PKDD'99 Discovery Challenge - Berka Dataset

Project by Siddhant Chauhan, Victor Ernoult, Ruturaj Mokashi

Problem Statement

Creating a Datamart to analyze the financial status of customers, segment the customers into a risky customer and potential customers (interested in bank products like cards, loans, etc) and create correlations between the features to track business trends.

Reference: https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm

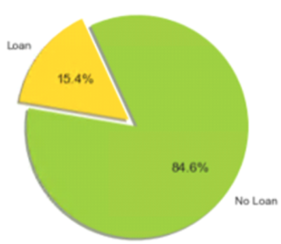
Data Exploration

We used Customer data from credit card, daily transactions, account, loan, demographics, disposition, orders and client information to create the data mart. There are 5,369 unique clients and observations with 49 columns in our data mart.

Loan Status

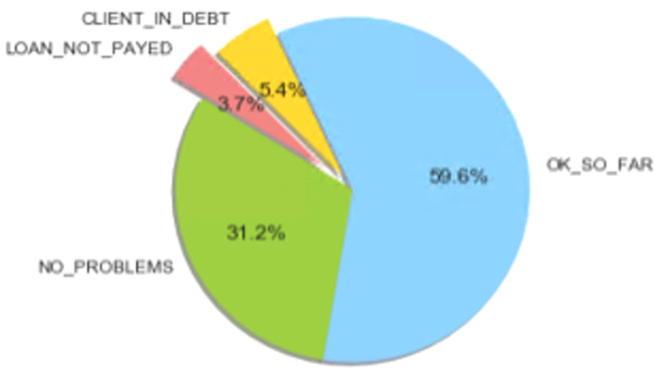
a. The loan status in the basetable was explored to analyze how many have taken a loan. The analysis was represented on a pie chart which shows around 15.4% took the loan and rest 84.6% didn't.

Loan	No Loan
827	4542

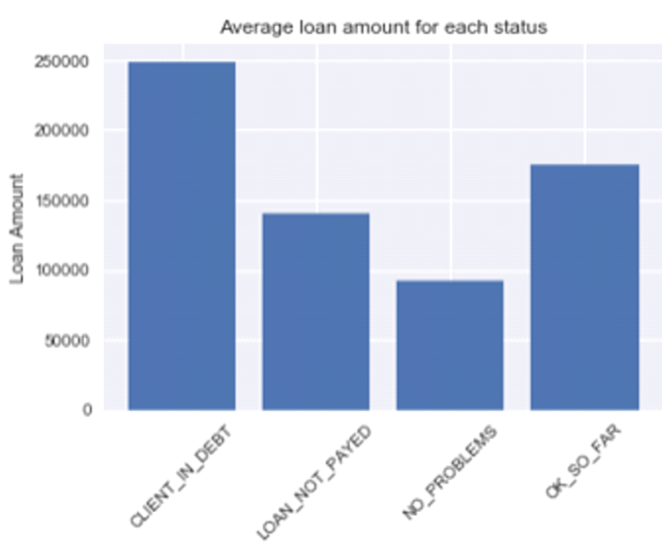


b. To further explore the loan status, the people who took the loan were further divided into 4 categories – 1) who re-payed the loan in time (NO_PROBLEMS), 2) who did not re-pay the loan (LOAN_NOT_PAYED) , 3) who are in the process and paying installments properly (OK_SO_FAR), 4) who are in the process and are in debt (CLIENT_IN_DEBT). The majority had the loan status as 'OK_SO_FAR'

