Capstone Project Milestone Report: Banking Product Recommendation

Problem definition

The project's focus is a business problem many financial institutions are currently facing. The traditional or as it is sometimes called the "bread and butter" business of "universal" banks is still built upon revenues from the corporate and retail segment. Thus, earnings generated from credit contracts and the selling of services remains a key stone. For the time being, many of the larger wholesale institutions throughout Europe face severe competition in these fields. Such competitive environment emerges, on the one hand, from smaller FinTech's and start-ups offering comparatively cheap online banking solutions. On the other hand, banking institutions suffer strong cost pressure, forcing them to reduce overhead costs. However, these changes are providing opportunities as well. Banking institutions do have still a large customer base mainly because of strong reputation that lasts for decades. Hence, they might be able to hold and even enlarge their client base by tailor made product and service recommendations. In a survey, 70 % of banking customers desired to receive automated and personalized. Moreover, especially younger clients who are more open to payment services provided by Google, Apple or Facebook (Gafa model) might get attracted again by highly responsive tailormade services (BEYOND DIGITAL: HOW CAN BANKS MEET CUSTOMER DEMANDS? 2017 Accenture).

Against this background the project's outcome is a recommendation engine which helps banks to identify appropriate products for specific customers or group of customers, respectively. From a business perspective, the outcome should enable banking institutions to a) increase their revenue stream by cross selling specific products, i.e. offering products to existing clients which in turn b) should lead to maintain and to intensify customer relationships.

Possible clients who could benefit from the outcome of this project are mainly financial institutions who want to improve customer relationships. Hence, a precise identification of likely product customer combinations might help to frame marketing initiatives which is key in customer targeting. Basically, unspecific marketing campaigns by recommending wrong products or by expensive and excessive advertising initiatives might lead to unwanted customer retention and could even lead them to leave the bank.

Data Wrangling – key steps

The dataset is from Bank Santander, a Spanish universal bank, and can be downloaded from https://www.kaggle.com/c/santander-product-recommendation. The dimension of the set is approximately 13 Mio rows and 48 columns. The columns have variables describing the client, for instance, her age, residence, sex, and the respective products in sparse matrix form. The dataset shows a chronological order over a time span of 1.5 years, and thus, contains readings of approximately 940,000 unique customers.

Some client features contain erroneous and missing entries, and thus, have to be cleaned. The sparse matrix part comprises 24 columns with each column representing a product. The target, the machine learning technique should be trained to predict, is the product the customer is going to buy the month ahead. So, the target variable is an index of the additionally purchased product. Therefore, in this dataset the target must be constructed implicitly by comparing the products of a customer in month (t) to the products in month (t+1).

After defining target variables various cleaning steps are necessary:

- **Cleaning features**

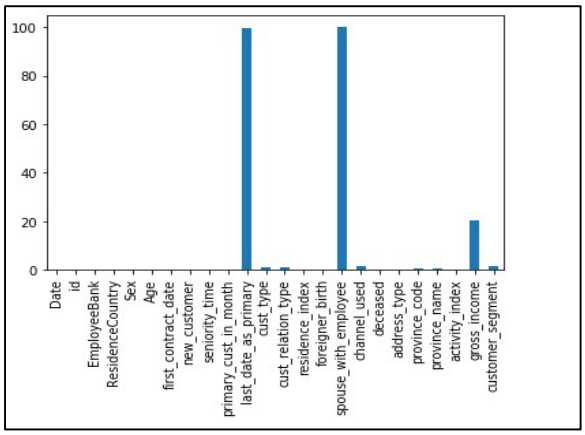
The dataset contains labels and feature names in Spanish; thus, translation and relabelling have been conducted. After assessing the targets in a sparse matrix format and translation to English, the datatypes of the respective columns have been controlled. Many features show mixed datatypes, as a result of mixing up float values with strings. An appropriate algorithm is applied to clean these datasets and to transform them to continuous types. Other features, mainly categorical ones, have inconsistent labels which are hard to understand logically. Accordingly, they have been replaced by more consistent and comprehensive labels. Features, which had integer values but are intended to represent categories are converted to categorical datatypes to enhance computational efficiency. In addition, one feature which comprises the dates, of when a customer has become a client, gets transformed to represent the time difference to respective records' dates to enhance informational content.

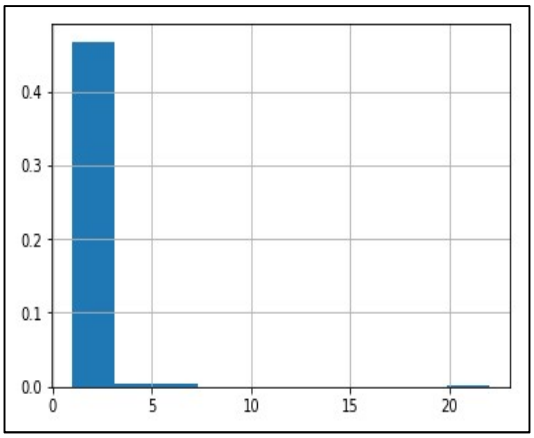
- **Outlier detection**

Based on the fact that most features are categorical, an outlier is defined as percentage of occurrence. Thus, if a categorical label does not occur more often than 2 % in this feature it is marked as a possible outlier. For continuous features, standard deviations from the mean are considered. Hence, outlier records of the features "Age" and "gross_income" are marked in case of values departing more than three standard deviations from their respective mean. An inspection of possible outlier values in categorical features is not providing a real clue, whether these values are the consequence of typos during data collection or if they are a real print. For instance, if customers' proveniences from a certain region are only about one percent, one cannot conclude this to represent a mistake. However, for continuous features a couple of outliers are detected that are dispersed widely from the feature's average. Nevertheless, considering the shape of these features' respective dispersion an additional step has been taken into account. Therefore, values of "Age" higher than 110 and lower than 12 have been replaced as missing values. Basically, if children are allowed to have accounts the control of these accounts is likely to be legally restricted to adults. Thus, in this case, age does not provide proper information. Moreover, "gross_income" which reflects the household's income of the customer has entries above 1.5 million Euro which is quite unlikely to occur that often. Therefore, beside upside deviations from the average, such values have been marked as missing values as well.

- **Completeness of the dataset**

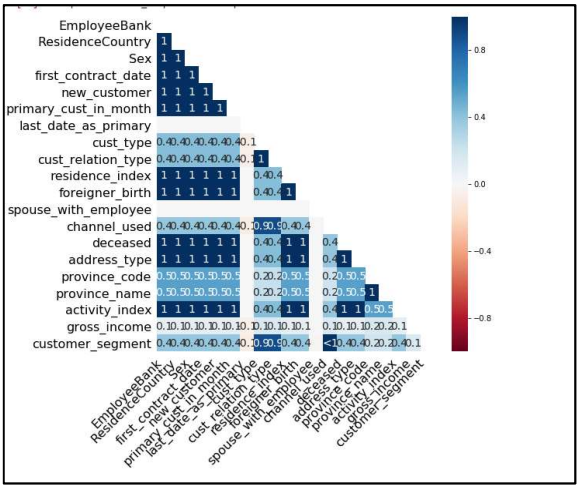
In this step, missing values are replaced and features which do not contain much informational content in terms of missing variance or arbitrary content have been removed. The following graphs show the occurrence of missing values (NAs) only containing the matrix with clients' features: 1) per feature (left) and 2) as frequency per row (right).





