

# Banking Service Recommendation System

Designing tailor-made  
recommendations

---

STEFAN MEMMER (JANUARY 2020)

SPRINGBOARD DATA SCIENCE CAREER TRACK

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# Executive Summary

---

Digitization changes banking services as well

Keeping and attracting customers are a major challenge against the background of cost pressure and low interest rates

Automated interaction with clients and tailor made product recommendations are a key in today's banking experience

Based on ensemble learning pairing of clients to specific products can be achieved

This represents the foundation for automated responses to clients when interacting with the bank and should spur success of cross selling and customer targeting


# First steps on Data

---

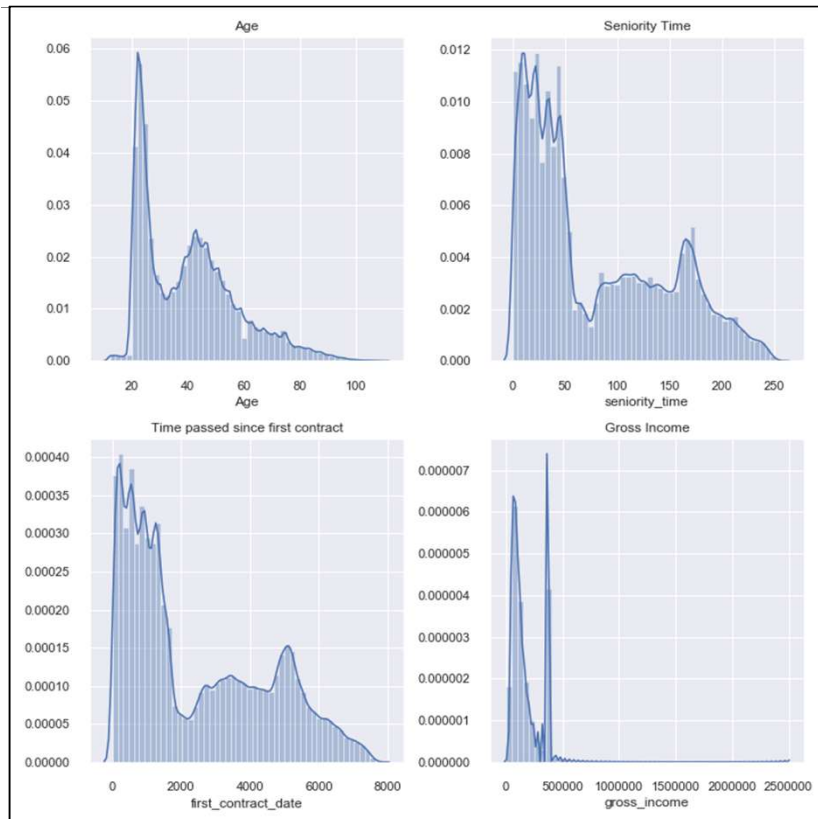
Data is obtained from a kaggle challenge sponsored by the spanish Bank Santander

Set contains purchased products and features of ~900,000 clients in chronological order over 1.5 years

Necessary cleaning steps:

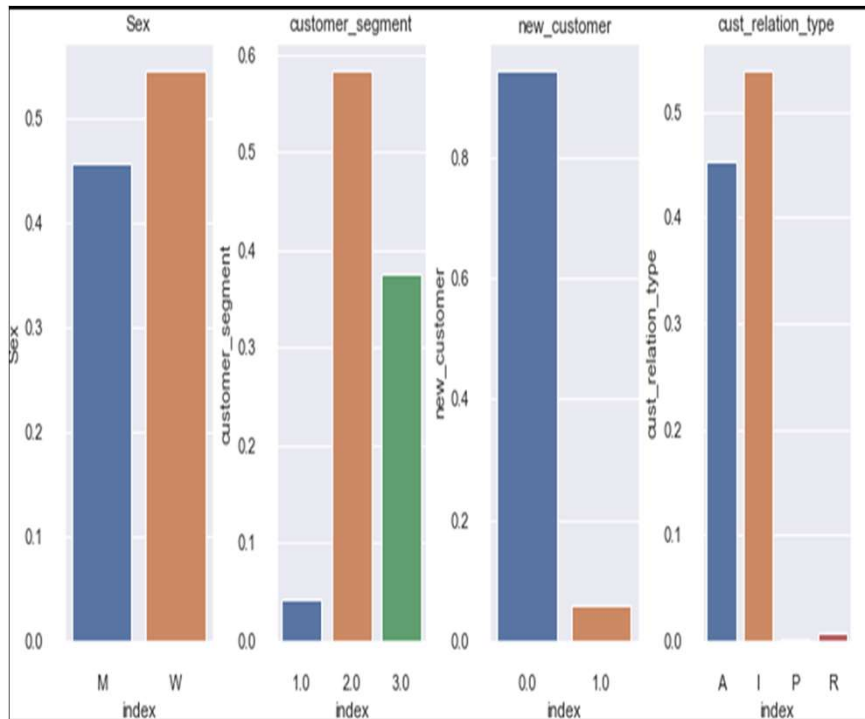
- Translation of feature names to English
  - Renaming labels consistently and tidying up of erroneously mixed data types
  - Checking outliers and possible systematic dependencies of missing values
  - Removing features showing only small variance or do not provide informational content
  - Replacing missing by cluster analysis and look-up tables
- 
- A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# Exploratory Analysis: the clients I



