End2End ML

Lior Sidi & Noa Lubin

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

What you will learn today

Formulating a business problem to a machine learning one
Best practices in data science

Creating real impact with your models by scoping it

Evaluating and hypertunning models
Understand your models and improving them

How to become a Lego-Master-Builder

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

What you will **NOT** learn today

Coding best practices (but you might see some nice python tricks)
MLOps and model deployment
Scaling learning on bid datasets
Learn how to do nice plots
New algorithm

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

You are a...

A Data scientist who works in a big electronic online retailer (i.e Best Buy).
Millions of users enter the company sites daily.

KPI is to improve sales on the site:

- Improve experience
- Improve offerings

You build superpowers

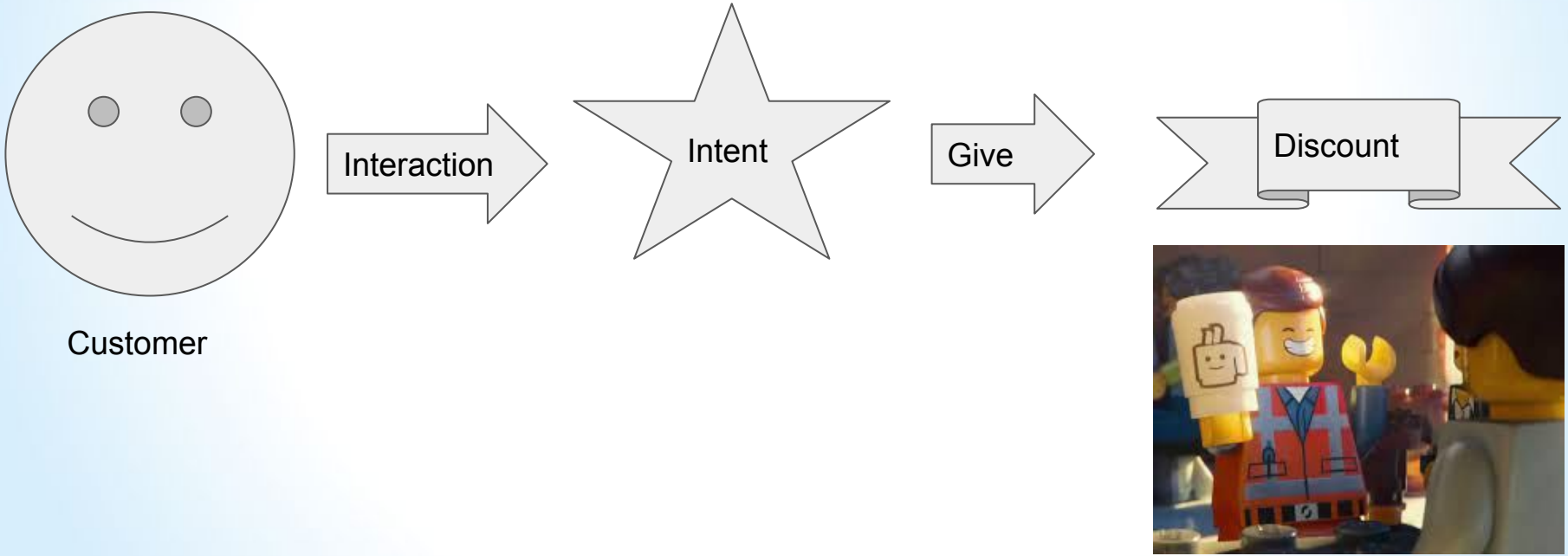
- Improve experience
  - Product recommendation
  - Chatbot & Sentiment analysis
  - Improve sites navigation
- Improve offerings
  - Optimize price with discounts, coupons and bundles
  - Identify new products to sell
  - Inventory prediction

Why personalization is important?

Each customer has different:

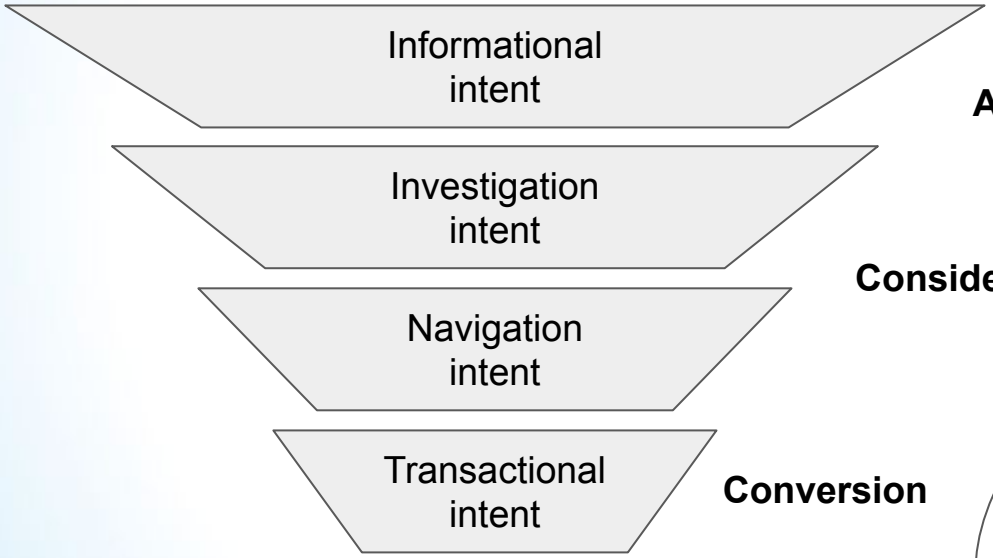
- Products needs
- Budget
- Intent

What is intent?



Customer Purchase Intent

*The willingness of a customer to buy a product or service in a certain condition**



