Optimisation of customer lists for communication based on contact history

Emil Akopyan, Yaroslav Ruban, Alexander Stepin, Бирюкова Ирина

Problem statement

- Predict whether the client will respond to us
- Contact policy: business logic and performance analysis according to it

Goals

- EDA: to gain insights of the data and visualise it;
- Modelling: to predict the clients' response;
- Web-interface: to provide accessibility for laymen.

Contact policy

Companies aim to contact only certain clients.

- To evaluate the performance of the model, we need to take into account the needs of our business!
- Recall represents how many potentially interested we contact. Probably, the most important metric!
- Precision: the fraction of clients that we contact who will respond.

EDA

19 features

10 categorical features

1. Ind_household
2. Ind_deposit
3. Ind_email
4. Ind_phone
5. Ind_salary
6. Region
7. Gender
8. District
9. Age_group
10. Segment

binary

ordinal

nominal

9 numerical features

1. Lifetime
2. Age
3. Income
4. trans_3_month
5. trans_6_month
6. trans_9_month
7. trans_12_month
8. amont_trans
9. amont_day_from

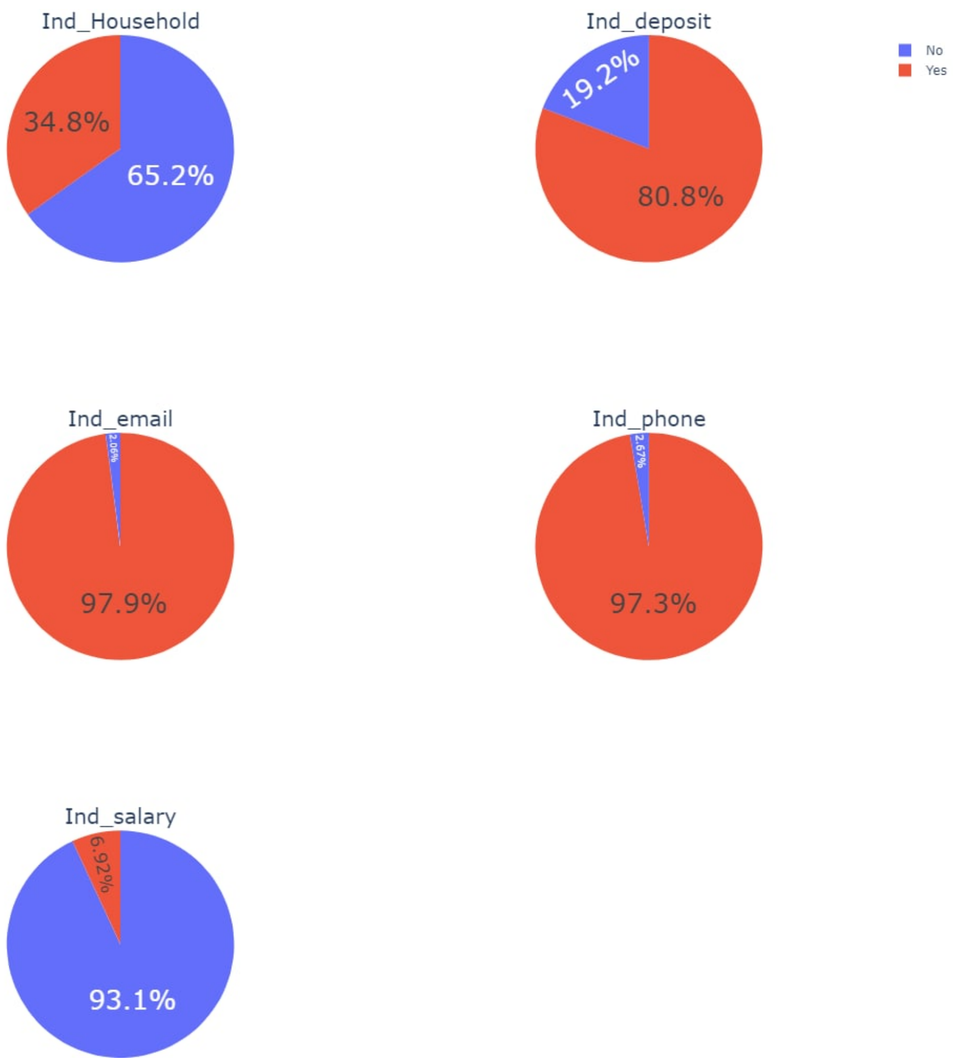
normally distributed

F-distributed

EDA.Categorical Features

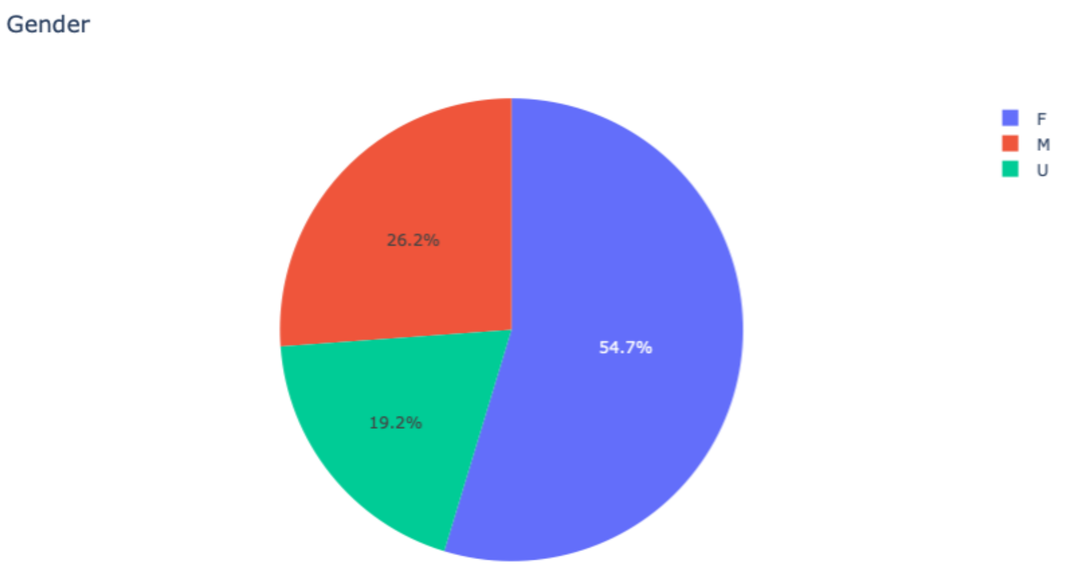
Nominal binary

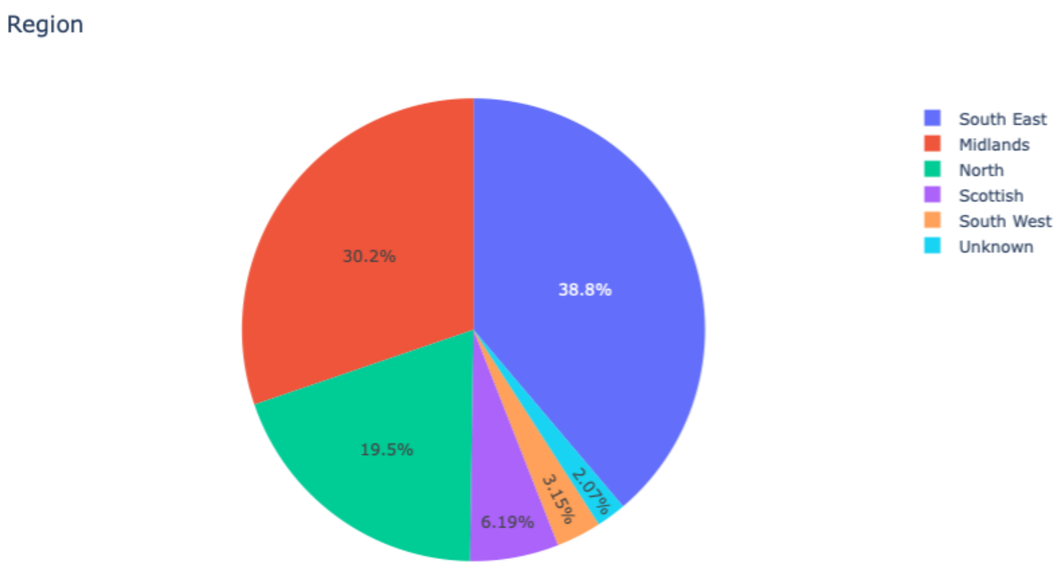
Highly imbalanced classes, difference between fractions leads to obstacles when a simple model is performed



