A predictor of Customers who will get into Default on Home Credit’s Mortgages Contracts

# Abstract

## Business Domain

The project is aimed to analyse if it is possible, with some level of confidence, to predict if a person will be able to pay his/her mortgage, based on data such as :

* Personal profile of who is contracting the mortgage, such marital status, age;
* Finance status of the customer : if He/She has a car, if has a mobile phone, is employed, his/her income level, if He/She has a previous loan, his/her credit score, how many enquires to the credit bureau were done by the customer.
* Contact history: if He/She answered the phone from the finance institution when contacted
* Some of the characteristics of the house been mortgaged, such as type, area measures and price.
* Some regional data, such as population from the region, city, level of Default in that region

As if it is possible to find good correlations between some of these variables and the actual data that shows if a customer got into Default, the finance institution that hold the data can effectively make actions to anticipate a Default situation, like make provisions or actively renegotiate a contract extension before the Default happens.

Extrapolating this model which is based on already customers who contracted the mortgage, the company could understand which variables are more impacting a customer to get into Default. By knowing these variables, the company can improve its risk evaluation methods for granting mortgages and setting proper interest of it.

## Framing the Problem

As by exploring the problem and potential datasets, this project will try to answer the following questions :

* Is there a set of variables which can be used, with some level of confidence, to predict if a mortgage contract customer will get into a Default situation ?
* Is a Default situation much more related to the customer payment history or it is affected by some other factors such as the type of the house been mortgaged, marital status of the customer, his/her occupation, his/her education level and some other non-finance data ?
* How confident a finance institution can be, by applying such method ?
* For a finance institution dealing with a large number of variables / data from its mortgage customers, which ones are more relevant to determine the outcome ?