All sessions

Awareness

Product views

Consideration

Add to Cart

Conversion

Purchase

*https://www.igi-global.com/dictionary/purchase-intention/24144

Y-DATA
SCHOOL OF DATA SCIENCE

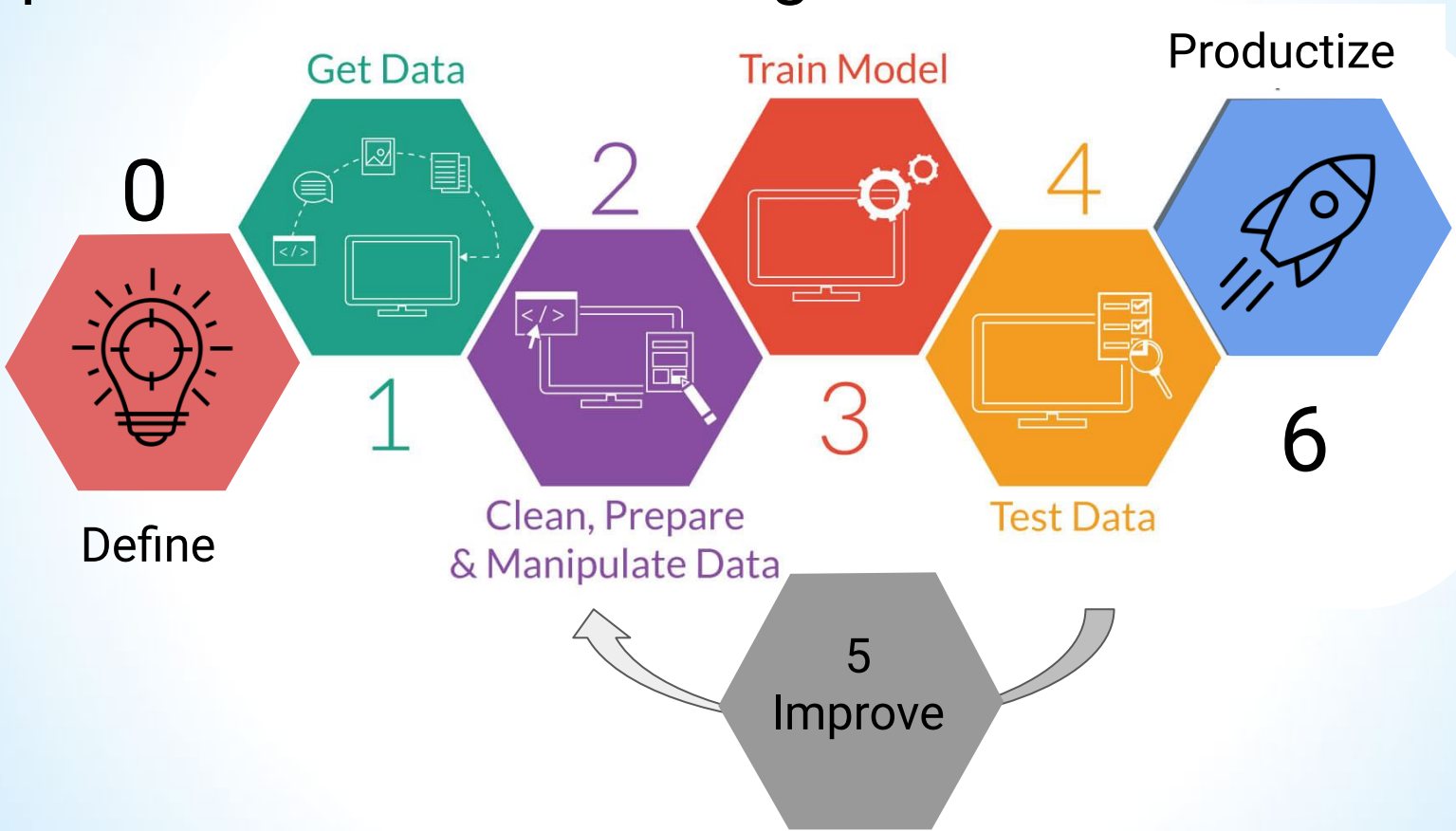
Yandex

How Customer interaction looks like?

user_session	user_id	event_time	category_code	product_id	brand	price	event_type
CxMKMQDRAN	1515915625610973155	2021-02-28 19:13:22 UTC	computers.peripherals.printer	3829912	pantum	122.86	view
CxMKMQDRAN	1515915625610973155	2021-02-28 19:15:40 UTC	computers.peripherals.printer	3829913	pantum	116.05	view
CxMKMQDRAN	1515915625610973155	2021-02-28 22:56:17 UTC	computers.peripherals.printer	3829913	pantum	116.05	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:01:14 UTC	computers.peripherals.printer	500058	pantum	67.00	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:02:38 UTC	computers.peripherals.printer	500058	pantum	67.00	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:03:42 UTC	computers.peripherals.printer	500058	pantum	67.00	cart
CxMKMQDRAN	1515915625610973155	2021-02-28 23:08:57 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:20:48 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:23:11 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:26:07 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:43:24 UTC	computers.peripherals.printer	3829912	pantum	122.86	view

https://www.kaggle.com/mkechinov/ecommerce-events-history-in-electronics-store

Steps to Predictive Modeling



Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Let's Start!!!!


EVERYTHING'S NOT AWESOME

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Wake up from LA LA LAND!

Split

- Based on time

Feature_extraction

- Deal with NA badly
- Aggregation keys on entire sessions and not windows
- Non relevant aggregations function
- Unique_event_type + session size- Leakage!!!
- Feature_extraction is also need to be fitted!

Wake up from LA LA LAND!

Split

- Based on time

```
train_X, train_y, test_X, test_y = split_data(df)
```

Feature_extraction

- Deal with NA badly

```
df_ = df_.dropna()
```

- Aggregation keys on entire sessions and not windows
- Non relevant aggregations function
- Unique_event_type + session size- Leakage!!!
- Feature_extraction is also need to be fitted!

```
df_agg = df_.groupby('user_session').agg({'price' : ['sum', 'min'], 'brand' : ['count'],
                                          'category_code' : ['count'],'product_id' : ['count'],
                                          'event_time' : ['count']}, axis="columns")
df_agg['unique_event_type'] = df_.groupby('user_session')['event_type'].nunique()
```

Wake up from LA LA LAND!

Model

- Didn’t hypertuned the model
- Didn’t tried other models - validation set?

Evaluation

- Using Accuracy for imbalance data
- Not using relevant metrics

Error analysis

- Understanding when your model mistakes and try to improve

Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

How to solve it?

- Gain EXPERIENCE
- Be skeptical
- Work with Product definition

Y-DATA
SCHOOL OF DATA SCIENCE

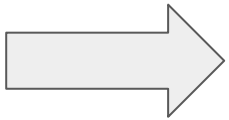
Yandex

Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

How to solve it?

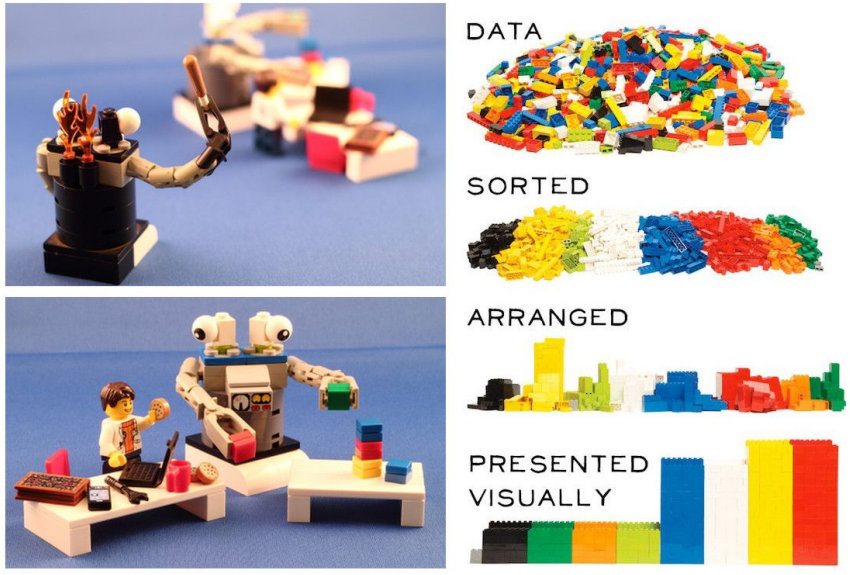
- Gain EXPERIENCE
- Be skeptical
- Work with Product definition



Do Exploratory Data Analysis

What we want to achieve in EDA

Understanding what and when we calling the model
Understanding how that data and features behave - do we have signal?
Identify possible pitfalls - evaluation, leakage, noise, imbalance