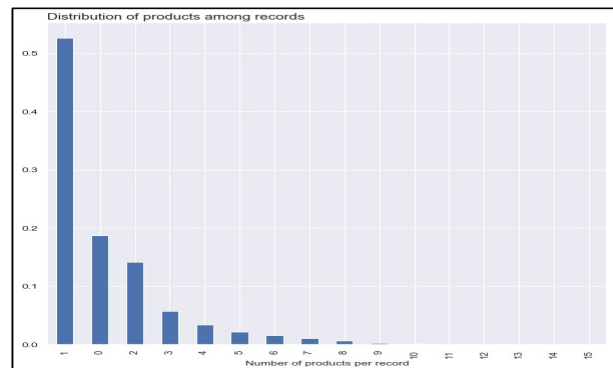
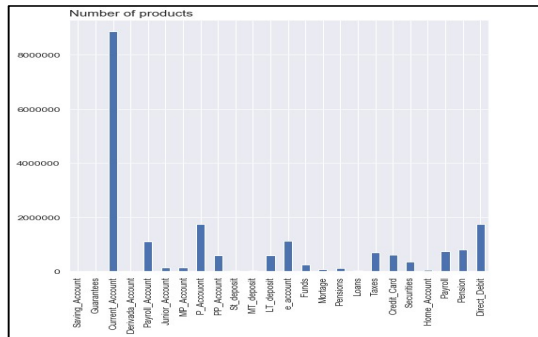
- Bimodal distributions hint at clustering
- Many clients are youngsters (Age) – likely to be attracted by automated solutions
- Strong outliers for income and wealth (gross income) – belong possibly to premium segment of clients
- Most have joined the bank or take services within the last 5 years (Seniority Time in month & Time passed since first contract in days)

# Exploratory Analysis: the clients II



- Slightly more females than males (Sex)
- Most customers belong to baseline segment of individuals (index 2) or college graduates (index 3) - not many VIPs
- Only small percentage joined the bank within the last 6 months (new\_customer index 1)
- Majority of customer relations classified as Inactive (index I) or Active (A); former (index P) or potential clients (index R) classification rather small

# Exploratory Analysis: Product basket



- Current Accounts, Direct Debit, Particular and electronic Accounts are most frequent services clients have purchased
- Overly, in the set clients are holding one or two products, while there are also records which use no service at all
- The number of customers holding more than 8 products is comparably small

