Introduction to Data Science

Zindi User Behaviour Birthday Challenge

Team 18

Skoltech

General problem and dataset description

- Problem: user churn
- Goal: prediction of the users' behavior
- Purpose: Increasing the quality of the Zindi website performance as a Data Science platform

Loading and data

	UserID	month	year	CompPart	Comment	Sub	Disc
100500	ID_000VV0KM	12	2	1	0	1	0
100501	ID_000VV0KM	1	3	0	0	0	0
100502	ID_000VV0KM	2	3	0	0	0	0
100503	ID_000VV0KM	3	3	0	0	0	0
100504	ID_000VV0KM	4	3	0	0	0	0

Train

	Target
100500	1
100501	0
100502	0
100503	0
100504	0

Target

Preparing the dataset

One-hot encoding:

```
total_data = pd.concat([total_data, pd.get_dummies(total_data['Points'], prefix='points_')], axis=1).drop('Points', axis=1)
total_data = pd.concat([total_data, pd.get_dummies(total_data['PublicRank'], prefix='public_')], axis=1).drop('PublicRank', axis=1)
```

✓ 0.0s

points__group 3	points__group 4	...	public__rank 10	public__rank 11	public__rank 2	public__rank 3	public__rank 4	public__rank 5	public__rank 6	public__rank 7	public__rank 8	public__rank 9
True	False	...	False	False	False	False	False	False	True	False	False	False
True	False	...	False	False	False	False	False	False	False	False	False	False
True	False	...	False	False	False	False	False	False	False	False	False	False
True	False	...	False	False	False	False	False	False	False	False	False	False
True	False	...	False	False	False	False	False	False	False	False	False	False

Preparing the dataset

Feature generation:

	UserID	PublicRank	year	month
0	ID_000VV0KM	rank 6	2	12
1	ID_00HKNVC0	rank 10	3	3
2	ID_00QSUS04	rank 5	2	2
3	ID_00W1WG4W	rank 8	2	7
4	ID_00WD4BRD	rank 11	3	6

‘Public rank’

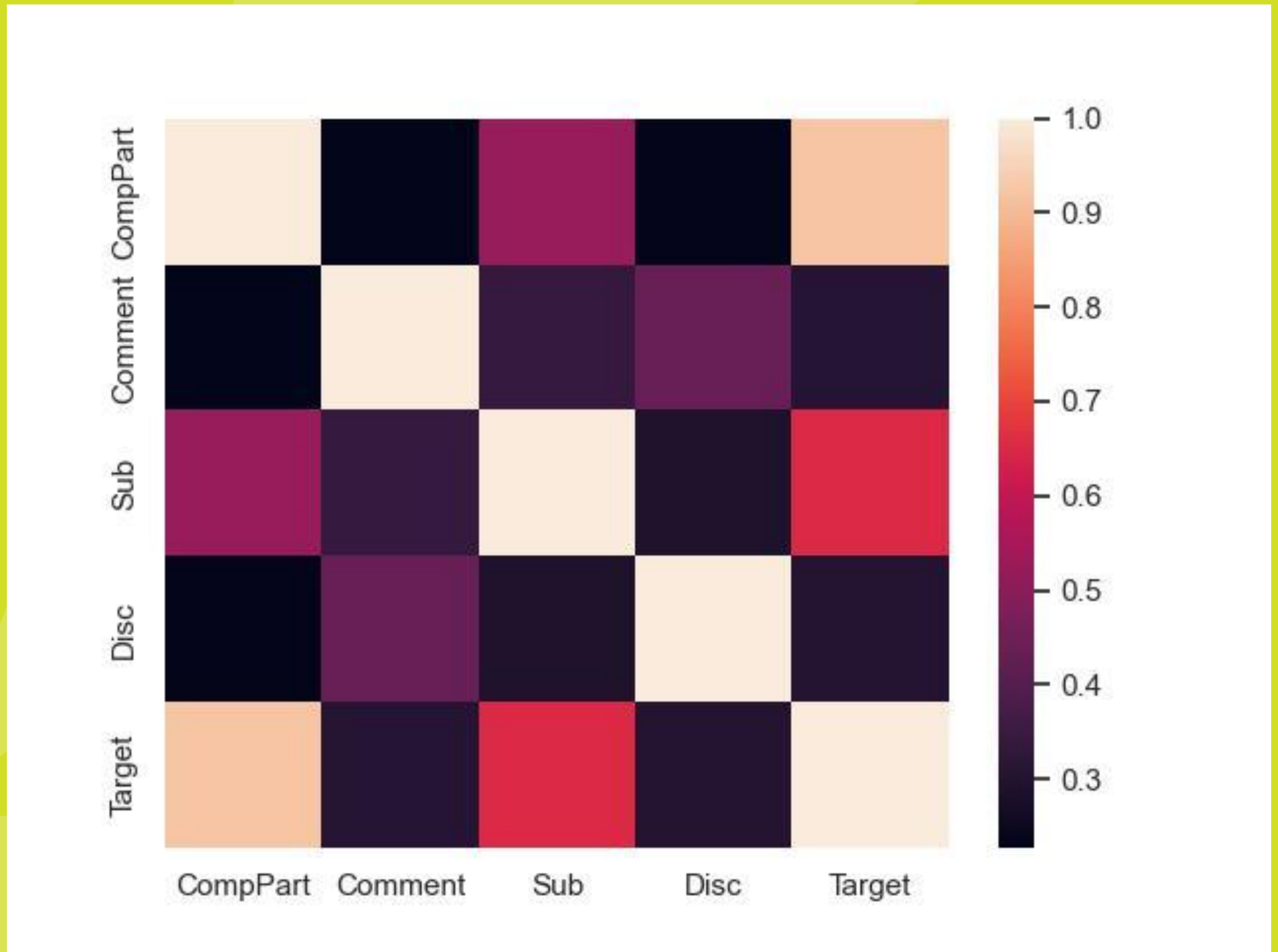
	UserID	year	month	account_age
0	ID_000VV0KM	2	12	0
1	ID_000VV0KM	3	1	1
2	ID_000VV0KM	3	2	2
3	ID_000VV0KM	3	3	3
4	ID_000VV0KM	3	4	4
...	...	...	...	...
259827	ID_ZZXDLYXB	3	8	4
259828	ID_ZZXDLYXB	3	9	5
259829	ID_ZZXDLYXB	3	10	6
259830	ID_ZZXDLYXB	3	11	7
259831	ID_ZZXDLYXB	3	12	8

