IÉSEG
SCHOOL OF MANAGEMENT

Individual Project Report

Elie Makhoul

Introduction

This report documents the entire process of creating a base table from scratch for the provided relation financial dataset.

The objective of this project is to create an understanding of the time window and create the relevant and necessary features that can be used to predict whether the client should or should not be granted a loan, and whether they should be granted a card.

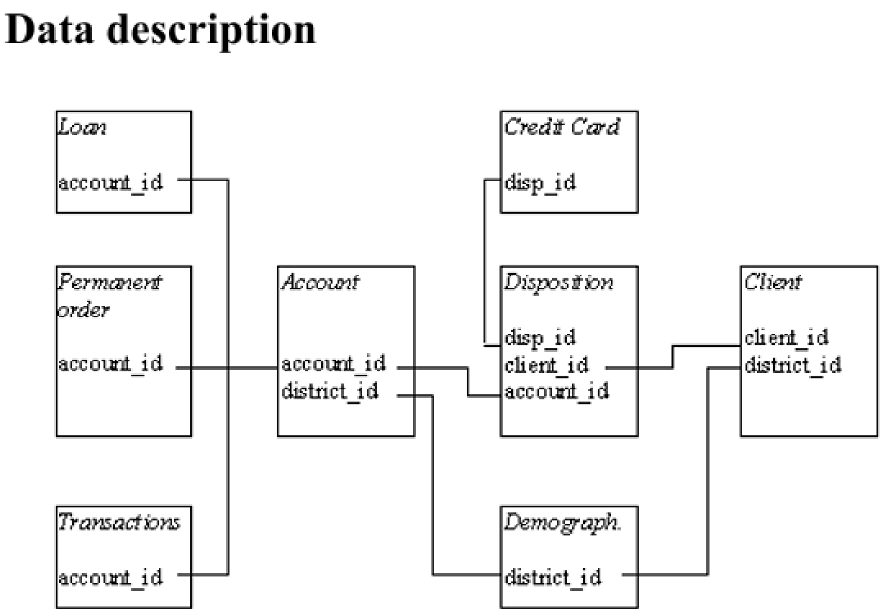


Figure 1

This scheme shows the relationship between the different tables used to create features to help predict the DVs: Target variable #1:

1. Client had granted loan in the dependent variables time window (i.e. 1997), binary value (0 = did not have granted loan, 1 = had granted loan).

2. Target variable #2: Client had credit card issued (for both account owner and disponent) in the dependent variables time window (i.e. 1997), binary value (0 = did not have credit card issued, 1 = had credit card issued).

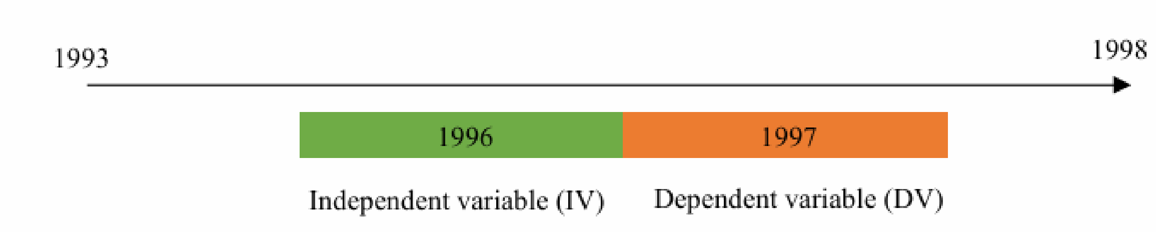


Figure 2

The time window for the IVs is 1996 and for the DVs 1997. Moreover, the granularity level is the client who is account owner. Therefore, each row in the base table represents one owner client.

Data Preparation and transformation

1. Full Explanation

At first, I understood each table, their data types, and their relationship with other tables. I checked for missing values in each table to understand how I am going to interpret them later for feature engineering.

The default base table is the merge between the client and the owner (disp table filtered by owner). Then I created basic features from my already created base table for the client. Features created were birth year, birth month, age.

Later I added accounts that were filtered before 1996(did not take accounts created in 1996 because there is not enough data for these accounts).

Then I added features from the transaction, district, loan, and card table using the 1996 IV-time window date.

After adding features from the loan and card table I added the two DVs *had_loan_97* and *had_card_97*. For the *had_card_97* column, I first merged the card and disp tables with all cards for each client including owner and disponent, then I filtered for date only 1997 and created the DV *had_card_97*. I then joined based on account_id and not on disp_id. merging and joining on account_id. will give the same result as without joining and merging on the disp_id. I adopted this logic because it states in the individual project description that we need the had_card_97 for both owner and disponent. **"Target variable #2: Client had credit card issued (for both account owner and disponent) in the dependent variables time window (i.e. 1997), binary value (0 = did not have credit card issued, 1 = had credit card issued)."**

Our two DVs:

- **got_loan_97(0= no loan, 1= got loan)**
- **got_card_97 (0 = no card, 1= got card)**

2. Data Dictionary

Feature	Data Type	Descripton	Missing	Replaced by
Birth year	Int	Birth month of client	NO	
Birth_month	Int	Birth month of client	NO	
Birth_day	Int	Birthday of client	NO	
gender	Object	"M" or "F"	NO	
age	Int	Age of client	NO	
age_group	Int	Age group of clients	NO	
Age_category	Object	Age category of client	NO	

