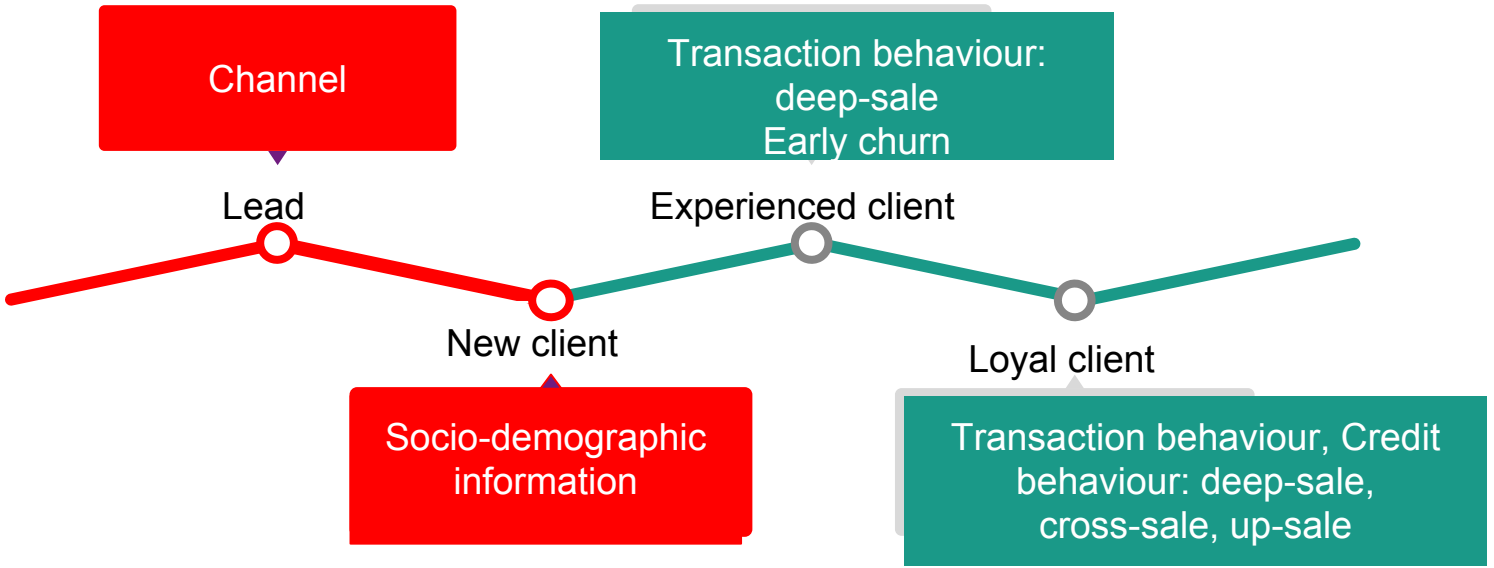
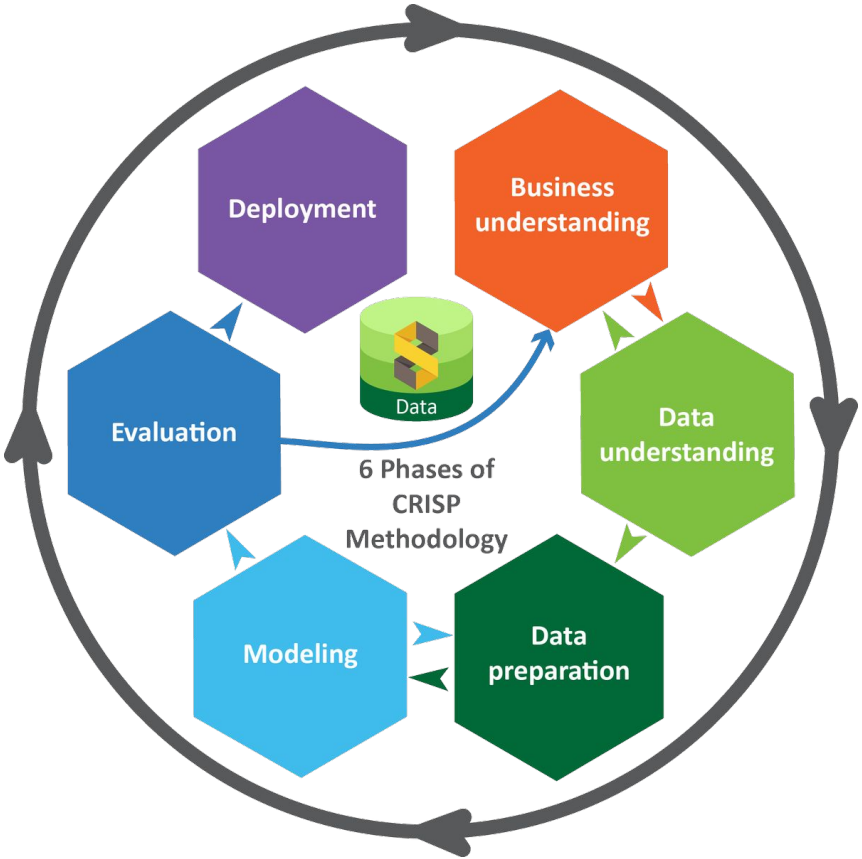
Predictive analysis: the way to get a satisfied customer