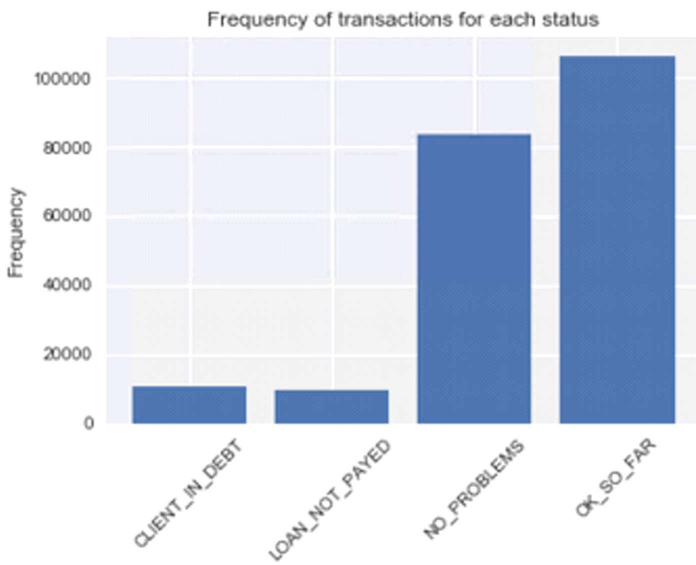
CLIENT_IN_DEBT	LOAN_NOT_PAYED	NO_PROBLEMS	OK_SO_FAR
45	31	258	493



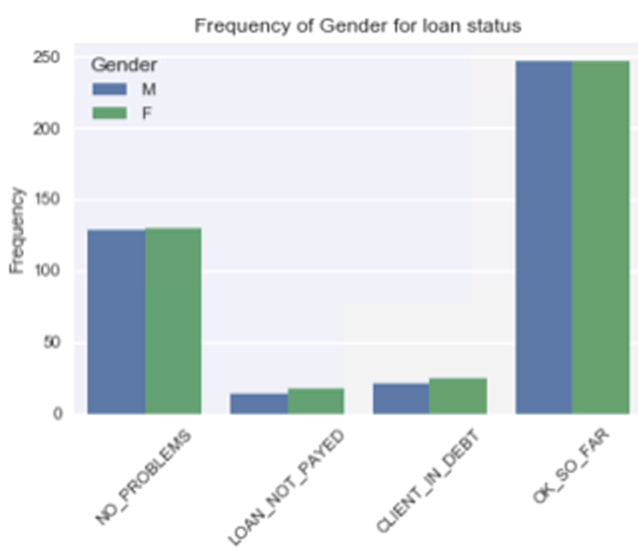
c. The average loan amount for each status was calculated to compare with the total loan amount. Following is the bar diagram which represents the analysis. The average loan amount for client in debt are more compared to other loan status



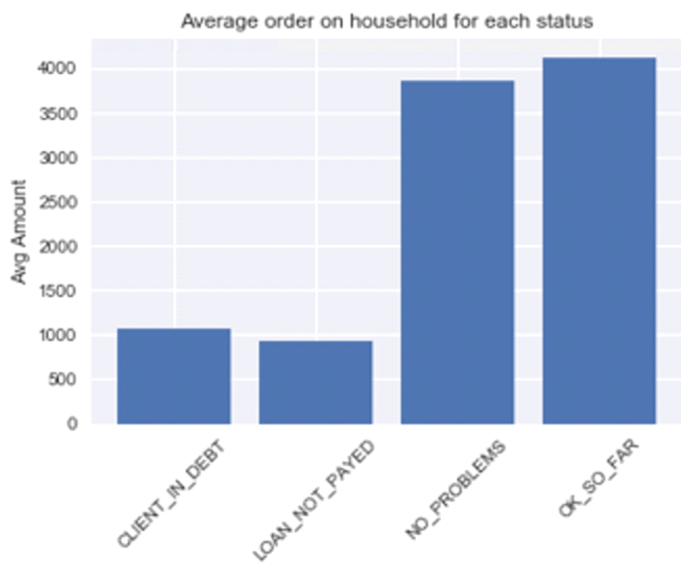
d. The loan status was compared with the frequency of transactions for last 3 years. People with loan status IN_DEBT and Loan_Not_Payed have not transacted much in the recent 3 years.



