CREDIT EDA CASE STUDY

BY: - MOHAMMED HASHMATH ALI KHAN

ROHIT THANVI

OVERVIEW

INTRODUCTION:-

→THIS CASE STUDY AIMS TO GIVE YOU AN IDEA OF APPLYING EDA IN A REAL BUSINESS UNDERSTANDING.

→IN BANKING AND UNDERSTAND HOW DATA IS USED TO MINIMIZE THE RISK OF LOSING MONEY WHILE LENDING TO CUSTOMERS.

BUSINESS UNDERSTANDING:-

→THE LOAN PROVIDING COMPANIES FIND IT HARD TO GIVE LOANS TO THE PEOPLE DUE TO THEIR INSUFFICIENT OR NON-EXISTENT CREDIT HISTORY.

→THIS WILL ENSURE THAT THE APPLICANTS CAPABLE OF REPAYING THE LOAN ARE NOT REJECTED.

→WHEN THE COMPANY RECEIVES A LOAN APPLICATION, THE COMPANY HAS TO DECIDE FOR LOAN APPROVAL BASED ON THE APPLICANT'S PROFILE. TWO TYPES OF RISKS ARE ASSOCIATED WITH THE BANK'S DECISION:

→IF THE APPLICANT IS LIKELY TO REPAY THE LOAN, THEN NOT APPROVING THE LOAN RESULTS IN A LOSS OF BUSINESS TO THE COMPANY

→IF THE APPLICANT IS NOT LIKELY TO REPAY THE LOAN, I.E. HE/SHE IS LIKELY TO DEFAULT, THEN APPROVING THE LOAN MAY LEAD TO A FINANCIAL LOSS FOR THE COMPANY

Business Objectives

These are following Business objectives:-

→This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

→This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

→In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

There are four types of Decisions:-

→Approved

→Cancelled

→Refused

→Unused offer

DATA SETS

- THREE DATASETS HAS BEEN USED THROUGHOUT THIS ANALYSIS (CREDIT CASE STUDY).

→ **'APPLICATION_DATA.CSV':** WHICH CONTAINS ALL THE INFORMATION OF THE CLIENT AT THE TIME OF APPLICATION. THE DATA EXPLAINS THE PAYMENT DIFFICULTIES OF THE CLIENT.

→**'PREVIOUS_APPLICATION.CSV':** WHICH CONTAINS INFORMATION ABOUT THE CLIENT'S PREVIOUS LOAN DATA. IT CONTAINS THE DATA WHETHER THE PREVIOUS APPLICATION HAD BEEN APPROVED, CANCELLED, REFUSED OR UNUSED OFFER. •

→**'COLUMNS_DESCRIPTION.CSV':** WHICH CONTAINS DATA DICTIONARY WHICH DESCRIBES THE MEANING OF THE VARIABLES.

DATA EVALUATION

→Reading the data set ‘‘application _ data. Csv”

→Structure of the dataset

→checking the missing values

→Percentage of missing values.

→Drop columns with unimportant missing values.

→Checked the datatypes of all the columns and change the datatype.

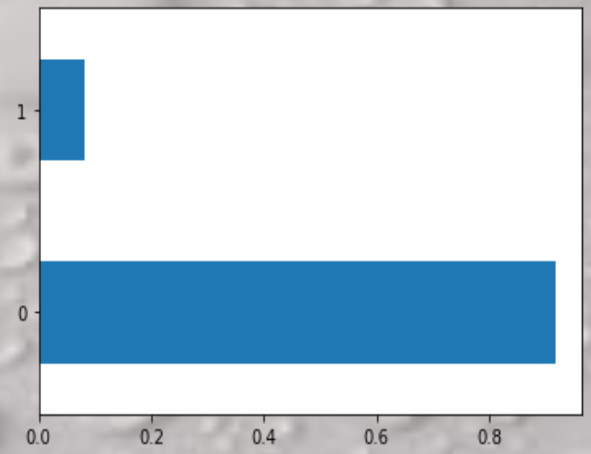
→For numerical columns check for the outliers and reported for 5 variables.

→Binning of the continuous variables.

