Home Credit Default Risk ADS Analysis

Eileen Cho (ec3636), Apurva Bhargava (ab8687)
DS-GA 3001.009, Responsible Data Science, Spring 2020

May 11, 2020

1 Background

Home Credit seeks to offer loans to unbanked people - people who struggle to to secure loans from banks because of insufficient credit history. Without the the ability to determine credit worthiness through the means of credit history, Home Credit needs to use alternative data to determine default risk for identifying potential customers. It is important for the company to accurately identify those who will be able to repay their loans in a timely manner as new customers, while at the same time minimizing default risk. We will evaluate and create a nutritional label for the Home Credit default risk automated detection system (ADS), in terms of recipe (LightGBM algorithm), ingredients (attributes serving as most important features), stability and fairness. The goal of this ADS, as stated by Home Credit on Kaggle, is to identify potential customers who are capable of repaying their loans. To this end, this ADS was created to predict the probability of the applicant defaulting on their loan. The evaluation metric used for the ADS is Area Under Curve (AUC for Receiver Operating Characteristic Curve) using the predicted probabilities and the observed target. Here, a higher AUC means a better performing model.

2 Input and Output

2.1 Data Source

The dataset used for training the ADS is provided by Home Credit on Kaggle.[1] This data set consists of seven tables, with information that either exists in Home Credit's own database or information from other financial institutions that reported to the Credit Bureau.

The main one is `application_train.csv` (also, `application_test.csv`), which contains loan applicant information from Home Credit's own database, with the target variable being 1 for high default risk and 0 for other cases. The data here is given by the applicant to Home Credit when applying for a new loan. This table can be useful for identifying sub-populations, since it contains information such as applicant gender, age, education, occupation, address, housing type, income, etc. Each row represents one loan in the data sample and is identified by an ID number. All other tables are joined with main files using the attribute `SK_ID_CURR`.

The table in `bureau.csv` contains data of previous loans from other financial institutions, available from the Credit Bureau. For every loan in data sample, there are as many rows as number of credits the client had in Credit Bureau before the application date. Another table, `bureau_balance.csv` contains monthly balances of previous credit reported to the Credit Bureau. This table has one row for each month of history of every previous credit reported to Credit Bureau, i.e the table has (number of loans in sample of relative previous credits $\times$ number of months where we have some history observable for the previous credits) rows. These two data tables are joined using `SK_ID_BUREAU`.

The `previous_application.csv` contains application information of previous loans with Home Credit. There is one row for each previous application related to loans in data. The `POS_CASH_balance.csv` table contains monthly balance and cash loan information that the applicant may have had previously with Home Credit. This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in data sample, i.e. the table has (number of loans in sample $\times$

[1] `https://www.kaggle.com/c/home-credit-default-risk/data`

number of relative previous credits × number of months in which we have some history observable for the previous credits) rows. The `credit_card_balance.csv` table has historic monthly credit balance from the applicant's previous credit at Home Credit. This table has (number of loans in sample × of relative previous credit cards × number of months where we have some history observable for the previous credit card) rows. Finally, `installments_payments.csv` shows the repayment history of the applicant's previous loans with Home Credit. One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in data sample. These (last) three tables are joined with previous application table on the attribute `SK_ID_PREV`.

2.2 Input Features

The solution ADS uses 660 features for making a risk prediction. Apart from the features from the original seven CSVs, there are also engineered features like Boolean indicators (late payments, document flags, etc.), ratios and differences of credit amounts, annuities, payments, durations, etc., and statistics (maximum, minimum, mean, median) of aggregations over groups. The non-numerical categorical attributes have been used in the ADS as either mapped numbers or one-hot encoded features.

The pairwise correlation heatmaps for numerical and non-numerical, categorical features are given in Appendix A. The most relevant/ interpretable features, their distribution, attribute types, data types and missing value counts are given in Table 1 under Appendix B. The attributes from main file that are most positively correlated and negatively correlated with `TARGET` (output attribute) are given in Figure 1.

TARGET	1.000000	EXT_SOURCE_3	-0.178919
DAYS_BIRTH	0.078239	EXT_SOURCE_2	-0.160472
REGION_RATING_CLIENT_W_CITY	0.060893	EXT_SOURCE_1	-0.155317
REGION_RATING_CLIENT	0.058899	DAYS_EMPLOYED	-0.044932
DAYS_LAST_PHONE_CHANGE	0.055218	FLOORSMAX_AVG	-0.044003
DAYS_ID_PUBLISH	0.051457	FLOORSMAX_MEDI	-0.043768
REG_CITY_NOT_WORK_CITY	0.050994	FLOORSMAX_MODE	-0.043226
FLAG_EMP_PHONE	0.045982	AMT_GOODS_PRICE	-0.039645
REG_CITY_NOT_LIVE_CITY	0.044395	REGION_POPULATION_RELATIVE	-0.037227
FLAG_DOCUMENT_3	0.044346	ELEVATORS_AVG	-0.034199
DAYS_REGISTRATION	0.041975	ELEVATORS_MEDI	-0.033863
OWN_CAR_AGE	0.037612	FLOORSMIN_AVG	-0.033614
LIVE_CITY_NOT_WORK_CITY	0.032518	FLOORSMIN_MEDI	-0.033394
DEF_30_CNT_SOCIAL_CIRCLE	0.032248	LIVINGAREA_AVG	-0.032997
DEF_60_CNT_SOCIAL_CIRCLE	0.031276	LIVINGAREA_MEDI	-0.032739
FLAG_WORK_PHONE	0.028524	FLOORSMIN_MODE	-0.032698
AMT_REQ_CREDIT_BUREAU_YEAR	0.019930	TOTALAREA_MODE	-0.032596
CNT_CHILDREN	0.019187	ELEVATORS_MODE	-0.032131
CNT_FAM_MEMBERS	0.009308	LIVINGAREA_MODE	-0.030685
OBS_30_CNT_SOCIAL_CIRCLE	0.009131	AMT_CREDIT	-0.030369
OBS_60_CNT_SOCIAL_CIRCLE	0.009022	APARTMENTS_AVG	-0.029498
REG_REGION_NOT_WORK_REGION	0.006942	APARTMENTS_MEDI	-0.029184
REG_REGION_NOT_LIVE_REGION	0.005576	FLAG_DOCUMENT_6	-0.028602
FLAG_DOCUMENT_2	0.005417	APARTMENTS_MODE	-0.027284
FLAG_DOCUMENT_21	0.003709	LIVINGAPARTMENTS_AVG	-0.025031
LIVE_REGION_NOT_WORK_REGION	0.002819	LIVINGAPARTMENTS_MEDI	-0.024621
AMT_REQ_CREDIT_BUREAU_DAY	0.002704	HOUR_APPR_PROCESS_START	-0.024166
AMT_REQ_CREDIT_BUREAU_HOUR	0.000930	FLAG_PHONE	-0.023806
AMT_REQ_CREDIT_BUREAU_WEEK	0.000788	LIVINGAPARTMENTS_MODE	-0.023393

Figure 1: Correlation of input features with class variable TARGET

The visualization of the distribution of numerical variables (boxplots) and categorical variables (piecharts), as well as table of missing values for both numerical and non-numerical features is given in `DataExploration.ipynb`.

Fraction of positives is the ratio of number of examples with `TARGET`=1 (high risk) to total number of examples (`TARGET`=1 and `TARGET`=0 cases) for a given subgroup based on some attribute value. The fraction of positives for certain values for a given attribute is higher, indicating that the examples from the subgroup have been assigned higher risk in the training data. It can be noted that the data is not biased against females as they have higher fraction of positives than males, however certain occupations, regions, housing types, ages, and employment durations show higher default risk being assigned more often to certain groups than others, for example, as age decreases, the fraction of high risk labels increases. `AGE_BIN` and

EMPLOYMENT_DURATION_BIN are created from DAYS_BIRTH and DAYS_EMPLOYMENT respectively by converting them into positive years and then binning them. The fractions of positives (high risks) for different features from the training set are given in Figure 2.

