TASK REPORT

Author: Tomas Kliment (08.03.2022)

Task: Customer targeting for marketing campaign based on available dataset.

Available data:
- customer-spending-1.csv
- customer-spending-2.csv
- customer-spending-3.csv

I understood the task, as follows: It's possible to create customers clusters that will be characterized by the same customer behavior. Based on similar features, it will be possible to design a marketing campaign directly for a given cluster i.e., directly for a specific customer. The available data needs to be preprocessed to determine the appropriate features on the basis of which I will create the clusters.

I approached the task as follows:
- Analysis of information in the available data, in the input .csv files (A1 part).
- Unique customers identification and their new features determination (A2 part).
- Data preprocessing of the new features (A2 part).
- Cluster algorithm and number of clusters identification (B1 part).
- Cluster algorithm fitting and unique customers classifying (to clusters) (C1 part).
- Visualization and data interpretation with consequent advices for marketing (C1 part).

More detailed information on solving these problems can be found below.

A1 – Analysis of delivered data

The dataset contains 537,577 records with 12 columns (dataset contains 12 unique original features). The following table describes the number of unique values in the specific columns:

Tab.1 – Specification of the columns

Column	Num. of unique features	Column	Num. of unique features
user_id	5891	years_in_residence	5
prod_id	3623	car_ownership	2
sex	2	prod_cat_1	18
age_cat	7	prod_cat_2	17
credit_status_cd	21	prod_cat_3	15
edcution_cat	3	revenue_usd	17,959

Missing values are only in 2 columns: prod_cat_2 (166,986 missing records) and prod_cat_3 (373,299 missing records).

One possible approach is to find out if there is a significant correlation between "revenue_usd" with any column characterizing a certain group of people (age, gender, education, etc.) The following "heatmap" characterizes the calculated correlations.

Note: There are columns with continues and discrete values as well. The "heatmap" was determined using dython package (more specifically: http://shakedzy.xyz/dython/modules/nominal/).

Fig. 1 – Correlations between original dataset features (columns). Columns "prod_cut_2" and "prod_cut_3" were removed because of missing data.

The correlation "heatmap" shows that there is no significant correlation between "revenue_usd" and the category that would determine the group of people. As a result, it is not clear which group of people to focus on.

A2 – Unique customers identification and their new features

The original dataset has been modified in order to identify unique users. The following features were then identified for each user:

1. Average credit status for user – averaging was introduced for a reason if some values don't match. Although this variable is not further specified, it was considered as a feature and would certainly be the subject of questions for the manager.

2. The number of unique products which user has purchased - this may be interesting information about whether the user likes to shop.
3. Average purchase price - can indicate whether a user buys more expensive products/services.

These new features have been preprocessed as follows:

- If a feature has an asymmetric (skewed) distribution, a logarithmic transformation was applied.
- Centering features with different means.
- Scaling features with different variance - scaling features were done by dividing them by standard deviation of each.

The following figure illustrates the histograms of the new features and new features preprocessed by the mentioned steps (the bottom three histograms are preprocessed features):