LOR	Int	Period between account creation and 1996	NO	
Max_trans_date	Date time	Maximum transaction date for each client	NO	
num_of_transactions	int	Frequency of transaction for each client	NO	
total_credit	float	Total credit	YES	0
total_withdrawal	float	Total debit	NO	
net_saving_rate	float	Check whether a client is a saver or a Spender. (totalcredit – total withdrawal) /total credit	YES	0
min_balance	float	Minimum balance	NO	
max_balance	float	Maximum balance	NO	
avg_balance	float	Avg_balance	NO	
balance_volatility	float	Balance_volatility	NO	
last_balance	float	Last_balance	NO	
neg_balance_days	int	Days with Negative balance	NO	
q1_credit_debit_ratio	float	Credit to debit ratio for Q1	YES	-1
q2_credit_debit_ratio	float	Credit to debit ratio for Q2	YES	-1
q3_credit_debit_ratio	float	Credit to debit ratio for Q3	YES	-1
q4_credit_debit_ratio	float	Credit to debit ratio for Q4	YES	-1
total_credit_card_withdrawal	float	Credit card withdrawal (withdrawal from ATM)	YES	0
total_credit_cash	float	Deposit in cash	YES	0
total_debit_cash	float	Withdrawal in cash (withdrawal from bank teller)	YES	0
total_collections	float	Collections from other banks	Yes	0
total_remittance	float	Remittance to other banks	YES	0
card_withdrawal_ratio	float	Card withdrawals (from ATM) from total withdrawa	YES	-1
cash_withdrawal_ratio	float	Cash withdrawals (from teller) from total withdraw	YES	-1
Cash_deposit_ratio	Float	Cash deposit (from teller) from the total credit ratic	YES	-1
collection_remittance_ratio	float	Total collections from total remittance ratio	YES	-1
num_loan_payments	float	Number of loan payments.	YES	0
sum_loan payments	float	Sum of loan payments	YES	0
num_insurance_payments	float	Number of insurance payments	YES	0

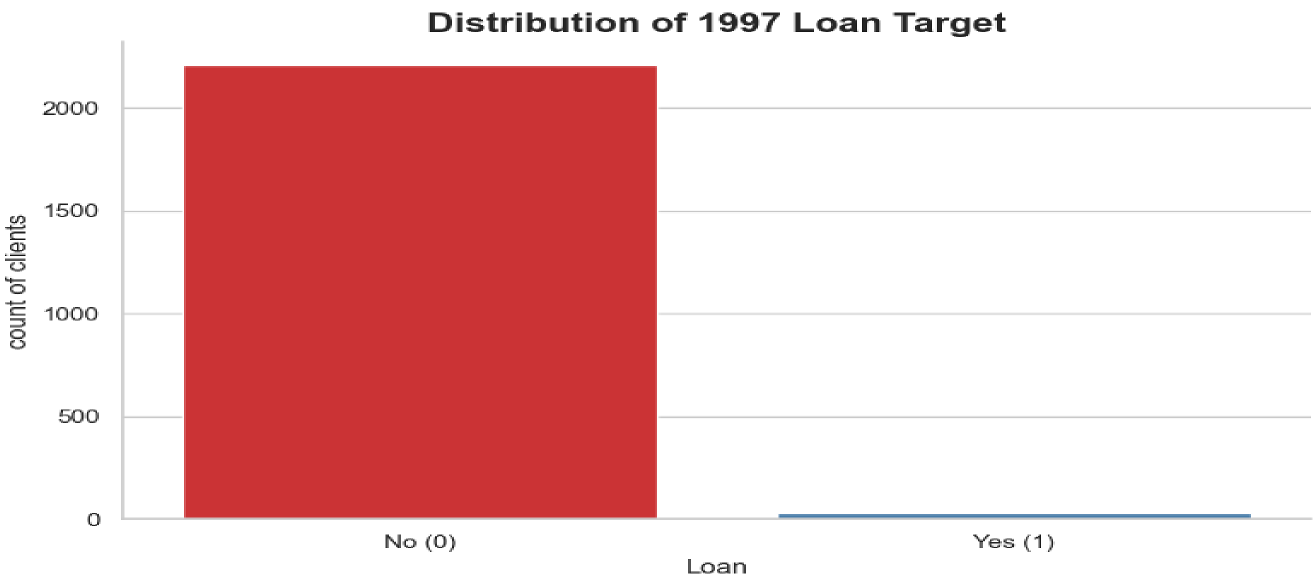
total_insurance_payments	float	Sum of insurance payments	YES	0
num_household_payments	float	Number of household payments	YES	0
total_household_payments	float	Sum of household payments	YES	0
total_pension	float	Total pension	YES	0
total_interest_credited	float	Sum of interest credited	YES	0
num_of_sanctions	float	Number of sanctions	YES	0
region	object	Region of district	NO	
inhabitants	int	Total inhabitants per district	NO	
Avg_salary	int	Avg salary per district	NO	
Crime_96	int	Total numbers of crimes of 1996 for	NO	
Crime_ratio_96_95	float	Ratio of crime rate from prev year	NO	
Unemp_rate_96	float	Total unemployment	NO	
Unemp_rate_96_95	float	Unemployment rate from prev year	NO	
Urban_ratio	float	Urban inhabitants from total inhabitants' ratio	NO	
Entrepreneurs_per_1000	float	Entrepreneurs per 1000 people	NO	
Had_prevoius_loan_96	float	Had a loan in 1996	YES	0
amount	float	Amount of loan	YES	0
duration	float	Duration of loan	YES	0
monthly payments	float	Monthly payment for a loan	YES	0
status	object	Status of loan	YES	-1
Debt_income_ratio	float	Debit to income ratio	YES	-1
Days_from_acct_to_loan	object	Days from account creation to loan granted	YES	-1
Avg_balance_bef_loan	Object	Avg balance before loan issuance	YES	0
Avg_balance_aft_loan	Object	Avg balance after loan issuance	YES	0
Aggressiveness_ratio	float	Show amount over duration ratio	YES	-1
Had_prev_card_96	float	Had previous card in 1996	YES	0
Card_type	object	Type of card	YES	-1
Days_from_acct_to_card	object	Days from account creation to card granted	YES	-1

