IDENTIFY PATTERNS INDICATING CLIENTS WHO MIGHT HAVE DIFFICULTY PAYING THEIR INSTALLMENTS

Project made by: Mai Huy Khang

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1. Context

Upon receiving a loan application, the company must assess the applicant's profile to decide on loan approval, which involves two types of risks:

- Not approving the loan for an applicant likely to repay would result in a loss of business for the company.
- Approving the loan for an applicant likely to default would lead to a financial loss

The provided data contains information about loan applications at the time of applying, categorized into two scenarios:

- Clients facing payment difficulties: These individuals have a history of late payments for more than X days on at least one of the first Y installments in the loan sample.

All other cases: This category includes clients who made timely payments.

When a client applies for a loan, there are four possible outcomes:

- Approved: The company approves the loan application.
- Cancelled: The client cancels the application either due to a change of mind or receiving unfavorable terms.
- Refused: The company rejects the loan application because the client does not meet their requirements.
- Unused offer: The client cancels the loan at different stages of the process.

1.2. Determine the main questions, input and output expectation

❖ The main objective:

To identify patterns indicating clients who might have difficulty paying their installments.

This information will be used to make informed decisions, such as denying loans, reducing loan amounts, or offering higher interest rates to riskier applicants. Understanding the factors driving loan defaults will help the company manage its portfolio and assess risks effectively

❖ Input data

The dataset comprises 3 files:

1. 'application_data.csv': This file contains all client information at the time of loan application, including whether the client experiences payment difficulties.

2. 'previous_application.csv': Contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. 'columns_description.csv': This file serves as a data dictionary, providing descriptions and meanings of the variables used in the dataset

❖ **Detail output and insight questions**

Question 1: What are contributors that had effect to the target output in the dataset? Listing out the contributors?

Question 2: Comparing and indicating the main features of those variances that lead to customers' difficulty in paying their installments?

Question 3: Is there any relevant to the previous application to cause the customers' difficulty in paying their installments?

Question 4: Ranking the level of influences and showing the top causes to the main output?

CHAPTER 2: DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

2.1. Data preprocessing

The main dataset table in this report has 122 features and over 300 thousand records about client information at the time of loan application.

Overall, all features fit their right datatype and there are no duplicates records. However, the statistics show that almost all the records have null values that need to be fixed.

For further information, I recognize that there are positive values in the DAY_EMPLOYEE which are supposed to be null or negative and so that not makes sense in this column context.

After summarizing errors, data preprocessing was performed as following:

- Removing SK_ID_CURR column (for correlation check later)
- Divide dataset into categorical (nominal or ordered data type) and numeric features
- Fill null value in categorical (by Unknown) and numeric (by mean value) features
- Remove records with condition of DAYS_EMPLOYED is '365243'
- Encode categorical features using the get_dummies function (for correlation check later)

2.2. Exploratory data analysis

2.2.1. Correlation analytic

The correlation relationship of the variables will give us an overview of the impact between variables in dataset. I use this relationship to classify which variables have an impact toward the output. And those results will support my further analysis.

TARGET	1.000000
CNT_CHILDREN	0.007741
AMT_INCOME_TOTAL	-0.007332
AMT_CREDIT	-0.040659
AMT_ANNUITY	-0.020040
AMT_GOODS_PRICE	-0.050320
REGION_POPULATION_RELATIVE	-0.040640
DAYS_BIRTH	0.065756
DAYS_EMPLOYED	0.074958
DAYS_REGISTRATION	0.036087
DAYS_ID_PUBLISH	0.039910
OWN_CAR_AGE	0.022366
CNT_FAM_MEMBERS	-0.001824
HOUR_APPR_PROCESS_START	-0.031099
EXT_SOURCE_1	-0.102369
EXT_SOURCE_2	-0.170356
EXT_SOURCE_3	-0.158249
APARTMENTS_AVG	-0.021957
BASEMENTAREA_AVG	-0.015161

The recausedicate that own_car_age, days_registration, days_id_publish, days_birth, days_employed, region_rating_client, region_rating_client_w_city, days_last_phone_change, def_30_cnt_social_circle, def_60_cnt_social_circle have positive correlation to the output, which means the increase trend in those elements might cause to customer's difficulty in paying their installments in general.

