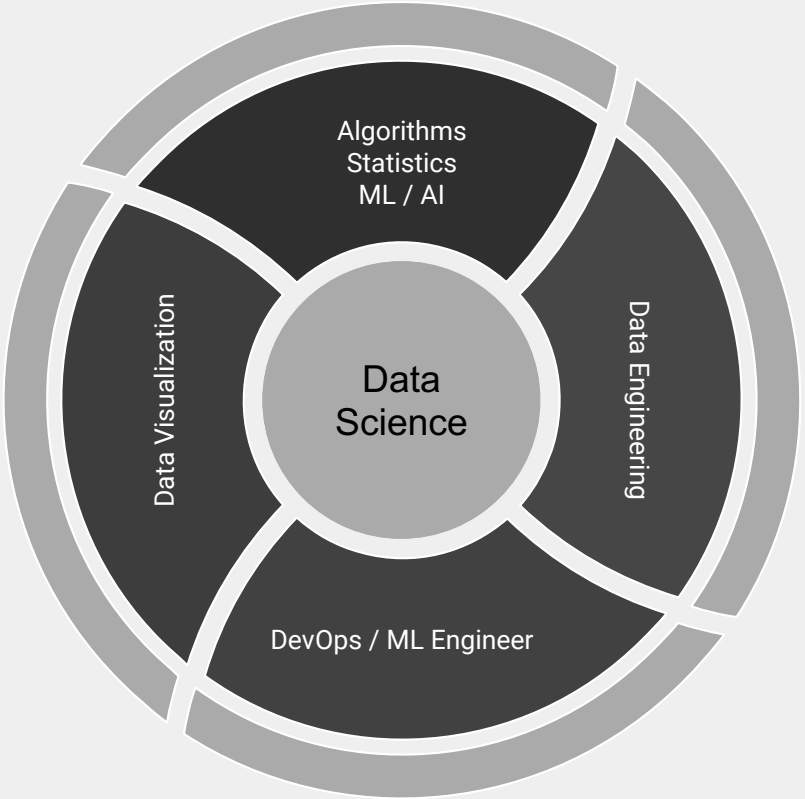
Grokking Data Science

DivyaJyoti (DJ) Rajdev
Aug 21, 2019

Contents

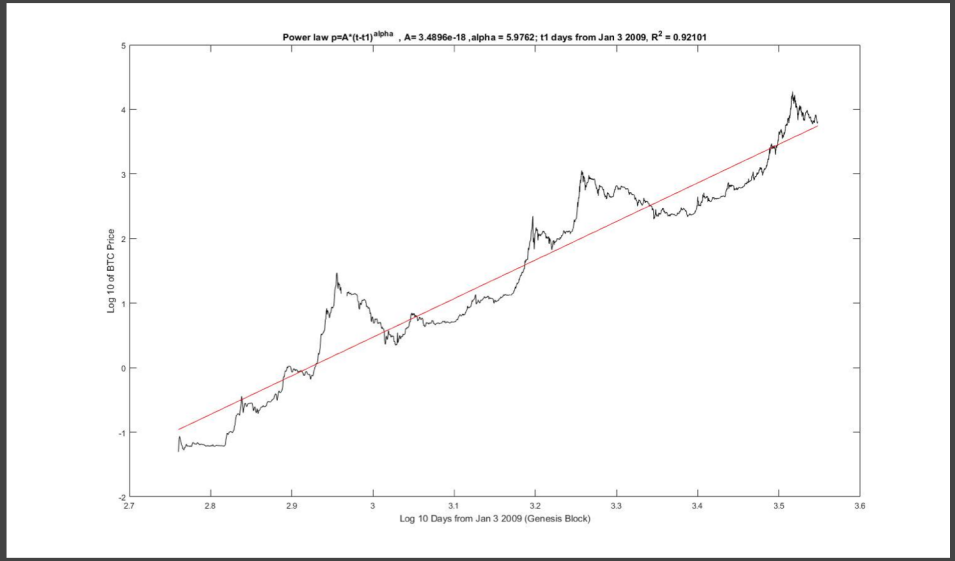
- what makes up data science?
- should I even solve it?
- data or dirt?
- is the model good enough?
- how can I improve it?
- what is driving my model?
- what’s the long term plan?
- references

What makes up data science?



Should I even solve it?

- **be wary of black swans**
- define the right problem
- build backwards from the ideal solution
- prediction or insight
- representative and sufficient data
- goal post for success



source

Should I even solve it?

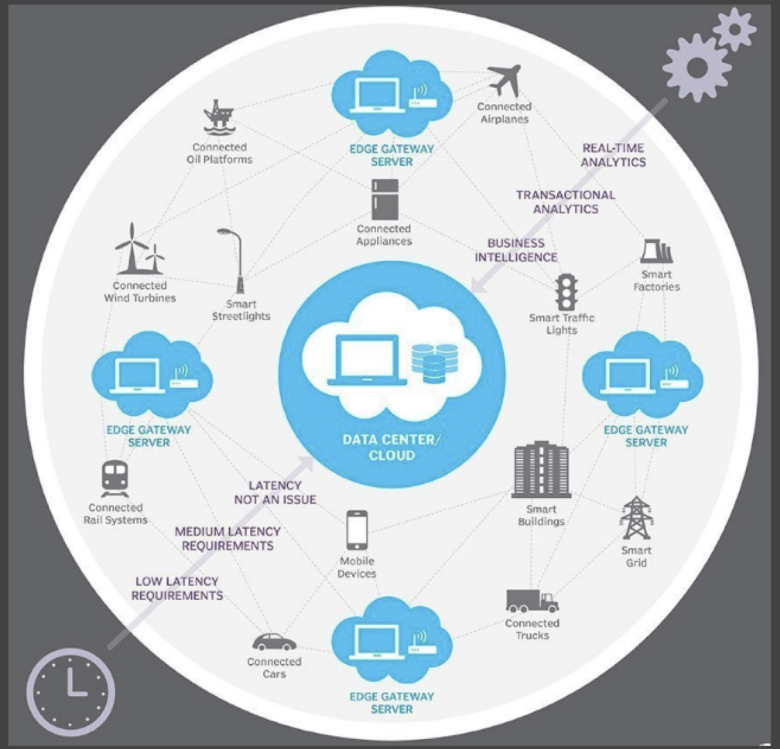
- be wary of black swans
- **define the right problem**
- build backwards from the ideal solution
- prediction or insight
- representative and sufficient data
- goal post for success

Customer is unhappy about service disruptions due to failures

improve product to have fewer failures ??

Should I even solve it?

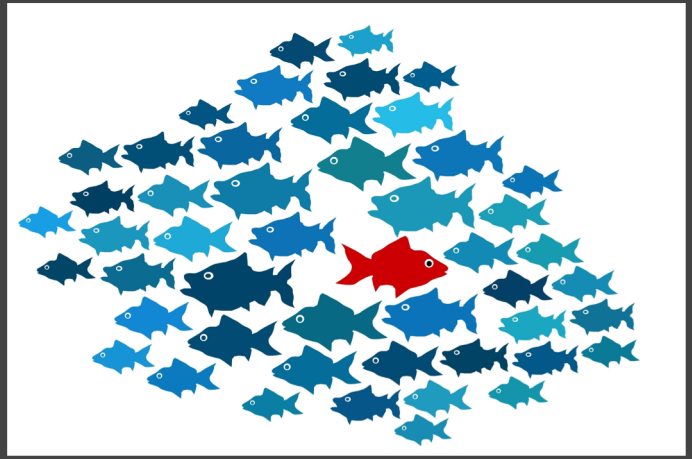
- be wary of black swans
- define the right problem
- **build backwards from the ideal solution**
- prediction or insight
- representative and sufficient data
- goal post for success



source

Should I even solve it?