Got_loan_97	float	Got loan in 1997	NO	
Got_card_97	float	Got Card in 1997	NO	

Note: I considered all the missing values of the ratios as **-1**.I can't have a negative credit or debit therefore all missing values for ratios cromputed are considered as **-1** except for the net_saving_rate which is imputed by **0**. Moreover, for the crime_ratio_96_95 and the unemp_ratio_96_95, I adjusted the crime 95 and unemp 95 missing value to the mean.

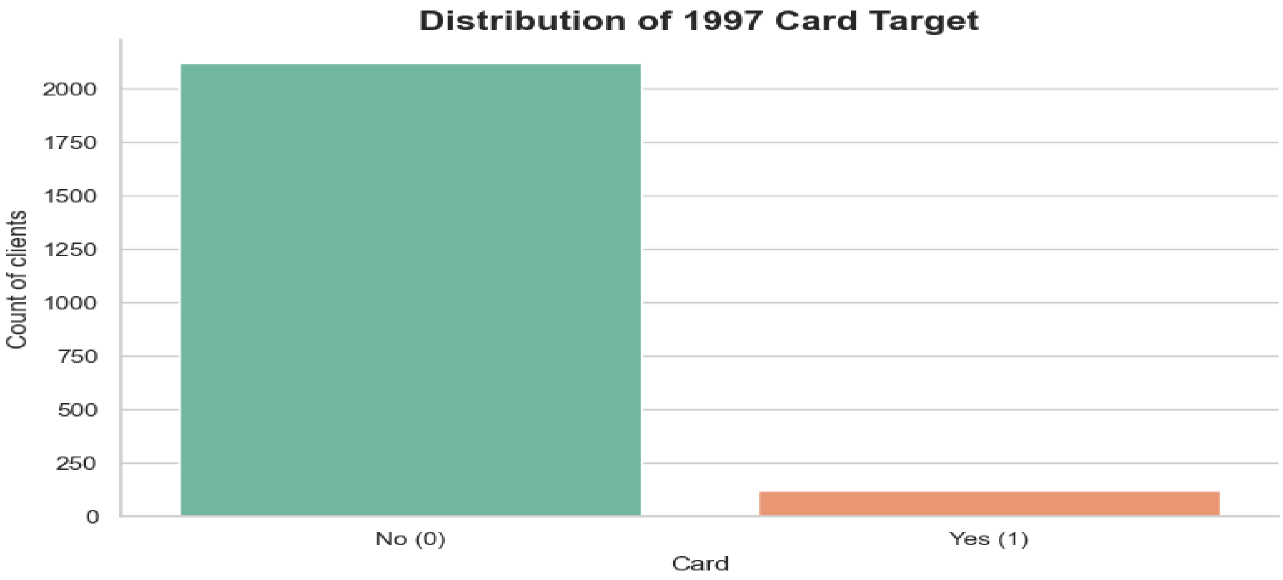
Visualization and Interpretation