→Checked if need to bin any variable in different categories

ANALYSIS OF “TARGET” VARIABLE

- The imbalance percentage for the 'target' variable with 1(defaulters) and 0(no defaulters) respectively is shown in the graph below:
- Therefore the data is imbalance
- We can say that 92% of data is non defaulter where as 8% are the defaulters.



UNIVARIATE ANALYSIS

FOR

CATEGORICAL VARIABLE

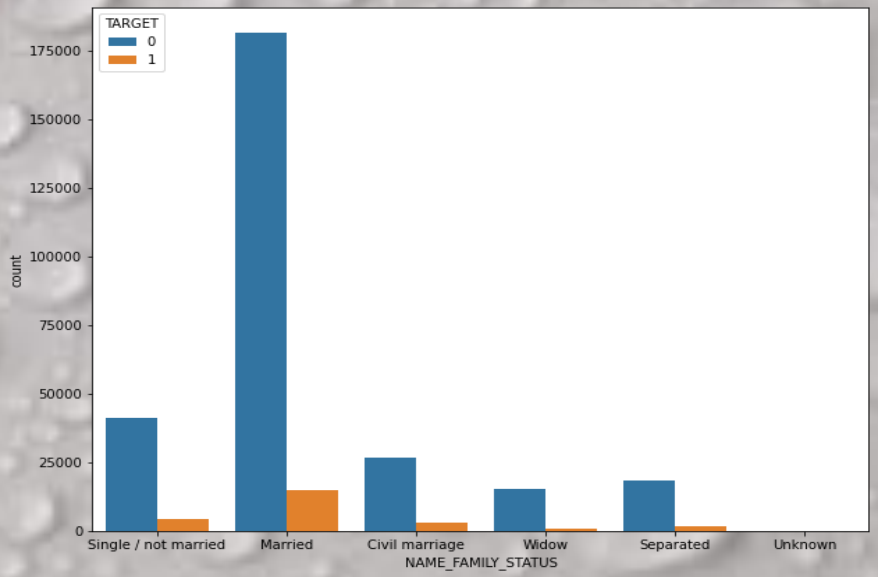
(NAME FAMILY STATUS: TARGET)

The following points are:

¬ Unmarried customers has less repayment difficulties

¬ But married customers are more favorable as we find the less ratio of defaulters in this category.

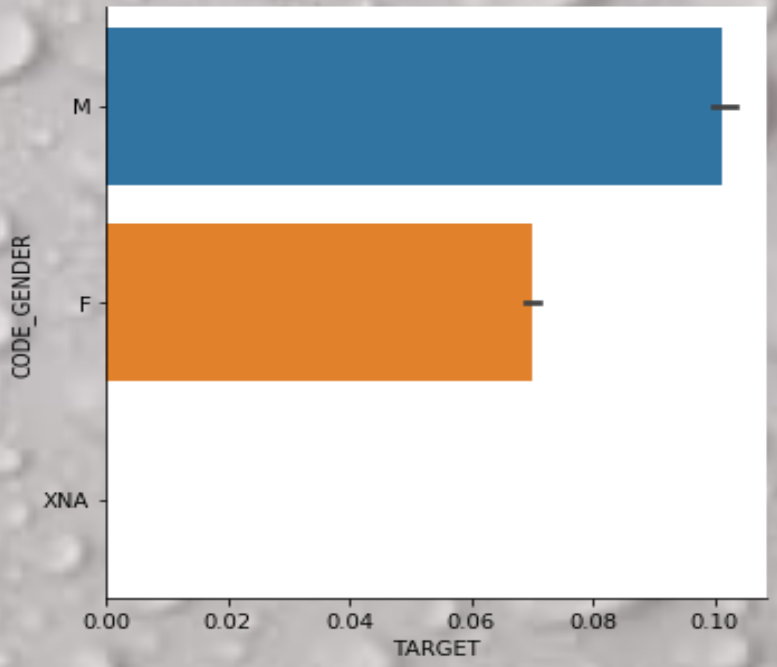
¬ Widow and separated are least favorable as their ratio of repayment is less



BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLE (CODE GENDER: TARGET)

→Blue Bar Represents TARGET-0 (Non-Defaulter) Orange Bar Represents TARGET-1 (Defaulter)

→When we take Gender into consideration, Male customers has less difficulties and repay loan on time. In contrast Female Customers have difficulties in repay the loan and they are the defaulters in this category.