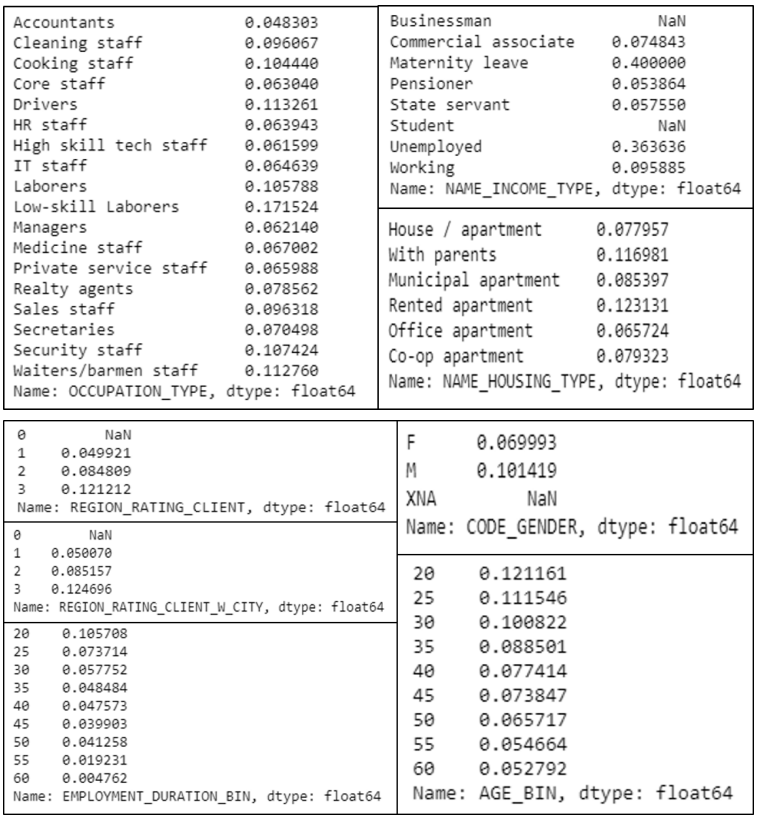


Figure 2: Fraction of risk assignments for various attributes

2.3 Output Variable

The output variable is TARGET, which is a binary categorical variable. Its values are 1 (risk label) and 0 (otherwise). Distribution-wise, it is extremely skewed. In the main training file, there are only 24825 rows where TARGET=1 and 282686 rows where TARGET=0. The ADS outputs a probability (between 0 and 1) for risk. A classification threshold (some probability value separating the two classes) can be chosen for classifying a given example as risk or no-risk.

3 Implementation and Validation

The entire pipeline for the ADS we are evaluating comes from the 7th place solution posted on Kaggle.[2]

3.1 Data Cleaning and Preprocessing

The data from the seven tables were preprocessed and joined into a single table to be read as input for the ADS. From the application_train.csv table, 10 categorical features are kept, including CODE_GENDER, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, to name a few. Applicant age was split into 5 bins: < 27, 27-40, 40-50, 50-65, 65-99. For the other numerical features in this table, some ratios and differences were calculated to produce some new features. These ratios (or differences) include ones such as credit to total income, and age of owning a car to total income. Statistics on these ratios grouped by ORGANIZATION_TYPE, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, AGE_RANGE, CODE_GENDER. From the other six tables, similar calculations of ratios between relevant features, and differences between features for time duration were calculated, as well as aggregations over specific loan types, and specific durations. These features are added to

[2] https://www.kaggle.com/jsaguiar/lightgbm-7th-place-solution

emphasize time-related features. Many features such as `CNT_CHILDREN`, `CNT_FAM_MEMBERS`, `FLAG_OWN_REALTY`, `OBS_30_CNT_SOCIAL_CIRCLE`, `COMMONAREA_MODE`, etc. have been dropped. The final preprocessed dataset that the model trains on contains 660 features.

3.2 ADS implementation

This ADS uses the LightGBM implementation of a Gradient Decision Boosted Tree (GDBT) with Gradient-based One-Side Sampling (GOSS), as described in Ke et al. (2017). GDBT is a tree ensemble model. Unlike the Random Forest model, which trains many independent decision trees and combines the models, GDBT trains many trees sequentially. At each iteration, the new tree tries to boost its performance by fixing the mistakes in prediction in the previous iteration. This way each of the weak models benefits from the error of the previous weak models. LightGBM is an implementation that uses GOSS to make training more efficient on large datasets. This efficiency is gained under the assumption that data instances with larger gradients have more information to be learned than the instances with smaller gradients. So when down sampling the data during training, the instances with larger gradients are kept and the instances with small gradients are randomly dropped.

3.3 ADS validation

This ADS is evaluated using AUC score. The model was trained using a k-fold of 10, with early stopping rounds set at 100. This way, the training data is split into 10 parts, with one part serving as validation and the other nine as training data. Models were trained over these 10 splits, and the best model iteration of each train-validation split gets used to predict probabilities on the test data set. The results of each sub model on the test set is averaged in order to get the final AUC score, which was 0.80. This fulfills the Kaggle competition's goal of having the model achieve a high AUC score.

4 Outcomes

4.1 Selection of Classification Threshold

Since the official ADS solution was only evaluated using AUC, the exact prediction (1/risk or 0/no-risk) is dependent on the choice of the classification threshold. In order to evaluate model performance in terms of correctness, fairness, and stability, the model's accuracy, target-wise accuracy and F1 scores were calculated for different threshold values, as shown in Figure 4. The confusion matrices for the outcomes are given in Figure 3. For any threshold, the accuracy of no-risk cases decreases when that of risk cases is attempted to be increased, and vice versa. Thus, when minimizing risk, there is a loss of potential no-risk applicants. And when trying to increase applicants, there is a potential increase of also admitting risky applicants.

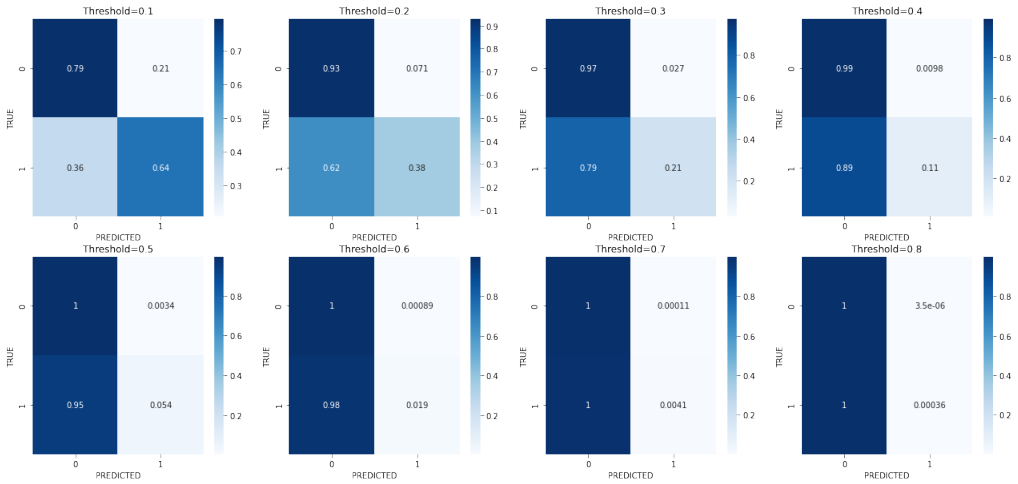
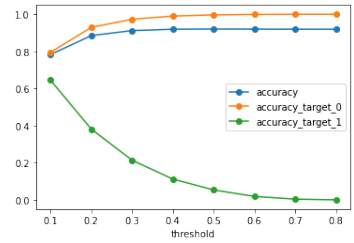