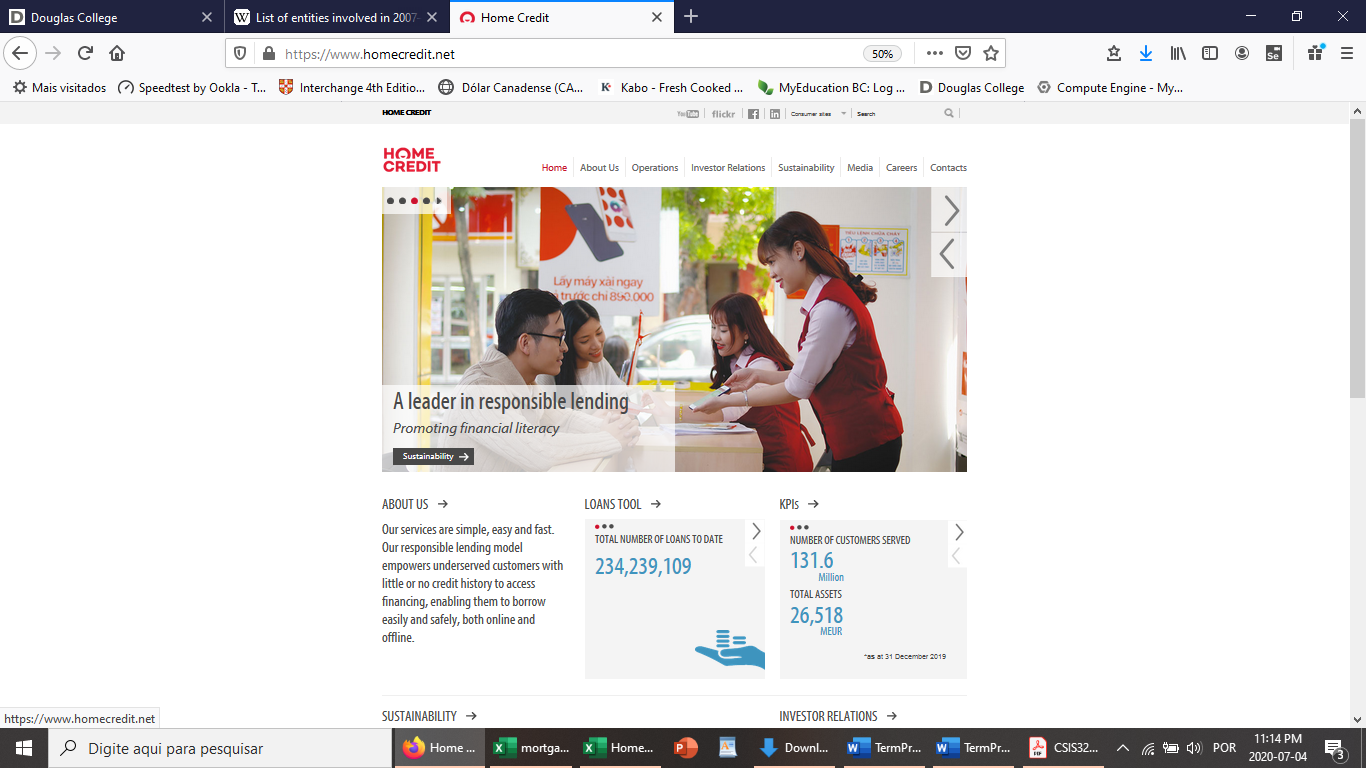
## Initial Hypothesis

There is a set of data from physical people, combined with the data of the house to be financed and its region, among other variables, that can be used to determine with a good level of confidence, that a person will be able to pay for his/her mortgage contract, without incurring into any Default situation.

## Data Source

From Kaggle it was possible to identify a dataset with a 307,511 rows which contains a variable that informs if a customer get into Default situation (24.825 Default cases, or 8.07 %) on his/her mortgage contract. This dataset was provided by Home Credit company and depicts actual customer from its database. The data set has 121 different variables, such as Customer Incoming, Assets and Occupation, as well as from the Home that was mortgaged, which can be explored as explanatory variables. It shall be also noted that, on the dataset, there are many missing data and one or more approach should be adopted to deal with these variables without risking losing valuable information.

[https://www.kaggle.com/c/home-credit-Default-risk/](https://www.kaggle.com/c/home-credit-default-risk/data?select=application_test.csv)



In this project it can be verified that many techniques were used to make feasible the dataset processing and make the model more accurate, by transforming the data which have been proved to be very complex. The technics used to simplify the dataset was.

* Features elimination based on domain knowledge
* Elimination of redundant features which presented a high correlation
* Features elimination based on statistic significance
* Features Scaling / Normalization
* Treatment of Unbalance frequency of the outcome
* Elimination of outliers

# Data Preparation

The first round of features elimination relied on the domain knowledge. It proves that a data scientist supported by somebody with good domain knowledge can save a lot of time by identifying a priori a set of variables which should no be impacting the outcome. Naturally, it can happen that some variables which are not considered by the domain expert could indeed have impact on the outcome and some valuable information could be lost. Considering some constrains of computational resources and time, though, relying on the domain expert can be adopted at first, and if there is the proper conditions to engage in a model optimization, some variables can be reintroduced continuously to the model.

Any combinations of dataset shapes and sizes were tried during this project. In some cases, the model performed poorly. At last, an optimal number of 50,000 rows to be processed proved to be sufficient to make good predictions (by adding more rows, the model could not be improved) and also made the dataset / model feasible to be processed with the computational resources I have at hand.

## Features Reduction by Domain Analysis

The following variables where discarded at first as part of the Domain Analysis.

* Dismiss all variables which provides median and mode data about the Building area. Since there are already the average values for the same variables, these are the ones that will be kept, cause the others seemed redundant.
* Dropped all features which the meaning is not clear in the domain context. These variables were simply related to the way the customer filled the Home Credit Form. It is not made any explicit reference about the meaning of this fields in the form, neither when the form was filled by the customer (is it an application form ?)
* The dataset has 3 normalized credit score from 3 different source / agencies. The customers can have 1, 2, 3 or none credit score information. An average of the existent credit scores is calculated and that is the unique credit score that will be used in the model.
* Wherever some variables seemed to be redundant and highly correlated, just one of them were kept. For instance, the dataset had Avg Rating of Clients for the *Region* the customer lives and the same information in the dataset about the *City* where the customer lives. In this case, it was kept just the last one.
* Some variables were summed up in order to keep just one meaningful variable. It was the case the variables which informed the quantity of Requirements done by the customer to the Service Bureau within the last hour, last day, last week, last month and last quarter. All of them were summarized in the quantity of the Last Quarter requirements done.
* Since there are still many categorical data remaining in the dataset, the GLM model ran with a lower quantity of rows in order to verify if these categorical variables were statistically significant. The GLM summary showed that there were some categorical variables which dummy variables P-Value almost reached 1.0. These variables much surpassed the limit of alpha 0.05 to determine if the variable is statistically significant, as shown below, and were removed from the model. In other words, these variables do no impact the Default outcome in at least 95% of the cases. If they do so, it is completely by chance. From the business perspective we can say that the fact the Gender is Male, Female or Other (XNA), it does not affect the fact that a customer will get into Default or not.