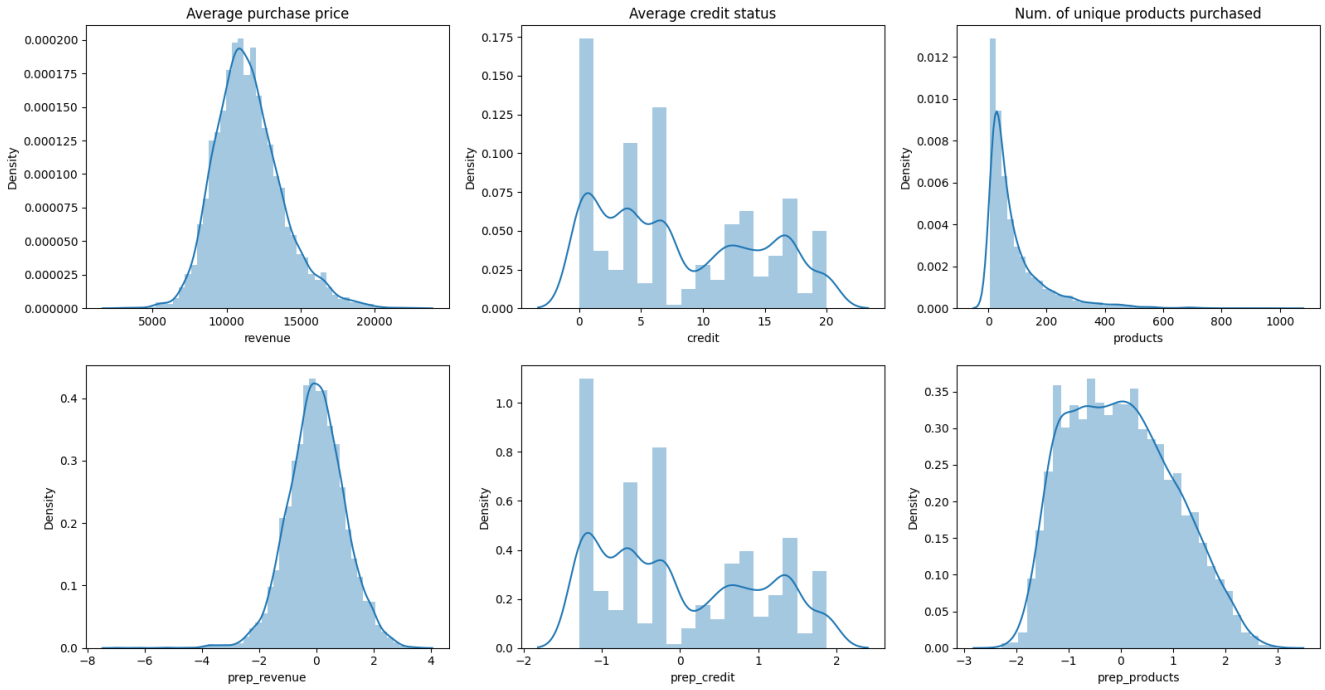

Fig. 2 – Histograms of the new features and preprocessed new features (logarithmic transformation was not applied to "Average credit status", because distribution is not skewed)

B1 – Cluster algorithm and number of clusters identification

To solve this problem, I decided to use k-means clustering. First, it was necessary to determine the number of cluster users are divided into. The elbow method was used to determine the correct K. To determine the optimal number of clusters, the value of K must be selected „at the elbow" i.e., the point after which the distortion/inertia start decreasing in a linear fashion.

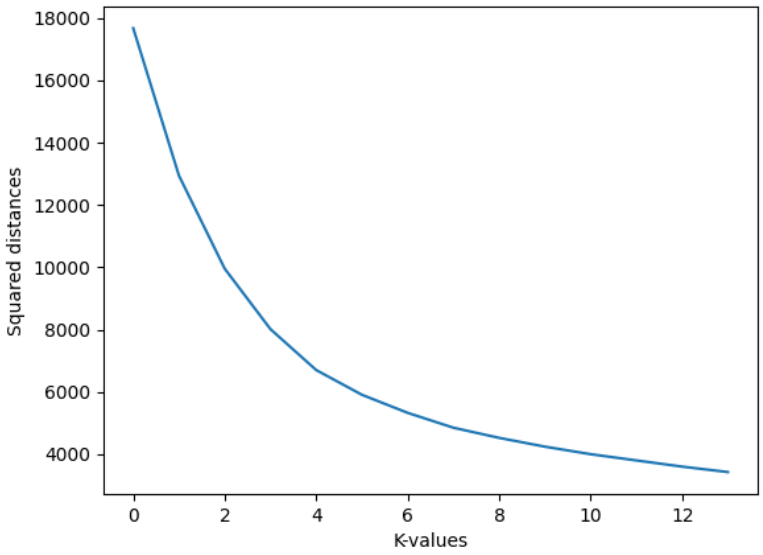

Fig. 3 – The results of the elbow method visualized

As an adequate k-value 7 was chosen.

C1 – clusters identification and results interpretation

The individual colors illustrate the affiliation of users (customers) to a particular cluster. Each axis in 3D visualization corresponds with determined feature. Fig. 4 and Fig. 5 visualize the same data from a different perspective/angle.