EDA.Categorical Features

Unknown value is presented and should not be dropped



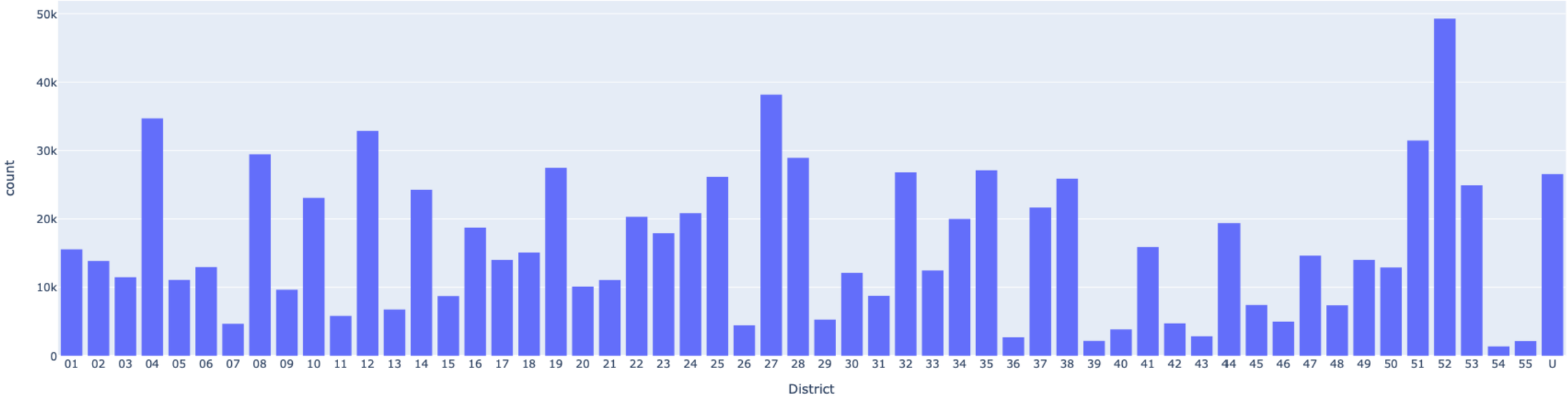


probably a significant share of men does not specify their gender

EDA.Categorical Features

Most of the clients based in 52 district. The sufficient number of clients districts are unknown.

It is possible that the bank is not that popular in certain regions. Both competition and underdevelopment might be the reasons for that



