

GRM: Credit Sciences Business Case

---- Bank Accounts Prediction

By Peter Chen

Case Overview

- I am provided with more than 1 year of customer product data. The objective is to predict what accounts a customer will acquire in the next month, 2016-05.

Table 1. Demographic Data

cust_id	country	sex	age	gross_income	...
12345	1	1	21	10000	
12346	2	0	45	85000	
12347	1	1	31	65000	
12348	3	1	11	1000	

Table 2. Customer Product Data

Cust_id	date	savings_account	mortgage	e_account	tax_account	...
12345	2016-01	0	0	0	1	
12345	2016-02	0	0	0	1	
12345	2016-03	1	0	0	1	
12345	2016-04	1	0	0	1	

Data size: 3G data files, 931,453 clients, 23 products, 12,715,856 rows, 47 columns

Characteristics

- **Multi-labels**

There are 23 products(labels). One client can have several labels in the same time.

Solution: Train 23 models for each of the labels.

- **Time series**

Date covers from 2015-01 to 2016-04.

Solution: Features of adding date

Avoid information leakage

- **Unbalanced Data**

The maximum positive rate of label is 60.522%.

The minimum positive rate of label is 0.002%.

Solution: Adjust 'class_weight' when training the model

Table 4. Percentage of Owning rate in 2016-04

Account_type	Percentage of Owning
current_accounts	60.522%
direct_debt	12.038%
particular_account	10.885%
e_account	7.943%
payroll_account	7.730%
payroll	4.971%
taxes	4.877%
credit_card	3.750%
particular_plus_account	3.606%
long_term_deposits	3.486%
securities	2.304%
funds	1.596%
more_particular_account	0.932%
junior_account	0.819%
pensions	0.793%
mortgage	0.490%
home_account	0.319%
loans	0.214%
medium_term_deposits	0.111%
short_term_deposits	0.035%
derived_account	0.034%
savings_account	0.008%
guarantees	0.002%


Target definition

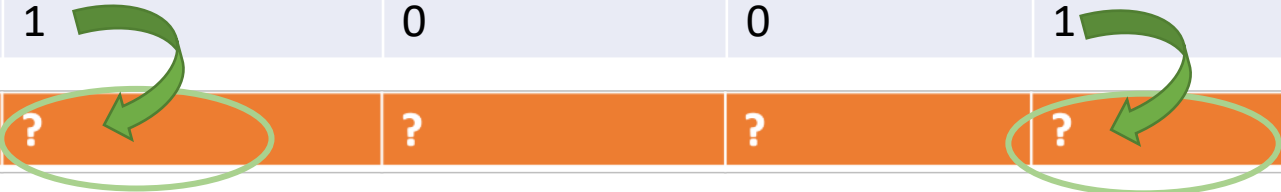
- What accounts a customer will **acquire** in the next month?

Table 3. Submmission table

cust_id	added_products
13344	savings_account
67821	mortgage; e_account
69405	payroll_account; pensions
...	...

Table 2. Customer Product Data

Cust_id	date	savings_account	mortgage	e_account	tax_account
12345	2016-01	0	0	0	1
12345	2016-02	0	0	0	1
12345	2016-03	1	0	0	1
12345	2016-04	1	0	0	1
					
12345	2016-05	?	?	?	?



If a customer opened the account before, the account status of next month will be 1.
Therefore, we only predict mortgage and e_account for this client.

- What accounts a customer will **acquire** in the next month?

-- Supervised Classification Problem

-- Definition of Y

If transition happened ($0 \rightarrow 1$): $Y = 1$

If NO transition happened ($0 \rightarrow 0$): $Y = 0$

If NO transition happened ($1 \rightarrow 1$): throw it away because we don't need to predict it

Note: There is no cancellation ($1 \rightarrow 0$) in the dataset

Cust_id	date	savings_account
A	2015-01	0
A	2015-02	0
A	2015-03	1
A	2015-04	1
B	2015-01	0
B	2015-02	1
B	2015-03	1
B	2015-04	1



Define Y

Table 4. Y label

Cust_id	date	savings_account_Y
A	2015-01	0
A	2015-02	0
A	2015-03	1
B	2015-01	0
B	2015-02	1

Account Features

Generate 23 features
(tax_first, mortgage_first, etc)



Cust_id	date	savings
A	2015-01	0
A	2015-02	1
A	2015-03	1
A	2015-04	1
B	2015-01	0
B	2015-02	0
B	2015-03	0
B	2015-04	0

Extract
'adding date'