1. Distribution of the loan target



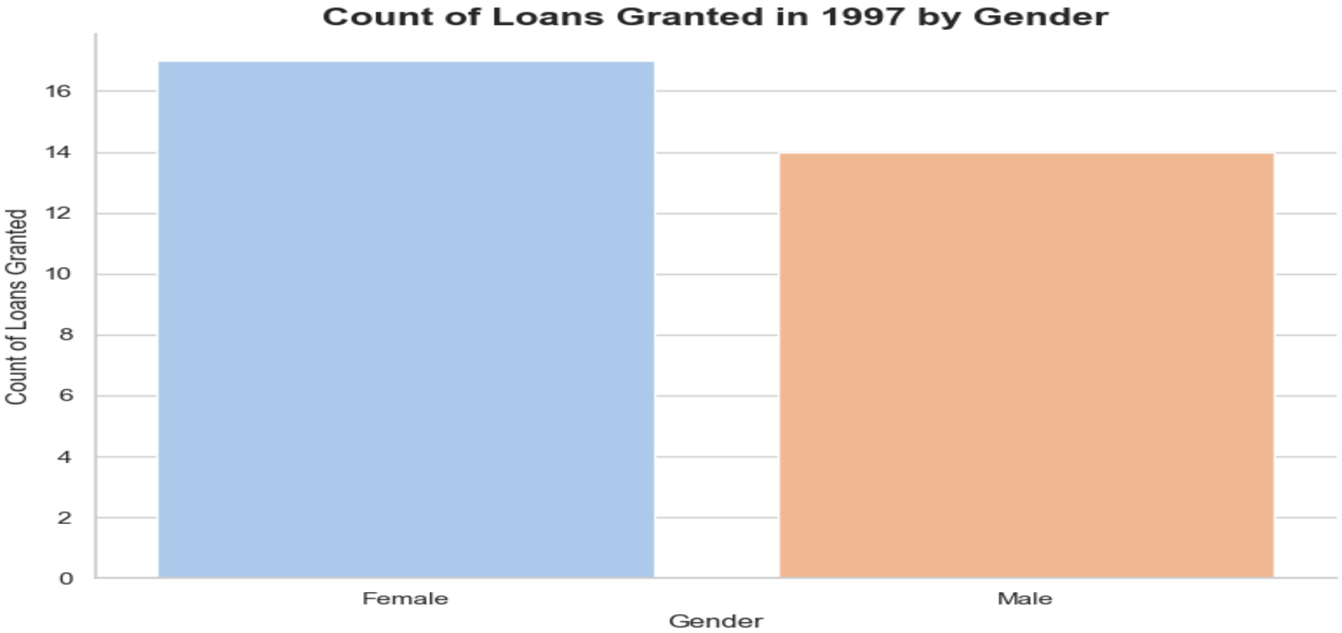
Most clients in 1997 did not get a loan, the reason is unknown. The client might have been rejected, or he did not apply for one in the first place

2. Distribution of card target



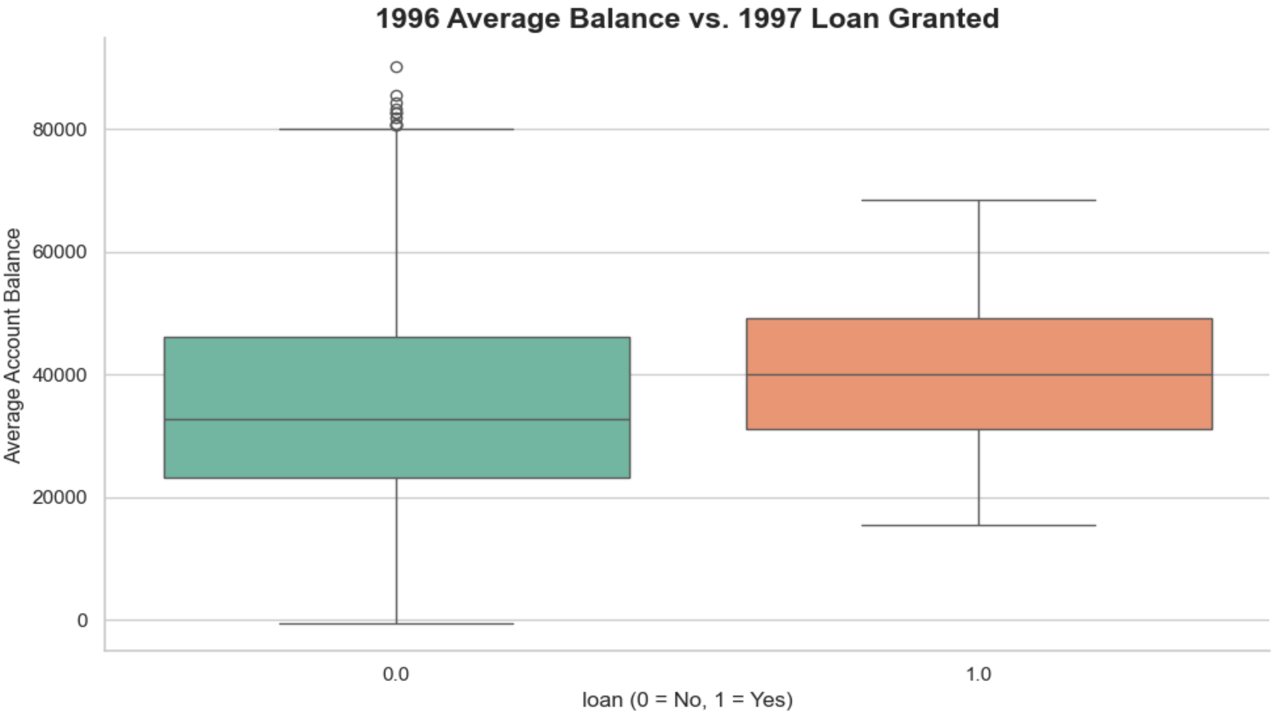
For Card it is the same thing, most clients were not granted a card in 1997. The reason might be that the duration of the card exceeds 1997; therefore, the client might reapply for a card after 1997.