UNIVARIATE ANALYSIS

FOR

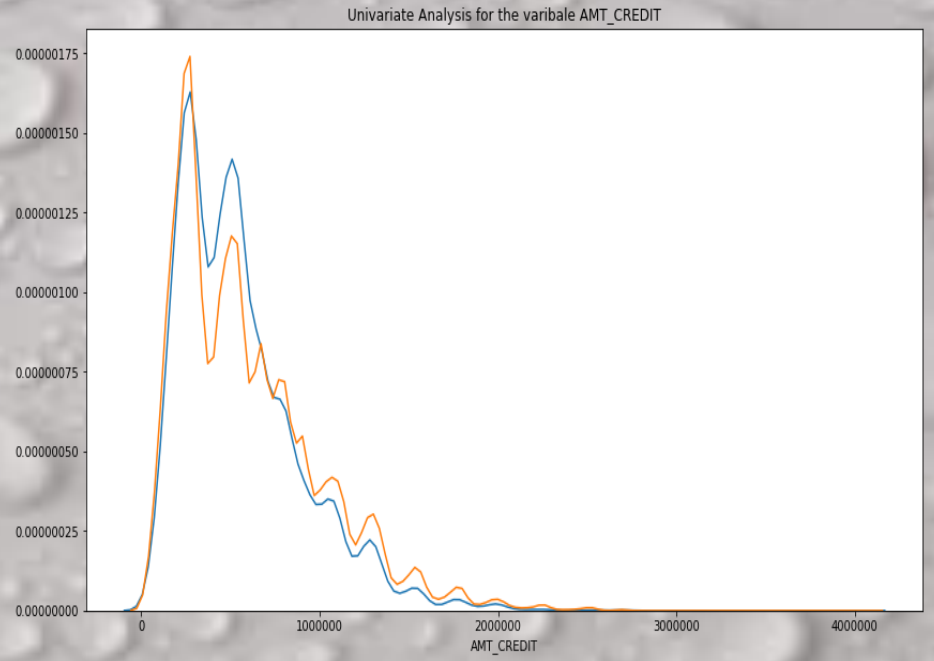
NUMERICAL VARIABLE (AMT CREDIT) W.R.T TARGET

→Blue line Represents TARGET-0

→Orange line Represents TARGET-1

On the base observation we can conclude that the frequency of credit amount between than 25, 00,000 - 70, 00,000 have more clients with payment difficulties compared to other clients.

For the credit amount less than 25, 00,000 or more than 70, 00, 00 have paid on time compared to clients who have payment difficulties.



BIVARIATE ANALYSIS

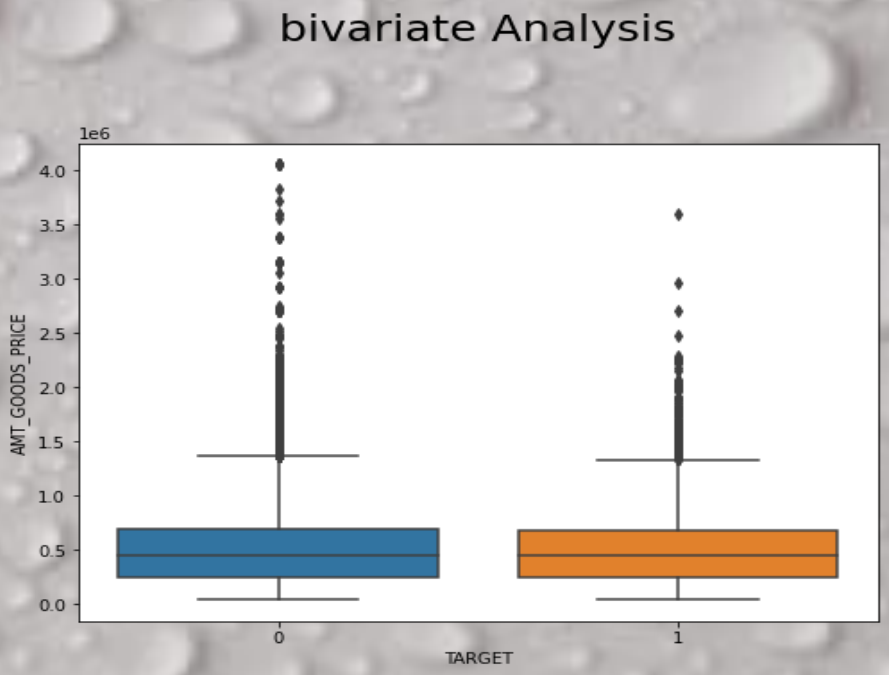
FOR

NUMERICAL VARIABLE

(AMT GOODS PRICE: TARGET)

