NEW YORK UNIVERSITY

Responsible Data Science
(DS_GA_1017)

Final Report - Spring 2022

Nutritional Labels for Automated Decision Systems by Home Credit Default Risk

Wonkwon Lee (wl2733),
Soowhan Park (sp6682)

1 Background

1.1 What is the purpose of this ADS? What are its stated goals?

The ADS aims to provide a comprehensive machine learning model for beginners, using Linear Regression and the Random Forest model. The goal is to train a model to learn all the features after encoding the categorical variables and predict the binary target label, 'TARGET' indicating 0: the loan was repaid or 1: the loan was not repaid. The code for the ADS can be found atStart Here: A Gentle Introduction.

1.2 If the ADS has multiple goals, explain any trade-offs that these goals may introduce.

The author has published multiple ADS that gradually get more complex and sophisticated for a better score on the Kaggle competition. The most sophisticated version, LightGBM, has the highest submission score of 0.754, but it includes many features out of scope for the project. Although the gradient boosting models are the leading ones for structured datasets like Kaggle competitions, they are more expensive to implement and are not appropriate for our project. Thus, our group focused on less sophisticated models : Logistic Regression and Random Forest. Among the two models, we finalized to use Random Forest as it has slightly higher submission score and is better to apply AIF360 and LIME libraries for the analysis. However, as we go through the project, we have found out there exists an imbalanced class problem where the total number of a class of data (positive) is far less than the total number of another class of data (negative).

2 Input and Output

2.1 Describe the data used by this ADS. How was this data collected or selected?

All the data is collected and provided by Home Credit. The access of data is granted under non-commercial purposes including academic research and education.

1. application_train|test.csv
    a) The main training (with TARGET) and testing (without TARGET) data with information about each loan application at Home Credit.
    b) Each row represents one loan, identified by the feature SK_ID_CURR.
    c) Only training data includes TARGET indicating 0: the loan was repaid or 1: the loan was not repaid. This is static data for all applications.
    d) Sensitive attributes are included such as gender, employment status, income status, education status, and family status.
2. bureau.csv
    a) Data concerning the client's previous credits from other financial institutions that were reported to the Credit Bureau before the application date.
    b) Each previous credit from the client in our sample has its own row.
3. bureau_balance.csv
    a) Data concerning the monthly balances about every previous credit in the Credit Bureau.
    b) Each row represents one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.

4. previous_application.csv
    a) Data concerning all the previous applications for loans at Home Credit of clients who have loans in the application data.
    b) Each current loan in the application data can have multiple previous loans.
    c) Each previous application has one row and is identified by the feature SK_ID_PREV.
5. POS_CASH_BALANCE.csv
    a) Data concerning the monthly balance about previous point of sale or cash loans clients had with Home Credit.
    b) Each row represents a month of a previous point of sale or cash loan, and a single previous loan can have multiple rows.
6. credit_card_balance.csv
    a) Data concerning the monthly balance of previous credit cards that clients had with Home Credit.
    b) Each row represents a month of a credit card balance, and a single credit card can have multiple rows.
7. installments_payment.csv
    a) Data concerning the repayment history for previous loans at Home Credit.
    b) There is one row for every payment that was made and one row for every missed payment.

2.2 For each input feature, describe its datatype, give information on missing values and on the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting and appropriate.

Data Types	Counts
float64	65
int64	41
object	16

Table 1: Datatypes for all input features

There are 106 columns with numerical values and 16 columns with categorical values that need to be encoded.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8

Figure 1: Snippet of missing values[1]

There are 67 columns that have missing values out of 122 columns. About 54.9% of the columns have missing values.