Tatyana Fursova 2018

Client journey



CRISP - DM


Deployment
Business understanding
Data
Evaluation
Data understanding
6 Phases of CRISP Methodology
Modeling
Data preparation

DATAMART in retail banking

Available data

- Socio-demographic information

Client_ID	Birth_date	Gender	Marital_status	Children	Income
23435	01.01.1985	M	Single	0	15000

- Customer-Product (ownership)

Client_ID	Prod_ID	Open_Dt	Close_Dt	Issue_loc	Cur_status
23435	25252	01.01.2018	01.01.2021	POS22	Active

- Product catalogue

Prod_ID	Prod_Name	Type_1	Type_2	Luonch_dt	Remove_dt	Rate
23435	Legko	Card	Credit card	01.01.2005	31.12.2010	0.4

- Transactions

TRX_ID	Prod_id	TRX_dt	Direction	Amount	Type	Channel	Place
12121	25252	01.01.2018	Credit/debit	200	Purchase	Terminal	Food market

- Balances

Balance_dt	Prod_id	Credit_limit	Debit	Credit	DPD_days	DPD_amnt
01.01.2018	25252	2000	0	1000	1	500

Aggregation

- Payments
- History of applications for purchasing products
- Campaign contacts and responses
- Recorded complaints and complaint outcomes
- Business segmentation (corporate, affluent, mass, etc.)
- History of BKI

Aggregation

Client_ID	Reference_date	Gender	Total prod cnt	Credit cards open cnt	Credit_trx_cnt	Credit_trx_cnt 1m
23435	01.01.2016	M	5	1	1500	100

For continuous variables:

Frequency/Monetary
- Total_cnt (amnt)
- Avg_cnt (amnt)
- Min_cnt (amnt)
- Max_cnt (amnt)
- Trend (amnt)

Ricency
- Min_days_between
- Max_days_between
- Avg_days_between
- Days_from_last
- Days_from_first

All needed categories transport to fields:

TRX_ID	Prod_id	TRX_dt	Direction	Amount	Type	Channel	Place
12121	25252	01.01.2018	Credit	200	Purchase	Terminal	Food market
12122	25252	01.01.2018	Credit	150	P2P	web	P2P
12123	25252	01.01.2018	Credit	500	On-line	web	Fashion

Client_ID	Reference_dt	CC_ trx_cnt	CC_Total_amount	CC trx_Food market_cnt	CC_Fashion_ cnt	CC_trx_P2P_cnt
23435	01.01.2016	3	850	1	1	1

