Classification of Sber Avtopodpiska website visitors

1.

Task

About project

- an alternative to carsharing and buying on credit
- long-term car rental service for individuals with a monthly payment
- insurance and repair
- additional services

Main tasks

- Develop binary classification model of users' actions
  - ROC AUC ≥ 0.65
  - Use only `utm`, `device` and `geo` features
- Deploy as API
  - Scalable
  - Response time < 3s

2.

EDA

Data format

sessions.csv

	geo_city	...	visit_date	session_id
0	Moscow	...	2021-12-09	1
1	None	...	2021-12-15	2
...	...	...	...	...
100000	Moscow	...	2021-12-28	100000
100001	Omsk	...	2021-12-16	100001

hits.csv

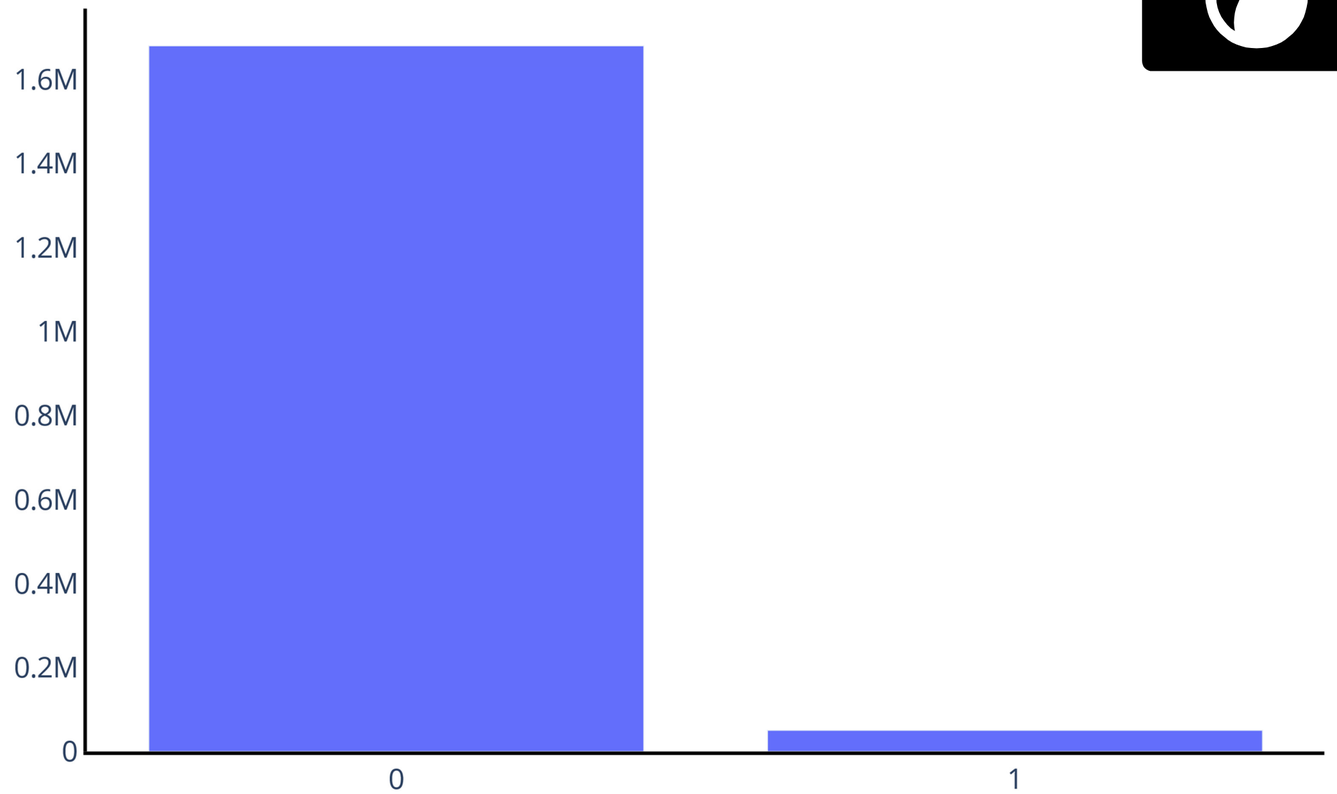
	session_id	...	event_label
0	1	...	quiz_show
1	1	...	view_card
2	1	...	sub_submit_success
...	...	...	...
100	1	...	view_card
101	1	...	sub_car_claim_click
102	1	...	sub_landing

Data format

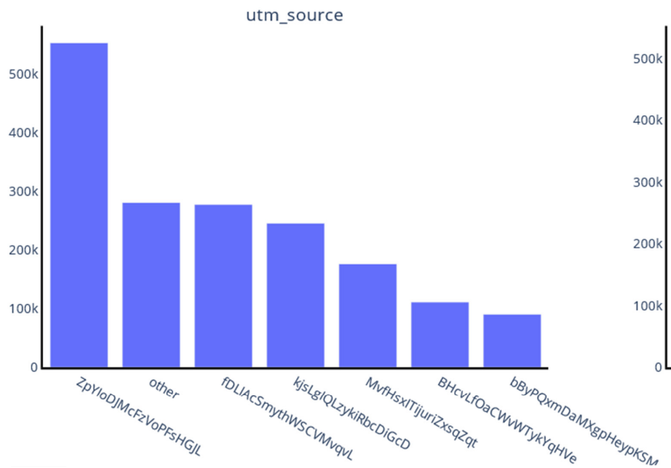
column	non-null count	percentage %	dtype	nunique
utm_source	1732190	100	object	281
utm_medium	1732266	100	object	55
utm_campaign	1536979	88.73	object	407
utm_adcontent	1428129	82.44	object	281
utm_keyword	711514	41.07	object	1193
device_category	1732266	100	object	3
device_os	718302	41.47	object	14
device_brand	1385070	79.96	object	201
device_model	15062	0.87	object	105
device_screen_resolution	1732266	100	object	4947
device_browser	1732266	100	object	55
geo_country	1732266	100	object	159
geo_city	1732266	100	object	2389

