OPTIMIZING YOUR COMMUNICATION CLIENT LIST

BASED ON TRANSACTION HISTORY

Ilshat Dineev, Oleg Bobylev, Alexandra Arkhipova DSBA213

TABLE OF **CONTENTS**

Project's Goal

Response Model

Build prediction response model for optimizing communication list and represent pricing model

Segmentation

Find and describe main segments of our client base

Data Research

1.

Exploratory data analysis

Dataset overview

Represented

Personal data, transaction history, responses

Consists of

22 variables, 13 categorical, 9 quantitative

Name	Type	Label
ID	Character	ID клиента
Ind_Household	Character	Факт домовладения
Age_group	Character	Возрастная группа
District	Character	Район
Region	Character	Регион
Segment	Character	Статус клиента
Ind_deposit	Character	Индикатор владения депозитом
Ind_email	Character	Индикатор наличия e-mail
Ind_phone	Character	Индикатор наличия телефона
Ind_salary	Character	Индикатор владения зарплатной картой
Gender	Character	Пол
Target1	Character	Отклик на коммуникацию по e-mail
Target2	Character	Отклик на коммуникацию по телефону
Age	Numeric	Возраст
Lifetime	Numeric	Время, проведенное с банком
Income	Numeric	Доход
trans_6_month	Numeric	Транзакции за 6 месяцев
trans_9_month	Numeric	Транзакции за 9 месяцев
trans_12_month	Numeric	Транзакции за 12 месяцев
amont_trans	Numeric	Кол-во транзакций
amont_day_from	Numeric	Количество дней с последней транзакции
trans_3_month	Numeric	Транзакции за 3 месяца

Dataset **distribution**

We tried to build graphs for response and non-response clients, but distribution was as in the general case