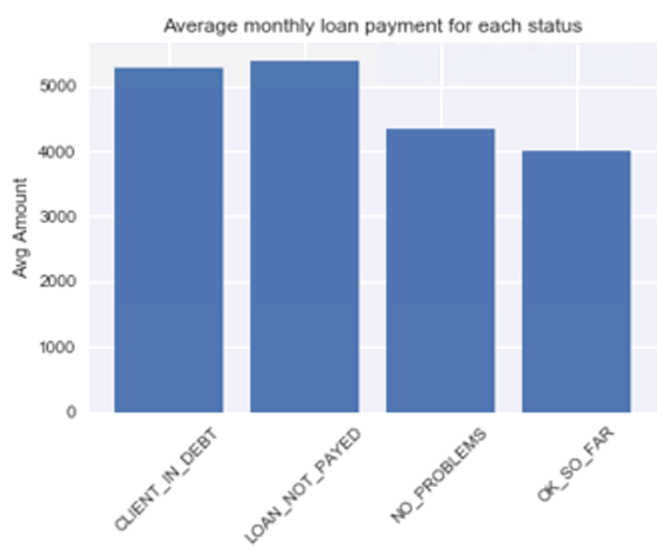
e. The frequency of loan and gender were analyzed to understand the frequency of gender for loan status. The following bar diagram shows both male and female had the high frequency for loans compared to other loan status.



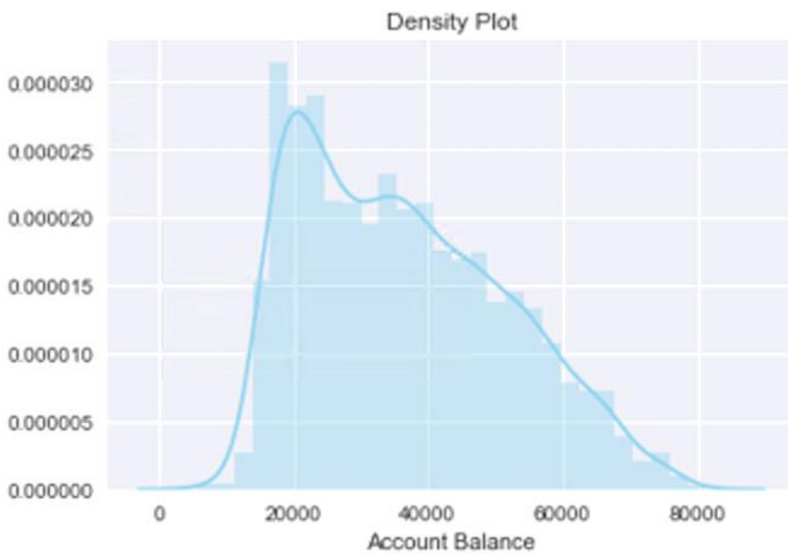
f. For each loan status, average order on household was analyzed. The loan status with 'Ok_So_Far' followed by 'No_Problems' had the high average amount compared to other loan status