Figure 3: Confusion Matrices for different thresholds

	accuracy	accuracy_target_0	accuracy_target_1	f1_score
0.1	0.782710	0.794819	0.644820	0.323923
0.2	0.885101	0.929327	0.381486	0.348983
0.3	0.911803	0.973101	0.213785	0.281277
0.4	0.919325	0.990205	0.112190	0.183356
0.5	0.920460	0.996576	0.053698	0.098286
0.6	0.919985	0.999109	0.018974	0.036873
0.7	0.919507	0.999890	0.004149	0.008254
0.8	0.919299	0.999996	0.000363	0.000725



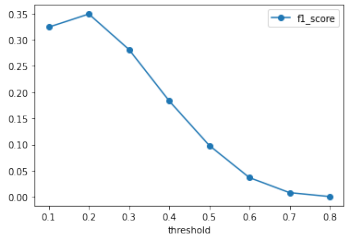


Figure 4: Accuracies and F1 Score for different thresholds

Assuming that Home Credit, as stakeholders invested in gaining new customers, chooses the threshold that prioritizes accuracy of low-risk cases over than that of risk cases, which are few in number, and that they also want to have a reasonable F1 score, a classification threshold of 0.3 may be selected. For all further analysis, 0.3 will be the chosen threshold such that `PREDICTED_TARGET`=1 if predicted probability of risk > 0.3, and 0 otherwise.

4.2 Performance Across Sub-populations

The performance of the model was tested for different subsets of the data. The performance measures are accuracy and AUC. The attributes used for subset (subpopulation) creation are:

- `CODE_GENDER`(gender: female and male). The division could be made based on whether the value of this attribute is F or M. While the data shows no bias against the protected group (F), however, it is prudent to check for the same in the results of the ADS.
- `DAYS_BIRTH` and `AGE_RANGE` (derived from `DAYS_BIRTH`). The training data showed higher risk assignment as age decreased. `AGE_RANGE` was used to generate subgroups and test this.
- `DAYS_EMPLOYMENT` or number of days the applicant has worked counted backwards from the day of application.
- `OCCUPATION_TYPE`. Occupations such as labourers, waiters, security staff have a higher fraction of positive labels than other classes in the main train file.
- `REGION_RATING_CLIENT`, `REGION_RATING_CLIENT_W_CITY`. Rating of the region where the applicant lives also affected the risk assignment.

Of these, we will discuss at `CODE_GENDER` (Male is privileged, Female is protected), `AGE_RANGE` (older people are privileged), `REGION_RATING_CLIENT` (1 is the best rating and 3 the worst) in this report. The other attributes are analysed in a similar manner in `FairnessMetrics.ipynb`.

4.3 Accuracy and AUC

The accuracies and AUC scores for the different subgroups are plotted in Figure 5. The classification accuracies are the highest for the privileged subgroup in case of `AGE_RANGE` and `REGION_RATING_CLIENT`, i.e., the model performs better for those groups. This does not hold for `CODE_GENDER`. This is likely because 66% of the applicants in the dataset are Female. The AUC score does not follow a discernable trend.

4.4 Fairness

For testing fairness in classification among different sub-populations, following fairness measures will be used to quantify bias and fairness. (Here, positive is 0 or low-risk label, and negative is 1 or high-risk label.)

- **Statistical Parity**. Statistical parity requires that the fraction of risk outcomes and no-risk outcomes in each group is the same. As shown in Figure 6, the ratios aren't equal for all sub-populations of a given category. Except for gender, where the protected subgroup FEMALE has lower predicted risk assignment, other categories have higher predicted risk assignments for protected classes (like lower age or lower-rated region of living get higher risk).

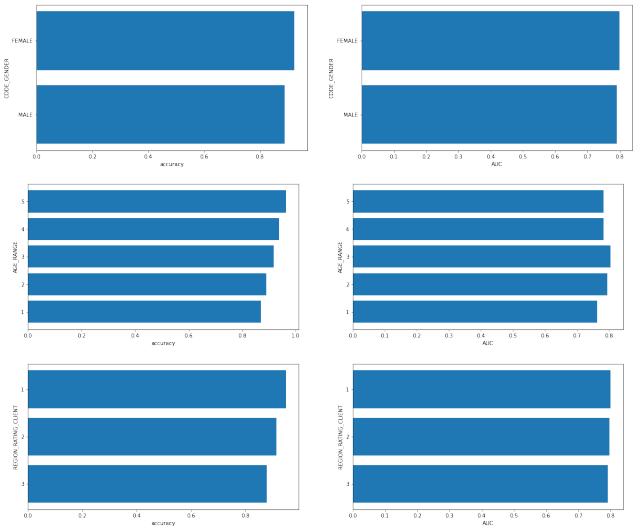


Figure 5: Accuracy and AUC for CODE_GENDER, AGE_RANGE and REGION_RATING_CLIENT

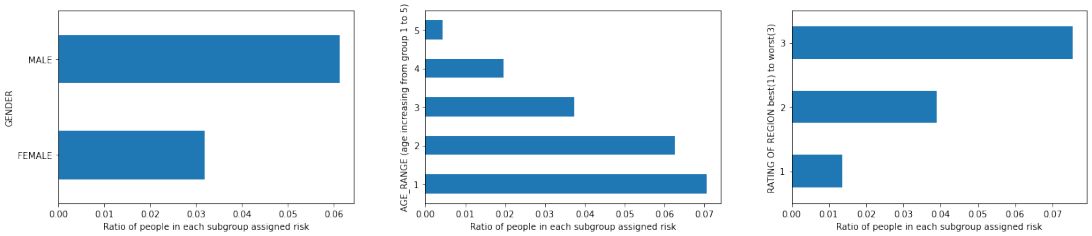


Figure 6: Statistical Parity for CODE_GENDER, AGE_RANGE and REGION_RATING_CLIENT

- **Disparate Impact**. It is the ratio of no-risk outcomes for the protected group and that for the non-protected group. In case of multiple groups, the denominator in the ratio could be for all other groups except for the one in numerator, or the less protected groups. Since this is a pairwise metric, age is divided into two groups and region ratings are selected pairwise for measurement. Results can be found in Table 1. For gender, the value of disparate impact is greater than 1. For all other categories, it is less than 1 (showing that protected groups are less favoured), however, the value is still above 0.93 in most cases. This is because at the chosen threshold, a large number of people are assigned no-risk and the training data is also largely skewed in favor of no-risk label.

Table 1: Disparate Impact

Attribute	Privileged Group	Unprivileged Group	Disparate Impact
Gender	Male	Female	1.031392078802175
Age	>=50	<50	0.9638316697820878
Region Rating	1	3	0.9372257744139441
Region rating	2	3	0.9618648938300324
Region rating	1	2	0.9743840121683013

- **Predictive Parity (equality of Positive Predictive Value) and Negative Predictive Value (NPV).** PPV is the probability indicative of whether a low-risk assignment is truly low-risk. This is larger for all privileged subgroups (except Male in gender). This means that low-risk individuals are more likely to be marked low-risk in privileged groups as compared to unprivileged groups. NPV, on the other hand, indicates the probability that a risk assignment is not actually a risk case. NPV is higher for unprivileged groups (except for Male).

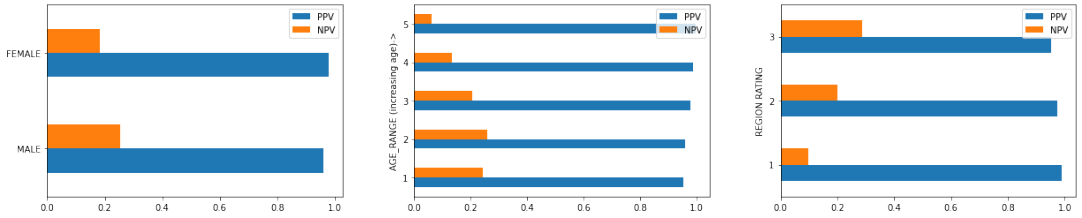


Figure 7: PPV and NPV for CODE_GENDER, AGE_RANGE and REGION_RATING_CLIENT.