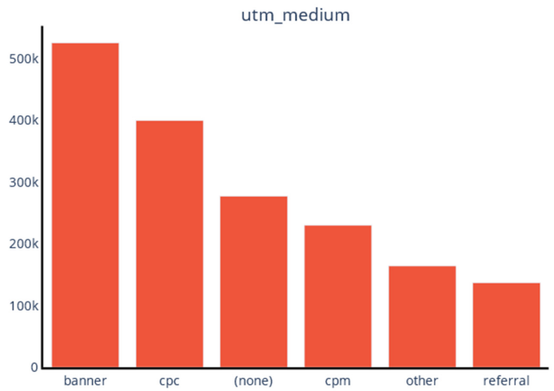
Target

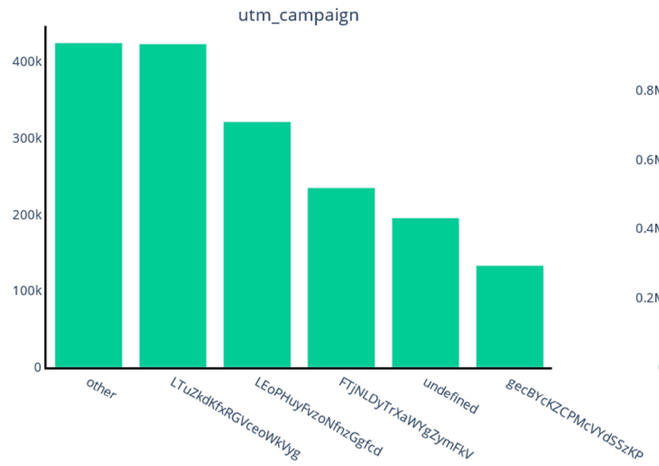
- 0 – 97.1%
- 1 – 2.9%

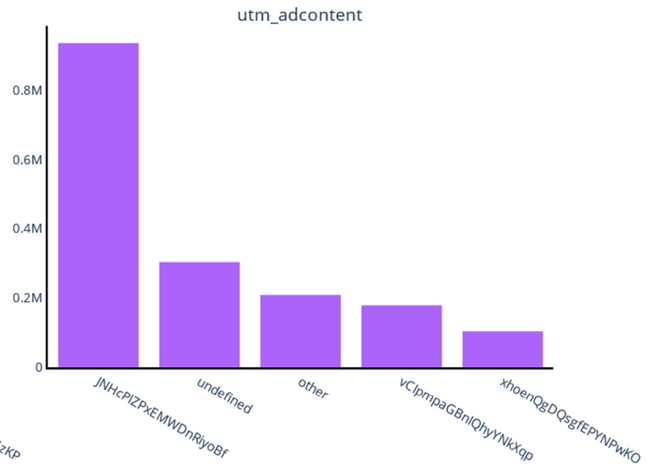


Distributions

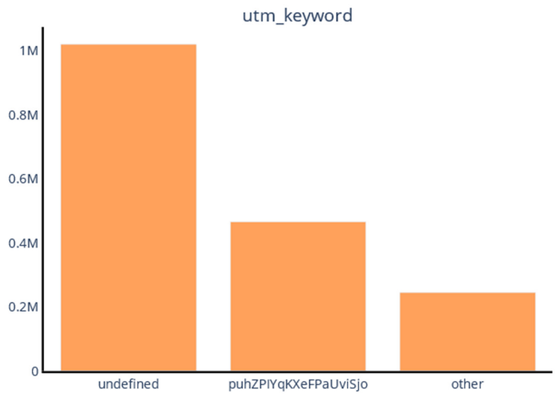


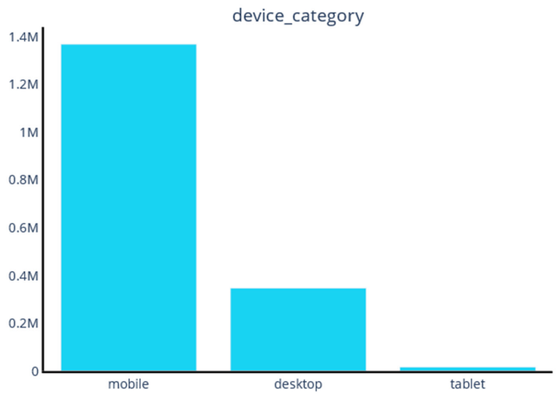


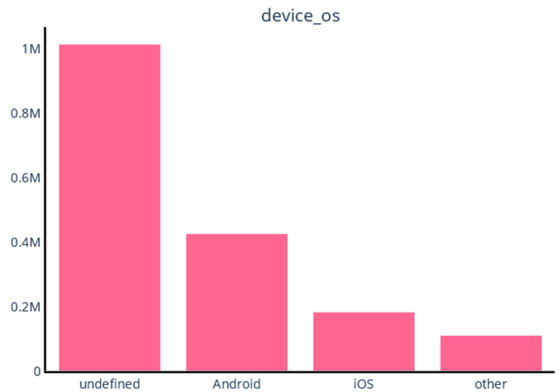


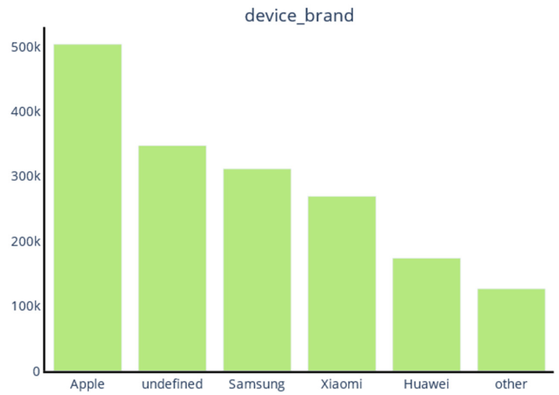


Distributions

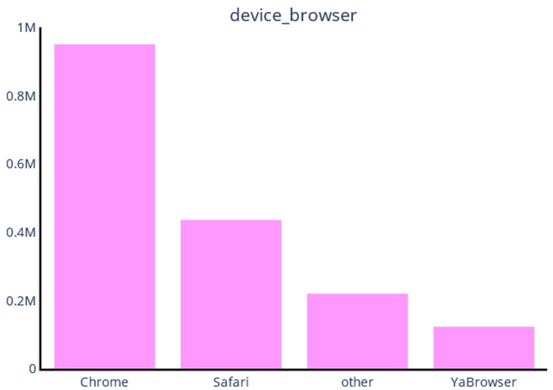


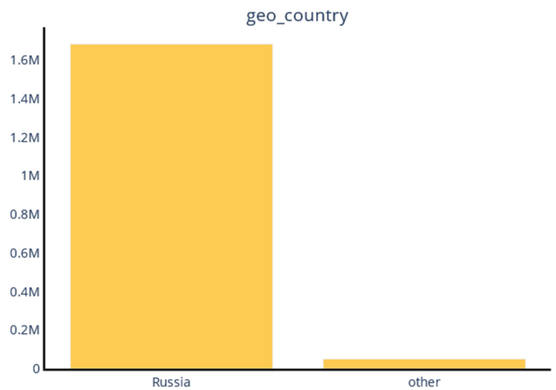


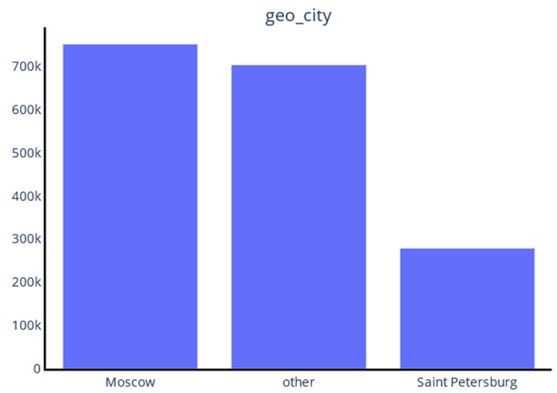




Distributions







3.