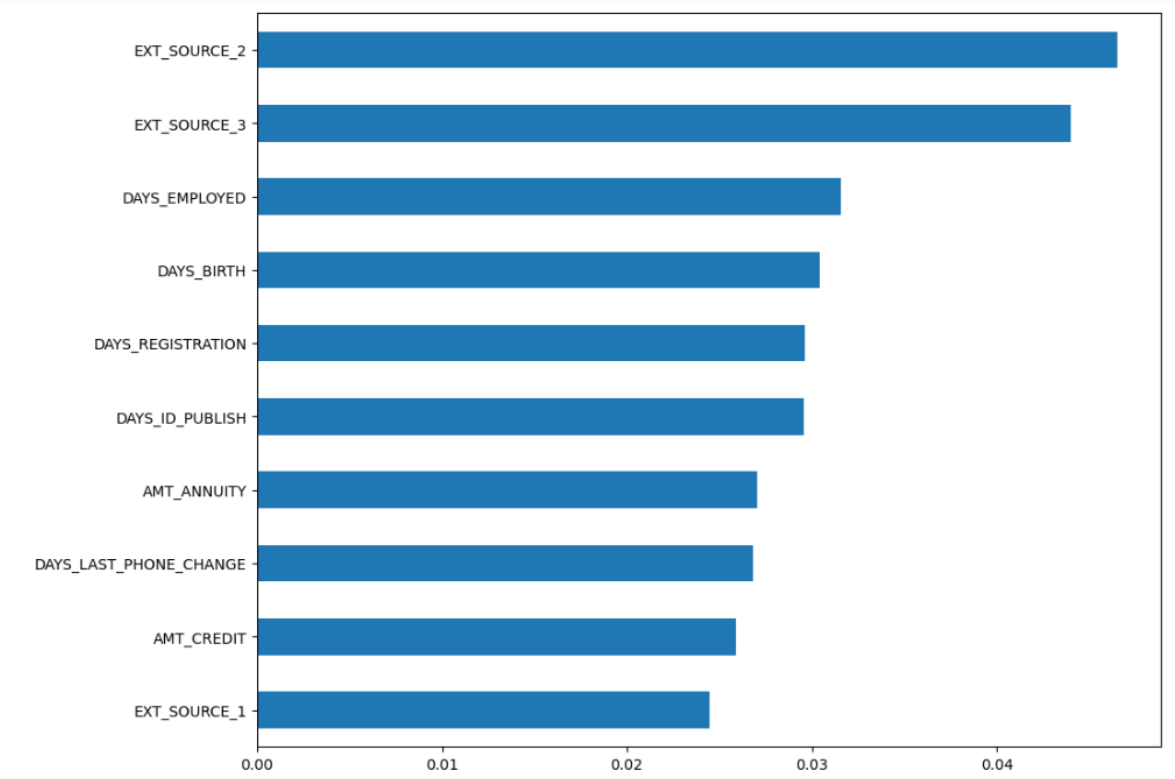
On customers'hand, some other contributors such as flag_own_car, amt_credit, amt_goods_price, region_population_relative, flag_phone, reg_city_not_live_city, reg_city_not_work_city, ext_source_1, ext_source_2, ext_source_3 seem to be negative impact to the output, which means the decrease trend in those might cause to customer's difficulty in paying their installments in general.

On top of that, name_contract_type, name_income_type, name_education_type, name_family_status, name_housing_type, occupation_type reveal the signal of some label patterns that can recognize customer's difficulty in paying their installments.

2.2.2. Main influence feature toward output in dataset

"Feature importance" refers to the process of determining the relative importance or contribution of different features in a dataset towards predicting a target variable.

In this report, I calculate feature importance using Random Forest Classifier for some benefits including low-linearity results above, handling irrelevant features and robustness to overfitting dataset. The picture below shows the descending effects toward output of top 10 features.



From the ranking list, although the level of effects is not clear enough, we can see that normalized score from external data sources 2 and 3 get the top peak of influencers with big gap compared to other features. Followed by customer's age and employee's experiences.

