Feature Selection and Outliers

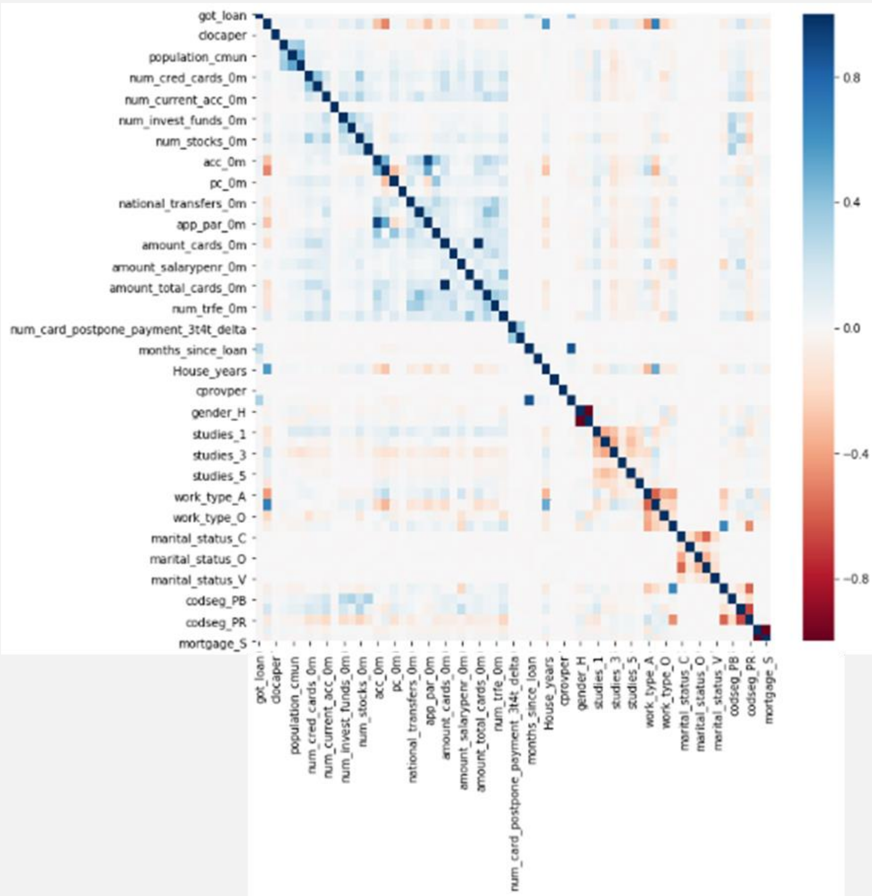
3. Feature Selection

Four different correlation and feature selection analysis have been defining the "final" features used.

3.1 Heatmap

A heatmap is very useful in visualizing the concentration of values between two dimensions of the dataset and the correlation between the two.

Graphic 6



Insights:

- A first contact revealed how some features are highly correlated with one other. Some of them also more correlated than others with the dependent variable. See further detail in 2.2.

3.2 Top varible absolute correlation

The following of the highest absolute correlation pairs from a correlation matrix:

- Unstack and order to get the most correlated pairs.
- Set_option allows displaying all instances.
- Drop all duplicates and ascending order to display the more correlated only.
- Eliminate variables that are highly correlated with others.

```
         Top absolute Correlation pairs

Out[123]: amount_total_cards_0m   amount_cards_0m      1.0000
          mortgage_S              mortgage_S           1.0000
          mortgage_N              mortgage_S           0.9989
          acc_0m                  app_par_0m           0.9088
          previous_loans          months_since_loan    0.8703
          work_type_J             age                  0.6861
          codseg_PI               codseg_PR            0.6683
          work_type_P             codseg_CN            0.6549
          work_type_A             work_type_J          0.5895
          codseg_PR               codseg_CN            0.5847
          marital_status_C        marital_status_S     0.5729
          House_years             age                  0.5660
          disporent_cmun          population_cmun      0.5182
          app_par_0m              mobile_0m            0.5151
          House_years             work_type_J          0.4966
          dtype: float64
```

Insights:

- Useful to avoid modelling problems by eliminating varibles that are highly correlated with others.For this exercise, > 0.7 is going to be considered high correlation.