Basic steps

- Business task (with KPI determination)
- Population determination
- Target determination
- Evaluation metrics selection
- Data preparation
- Algorithm selection
- Evaluation

Likelihood to become potencial credit card user

Business task (with KPI determination):

Person who came for consumer credit to the POS-network. We are looking for potential user of Credit Card (he has use his credit limit). We need to improve our KPI: Utilization rate= (number of utilized cards)/(Number of issued cards) and Total Portfolio (Credit limit used in money)

Population determination:

All clients who take a POS-credit (date between ‘01.01.2017’ and ‘01.01.2018’)

Target determination:

We need to generate the target!

Evaluation metrics selection:

Auc Roc and Lift

Data preparation:

We have some limitations. Those clients are new so we have only Socio-Demo and BKI

Algorithm selection

The main limitation - easy model interpretation

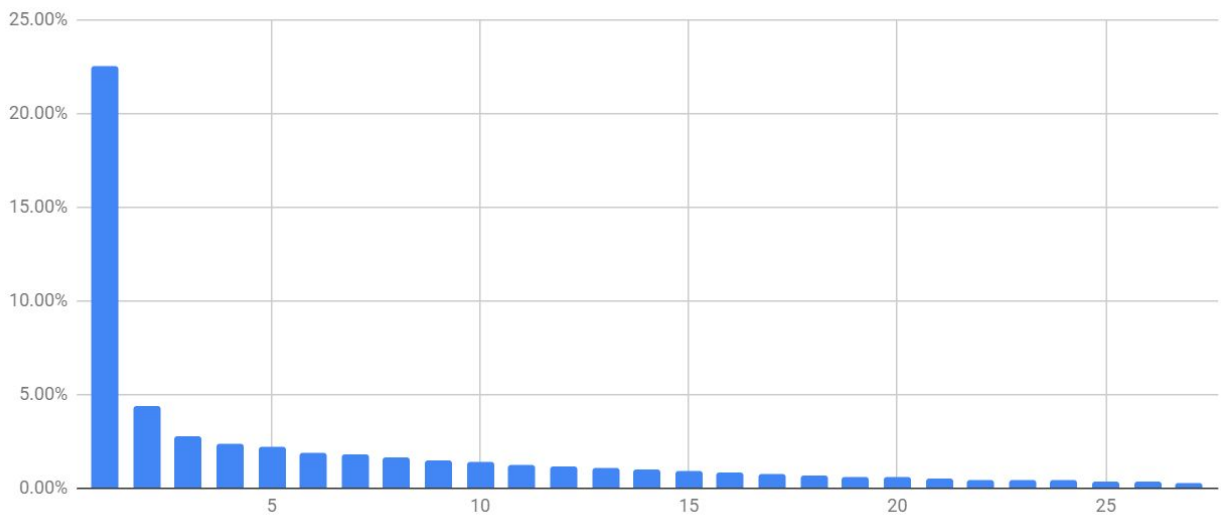
Target

What is “card utilization’?

TRX_cnt	Clients_Share%
0	67%
1	7%
2 and more	26%

- 23% of clients has only 1trx per year
- 4% - 2 transactions
- 3% - 3 transactions
- etc.

Clients with one and more transactions distribution (for 1Year)



Target - clients who make 2 and more transactions in the period (1 year)