# The variables to be removed before being converted to Dummies were : NAME\_EDUCATION\_TYPE, CODE\_GENDER, NAME\_FAMILY\_STATUS

# NAME\_EDUCATION\_TYPE\_Academic degree -1.3275 817.8400 -0.0016 0.9987 -1604.2645 1601.6095

# NAME\_EDUCATION\_TYPE\_Higher education -1.0479 817.8398 -0.0013 0.9990 -1603.9845 1601.8887

# NAME\_EDUCATION\_TYPE\_Incomplete higher -0.9002 817.8398 -0.0011 0.9991 -1603.8368 1602.0363

# NAME\_EDUCATION\_TYPE\_Lower secondary -0.5536 817.8398 -0.0007 0.9995 -1603.4902 1602.3829

# NAME\_EDUCATION\_TYPE\_Secondary / secondary special -0.6566 817.8398 -0.0008 0.9994 -1603.5932 1602.2800

# CODE\_GENDER\_F 4.8386 5724.8786 0.0008 0.9993 -11215.7173 11225.3945

# CODE\_GENDER\_M 5.1230 5724.8786 0.0009 0.9993 -11215.4328 11225.6789

# CODE\_GENDER\_XNA -14.4475 15538.9562 -0.0009 0.9993 -30470.2419 30441.3470

# NAME\_FAMILY\_STATUS\_Civil marriage -0.8078 817.8398 -0.0010 0.9992 -1603.7443 1602.1288

# NAME\_FAMILY\_STATUS\_Married -1.0031 817.8398 -0.0012 0.9990 -1603.9397 1601.9334

# NAME\_FAMILY\_STATUS\_Separated -0.7674 817.8398 -0.0009 0.9993 -1603.7039 1602.1692

# NAME\_FAMILY\_STATUS\_Single / not married -0.8544 817.8398 -0.0010 0.9992 -1603.7909 1602.0822

# NAME\_FAMILY\_STATUS\_Widow -1.0532 817.8398 -0.0013 0.9990 -1603.9898 1601.8834

## Features Reduction – Statistical Criteria

* There were many columns also in the model which were sparsely filled / informed, so these features could not be considered as reliable to predict the outcome and were dismissed. All columns which have more than 20% of null values were removed from the model. Considering the dataset with the first 50,000 rows from the original dataset (in this document it will be explained before why the model, at last, ran with 50,000 rows instead of the whole dataset), these were the columns that were dismissed.