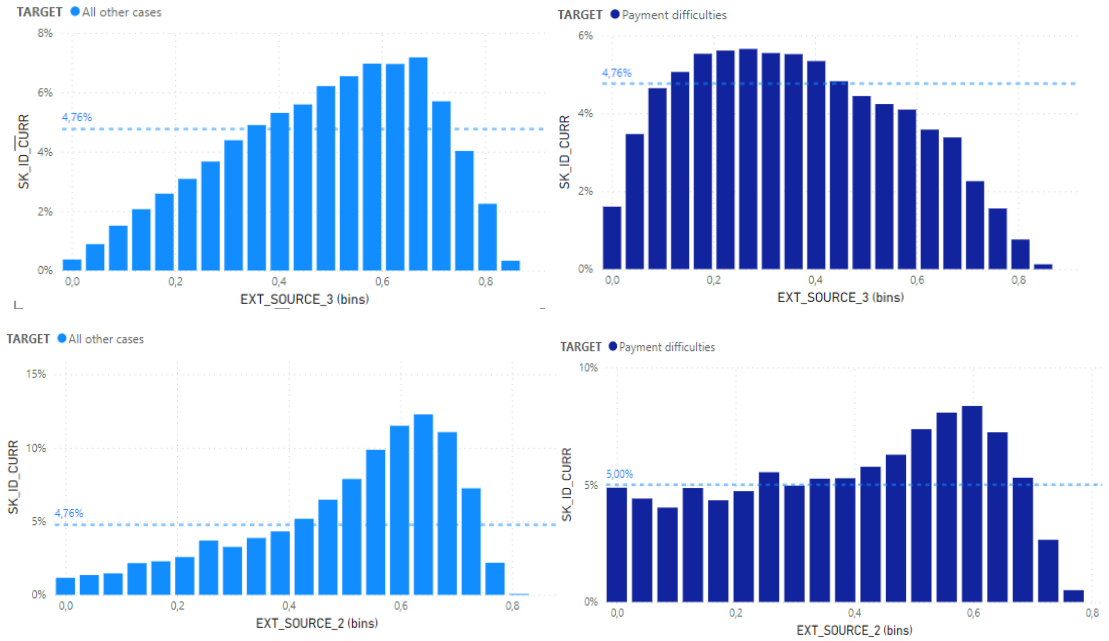
2.2.3. Detail feature patterns that lead to clients' difficulty paying their installments

From the result in feature importance list, we can identify all influencer to the output. However, there are just some factors that show the impacts clearly and so that can be recognize from visualization.

❖ **Contract status of previous application**

From the graph above, we can figure out that although both outputs have all kinds of contract status compared to previous data, customers facing payment difficulty have a greater proportion on refused and canceled contract. It turns down that when customers were refused and canceled in the past, they might have payment difficulty at the current application.

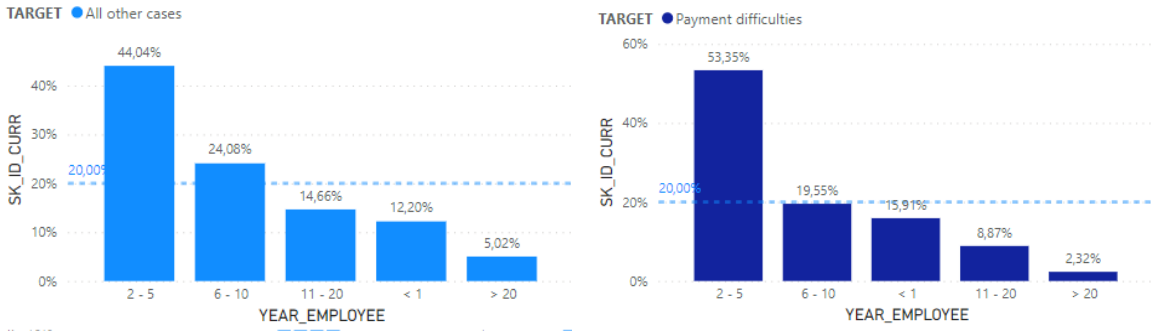
❖ Normalized score from external data source



The chart above shows that, from ext source 3, customers facing payment difficulty is usually only rated below 0.4 while others are rated between 0.45 and 0.72.