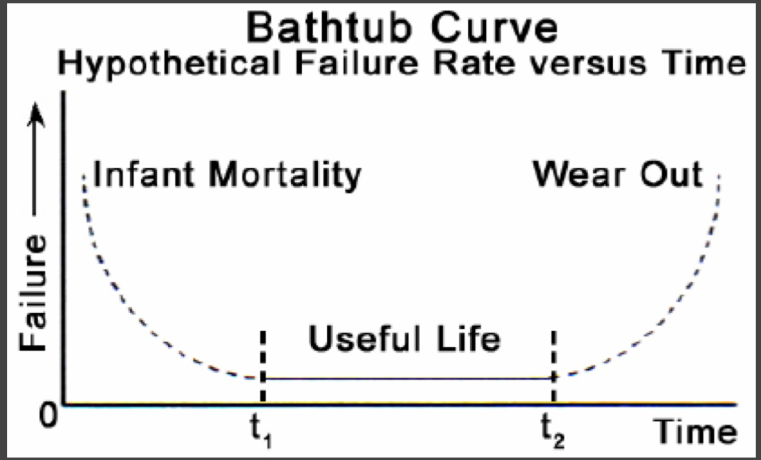
- be wary of black swans
- define the right problem
- build backwards from the ideal solution
- **prediction or insight**
- representative and sufficient data
- goal post for success



source

Should I even solve it?

- be wary of black swans
- define the right problem
- build backwards from the ideal solution
- prediction or insight
- **representative and sufficient data**
- goal post for success



source

Should I even solve it?

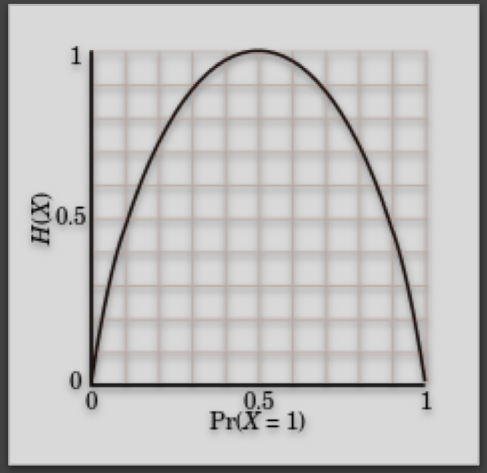
- be wary of black swans
- define the right problem
- build backwards from the ideal solution
- prediction or insight
- representative and sufficient data
- **goal post for success**

	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Tablet and Desktop Traffic 63,733 users	**99.99%**	**3.41%**	**1.99%**	**1.55%**	**1.05%**	**0.63%**	**0.16%**
Aug 13, 2017 - Aug 19, 2017 10,458 users	99.96%	3.35%	2.21%	1.67%	1.43%	1.08%	0.16%
Aug 20, 2017 - Aug 26, 2017 10,444 users	100.00%	3.54%	1.96%	1.77%	1.55%	0.18%	
Aug 27, 2017 - Sep 2, 2017 10,835 users	99.96%	3.78%	2.44%	2.41%	0.18%		
Sep 3, 2017 - Sep 9, 2017 9,735 users	100.00%	4.04%	2.96%	0.22%			
Sep 10, 2017 - Sep 16, 2017 9,876 users	100.00%	4.92%	0.32%				
Sep 17, 2017 - Sep 23, 2017 12,385 users	100.00%	1.32%					
Mobile Traffic 32,335 users	**99.97%**	**2.66%**	**0.90%**	**0.69%**	**0.58%**	**0.46%**	**0.00%**
Aug 13, 2017 - Aug 19, 2017 5,302 users	99.94%	2.81%	0.98%	0.64%	0.85%	0.81%	0.00%
Aug 20, 2017 - Aug 26, 2017 5,346 users	99.96%	2.49%	1.01%	1.12%	0.75%	0.11%	
Aug 27, 2017 - Sep 2, 2017 4,511 users	99.93%	2.33%	1.13%	0.84%	0.07%		
Sep 3, 2017 - Sep 9, 2017 4,541 users	99.98%	3.28%	0.99%	0.09%			
Sep 10, 2017 - Sep 16, 2017 4,915 users	99.98%	4.46%	0.39%				
Sep 17, 2017 - Sep 23, 2017 7,720 users	100.00%	1.36%					

source

Data or dirt?

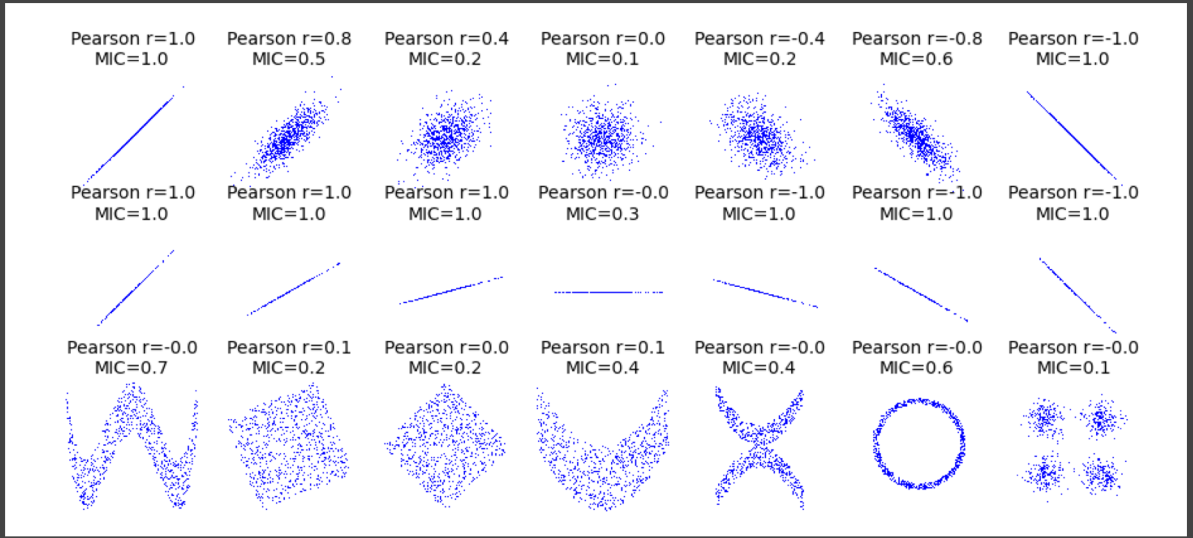
- **explore data summaries**
- understand (non parametric) data relationships
- think about featurization and transforms
- treatment for outliers and high leverage points
- missing value imputation
- decompose time series



source

Data or dirt?