Obviously, the features “last_date_as_primary” and “spouse_with_employee” have overly missing values, whereas record-wise most have about 2 NAs and only a few are covered completely by missing values, i.e. values of 20. The feature “gross_income” and “customer segment” have also a remarkable percentage of missing values.

In order to detect some kind of patterns among the missing values of different features a heatmap is shown below:



On the heatmap above, contrary to the common understanding, red colour contour hints at dependence between NA occurrence of feature X to feature Y. Accordingly, not systematic dependence of missing values is being detected. Regarding the sparse product basket, two features have missing values. In the following, these values are being cleaned as well.

- **Tidying up**

The feature “deceased” marks whether a customer is leaving the bank or has died. Since with deceasing accounts one cannot predict any future change, all records with positive “deceased” entries are removed, then the whole column has been dropped from the set

Because of the time chronological order of the dataset, a customer has many entries in the dataset. Hence, some features are expected not to change, i.e. like “Sex”. Thus, a look-up procedure is applied to find replacements for missing values among records of the same customer. For instance, missing values in gender or Age are not likely to change at all or over a couple of months.

Records which show missing entries in the product basket are removed completely. The features “spouse_with_employee” and “last_date_as_primary” which have entirely missing values are dropped. The features “address_type” and “province_code” are removed as well. These features do not provide any important content at all, i.e. province_code is a mirror of the feature “province_name”.

For the continuous features “Age” and “gross_income” a replacement procedure is designed. This takes other features which show appropriate variance and constructs clusters. According to the cluster to which a missing value record belongs, the clusters’ median of the feature in question is taken as a replacement value. For categorical features the respective clusters’ most common value has been taken.

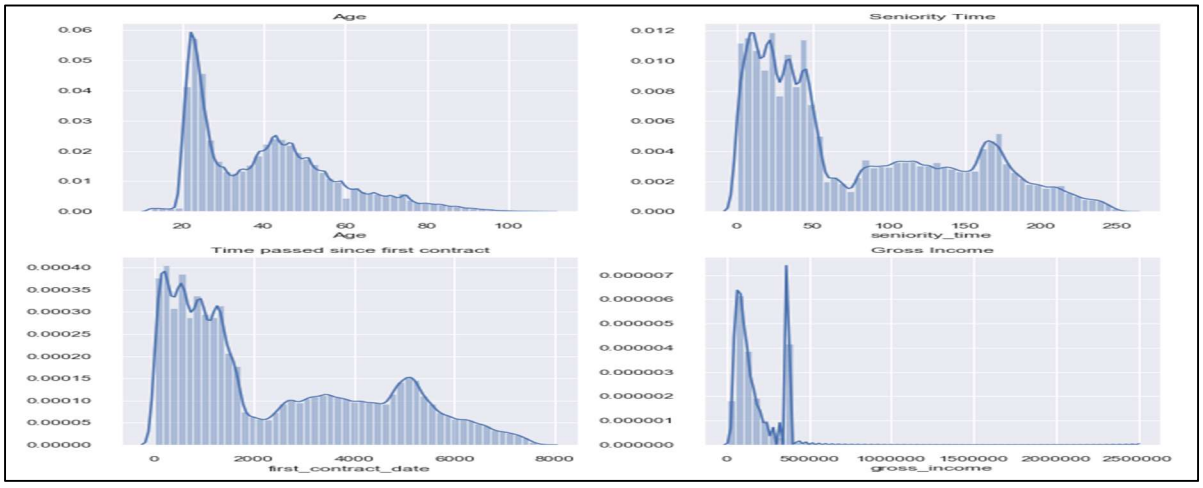
Finally, two new features are being created. Following the first step, where the target variables have been calculated, the numbers of products that have been purchased and recalled with respect to the recent month are assessed. Thus, for customer i at datetime T the feature “new_ones” shows the number of products the customer i has purchased in comparison to datetime T-1. The feature “leave_ones” shows instead, the number of products which have been cancelled from this customer’s product basket from datetime T-1 to datetime T. Possibly, these two features might provide valuable features for an appropriate machine-learning algorithm.

A journey through various features

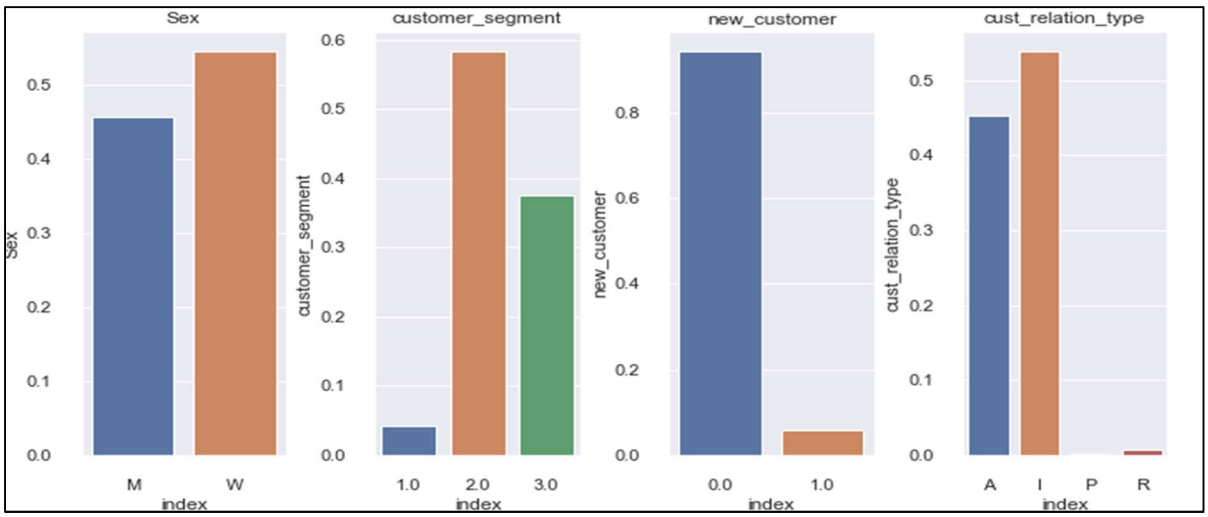
Exploratory data analysis reveals some insights about the bank’s customers. A more extended data visualization is provided in the Appendix showing all features with respective distributions and changes over time. However, here only a small subset of obviously more important features is presented.

- **Features describing the clients**

All continuous features show a bimodal distribution with peaks either in the lower and middle (Age, gross income) or upper range (Seniority Time and first contract date). Thus, there is possibly some kind of clustering among the clients: one group of youngsters having less income and another group with higher Age and possibly better income numbers.



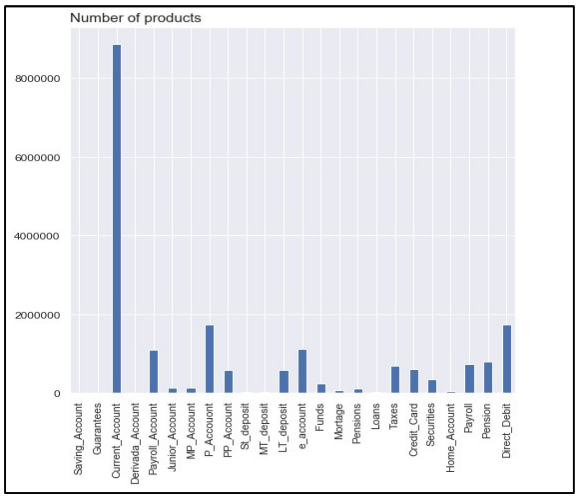
Regarding categorical features which seem to be important and show enough variation: gender, novelty of clients, client segmentation on part of the bank and various client-bank relation types are plotted below.