- Blue line =TARGET-0
- Orange line =TARGET-1

- Min, Mean and Median , 1st Quartile, 2nd Quartile ,3rd Quartile and whiskers are the same but Max for the non-defaulters(0) are high. I.e. if Amount of the good price is high, it might be payment difficulty for the client.
- And The Defaulter -1 has outliers, which means it is difficult for the client to repay because of increase in good prices.



CORRELATION FOR THE CLIENT TARGET=0(NON-DEFAULTER)

Points to be concluded from the graph presented before.

Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.

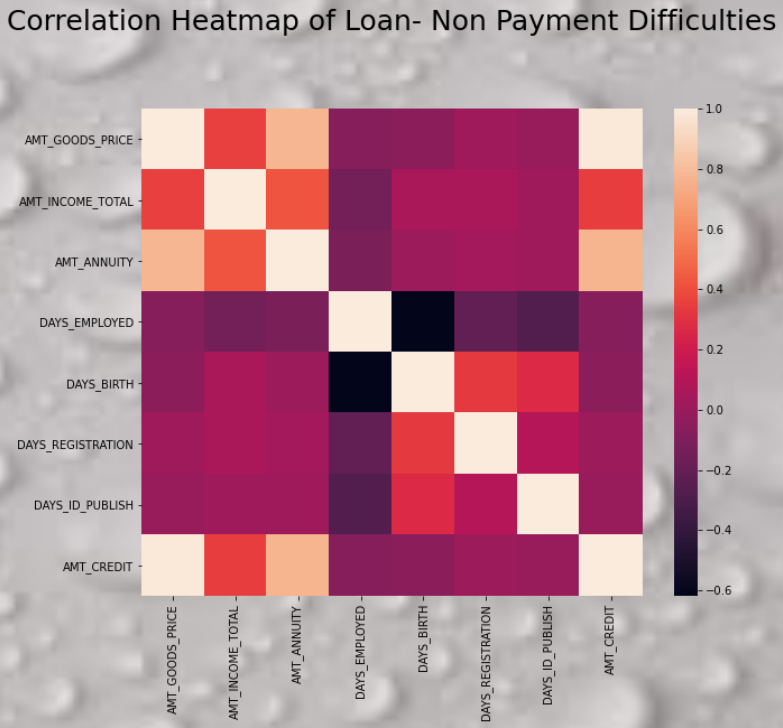
Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.

Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

Less children client have in densely populated area.

Credit amount is higher to densely populated area.

The income is also higher in densely populated area.



