2019

IDM PROJECT

INSTRUCTOR: DR. SAJJAD HAIDER

MUHAMMAD FARMAL KHAN
MOHAMMAD SAAD MAQSOOD
NIKHIL SATIANI

Problem Description

Home Credit B.V. is an international non-bank financial institution founded in 1997 in the Czech Republic and headquartered in Netherlands. The company operates in 10 countries and focuses on lending primarily to people with little or no credit history.

The company has posted their dataset on Kaggle with the challenge to predict how capable a current applicant is of repaying a loan which will help them decide whether to give them a loan or not.

As an additional feature, we have performed customer profiling as well through clustering like differentiating between First Time Defaulters, Lazy customers, Point of No Return Customers etc. because it helps the company to divide the customers into different segment and recommend differential treatment to each segment to maximize their recoveries and reduce cost.

Data Description:

For our ease, they have divided the dataset into two groups, train and test, to build and improve our model around them.

The datasets contain:

Column	Description
SK_ID_CURR	ID of loan in our sample
TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_INCOME_TOTAL	Income of the client
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave…)
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents? ...)
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
DAYS_BIRTH	Client's age in days at the time of application

DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_REGISTRATION	How many days before the application did client change his registration
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
OWN_CAR_AGE	Age of client's car
FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL	Did client provide email (1=YES, 0=NO)
OCCUPATION_TYPE	What kind of occupation does the client have
CNT_FAM_MEMBERS	How many family members does client have
REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
ORGANIZATION_TYPE	Type of organization where client works
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
APARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

ENTRANCES_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

FONDKAPREMONT_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
HOUSETYPE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
TOTALAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
WALLSMATERIAL_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
EMERGENCYSTATE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
OBS_30_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings with observable 30 DPD (days past due) default
DEF_30_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings with observable 60 DPD (days past due) default
DEF_60_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings defaulted on 60 (days past due) DPD
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
FLAG_DOCUMENT_2	Did client provide document 2
FLAG_DOCUMENT_3	Did client provide document 3
FLAG_DOCUMENT_4	Did client provide document 4
FLAG_DOCUMENT_5	Did client provide document 5
FLAG_DOCUMENT_6	Did client provide document 6
FLAG_DOCUMENT_7	Did client provide document 7
FLAG_DOCUMENT_8	Did client provide document 8
FLAG_DOCUMENT_9	Did client provide document 9
FLAG_DOCUMENT_10	Did client provide document 10
FLAG_DOCUMENT_11	Did client provide document 11
FLAG_DOCUMENT_12	Did client provide document 12
FLAG_DOCUMENT_13	Did client provide document 13
FLAG_DOCUMENT_14	Did client provide document 14
FLAG_DOCUMENT_15	Did client provide document 15
FLAG_DOCUMENT_16	Did client provide document 16
FLAG_DOCUMENT_17	Did client provide document 17
FLAG_DOCUMENT_18	Did client provide document 18
FLAG_DOCUMENT_19	Did client provide document 19
FLAG_DOCUMENT_20	Did client provide document 20
FLAG_DOCUMENT_21	Did client provide document 21
AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)

AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

Data Preparation

- Used Label Encoding and One-Hot Encoding for categorical variables as machine learning models work best on numbers.
- The missing values are replaced by the calculated mean of the respective column.
- The values NOT normalized are scaled between 0-1.
- Applied PCA to reduce the dimensionality of the dataset.

Model Building and Evaluation

For **classification** the models used were:

1. Gradient Boosting
2. Random Forrest
3. Decision Tree
4. Logistic Regression

Best model

- Logistic Regression
- Accuracy: 0.729 (on test data)
- Parameters: c = 1
- Time: approximately 1 min (for both training and predictions)

For **clustering:**

- Algorithm: K-Means
- Evaluation: Elbow Method (to find optimal number of clusters)
- Time: 2-3 mins

Device Specifications:

- Model: Haier Y11C
- Processor: Intel ® Core ™ m3-7Y30 CPU @ 1.00 GHz
- RAM: 8 GB
- System Type: 64-bit OS, processor