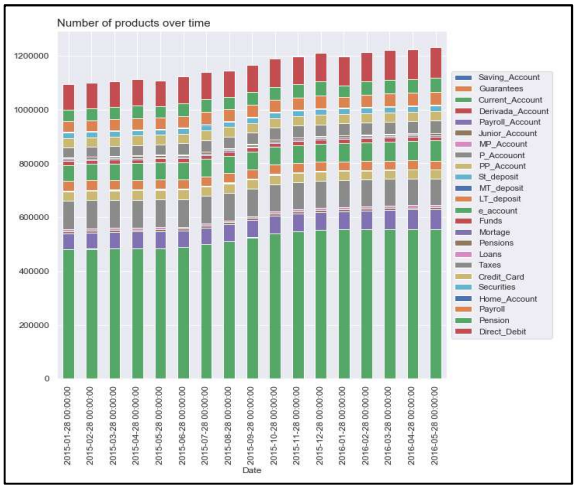


Actually, slightly more females (W) are in the client base of the bank. Most customers belong to the segment “individuals” (index 2) or are college graduates (index 3), while the “VIP” (index 1) segment is comparably small. Overly, there are “old” customers, i.e. which have not signed up within the last six months (new_customer. Though, many clients are considered to have either an active relationship (index A) or inactive relationship (I). The percentage of former (index P) clients and potential clients (index R) who are considered to make purchases potentially is comparatively small.

- **Product basket**

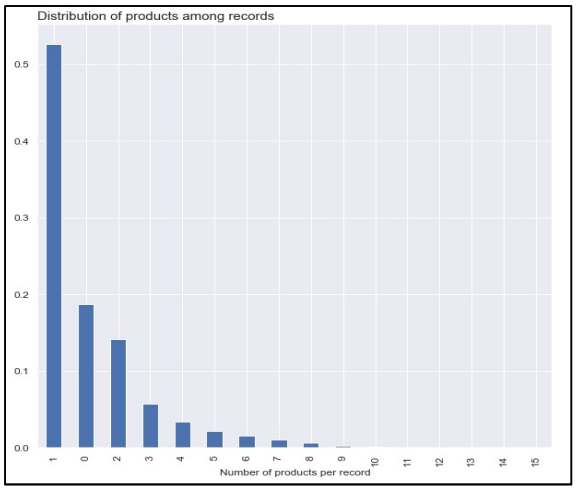
Over the whole-time span “Current_Accounts” are the most frequent services clients required, followed by “Direct_Debit”, “P_Accounts” (particular/special accounts) and “e_accounts” as it gets depicted by the plot below.





Comparing the number of the different products against time, there is not a clear decisive pattern hinting at seasonal dependencies. Rather there is a slight upside drift of overall frequencies in the last quarter of 2015.

Mostly, clients are holding only one or not any service in their basket, while there are very few records showing subscriptions of more than 8 products (lower plot).



Clusters: A more comprehensive view

The previous analysis reveals some information about the characteristics of clients and the product basket. Moreover, especially, continuous features indicate the presence of groupings among the bank's clients. Therefore, a cluster analysis is conducted. On the one hand, such dimensionality reduction might provide a better overview against the background of such a large feature space, on the other hand, grouping clients to clusters corresponds well to common approaches of recommendation systems.

Clustering is done separately, for features describing the client base of the bank and for the product basket. This approach is suitable, because descriptive features belong to specific clients which of course appear many times in the dataset, i.e. clients are showing up over a period of subsequent months. So, for this analysis of the dataset is reduced in order to have unique client records. Then categorical features are hot-encoded and continuous ones are binarized through quintiles of their respective distribution. A non-negative matrix factorization is applied to the resulting sparse matrix to reduce the dataset to four major components. However, more components would provide a better approximation but the additional gain in explained variance starts decreasing after four components. Thus, for the sake of having a less complex overview, four components are selected. In a final step, the "loadings" of the four components are grouped by K-Means clustering. This is necessary for a better clustering which would not be possible due to ambiguous component loadings.

As a result, with regard to the features whose labels show the most variation, the four groups are easily characterized:

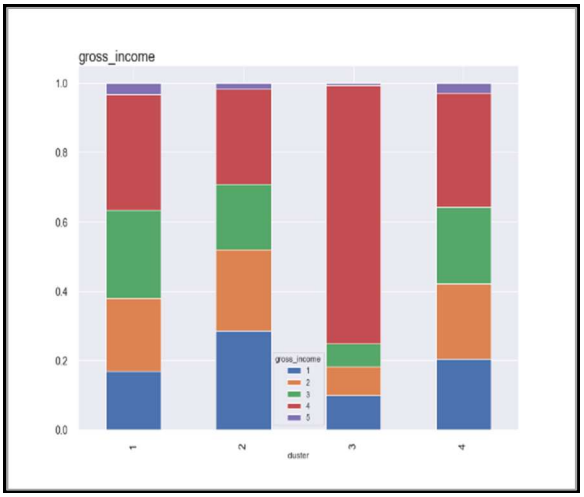
Group 1: Rather older customers whose income is mixed. This group shows also mixed duration of client history, but they are considered as very active. Most belong to the baseline segment of individuals, though this cluster has the largest part of so-called VIPs.

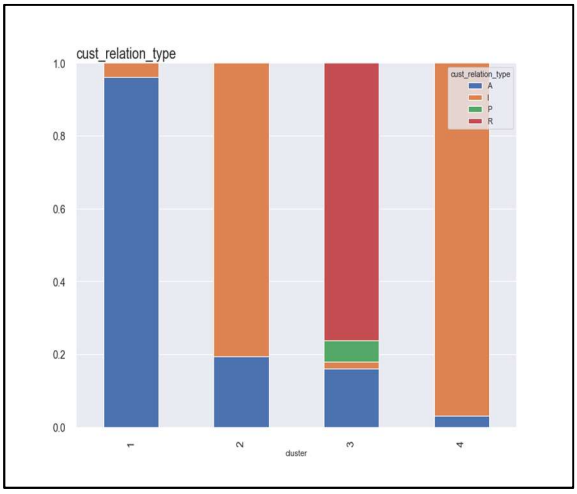
Group 2: This cluster is overly regarded as inactive by the bank and most clients have graduated from college. They are younger and most of them show an average time being a client to this bank, moreover, income is mixed.

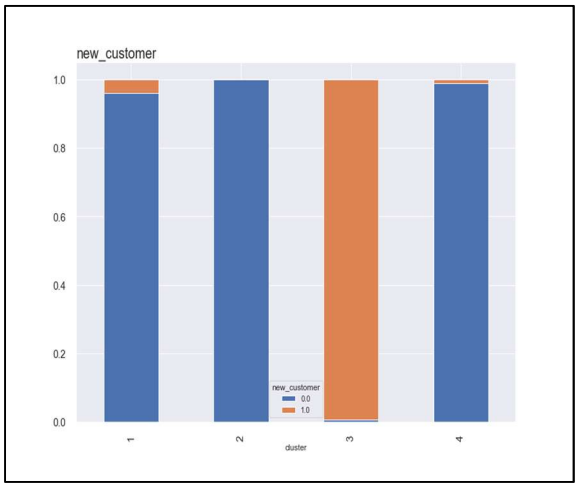
Group 3: These are the clients with the highest potential for future business according to the bank's segmentation, in addition, their income is above average. Most of them have graduated from college and have opened accounts or became clients recently. Furthermore, they are quite active and are younger on average.