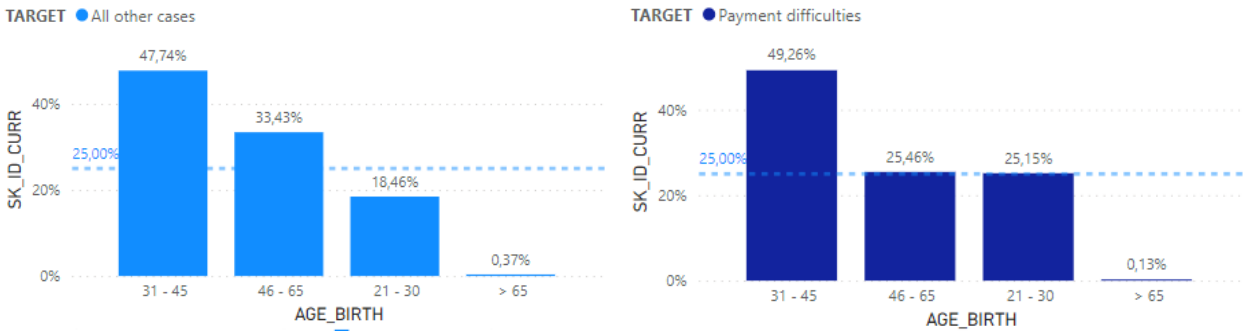
On the top of that, from ext source 2, customer facing payment difficulty has a fairly uniform distribution but still accounts for over 60% below 0.4 points, otherwise customers in other cases accounted for more than 66% scored from 0.4 up to 0.7

❖ Client's experiences in years at the time of application



In general, the majority of customers at the company have less than 10 years of experience. Furthermore, when analyzing in more detail, it can be seen that customers with payment difficulty have the below 1-year employee group accounts for nearly 16%. Whilst the other cases make up that element just around 12% and have high percentage of 11-20 years group at almost 15%

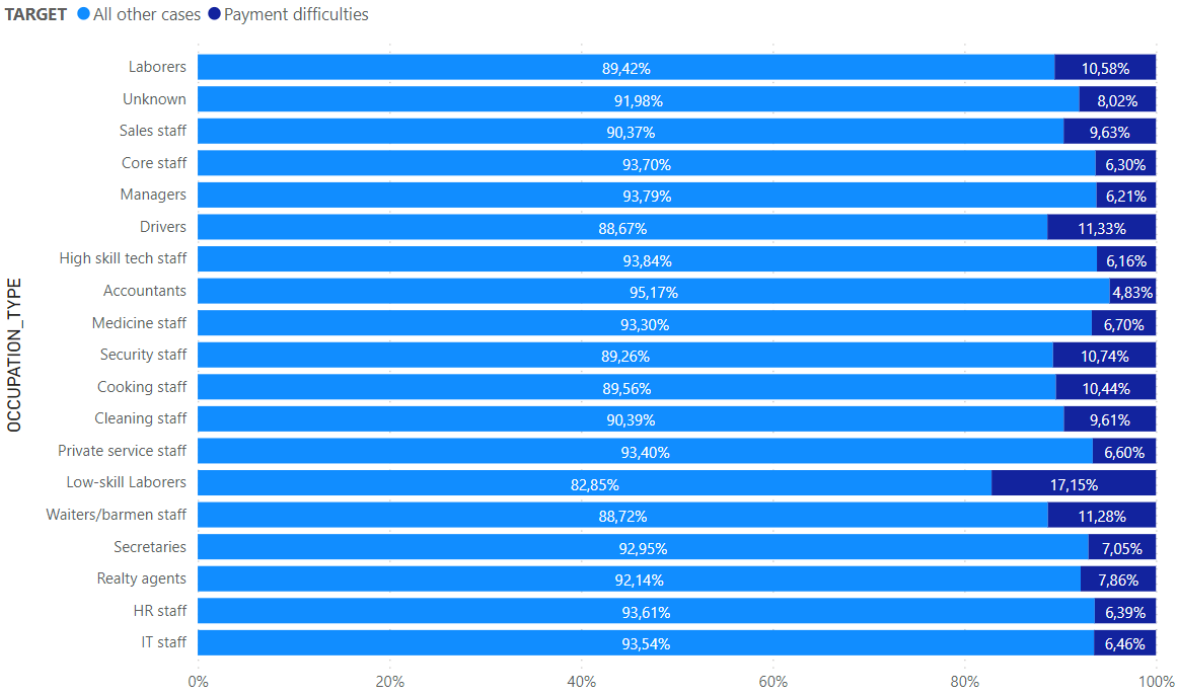
❖ Client's age in years at the time of application



For client's age, customers have payment difficulty take slightly greater proportion in 21-30 age and downtrend in 46-65 age group

