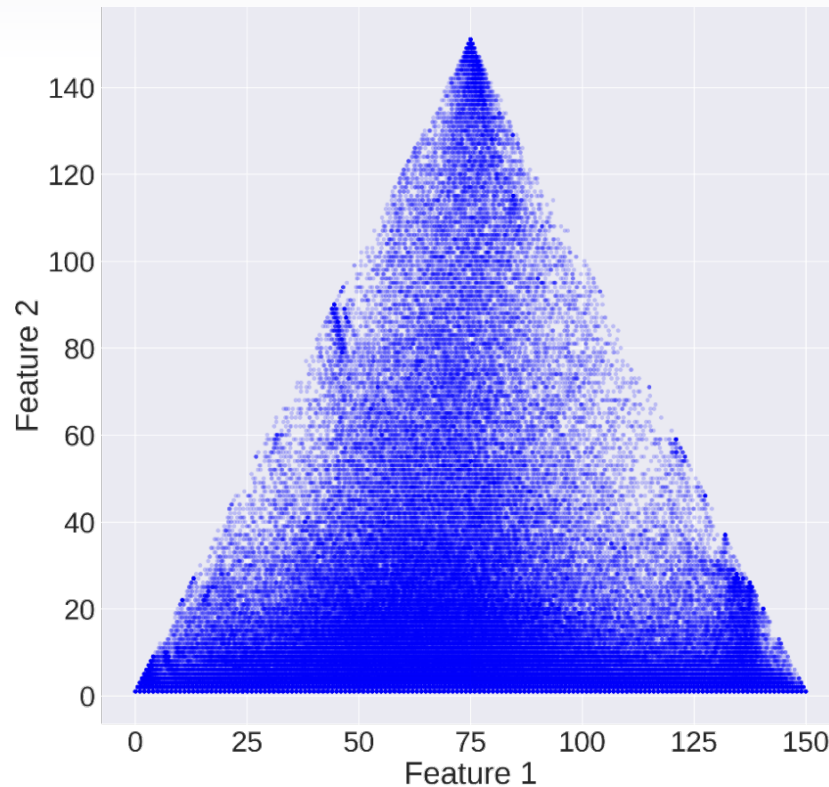
Test



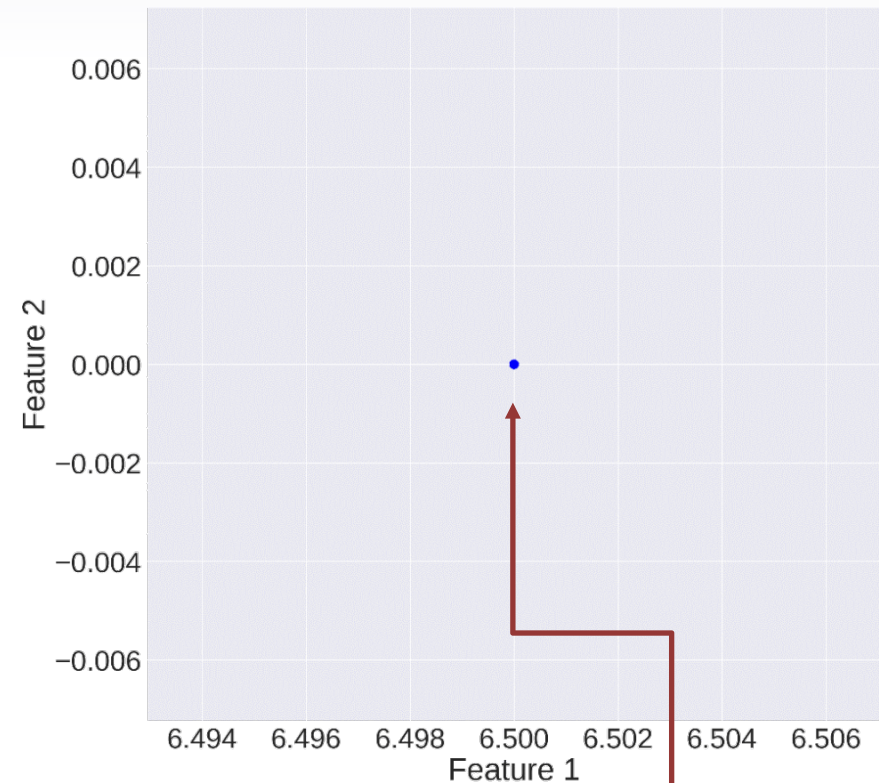
Dot!

Understand how the data was generated

Train



Test



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

Dot!

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