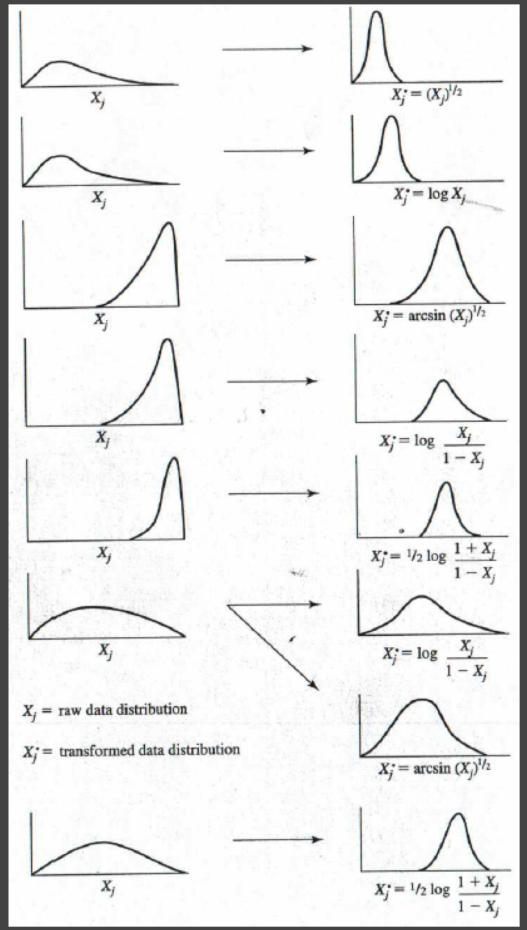
- explore data summaries
- **understand (non parametric) data relationships**
- think about featurization and transforms
- treatment for outliers and high leverage points
- missing value imputation
- decompose time series



source

Data or dirt?

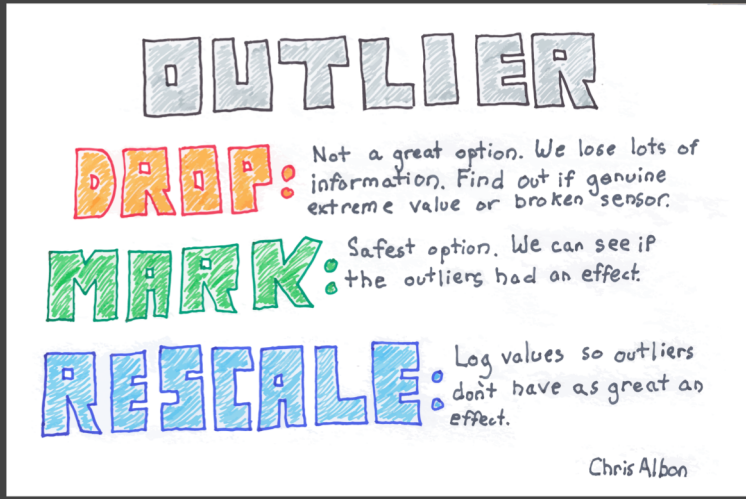
- explore data summaries
- understand (non parametric) data relationships
- **think about featurization and transforms**
- treatment for outliers and high leverage points
- missing value imputation
- decompose time series



source

Data or dirt?

- explore data summaries
- understand (non parametric) data relationships
- think about featurization and transforms
- **treatment for outliers and high leverage points**
- missing value imputation
- decompose time series

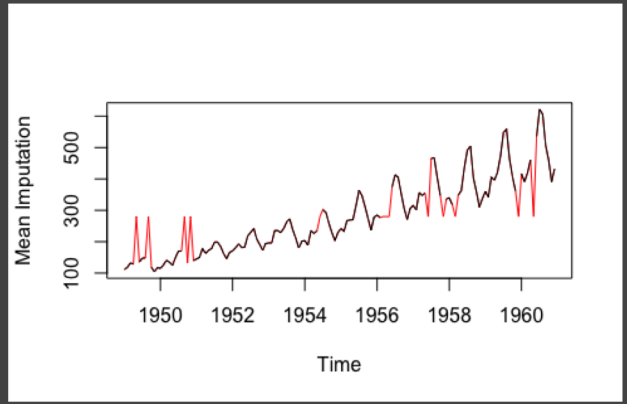


source

Data or dirt?

- explore data summaries
- understand (non parametric) data relationships
- think about featurization and transforms
- treatment for outliers and high leverage points
- **missing value imputation**
- decompose time series

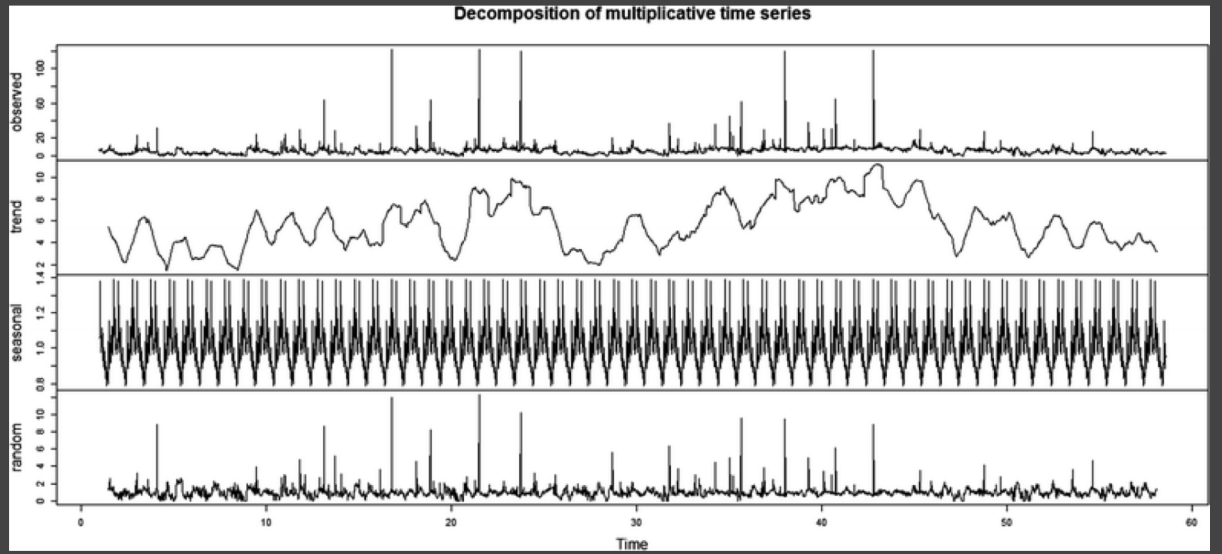
filter out
zero
mean, median
prev, next
cluster centroid, knn
groupwise aggregate
add new row



source

Data or dirt?

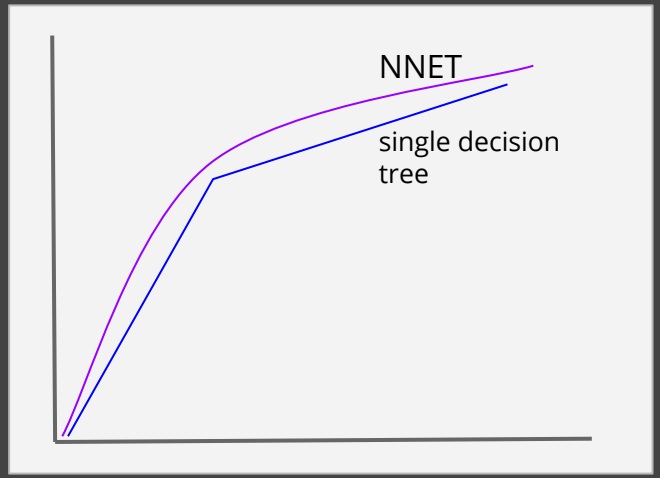
- explore data summaries
- understand (non parametric) data relationships
- think about featurization and transforms
- treatment for outliers and high leverage points
- missing value imputation
- **decompose time series**



source

Is the model good enough?