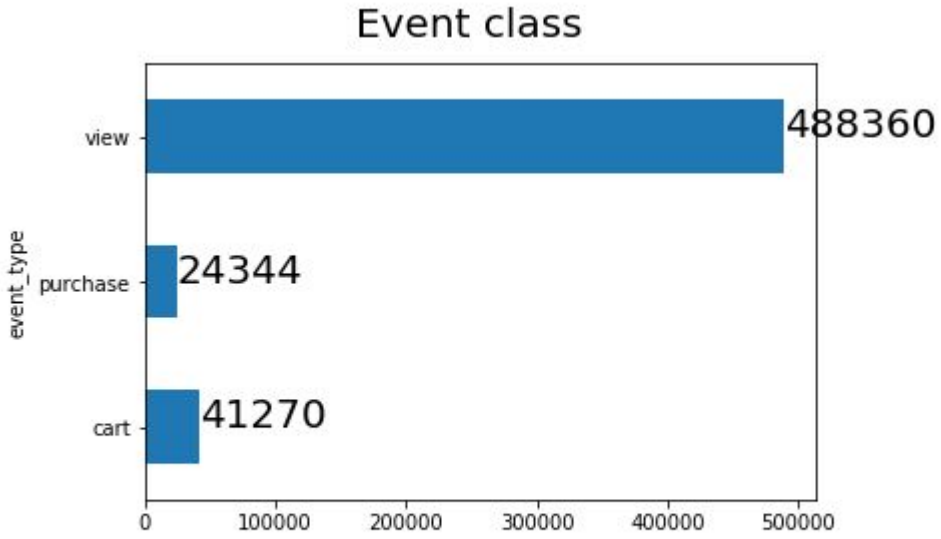


Back to reality - EDA

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Label distribution

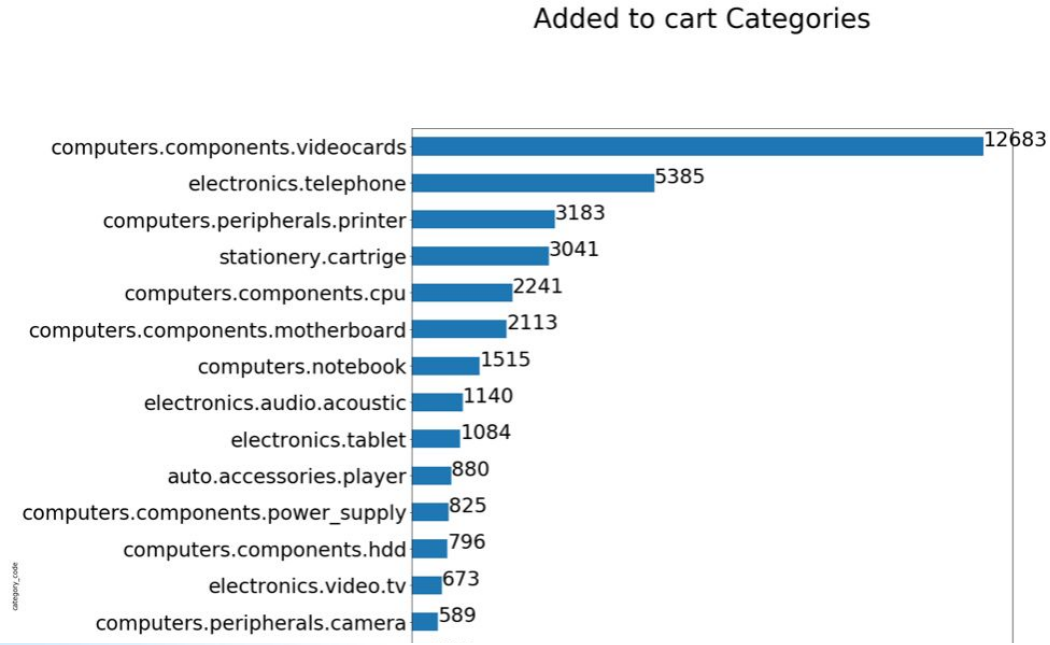


Labeling insights: there is imbalance, also might consider using cart over purchase

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Categories correlation with label



Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Catrgories add to cart ratio

category_code	view_count	cart_count	cart_ratio
computers.peripherals.camera	4369	589.0	0.134813
computers.components.videocards	97141	12683.0	0.130563
computers.peripherals.scanner	1600	200.0	0.125000
computers.components.hdd	7549	796.0	0.105444
computers.components.cpu	21314	2241.0	0.105142
computers.components.power_supply	8050	825.0	0.102484
computers.ebooks	2827	268.0	0.094800
stationery.cartrige	32939	3041.0	0.092322
computers.components.motherboard	23221	2113.0	0.090995
electronics.video.projector	1372	118.0	0.086006
computers.peripherals.printer	37479	3183.0	0.084928
construction.tools.painting	451	35.0	0.077605
electronics.video.tv_remote	930	68.0	0.073118
computers.peripherals.wifi	6235	455.0	0.072975
electronics.audio.music_tools.piano	370	27.0	0.072973
electronics.telephone	74839	5385.0	0.071954

Products add to cart ratio

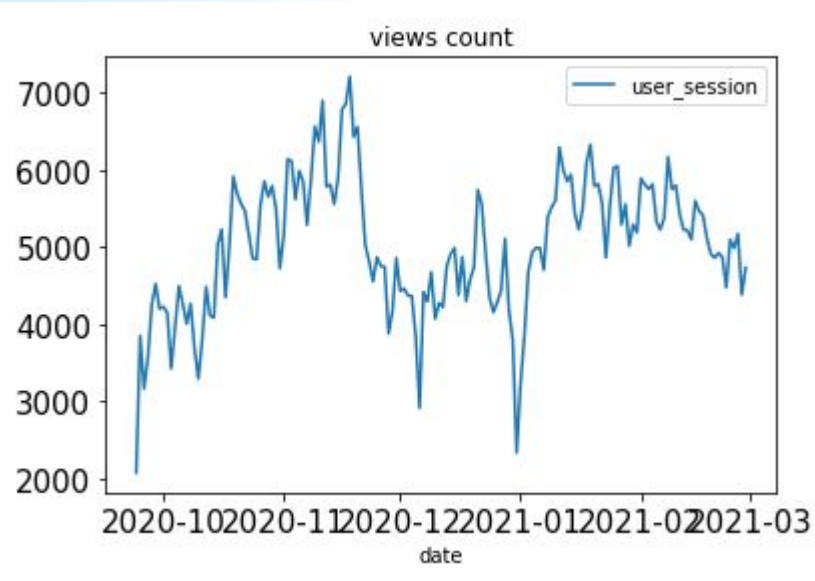
product_id	view_count	cart_count	cart_ratio
623426	106	51.0	0.481132
1586461	78	30.0	0.384615
1856480	181	66.0	0.364641
4013214	328	119.0	0.362805
4171147	167	52.0	0.311377
8093	178	54.0	0.303371
3581576	221	64.0	0.289593
1038724	194	55.0	0.283505
672145	99	28.0	0.282828
4013582	428	111.0	0.259346
866570	110	28.0	0.254545
4171037	143	36.0	0.251748
886023	169	42.0	0.248521
3829374	280	69.0	0.246429
665345	268	66.0	0.246269
3606492	394	93.0	0.236041
3699150	137	32.0	0.233577
841972	160	37.0	0.231250
893196	2866	662.0	0.230984
885572	299	69.0	0.230769
821773	126	29.0	0.230159
821628	158	36.0	0.227848

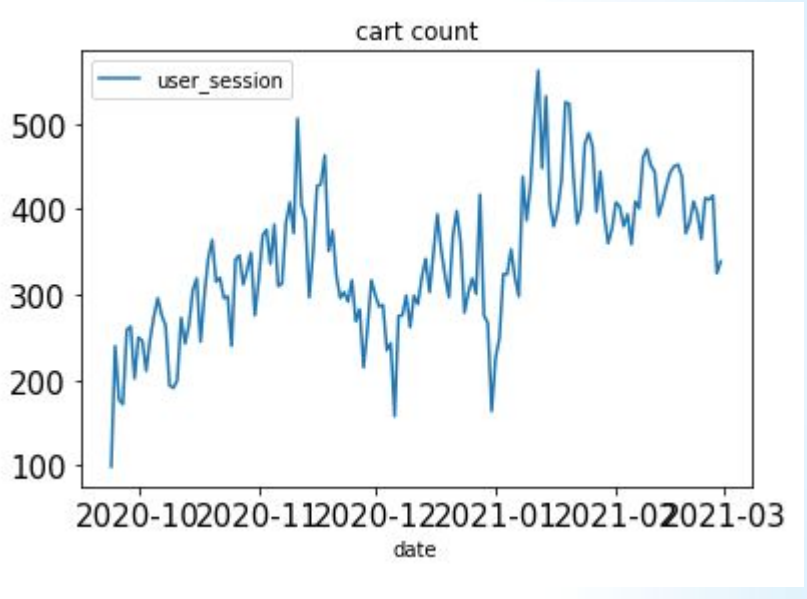
Categories insights: categories do matter for classification, need to incorporate them correctly in the features. like using the prior.