New client → understand his profile

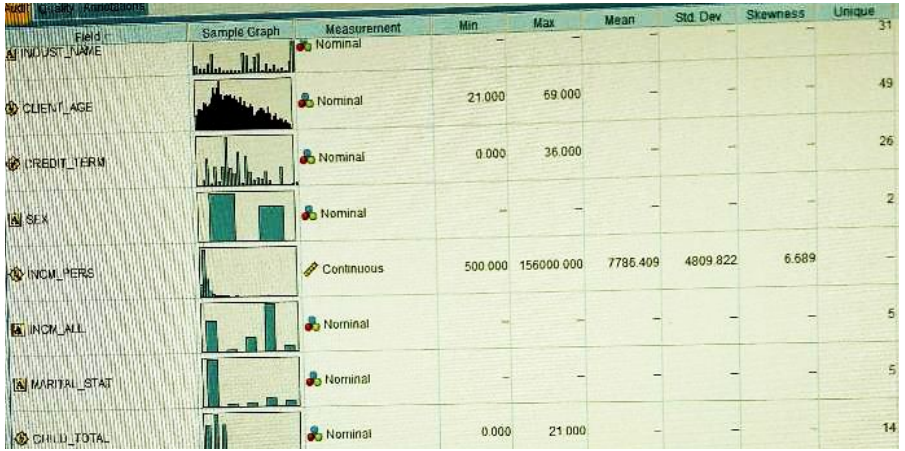
Acquisition. Inputs audit

Input Name	Input Type
Marital_status	Nominal
LiveRegion	Nominal
RegRegion	Nominal
Work_expirience	Continuous
Industry	Nominal
Age	Continuous
Sex	Flag
Credit_amount	Continuous
Desired_1st_amount	Continuous
Credit_term	Nominal
Owned_Realty_type	Nominal
Good1_Category	Nominal
Income source	Nominal
Children	Continuous
Goods_quantity	Nominal
Old_client	Flag
Education	Nominal
Income	Continuous
Position	Nominal
BKI debt	Continuous

Make summary():

- Min
- 1st quartile
- Median
- Mean
- 3d quartile
- Max
- Unique
- NA
- Total_cnt
- Cross-correlation

Visualize:



- NA?
- Outliers?
- Skewness?

New client → understand his profile

1 Unbalanced data

- We get 74% (Bads) vs 26% (Goods)

Data balancing:

- Under - sampling (Risk - to lose information)
- Over- sampling (Risk - overfitting)

Data partitioning

Algorithms

Classification problem:

1. Decision tree
2. Random forest (interpretation limits)
3. XGBoost (interpretation limits)
4. LogRegression

Evaluation (1st iteration):

1. Decision tree: Auc=0.65, Lift=1.4
2. LogRegression: Auc=0.65, Lift=1.5

2 Feature engineering

Live_Reg_flg: If LiveRegion=RegRegion then 1

Income_credit_share: Income/Credit_amount

Binning for continuous variables

Age_Optimal

Age group	Good share	Clients_cnt_share
25-35	39%	25%
36-45	30%	45%
46-55	22%	20%
>55	17%	10%

LogRegression on nominal inputs

Evaluation (2st iteration):