3. **Loan by gender**



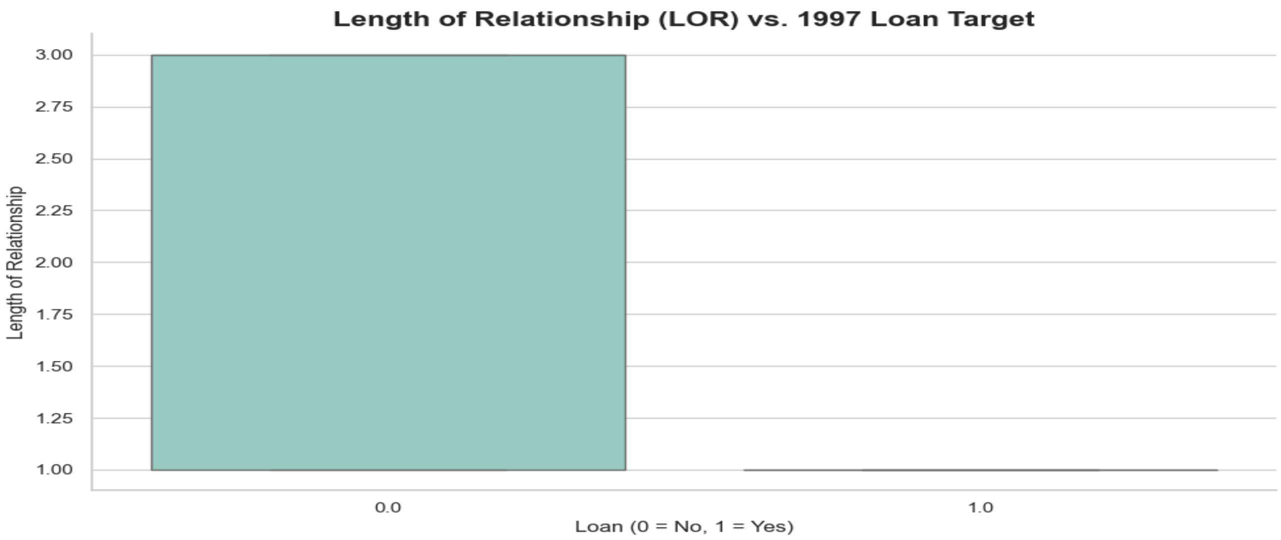
Out of the total 31 loans, 17 loans were given to females, and 14 loans were given to males it is not a big difference, but this difference may be caused by the net saving rate. Maybe Women tend to save more and spend less than men.

4. **Balance for loan**



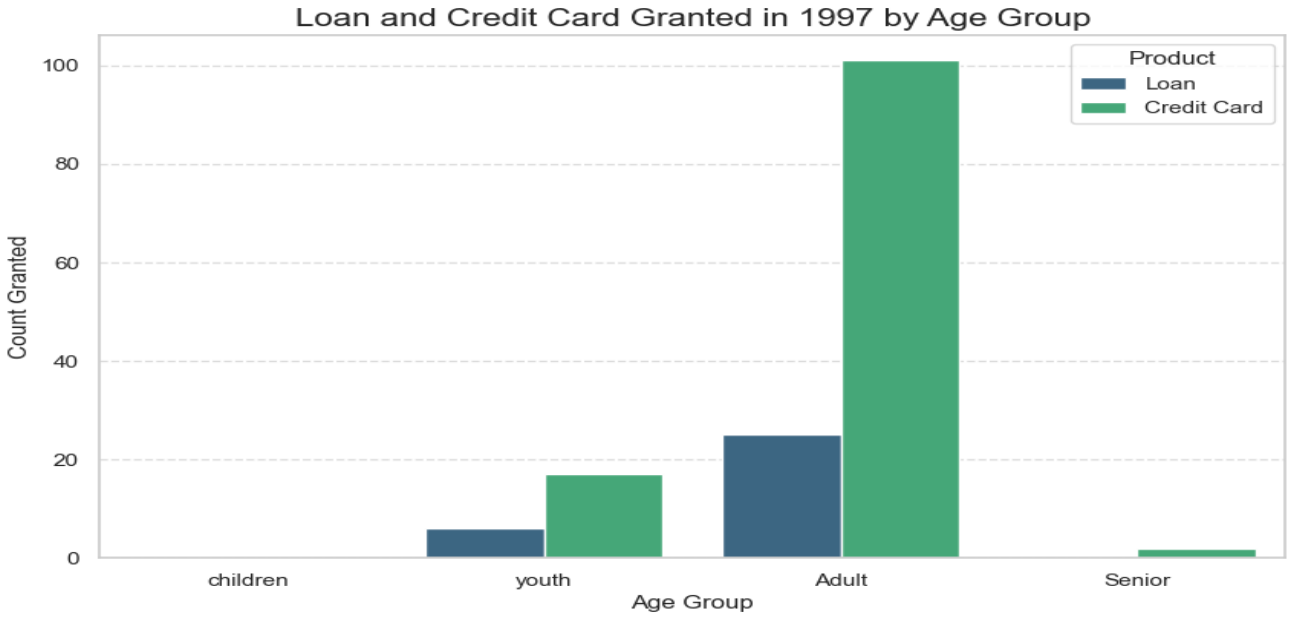
We can see that the median of avg balance for clients that are granted a loan is higher than that of clients that were not granted a loan for 1997.