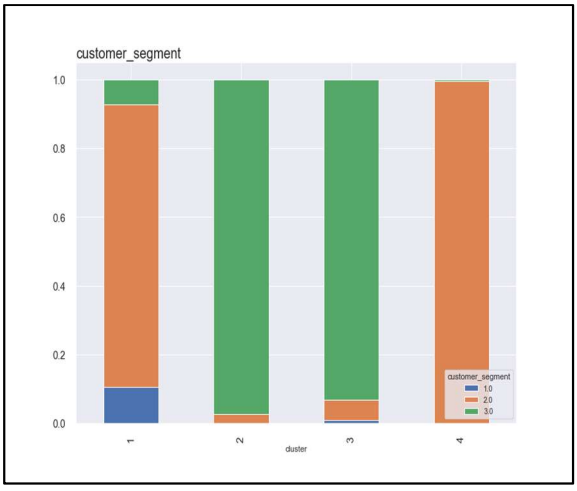
Group 4: Customers of this cluster have relationships to the bank for quite a while and are mostly considered as inactive. However, income and age are mixed in this group.

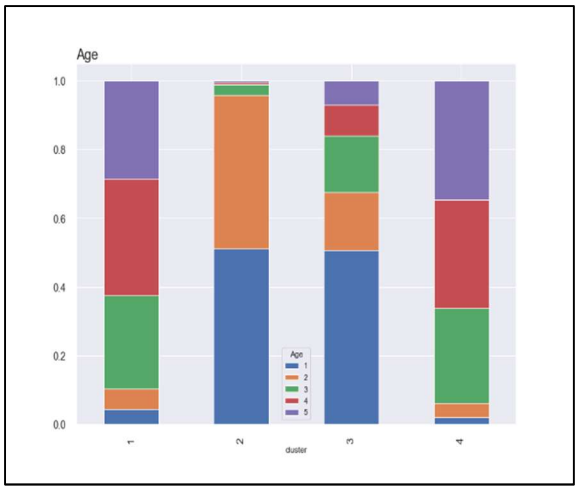
The charts below show frequencies of some key features for each cluster:



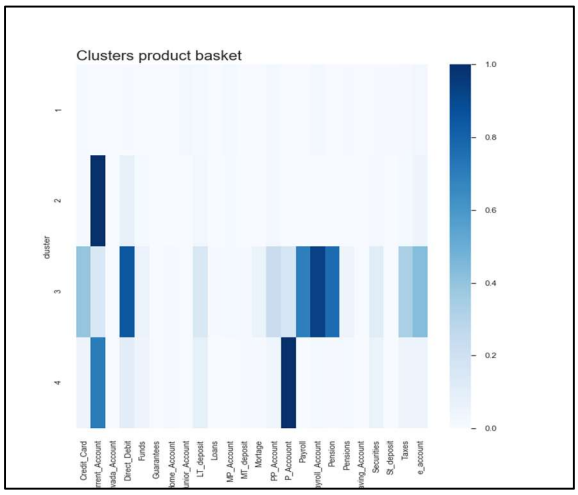








In order to segment the product basket, the same clustering approach is applied as it is done for the descriptive features, except that the whole dataset is taken into account. The chart below shows the four clusters with respective frequencies of products as a heatmap. Obviously, there is a distinction between the clusters: The first cluster shows low frequency across all products while the second one is dominated by Current Accounts. The third cluster might be called the Payroll products basket showing some considerable occurrences of Credit Card and e-account. The fourth cluster is dominated by so called special accounts.

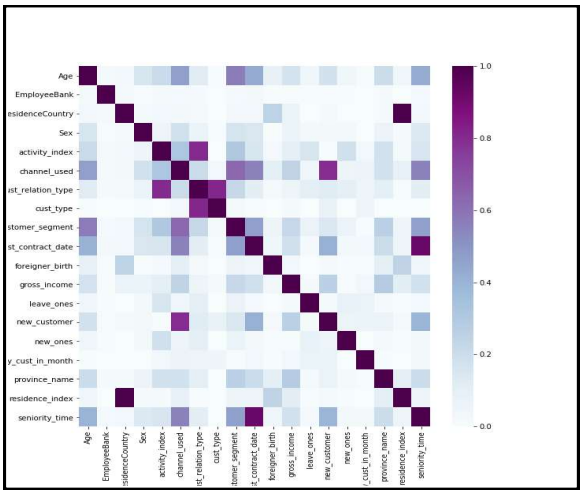


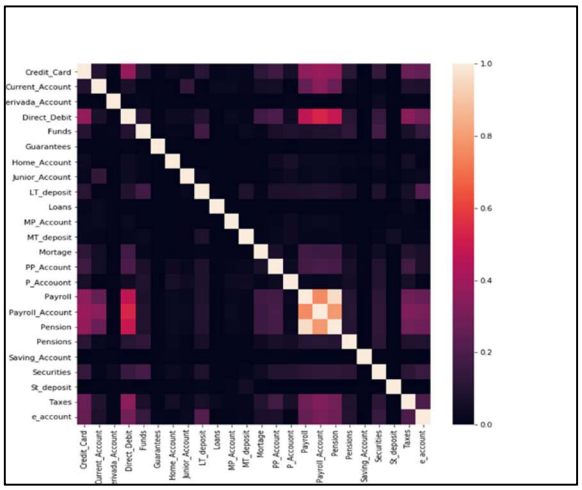
Feature dependencies?

Although, the clustering analysis is helpful to segment clients and records of the product basket, the dependencies among different features are likely to yield some further insights.

The first heatmap below on the left depicts correlations among categorical features describing clients. The correlation coefficients are based on Cramers V which extends chi-square analysis. This reveals some dependencies between features: “channel_used” to “new_customer”, “customer_segment” to “Age” or “activity index” to “relation_type”. Obviously, the way a customer joins the bank is strongly related to whether she is a new client or has a long client history. Moreover, the time how long someone is already a client relates to the segment he is characterized being part of; in addition, there

is also some kind of dependency between the type of relation she has to the bank and if she is considered as an active client. However, most correlations are rather weak, negating a strong level of overall collinearity. But which products are frequently bought together?





Hence, the heatmap above shows again Cramers V correlation coefficients, now between pairs of products. “Direct Debit” has a considerable relation to “Payroll” and “Pension” services. These in turn are corresponding to some extent to “Credit Cards”. Obviously, between these products there is some kind of association.

Some first advices to the bank

For the marketing team in the bank, these analysis reveals some important views: First, there is a group of clients which has just joined the bank recently and which is very active. This group should be easy to access employing merely passive marketing strategies (cluster 3). Secondly, there is a segment of “seniors” which is quite active (cluster 1) as well and might require more “tailor-ship” product recommendations. Instead, the other two groups should be taken care off. They keep relationships with the bank for longer and are considered as inactive (cluster 2 and 4). Thus, they could be at risk to leave the bank and might require more attention on part of the bank. Active marketing initiatives could get them back on track to subscribe to new services.

With regard to the product basket, clients who have Payroll accounts might get offered Direct Debit and Credit Cards and vice versa (cluster 3 product basket). Moreover, Payroll, Direct Debit and Pension accounts show some considerable dependences which are exploitable. Clients who have Particular accounts in their basket – as in cluster 4 – might have some desire for Taxes (accounts) and Pension accounts. Based on these co-occurrences the marketing team could frame the contents of their campaigns to enable more specific customer targeting.