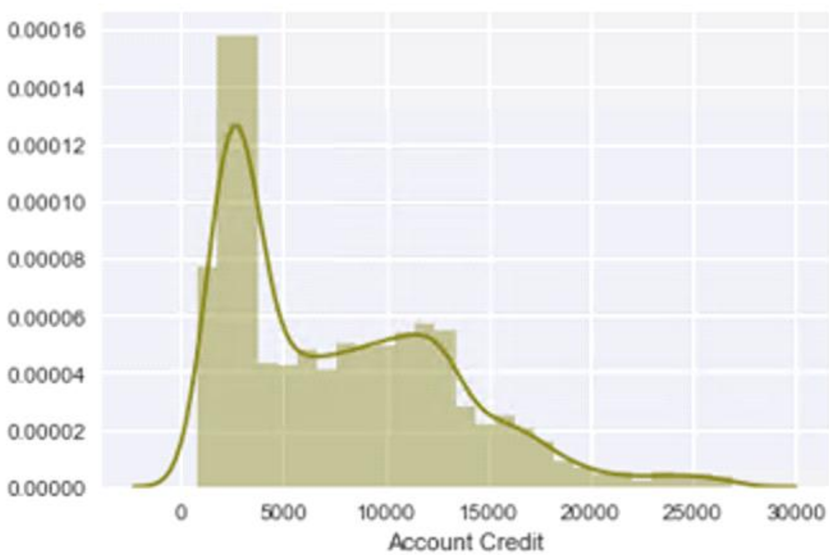


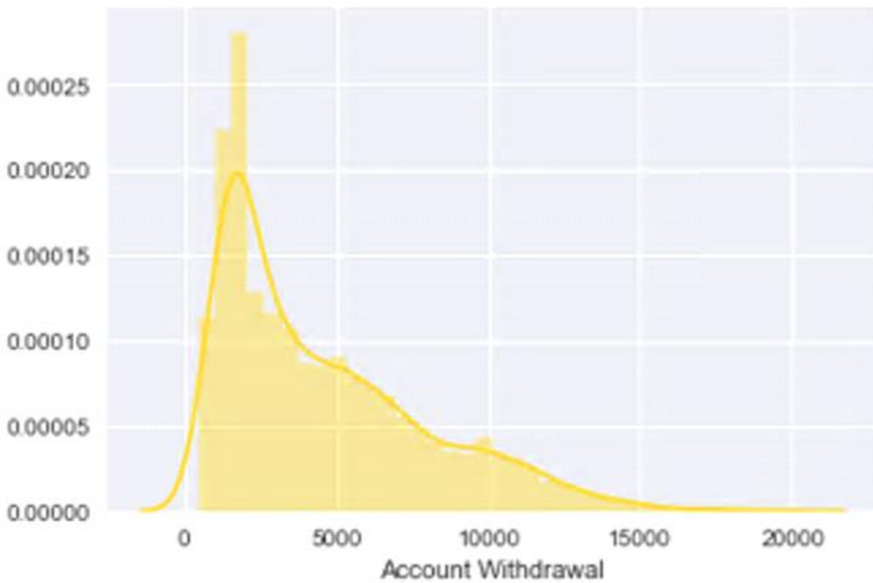
g. The average monthly loan payment for each status shows that average amount for 'Clients_In_Debt' and 'Loan_Not_Payed' are more compared to other status.