- **build baseline models**
- think about error metric
- use sampling
- choose the right algorithm family
- understand the breaking point of models
- explore domain specific methods
- develop intuition about statistics



Is the model good enough?

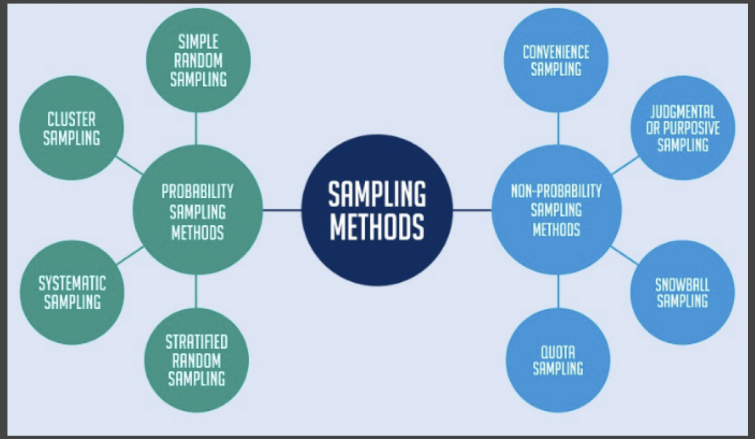
- build baseline models
- **think about error metric**
- use sampling
- choose the right algorithm family
- understand the breaking point of models
- explore domain specific methods
- develop intuition about statistics

Mean estimator:
MSE

Median estimator:
MAE

Is the model good enough?

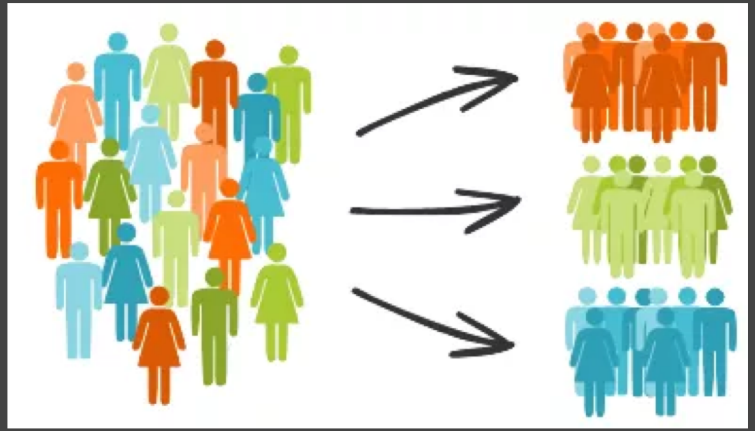
- build baseline models
- think about error metric
- **use sampling**
- choose the right algorithm family
- understand the breaking point of models
- explore domain specific methods
- develop intuition about statistics



source

Is the model good enough?

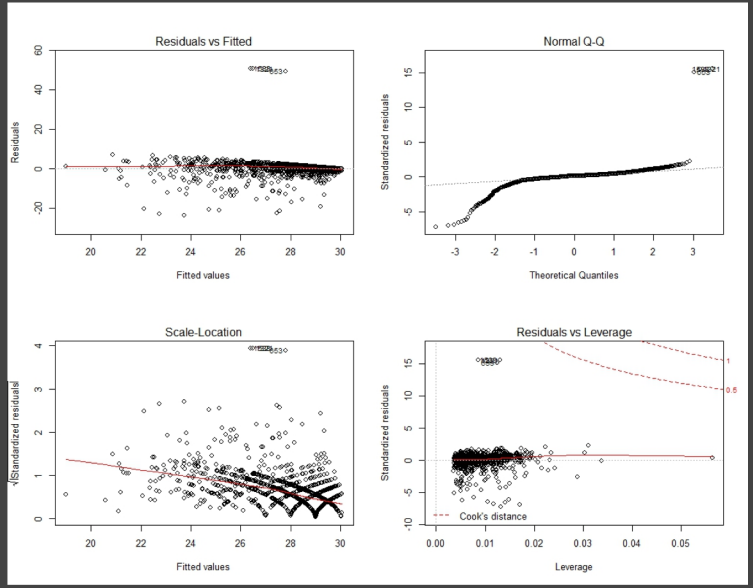
- build baseline models
- think about error metric
- use sampling
- **choose the right algorithm family**
- understand the breaking point of models
- explore domain specific methods
- develop intuition about statistics



source

Is the model good enough?

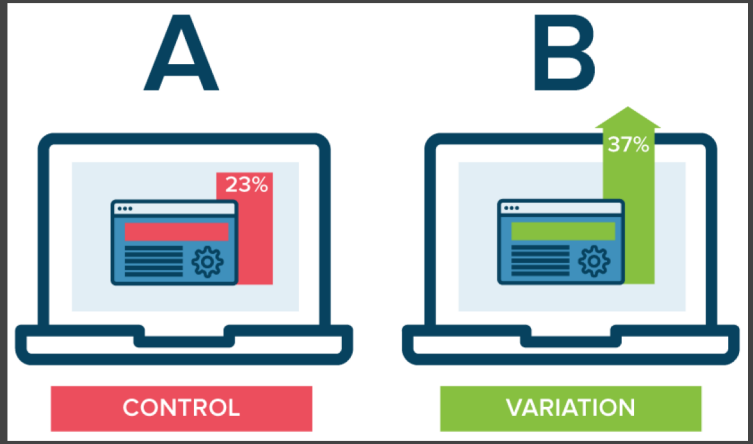
- build baseline models
- think about error metric
- use sampling
- choose the right algorithm family
- **understand the breaking point of models**
- explore domain specific methods
- develop intuition about statistics



source

Is the model good enough?

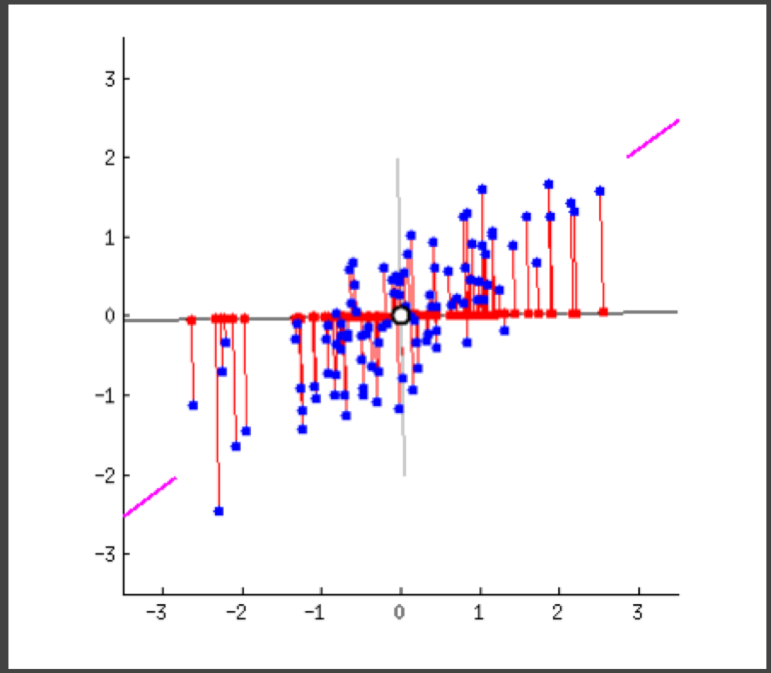
- build baseline models
- think about error metric
- use sampling
- choose the right algorithm family
- understand the breaking point of models
- **explore domain specific methods**
- develop intuition about statistics



source

Is the model good enough?