More on feature engineering

So far, the dataset has been cleaned and explored delivering some actionable insight for the bank’s marketing team. However, we need to get some ideas how different features might explain and contribute to the probability of future purchases. But first, two additional features get explored again which have been created from the product basket: “leave_ones” and “new_ones”. These features

show, how many products have been bought or kicked out by a client from one month to the other. Such information seems to be crucial, because clients' recent purchases or terminations might affect their future behaviour, i.e. a client who has recently purchased many products might be less likely to subscribe to new services in the future. Both features represent counts and are treated as categorical variables.

- **Predictive power of single features**

Although the prediction task focuses on the purchase of specific products a simplification through a dummy variable is introduced. This variable simply shows if a client will buy at least one product or not in the subsequent month. Hence, the variable is called "Target" and has binary labels. In order to investigate the predictive relationship of features to the outcome of this target variable correlations and their respective p-values have been assessed.

	feature	p-value	Cramers
0	EmployeeBank	0.0	0.0079
1	ResidenceCountry	0.0	0.0086
2	Sex	0.0	0.0235
3	Age	0.0	0.0665
4	first_contract_date	0.0	0.0391
5	new_customer	0.0	0.0598
6	seniority_time	0.0	0.0372
7	primary_cust_in_month	0.0	0.0021
8	cust_type	0.0	0.0351
9	cust_relation_type	0.0	0.1783
10	residence_index	0.0	0.0054
11	foreigner_birth	0.0	0.0014
12	channel_used	0.0	0.1045
13	province_name	0.0	0.0465
14	activity_index	0.0	0.1623
15	gross_income	0.0	0.0151
16	customer_segment	0.0	0.082
17	new_ones	0.0	0.0814
18	leave_ones	0.0	0.3539

The table above shows Cramers V correlation coefficient of descriptive features to the target variable. Obviously, all p-values hint at statistical significance while the coefficients do not reveal strong impacts, i.e. dependency. Expect for the feature leave_ones, i.e. measures how many products a customer has sold with respect to the last month. Obviously, this feature variation shows a stronger dependency to whether a customer will buy a product next month or not.

The occurrence of specific products might have an impact if a client is going to purchase another product, thus the table below shows correlations between the items in the product basket and the target variable:

	feature	p-value	Cramers
0	**Saving_Account**	**0.0158**	**0.0007**
1	**Guarantees**	**0.0**	**0.0019**
2	**Current_Account**	**0.0**	**0.0555**
3	**Derivada_Account**	**0.0**	**0.0046**
4	**Payroll_Account**	**0.0**	**0.1624**
5	**Junior_Account**	**0.0**	**0.014**
6	**MP_Account**	**0.0**	**0.0439**
7	**P_Accouont**	**0.0**	**0.0073**
8	**PP_Account**	**0.0**	**0.0479**
9	**St_deposit**	**0.0**	**0.034**
10	**MT_deposit**	**0.0**	**0.0074**
11	**LT_deposit**	**0.0**	**0.0475**
12	**e_account**	**0.0**	**0.0902**
13	**Funds**	**0.0**	**0.0324**
14	**Mortage**	**0.0**	**0.0227**
15	**Pensions**	**0.0**	**0.0268**
16	**Loans**	**0.0**	**0.0024**
17	**Taxes**	**0.0**	**0.0855**
18	**Credit_Card**	**0.0**	**0.0659**
19	**Securities**	**0.0**	**0.0433**
20	**Home_Account**	**0.0**	**0.0092**
21	**Payroll**	**0.0**	**0.0804**
22	**Pension**	**0.0**	**0.0847**
23	**Direct_Debit**	**0.0**	**0.108**

Accordingly, all p-values signal statistical inference while the effect of dependence is overly weak. Only, Payroll Account and Direct Debit show numbers higher than 0.100.

Based on these correlations single features do not seem to be sufficient to explain future purchases, moreover, the prediction task is more challenging than simply assessing the probability of buying at least one product or not. Instead, the purchase of specific products must be predicted.

- **Target Mean Encoding and model considerations**

Up to now, against the background of high dimensionality of the feature space, overly categorical feature types and a high degree of required selectivity in the prediction task two major steps are required.

1) To overcome the problems associated with categorical features Target Mean Encoding is conducted. This technique assigns to the different labels of a categorical variable the corresponding frequencies of target labels. For instance, if a feature has two labels, i.e. like gender, then for each label the mean of the target variable is calculated. First, this transforms categorical feature to continuous ones and a relation to the target feature is being established. This encoding might be very helpful because features might show different importance predicting the purchase of specific products.
2) In order to predict the purchase of specific products, different subsets of features might be necessary. Moreover, for the purchase of each product a specific model could provide a better solution than one single model. Thus, for every single product feature selection should be conducted and then a model will be trained for prediction. This is likely to increase specificity to a level which would not be possible developing a single model for all products.

The Pipeline

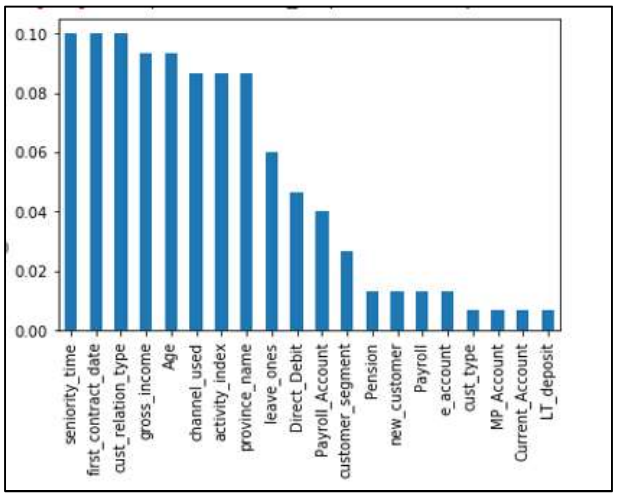
In order to account for both requirements a machine learning pipeline gets designed. This pipeline is used to develop and test models for the prediction of each product respectively. In addition, the pipeline must handle relatively imbalanced data as well. Essentially, 8 targets are very sparse, and thus, require special algorithms and designs for treating extremely strong imbalances. In this project, these special cases remain unsolved.

1) The whole data set is randomly split in a 40 % hold-out and 60 % training set
2) In the training set two steps are performed: Feature selection and parameter grid search with Random Forest Models (RF).
3) Assessing variable importance:
    a. Subsampling of training set with Kfolds
    b. Target Encoding of categorical variables
    c. Subsampling each fold in equally balanced sets
    d. Fitting RF
    e. Assessing Feature Importance (FI)
    f. Repeating steps b to e and averaging Feature Importance over Kfolds
    g. Features are sorted to decreasing importance; those from this order are selected that sum up to 80 % of FI. Most times this leaves one quarter of features as predictors.
4) Model Training:
    a. Grid search of parameter space: number of estimators and max depth
    b. 10-fold cross validation of the training set
    c. Reducing samples to more balanced target relations
    d. Target Encoding of categorical variables
    e. Fitting models with feature subsample from step 3

f. Evaluating model by area under ROC, respectively, weighted F1 score
g. Selecting parameters of the model

Results

The outcome of this pipeline are specific models which contain selected features and parameters for the Random Forest. Hence, for each specific product, a model is developed to deliver the probability of getting purchased. The bar plot below shows the frequency of features after respective selections:



Thus, considering feature selections for all models, 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) and the number of products that are terminated with respect to last month (leave ones) seem to be most useful.