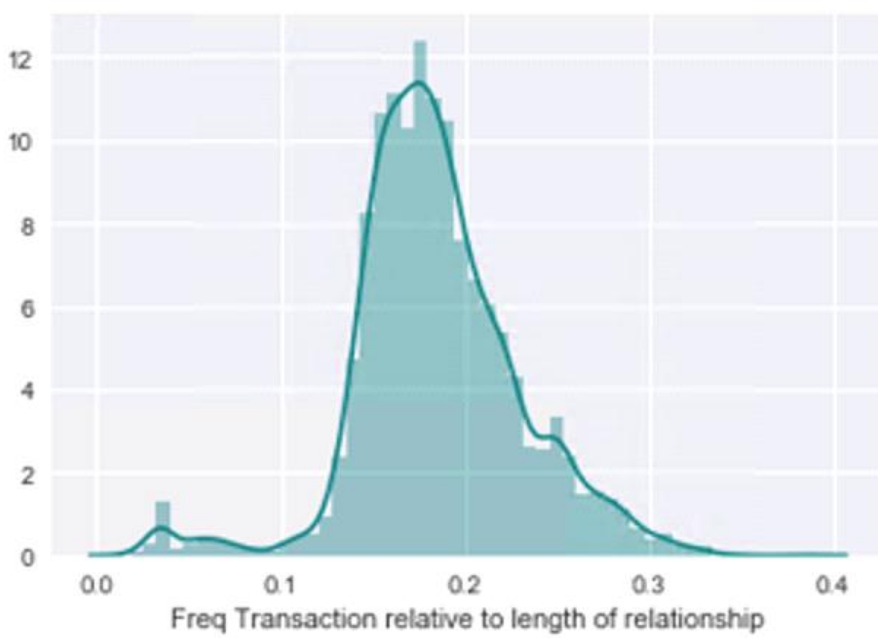


h. The transactional data was explored, to analyze the monthly average balance, account of credit, account withdrawn and frequency of transactions relative to length of relationship. The following density plot shows the analysis,



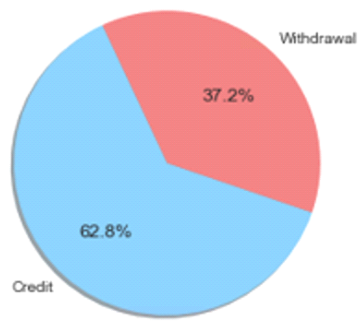



0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
20000
Account Withdrawal


12
10
8
6
4
2
0
0.0
0.1
0.2
0.3
0.4
Freq Transaction relative to length of relationship

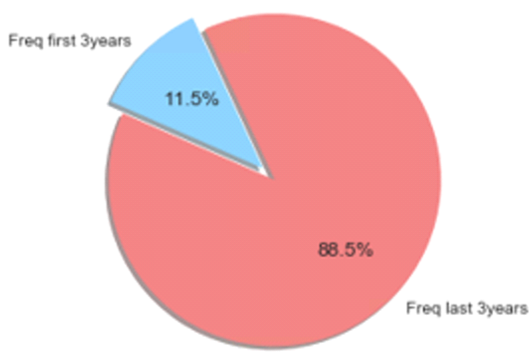
i. The average monthly credited transaction and withdrawal transaction were analyzed using the pie chart. The total credited transactions were more compared to withdrawals.

Credit	Withdrawal
7710.909427181275	4570.141979313054



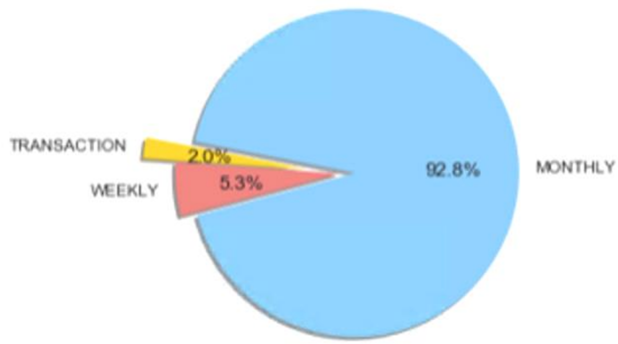
j. The frequency of transactions for first three years and last three years were compared. The following pie chart shows that the freq. of transactions for last years was 88.5% (11,17,655) and the first three years was 11.5% (1,44,970).

Freq first 3years	Freq last 3years
144970	1117655



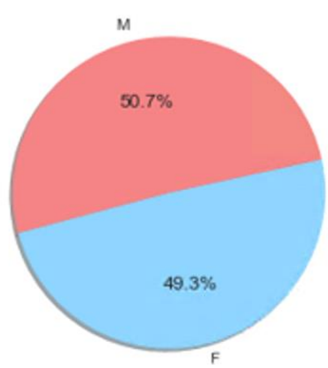
k. The frequency of monthly, weekly was compared.

MONTHLY	TRANSACTION	WEEKLY
4980	107	282



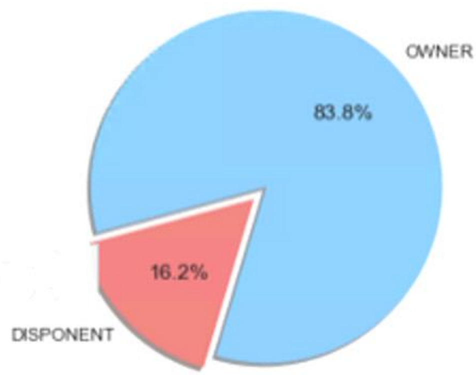
l. The client gender was compared. The number of males is slightly more than the females.