‘Account age’

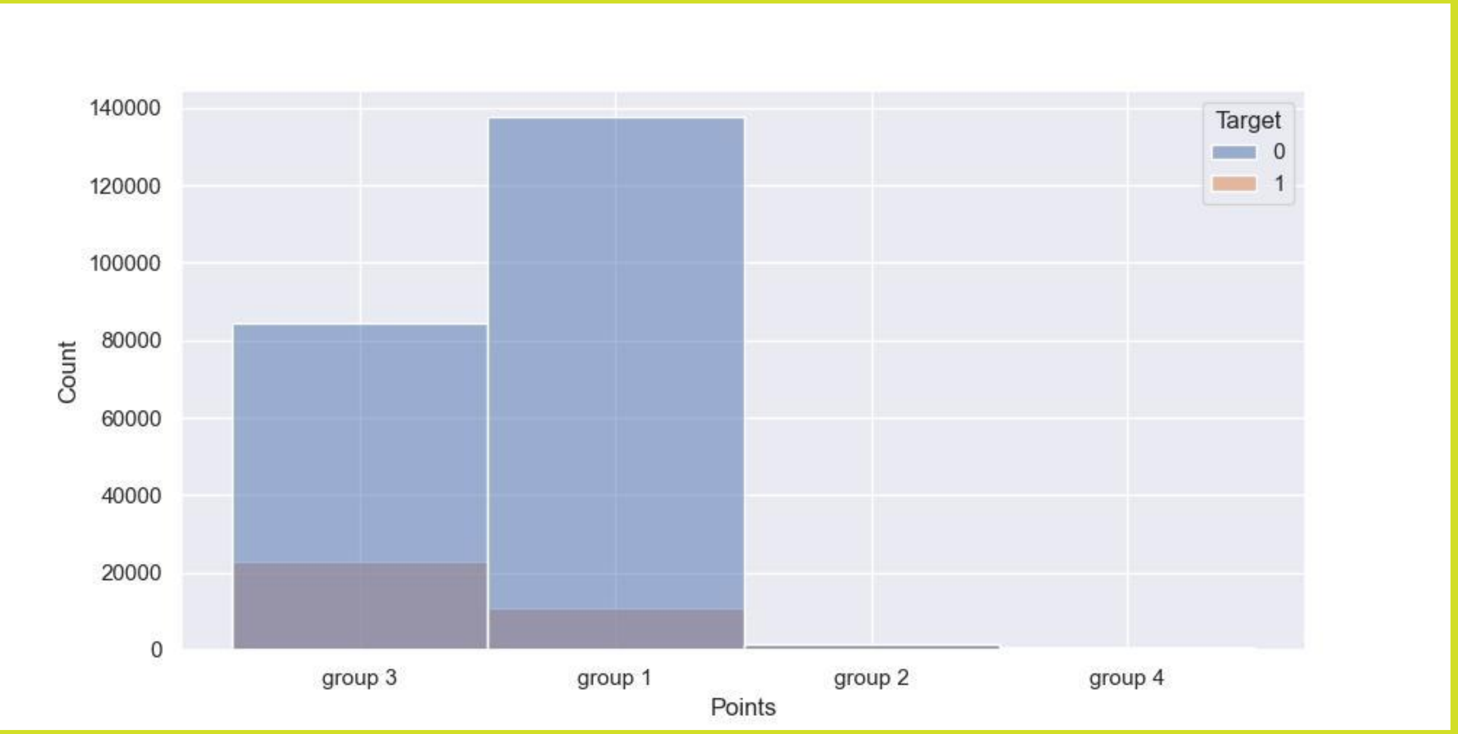
	UserID	prev_m_act
0	ID_000VV0KM	0.0
1	ID_000VV0KM	2.0
2	ID_000VV0KM	0.0
3	ID_000VV0KM	0.0
4	ID_000VV0KM	0.0