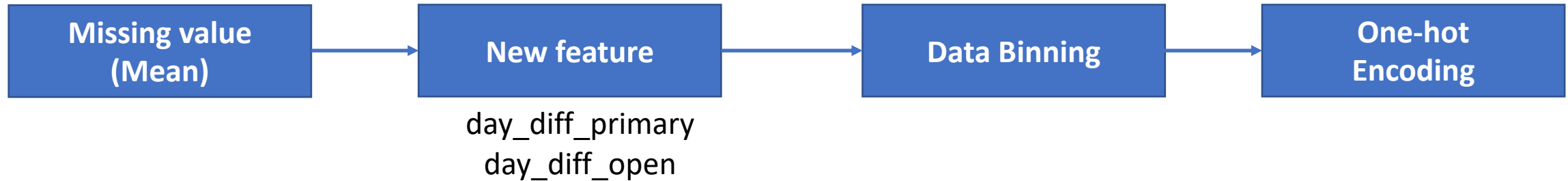


Generate

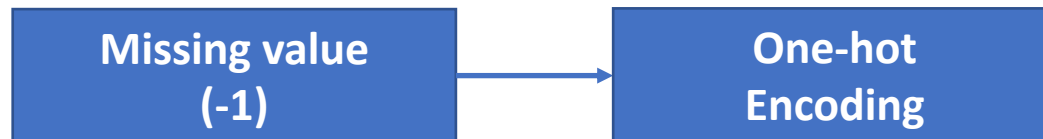
Cust_id	savings_add_date	date	savings_first	savings_add_Y
A	2015-02	2015-01	0	0
A	2015-02	2015-02	0	1
A	2015-02	2015-03	1	0
A	2015-02	2015-04	2	0
B	Null	2015-01	0	0
B	Null	2015-02	0	0
B	Null	2015-03	0	0
B	Null	2015-04	0	0

Demographic Features

-- Numerical Variables: (Age, seniority, gross_income, etc)



-- Categorical Variables: (Country, Sex, channel , etc)



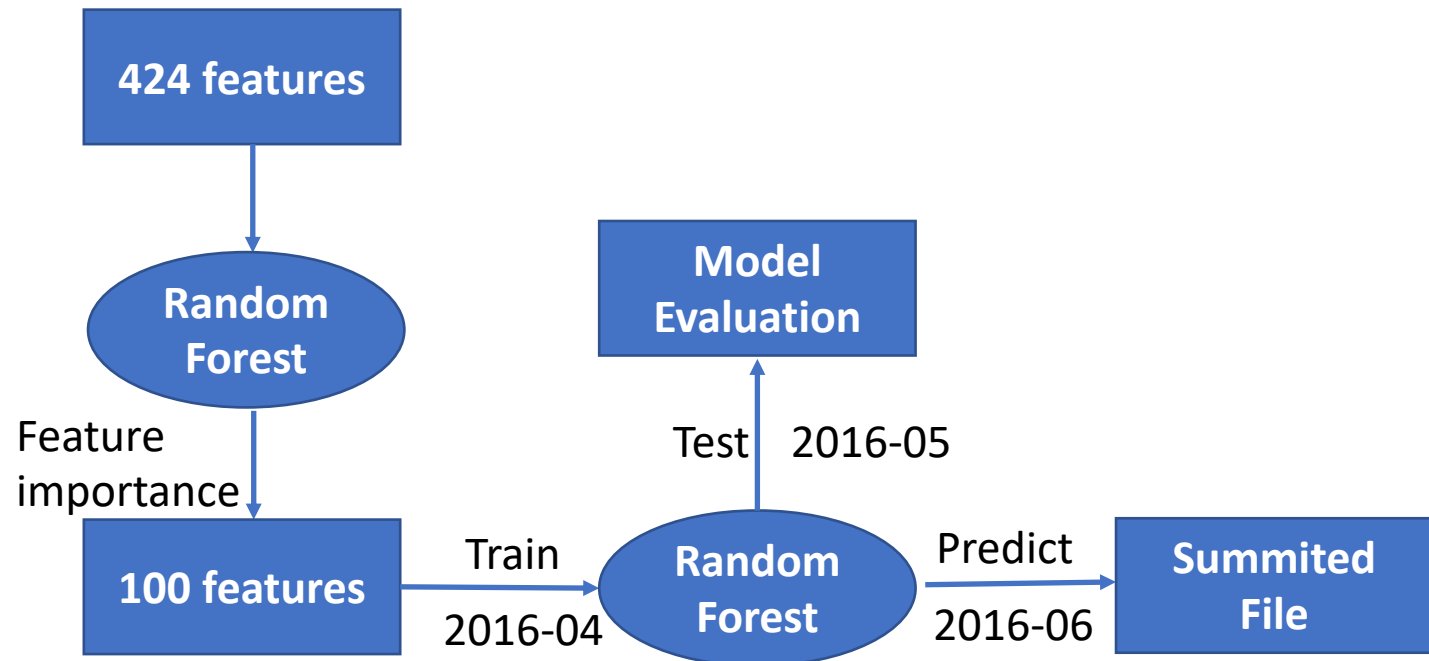
Feature Selection & Modelling

We have 424 features.