Deleted column OWN\_CAR\_AGE, 65.91% of nulls > 20.00%

Deleted column OCCUPATION\_TYPE, 31.30% of nulls > 20.00%

Deleted column APARTMENTS\_AVG, 50.77% of nulls > 20.00%

Deleted column BASEMENTAREA\_AVG, 58.40% of nulls > 20.00%

Deleted column YEARS\_BEGINEXPLUATATION\_AVG, 48.79% of nulls > 20.00%

Deleted column YEARS\_BUILD\_AVG, 66.47% of nulls > 20.00%

Deleted column COMMONAREA\_AVG, 69.92% of nulls > 20.00%

Deleted column ELEVATORS\_AVG, 53.30% of nulls > 20.00%

Deleted column ENTRANCES\_AVG, 50.39% of nulls > 20.00%

Deleted column FLOORSMAX\_AVG, 49.75% of nulls > 20.00%

Deleted column FLOORSMIN\_AVG, 67.79% of nulls > 20.00%

Deleted column LANDAREA\_AVG, 59.44% of nulls > 20.00%

Deleted column LIVINGAPARTMENTS\_AVG, 68.45% of nulls > 20.00%

Deleted column LIVINGAREA\_AVG, 50.27% of nulls > 20.00%

Deleted column NONLIVINGAPARTMENTS\_AVG, 69.43% of nulls > 20.00%

Deleted column NONLIVINGAREA\_AVG, 55.14% of nulls > 20.00%

Deleted column FONDKAPREMONT\_MODE, 68.38% of nulls > 20.00%

Deleted column HOUSETYPE\_MODE, 50.15% of nulls > 20.00%

Deleted column TOTALAREA\_MODE, 48.30% of nulls > 20.00%

Deleted column WALLSMATERIAL\_MODE, 50.92% of nulls > 20.00%

Deleted column EMERGENCYSTATE\_MODE, 47.40% of nulls > 20.00%

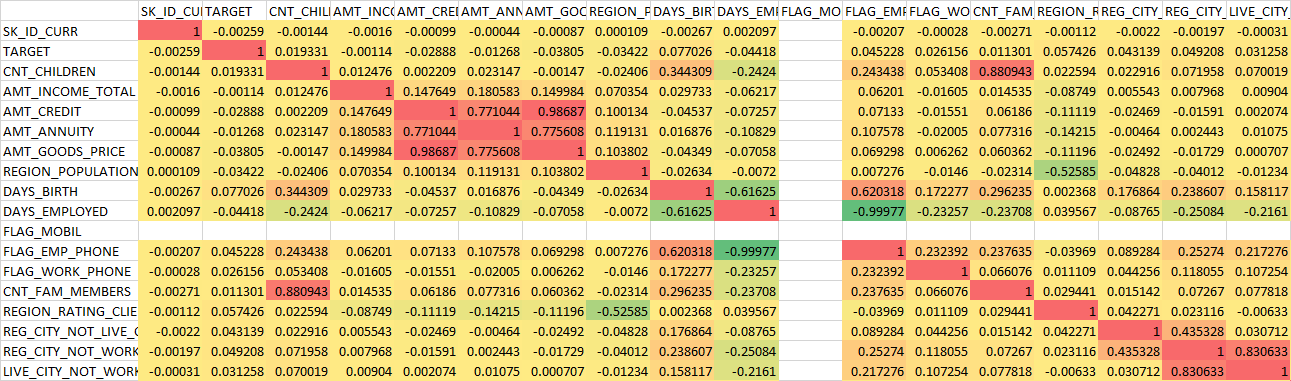
Rows before drop nulls

49965

Rows after drop nulls

42869

* By running the correlation matrix was possible to remove some columns that were redundant in the model since they were very correlated
  + AMT\_CREDIT and AMT\_GOODS\_PRICE - 98.7%
  + CNT\_FAM\_MEMBERS and CNT\_CHILDREN - 88.1%



## Elimination of Rows with Null Values

* After the reduction of features from the model, all rows which remained with any Null value were eliminated from the model.

The dataset became with 42869rows (from the 50,000 originally inputted) and 24 variables

## Data Transformation

* *Categorical Variables into Dummy Variables.* After discarding several categorical variables which had a great quantity of categories (impacting the model execution) or which were not statistically significant, it remained 3 categorical variables as below.
  + 'FLAG\_OWN\_CAR', 'FLAG\_OWN\_REALTY','NAME\_CONTRACT\_TYPE'
* *Unbalanced Dataset* - One issue that was impacting heavily the model was the fact that the dataset is highly unbalanced. Just 24.825 Default cases out of 307,511 rows, or just 8.07 %. This situation was making the 100% of prediction as Non Default (0). A high accuracy of 92% was masking the fact that these 92% was related to the all 92% of Non Default cases, while the model was predicting as False Negative 100% of the cases of Default. The ROC graph showed the True Positive curve almost touching the line that shows predictions by chance. It was clear, based on this, that the accuracy itself cannot be used alone as performance indicator of the model. Area Under ROC Curve (AUROC) can be used in order to certify if the Accuracy is good or not for unbalanced datasets. A routine to make the dataset balanced with the same quantity of Default (1) and Non Default (0) rows was implemented, resampling the dataset to have 3 times the original quantity of Default cases and make the Non Default cases as the same quantity. The final number of rows that was used to do the training / testing process was as below with the outcome distributions. Some tries were made to do the upscaling of Default rows as 4 times did not introduced any improvement on the model. Also downscaling the Non Default to number of cases (downscale) proved to be less efficient than upscaling the Default cases.

Originally

0 – No Default – 39517 rows

1 – Default – 3352 rows

After Resample

0 – No Default – 10056rows

1 – Default – 10056rows

* Scaling – All numeric, non categorical dummy variables were scaled before training. It was tried MinMax, Standard and Robust scalers. *Robust scaler*, which reduce the effects of outliers, presented the better performance, improving the accuracy of the final model.

## Logistic Binomial Regression Analysis (GLM – Generalized Linear Model)

By Analyzing the dataset summary below it is possible to verify that some features have P-Value > 0.05, which means that these features are not statistically significant (the features in yellow shows the ones that are statistically significant). In other words, these variables do not impact consistently the predicted outcome and if they so, it is just by chance. More precisely, the other variables that are indeed statistically significant impact the outcome in at least 95% of the predictions.

The features that are not statistically significant can be removed from the training and test datasets since they have no relevant impact on the results.

Some business insights could be obtained by verifying the Generalized Linear Model and the Odds Ratio as below.

Summary of Logistic Binomial Regression Analysis (GLM function) for the Balanced Dataset

Results: Generalized linear model

===================================================================================

Model: GLM AIC: 24280.0674

Link Function: logit BIC: -174837.1877

Dependent Variable: y Log-Likelihood: -12118.

Date: 2020-08-02 22:28 LL-Null: -13941.