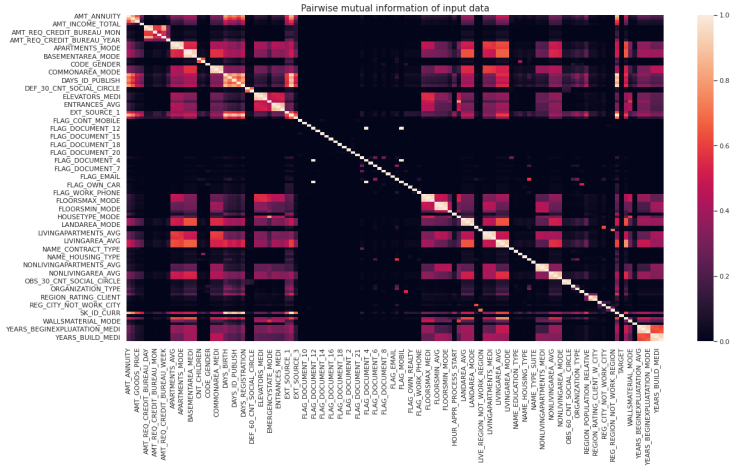

Figure 2: Pairwise correlations in application_train dataset

Pairwise correlation between features are computed and visualized with a heatmap. Because the training data contains 122 columns with large size, 10 percent of the training data was randomly sampled and computed.

```
Most Positive Correlations:
 OCCUPATION_TYPE_Laborers                                0.043019
FLAG_DOCUMENT_3                                          0.044346
REG_CITY_NOT_LIVE_CITY                                   0.044395
FLAG_EMP_PHONE                                           0.045982
NAME_EDUCATION_TYPE_Secondary / secondary special        0.049824
REG_CITY_NOT_WORK_CITY                                   0.050994
DAYS_ID_PUBLISH                                          0.051457
CODE_GENDER_M                                            0.054713
DAYS_LAST_PHONE_CHANGE                                   0.055218
NAME_INCOME_TYPE_Working                                 0.057481
REGION_RATING_CLIENT                                     0.058899
REGION_RATING_CLIENT_W_CITY                              0.060893
DAYS_EMPLOYED                                            0.074958
DAYS_BIRTH                                               0.078239
TARGET                                                   1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                          -0.178919
EXT_SOURCE_2                           -0.160472
EXT_SOURCE_1                           -0.155317
NAME_EDUCATION_TYPE_Higher education   -0.056593
CODE_GENDER_F                          -0.054704
NAME_INCOME_TYPE_Pensioner             -0.046209
DAYS_EMPLOYED_ANOM                     -0.045987
ORGANIZATION_TYPE_XNA                  -0.045987
FLOORSMAX_AVG                          -0.044003
FLOORSMAX_MEDI                         -0.043768
FLOORSMAX_MODE                         -0.043226
EMERGENCYSTATE_MODE_No                 -0.042201
HOUSETYPE_MODE_block of flats          -0.040594
AMT_GOODS_PRICE                        -0.039645
REGION_POPULATION_RELATIVE             -0.037227
Name: TARGET, dtype: float64
```

Figure 3: Top 15 positive and negative correlations

The figure 3 demonstrates the most significant correlations in the dataset: the 'DAYS BIRTH' with the highest positive correlation. 'DAYS BIRTH' represents the age in days of the client at the time of the loan in negative days. In other word, a strong positive correlation in 'DAYS BIRTH' indicates that the age has a strong negative correlation: the older the client are, the more likely they will repay the loan.

2.3 What is the output of the system (e.g., is it a class label, a score, a probability, or some other type of output), and how do we interpret it?

The ADS produces a binary classification label, predicting 0 for applicants who will repay the loan or 1 for applicants who will not repay the loan.

3 Implementation and Validation

3.1 Describe data cleaning and any other pre-processing

1. Encoding categorical variables

   - This ADS' machine learning model cannot handle the categorical values, so it is necessary to convert them to numeric values. Each unique category in a categorical variable will be replaced with an integer. Note this does not create new column.

2. Imputing missing values

   - There are 67 columns out of 122 columns that contain missing values. Although it is possible to drop the columns with high percentage of missing values, it is not an ideal approach since we do not know if the column is informative ahead of the time. Thus, the missing values are imputed with median.