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

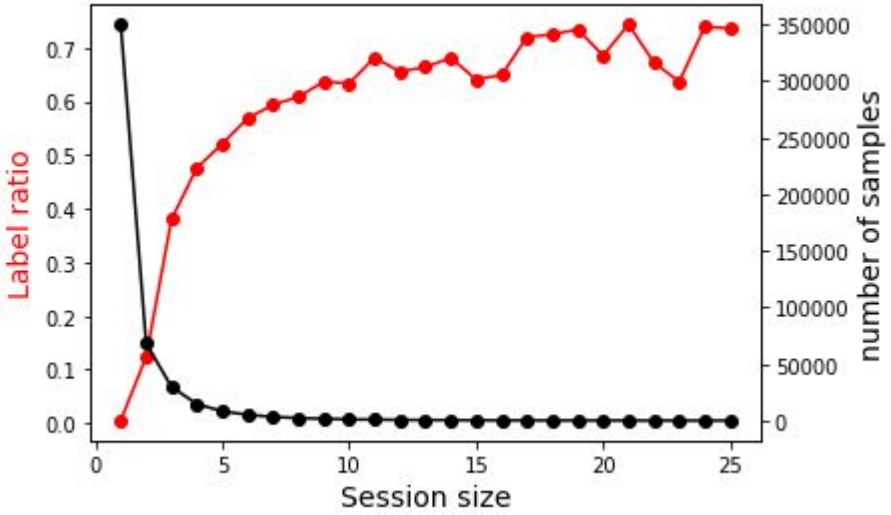
Macro understanding





Macro insights : there is seasonality in the data which requires from us to evaluate properly by time.

Session size analysis

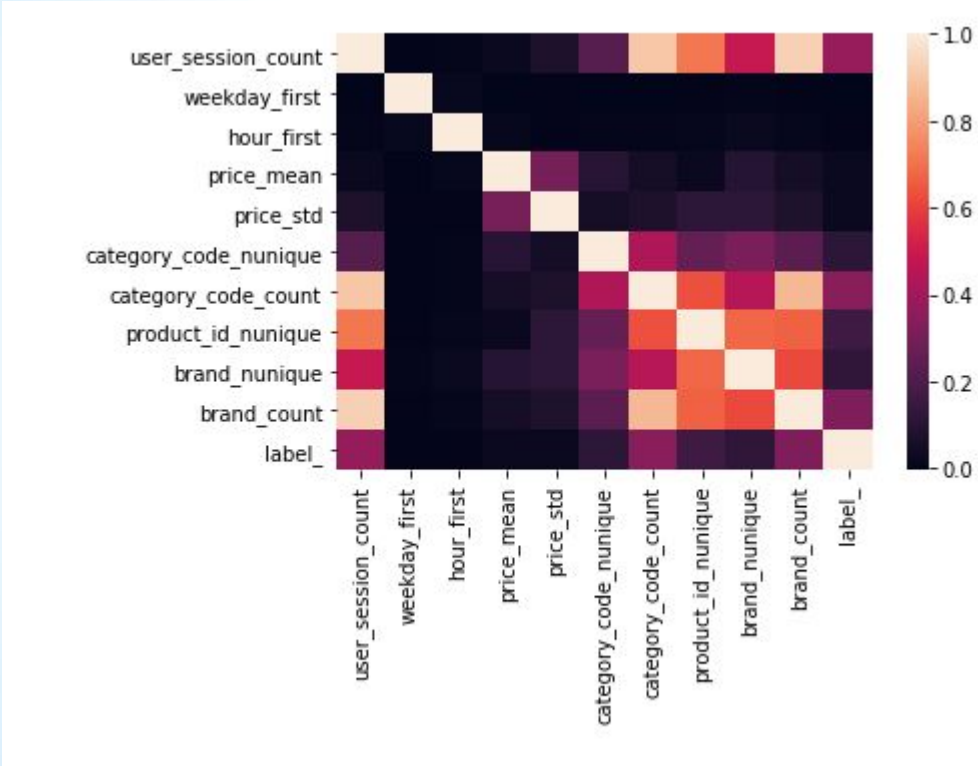


Session size insights: we can use 3 events that has signal of 40% target labels, this is a great start and will allow us to detect intent in very early stage.

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Features understanding



Features insights: there is a small correlation between the session count, barnd, and category code with the label. other features are not correlate and can be used together

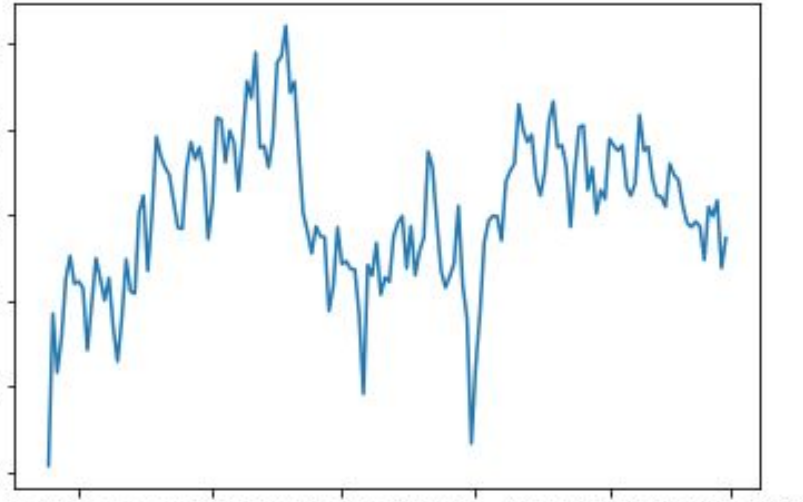
What did we learn?

The problem is imbalance
Some categories and product has strong prior
There is seasonality in the data
The session size is critical for framing the solution

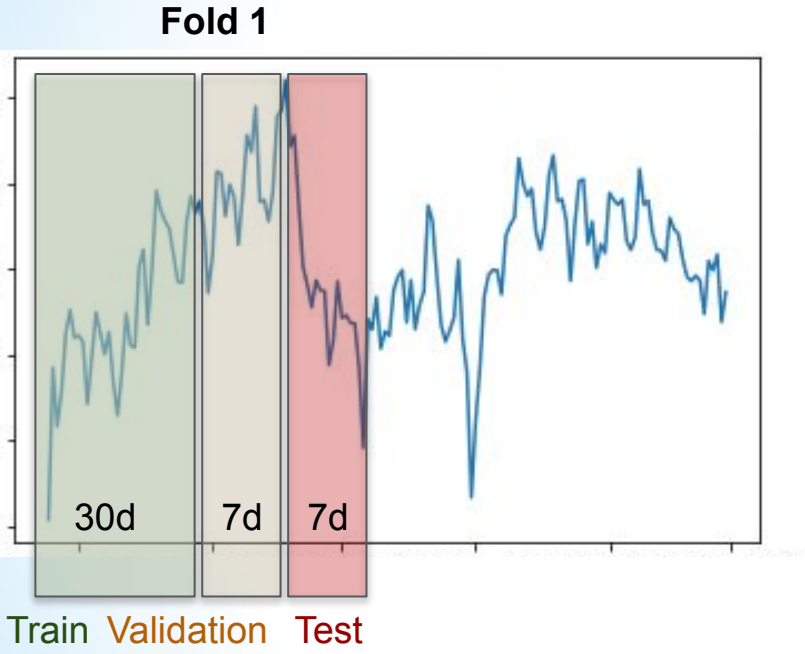
Also Try Pandas GUI -
https://towardsdatascience.com/pandasgui-analyzing-pandas-dataframes-with-a-graphical-user-interface-36f5c1357b1d