# Clustering the client base

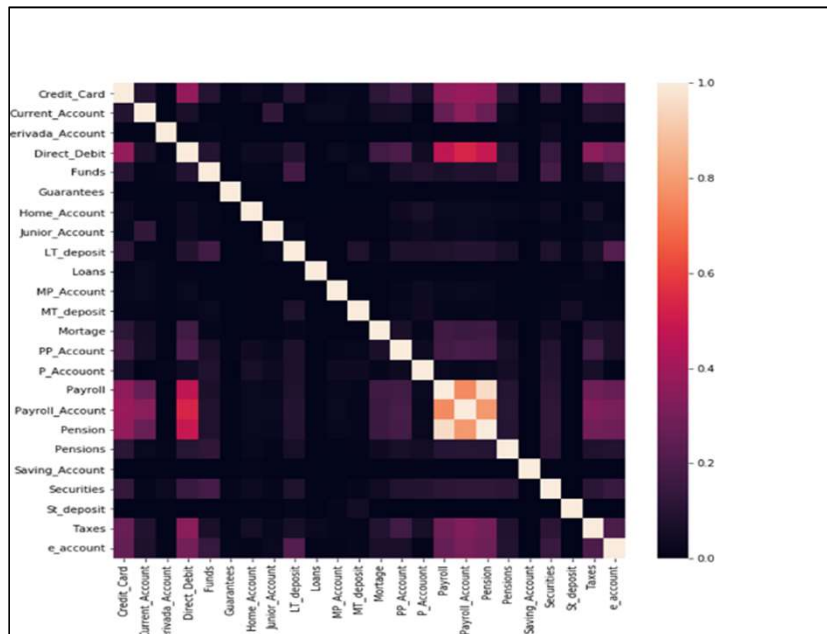
---

Approach: Non Negative Matrix Factorization and subsequent K-Means clustering group customers in four distinct segments:

- **Group 1 – older but active:** older ones with mixed income - belong to the baseline segment of individuals, but cluster with largest part of so-called VIPs
- **Group 2 – inactive college graduates:** overly regarded as inactive with many college graduates - younger group with mixed income figures
- **Group 3 – high potentials:** above average income, joined rather recently, are considered as very active – quite young
- **Group 4 – regulars:** mostly inactive, but are clients quite for a while with mixed income and age

# Which products are frequently bought together?

Correlations among product occurrences shows some relations:



- between pairs “Direct Debit” and “Payroll” products as well as to Pension services
- these correspond well to occurrences of Pension accounts
- between e-accounts, mortgages and particular accounts seems to be an association as well



# First advices to the bank's sales team

---

A special group of clients (cluster 3) with high potential and high activity:

-> this cluster should be easily accessible sufficiently by passive marketing strategies

Segment of "seniors" which is quite active (cluster 1)

-> might require more "tailor-ship" product recommendations

Clients of group 2 (inactive college graduates ) and group 4 (regulars) are possibly at risk to leave the bank

-> keep relationship for longer but are considered as inactive; thus more aggressive marketing strategies seem to be required

With regard to the product basket:


- clients who have Payroll accounts might get offered Direct Debit and Credit Cards and vice versa
- Payroll, Direct Debit and Pension accounts show some considerable dependences which are exploitable
- Clients who have Particular Accounts in their basket could have desire for tax and pension accounts

# Diving into the prediction problem

---

First actionable insights with clustering and dependence analysis

However, for a recommendation engine some key problems are getting solved :

- Multilabel task – out of 24 services customer can buy more than one product the month ahead
  - High dimensionality of dataset, i.e. 18 descriptive features and combinations out of 24 products in each clients basket
  - Correlation analysis do not suggest any remarkable dependency between single features and target labels
- 
- A solid orange horizontal bar spanning the width of the slide, located at the bottom.

# Going Ahead: Feature Engineering

---

Introduction of two new features:

- Leave ones: number of products a client un-subscribed since last month
- New ones: number of services purchased with respect to last month

Feature Target Encoding:

- Most features are categorical and cannot be handled by Machine Learning Algorithms
- Establishing a relation between categorical features and targets by assigning target frequencies to labels
- Encodes categorical to continuous features

Likely, for any specific product only a subset of features for a specific model could be useful:

- Feature selection approach is necessary
- Training and evaluating a unique model for each product

# The Pipeline

---

The whole data set is randomly split in a 40 % hold-out and 60 % training set

Training Set : Feature selection and parameter grid search with Random Forest Models (RF)