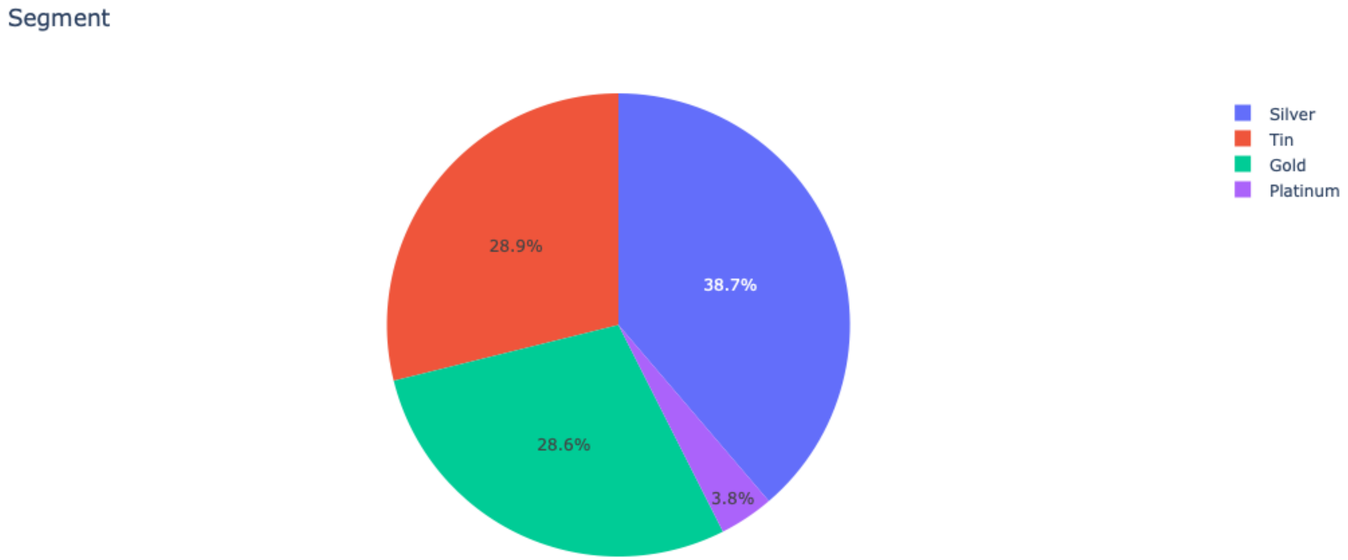
EDA.Categorical Features

Connection between Age_group and Age

In general, it occurs that most bank's clients are not very young

EDA.Categorical Features

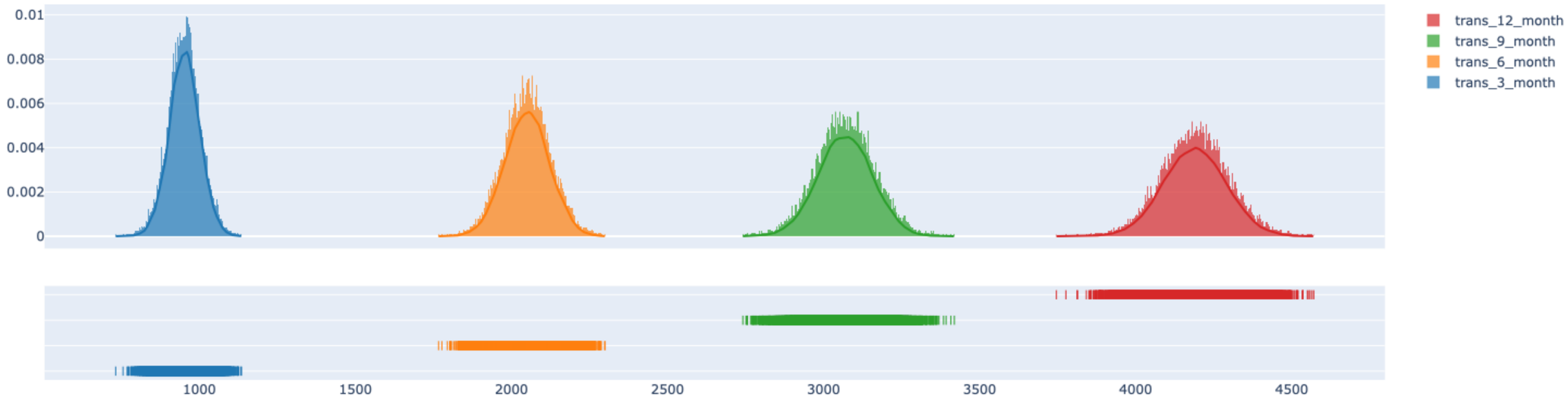
All segments share more or less the same value except for "platium". Probably, platinum clients are clients which make huge transactions or have significant savings in the bank



EDA.Continuos data.Normal distribution

trans_X_month distribution

Normally distributed features,
each iteration being shifted by 1000

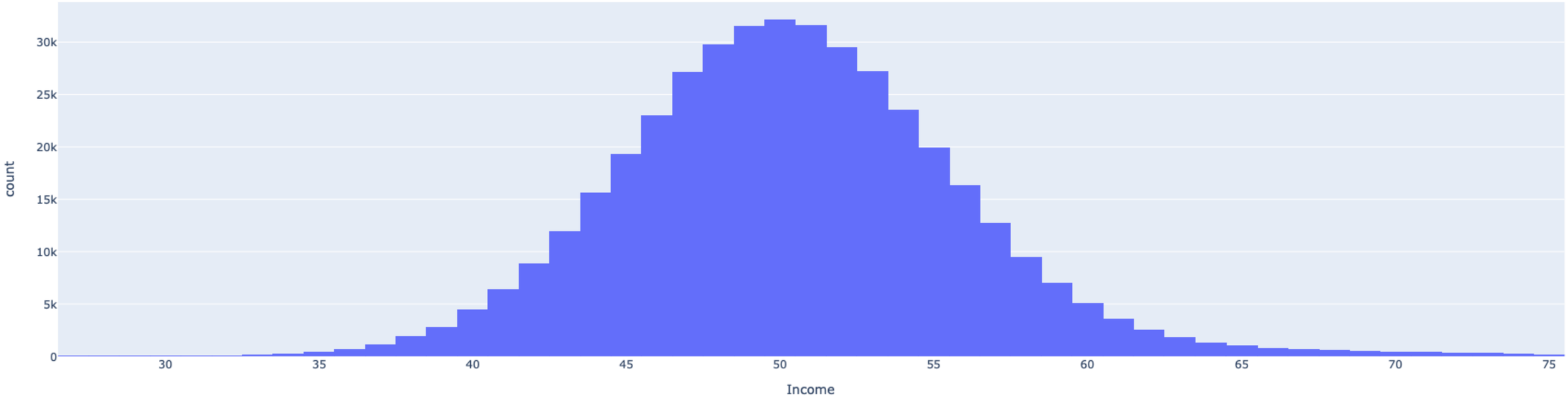


EDA.Continuos data.Normal distribution

Income distribution

Normally distributed features

mean	50.351041
std	5.442882
min	27.000000
25 %	47.000000
50 %	50.000000
75 %	54.000000
max	75.000000
median	50.000000