- **False Positive Rate Ratio (FPRR) and False Negative Rate Ratio (FNRR).** FPR is the probability of false no-risk assignment to a group. FPRR is ratio of FPR for unprivileged and privileged subgroups. As shown in Table 2, it is lower than 1 in most cases (excluding gender), indicating that privileged subgroups are more likely to be marked low-risk even when they are risky. FNR is the probability of false risk assignment to a group. FNRR is ratio of FNR for unprivileged and privileged subgroups. It is greater than 1 in most cases (excluding gender), indicating that unprivileged subgroups are more likely to be marked high risk even when they are not risky.

Table 2: False Positive Rate Ratio and False Negative Rate Ratio

Attribute	Priv. Group	Unpriv. Group	FPRR	FNRR
Gender	Male	Female	1.0273	0.7316
Age	>=50	<50	0.9734	1.4676
Region Rating	1	3	0.8836	1.9405
Region rating	2	3	0.9767	1.3054
Region rating	1	2	0.9047	1.4865

- **Conditional ratio of risk/ no-risk given `CREDIT_TO_INCOME_RATIO`.** The sub-populations are divided into various groups according to the value of `CREDIT_TO_INCOME_RATIO` (CTIR) into CTIR bins. It is expected that the higher the credit to income ratio, the higher the chances of being assigned as high risk. As shown in Figure 8, this trend plays as expected in the case of the privileged subgroups, but deviates for certain CTIR bins in the case of unprivileged subgroups. Also, for a given CTIR bin, the unprivileged group is assigned high risk more than the privileged group. In the plot, the absence of the unprivileged group for some CTIR bins is not because of 0 risk assignment, but because no individual from the group belongs to that CTIR bin. The other conditions used for checking equality of outcome and generating the plots were `ANNUITY_TO_INCOME_RATIO` and `CREDIT_TO_ANNUITY_RATIO`, and are given in `FairnessMetrics.ipynb`.

From the above, it can be concluded that there is no bias against Female group, i.e., no bias against the unprivileged subgroup of `CODE_GENDER`. There is, however, bias in decisions given a region or age (or occupation or employment duration, as explored in `FairnessMetrics.ipynb`).

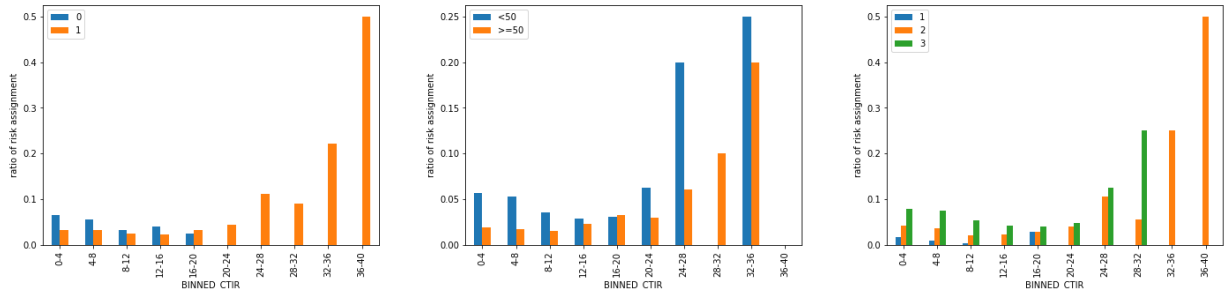


Figure 8: Conditional Risk Assignment for `CODE_GENDER`, `AGE_RANGE` and `REGION_RATING_CLIENT` given Credit To Income Ratio bins.

4.5 Stability

As mentioned in Section 3.3, this ADS was assesed with k-fold cross validation. As can be seen in Figure 9, the ROC curves for each fold almost completely overlap each other, nearly indistinguishable from the mean ROC curve. This shows that regardless of the train/validation split, the model does not overfit the data, indicating that the model is very stable. Given this result, the model will likely achieve similar performance on the unseen test set. The model achieves an AUC of 0.8, indicating that it is not just guessing at random.

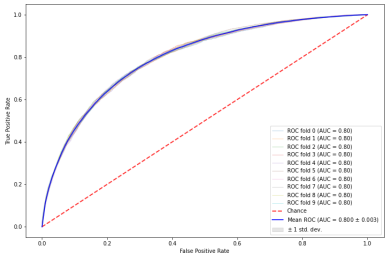


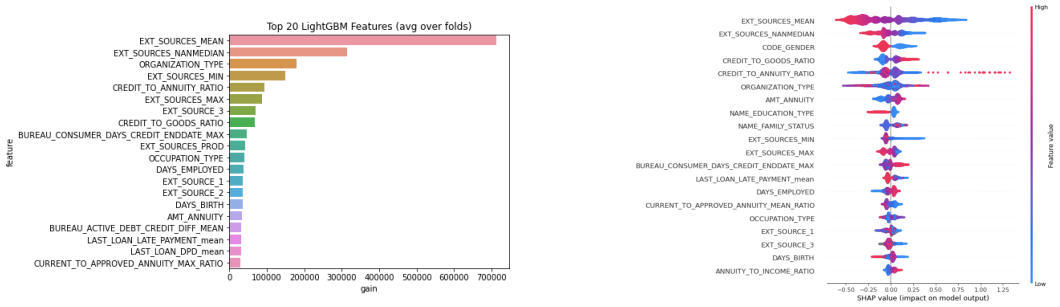
Figure 9: ROC curves plotted with k-fold cross validation for $k = 10$.

4.6 Explainability

As an ADS that aims to assign credit default risk to loan applicants, it is very important that the model predictions can be explained. In Figure 10a, we see the top 20 most predictive features, sorted by information gain. We see that quite a few of the top features are some form of `EXT_SOURCE`, whether they be features originally present in the application data, or a hand-crafted feature created by some form of aggregation. The description file for the dataset merely states that these are "Normalized score from external data source", without offering much other information about what exactly the score represents. We also see some other expected features, such as `CREDIT_TO_ANNUITY_RATIO` and `DAYS_EMPLOYED`, but these are not ranked anywhere near `EXT_SOURCES`.

In addition to using feature importance, we also use SHAP (Lundberg et al. 2020) to help better explain the model prediction of assigning individuals with probability of high default risk. SHAP values are a way that help explain model predictions for individuals by assigning importance to each feature. Here we see that as expected, `CREDIT_TO_ANNUITY_RATIO` is very important for predicting applicant assignment as high risk. Another observation is that `CODE_GENDER` is more important in predicting applicant as not high risk. We also see a similar behavior for `EXT_SOURCES`, which contribute a lot in model prediction, but mostly for assigning the probability that `TARGET`=0. This is an interesting point, because this highly predictive feature may be very relevant or not so much depending on whether the the user of this ADS is more invested in identifying low default risk applicants or high default risk applicants. In fact, many of the features listed in Figure 10b, when colored as having high value, tend to have a a lower SHAP value, contributing more to predicting the applicant as having low default risk. There are only a few features,

such as CREDIT_TO_ANNUITY_RATIO, CREDIT_TO_GOODS_RATIO, and ANNUITY_TO_INCOME_RATIO that are very important in contributing to predicting an applicant as high risk.

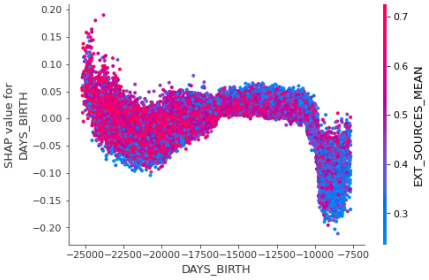


(a) Sorted by information gain.