Back to reality - Modeling

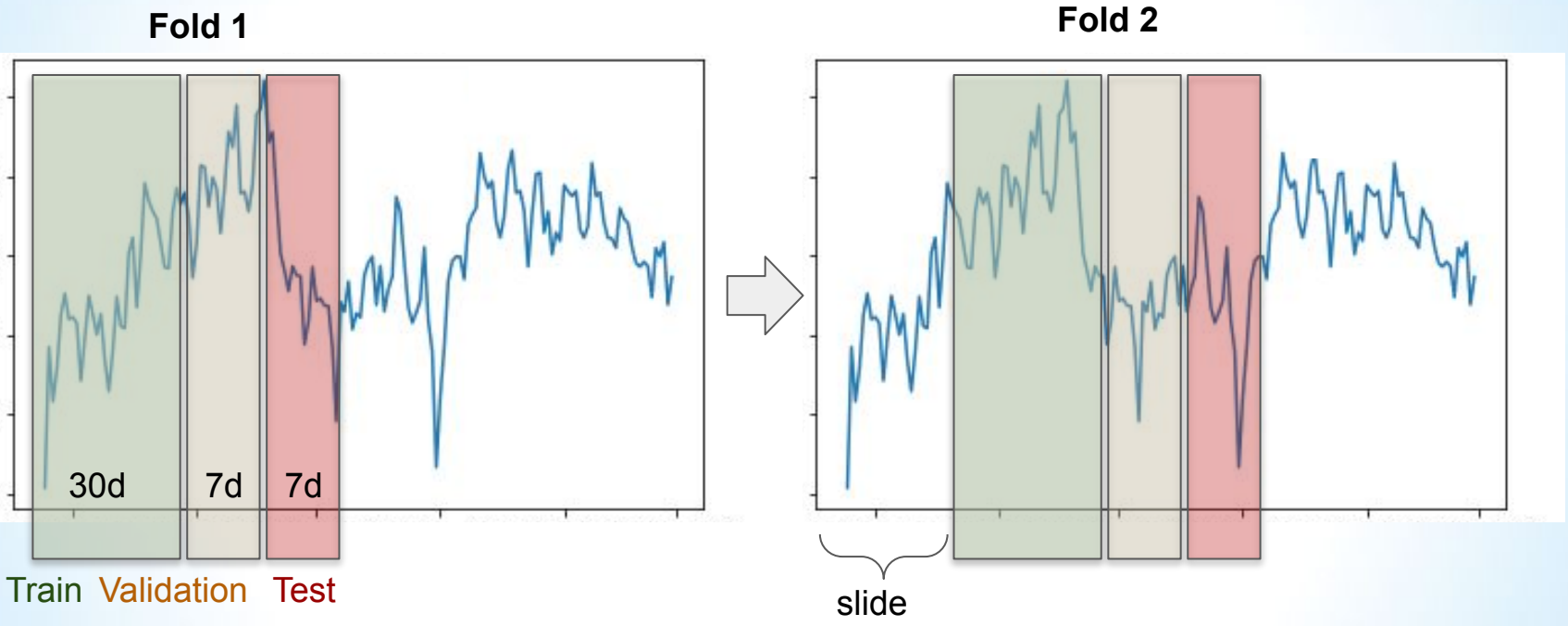
Temporal cross validation

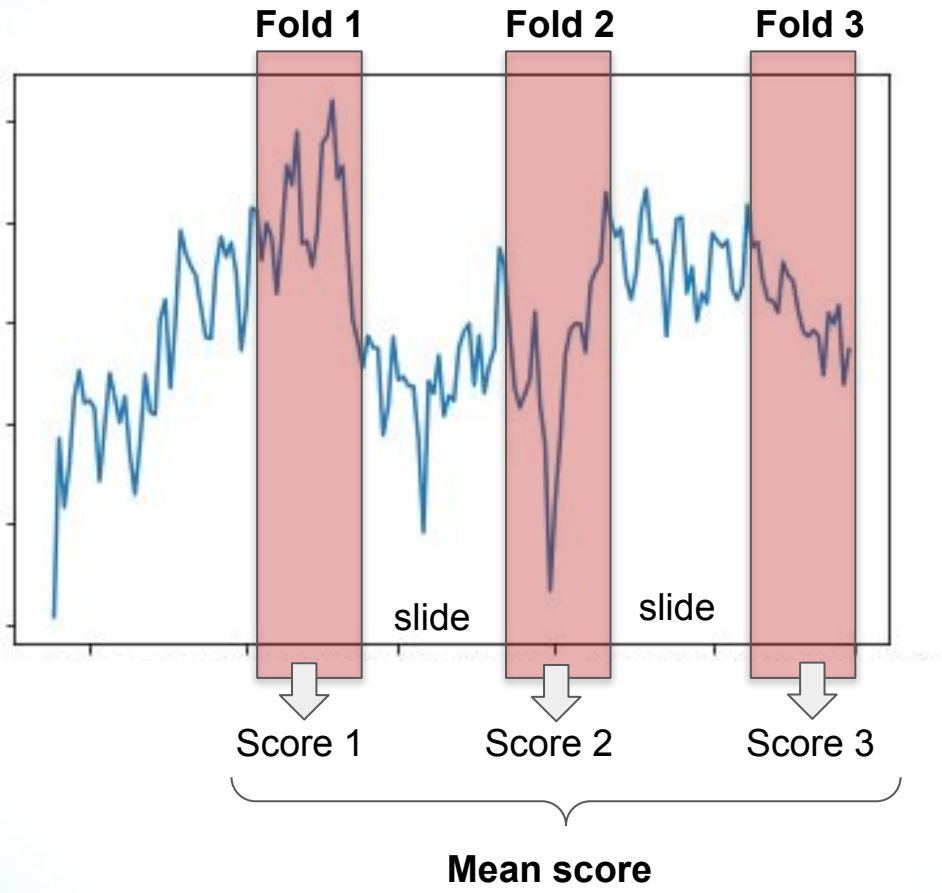


Temporal cross validation

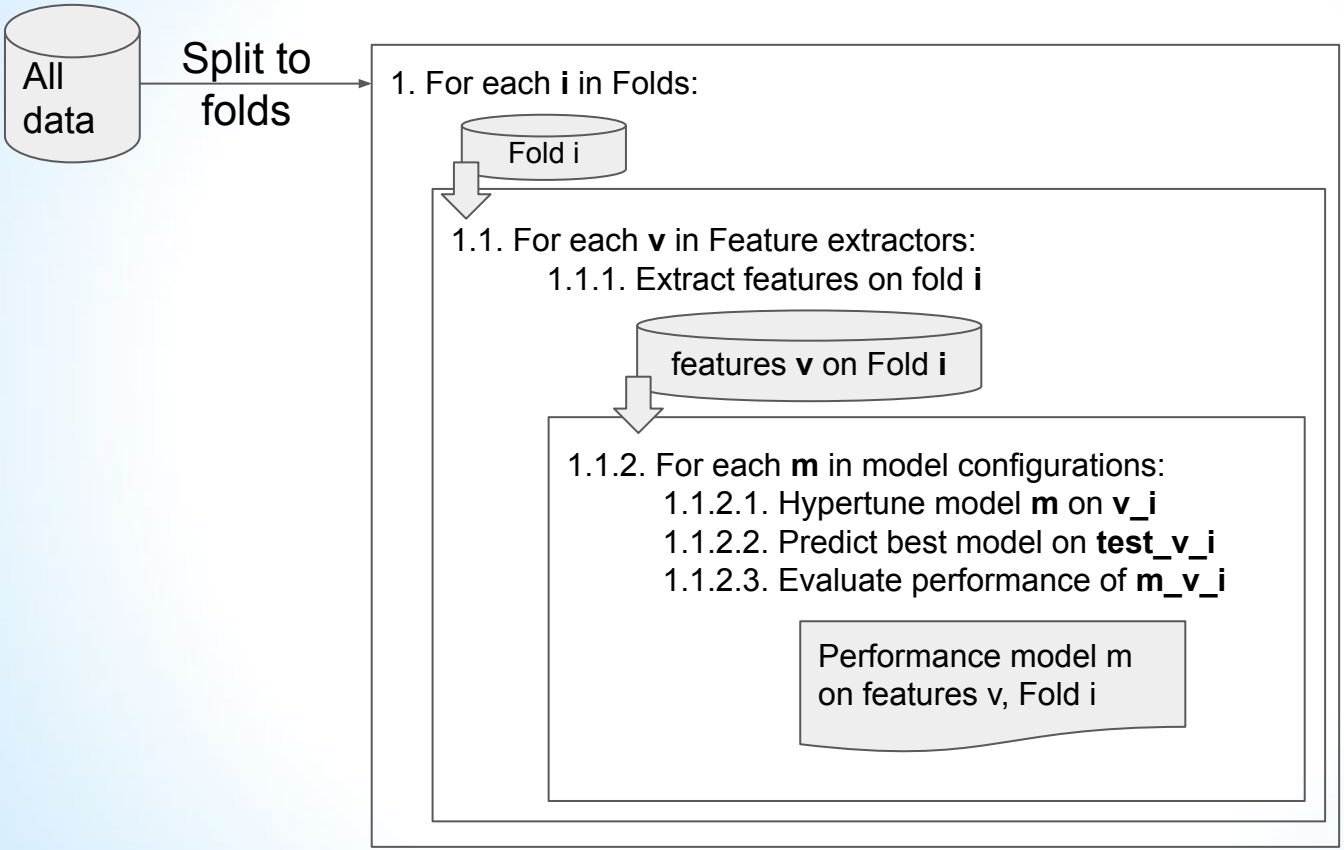


Temporal cross validation




Fold 1
Fold 2
Fold 3
slide
slide
Score 1
Score 2
Score 3
Mean score

The Evaluation loop

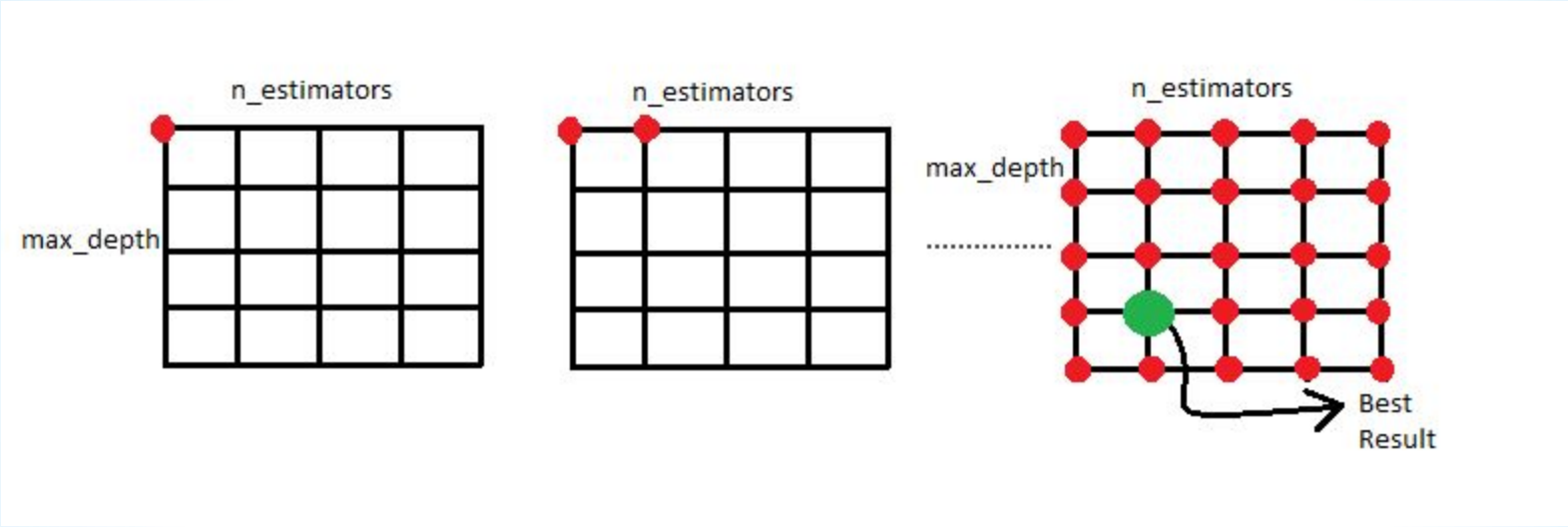


Y-DATA
SCHOOL OF DATA SCIENCE

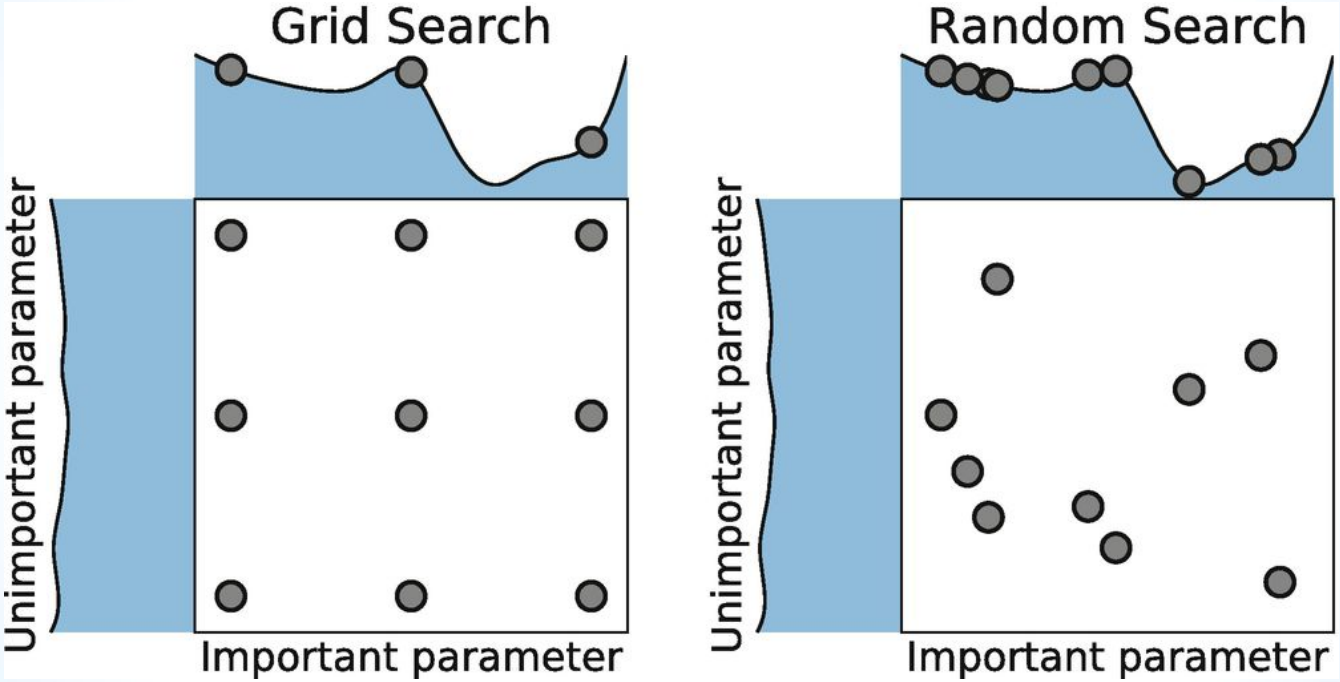
Yandex

Optimizing

Grid Search


n_estimators
max_depth
n_estimators
n_estimators
max_depth
Best
Result

Hyper parameter tuning

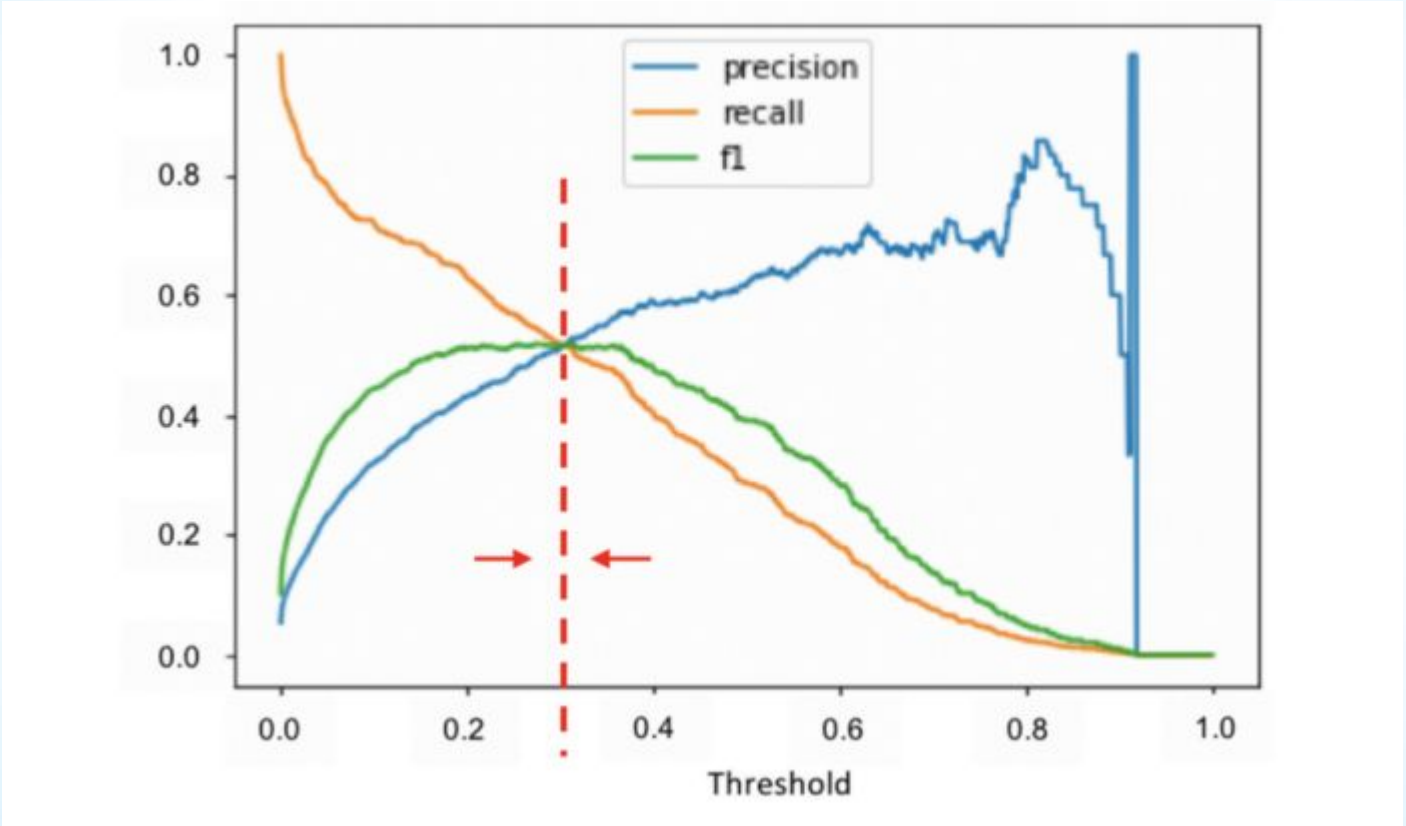


Optimization configuration