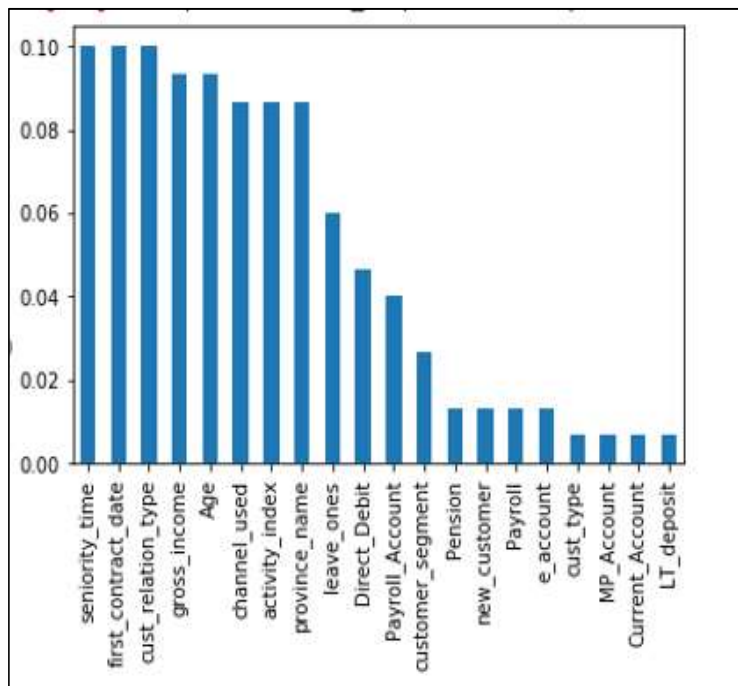
Assessing variable importance:

- Subsampling of training set in Kfolds
- Subsampling each fold in equally balanced sets
- Target Encoding of categorical variables
- Fitting RF
- Assessing Feature Importance (FI)
- Repeating steps b to e and averaging Feature Importance over Kfolds
- Features with large importance that sum up to 80 % of FI are selected

Model Training:

- Grid search of parameter space: number of estimators and max depth
- 10 Kfolding training set
- Reducing samples to more balanced target relations
- Target Encoding of categorical variables
- Fitting models with feature subsample from step 3
- Evaluating model by area under ROC, respectively, weighted F1 score
- Selecting parameters of the model

# Important Features:



Considering feature selections for all models most useful are:

- client history (seniority\_time & first\_contract\_date)
- the relation to the client (cust\_relation)
- Age
- wealth defined by income (gross\_income)
- level of activity (activity\_index)
- provenience (province\_name),
- the way a client joins the bank (channel\_used)
- number of product that are terminated with respect to last month (leave ones)

# Results:

Names	mean_test_ score	std_tests core	AUC_out _of_bag	upper_ band	lower_ band
Credit_Card_target	0.94	0.0	0.92	0.94	0.93
Current_Account_target	0.89	0.0	0.87	0.9	0.89
Direct_Debit_target	0.9	0.0	0.88	0.91	0.9
e_account_target	0.92	0.0	0.91	0.93	0.92
Funds_target	0.9	0.01	0.89	0.91	0.89
LT_deposit_target	0.88	0.0	0.86	0.89	0.87
MP_Account_target	0.96	0.0	0.97	0.97	0.96
Payroll_Account_target	0.92	0.0	0.89	0.93	0.92
Payroll_target	0.95	0.0	0.94	0.95	0.95
Pension_target	0.96	0.0	0.94	0.96	0.95
PP_Account_target	0.96	0.0	0.95	0.97	0.95
P_Accouont_target	0.96	0.0	0.94	0.96	0.95
Securities_target	0.89	0.01	0.88	0.9	0.88
St_deposit_target	0.98	0.0	0.98	0.99	0.98
Taxes_target	0.88	0.01	0.87	0.9	0.87

Measure: Area under the ROC curve

- Out-of-bag: achieved after applying the models on the test set
- Mean\_test & std\_test: values of validations from Kfold testing
- Lower\_ & upper\_band:  $\text{Mean\_test} \pm 2 * \text{std\_test}$
- Overall, the area under ROC of the test set is quite in-line with results of the validation sets in the cross validation
- Generally, the numbers indicate strong predictive power of the respective models.

# Discussion

---

- To predict the purchase for any single product a specific model has been developed relying on different feature combinations
- RF models provide predictions of probability
- According to these probabilities for each client products can be ranked rendering a „table of preferences“
- Upon this ranking the tailor made product recommendations and marketing initiatives can be designed

# Further steps and critiques

---

- In spite of compelling performance figures, approach is computationally expensive – might be problematic for fast online recommendations
- Cold start problem: most models rely on information from product basket as well – problems to get recommendations for completely new clients
- Common approaches for recommendation systems like collaborative filtering could provide additional insights
- For 8 targets the frequencies are too low to solve it with this approach, thus further techniques for strongly imbalanced datasets need to be tested