The table below depicts the area under the ROC curve achieved after applying the models on the test set against the averages and standard deviations of the Kfold validation sets. Moreover, upper and lower bands around these averages have been constructed by adding and subtracting two times the respective standard deviations. Overall, the area under ROC of the test set is quite in-line with the results of the validation sets of the cross-validation procedure. Generally, the numbers indicate strong predictive power of the respective models.

Names	mean_test_score	std_test score	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

As a result, the probability of specific product purchases might be assessed which in turn could provide valuable information to the bank's marketing team. Now, they would be able to achieve rankings of future purchases for each client and to start specific product targeting campaigns.

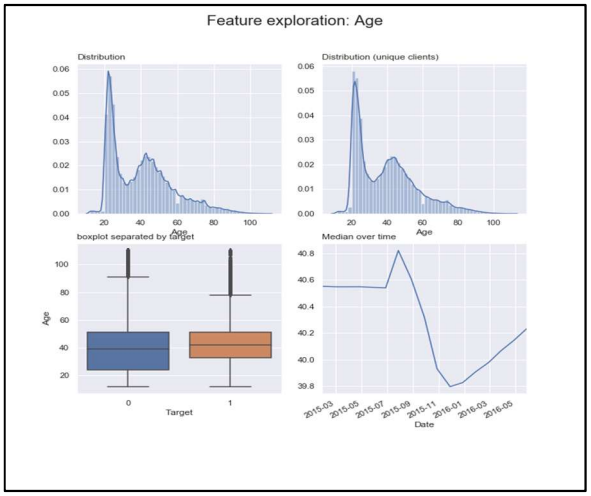
Conclusion and further steps

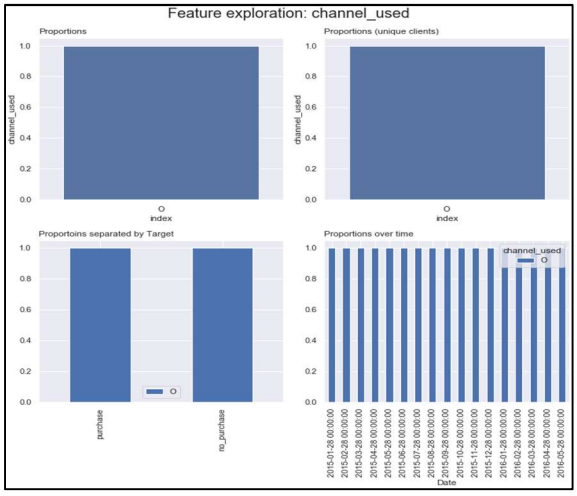
The outcome of this project provides strong support to sales and marketing teams in the banking sector. Essentially, it might help to sharpen customer targeting initiatives which represent a key stone of success in nowadays banking landscape. Thus, the models developed within this project deliver probabilities for the purchase of each product, accordingly, for every customer a ranking of products she is interested in or likely to purchase could get established. Moreover, the pursued approach differs to some extent to the ones which are often applied designing recommendation engines. Here, the focus has been on developing models which predict probabilities for different products, instead of finding similarities among clients. These latter approaches would build upon the cluster analysis and would provide a valuable extension to the presented solution. Overall, the dataset is quite challenging in terms of its size and dimensionality. Hence, dimension reduction techniques could have been conducted in feature engineering in first place or as additional step. Furthermore, other techniques for multilabel classification like classifier chains or Binary Relevance could provide computationally more efficient solutions. However, the classification task is confronted by class imbalances. For this project, under sampling the majority class is intended to train the classifier on more balanced datasets. Though, this leads to information loss which cannot be entirely compensated by cross validating, therefore, other approaches to treat imbalanced datasets could have been conducted and compared. In addition, regarding the eight targets which are heavily imbalanced, special designs and algorithms are necessary. So, in order to complement the recommendation system for these cases, one-class Support Vector Machines or Isolation Forests would provide a solution. Another problem which is unfortunately common to many other recommendation systems is the cold start. Because most models rely on features of the product basket it is not possible to apply them to new clients. Possibly, by means of clustering techniques services could get assigned to new clients in order to "fill" the required feature space.

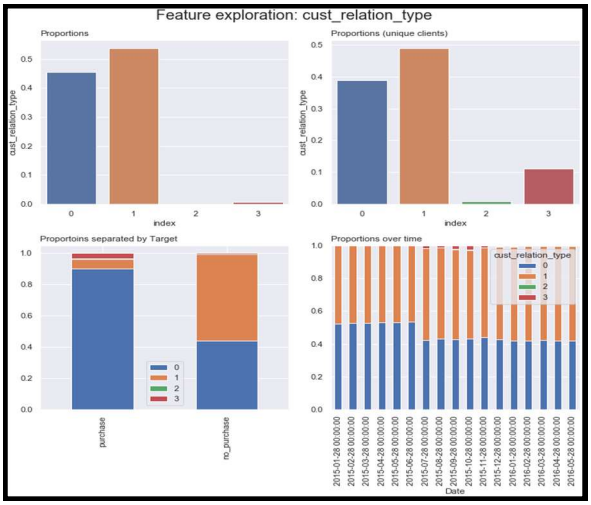
Appendix

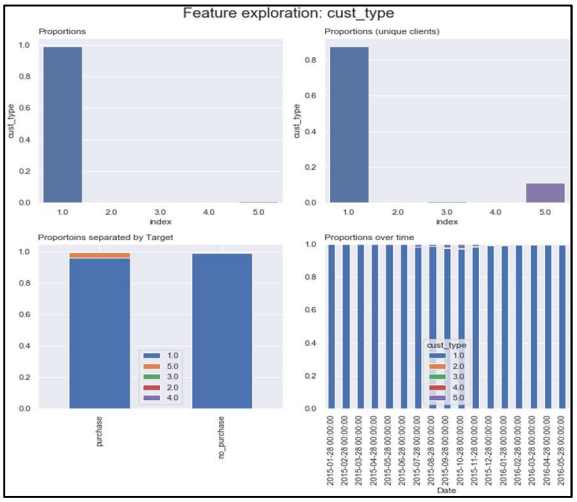
The following graphics aim to support exploration of features describing the clients. Although, some of these features do not provide much insight they are presented for sake of completeness. Hence, for every feature there are four plots:

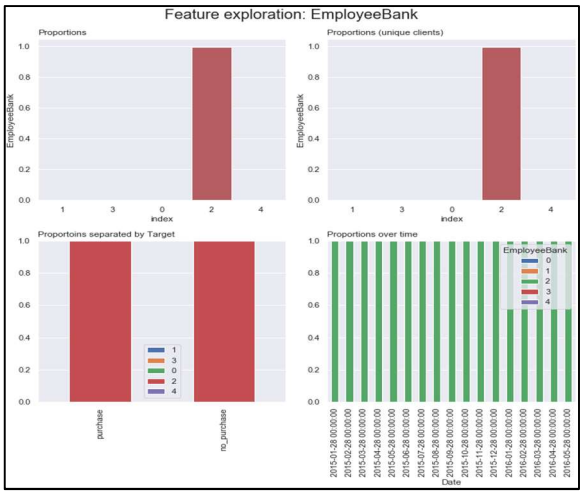
1. Proportions: shows the percentage proportion of labels for categorical and the histogram for continuous variables.
2. Proportions (unique clients): this graph shows the similar content to the first, except that here only unique records are taken into account. This give a more precise picture of distributions and percentage proportions, because there is not any distortion caused by different client frequencies. For instance, if one client is male and shows up 10 times in the dataset and another one is female and appears only 4 times averaging the values would not lead to equal frequency (one man & one woman) in this case.
3. This plot depicts percentage proportions of labels for categorical features and boxplots for continuous ones with respect to whether at least one product will be purchased or not. So, a high percentage of a label occurrence in the “purchase” group indicates predictive ability.
4. Over time proportions of categorical labels or the median are plotted to show eventual time dependencies or seasonality.