```
model_packs = {
    "BernoulliNB" : {"class" : BernoulliNB,
                     "args" : {'fit_prior' : [True, False],
                               'alpha' : [0.1,0.5,1.0]}},
    "LogisticRegression" : {"class" : LogisticRegression,
                            "args" : {'penalty' :['l1', 'l2', 'elasticnet', 'none'] }},
    "DecisionTreeClassifier" : {"class" : DecisionTreeClassifier,
                                "args" : {"criterion" :  ["entropy", "gini"],
                                          "min_samples_leaf" : [5,15,20]}},
    "RandomForestClassifier" : {"class" : RandomForestClassifier,
                                "args" : {"criterion" : ["entropy", "gini"],
                                          "n_estimators" :[50,100],
                                          "max_depth" : [None, 10]}}
}

feature_extractors_pack = {
    "V1" : {"class" : FeatureExtractorV1, "args" : {"session_size" : 3}},
    "V2" : {"class" : FeatureExtractorV2, "args" : {"session_size" : 3}}
}
```

```
clf_cv = GridSearchCV(model_params['class'](),model_params['args'], cv = cv_for_grid_cv)

clf_cv.fit(X_for_grid_cv,y_for_grid_cv)
preds_test = clf_cv.predict(test_features)
probs_test = clf_cv.predict_proba(test_features)[:,1]
```