F	M
2645	2724



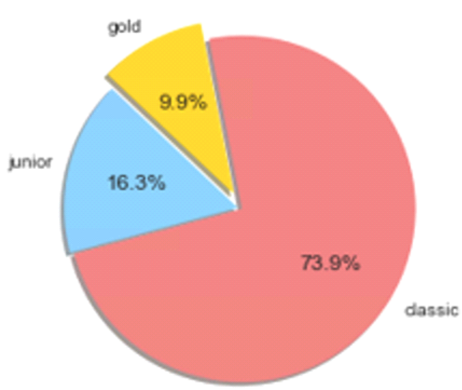
j. The number of disponents and owners were compared. The owners are more than disponent by 83.8%

```
---------  -----
DISPONENT  OWNER
869        4500
---------  -----
```



k. The card types were analyzed to determine the total count. The card types are classic, gold and junior. The number of classic cards is 73.9% more than other card types.

```
-------  ----  ------
classic  gold  junior
659      88    145
-------  ----  ------
```



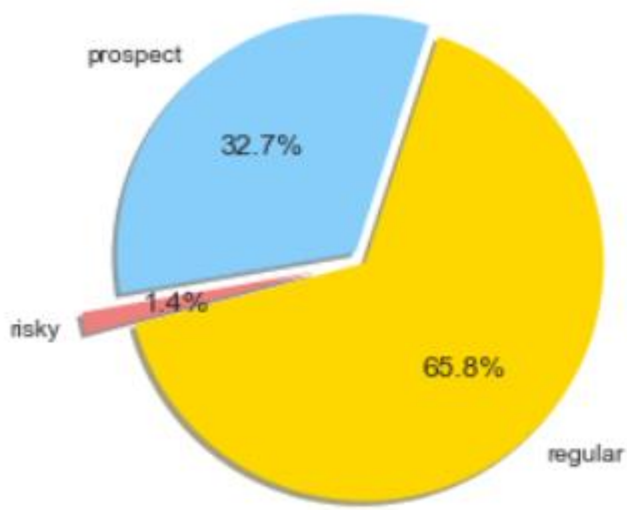
Identification of risks & opportunities

To have a glimpse at the potential of our dataset, we attempted to flag the customers deemed prospects or on the contrary, those bearing risk.

A set of rules was used to create the flags. Potential customers are clients under 70 who own an account with a positive balance and who spend at least half of their income, showing a potential interest for a consumption credit. Moreover, their loan history must be clear and have had an increased activity to qualify as prospects.

On the other hand, customers are considered risky when they have had issue repaying a loan, or when they tend to be in the red, balance-wise.

regular	prospect	risky
3535	1758	76



Data Preparation

1. The required libraries numpy, pandas, matplotlib, datetime were imported. The age and card duration were calculated using a reference date.

2. The datasets were read, and each record describes static characteristic of an account.

3. The disp, card, client, district, account, order, loan, transaction, dataset was preprocessed. The disp dataset column were renamed to disp_type. The card issued format is specified by 'ymd' date function and the type is renamed as card type and for issue, it is renamed as 'days since card issuance'.

3. In the client dataset, the function was written to return nth digits which is an index or list of indexes for which to retrieve the digits. Also, the month of birth number, gender by birth number are returned in subsequent steps. The birth number is converted into a date

4. In the district dataset, the columns A1 to A16 were renamed. The '?' were replaced with proper missing values. The columns were converted from string to floats. To deal with the missing values, we replaced with mean of the region.

5. In the account dataset, the columns district id, frequency, date was renamed. The account opening date was converted to normal date.