CLIENTS **PROFILE**

GENDER

67%

33%

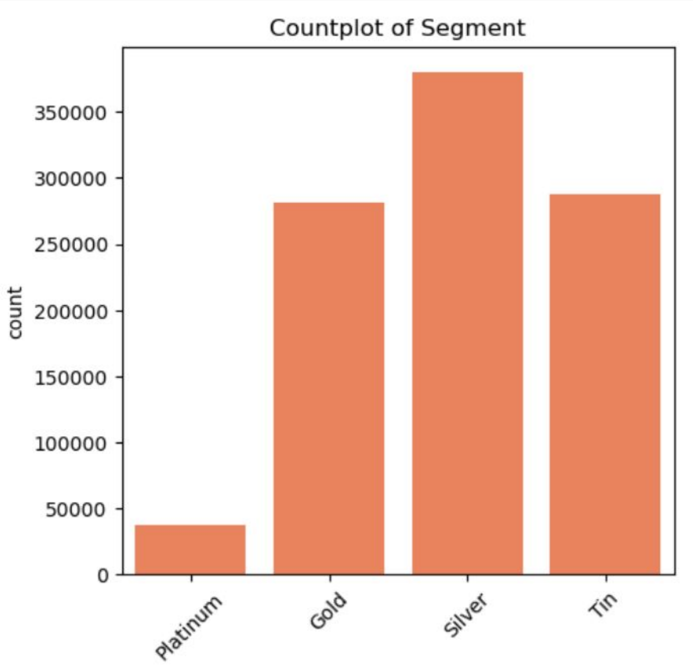
AGE

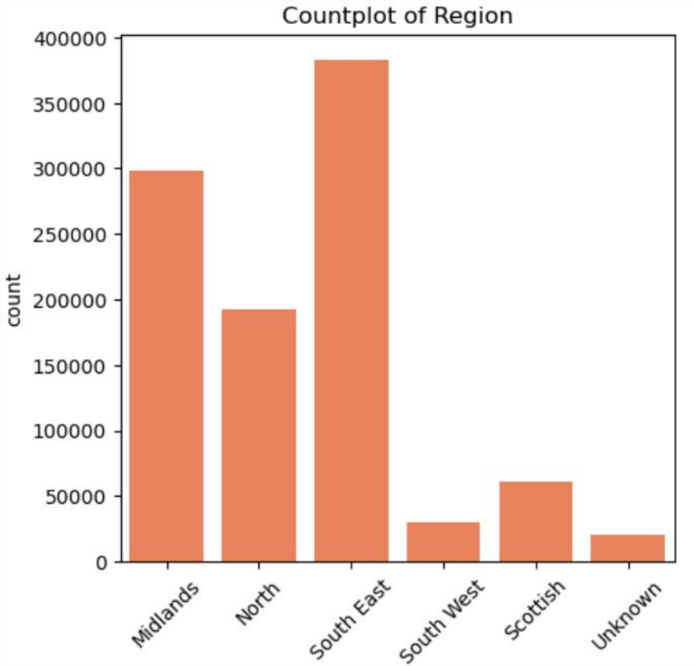
18-30 3%

30-60 63%

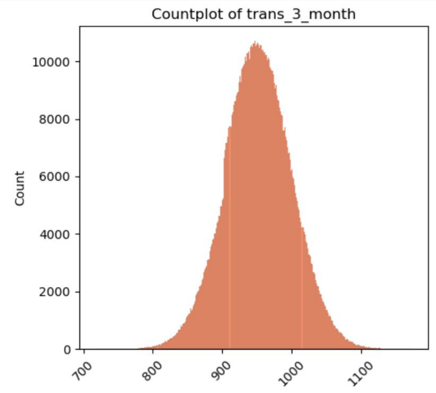
60-79 33%

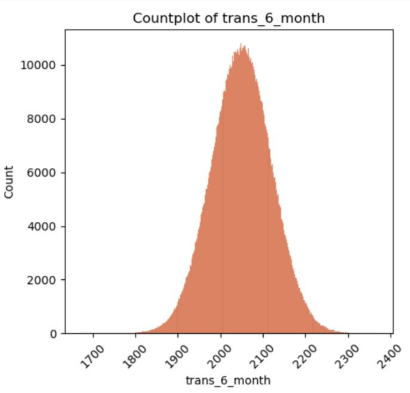
Dataset **overview**

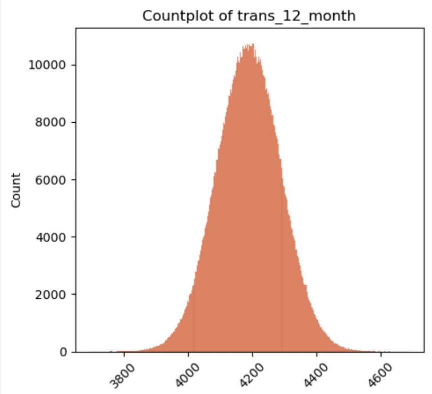


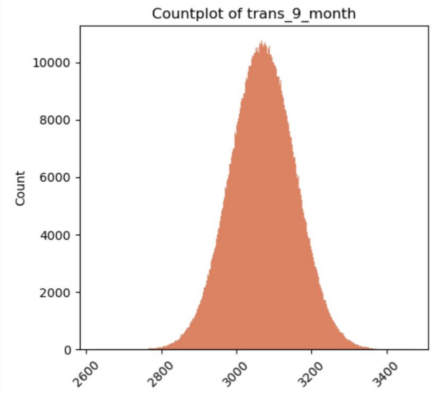


ACTIVITY OF **Transaction**

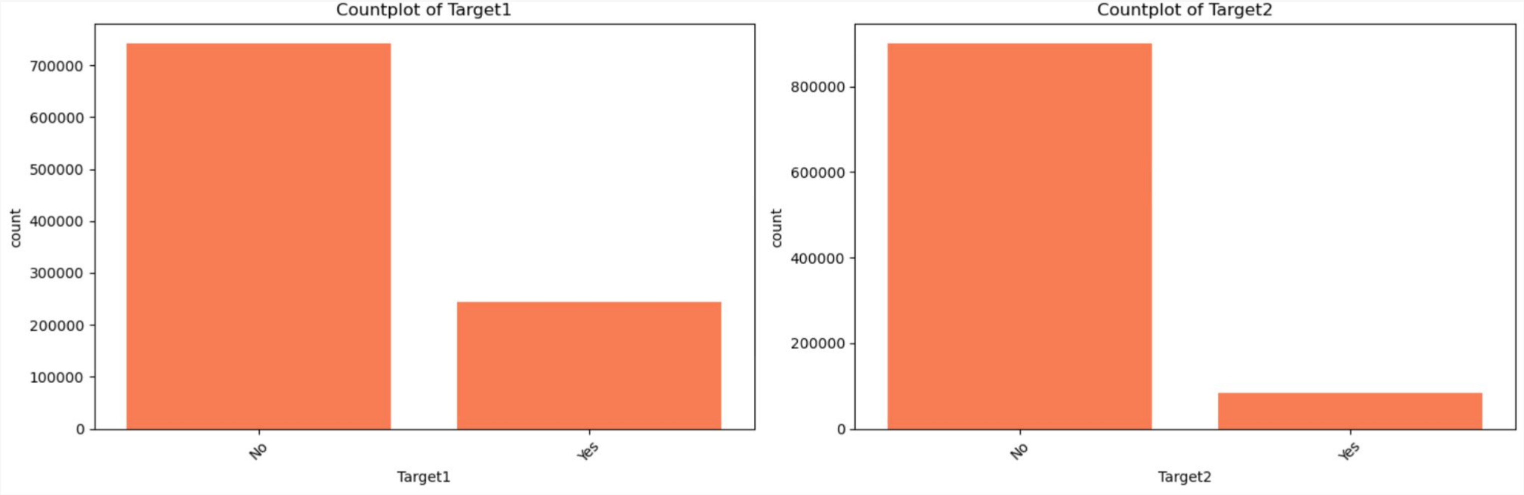








DISTRIBUTION OF **TARGET**



Disbalance of classes - we can't use accuracy metric

client average

Age	53.792
Lifetime	6.562
Income	50.356
trans_6_month	2049.943
trans_9_month	3069.965
trans_12_month	4189.979
amont_trans	7.240
amont_day_from	19.742
trans_3_month	952.514

Ind_Household	No
Age_group	middle
District	52
Region	South East
Segment	Silver
Ind_deposit	Yes
Ind_salary	No
Gender	F
Target1	No
Target2	No

Average of responded

Age	46.799
Lifetime	6.127
Income	51.443
trans_6_month	2050.057
trans_9_month	3070.120
trans_12_month	4190.191
amont_trans	7.240
amont_day_from	23.652
trans_3_month	953.352

Ind_Household	No
Age_group	middle
District	52
Region	South East
Segment	Silver
Ind_deposit	Yes
Ind_salary	No
Gender	F
Target1	Yes
Target2	No

2.

Data cleaning preprocessing

MISSING VALUES

MISSING VALUES

```
Age_group     0
Region        0
District      0
Segment       0
dtype: int64
Ind_Household     0
Ind_deposit       0
Ind_salary        0
Gender            0
dtype: int64
Age               66958
Lifetime          12608
```

```
Income                 0
trans_6_month          0
trans_3_month          0
trans_9_month          0
trans_12_month         0
amont_trans            0
amont_day_from         0
dtype: int64
Target1     0
Target2     0
```

SOLUTION

- Imputation with average of numerical variables
- Leave unknown categorical variables as they can have impact