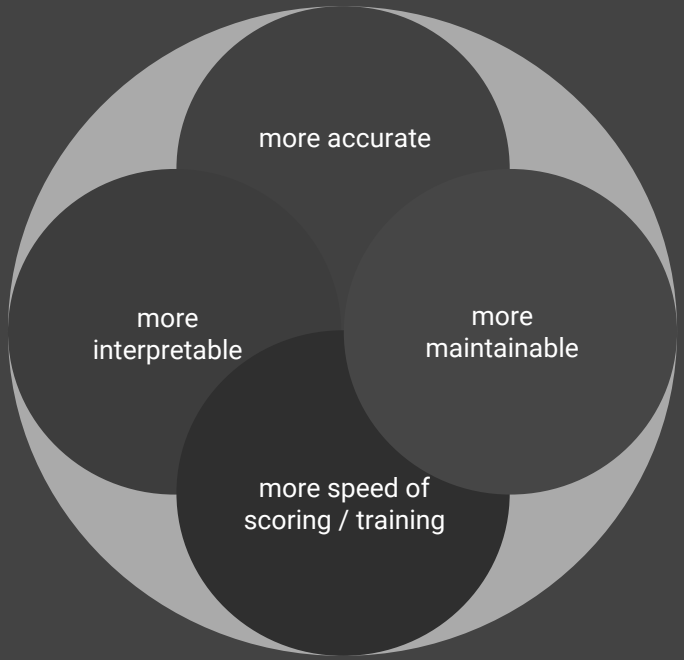
- build baseline models
- think about error metric
- use sampling
- choose the right algorithm family
- understand the breaking point of models
- explore domain specific methods
- **develop intuition about statistics**



source

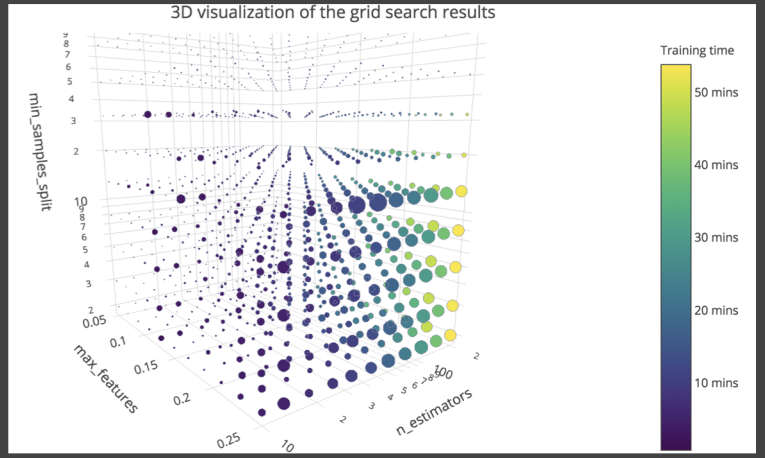
How can I improve the model?

- **define improvement**
- automate tuning via grid search
- try using a subset of predictors
- model stacking for better performance
- residuals should look like a cloud
- try semi supervised approaches



How can I improve the model?

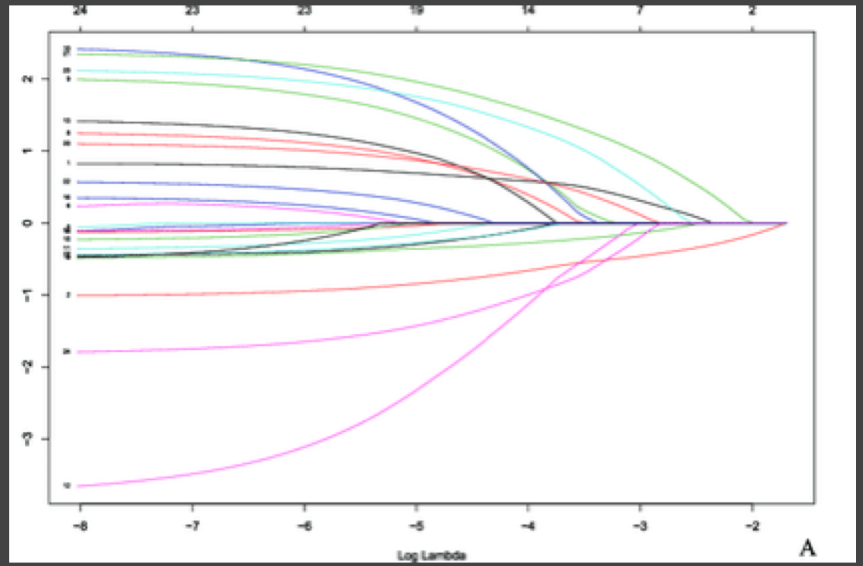
- define improvement
- **automate tuning via grid search**
- try using a subset of predictors
- model stacking for better performance
- residuals should look like a cloud
- try semi supervised approaches



source

How can I improve the model?

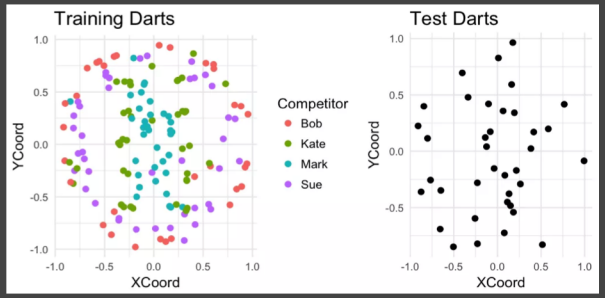
- define improvement
- automate tuning via grid search
- **try using a subset of predictors**
- model stacking for better performance
- residuals should look like a cloud
- try semi supervised approaches

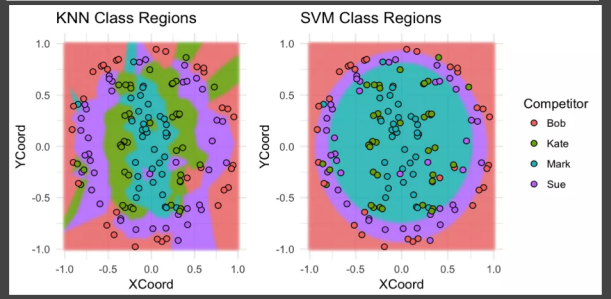


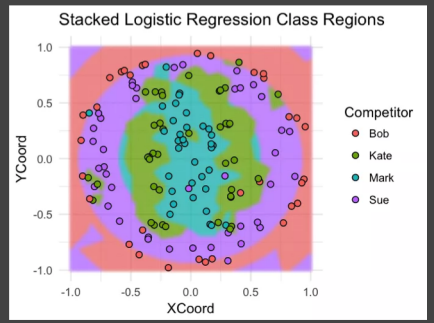
source

How can I improve the model?