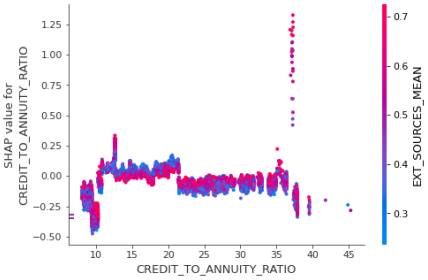
(b) SHAP summary plot

Figure 10: Feature Importance

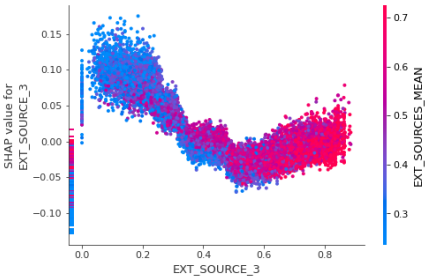
We further explore how a given feature affects the the model predictions for assigning a higher probability that the applicant has high default risk, and the interaction with other features. As shown in Figure 11, we select several features from 10b to examine. The DAYS_BIRTH variable is negative, given as the number of days before date of application that the applicant was born, so the more negative the number, the older the applicant is. We observe in Figure 11a that the older applicant age tend to also have higher scores in the EXT_SOURCE_MEAN, and applicants that are high on both also trend towards having a higher SHAP value for predicting high default risk. In Figure 11b, we see that actually, most of the values for credit to annuity ratio produce a rather low SHAP value. However, there is a group of applicants that score about 37 for the CREDIT_TO_ANNUITY_RATIO, and score high on EXT_SOURCES_MEAN producing a high SHAP value. We further examine EXT_SOURCES in Figure 11c. We see here that having a lower score in EXT_SOURCE_3 also makes this feature more important in classifying an applicant as high risk, while a higher score will result in a lower SHAP value. In Figure 11d, we observe that the number of days the applicant has been employed does not really have much interaction effect with EXT_SOURCES_MEAN. The less days the applicant has been employed, the higher the SHAP value is for the DAYS_EMPLOYED feature for assigning the applicant with high default risk. From these graphs we see that although EXT_SOURCES_MEAN is a rather nondescript feature shown as a very important for this model's predictions, its interactions with some other more easily accessible features help explain the model predictions.



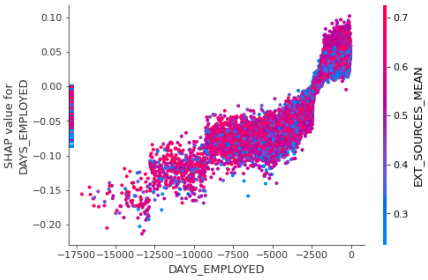
(a) Days of birth.



(b) Credit to annuity ratio



(c) External source 3



(d) Days employed

Figure 11: SHAP feature dependence and interaction with EXT_SOURCE_MEAN

5 Summary

We have evaluated the ADS created for the Home Credit in terms of the input data, the LightGBM model, and the predictions produced by this ADS. The dataset provided for this ADS is extremely comprehensive for each applicant. Since Home Credit is meant to help people with insufficient credit histories, it makes sense that the dataset includes information beyond traditional income and credit information. However, a concern is that although the applicant's name is anonymized, and other sensitive information not displayed in the data, the detailed information provided on each applicant may potentially be deanonymized by a determined attacker, raising privacy issues. During feature engineering, many of the less relevant features, such as the booleans about whether or not an applicant provided a certain piece of information or document are dropped, indicating that there is far more information provided than needed for the ADS. Furthermore, as seen in Section 4.6, much of the ADS prediction is driven by the scores given by the `EXT_SOURCES` features. In light of this finding, future analysis can be done to better refine and prune out the less relevant features. This way, data for only the most important features need to be collected, reducing privacy concerns, and can greatly speed up processing. With the less relevant features out of the way, the ADS will also become more transparent.

This model was trained with k-fold cross-validation, and the results do not vary much from fold to fold, indicating that the model is very stable. While the model does not discriminate against women (owing to the skew towards female population), it discriminates on other bases such as the region where the client lives, his/her age, occupation and years of experience. Since only a small part of the population is assigned high risk the impact is not widespread, but in that small segment, unprivileged groups are more compromised. The model is also slightly less accurate for unprivileged groups. This can be harm for members in the unprivileged group seeking a loan and for Home Credit, who is looking to profit. As mentioned in Section 4.1, the choice of threshold is based on the balance of minimization of risk (benefiting Home Credit) and increasing the number of accepted applications (benefiting applicants, but potentially causing losses for Home Credit). For this model, there is no threshold where the target-wise accuracies for both risk and no-risk targets are high. The overall accuracy is over 91%, however, the accuracy for risk target is very low as compared to no-risk target.

Although the Kaggle solution we evaluate here achieved 7th place on the Kaggle leaderboard with an overall AUC of 0.80, we would not recommend directly deploying this ADS for use by Home Credit. This ADS may be a good supporting tool to identify new customers in the loan applicants with the potential to repay their loans, but the final decision would still need to be determined by a human, to have an appropriate classification threshold based on whether they care more about getting a new customer or avoiding giving a loan to an applicant with high default risk.

References

Ke, Guolin et al. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems*, pp. 3146–3154.

Lundberg, Scott M. et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 2522–5839.