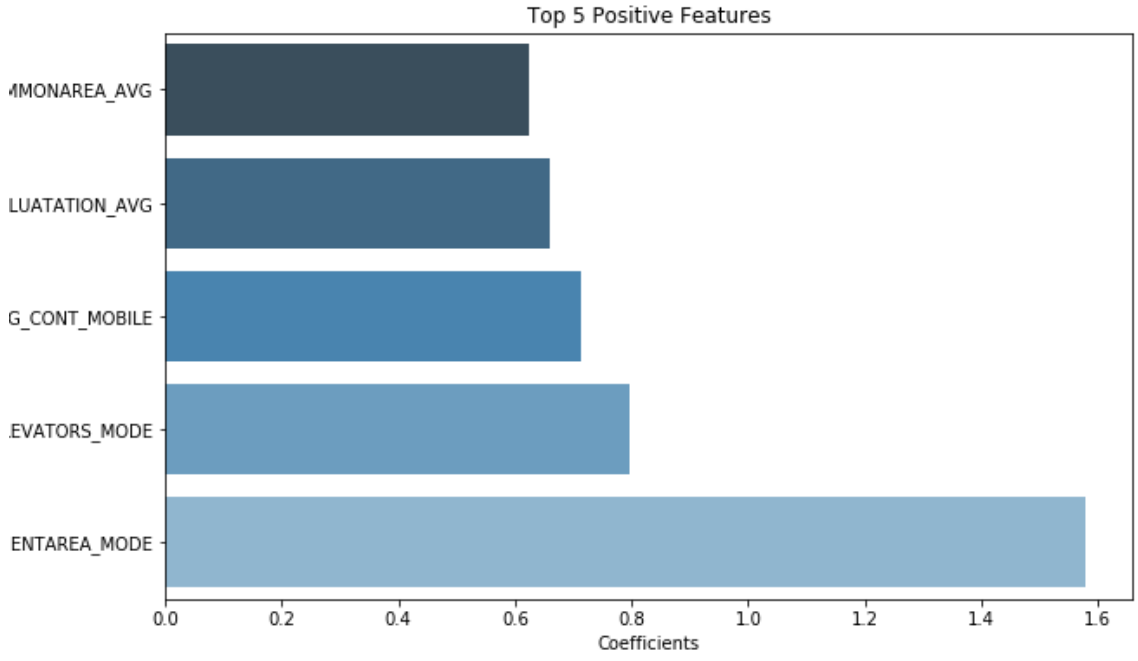
Software and Library:

- Python and Jupyter Notebook
- Libraries: pandas, sklearn, matplotlib (for visualizations)
- Knime

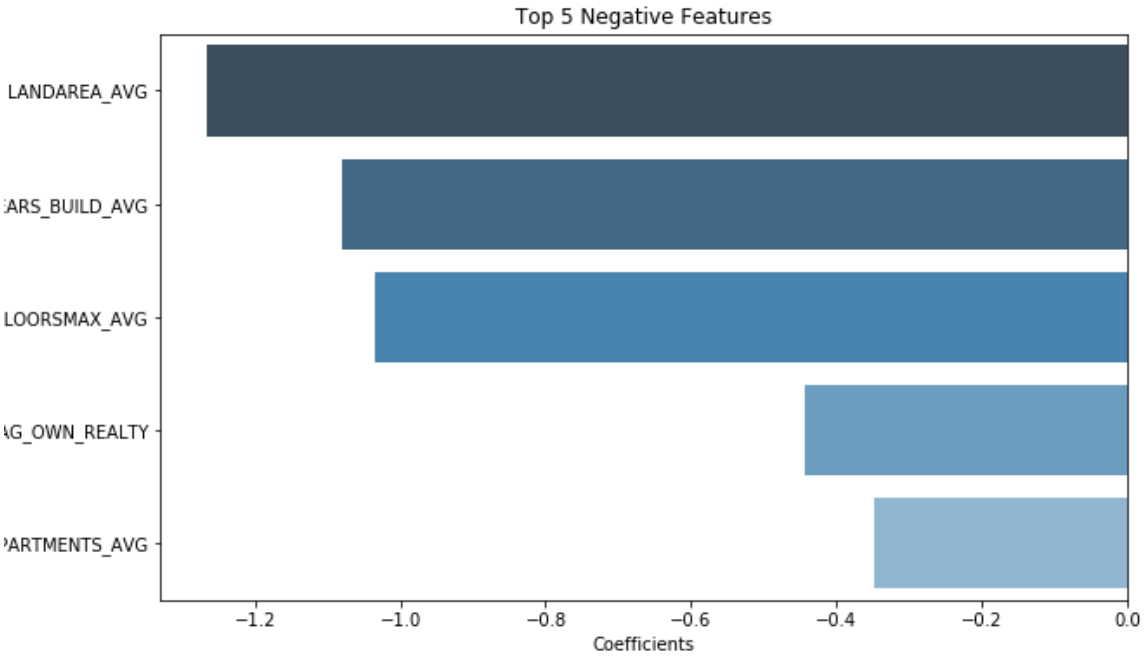
Data Insights

Classification:

Since the logistic regression model provided as with the best accuracy score for classification, we used its descriptive features to identify which factors influence the most on the ability of a customer to repay his/her loan.