CORRELATION

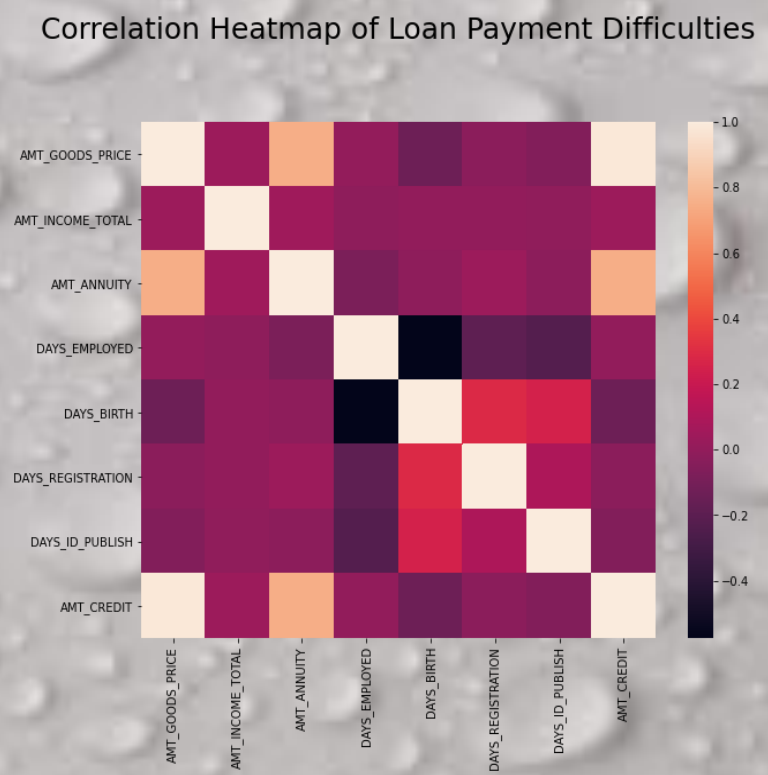
FOR

THE CLIENT TARGET=1(DEFAULTER)

→ This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different. They are listed below.

The client's permanent address does not match contact address are having less children and vice-versa

The client's permanent address does not match work address are having less children and vice-versa



CONTINUOUS UNIVARIATE ANALYSIS (W.R.T NAME CONTRACT STATUS)

- INFERENCES FROM THE PLOT:

Blue Bar Represents Loan Approved

Orange Bar Represents Loan Cancelled

Green Bar Represents Loan Refused

→Client with no children has high possibility of loan getting approved, it can be observed 5-6 Lakhs

Client who has no children got their loan approved from the bank.

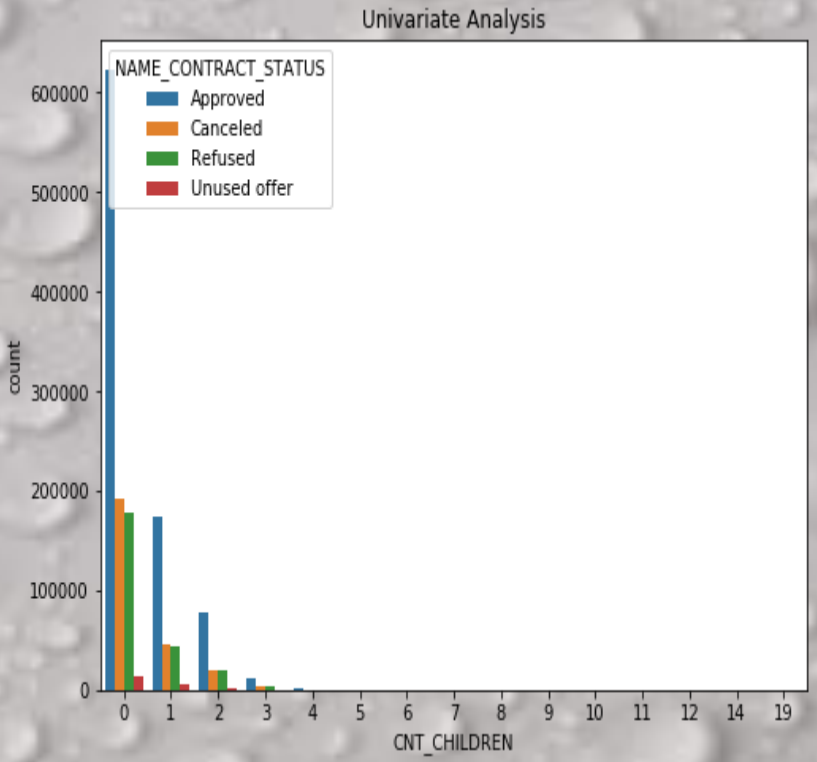
→Whereas Client having at least 1 or 2 Child has less

Frequency of approval i.e. 1-2 Lakhs Clients got their loan Approved.

→Therefore, It can be stated that clients with more

Number of children will have more payment difficulties

And less changes of loan being approved.