Feature Engineering

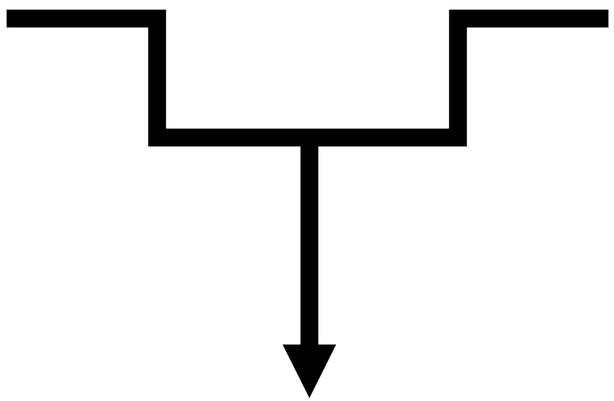
Basic transformation

- combine `os` & `keyword`
- extract `device_screen_height` from `device_screen_size`

RU Cities

Geonames ru cities data [1]

	city	population	timezone
0	Moscow	10381222	Europe/Moscow
1	None	5351935	Europe/Moscow
...	...	...	...
100000	Trakow	0	Europe/Moscow
100001	Akatnov	0	Europe/Moscow



cities.csv

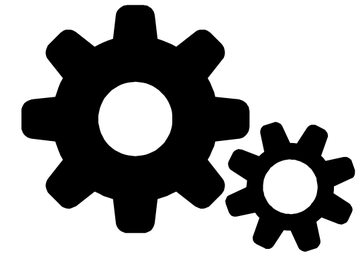
RU cities by population [2]

	city	federal_subject	federal_district
0	Moscow	Moscow (federal city)	Central
1	Saint Petersburg	Saint Petersburg (federal city)	Northwest
...	...	...	...
308	Snezhinsk	Chelyabinsk Oblast	Ural
309	Zhigulyovsk	Samara Oblast	Volga

4.

Preprocessing

Preprocessing I



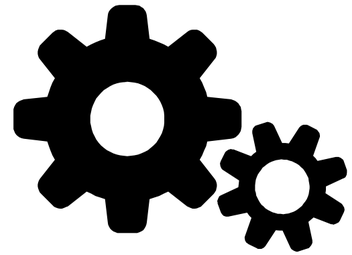
`dev/train/query.sql`

CONVERT
`event_action` to binary
`target_action`

MAX
`target_action` groped by
`session_id`

JOIN
with sessions on
`session_id`

Preprocessing II



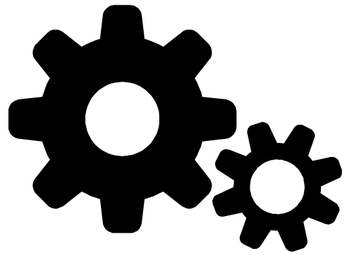
`pandas_preprocess()` at `dev/train/train.py`

Assign*

Merge
with `cities.csv` on city name

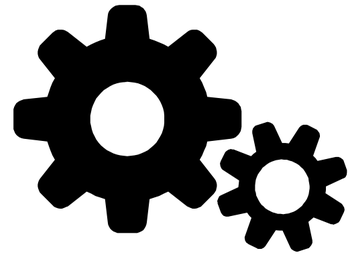
Drop
`device_model`
`device_screen_resolution`
`device_os`
`utm_keyword`

Assign



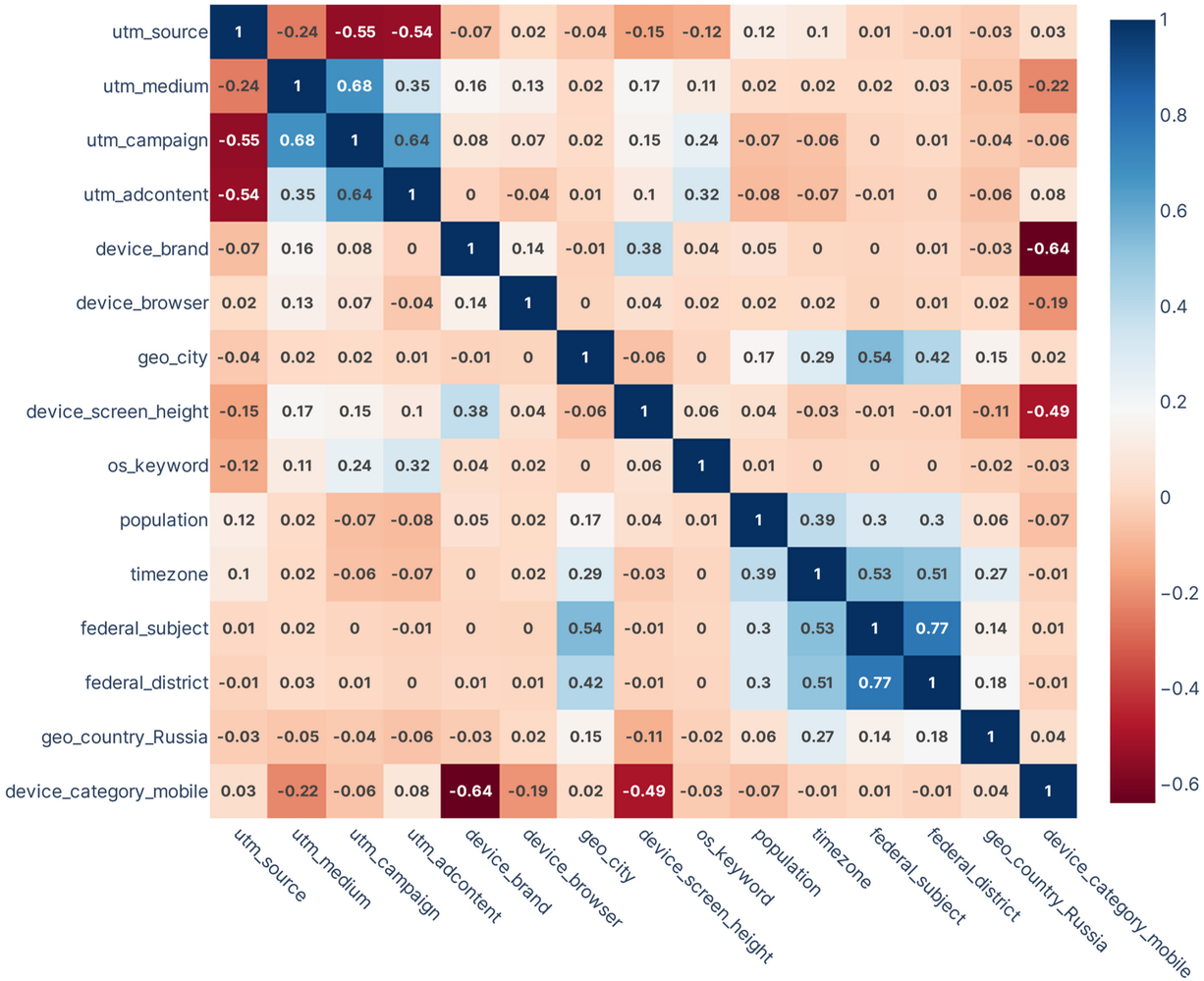
column	action
utm_campaing	fillna
utm_adcontent	fillna
utm_keyword	fillna
device_category	merge tablet & mobile
device_os	fillna convert to Windows/Macintosh/Linux
device_brand	fillna
device_screen_resolution	new column device_screen_height
device_browser	merge Safari & Safari (in-app)
os_keyword	combine os & keyword columns