Partition	Training	Testing
Auc	0.71	0.70
Lift	1.95	1.90

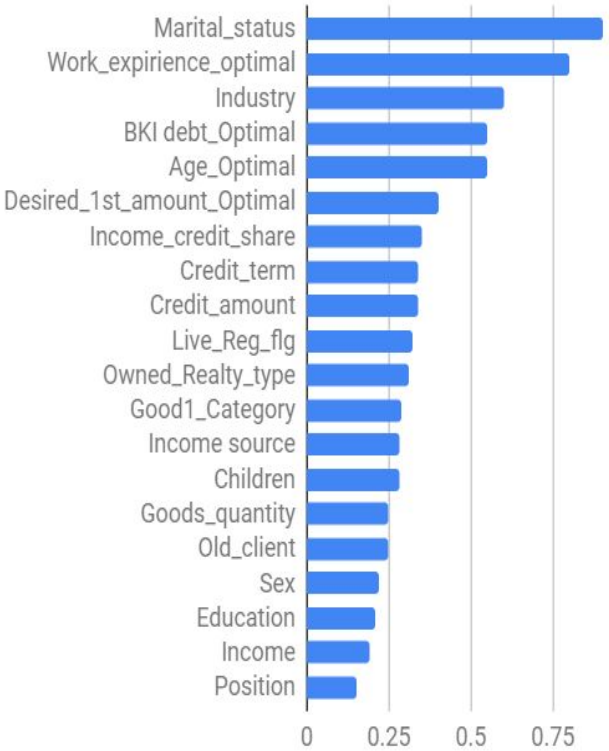
New client → understand his profile

Model

Predictor	Weight	Client
Marital_status_1	1.32	1
Marital_status_2	-0.1	
Marital_status_3	-0.05	
Work_expirience_optimal_1	1.2	
Work_expirience_optimal_2	0.5	
Work_expirience_optimal_3	0.32	1
Industry_1	0.99	
Industry_10	-2.5	1
BKI debt_Optimal_1	1.1	
BKI debt_Optimal_2	0.5	
Age_Optimal_1	0.56	1
	y=	**-0.3**

$$\text{Score} = \frac{1}{1 + e^{-y}} = 0.90$$

Importance



Profile

Factor	Not Likely to	Likely to
Age	More	Less
Work_expirience	More	Less
Desired_1st_amount	More	Less
Credit_term	Less	More
Live_Reg_flg	Yes	No
Children	Less	More
Goods_quantity	Less	More
Education	More	Less
Income	More	Less

Deep sail. Event Sequences

Business task: Deep-sale: Predict the next location for client's transaction

Algorithm selection: Association rules

Data preparation:

Client_ID	Trx time	Trx location
11111	week1	EntertaimentService
11111	week2	RestoranService
11111	week3	Transport
11111	week4	FoodMarket
22222	week1	RestoranService
22222	week2	Transport
33333	week1	Transport
33333	week2	FoodMarket
33333	week3	ATM
33333	week4	FoodMarket
44444	week1	Transport
44444	week2	FoodMarket
44444	week3	ATM
44444	week4	RestoranService
44444	week5	Transport

Rule of an association detection model:

Rules	Consequent	Antecedents	Support %	Confidence %
Rule 1	Transport =>	RestoranService	75%	100%
Rule 2	ATM=>Transport	FoodMarket	75%	67%

Evolution.

Business task:

Deep-Sale: Credit Customers segmentation by their transaction behaviour

Population determination:

All clients with Credit Cards

Target determination:

Unsupervised learning

Evaluation metrics selection:

Elbow method

Data preparation:

We have Socio-Demo, Transactions

Algorithm selection

The main limitation - easy model interpretation

Experienced client→ deep sail

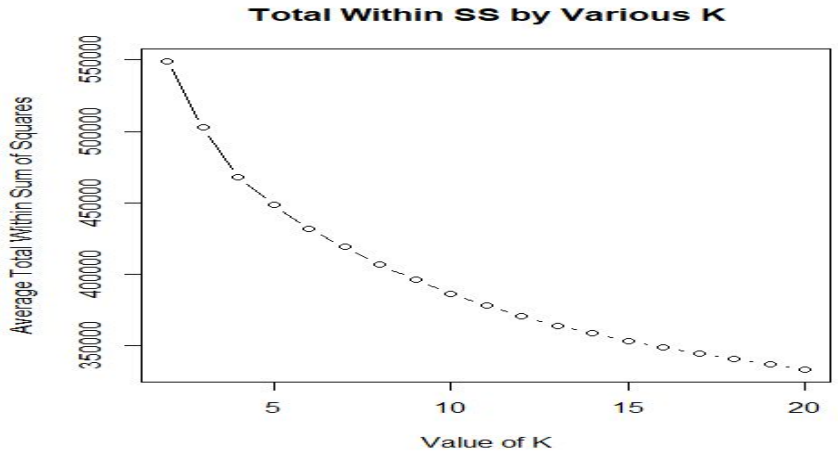
Inputs

Input Name	Input type
Weekly_amount	num
Weekly_trx_cnt	num
Week_unique_locations	num
TRX_amnt_mean	num
TRX_cnt_avg	num
Total_amount	num
Total_cnt	num
Total_unique_locations	num
Count_city	num
Gender	factor
Age	num
Days_btwn_transactions	num
Foodmarket_share	num
Dragstore_share	num
CarService_share	num
Airlines_share	num
Restorans_share	num
TaxiServie_share	num