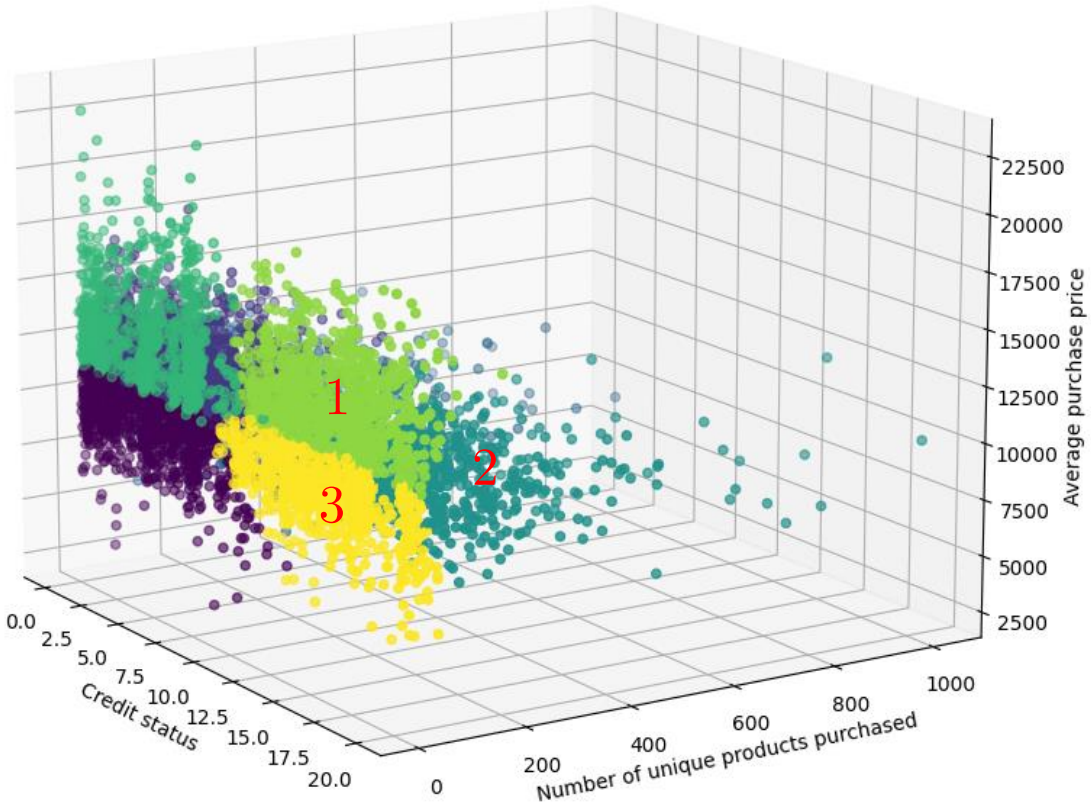

Fig. 4 – Clusters visualized (visualization without preprocessed feature space)

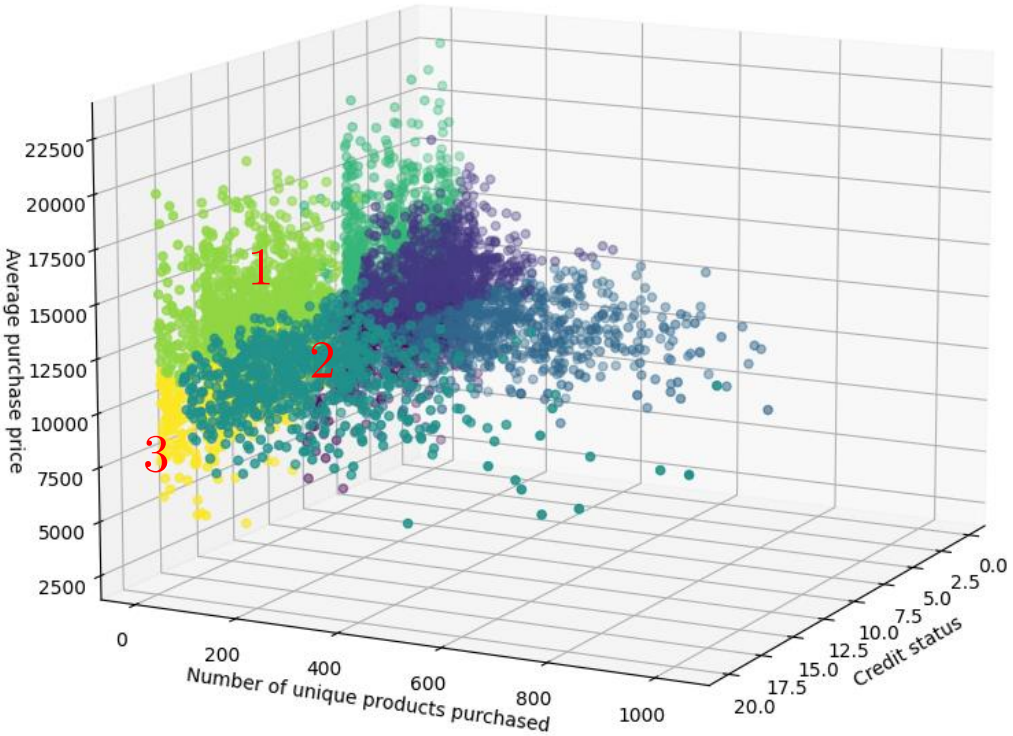

Fig. 5 – Clusters visualized (visualization without preprocessed feature space)