‘Previous month activity’

Exploratory analysis

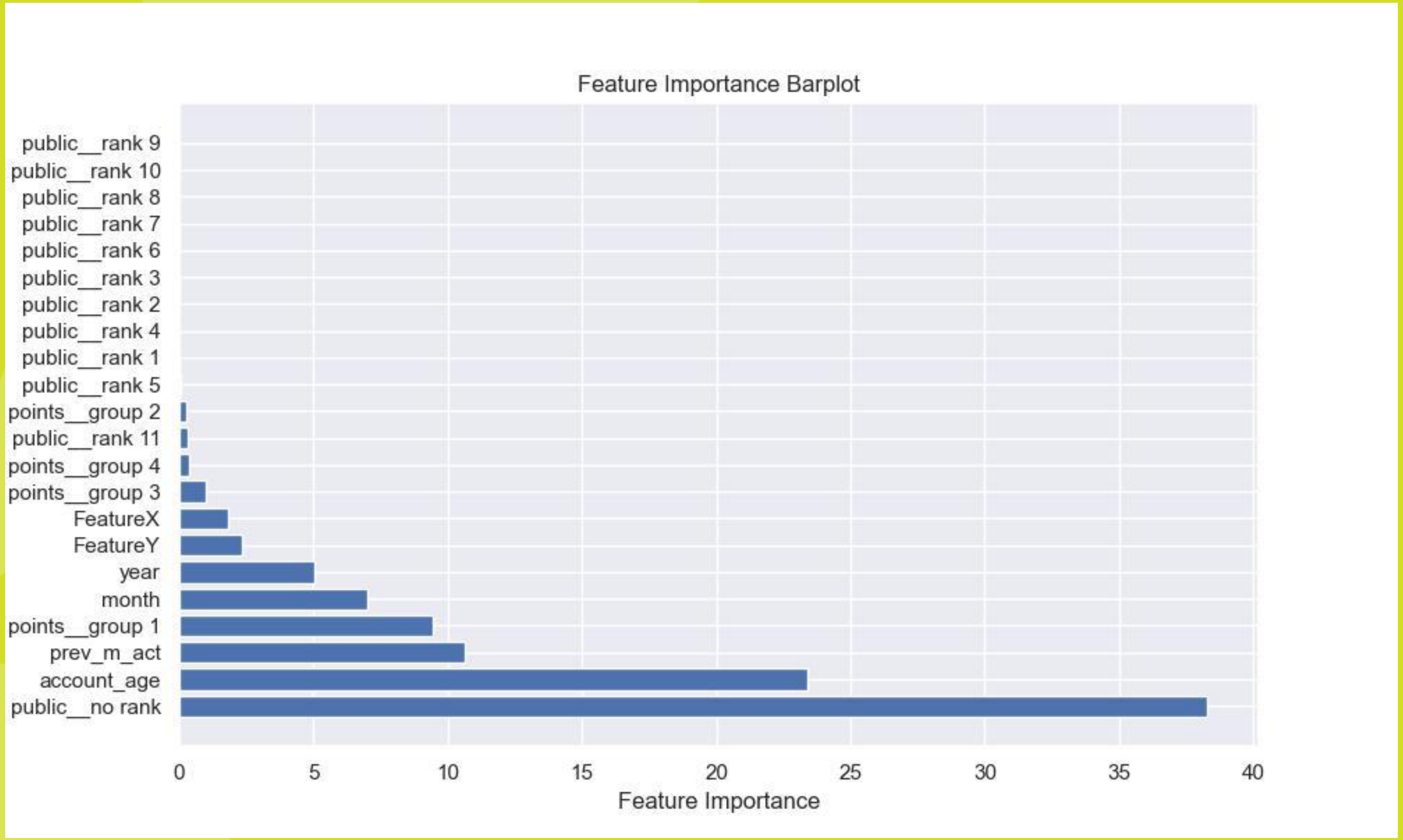


Correlation of selected features and Target



Distribution of observations among groups (by points)

Feature importance



Feature importance based on CatBoost

Choosing a ML model

- LogReg
  - roc-auc: `0.819 +- 0.001`
  - f1-score: `0.701 +- 0.001`
- RandomForest
  - roc-auc: `0.858 +- 0.002`
  - f1-score: `0.791 +- 0.003`
- CatBoost
  - roc-auc: `0.897 +- 0.002`
  - f1-score: `0.820 +- 0.002`

Best parameters

- CatBoost
  - depth: 8
  - l2_leaf_reg: 5
  - learning_rate: 0.03

ROC-AUC: `0.899 +- 0.001`
F1-score: `0.821 +- 0.001`

```
roc_auc :  0.898279897429725
f1 score :  0.8222079140013717
-------