5. LOR per loan granted in 1997



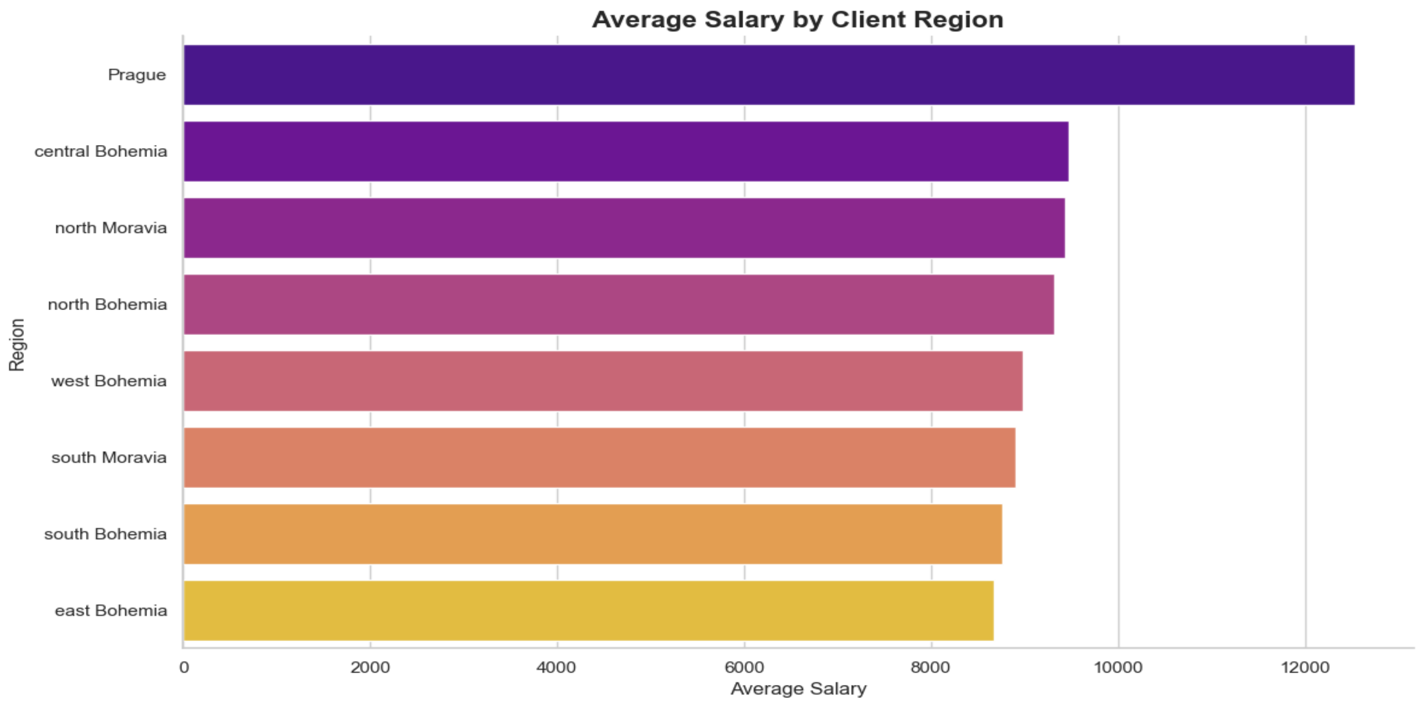
As shown earlier, most clients did not receive a loan in 1997. This graph further shows the clients that received a loan in 1997. Clients that created their accounts in 1995 are clients that were granted loans in 1997. The reason might be that they demanded loans after one year of opening their account. Moreover, the reason why the rest of the clients that have a high LOR were not granted a loan might be that they already have a loan before 1997 or they did not demand one.

6. Loan and Credit Card Granted in 1997 by age category



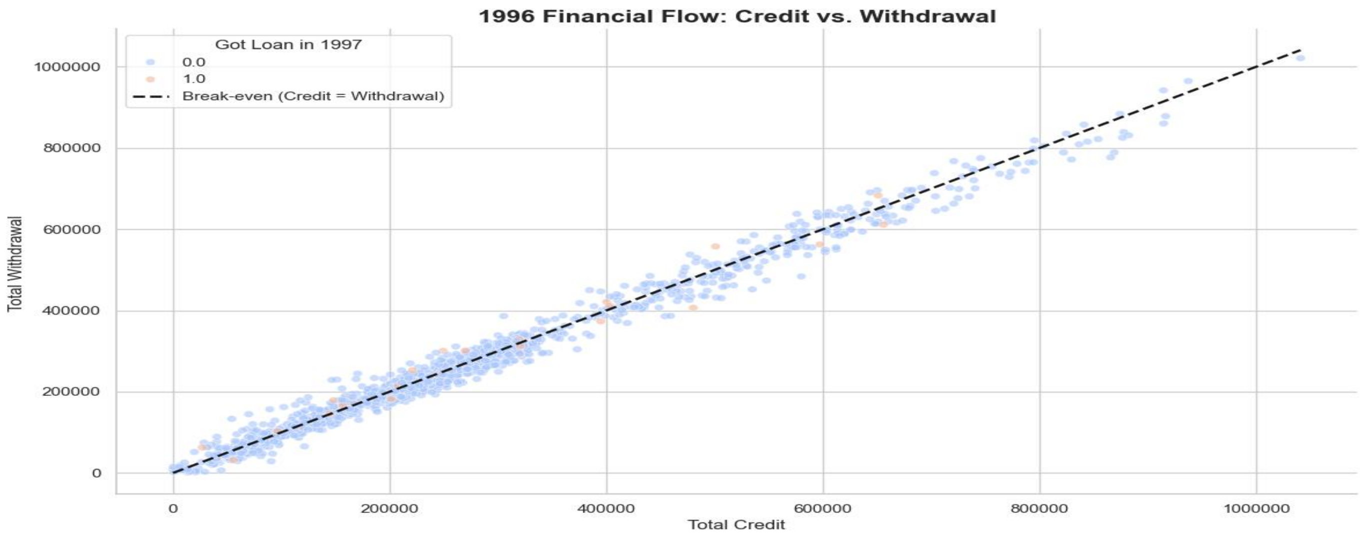
This shows that most loans and credit cards granted in 1997 are for adults followed by youth followed by seniors. We can further say that people that are aged between 25 and 65 are granted more loans and cards than other groups.