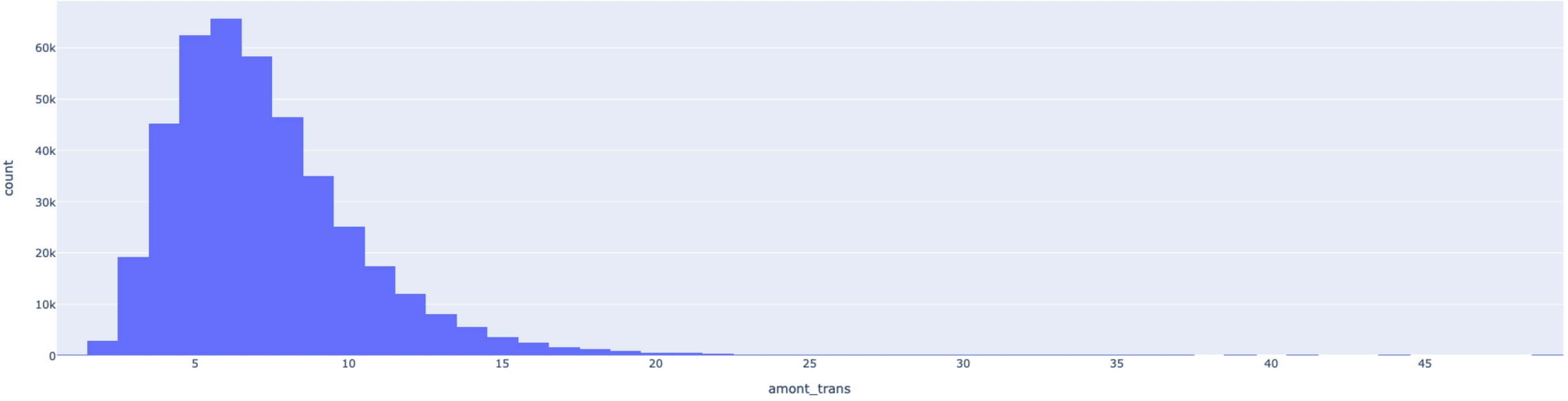


EDA.Continuos data

amont_trans distribution

Chi-squared distribution or F-distribution
with small number of freedom degrees.

mean	7.243598
std	3.035935
min	1.000000
25 %	5.000000
50 %	7.000000
75 %	9.000000
max	49.000000
median	7.000000

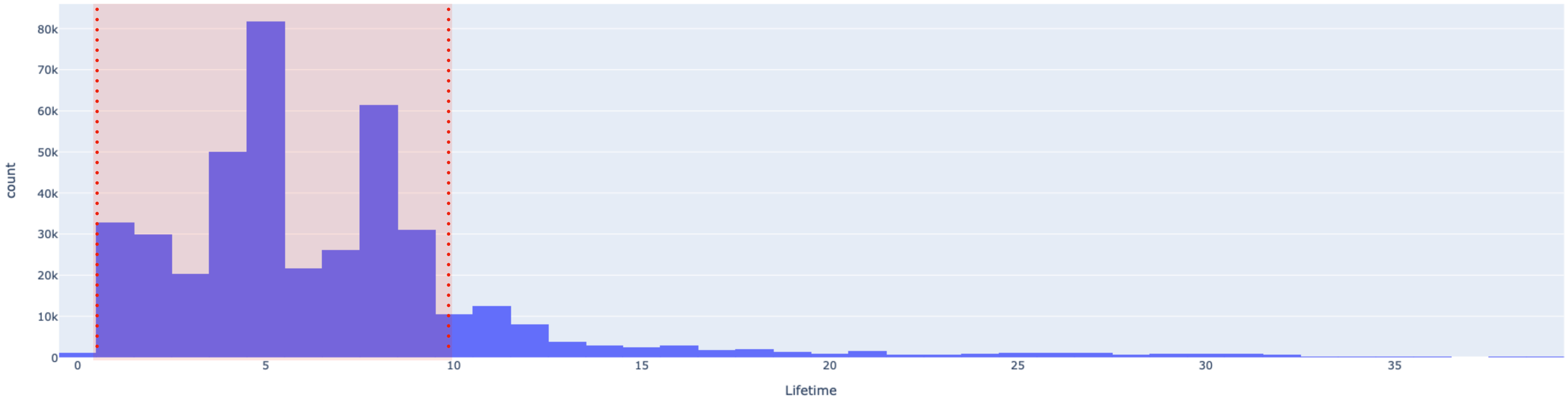


EDA.Continuos data.Non-standart

Lifetime

Data is right-skewed,
with most data centred around value 5.

mean	6.567161
std	4.681223
min	0.000000
25 %	4.000000
50 %	5.000000
75 %	8.000000
max	39.000000
median	5.000000