3. Scaling features to a range

   - Each features are scaled into range between 0 and 1.

3.2 Give high-level information about the implementation of the system

After data profiling and exploratory data analysis, the author performed necessary prepossessing steps such as encoding a categorical variables, imputing missing values, and scaling features to a range. The logistic regression model was selected as a baseline model. The default regularization coefficient, C, is lowered to increase the accuracy. Using the trained model, the author calculated probabilities of not repaying a loan and the accuracy of the model was 0.671.

3.3 How was the ADS validated? How do we know that it meets its stated goal(s)?

The result of ADS is validated on the area under the Receiver Operating Characteristic (ROC) curve between the predicted probability and the observed target. The ROC curve graphs the true positive rate versus false positive rate. The goal of this ADS is to use historical loan application data to predict whether or not an applicant will be able to repay a loan. The final accuracy of this ADS using logistic regression is 0.67, which is a decent performance considering the accuracy of the top of the leader-board on Kaggle is around 0.79.

4 Outcomes

4.1 Analyze the effectiveness (accuracy) of the ADS by comparing its performance across different subpopulations.

Among the several protected groups in the dataset, we chose age, gender, and education status to evaluate the performance of the ADS.

1. Age

   - Applicants older or equal to 30 years old are assigned to the privileged group, and the applicants younger than 30 years old are assigned to the unprivileged group. The overall accuracy and ROC AUC are 0.919 and 0.711, respectively. Separately, the privileged group has an accuracy of 0.925, and the unprivileged group has 0.886.

2. Gender

   - Male applicants are assigned to the privileged group, and female applicants are assigned to the unprivileged group. The overall accuracy and ROC AUC are 0.918 and 0.703, respectively. Separately, the privileged group has an accuracy of 0.895, and the unprivileged group has 0.931.

3. Education

   - Applicants who attended above or equal to higher education are assigned to the privileged group, and the applicants who attended below higher education are assigned to the unprivileged group. The overall accuracy and ROC AUC are 0.92 and 0.709, respectively. Separately, the privileged group has an accuracy of 0.921, and the unprivileged group has 0.891.

4.2 Select one or several fairness or diversity measures, justify your choice of these measures for the ADS in question, and quantify the fairness or diversity of this ADS.

	Age	Gender	Education
Overall Accuracy	0.9193	0.9185	0.9201
Overall AUC	0.7113	0.7027	0.7098
Accuracy (privileged)	0.925	0.8956	0.9207
Accuracy (unprivileged)	0.8861	0.9305	0.8911
Mean Difference	0.0396	-0.031	0.0291
Disparate Impact	1.5288	0.6639	1.3623
Error rate difference	0.0394	-0.0351	0.0187
FPR (privileged)	0.000103	0.000160	0.000352
FPR (unprivileged)	0.00028	0.000132	0.0000
FPR difference	-0.000028	-0.000028	-0.000352
FNR (privileged)	0.998492	0.997714	0.999367
FNR (unprivileged)	0.997082	0.998581	1.000000
FNR ratio	0.998588	1.000869	1.000633

Table 2: Measures for fairness or diversity

For all three protected attributes, the FPR difference is nearly 0, and the FNR ratio is close to 1. The outcome is favorable to the unprivileged group for 'Age' and 'Education,' and is favorable to the privileged group for gender group. According to the disparate impact, the 'Age' group is the most biased among the other attributes, but all three attributes are relatively close to 1, so the bias is not extreme. In other word, our ADS are hardly affected by the sensitive attributes, such as age, education status, or gender. Rather, as we observe through LIME explainer later, the ADS is rather strongly affected by the external financial source or associated financial documents.