7. Avg Salary Per client region



This shows the avg salary per region of a client. Clients that has a higher avg in their region might be of a less risk. These kinds of clients are more likely to be rich and be able to repay the loan.

8. Credit vs withdrawal

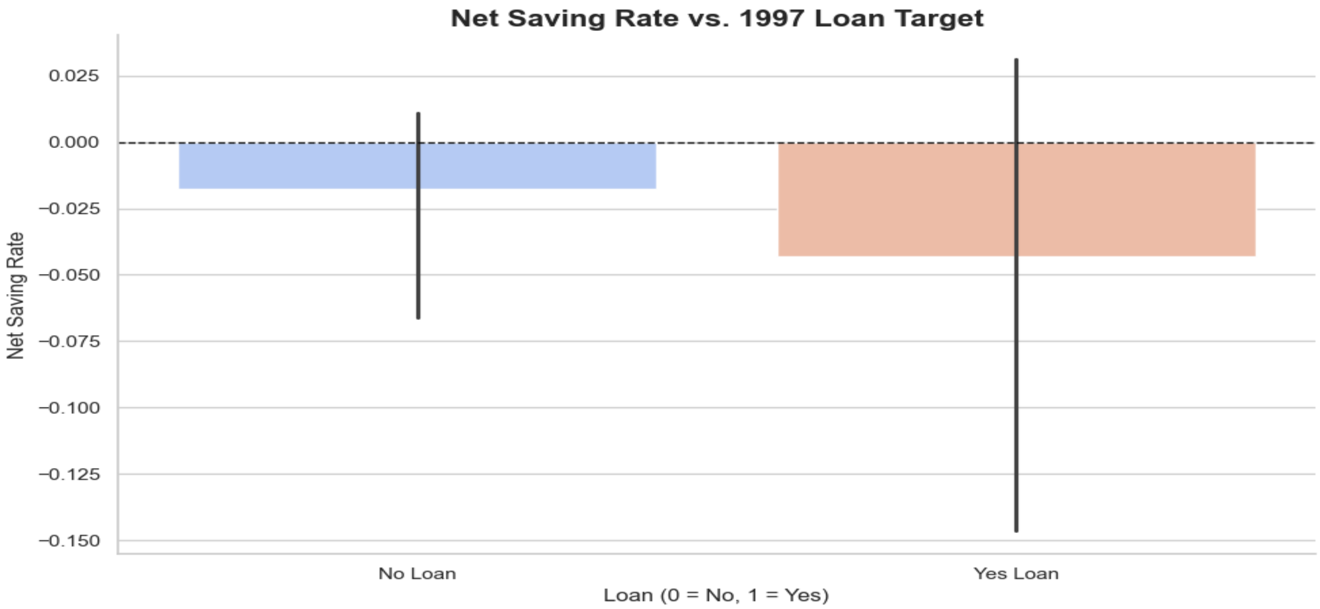


Clients that have higher credit tend to have higher withdrawal. There is a clear correlation between credit and debit.

The 45-degree Line (Break-even): The dashed black line represents where Money In = Money Out.

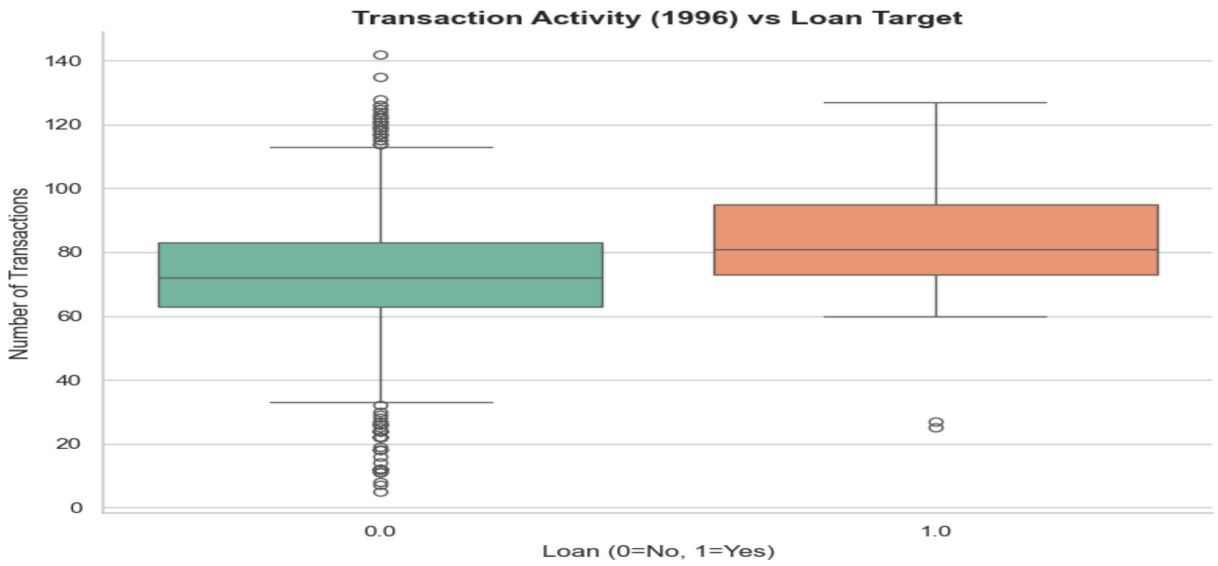
- **Points Below the Line:** Clients who saved money in 1996 (Credit > Withdrawal).
- **Points Above the Line:** Clients who spent more than they earned in 1996 (Withdrawal > Credit).