Appendix A

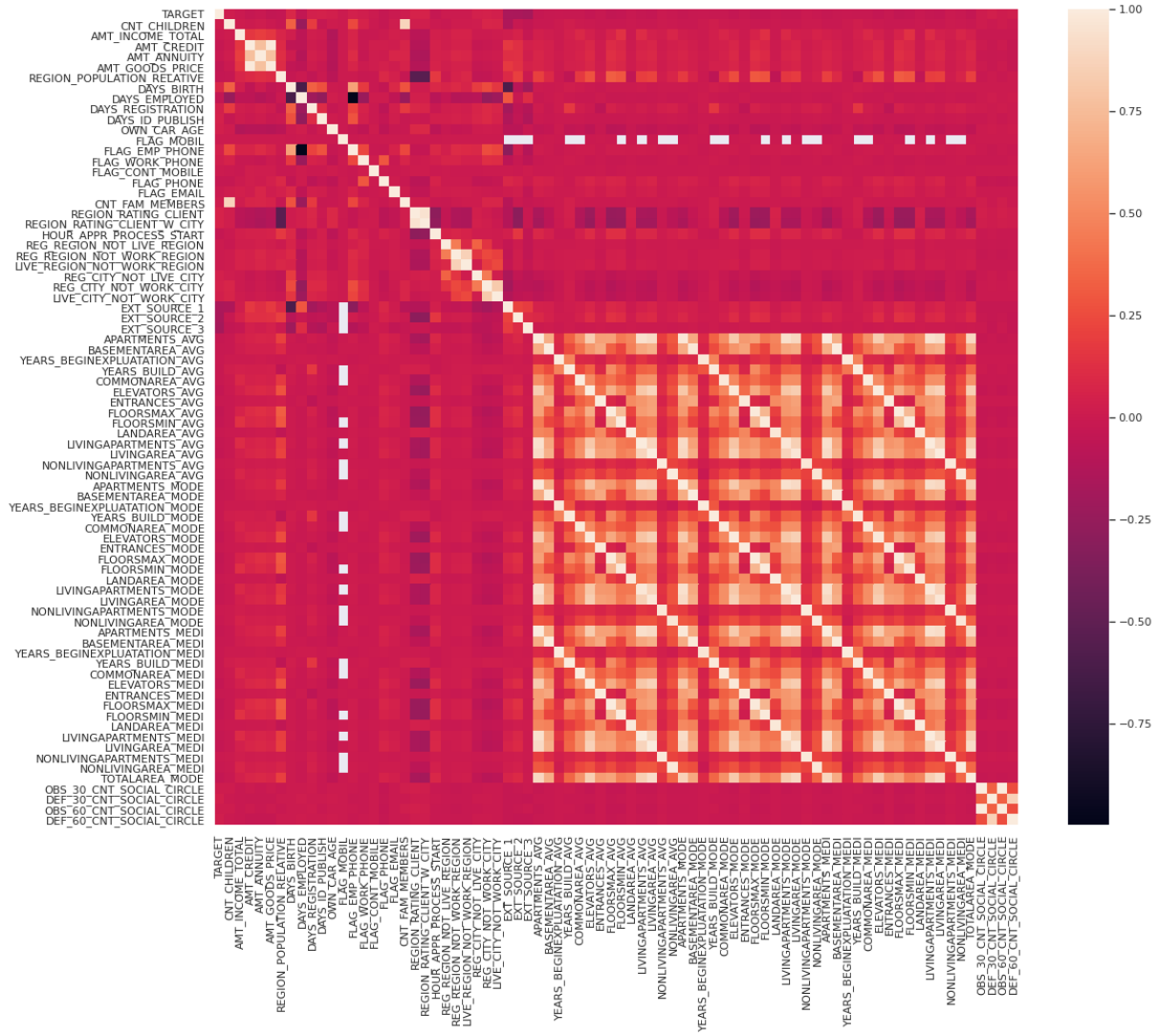


Figure 12: Pairwise Correlation Heatmap for Numerical Attributes

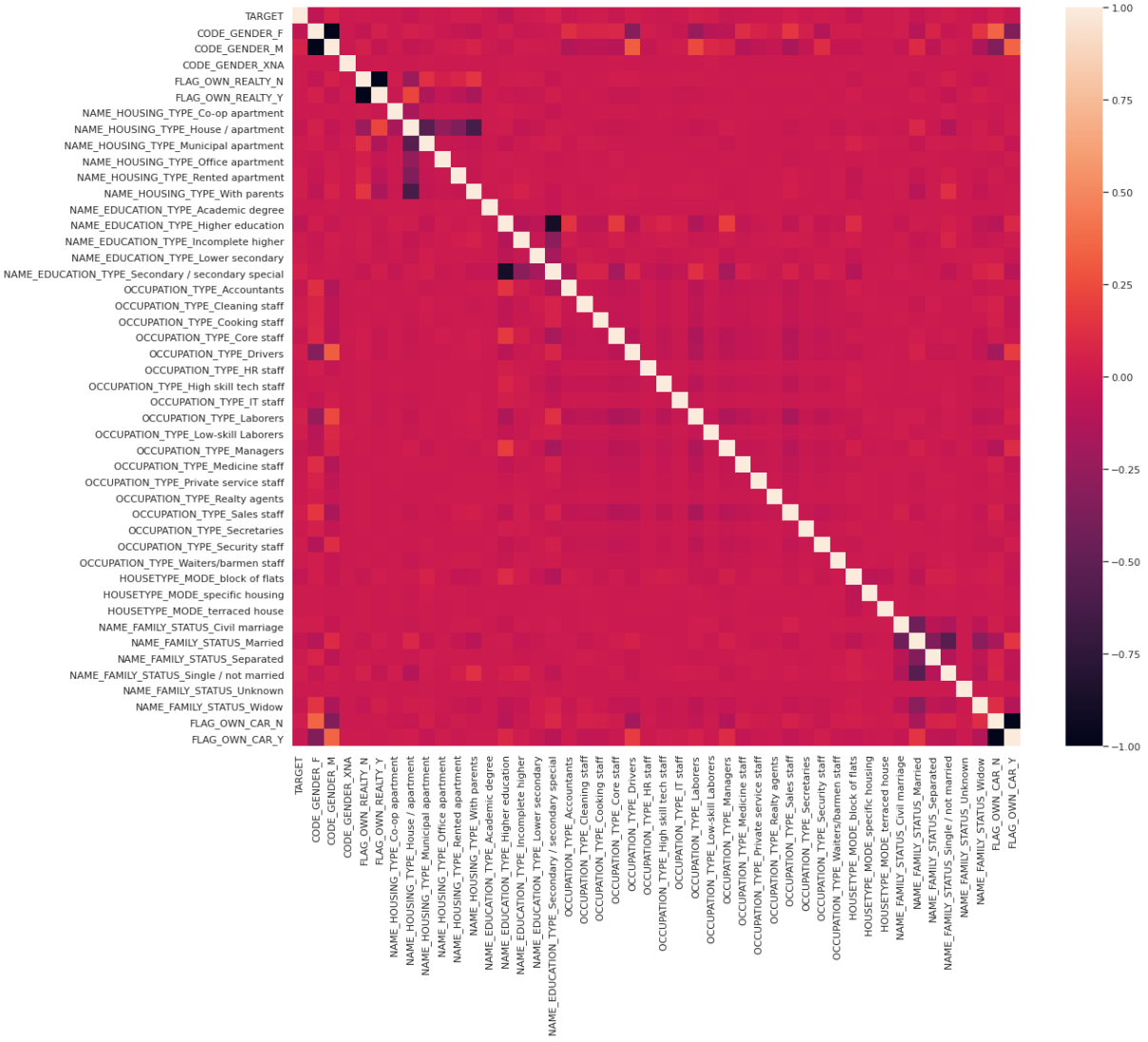


Figure 13: Pairwise Correlation Heatmap for Non-numerical, Categorical Attributes

Appendix B

Table 3: Input Features

FEATURE	DESCRIPTION
CODE_GENDER	Gender code F has 2024484 (66%) and M has 105059 (34%) rows. 4 rows with value XNA were dropped. Non-numerical, Categorical
FLAG_OWN_CAR	202924 (66%) N rows and 104587 (34%) Y rows. Non-numerical, Categorical
NAME_CONTRACT_TYPE	278232 (90.48%) cash loans and 29279 (9.52%) revolving loans. Non-numerical, Categorical
NAME_EDUCATION_TYPE	5 categories: Secondary / secondary special (71%), Higher education (24%), Incomplete higher (3%), Lower secondary, (1%), Other; no ordinality in distribution.Non-numerical, Categorical
Continued on next page	

Table 3 – continued from previous page

FEATURE	DESCRIPTION
NAME_FAMILY_STATUS	Married (64%), Single (15%), Civil Marriage (10%), Separated (6%), Widow(5%) and Unknown. Non-numerical, Categorical
NAME_HOUSING_TYPE	House / apartment (89%), With parents (5%), Municipal apartment (4%), rented (2%), office, co-op (1%).Non-numerical, Categorical
NAME_INCOME_TYPE	Working (52%), Commercial associate (23%), Pensioner (18%), State servant (7%), Unemployed, Student, Businessman, Maternity leave. Non-numerical, Categorical
OCCUPATION_TYPE	18 different occupations. 96400 (31%) missing values. Non-numerical, Categorical
ORGANIZATION_TYPE	58 different organization types. 22% business entity type, 18% self-employed, 18% XNA values. No other missing values. Non-numerical, Categorical
WEEKDAY_APPR_PROCESS_START	Day for starting process. 7 days of the week, 5 weekdays have nearly the same number of rows, weekend days have fewer. Non-numerical, Categorical
NAME_TYPE_SUITE	Unaccompanied (81%), Family (13%), Spouse (4%), Children (1%), Other_B, Other_A, Group of people. 1292 missing values. Non-numerical, Categorical
WALLSMATERIAL_MODE	Material of walls. Panel (21%), Stone/brick (21%), Block (3%), Wooden, Mixed, Monolithic (4%). 156000 (51%) rows have missing values. Non-numerical, Categorical
AMT_INCOME_TOTAL	Total income amount. Range=[25.6k, 117m], Mean=169k, StdDev= 237k, 25/50/75 Quantiles=[113k, 147k, 203k]. Outliers were excluded in data cleaning. Numerical, Continuous, Integer
DAYS_EMPLOYED	Days employed counted from day of application (negative). Replaced 365243 (invalid value since attribute here has negative values) with NaN. 25/50/75 Quantiles=[-2760, -1213, -289]. Numerical, Continuous, Integer
DAYS_LAST_PHONE_CHANGE	Day since last phone was changed from day of application (negative). Replaces 0 with NaN. Numerical, Continuous, Integer
DAYS_BIRTH	Days since birth counted from day of application (negative). Range=[-25.2k, -7489], Mean=-16k, StdDev=4.36k, 25/50/75 Quantiles=[-19.7k, -15.8k, -12.4k]. A rough, spread-out bell distribution. Numerical, Continuous, Integer
DAYS_REGISTRATION	Days since registration counted from day of application (negative). Range=[-24.7k, 0], Mean=-4.99k, StdDev=3.52k, 25/50/75 Quantiles=[-7.48k, -4.5k, -2.01k]. Increasing as days counted from application decrease. Numerical, Continuous, Integer
DAYS_ID_PUBLISH	Days since id piblished counted from day of application (negative). Range=[-7197, 0], Mean=-2.99k, StdDev=1.51k, 25/50/75 Quantiles=[-4299, -3254, -1720]. Numerical, Continuous, Integer
Continued on next page	