- 23 account features (How many months ago did clients open the account)
- 15 binning dummied features from numerical variables
- 386 dummied features from categorical variables

Random Forest

- Good performance
- Feature importance
- Unbalanced data



Model evaluation

F1 Score

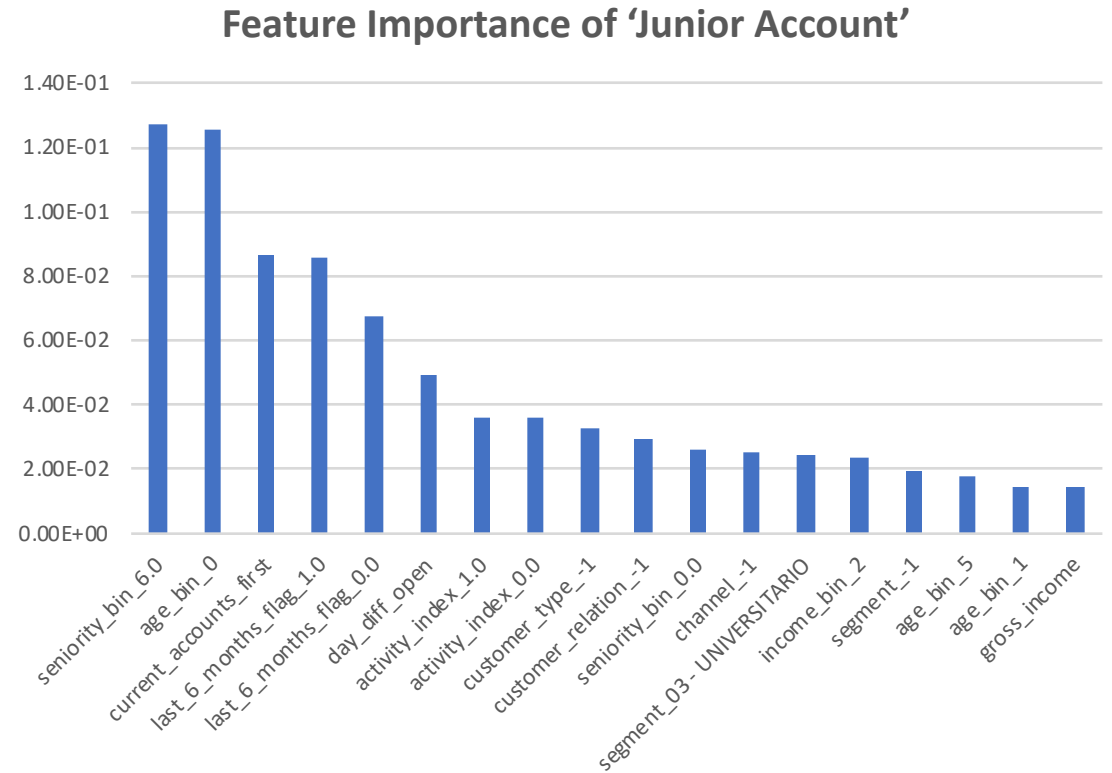
Product_Name	Test_size	Actual_1	recall	precision
more_particular_account	912,716	814	96.31%	4.71%
e_account	852,247	1,921	93.81%	0.85%
credit_card	871,258	598	92.64%	0.26%
payroll	861,506	1,476	91.40%	0.76%
payroll_account	846,820	1,448	89.71%	0.89%
direct_debt	778,566	2,749	89.52%	1.48%
current_accounts	322,484	3,416	88.93%	30.01%
junior_account	919,836	67	85.07%	17.98%
long_term_deposits	881,569	34	76.47%	0.03%
securities	904,534	80	62.50%	0.03%
taxes	883,802	809	60.32%	0.45%
particular_account	821,514	72	55.56%	0.04%
funds	911,190	69	49.28%	0.05%
particular_plus_account	891,604	40	20.00%	0.04%
pensions	920,294	22	4.55%	0.02%
savings_account	928,188	0	0.00%	0.00%
guarantees	928,251	0	0.00%	0.00%
derived_account	927,873	3	0.00%	0.00%
short_term_deposits	922,312	0	0.00%	0.00%
medium_term_deposits	926,446	0	0.00%	0.00%
mortgage	923,259	5	0.00%	0.00%
loans	926,110	3	0.00%	0.00%
home_account	925,046	2	0.00%	0.00%

Uncommon

Summary

Advantage

- Transfer a time-series problem into classification problem.
- Save the computation resource.
- Have Interpretability.
- Each product has a unique model.
Easy to maintain.



Summary

Advantage

- Transfer a time-series problem into classification problem.
- Save the computation resource.
- Have Interpretability.
- Each product has a unique model.
Easy to maintain.

Improvement

- Hyper-parameter tuning
- More feature engineering
- Model Stacking.
- Other data source

Application

Business

- Marketing & Promotion
- Improve customer experience
(Cold start, Call center, etc)

Strategy

- High precision labels: cold call to improve the response rate of clients.
- Low precision labels:, send email or push app notification to save cost of marketing.
- A/B Test