- define improvement
- automate tuning via grid search
- try using a subset of predictors
- **model stacking for better performance**
- residuals should look like a cloud
- try semi supervised approaches



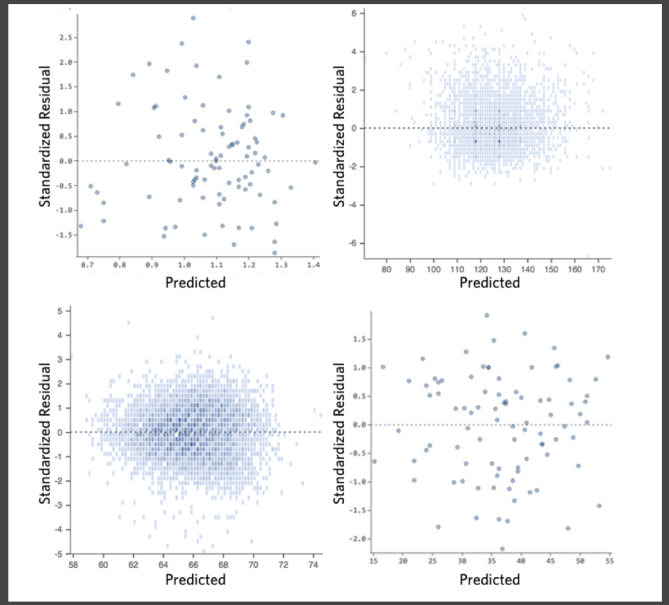




source

How can I improve the model?

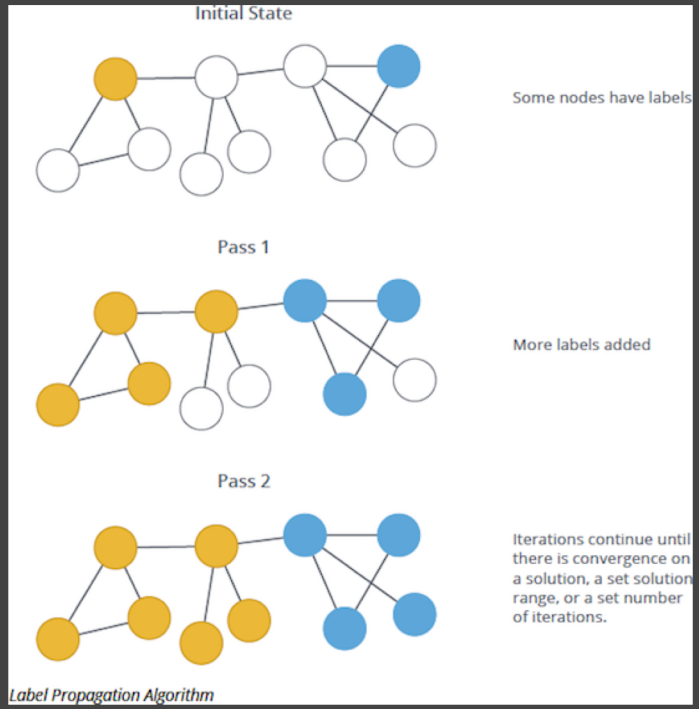
- define improvement
- automate tuning via grid search
- try using a subset of predictors
- model stacking for better performance
- **residuals should look like a cloud**
- try semi supervised approaches



source

How can I improve the model?

- define improvement
- automate tuning via grid search
- try using a subset of predictors
- model stacking for better performance
- residuals should look like a cloud
- **try semi supervised approaches**

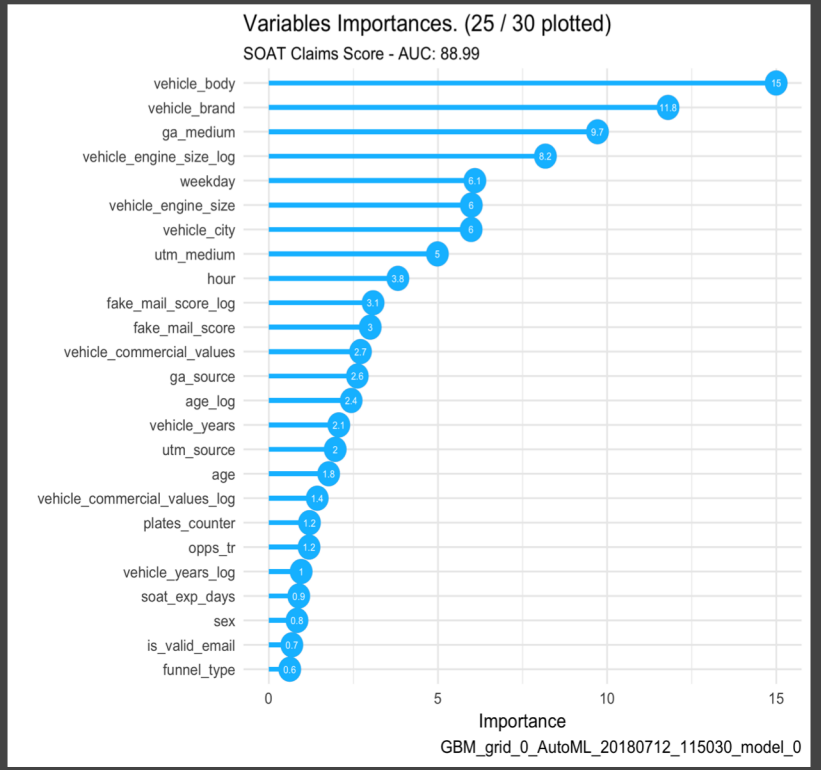


Label Propagation Algorithm

source

What is driving my model?

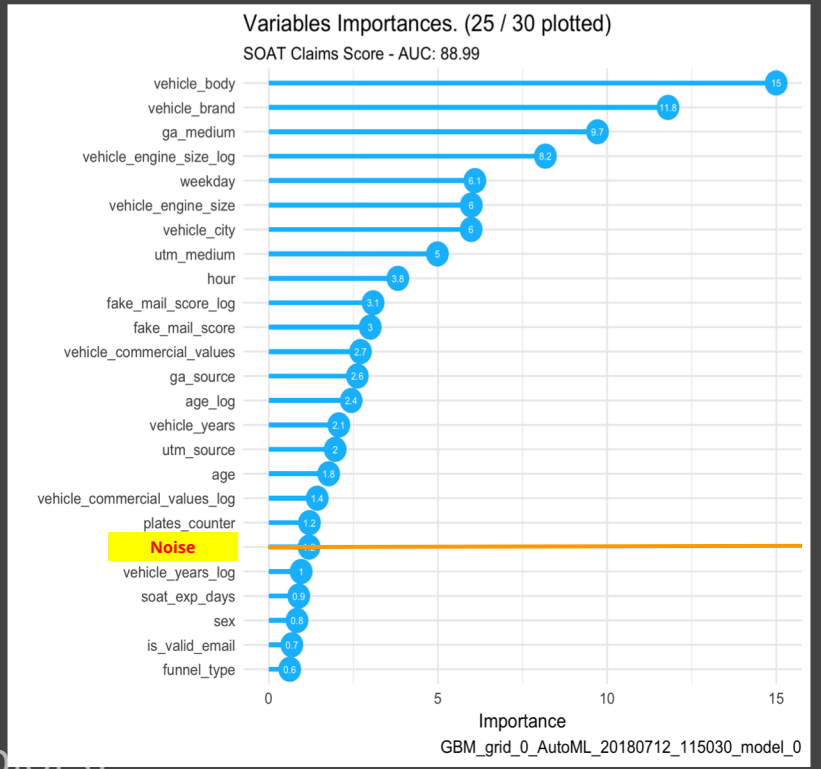
- **variable importance**
- variable importance with dummy noise
- the algorithm might be cheating
- time based split
- local interpretation: LIME
- do sensitivity analysis
  (partial dependence plots)



source

What is driving my model?