No. Observations: 20112 Deviance: 24236.

Df Model: 21 Pearson chi2: 2.02e+04

Df Residuals: 20090 Scale: 1.0000

Method: IRLS

-----------------------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------

AMT\_INCOME\_TOTAL -0.0000 0.0000 -2.0241 0.0430 -0.0000 -0.0000

AMT\_CREDIT -0.0000 0.0000 -6.3347 0.0000 -0.0000 -0.0000

AMT\_ANNUITY 0.0000 0.0000 6.5842 0.0000 0.0000 0.0000

REGION\_POPULATION\_RELATIVE 3.6006 1.3813 2.6066 0.0091 0.8933 6.3080

DAYS\_BIRTH 0.0000 0.0000 2.9029 0.0037 0.0000 0.0000

DAYS\_EMPLOYED -0.0001 0.0000 -10.4393 0.0000 -0.0001 -0.0001

FLAG\_MOBIL 1.5479 0.1641 9.4310 0.0000 1.2262 1.8696

FLAG\_EMP\_PHONE 0.3022 0.0323 9.3424 0.0000 0.2388 0.3656

FLAG\_WORK\_PHONE 0.0501 0.0392 1.2784 0.2011 -0.0267 0.1268

CNT\_FAM\_MEMBERS -0.0057 0.0171 -0.3335 0.7388 -0.0393 0.0279

REGION\_RATING\_CLIENT\_W\_CITY 0.1742 0.0371 4.6994 0.0000 0.1016 0.2469

REG\_CITY\_NOT\_LIVE\_CITY 0.1851 0.0758 2.4421 0.0146 0.0365 0.3336

REG\_CITY\_NOT\_WORK\_CITY -0.0344 0.0862 -0.3988 0.6900 -0.2034 0.1346

LIVE\_CITY\_NOT\_WORK\_CITY 0.0774 0.0838 0.9228 0.3561 -0.0870 0.2417

OBS\_30\_CNT\_SOCIAL\_CIRCLE 0.0040 0.0070 0.5735 0.5663 -0.0097 0.0178

DEF\_30\_CNT\_SOCIAL\_CIRCLE 0.0827 0.0337 2.4577 0.0140 0.0167 0.1487

DAYS\_LAST\_PHONE\_CHANGE -0.0001 0.0000 -2.8871 0.0039 -0.0001 -0.0000

AVG\_CREDIT\_SCORE -5.3778 0.1206 -44.5948 0.0000 -5.6142 -5.1414

AMT\_REQ\_CREDIT\_BUREAU\_QRT\_TOTAL -0.0010 0.0075 -0.1322 0.8949 -0.0157 0.0137

FLAG\_EMPLOYED 0.3022 0.0323 9.3424 0.0000 0.2388 0.3656

FLAG\_OWN\_CAR\_Y -0.1570 0.0339 -4.6283 0.0000 -0.2235 -0.0905

FLAG\_OWN\_REALTY\_Y 0.0146 0.0341 0.4268 0.6695 -0.0523 0.0814

NAME\_CONTRACT\_TYPE\_Revolving loans -0.5984 0.0636 -9.4147 0.0000 -0.7229 -0.4738

===================================================================================

## New Model Simplification – Features reduction based on Variables that are statistically significant

A second round of features reduction was done to reduce the number of features and improve the model accuracy and AUROC. Two methods were applied, and it was chosen that presented the better accuracy. By the Logistic Binomial Model, it was eliminated 7 variables more.

Columns to be eliminated using Logistic Binomial Model

['FLAG\_WORK\_PHONE', 'CNT\_FAM\_MEMBERS', 'REG\_CITY\_NOT\_WORK\_CITY', 'LIVE\_CITY\_NOT\_WORK\_CITY', 'OBS\_30\_CNT\_SOCIAL\_CIRCLE', 'AMT\_REQ\_CREDIT\_BUREAU\_QRT\_TOTAL', 'FLAG\_OWN\_REALTY\_Y']

Number of Columnns to be KEPT after eliminating features using Logistic Binomial Model : 16

Number of Columnns to be REMOVED using Logistic Binomial Model : 7

Accuracy of Linear Regression Classifiers used to Reduce Features from Dataset:

Model\_Name ... Model\_Dataset\_Accuracy

0 Logistic Regression ... 0.669871

1 LinearSVC ... 0.671362

2 SelectKBest / Mutual Info Classification ... 0.668213

3 Logistic Binomial Model ... 0.671859

Model used to reduce Dataset features : Logistic Binomial Model

Columns kept in the dataset after Features reduction using Logistic Binomial Model:

Index(['AMT\_INCOME\_TOTAL', 'AMT\_CREDIT', 'AMT\_ANNUITY',

'REGION\_POPULATION\_RELATIVE', 'DAYS\_BIRTH', 'DAYS\_EMPLOYED',

'FLAG\_MOBIL', 'FLAG\_EMP\_PHONE', 'REGION\_RATING\_CLIENT\_W\_CITY',

'REG\_CITY\_NOT\_LIVE\_CITY', 'DEF\_30\_CNT\_SOCIAL\_CIRCLE',

'DAYS\_LAST\_PHONE\_CHANGE', 'AVG\_CREDIT\_SCORE', 'FLAG\_EMPLOYED',

'FLAG\_OWN\_CAR\_Y', 'NAME\_CONTRACT\_TYPE\_Revolving loans'],

dtype='object')

Dataset shape after Features reduction using Logistic Binomial Model:

(20112, 16)

## Training the Model

Using a Scikit Pipeline the following Classifiers were used and the accuracy obtained for each model are below.

The model was not possible to run with the following models which struggled to run with 50,000 rows as input : "RBF SVM", "MPL Neural Net" and "Gaussian Process".

“QuadraticDiscriminantAnalysis” ran so poorly during the first runs with a smaller datasets that were dismissed for the final tests with 50,000 rows as input.

The Pipeline ran progressively with 1,000, 10,000, 50,000 and 100,000 rows. By 100,000 rows it was noticed no improvements in the models accuracy, so 50,000 rows as inputs was set as the optimal input dataset set.

Classifier\_Name ... AccuracyScore

6 Bagging Classifier ... 0.854144

8 XGBoost ... 0.801657

1 Nearest Neighbors ... 0.730801

7 AdaBoost ... 0.684530

2 Linear SVM ... 0.676519

3 Decision Tree ... 0.675967

0 Logistic Regression ... 0.674171

5 Random Forest ... 0.660773

4 Naive Bayes (GaussianNB) ... 0.647652

[9 rows x 3 columns]