3.3 Top correlated features with the dependent variable

- The highest correlated variable is ‘previous_loans’.
- The features to be later included in the model will be plotted to clearly see the correlation between them (Pairplot).
- As previously seen ‘months_since_loan’ and ‘previous_loans’ are highly correlated with one other.

```
Out[40]: got_loan             1.000000
         months_since_loan    0.278317
         previous_loans       0.314747
         Name: got_loan, dtype: float64
```

Insights:

- There are no highly correlated variables with the output variable ‘got_Loan’. However, correlation does not mean causality so continued analysis of the variables was required.

3.4 Backward feature elimination [2]

All features were included in this model. Afterwards, the performance of the model was verified. The worst performing features were removed iterativilly until the performance of the model was acceptable. The performance metric used was ‘pvalue’.

- Created loop with *while* to build the model in every new iteration
- Adding constant column of ones, mandatory for sm.OLS model.
- Features with pvalues > 0.05 were elimintated.

The final features number was: 33

```
features_to_model = ['got_loan',
                     'age',
                     'population_cmun',
                     'num_cred_cards_0m', 'aveg_amount_cred_card_0m', 'amount_cards_0m', 'amount_total_cards_0m',
                     'num_stocks_0m',
                     'acc_0m', 'mobile_0m', 'pc_0m', 'tablet_0m', 'app_par_0m', 'web_par_0m',
                     'amount_rec_0m', 'num_rec_0m',
                     'aved_amount_postpone_payment_3t4t_delta',
                     'months_since_loan',
                     'car_years',
                     'House_years',
                     'cprovper',
                     'previous_loans',
                     'gender_V',
                     'studies_1', 'studies_2', 'studies_3',
                     'work_type_A', 'work_type_J', 'work_type_O', 'work_type_P',
                     'codseg_PB', 'codseg_PI', 'mortgage_N']
```

Insights:

- Since the final work explains why a specific customer is selected by the model and it is also known that <<Global features important>> may not be important in the local context, the features selected previously (Backward Elimination) were used to prove this and to give more information to sales representatives.

4. Outliers [3]

2 methods to study outliers were carried out.(IQR & Zscore)

4.1 IQR (mean)

<< The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers>> *SciPy.org*

<<The **first quartile (Q1)**, is defined as the middle number between the smallest number and the median of the data set, the **second quartile (Q2)** – **median** of the given data set while the **third quartile (Q3)**, is the middle number between the median and the largest value of the data set. **IQR = Q3 – Q1** covers the center of the distribution and contains 50% of the observations>> *Geeksforgeeks*

Followed process:

- Drop features factorized.
- First and third quartile were defined.
- 1.5 times the IQR is a suspected outlier and 3 times the IQR above or bellow the Q1 or Q3 accodingly is considered a definitive outlier.
- A dataset was created with the instances with outliers in any feature.
- Dropped columns, included in the outliers´ review dataset.
- Both dataset were concated to get all the columns back without outliers.

```
Dataset with outliers (395937, 61)
Dataset without outliers (163467, 61)
```

Insigths:

- IQR does not work for outliers in this particular dataset due to the high data dispersion.Relevant Info would be lost for the model if used.

4.2 zScore (Standard deviation)

zscore tells us how many standard deviations away a value is from the mean.

- If the zscore is > than 3, that point can be classified as an outlier. Any point +/- 3 standard deviations would be an outlier.

- Features not factorized previously used.
- Created a dataset with the instances with outliers in any featured.
- Dropped columns included in the outliers´ review dataset.
- Concated both dataset to get all the columns back without outliers

```
Dataset with outliers (395937, 61)
Dataset without outliers (371986, 61)
```

Insigths:

- zccore is a better solution than IQR for this dataset, however has not been used for the model due to the fact that most sophisticated models used are prepared to deal with outliers. The tests that have been run, do not improve the results without outliers.

[2]Backward elimination: https://towardsdatascience.com/p-value-basics-with-python-code-ae5316197c52
[3]Outliers: https://medium.com/datadriveninvestor/finding-outliers-in-dataset-using-python-efc3fce6ce32

Although the initial information included in the dataset is real, in order to present this work and to comply with current regulations, all data has been anonymized.