9. Net saving rate



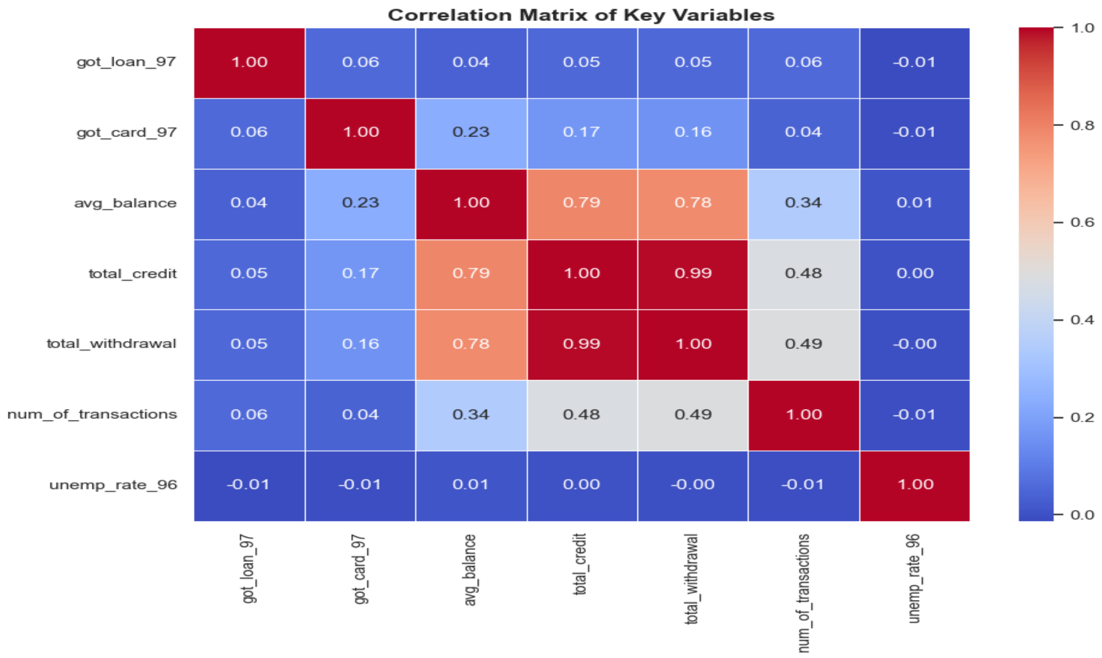
We can see that clients that actually received loans in 1997 tend to spend more than they save. So, this shows that people getting a loan tend. This might signify that People who save money (positive rate) generally don't need loans while people who overspend (negative rate) are demanding for
personal loans.

10. Transaction vs loan



The median for clients that were granted loans is higher than that for clients that were not granted a loan. Clients that were granted a loan in 1997 tend to have higher frequency in transactions in 1996.

11. Correlation Matrix



This correlation matrix shows a weak link between the two target variables. got_loan_97 and got_card_97 variables are not correlated with each other.

Got_loan_97 has week correlations with other variables (highest = 0.5).

got_card_97: Slightly better correlated with avg_balance (0.23); there is a clear correlation between avg_balance and got_card_97.

Balance vs. Credit/Withdrawal (~0.79): Wealthier accounts naturally have higher money flow. This is expected but confirms these variables are redundant.

Unemployment doesn't matter (statistically)
unemp_rate_96: The correlations are near zero (-0.01) for both got_loan_97 and got_card_97. This suggests that at an individual level, the regional unemployment rate has almost no impact on whether a specific client gets a product.

The key point to take from this correlation matrix is that there is a multicollinearity between total credit and total withdrawal. So, we can create one variable like credit/ debit or net saving ratio (like I did) to encompass both variables and drop one of the two columns. Adopting this method can lead to less distortion, which can allow for better training.

Summary

This Project highlights 70% of the work done by a data scientist which is understanding, preprocessing data and creating the base table. Creating the right features and imputing the missing values in the correct manner is crucial for the training of a model. I Created a base table that has a granularity level of client(owner), and added relevant features from different tables like transaction, district, loan, and card in the 1996-time window to help predict the granted loan and granted card in 1997.

I created visualizations that helped explain and represent the features that were created to further elaborate my work and check the correlation and significance between the IVs created and the DVs.