Preprocessing III



`preprocess` variable at `dev/train/train.py`

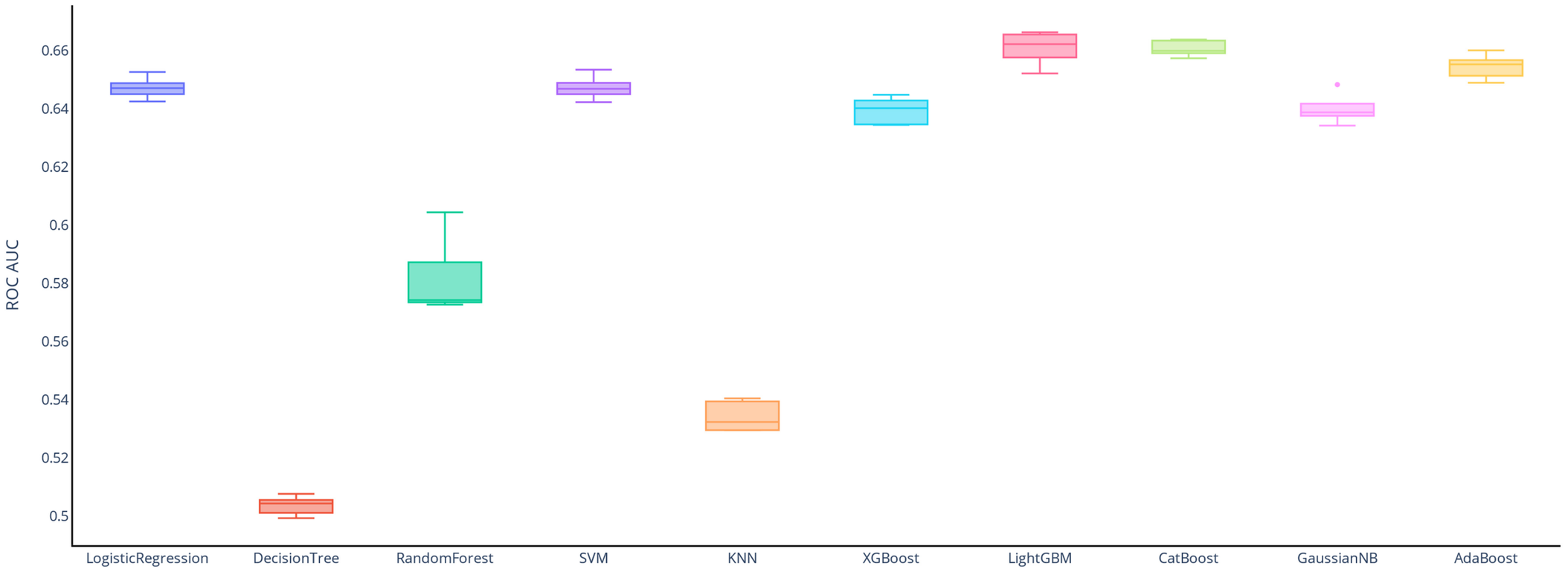
Correlation



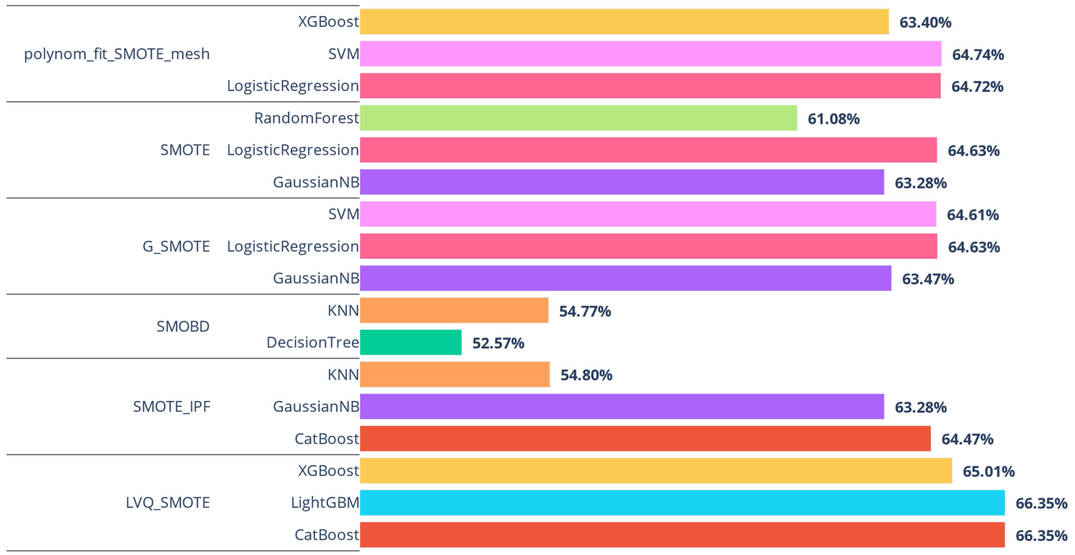
5.

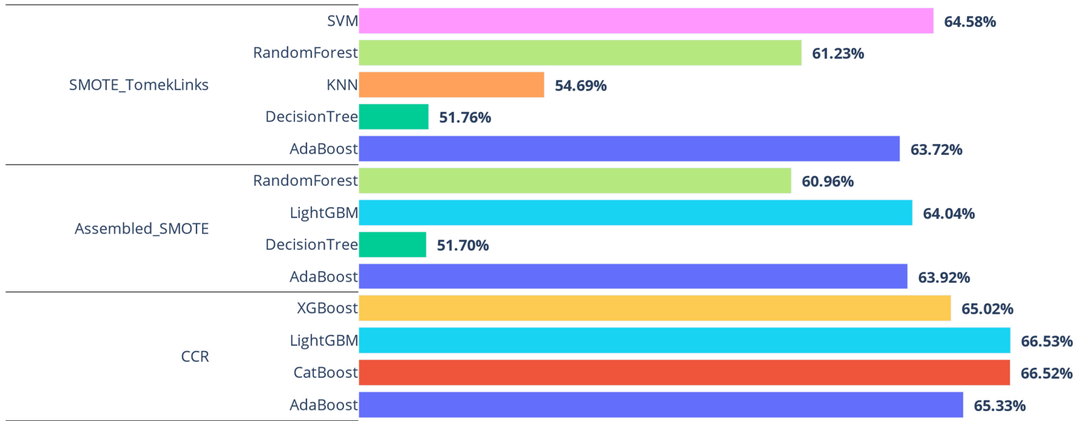
Model Selection

Baseline

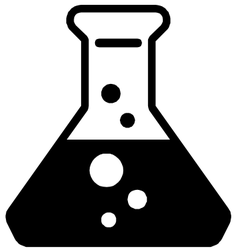

ROC AUC
0.66
0.64
0.62
0.6
0.58
0.56
0.54
0.52
0.5
LogisticRegression
DecisionTree
RandomForest
SVM
KNN
XGBoost
LightGBM
CatBoost
GaussianNB
AdaBoost