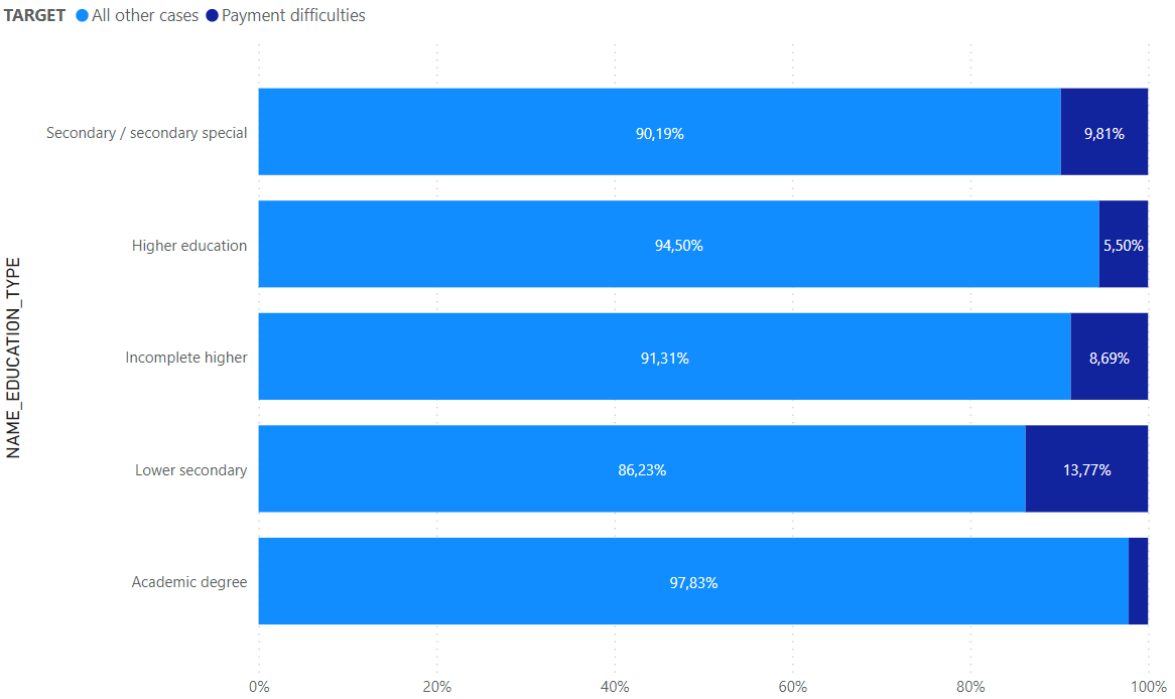
❖ Other patterns from geographical features

Occupation type



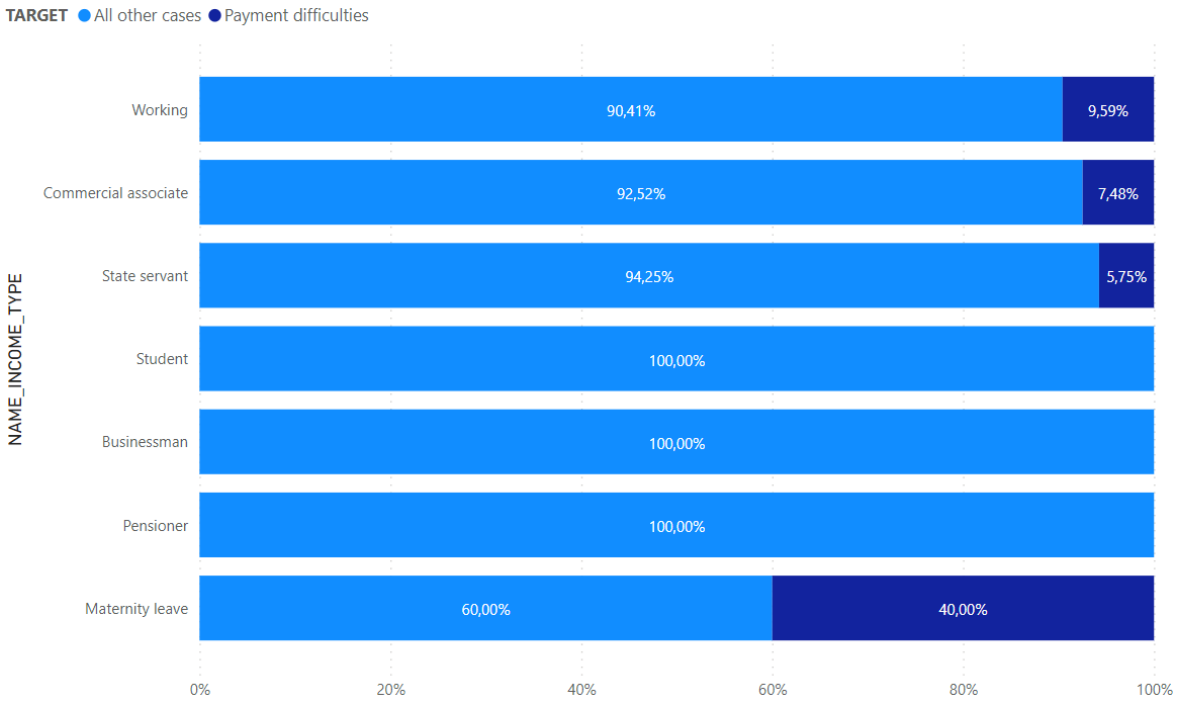
In some context, the company can recognize customers with payment difficulty through occupation such as low-skill laborers, drivers and waiters/barmen staffs as a signal for the output