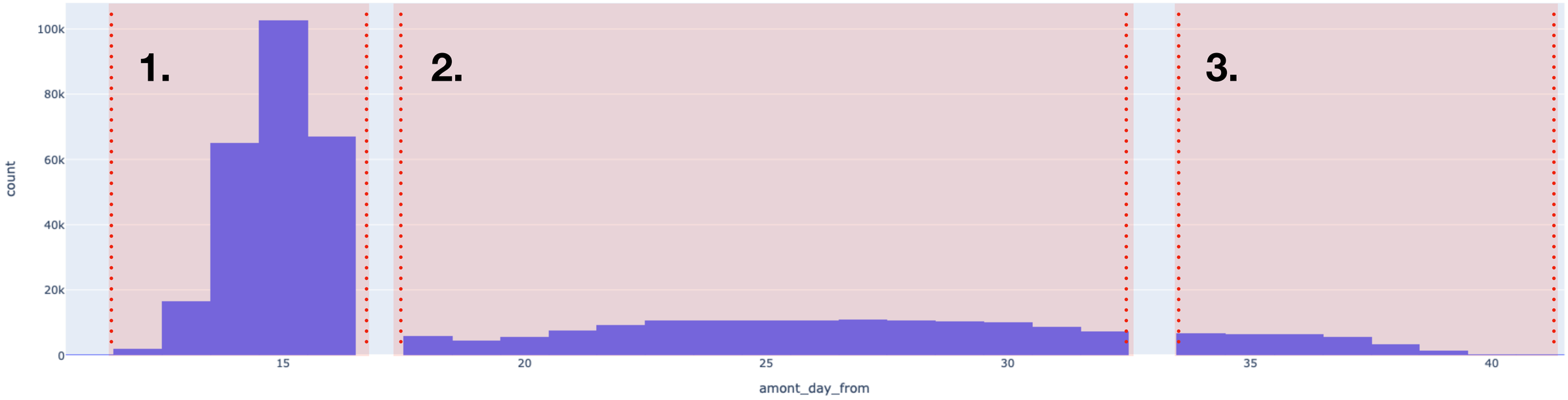


EDA.Continuos data.Non-standart

Amont_day_from

mean	19.742113
std	7.019663
min	11.000000
25 %	15.000000
50 %	16.000000
75 %	25.000000
max	41.000000
median	16.000000

Data is right-skewed,
The data can be presented as three clusters,
Cluster 1 being the most largest.