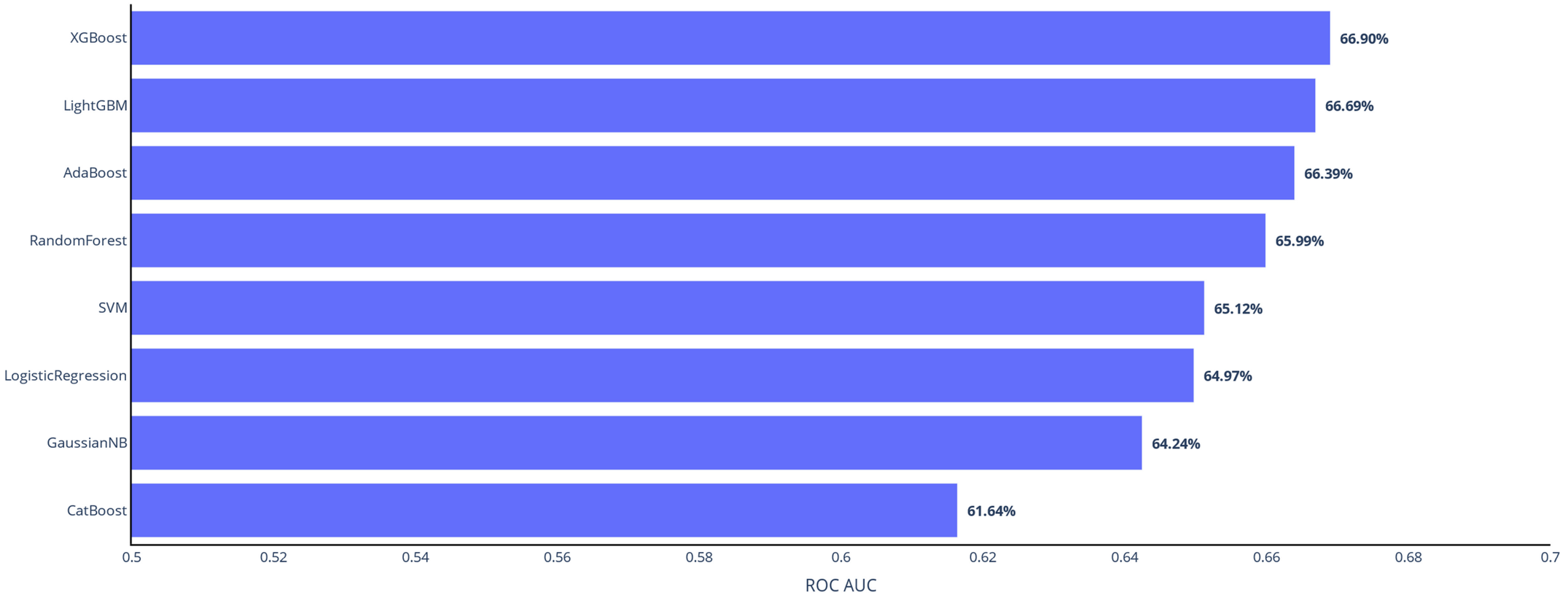
Resampling [4]