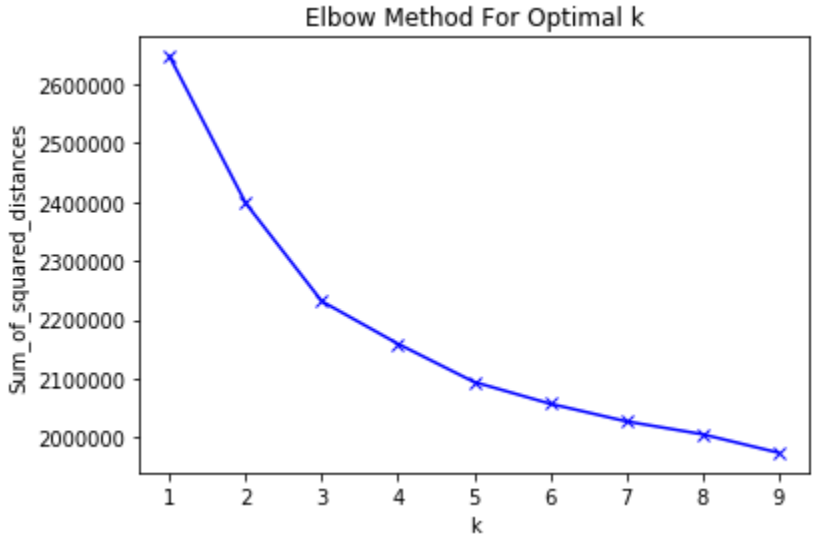
These are the top 5 features that contribute to target value – 1 (will have difficulty repaying loan). The company can make these features as red flags for future customers as a warning if these feature values are too high.

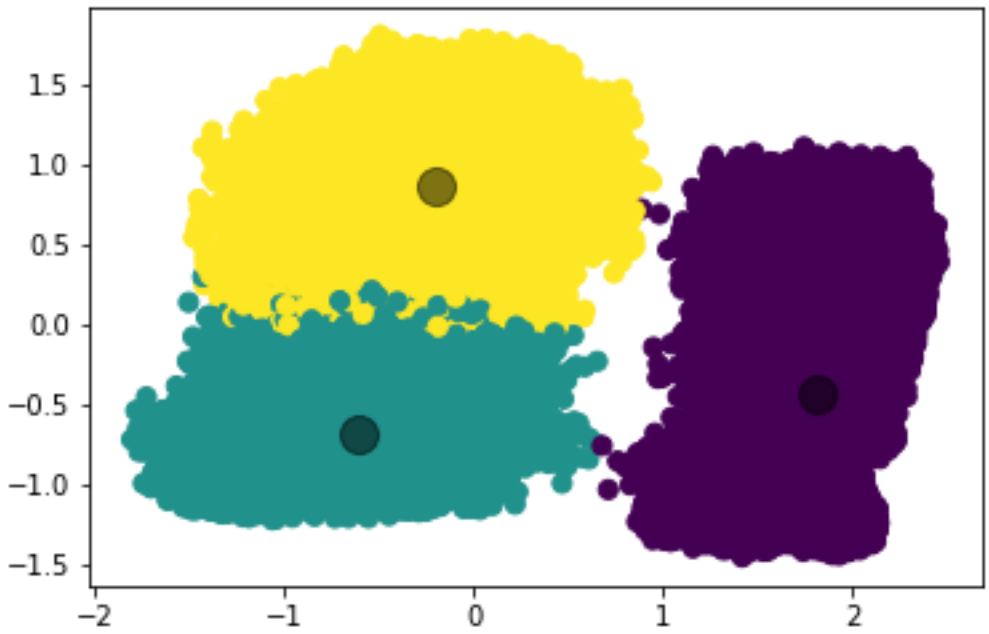


These features contribute the most to the target value – 0 (will repay on time) and can be deemed as good signs in a customer.

Clustering:

In K-Means clustering, the elbow method showed k=3 is the optimal number of clusters.





Conclusion: The current applicants of Home Credit B.V can be divided into three types of customers according to their past and present data provided.

Limitations of Study

Dataset Problem:

The amount of missing data in columns was huge. For future reference the company can provide some reason for the missing values, so we can prepare data accordingly.

Model Expiration:

Some of the features in the data are associated with their time of recording. For example AMT_INCOME_TOTAL, REGION_RATING_CLIENT, REGION_POPULATION_RELATIVE that indicate income of the clients, rating of the regions they live in and relative population of the regions they belong to respectively, are expected to change with time like average income might go up with inflation, population might increase or decrease drastically in a region in just a few years etc. So, when this data becomes obsolete or a less accurate representation of reality then our model will expire as well. There are also some features that reflect the culture and trends of the society like FLAG_PHONE, CNT_FAM_MEMBERS, OCCUPATION_TYPE that indicate whether the

clients have given their home phone numbers, how many members are there in their family and their job category respectively, will change with a shift in culture, norms and trends. It is tough to say that if a client with a home phone number today is likely to repay the loan then the same would be true 10 years later. Home phones may not even be that common by that time given the extensive use of cell phones. Thus our model is dependent on how long does the data representing such factors remains valid.