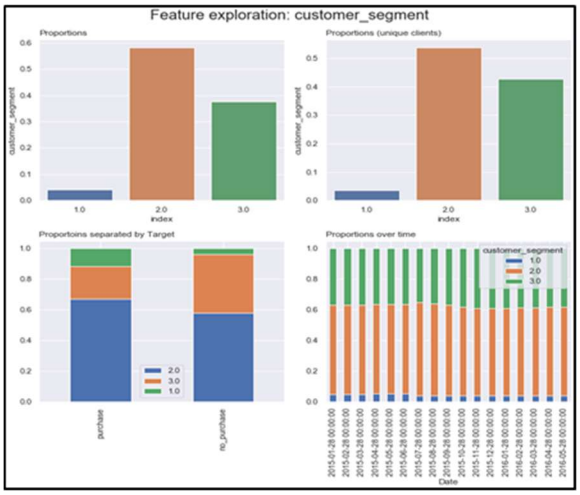


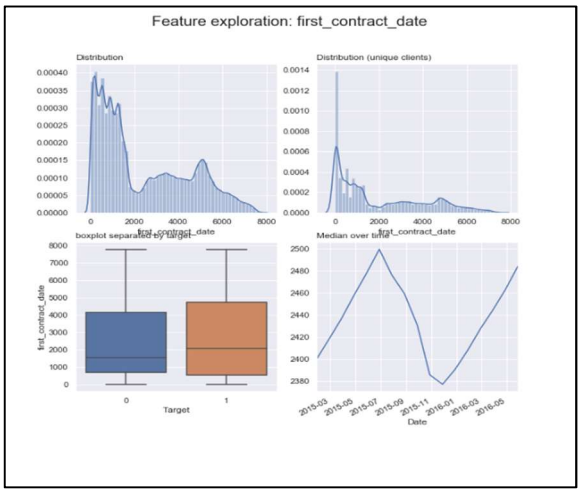


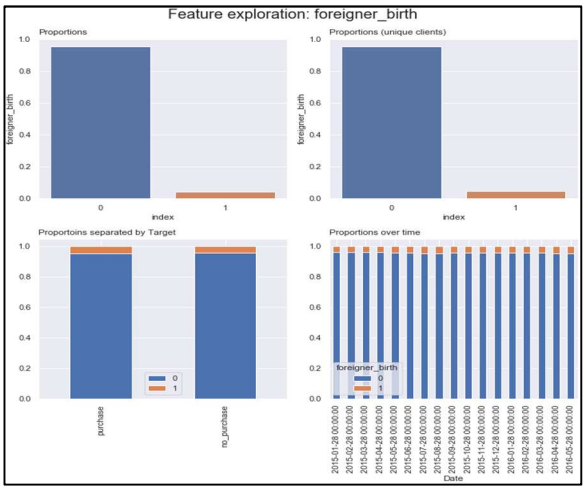


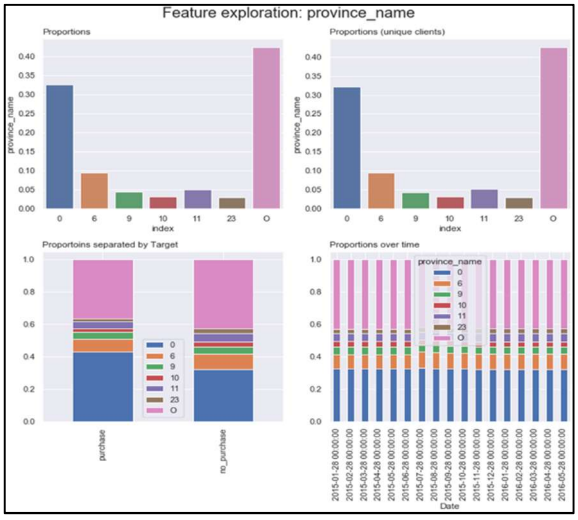


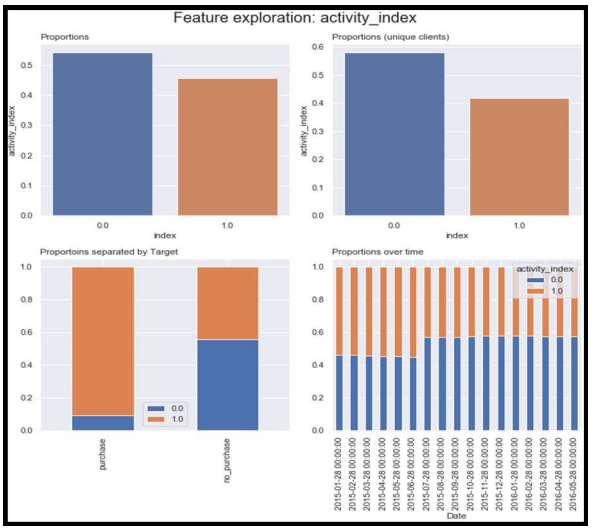

Feature exploration: EmployeeBank
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
EmployeeBank
index
Date
purchase
no_purchase

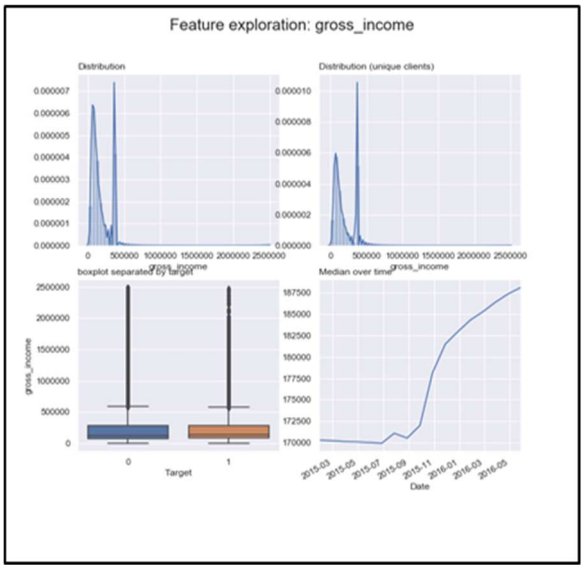

Feature exploration: customer_segment
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
customer_segment
index
Date
purchase
no_purchase

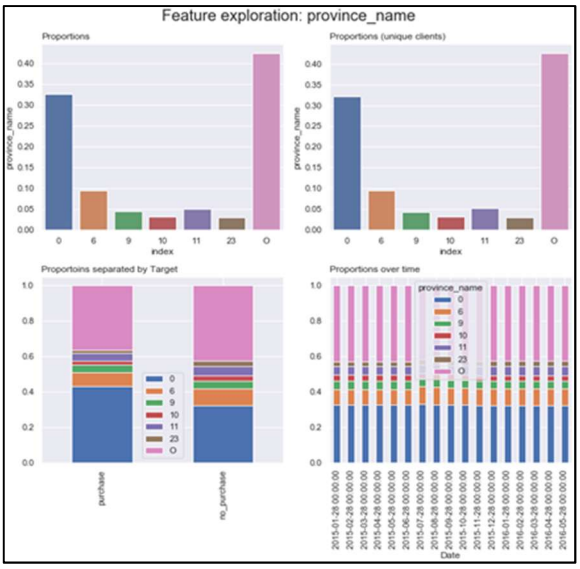

Feature exploration: first_contract_date
Distribution
Distribution (unique clients)
boxplot separated by target
Median over time
first_contract_date
Target
Date