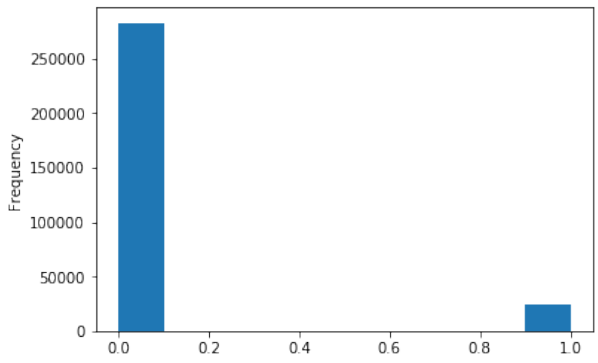

Figure 4: The frequency of target label

Further investigation was challenging as the dataset is overly imbalanced. As seen in the figure 4 , the dataset is imbalanced in the target attribute: the number of loans that were repaid on time far exceeds those were not repaid. The author points out the imbalanced class problem and suggests that he will improve the problem in the later notebooks, but he did not introduce nor apply on this system.

4.3 Develop additional methods for analyzing ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME), or any other property that you believe is important to check for this ADS.

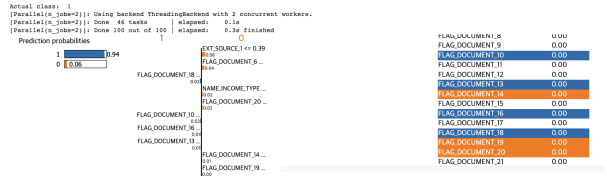

Figure 5: An example of correctly classified and its explanation

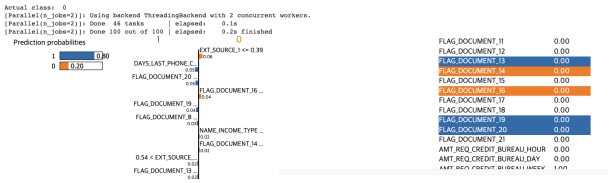

Figure 6: An example of misclassified and its explanation

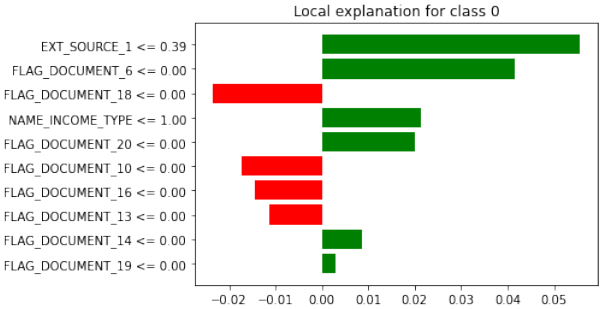

Figure 7: Snippet of LIME explanation

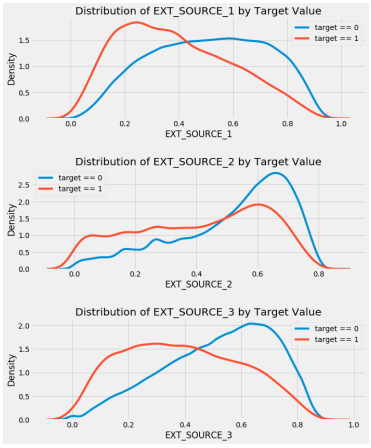

Figure 8: The distribution of external source

After investigating the fairness in the ADS, our group delve into more sophisticated aspects of the system using LIME library. The tabular explainer is used to interpret how the ADS classifies each individual. The figure 5 and 6 represent the correctly classified and the incorrectly classified, respectively. For the correctly classified individual, the ADS correctly predicts that the individual is unlikely to repay the loan (TARGET value of 1) based on the flag documents. The flag documents are the client provided documents, which contain client's previous financial history and financial assets. On the other hand, the EXT SOURCE 1 feature, which represents a normalized score from external data source (figure 8). It is not explicitly explained on the provided Kaggle explanation, but according to the author, external source features are cumulative sort of credit rating of each client. The ADS focuses a great weight for EXT SOURCE 1 value of less than 0.39, but based on the aggregated flag documents it correctly predicts that the individual is unlikely to repay.