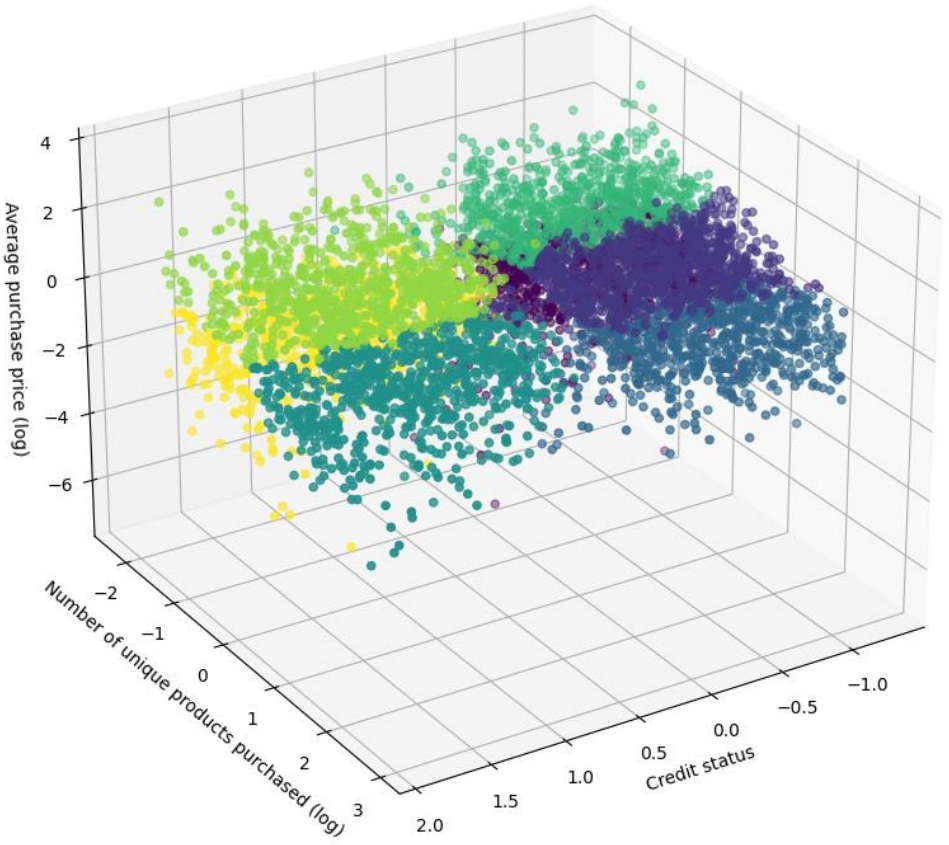

Fig. 6 – Clusters visualized (visualization with preprocessed feature space: logarithmic transformation, centering, scaling)

The interpretation of identified clusters and the way of the customer targeting could be as follows:

Cluster 1:
Characterization: Customers with a smaller number of purchases but a high average purchase price.
Possible marketing strategy: The campaign can focus on the quality and design of the product/service. It is assumed that the customer has no problem paying extra.

Cluster 2:
Characterization: The customer is characterized by a higher number of purchased unique products/services.
Possible marketing strategy: Offer to the customer additional (complementary) products/services to the already purchased products/services. This type of customer is supposed to love shopping.

Cluster 3:
Characterization: These customers made a lower number of purchases with a lower average purchase price.
Possible marketing strategy: Advertising should be better targeted to increase the number of purchases. Discounts may be interesting for this group.

The example above showcases how to differentiate between clusters of customers and how to address them specific marketing campaign. Each customer would first be assigned to a certain cluster based on their behavior and then the appropriate strategy would be chosen.

Once the meaning of „credit status_cd“ is clarified, I would adapt the interpretation of the clusters and choose the appropriate strategy.