Steps

k-means

- There is factor variable - exclude it
- Convert to matrix
- Check for correlation
- Standardize
- Determine start-number of clusters
- Build the model
- Visualize withins by Elbow-method



MODELING. Evolution. Case 2.3

Profile

CNT	clust	Weekly amount	Weekly trx_cnt	TRX amnt mean	TRX cnt_avg	Week unique locations	AVgAge	Days btwn transactions	Dragstore share	Airline share	Foodmarket share	Car Service share	Restorans share	TaxiService share
2290	1	223.86	30.45	119.03	60.96	3.59	45.72	15.42	0.24	0.05	0.33	0.09	0.04	0.11
1992	2	621.47	18.58	428.85	28.08	2.80	38.65	20.57	0.11	0.12	0.29	0.16	0.08	0.03
7336	3	346.68	44.00	73.95	223.26	4.71	40.22	12.24	0.15	0.08	0.32	0.11	0.04	0.11
3040	4	294.13	27.17	167.18	52.45	4.05	29.04	18.22	0.19	0.02	0.23	0.26	0.14	0.01
3108	5	281.41	16.21	202.32	23.28	2.72	44.97	22.16	0.18	0.02	0.35	0.12	0.06	0.09
4329	6	242.65	39.91	76.23	131.02	4.42	42.12	13.45	0.24	0.08	0.34	0.09	0.04	0.11
5413	7	399.10	36.54	189.57	82.70	4.34	41.67	14.19	0.18	0.09	0.28	0.13	0.06	0.03

Elder than others, **Drugstore-amateurs** low number of transactions, low transaction amount.

Highest transaction amount, highest share in airlines, young enough - **Trip amateurs**

High shae in Foodmarket, and taxiService - **massMarket average**

CarService, Restoran, the most younger group, comparatively low TRX amount_mean - **Golden Youth**

Foodmarket, low trx_cnt, low TRX amount, elder age - **Econom massMarket**

Evolution

Business task (with KPI determination):

Coss-sale: Client who has at least one closed consumer Credit likely to buy another consumer Credit. KPI - number of Credits sold

Population determination:

All clients who take a POS-credit and it is closed

Target determination:

Person who had at least one closed Consumer Credit and has one or more Credits after first Credit have been closed

Evaluation metrics selection:

Auc and Lift

Data preparation:

We have Socio-Demo, BKI, Credit Payments and DPD

Algorithm selection

The main limitation - easy model interpretation

Loyal client → deep sale, cross sale, up sale

Evolution

Target	Clients_Share%
0	90%
1	10%

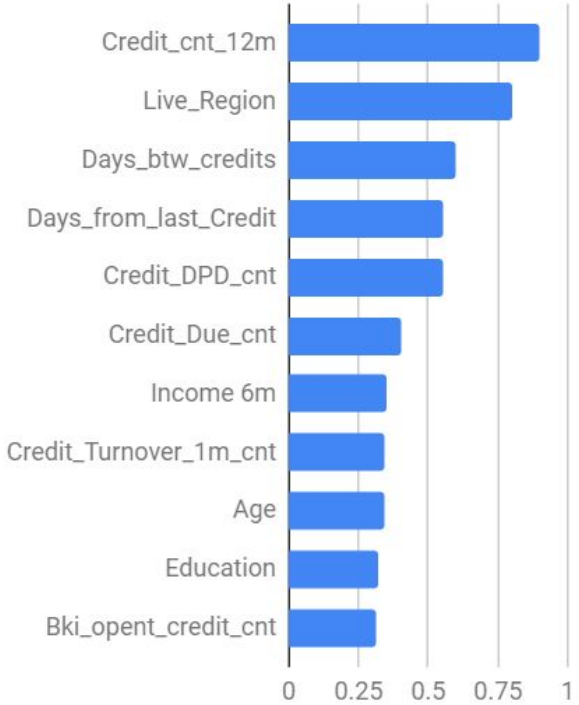
Classification problem:

1. **Decision tree**
2. Random forest (interpretation limits)
3. XGBoost (interpretation limits)

Evaluation:

Partition	Training	Testing
Auc	0.82	0.80
Lift	4.8	4.5

Importance



Profile

Factor	Not Likely to	Likely to
Credit_cnt_12m	More	Less
Days_btw_payments	More	Less
Days_from_last_Credit	More	Less
Credit_DPD_cnt	More	Less
Credit_Due_cnt	Less	More
Income 6m	Less	More
Credit_Turnover_1m_cnt	More	Less
Age	More	Less
Education	Less	More
Bki_opent_credit_cnt	More	Less

Retention

Business task (with KPI determination):

Find clients tended to churn.

Population determination:

All clients with Credit Card

Target determination:

Client who make one or more transaction in the month=(current month - 90 days) and does not make any transaction in the period between (current month - 90 days) and current month

Evaluation metrics selection:

Auc and Lift

Data preparation:

We have Socio-Demo, BKI, DPD, Credit Card transactions

Algorithm selection

The main limitation - easy model interpretation

Loyal client → retention

Retention

Target	Clients_Share%
0	82%
1	18%

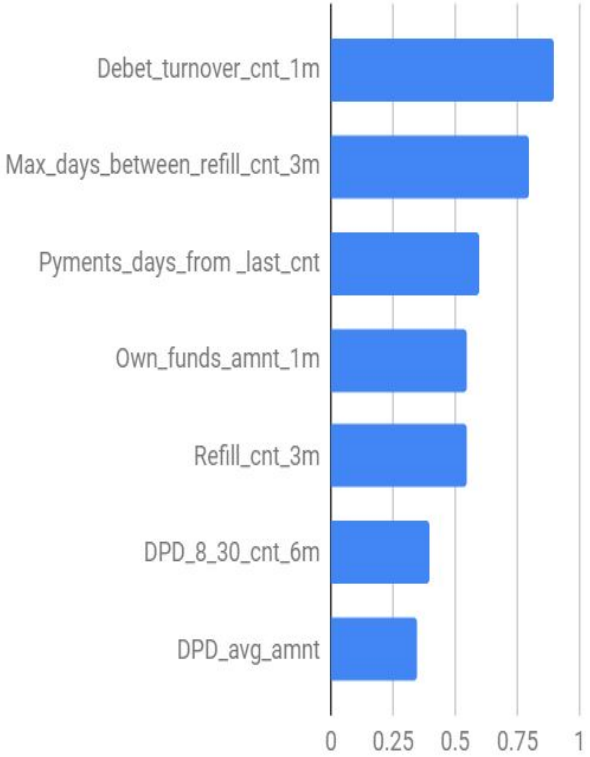
Classification problem:

1. **Decision tree**
2. Random forest (interpretati limits)
3. XGBoost (interpretation lim

Evaluation:

Partition	Training	Testing
Auc	0.82	0.81
Lift	4.7	4.5

Importance



Profile

Factor	Not Likely to	Likely to
Debet_turnover_cnt_1m	Less	More
Max_days_between_refill_cnt_3m	More	Less
Pyments_days_from _last_cnt	More	Less
Own_funds_amnt_1m	Less	More
Refill_cnt_3m	Less	More
DPD_8_30_cnt_6m	Less	More
DPD_avg_amnt	Less	More

Questions?