Table 3 – continued from previous page

FEATURE	DESCRIPTION
AGE_RANGE	Labels 1 to 5 generated from DAYS_BIRTH; 5 groups from age 0 to 99 (0 fo other cases). Numerical, Categorical, Integer
FLAG_DOCUMENT_x	Indicator for having a document (1) or not (0).x=2,3,4,... Numerical, Categorical, Boolean
DOCUMENT_COUNT	Counts documents by taking sum of FLAG_DOCUMENT_x values for all x. Numerical, Continuous, Integer
EXT_SOURCE_x	External Sources; x=1,2,3; bell distributions, range=[0, 1]. Missing values present. Numerical, Continuous, Float
EXT_SOURCES_PROD	Product of EXT_SOURCE_x for all x. Numerical, Continuous, Float
EXT_SOURCES_WEIGHTED	Weighted sum of EXT_SOURCE_x for all x. Numerical, Continuous, Float
AMT_CREDIT	Credit amount. Range=[45k, 4.05m], Mean=599k, StdDev=402k, 25/50/75 Quantile=[270k, 514k, 809k]. Numerical, Continuous, Float
AMT_ANNUITY	Annuity amount. Range=[1.62k, 258k], Mean=27.1k, StdDev=14.5k, 25/50/75 Quantile=[16.5k, 24.9k, 34.6k]. 12 missing values.Numerical, Continuous, Float
AMT_GOODS_PRICE	Goods price amount. Range=[40.5k, 4.05m], Mean=538k, StdDev=369k, 25/50/75 Quantile=[239k, 450k, 680k]. 278 missing values. Numerical, Continuous, Float
OWN_CAR_AGE	Car ownership age. Range=[0, 91], Mean=12.06, StdDev=11.94, 25/50/75 Quantiles=[5, 9, 15] Numerical, Categorical, Integer
FLAG_WORK_PHONE	Indicator for having work phone (1) (20%) or not (0) (80%). Numerical, Categorical, Boolean
REGION_RATING_CLIENT	Rating for region where client lives, from 1 to 3 (73.8%, 15.7%, 10.5%). Numerical, Categorical, Integer
REGION_RATING_CLIENT_W_CITY	Rating for region where client lives with respect to city, from 1 to 3 (74.63%, 14.26%, 11.11%). Numerical, Categorical, Integer
LIVE_REGION_NOT_WORK_REGION	Indicator for work and living regions being different (1) (99.96%) or not (0) (0.04%). Numerical, Categorical, Boolean
LIVE_CITY_NOT_WORK_CITY	Indicator for work and living cities being different (1) (99.82%) or not (0) (0.18%). Numerical, Categorical, Boolean
APARTMENTS_AVG	Normalized average apartment count. Range=[0, 1], Mean=0.117, StdDev=0.108, 25/50/75 Quantiles=[0.0577, 0.0876, 0.1485]. 156k values missing. Numerical, Continuous, Float
ENTRANCES_AVG	Normalized average entrance count. Range=[0, 1], Mean=0.1497, StdDev=0.1, 25/50/75 Quantiles=[0.069, 0.1379, 0.2069]. 155k values missing. Numerical, Continuous, Float

Continued on next page

Table 3 – continued from previous page

FEATURE	DESCRIPTION
FLOORSMAX_AVG	Normalized average floor max. Range=[0, 1], Mean=0.226, StdDev=0.144, 25/50/75 Quantiles=[0.1667, 0.1667, 0.3333]. 153k values missing. Numerical, Continuous, Float
LIVINGAREA_AVG	Normalized average living area. Range=[0, 1], Mean=0.107, StdDev=0.11, 25/50/75 Quantiles=[0.0453, 0.0745, 0.1299]. 154k values missing. Numerical, Continuous, Float
DEF_30_CNT_SOCIAL_CIRCLE	Number of observations of client's social surroundings who defaulted on 30 DPD (days past due). Value = 0, 1, 2, 3, Number of rows for each value decreases as value increases (88.5 for 0, 0.092 for 1, 0.017 for 2 and so on). Numerical, Categorical, Integer
DEF_60_CNT_SOCIAL_CIRCLE	Number of observations of client's social surroundings who defaulted on 60 DPD (days past due). Value = 0, 1, 2, 3, Number of rows for each value decreases as value increases (91.59 for 0, 0.071 for 1, 0.01 for 2 and so on). Numerical, Categorical, Integer
CREDIT_TO_ANNUITY_RATIO	AMT_CREDIT / AMT_ANNUITY. Numerical, Continuous, Float
CREDIT_TO_GOODS_RATIO	AMT_CREDIT / AMT_GOODS_PRICE. Numerical, Continuous, Float
ANNUITY_TO_INCOME_RATIO	AMT_ANNUITY / AMT_INCOME_TOTAL. Numerical, Continuous, Float
CREDIT_TO_INCOME_RATIO	AMT_CREDIT / AMT_INCOME_TOTAL. Numerical, Continuous, Float
INCOME_TO_EMPLOYED_RATIO	AMT_INCOME_TOTAL / DAYS_EMPLOYED. Numerical, Continuous, Float
INCOME_TO_BIRTH_RATIO	AMT_INCOME_TOTAL / DAYS_BIRTH. Numerical, Continuous, Float
EMPLOYED_TO_BIRTH_RATIO	DAYS_EMPLOYED / DAYS_BIRTH. Numerical, Continuous, Float
ID_TO_BIRTH_RATIO	DAYS_ID_PUBLISH / DAYS_BIRTH. Numerical, Continuous, Float
CAR_TO_BIRTH_RATIO	OWN_CAR_AGE / DAYS_BIRTH. Numerical, Continuous, Float
CAR_TO_EMPLOYED_RATIO	OWN_CAR_AGE / DAYS_EMPLOYED. Numerical, Continuous, Float
PHONE_TO_BIRTH_RATIO	DAYS_LAST_PHONE_CHANGE / DAYS_BIRTH. Numerical, Continuous, Float
(from bureau and bureau_balance CSVs)	
DAYS_CREDIT	Number of days before current application that client applied for Credit Bureau credit (negative). Numerical, Continuous, Integer
CREDIT_DAY_OVERDUE	Number of days past due on CB credit at the time of application for related loan in our sample. Numerical, Continuous, Integer
DAYS_CREDIT_ENDDATE	Remaining duration of CB credit (in days) at the time of application in Home Credit (negative). Numerical, Continuous, Integer
DAYS_ENDDATE_FACT	Days since CB credit ended at the time of application in Home Credit (only for closed credit) (negative). Numerical, Continuous, Integer