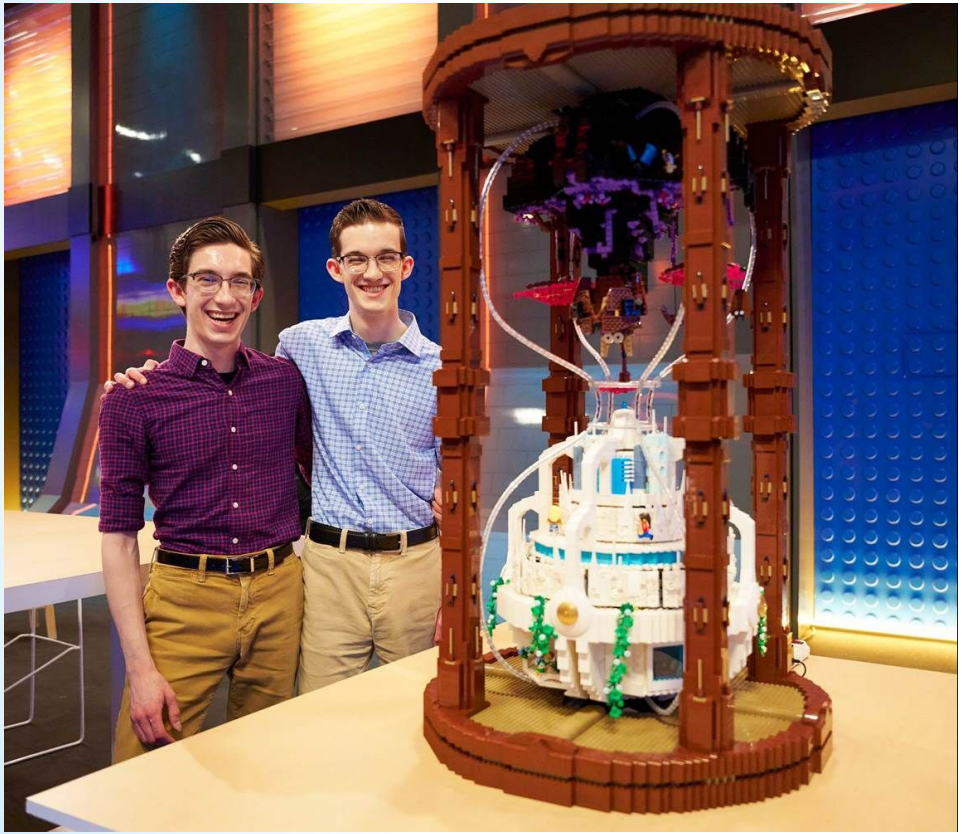
Threshold optimization



Batch Evaluation

batch	model_name	f1_score	f1_score_optimized
2020-10-31 00:00:00	BernoulliNB	0.645727	0.645445
2020-10-31 00:00:00	DecisionTreeClassifier	0.619484	0.688262
2020-10-31 00:00:00	LogisticRegression	0.695299	0.000000
2020-10-31 00:00:00	RandomForestClassifier	0.642916	0.704264
2020-11-14 00:00:00	BernoulliNB	0.635775	0.639161
2020-11-14 00:00:00	DecisionTreeClassifier	0.599850	0.679037
2020-11-14 00:00:00	LogisticRegression	0.689349	0.697489
2020-11-14 00:00:00	RandomForestClassifier	0.629699	0.691538

Show your model to the world!



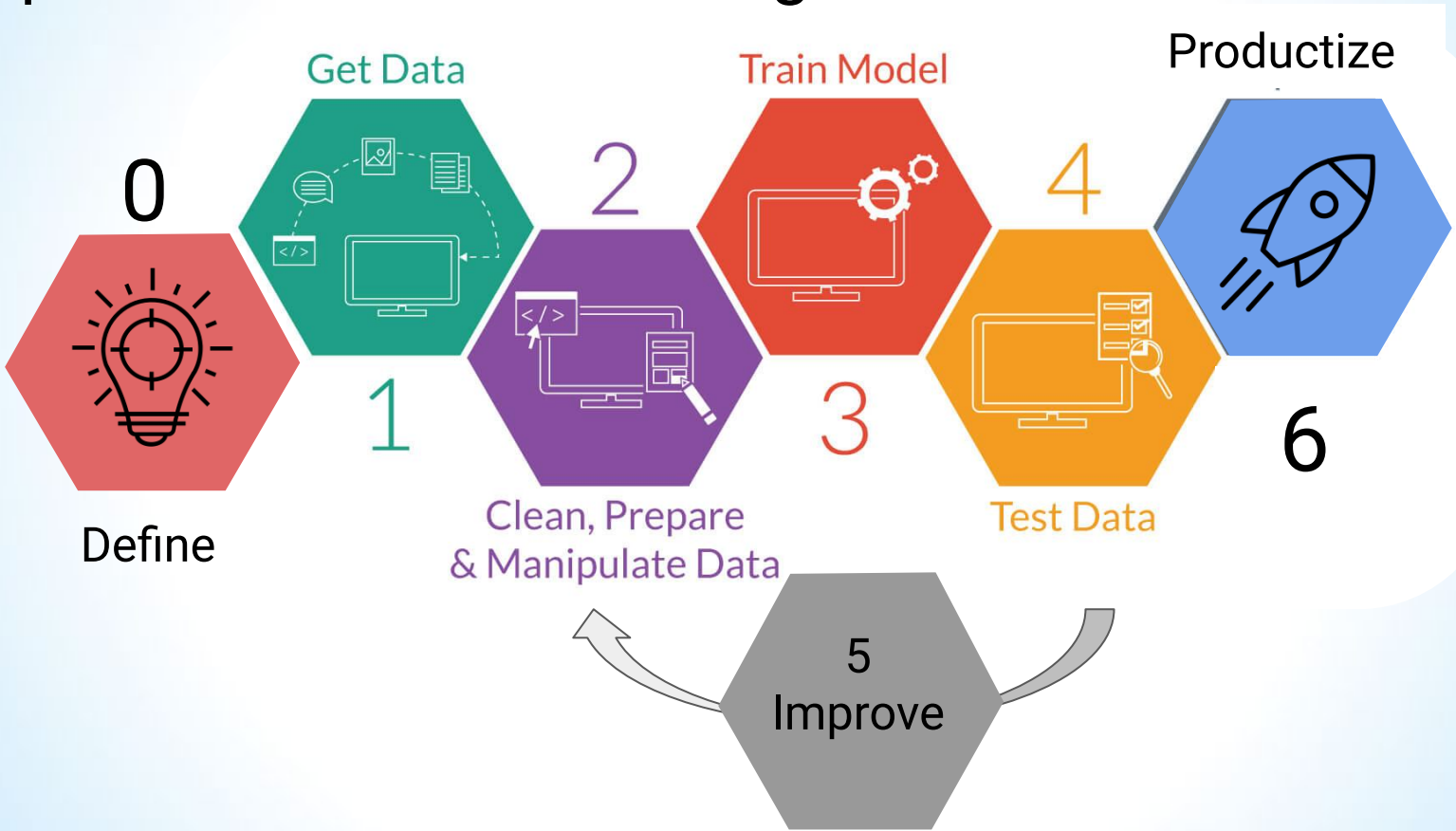
Streamlit

https://streamlit.io/gallery

Back to reality - Error Analysis

Back to reality - Improve your signal!

Steps to Predictive Modeling



Code Tricks Summary

Save prepared data and load if not os.path.isfile.

Use kwargs - class(**kwargs_dict)

Use classes to separate code ares

Use list of dictionaries to build data frame.

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

How to Dev?

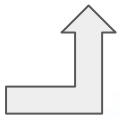
Jupyter

- Analysis and demoing
- In memory context - fast but cause mistakes with variables

Pycharm

- Organized code
- Debuggable
- Auto-complete
- Argument with bash

```
In [1]: %load_ext autoreload
        %autoreload 2

In [2]: from my_code.my_cool_utils import my_sum

        my_sum(x=5,y=5) # Should return the sum of x+y

Out[2]: 10

In [3]: # Change my_sum to print some wrapping text and then return the sum
        my_sum(x=5,y=5)

        The sum of 5 and 5 is:

Out[3]: 10
```



Combine

- Use Classes developed in pycharm at jupyter

Future steps

More Error analysis - going back the the raw data and aggregations
Extract session features: duration, average time between views
Reduce Features with Chi2 selection or PCA
Undersample / oversample the data

Y-DATA
SCHOOL OF DATA SCIENCE

Yandex

Q&A

Pandas GUI - https://towardsdatascience.com/pandasgui-analyzing-pandas-dataframes-with-a-graphical-user-interface-36f5c1357b1d