roc_auc :  0.9006989814422663
f1 score :  0.8197642958038469
-------
roc_auc :  0.9002162754797005
f1 score :  0.8220981085824118
-------
Mean roc auc :  0.8997317181172306  +-  0.0010453352709620895
Mean f1 scores :  0.8213567727958768  +-  0.0011269432210236432
```

Metrics' values for every model

Analyzing the obtained results

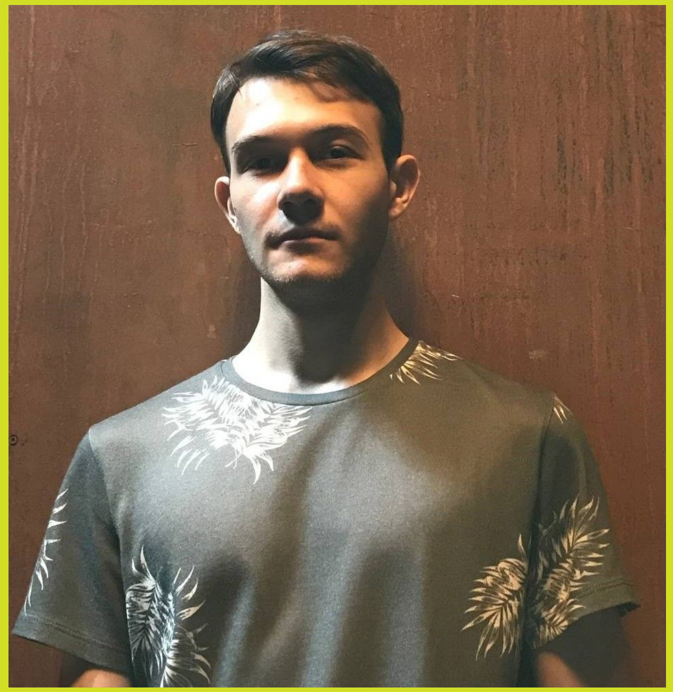
- An interesting finding: correctness of the targets components
- Used oversampling and undersampling

Conclusion: with the power of feature engineering CatBoost was the most powerful algorithm for solving our task.

Our team

Sabitov Elfat
Model selection

Fokin Alex
EDA



Osipenko Maksim
Feature engineering

Thx

Skoltech