Variables encoding

Categorical

One-hot encoding with dropping one variable to avoid multicollinearity

Binary and Target

LabelEncoder for binary and target variables

Variables **scaling**

scaling

We applied StandardScaler
form scikit-learn

Variables Deleting

Deleted

We deleted Ind_email and Ind_phone as they are omitted in test dataset

VARIABLES **INCORRECTNESS**

INCORRECT

```
mask = (
    (data['trans_6_month'] < data['trans_3_month']) |
    (data['trans_9_month'] < data['trans_6_month']) |
    (data['trans_9_month'] < data['trans_3_month']) |
    (data['trans_12_month'] < data['trans_9_month']) |
    (data['trans_12_month'] < data['trans_6_month']) |
    (data['trans_12_month'] < data['trans_3_month'])
)
```

DATASET **FINAL**

	ID	Age	Ind_Household	Lifetime	Income	Ind_deposit	Ind_salary	trans_6_month	trans_9_month	trans_12_month	...	District_50	District_51	District_52	District_53	District_54	District_55	District_U	Seg
0	1200000001	-0.219	0	-0.770	0.486	0	0	-0.335	-1.221	-0.491	...	0	0	0	0	0	0	0	
1	1200000002	-0.533	0	-0.986	0.118	0	0	-0.238	-1.162	0.126	...	0	0	0	0	0	0	0	
2	1200000003	-0.690	0	-0.122	-0.065	0	0	0.505	0.117	0.875	...	0	0	0	0	0	0	0	
3	1200000004	1.900	0	1.176	0.302	1	0	-0.330	-0.792	-0.844	...	0	0	0	0	0	0	0	
4	1200000005	0.252	0	0.311	0.302	1	0	1.305	1.365	2.120	...	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
985472	1201048571	-0.000	0	-0.122	-1.536	1	0	1.260	0.987	1.020	...	0	0	0	0	0	0	0	
985473	1201048572	0.723	0	0.527	0.118	0	0	-1.881	-0.808	-1.448	...	0	0	0	0	0	0	0	
985474	1201048573	1.115	0	-0.770	1.037	1	0	0.043	0.495	1.022	...	0	1	0	0	0	0	0	
985475	1201048574	-0.690	0	-1.203	0.670	1	0	0.465	0.852	0.175	...	0	0	0	0	0	0	0	
985476	1201048575	1.351	1	0.311	-1.720	1	0	1.262	1.629	2.014	...	0	0	0	0	0	0	0	

3.