CATEGORICAL UNIVARIATE ANALYSIS (W.R.T NAME_CONTRACT_STATUS)

INFERENCES FROM THE PLOT:

Blue Bar Represents Loan Approved

Orange Bar Represents Loan Cancelled

Green Bar Represents Loan Refused

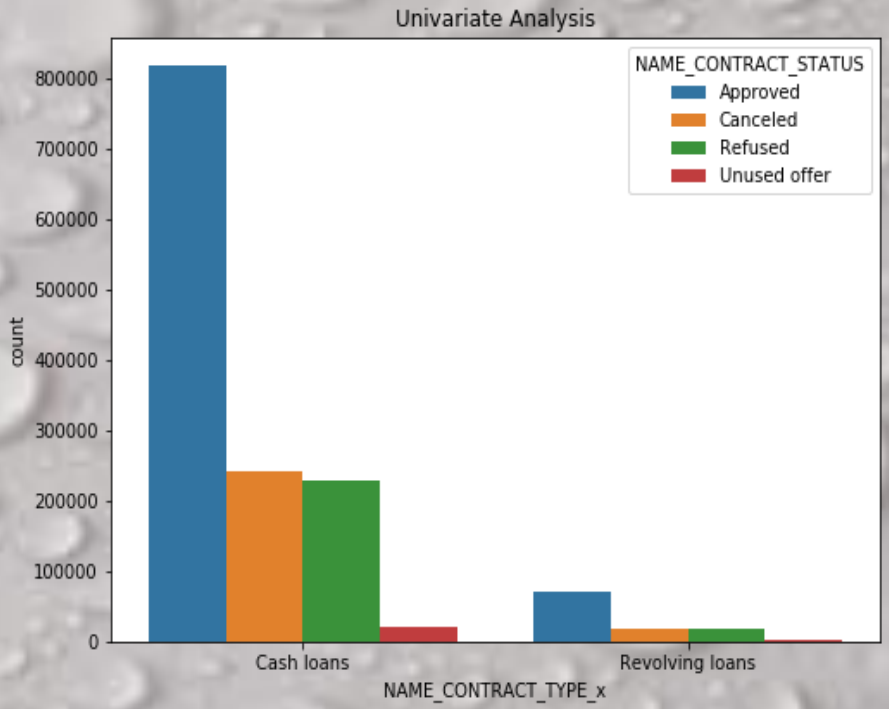
→It can be observed there are two types of loans

Cash Loans & Revolving Loans

Cash Loans: Which are fixed term loan have high

Possibility of getting approved, it can be inferred that 7-8 lakh of

Client are getting approved for cash loans.



Revolving Loans: Which are like credit card loans need to paid back on usage and reset the limit, has less frequency of Clients getting approval for it. Therefore Client opting for cash loans gets approval, other loans get rejected more Frequently

Conclusions:-

1.Bank can decrease the interest on the loan amount for the customers whose income range is low

2. For the consumers having high credit has less payment difficulties for which bank won't need to bear losses and can approve their loans.

3. Occupation type such as labors and sales staff are not recommended for loans whereas Accountants are recommended.4.Bank can see the family status i.e. families with more children will have high payment difficulties so can approve loans to unmarried or only couples.

5. Banks should focus more on contract type Student, pensioner and Businessman with housing 'type other than 'Co-op apartment' for successful payments.

6.Banks should focus less on income type 'Working' as they are having most number of unsuccessful payments.

INSIGHTS:-

1. Amount credit- Previous has highest refused cases and amt_credit_Current is similar for all 4 cases.
2. Time spent in unused offer is higher as compared to other categories.
3. So bank should reduce time spent on unused offer.
4. Nuclear family (2-3 people in family) get highest approval.
5. Previously most of the applications were cancelled or refused

Recommendation:-

Variable that can be used by the bank to approve or reject the loan from the data set.

1. TARGERT
2. AMT_INCOME_TOTAL
3. AMT_CREDIT
4. AMT_ANNUITY
5. NAME_FAMILY_STATUS
6. NAME_INCOME_TYPE
7. NAME_HOUSING_TYPE
8. AMT_REQ_CREDIT_BUREAU_MON
9. AMT_REQ_CREDIT_BUREAU_YEAR
10. NAME_CONTRACT_TYPE
11. AMT_ANNUITY
12. AMT_APPLICATION
13. AMT_CREDIT
14. AMT_DOWN_PAYMENT
15. NAME_CONTRACT_STATUS
16. NAME_CLIENT_TYPE

THANK YOU