Education type



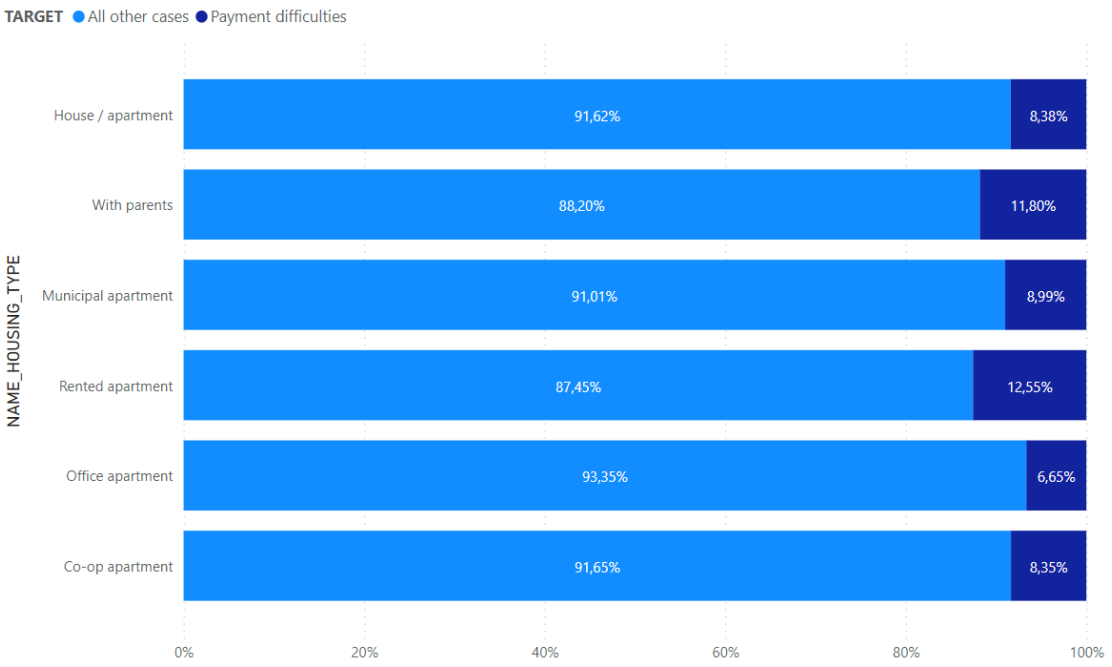
Similarly, education types also reveal some signals to point out the customers with payment difficulty including who from lower secondary and secondary background.

Income type



Customers with payment difficulty show a significant proportion in maternity leave by almost 40% compared to the other cases

Housing type



The rented apartment and with parents group show a better scaler in their group compared to other types at 12.5% and 11,8% respectively. So, customers with payment difficulty might fall into those types.

CHAPTER 3: CONCLUSION

After analyzing, I must affirm that the impacts of features to the output is not clear enough to classify, except ext_sources rating, which are greatest measure for company to define customers facing payment difficulty.

Feature	Patterns to indicate customers facing payment difficulty
Ext_sources rating	○ Lower 0.4 score
Contract status of previous application	○ Refused ○ Canceled
Client's experiences in years at the time of application	○ Below 1-year
Client's age in years at the time of application	○ 21-30 age
Occupation type	○ Low-skill laborers ○ Drivers ○ Waiters/barmen staffs
Education type	○ Lower secondary ○ Secondary school
Income type	○ Maternity leave
Housing type	○ Rented apartment ○ With parents