SEGMENTATION

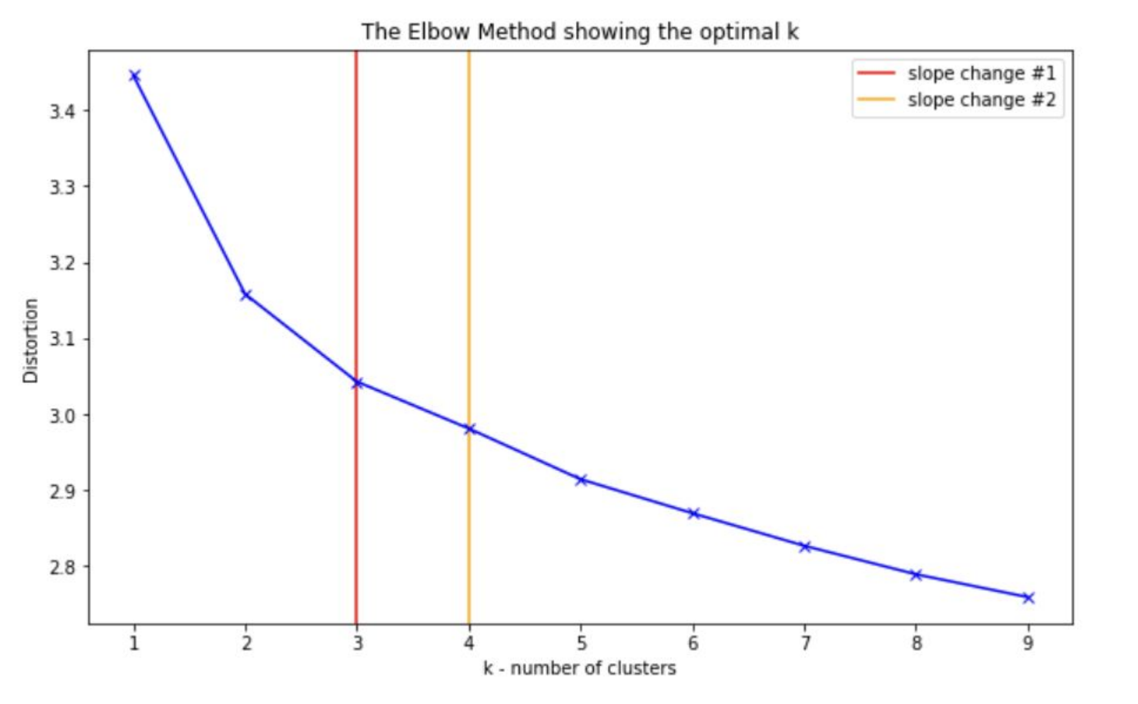
Clusterization

MODEL CHOICE

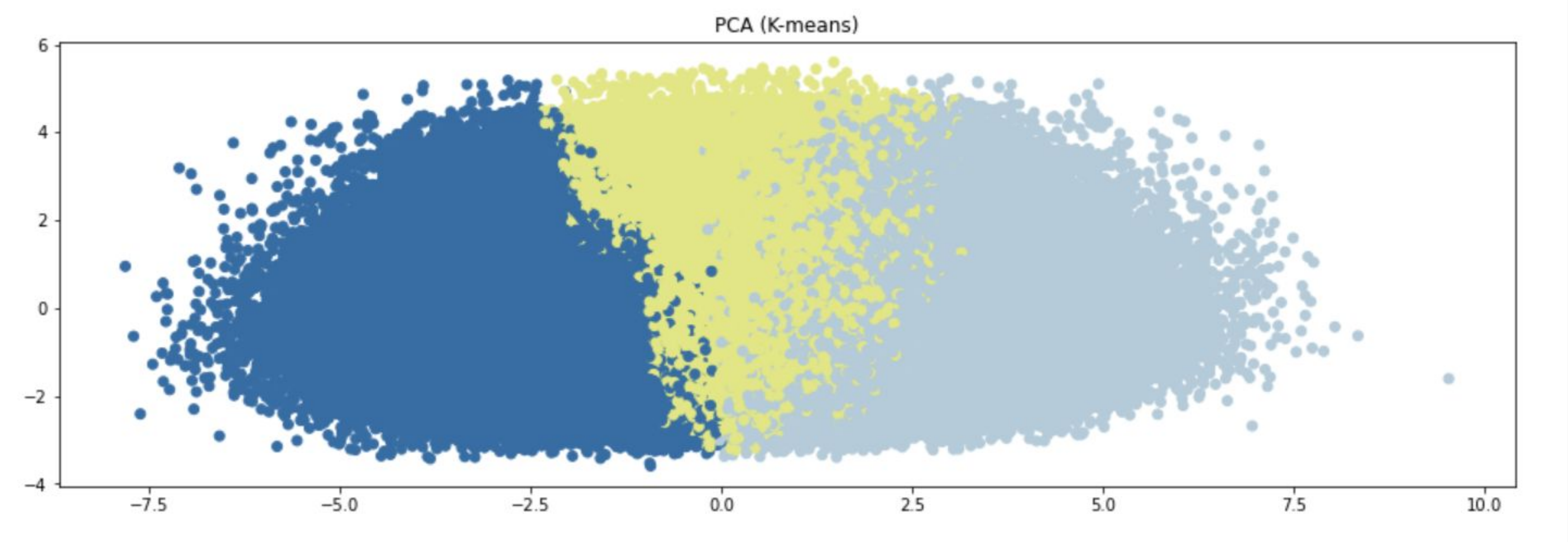
K-means

This method is excellent for market segmentation, including customer segmentation

ELBOW METHOD



CLUSTERIZATION



4. Prediction of e-mail and phone responses

Model building

Model choice

LogReg

Simple linear model based on mathematical regression

Random Forest

Ensemble of several Decision Trees that vote for class

Gradient Boosting

Builds on what was learned before, correcting mistakes from past to get better and better