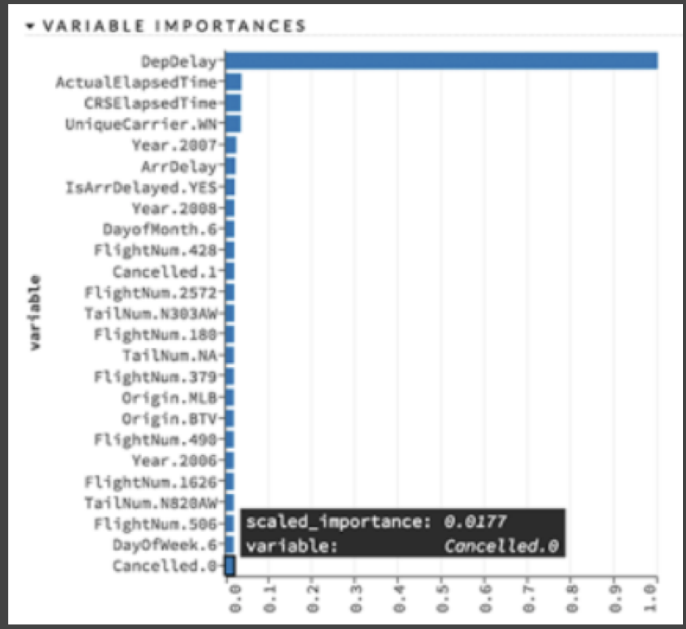
- variable importance
- **variable importance with dummy noise**
- the algorithm might be cheating
- time based split
- local interpretation: LIME
- do sensitivity analysis (partial dependence plots)



source with modifications

What is driving my model?

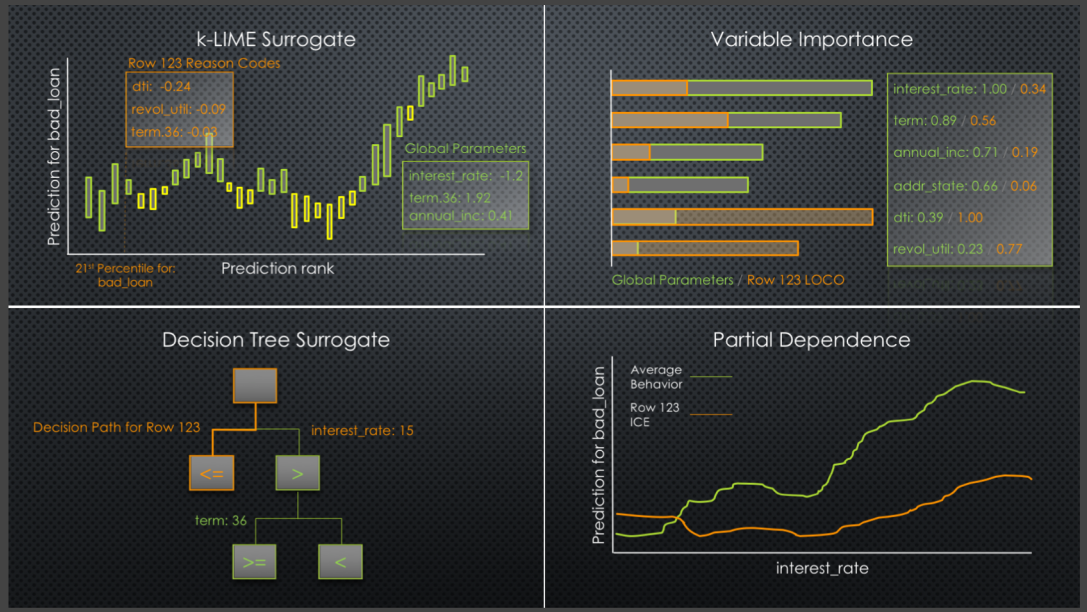
- variable importance
- variable importance with dummy noise
- **the algorithm might be cheating**
- **time based split**
- local interpretation: LIME
- do sensitivity analysis (partial dependence plots)



source

What is driving my model?

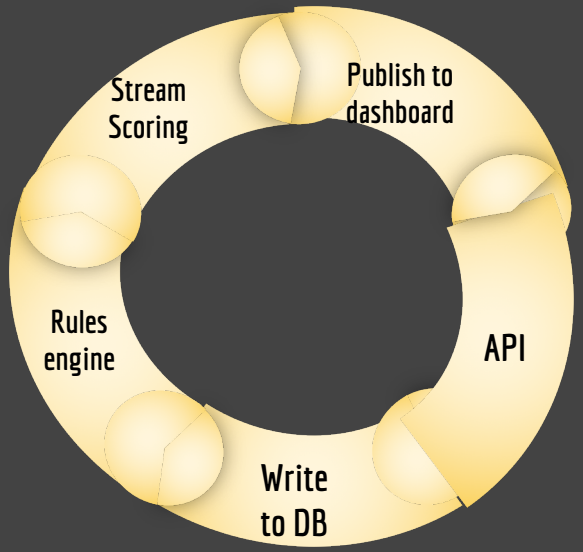
- variable importance
- variable importance with dummy noise
- the algorithm might be cheating
- time based split
- **local interpretation: LIME**
- **do sensitivity analysis (partial dependence plots)**



source

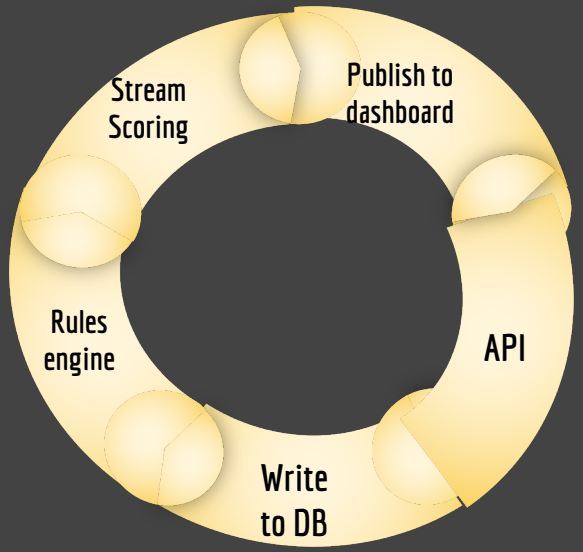
What's the long term plan?