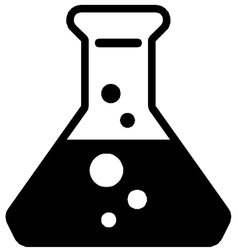


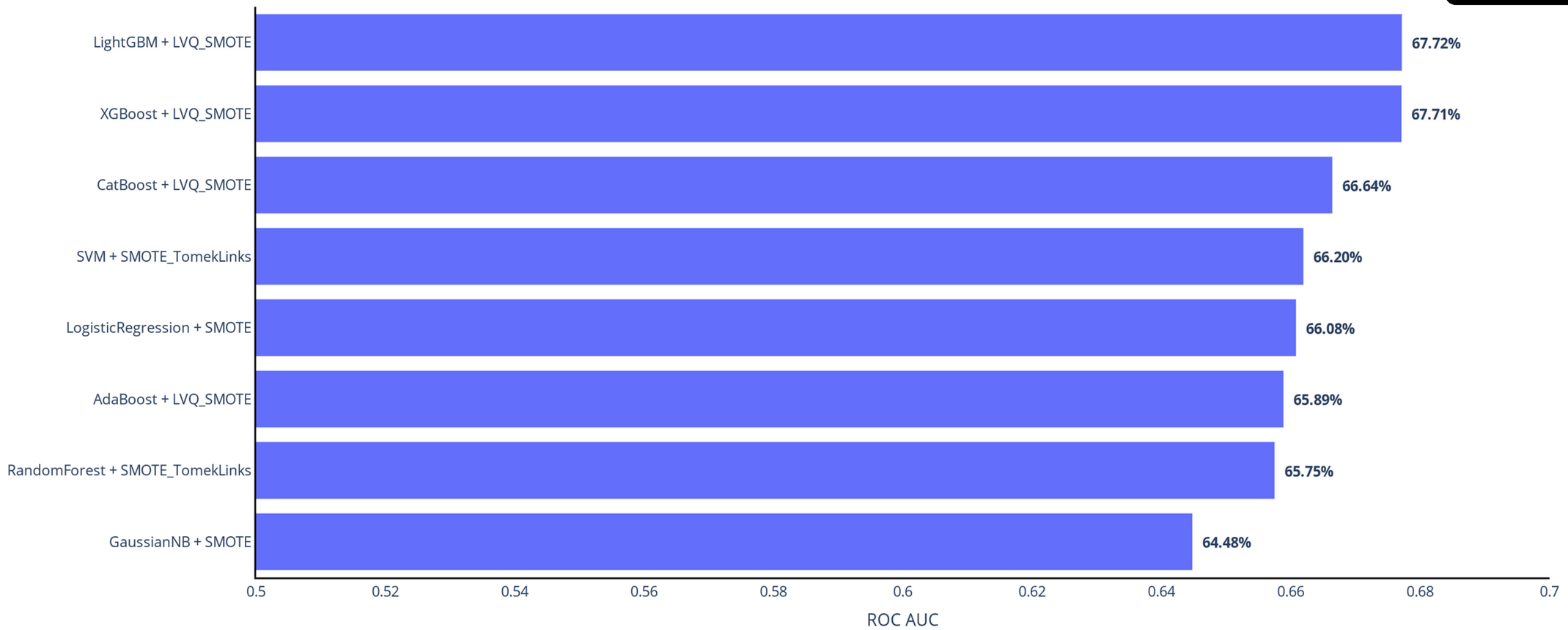
Tuned models



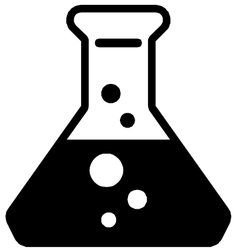


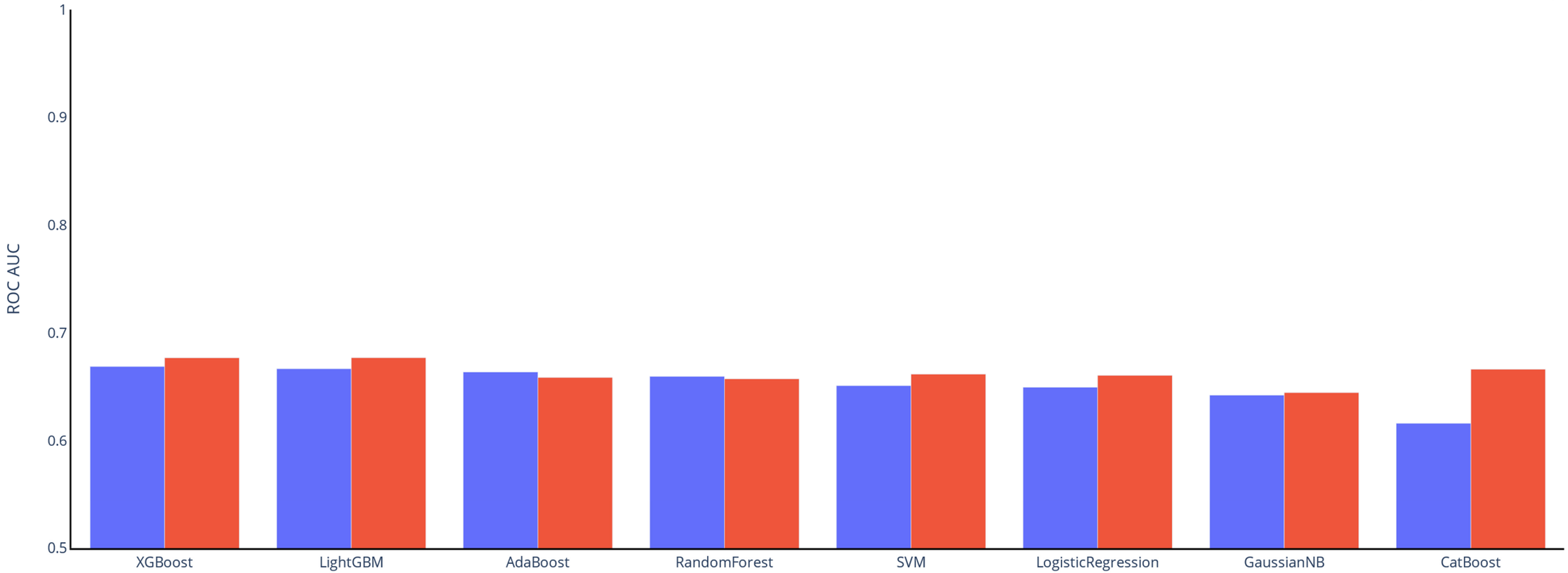
Tuned with Resampling





Weighted vs Resampled

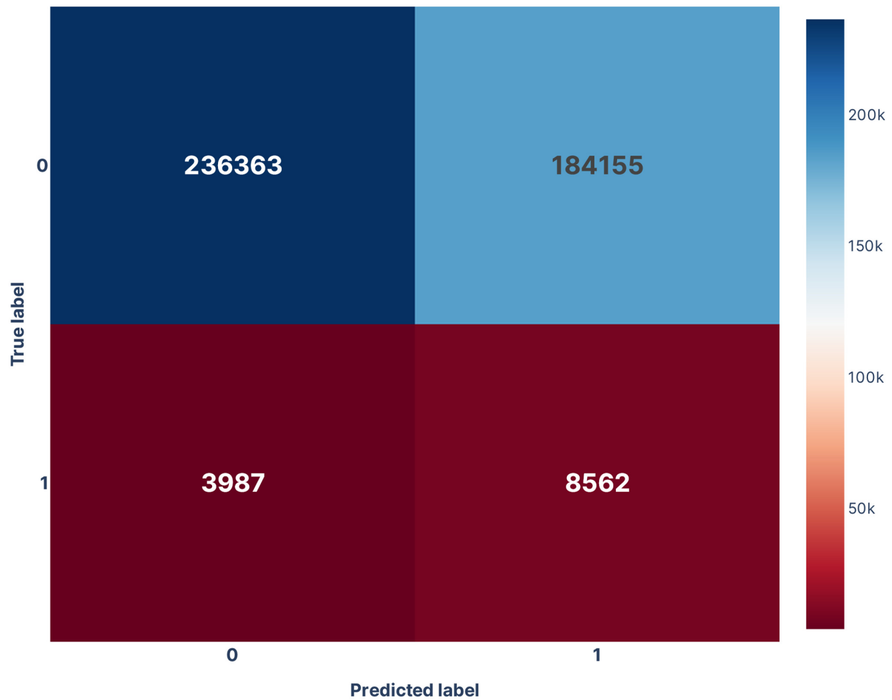




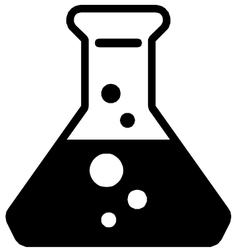
Model

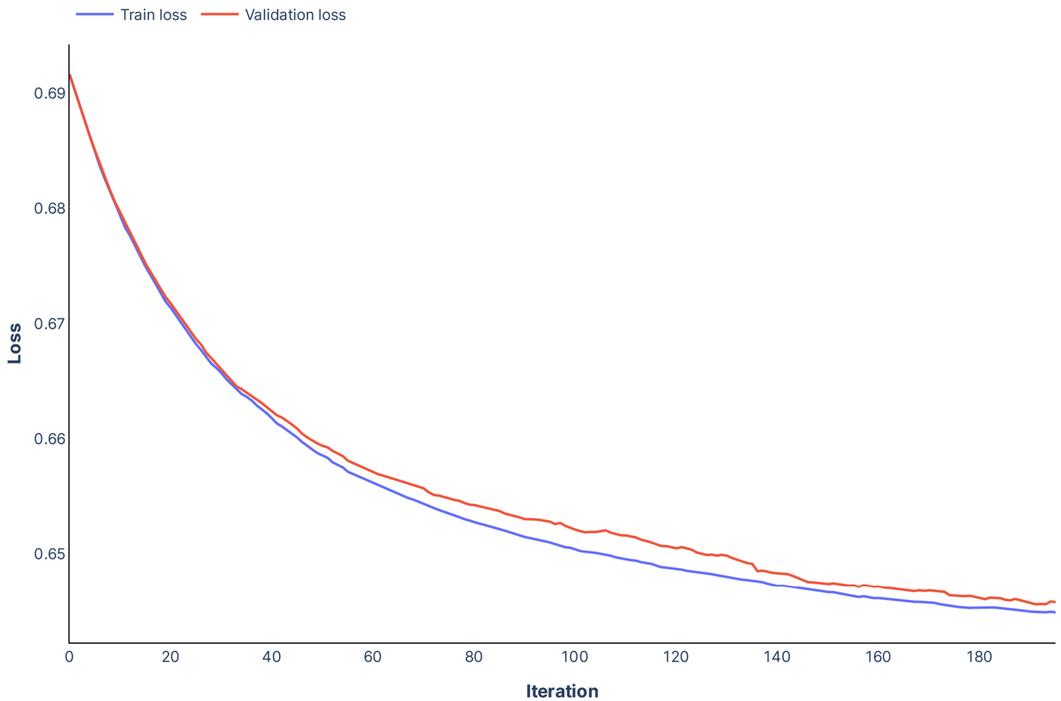
- XGBoost
- Tuned