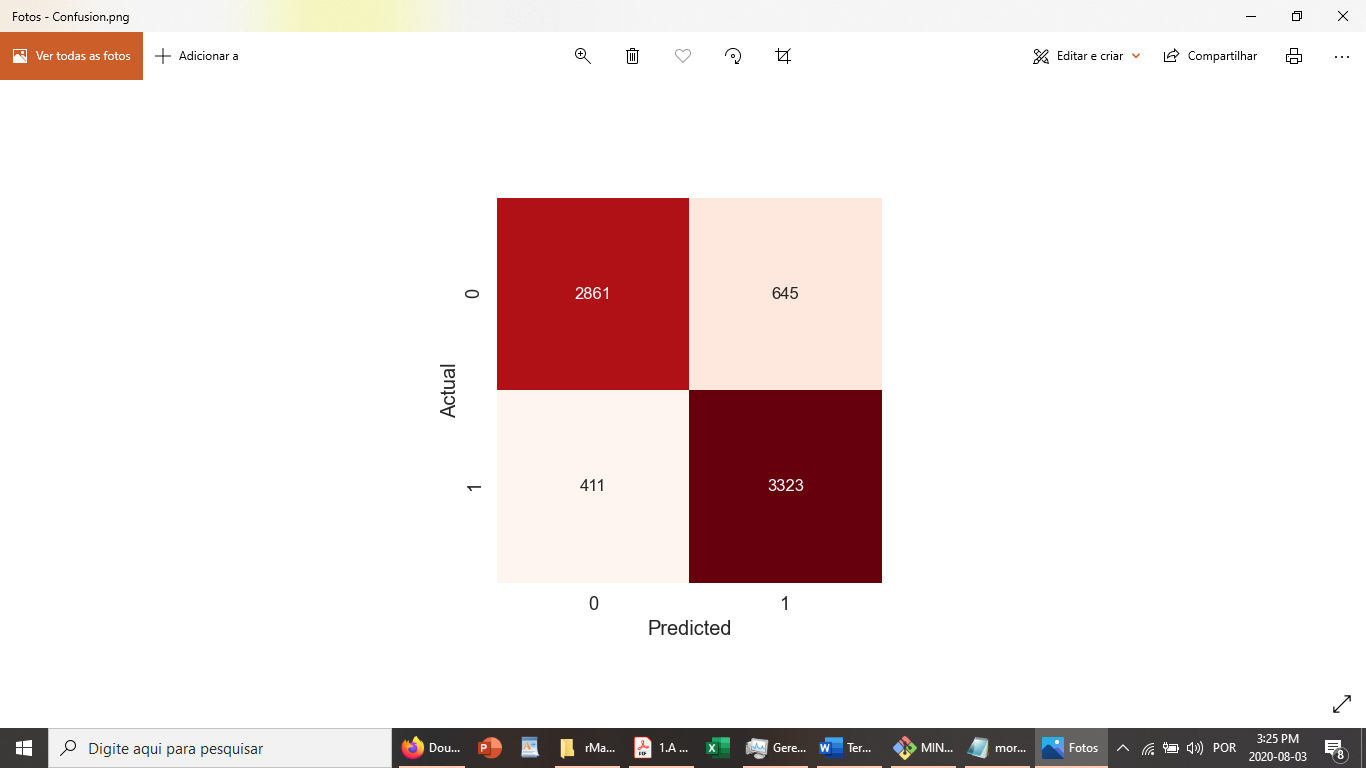
Selected Model is Bagging Classifier

BaggingClassifier()

## Assess the predictive performance of Linear SVM selected

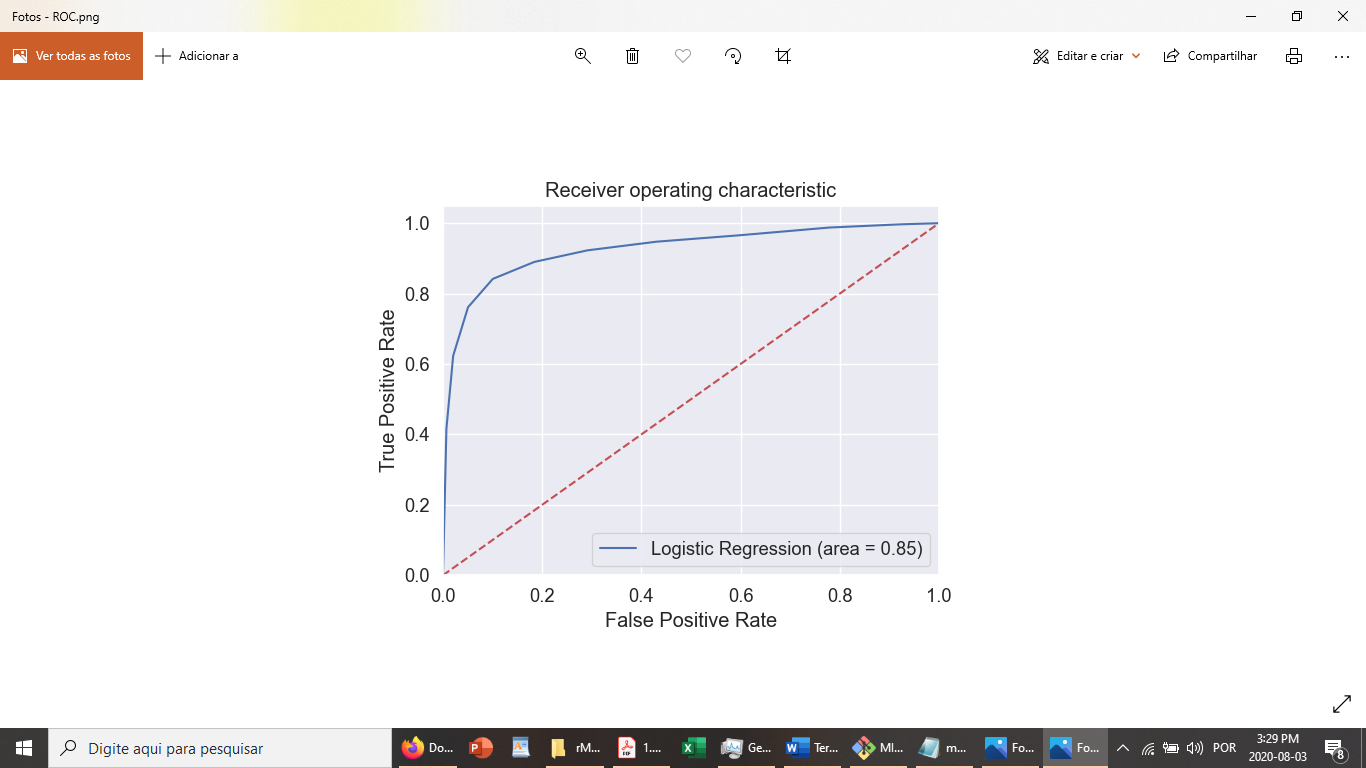
### Confusion Matrix

As we can see on the Confusion Matrix, the Bagging Classifier model made fair predictions for the Default cases with an accuracy of 85%. Also, it can be noted a low level of False Negatives, 18%, and only 11% of False Positives.