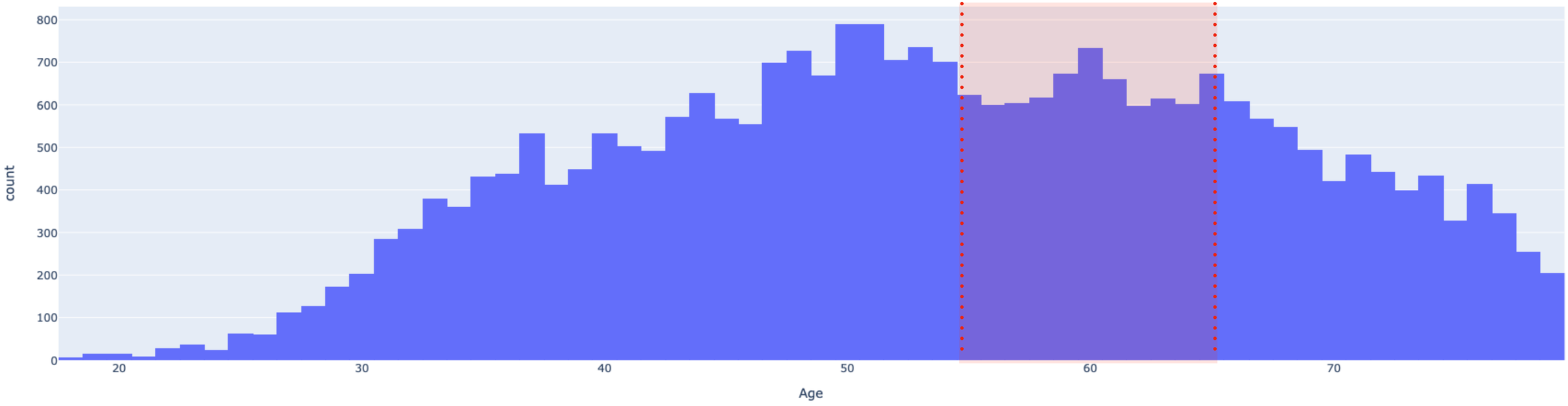


EDA.Continuos data.Non-standart

Age

Data is left-skewed,
With average age being 54

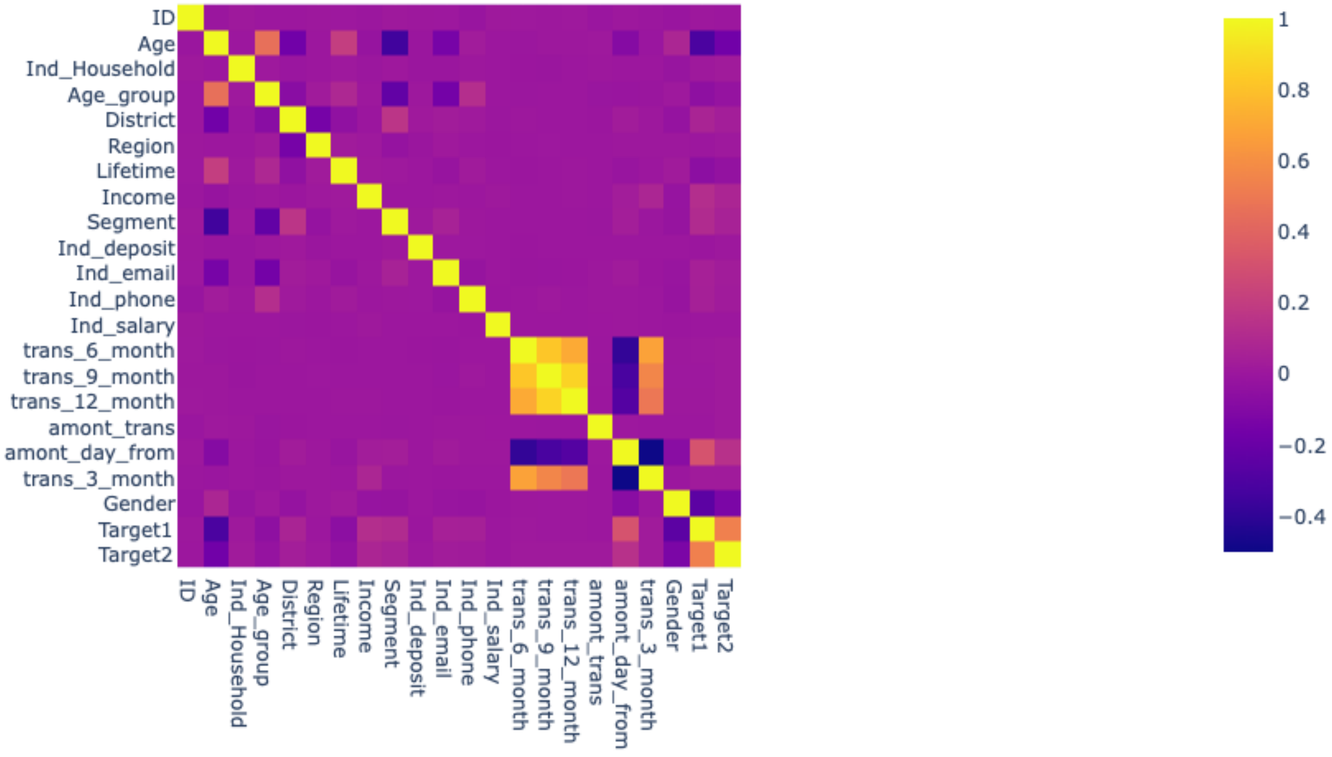
mean	53.782686
std	13.169368
min	18.000000
25 %	44.000000
50 %	54.000000
75 %	64.000000
max	79.000000
median	54.000000



EDA.Correlation table

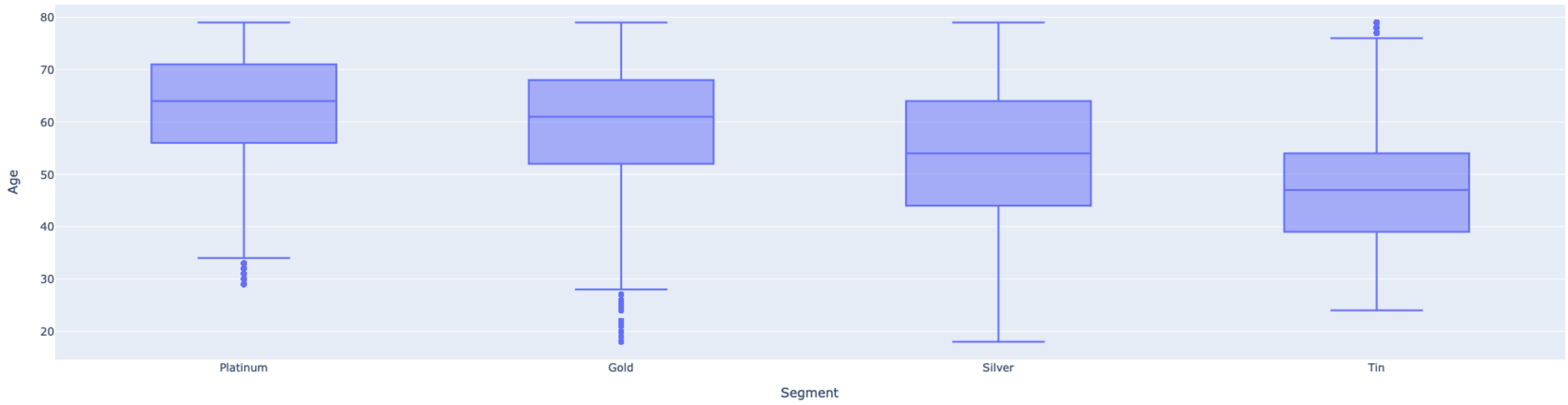
Significant correlation:

1. Age and Segment
   positive
2. trans_X_month
   positive
3. trans_X_month and amont_day_from
   negative



EDA.Correlation

Connection between Segment and Age



Modelling

- Gradient boosting!
  - Probably, the most powerful algorithm for tabular data
  - Computationally expensive
  - Sometimes prone to overfitting
  - Not interpretable

- Hyperparameter search:?
  - GridSearch Parameter search with (GridSearchCV), expensive yet comprehensive

Model Evaluation

- Cross Validation, producing standard classification reports with averaged metrics.