- Weighted

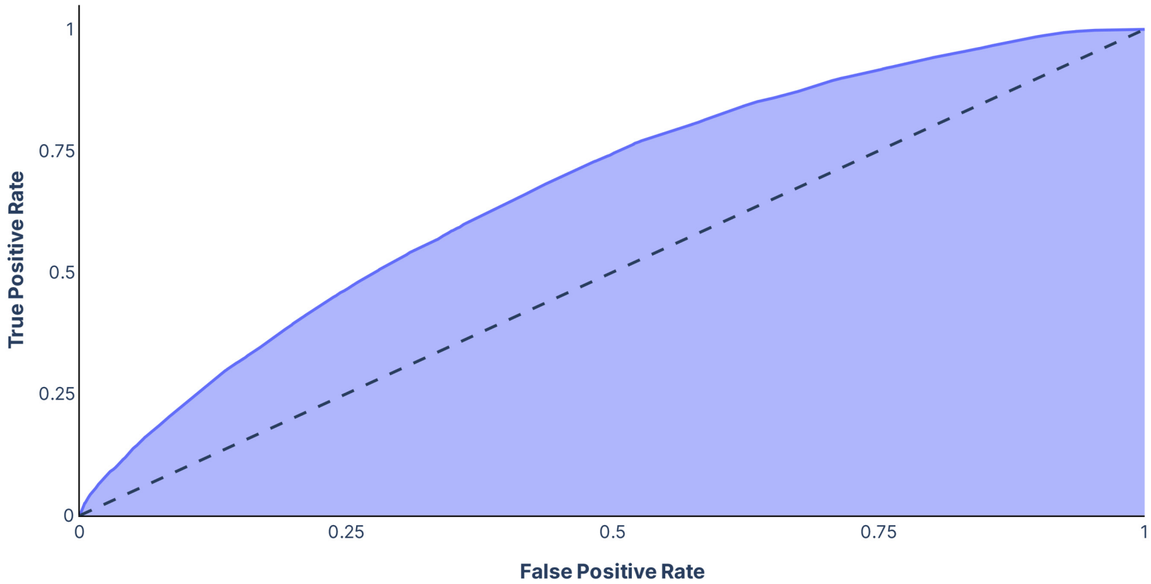
TNR 56%	FPR 44%	FDR 96%
FNR 32%	TPR 68%	NPV 98%
ROC AUC 67%	F1 67%	Precision 4%



Model

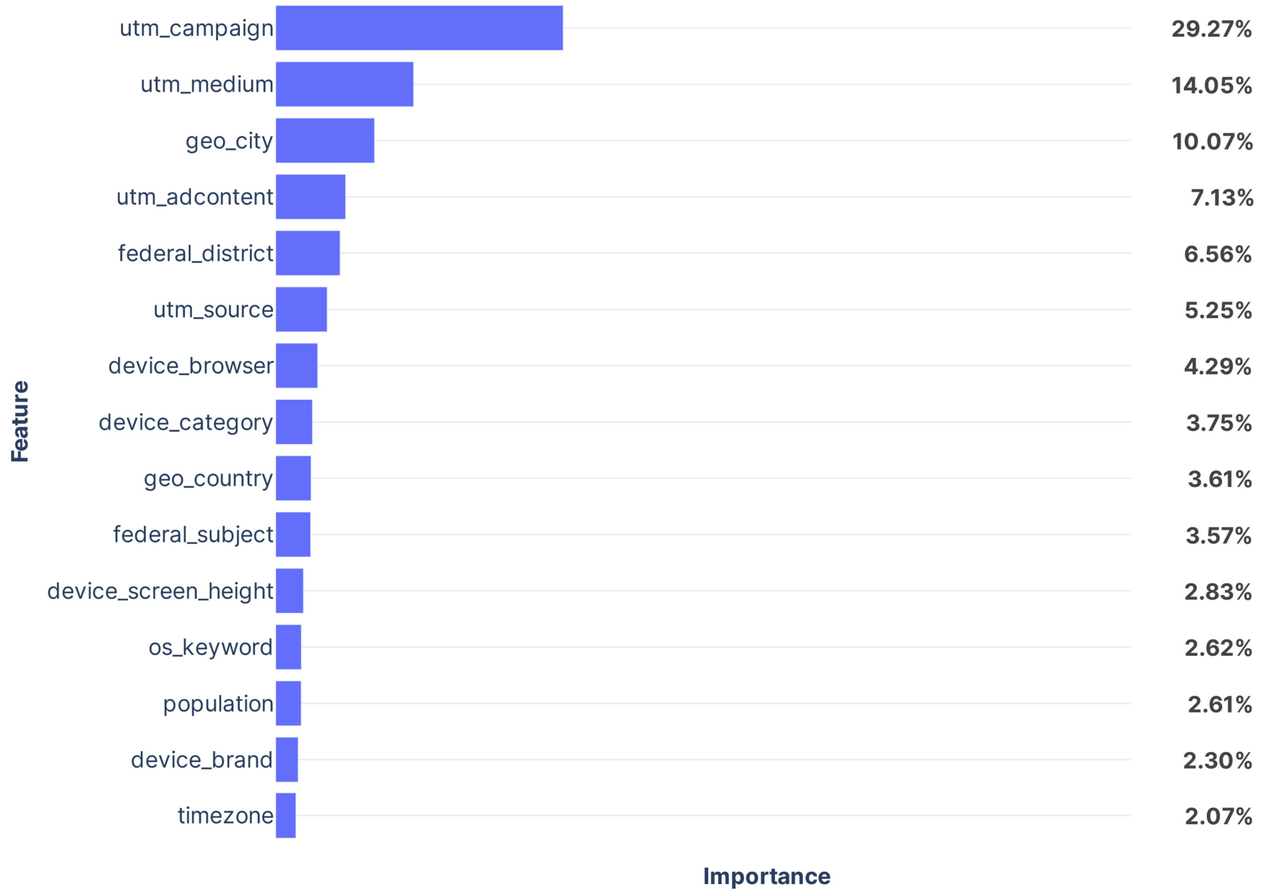



Train loss
Validation loss
Loss
0.69
0.68
0.67
0.66
0.65
0
20
40
60
80
100
120
140
160
180
Iteration


True Positive Rate
1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
False Positive Rate