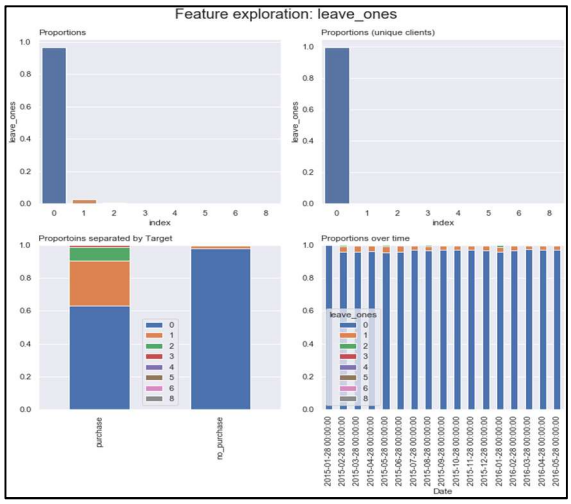

Feature exploration: foreigner_birth
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
foreigner_birth
index
Date
purchase
no_purchase

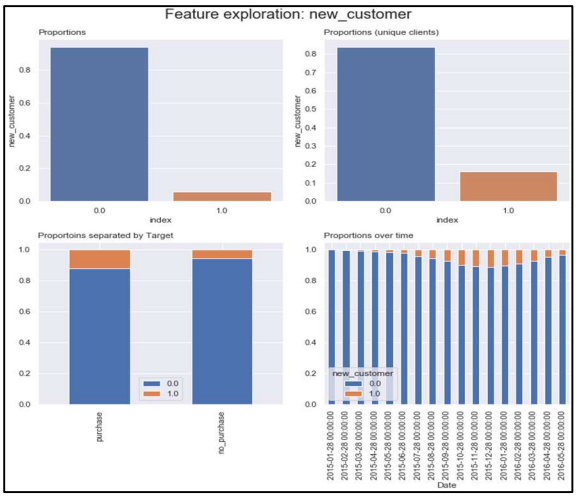

Feature exploration: province_name
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
province_name
index
Date
purchase
no_purchase

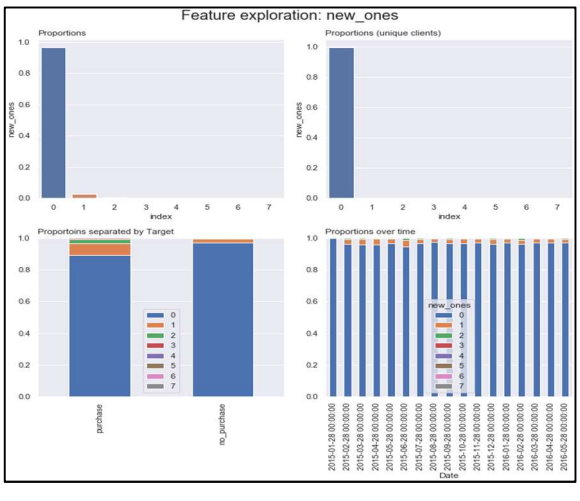

Feature exploration: activity_index
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
activity_index
index
Date
purchase
no_purchase

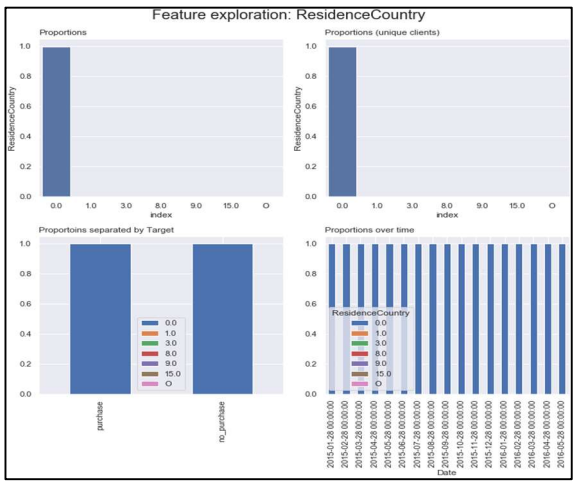

Feature exploration: gross_income
Distribution
Distribution (unique clients)
boxplot separated by Target
Median over time
gross_income
Target
Date

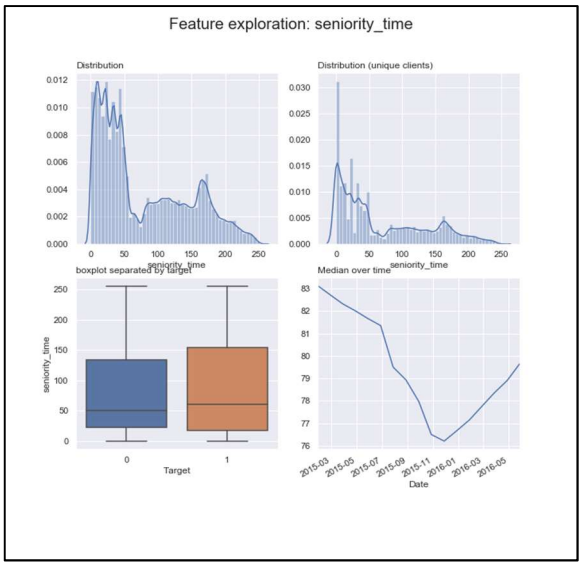

Feature exploration: province_name
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
province_name
index
purchase
no_purchase
Date

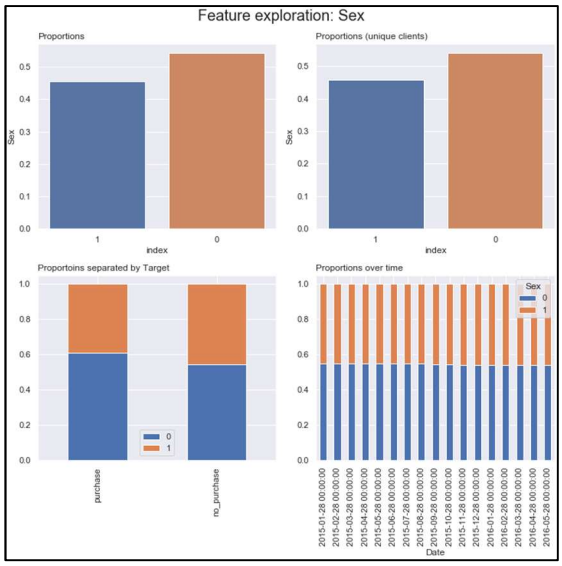

Feature exploration: leave_ones
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
leave_ones
index
purchase
no_purchase
Date


Feature exploration: new_customer
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
new_customer
index
purchase
no_purchase
Date


Feature exploration: new_ones
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
new_ones
index
purchase
no_purchase
Date


Feature exploration: ResidenceCountry
Proportions
Proportions (unique clients)
Proportoins separated by Target
Proportions over time
ResidenceCountry
index
purchase
no_purchase
Date


Feature exploration: seniority_time
Distribution
seniority_time
Distribution (unique clients)
seniority_time
boxplot separated by target
seniority_time
Target
Median over time
seniority_time
Date


Feature exploration: Sex
Proportions
Sex
index
Proportions (unique clients)
Sex
index
Proportoins separated by Target
purchase
no_purchase
Proportions over time
Sex
Date