### ROC graph

By the Receiver Operator Characteristic (ROC) curve below it can be verified that the curve passes on the True Positive band of the graph, showing that, as overall, the model performed well, with a reasonable AUROC (A random classifier would give points lying along the diagonal)



## Business Considerations and Conclusions

The model has proven that Home Credit’s set of variables can be used, with some level of confidence, to predict that mortgage contract customers will get into a Default situation soon. Since the statistically significant variables dataset can be changed along the time, it is necessary, though, to continuously update the dataset and reassess the data model. It was possible to verify also that is likely that it is not necessary to run the entire dataset in order to get good predictors. Upon a dataset size and beyond, it is shown that the model converged to an average (as an example the accuracy could not be improved by having 100,000 rows samples instead of 50,000 samples). Corporations like Home Credit though can afford for powerful computing resources which could make possible to explorer larger datasets.

The project showed that the Default situation has a very strong correlation with the credit score. So Home Credit have to assess if mortgages were granted to Default customers with low credit score and change its polices if it was the case or if the customer had his/her credit score deteriorating along the time.

The model also showed us that the Default situation is not only correlated to credit score and there are some other features that make the customer more likely to get into Default. But in general some personal data from the customer such as gender, education level, his/her family size does not impact at all the odds of get into Default.

Among those situations which can evolve along the mortgage contract period, the customer loosing his/her job, is the one that impact more the ability of mortgage installment payment. A customer who loose a job is 30% more likely to get into Default. So Home Credit should be actively monitoring theirs customers employment status. Some option for customers that lots job could be a renegotiation of the annuity and making the contract longer, so it would permit the customer still pay for the mortgage even with temporarily lower installments. This situation could be much better for Home Credit than simply does not receive any payment.

It was interesting to notice that more the region the customer lives is populated more the chance of Default. It should be assessed better by Home Credit. A Clustering process should be applied to investigate some regions separated and understanding if there some Regions in which these situations are happening and what would be the cause. Some regions could have higher interest on the mortgage based on this data.

One variable that presented a strong correlation with the Default situation is if the contract is Revolving Loans or not. (From Investopedia – “A revolving loan facility is a form of credit issued by a financial institution that provides the borrower with the ability to draw down or withdraw, repay, and withdraw again. A revolving loan is considered a flexible financing tool due to its repayment and re-borrowing accommodations. It is not considered a term loan because, during an allotted period of time, the facility allows the borrower to repay the loan or take it out again. In contrast, a term loan provides a borrower with funds followed by a fixed payment schedule”). Customers with Revolving Loan contracts, because of the flexibility of the contract to be adjusted as necessary, have less 60% chance of Default than the regular contract. Thus, Home Credit should increase the number of this contract modality if they wants to reduce risks, despite the contract can be less advantageous for the borrower. In situations such as COVID-19 pandemic when many customers faced difficulties to pay for the mortgage, though, such contract should be good to work as cushion for Home Credit’s cash flow issues.

The Accuracy of 85% showed that the model can be fairly applied to anticipate Default situations. This could be used extensively by Home Credit to assess their risk level and make the necessary provisions in proportion of the accuracy. The model presented a low level of False Negatives, 18%, and about 11% of False Positives.

An important finding from this project is that, working properly, statistically assessing the variables, a finance institution, like Home Credit, obtain relevant outcomes from a limited set of data. Extrapolating this approach for other companies with also suffers with non payments and which have huge databases, like Telco companies, can have reliable models running over less variables.

Another finding is the importance of the Domain Expert. His/Her knowledge can help to get valuable shortcuts when preparing the dataset, understanding the features, and interpreting the results. A combination of skilled programmer, the domain expert and somebody versed on statistics is a key success factor for machine learning projects such this.

Home Credit can also continuously make adjustments in their risk assessment policies and systems. Although this model is based on already customers who contracted the mortgage, the company assess of the employment status, the region, the age of the customer among other 8 variables before granting mortgages contract is worthy it to decide for the granting and setting proper interest, accordingly to the customer risk.

## Out of Sample Predictions

The table below, with the significant variables that were kept by the model and used as input to make the predictions, shows that changes on some variables truly can make changes on the outcome / prediction, as expected.

By increasing dramatically, the number of days employed (from 2,7 years, to 27 years), the prediction changes from “Default” no “Not Default” which proves the correlation between “Not Default” and customers who have more stable jobs.

The same in relation to the Credit Score. All else being equal, a customer that at first was predicted as Default, with a higher Credit Score is predicted as non Default.