Feature importance

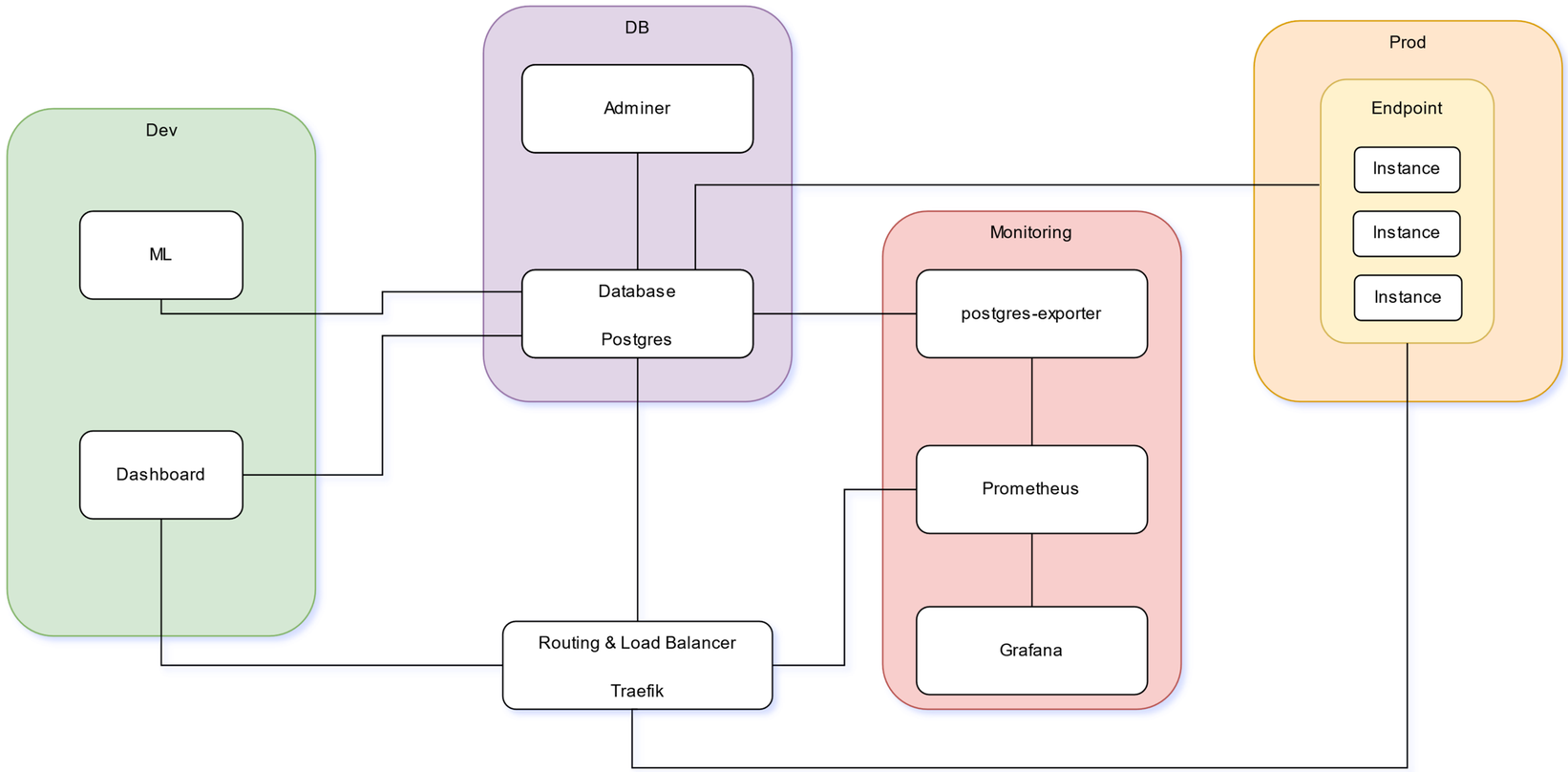
`utm` features cumulatively made the most impact



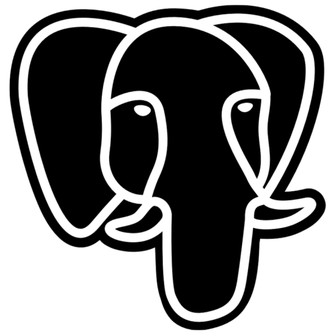
6.

Deployment

Services architecture



Database architecture



sber_avtopodpiska

raw_sessions		
PK	**id**	**SERIAL**
	session_id	VARCHAR
	client_id	VARCHAR
	visit_date	DATE
	...	...
	...	...
	...	...
	device_browser	VARCHAR
	geo_country	VARCHAR
	geo_city	VARCHAR

raw_hits		
PK	**id**	**SERIAL**
	session_id	VARCHAR
	event_action	TIMESTAMP

scores

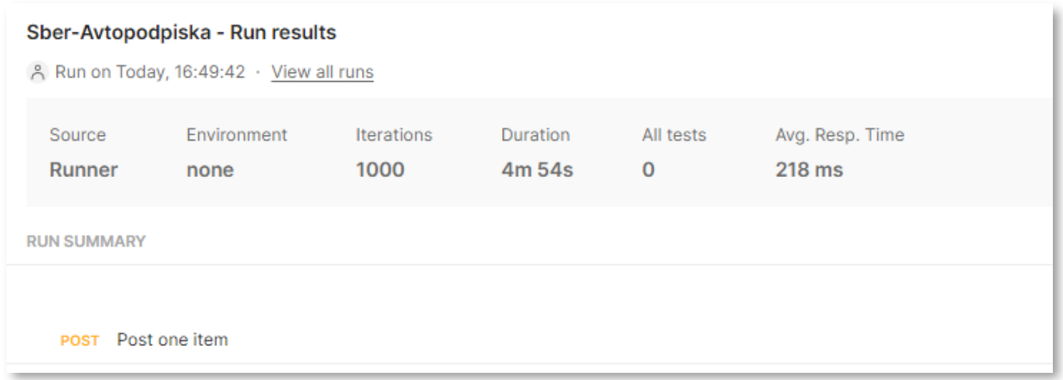
metrics		
PFK	**model_id**	**SERIAL**
	tn	FLOAT
	fp	FLOAT
	fn	FLOAT
	tp	FLOAT
	roc_auc	FLOAT
	f1_beta	FLOAT
	feature_importance	BYTEA
	corr	BYTEA
	evals_result	BYTEA

models		
PK	**model_id**	**SERIAL**
	model	BYTEA
	dt	TIMESTAMP

train_pred		
PK	**pred_id**	**SERIAL**
PFK	**model_id**	**SERIAL**
	fact	INTEGER
	pred_proba	FLOAT

API


```
{
    "items": [
        {
            "utm_source": false,
            "utm_medium": false,
            "utm_campaign": "isYoUwVPnRHJ",
            "utm_adcontent": "JNHcPlZPxEM",
            "utm_keyword": null,
            "device_category": "mobile",
            "device_os": null,
            "device_brand": "Nokia",
            "device_model": null,
            "device_screen_resolution": "412x823",
            "device_browser": "Chrome",
            "geo_country": "Russia",
            "geo_city": "Stavropol"
        }
    ]
}
```

Results

- Model with ROC AUC ≈ 0.67
- Deployed with response time ≈ 200 ms

Thanks

[0] Source code – https://github.com/AlimU11/Sber-Avtopodpiska

[1] “GeoNames Russian Cities Data.” download.geonames.org, https://download.geonames.org/export/dump/

[2] Wikipedia Contributors. “Federal Subjects of Russia.” Wikipedia, Wikimedia Foundation, 6 Oct. 2019, https://en.wikipedia.org/wiki/Federal_subjects_of_Russia

[3] Vorotyntsev, Denis. Benchmarking Categorical Encoders. 14 July 2019, https://towardsdatascience.com/benchmarking-categorical-encoders-9c322bd77ee8

[4] Kovács, György. “An Empirical Comparison and Evaluation of Minority Oversampling Techniques on a Large Number of Imbalanced Datasets.” Applied Soft Computing, vol. 83, Oct. 2019, p. 105662, https://www.sciencedirect.com/science/article/abs/pii/S1568494619304429, 10.1016/j.asoc.2019.105662.