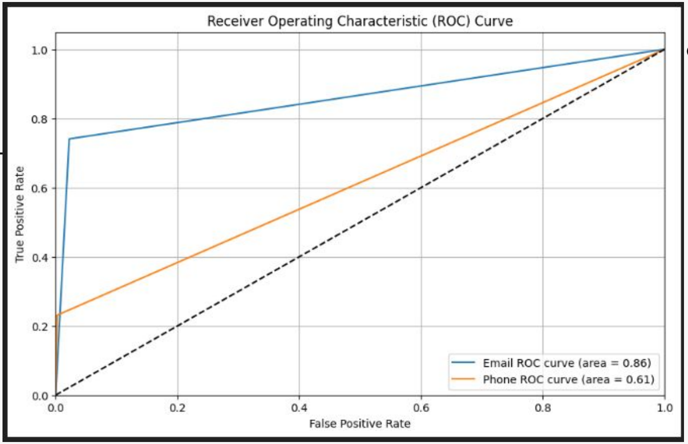
RESULTS

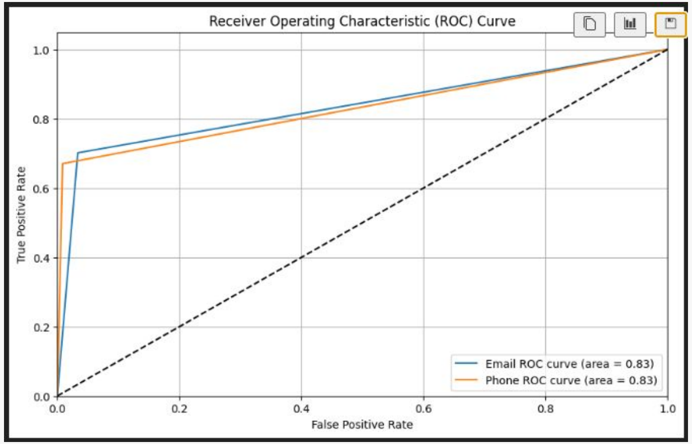
Model		LogReg	Random Forest	Gradient Boosting
Email Response	ROC-AUC	0,5	0,85	0,84
	MSE	0,25	0,08	0,09
	F1-Score	0	0,81	0,78
Phone Response	ROC-AUC	0,5	0,62	0,83
	MSE	0,08	0,06	0,0362
	F1-Score	0	0,36	0,7607

```
Logistic Regression - Email Response ROC AUC Scores: [0.49967766 0.49858098 0.50054241 0.5009195  0.5001732 ]
Logistic Regression - Email Response Mean ROC AUC: 0.49997875051871443
Logistic Regression - Phone Response ROC AUC Scores: [0.49711913 0.5014785  0.49883814 0.50031121 0.49530158]
Logistic Regression - Phone Response Mean ROC AUC: 0.49860971325246767

Random Forest - Email Response ROC AUC Scores: [0.97022611 0.96982712 0.96837659 0.96753044 0.96817315]
Random Forest - Email Response Mean ROC AUC: 0.968826680981574
Random Forest - Phone Response ROC AUC Scores: [0.93573472 0.93935581 0.94008771 0.93808432 0.9362528 ]
Random Forest - Phone Response Mean ROC AUC: 0.9379030723363329

Gradient Boosting - Email Response ROC AUC Scores: [0.93127361 0.93174453 0.93143939 0.93079123 0.93175575]
Gradient Boosting - Email Response Mean ROC AUC: 0.9314009020084605
Gradient Boosting - Phone Response ROC AUC Scores: [0.96868991 0.9710062  0.96966098 0.97032798 0.96893355]
Gradient Boosting - Phone Response Mean ROC AUC: 0.9697237227120207
```





Final Product

E-mail cost: 1$

Phone cost: 10$