- **accessibility of the model / results**
- data science needs testing too
- turn one time insights to rules
- present for the right audience
- plan for retraining model
- side effects on other initiatives



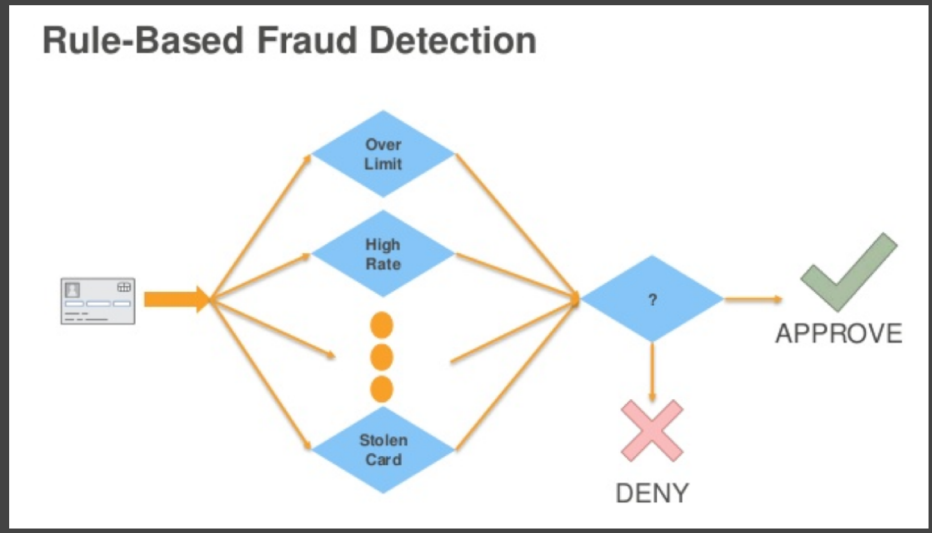
What's the long term plan?

- accessibility of the model / results
- **data science needs testing too**
- turn one time insights to rules
- present for the right audience
- plan for retraining model
- side effects on other initiatives



What's the long term plan?

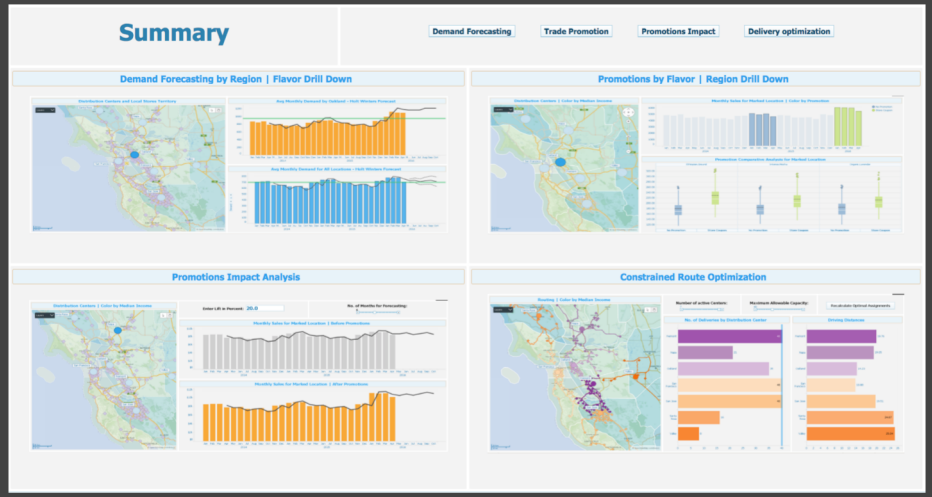
- accessibility of the model / results
- data science needs testing too
- **turn one time insights to rules**
- present for the right audience
- plan for retraining model
- side effects on other initiatives

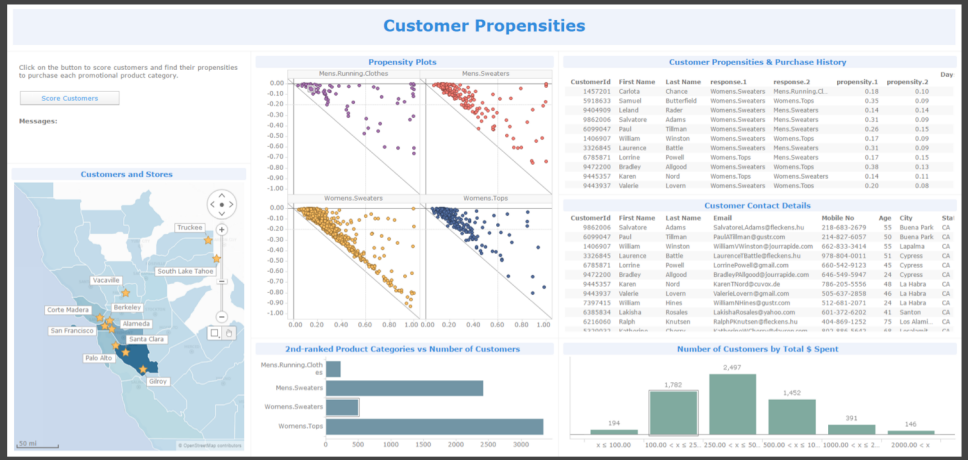


source

What's the long term plan?

- accessibility of the model / results
- data science needs testing too
- turn one time insights to rules
- **present for the right audience**
- plan for retraining model
- side effects on other initiatives

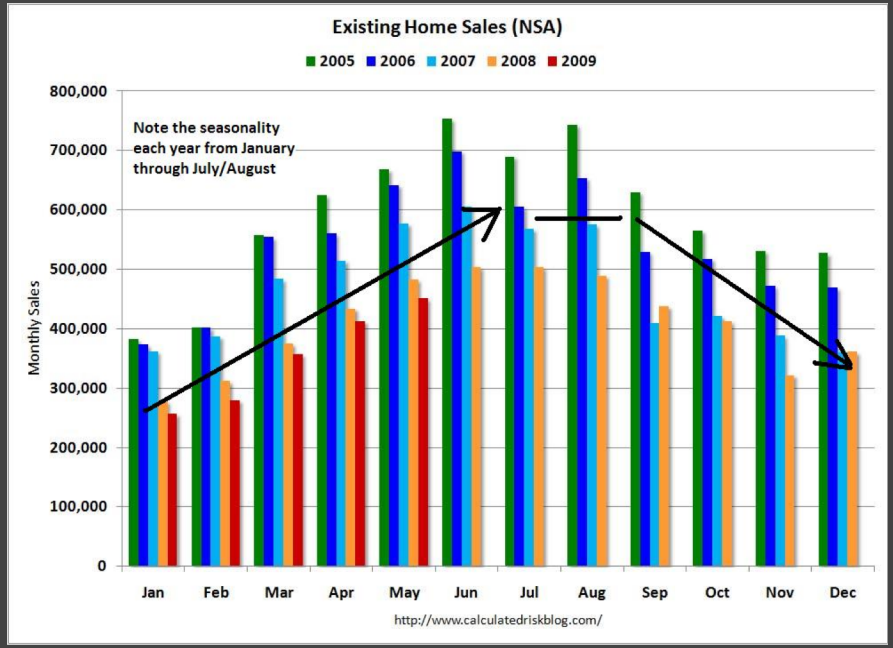




source

What's the long term plan?

- accessibility of the model / results
- data science needs testing too
- turn one time insights to rules
- present for the right audience
- **plan for retraining model**
- side effects on other initiatives



source

What's the long term plan?

- accessibility of the model / results
- data science needs testing too
- turn one time insights to rules
- present for the right audience
- plan for retraining model
- **side effects on other initiatives**

source

Questions?

DivyaJyoti (DJ) Rajdev

dj@djrajdev.com

twitter: @DivjyotiRajdev