Continued on next page

Table 3 – continued from previous page

FEATURE	DESCRIPTION
AMT_CREDIT_SUM	Current credit amount for the Credit Bureau credit. Numerical, Continuous, Float
AMT_CREDIT_SUM_DEBT	Current debt on Credit Bureau credit. Numerical, Continuous, Float
AMT_ANNUITY	Annuity of the Credt Bureay credit. Range=[0, 118m], Mean=15.7k, StdDev=326k, 25/50/75 Quantile=[0, 0, 13.5k]. 12 missing values. 1.23m missing values. Numerical, Continuous, Float
CREDIT_DURATION	(-DAYS_CREDIT) + DAYS_CREDIT_ENDDATE. Numerical, Continuous, Float
ENDDATE_DIF	DAYS_CREDIT_ENDDATE - DAYS_ENDDATE_FACT. Numerical, Continuous, Float
DEBT_PERCENTAGE	AMT_CREDIT_SUM / AMT_CREDIT_SUM_DEBT. Numerical, Continuous, Float
DEBT_CREDIT_DIFF	AMT_CREDIT_SUM - AMT_CREDIT_SUM_DEBT. Numerical, Continuous, Float
CREDIT_TO_ANNUITY_RATIO	AMT_CREDIT_SUM / AMT_ANNUITY. Numerical, Continuous, Float
STATUS_12345	Flag months with late payments by adding STATUS_1 to STATUS_5 values. Numerical, Continuous, Integer
(from previous$_{a}$*pplication.csv*)	
NAME_CONTRACT_STATUS	Contract status during the month: Approved (62%), Canceled (19%), Refused and other (19%). Non-numerical, Categorical
NAME_CONTRACT_TYPE	Cash loans (45%), Consumer loans(44%), Revolving loans and other (12%). Non-numerical, Categorical
NAME_YIELD_GROUP	Missing/ XNA (31%), middle (23%), low_action, high, low_normal (46%). Non-numerical, Categorical
PRODUCT_COMBINATION	Cash (17%), POS household with interest (16%), Others (Cash X-sell: low/middle/high, POS mobile with interest, etc.) (67%). Non-numerical, Categorical
NAME_PRODUCT_TYPE	Missing (XNA) (64%), x-sell (27%), walk-in (9%). Non-numerical, Categorical
NAME_CLIENT_TYPE	Repeater (74%), New (18%), Refreshed, etc. (8%). Non-numerical, Categorical
AMT_APPLICATION	Credit amount applied for on the previous application. Numerical, Continuous, Float
AMT_ANNUITY	Annuity of previous application. Numerical, Continuous, Float
AMT_CREDIT	Final credit amount on the previous application. This differs from AMT_APPLICATION in a way that the AMT_APPLICATION is the amount for which the client initially applied for, but during the approval process he could have received different amount - AMT_CREDIT. Numerical, Continuous, Float
AMT_DOWN_PAYMENT	Range=[-0.9, 3.06m], Mean=6.7k, StdDev=20.9k, 25/50/75 Quantiles=[0, 1.64k, 7.74k]. 896k values missing. Numerical, Continuous, Float
APPLICATION_CREDIT_DIFF	AMT_APPLICATION - AMT_CREDIT. Numerical, Continuous, Float
APPLICATION_CREDIT_RATIO	AMT_APPLICATION / AMT_CREDIT. Numerical, Continuous, Float

Continued on next page

Table 3 – continued from previous page

FEATURE	DESCRIPTION
CREDIT_TO_ANNUITY_RATIO	AMT_CREDIT/AMT_ANNUITY. Numerical, Continuous, Float
DOWN_PAYMENT_TO_CREDIT	AMT_DOWN_PAYMENT / AMT_CREDIT. Numerical, Continuous, Float
REMAINING_DEBT	AMT_CREDIT - AMT_PAYMENT. Numerical, Continuous, Float
REPAYMENT_RATIO	AMT_PAYMENT / AMT_CREDIT. Numerical, Continuous, Float
LATE_PAYMENT	1 if DAYS_ENTRY_PAYMENT - DAYS_INSTALMENT >0, 0 otherwise. Numerical, Categorical, Boolean
(from POS_CASH_balance.csv)	
SK_DPD	DPD (days past due) during the month of previous credit. Numerical, Continuous, Integer
LATE_PAYMENT	1 if SK_DPD>0, 0 otherwise. Numerical, Categorical, Boolean
MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date). Numerical, Continuous, Integer
CNT_INSTALMENT	Term of previous credit (can change over time). Numerical, Continuous, Integer
CNT_INSTALMENT_FUTURE	Installments left to pay on the previous credit. Numerical, Continuous, Integer
NAME_CONTRACT_STATUS	Contract Status: Active (91%), Completed (7%), Signed (1%), Demand, Other. Non-numerical, Categorical
POS_LOAN_COMPLETED_MEAN	mean(NAME_CONTRACT_STATUS = Completed). Numerical, Continuous, Integer
POS_REMAINING_INSTALMENTS	last CNT_INSTALMENT_FUTURE. Numerical, Continuous, Integer
(from installments_payments.csv)	
AMT_INSTALMENT	Prescribed installment amount of previous credit on this installment. Numerical, Continuous, Float
AMT_PAYMENT	Amount the client actually paid on previous credit on this installment. Numerical, Continuous, Float
DAYS_ENTRY_PAYMENT	Installments of previous credit paid actually (relative to application date of current loan; negative number). Numerical, Continuous, Integer
DAYS_INSTALMENT	Time when installment of previous credit was supposed to be paid (relative to application date of current loan; negative number). Numerical, Continuous, Integer
AMT_PAYMENT_GROUPED	ID wise sum of AMT_PAYMENT. Numerical, Continuous, Float
PAYMENT_DIFFERENCE	AMT_INSTALMENT - AMT_PAYMENT_GROUPED. Numerical, Continuous, Float
PAYMENT_RATIO	AMT_INSTALMENT / AMT_PAYMENT_GROUPED. Numerical, Continuous, Float
PAID_OVER_AMOUNT	AMT_PAYMENT - AMT_INSTALMENT. Numerical, Continuous, Float
PAID_OVER	1 if PAID_OVER_AMOUNT >0, 0 otherwise. Numerical, Categorical, Boolean

Continued on next page

Table 3 – continued from previous page

FEATURE	DESCRIPTION
DPD	Days past due date; max(0, DAYS_ENTRY_PAYMENT - DAYS_INSTALMENT). Numerical, Continuous, Integer
DBD	Days before due date; max(0, DAYS_INSTALMENT - DAYS_ENTRY_PAYMENT). Numerical, Continuous, Integer
LATE_PAYMENT	1 if DBD>0, 0 otherwise. Numerical, Categorical, Boolean
INSTALMENT_PAYMENT_RATIO	AMT_PAYMENT / AMT_INSTALMENT. Numerical, Continuous, Float
LATE_PAYMENT_RATIO	INSTALMENT_PAYMENT_RATIO if LATE_PAYMENT = 1, else 0. Numerical, Continuous, Float
SIGNIFICANT_LATE_PAYMENT	1 if LATE_PAYMENT_RATIO >0.05 else 0. Numerical, Categorical, Boolean
(from credit_card_balance.csv)	
AMT_RECEIVABLE	Amount receivable on the previous credit. Numerical, Continuous, Integer
AMT_BALANCE	Balance during the month of previous credit. Numerical, Continuous, Float
AMT_CREDIT_LIMIT_ACTUAL	Credit card limit during the month of the previous credit. Numerical, Continuous, Float
AMT_PAYMENT_CURRENT	Amount paid by the client during the month on the previous credit. Numerical, Continuous, Float
AMT_INST_MIN_REGULARITY	Minimal installment for this month of the previous credit. Numerical, Continuous, Float
AMT_DRAWINGS_ATM_CURRENT	Amount drawing at ATM during the month of the previous credit. Numerical, Continuous, Float
LIMIT_USE	AMT_BALANCE - AMT_CREDIT_LIMIT_ACTUAL. Numerical, Continuous, Float
DRAWING_LIMIT_RATIO	AMT_DRAWINGS_ATM_CURRENT - AMT_CREDIT_LIMIT_ACTUAL. Numerical, Continuous, Float