6. In the order dataset, the loan columns are renamed.

7. In the loan dataset, the columns amount, duration, payments, status and date were renamed. The loan date was converted to normal date.

8. In the transaction dataset, the columns were renamed. The transaction date was converted to normal date. The 'withdrawal in cash' has the transaction type unknown or withdrawal. We replaced all unknown to withdrawal.

```
Out[14]: trans_operation             trans_type
         CC_WITHDRAWAL               WITHDRAWAL      8036
         COLLECTION_FROM_OTHER_BANK  CREDIT         65226
         CREDIT_IN_CASH              CREDIT        156743
         REMITTANCE_TO_OTHER_BANK    WITHDRAWAL    208283
         UNKNOWN                     CREDIT        183114
         WITHDRAWAL_IN_CASH          UNKNOWN        16666
                                     WITHDRAWAL    418252
         Name: trans_id, dtype: int64
```

Finally, the datasets were merged.

	client_id	district_id	client_age	client_gender	disp_id	account_id	disp_type	card_id	card_type	days_since_card_is
0	1	18	48.04	F	1	1	OWNER	NaN	NaN	NaN
1	2	1	73.91	M	2	2	OWNER	NaN	NaN	NaN
2	3	1	78.23	F	3	2	DISPONENT	NaN	NaN	NaN
3	4	5	62.08	M	4	3	OWNER	NaN	NaN	NaN
4	5	5	58.49	F	5	3	DISPONENT	NaN	NaN	NaN

5 rows × 49 columns

Appendix: Basetable Variable explanation

The basetable has 49 variables. Following is the table which describes each column name

Column Names	Description
Client_id	Client identifier
District_id	Client district identification
Client_age	Age of the client
Client_gender	Gender of the client Male or Female
Disp-id	Disposition to the account
Account_id	Client Account Number
Disp_type	Type of disposition
Card_id	Client card identification
Card_type	Type of Card
Date_since_card_issuance	Number of days since the first issuance of card
District_name	Name of Client District
region	Client region
Num_inhabitants	Number of Inhabitants
Num_munipalities_gt499	Number of municipalities greater than 499
Num_municipalities_500to1999	Number of municipalities from 500 to 1999
Num_municipalities_2000to9999	Number of municipalities from 2000 to 9999
Num_municipalities_gt10000	Number of municipalities greater than 10000
Num_Cities	Number of cities
Ratio_urban	Urban ratio
Average_salary	Average salary
Unemp_rate95	Unemployment rate 95
Unemp_rate96	Unemployment rate 96
Num_entrep_per1000	Number of entrepreneurs per 1000
Num_crime95	Number of crimes 95
N96um_crimes	Number of crimes 96
Account_freq	Frequency of accounts
Account_date_opened	Date the account was opened
Freq_order	Frequency of orders
Freq_order_insurance	Frequency of order insurance
Freq_order_household	Frequency of order household
Freq_order_leasing	Frequency of order leasing
Mon_order_insurance	Monetary order insurance
Mon_order_household	Monetary order household
Mon_order_leasing	Monetary order leasing
Loan_id	Loan identification
Loan_date	Date of loan
Loan_amount	Amount of Loan
Loan_duration	Duration of the loan
Monthly_loan_payment	Monthly loan payment
Loan_status	Status of loan
Account_district_id	Account district identification
Monthly_Loan_Payment	Monthly loan payment amount

Loan_Status	Status of the loan
Recent_transaction	Recent transactions by client
Length_of_relationship	Period of client relationship
Mon_avg_balance	Monetary average balance
Freq_transaction	Frequency of transactions
Mon_trans_cred	Monetary transaction credited
Mon_trans_withdraw	Monetary transaction withdrawn
Freq_first_3years	Frequency of transaction for first 3 years
Freq_last_3years	Frequency of transaction for last 3 years
Type_of_customer	Measure of attention to be allocated to the client