For the misclassified example, the ADS incorrectly predicts that the individual will not pay the loan. Along with the client provided flag documents and DAYS LAST PHONE CHANGED feature, which represents how many days before application did client change phone, and EXT SOURCE of greater than 0.57, the ADS incorrectly predicts that the individual is unlikely to repay the loan.

Based on the LIME explanation, our Automated Decision System attaches more importance on the financial status or previous credit history features, rather than sensitive attributes such as gender, age, or education. Therefore, we can cautiously assume that the ADS is stable to be deployed and used for a short period of time. For robustness, our ADS is implemented using the random forest model; they are robust to outliers as they are averaged out by the aggregation of multiple decision tree output. Unfortunately, our baseline random forest model is simple and not complex enough to be used in real life. The submission accuracy was below 0.70, and the model was implemented on the imbalanced dataset. Therefore, our baseline ADS model is imperfect and inappropriate to be used in real life.

5 Summary

5.1 Do you believe that the data was appropriate for this ADS?

The data is imbalanced in the target attribute, which is heavily skewed to 0 representing applicants who repaid the loan. Thus, another preprocessing is needed to weight the classes by the representation in the data to reflect this imbalance. In fact, the author provided multiple Kaggle solutions for our Home Credit Default Credit Risk Kaggle competition[2]. He applies over-sampling or under-sampling on later solutions, but our baseline ADS is not implemented with such techniques. Furthermore, the Kaggle competition also offers detailed external dataset, regarding credit card history, bureau balance, installments, and others, but our model did not use such dataset for training. Therefore, our imbalanced training dataset was inappropriate for the ADS.

5.2 Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.

One of the greatest problem in the implementation of the ADS is that the model simply imputes all missing values with the median feature value for simplicity. All features like flag for owning own a car, a house, a realty, or credit card history are all replaced with the median. Inappropriate handling of missing data can trigger a potential bias in the system or fairness. We could not observe explicit bias between subpopulations by gender, age, and education from previous analysis. Also, the difference between the accuracy of privileged and unprivileged is hard to notice. Rather, our ADS uses features that are directly related to client's financial status for its decision

making process. Thus, in terms of sensitive attributes, our ADS is relatively accurate and fair. With further modification in model implementation and dataset preprocessing, the system would work even better. The stakeholders would be any clients who use the dataset, the supplier, and the vendor as our model is hardly affected by the protected attributes and rather use financial information to make a decision.

5.3 Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?

Although the ADS is relatively fair and accurate for different subpopulations, it is still inappropriate to be deployed neither in the public sector nor in the industry, unless it is solely for educational purposes. Missing values replaced with the median value, imbalanced class problem, and the low accuracy of our baseline ADS demonstrates that the ADS needs far more extension to be deployed. Further, most of data science competitions are only concerned about achieving the best performance on a single metric; this can potentially cause issues. Yet, the explanation from LIME shows that the ADS put weight on financial attributes than the sensitive attributes. This gives a hint of hope that the ADS is relatively fair, and with sufficient modification, it can be possibly used in the industry or in the public sector.

5.4 What improvements do you recommend to the data collection, processing, or analysis methodology?

Because of the imbalanced target class in the dataset, the author suggests over-sampling or under-sampling in his later Kaggle solutions. Such techniques are useful to improve the performance of the system, but they may fail to reflect the real distribution of the dataset. The Kaggle competition did not provide sufficient information about external sources or flag documents data, as they contain private information. Without sufficient domain knowledge about our feature may cause dropping or undermining important feature during processing. For simplicity, our baseline model did not use additional data like credit history, bureau balance, or installments. The model can be improved by using more balanced dataset and external client's financial data.

References

[1] W. Koehrsen, "Start here: A gentle introduction," *Home Credit Default Risk Competition*, 2018. [Online]. Available: https://www.kaggle.com/code/willkoehrsen/start-here-a-gentle-introduction/notebook

[2] Home_Credit, "Can you predict how capable each applicant is of repaying a loan?" *Home Credit Default Risk*. [Online]. Available: https://www.kaggle.com/competitions/home-credit-default-risk/data