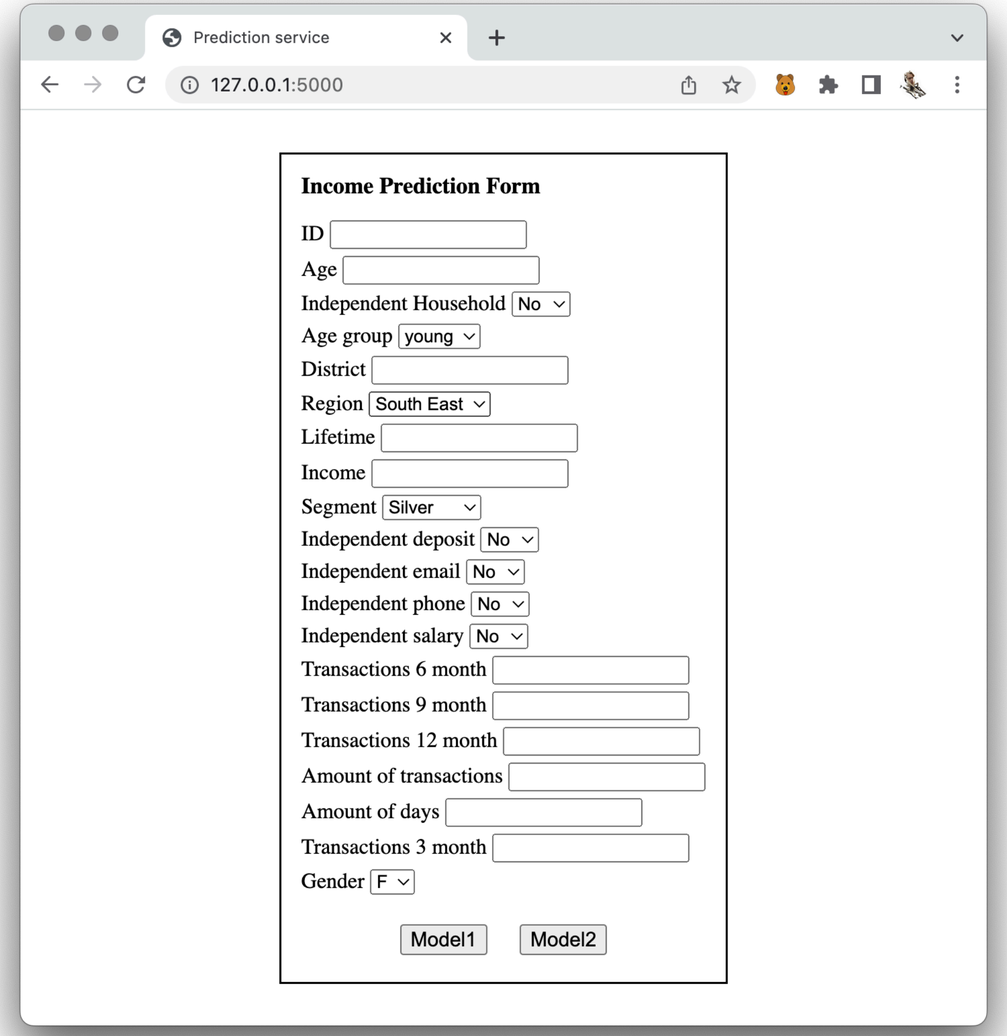
Email

	precision	recall	f1-score	support
0	0.98	0.99	0.99	185563
1	0.98	0.93	0.96	60807
accuracy			0.98	246370
macro avg	0.98	0.96	0.97	246370
weighted avg	0.98	0.98	0.98	246370

Sms

	precision	recall	f1-score	support
0	0.98	0.99	0.99	225301
1	0.93	0.86	0.85	21069
accuracy			0.98	246370
macro avg	0.96	0.89	0.92	246370
weighted avg	0.98	0.98	0.98	246370

User Interface



Simple

Build on Flask

Compatible with backend

Role assignment

- Explanatory Data Analysis
  1. Categorical Data — Бирюкова Ирина, Alexander Stepin
  2. Numerical Data — Yaroslav Ruban
  3. Visualisation — Yaroslav Ruban, Бирюкова Ирина

- Modelling and validation — Emil Akopyan

- Web-interface
  1. Web-interface for individual client — Alexander Stepin
  2. Web-interface for datase — Alexander Stepin

- Presentation — Бирюкова Ирина