Ethical Audit of the Home Credit Default Risk Algorithm

MinJoo Kim and Azrael Ning

May 9, 2024

1 Background

We will explore the realm of Home Credit, a financial service that "strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience".[1] Home credit default risk, the probability that borrowers may fail to meet their loan obligations, poses significant challenges and opportunities in ethical lending practices. By choosing the Home Credit Default Risk Automated Decision System (ADS) for our audit, we aim to critically assess the fairness and accuracy of credit scoring systems, bridging data science with ethical considerations and advocating for enhanced financial inclusivity through more equitable lending practices.

The Home Credit Default Risk ADS aims to develop predictive models that accurately assess default risk, using advanced statistical and machine learning techniques to analyze detailed data provided by Home Credit. The ADS further intends to provide insights into the factors influencing credit risk, allowing Home Credit to tailor its lending policies and strategies effectively. The Home Credit Default Risk ADS encompasses multiple goals, including but not limited to:

1. Improving accuracy in credit risk assessment
2. Minimizing default risk
3. Optimizing lending decisions

In auditing the Home Credit Default Risk ADS, our primary goals are to enhance fairness and accuracy in credit scoring, which, as with most goals, introduce several trade-offs and risks.

1. **Accuracy vs. Interpretability**: A more complex predictive model may achieve higher accuracy in identifying potential defaulters. However, it could potentially still sacrifice interpretability, making it challenging to understand and explain the factors driving the predictions.

2. **Risk Minimization vs. Inclusivity**: While the primary aim is to minimize default risk, striving for greater inclusivity might lead to approving more high-risk loans, potentially increasing the default rate, which

[1] Anna Montoya, KirillOdintsov, and Martin Kotek, “Home Credit Default Risk” (https://kaggle.com/competitions/home-credit-default-risk, 2018).

could affect the financial stability of lending institutions and their willingness to participate in such programs. Overly conservative risk assessment strategies might also result in the exclusion of potentially creditworthy applicants.

3. **Cost vs. Benefit**: Implementing sophisticated predictive models and data analysis techniques can incur significant costs in terms of resources, time, and infrastructure. However, the benefits of accurate risk assessment and optimized lending decisions, such as reduced defaults and improved profitability, may outweigh these costs in the long run.

4. **Privacy vs. Data Utility**: Utilizing extensive datasets to train predictive models raises privacy concerns regarding the sensitive information collected from loan applicants. Balancing data utility with privacy protection is essential to ensure compliance with regulatory requirements and maintain trust with customers.

5. **Short-Term vs. Long-Term Objectives**: While the immediate goal may be to minimize default risk and optimize lending decisions, it's essential to consider the long-term implications of these decisions. Prioritizing short-term gains over long-term sustainability could lead to adverse consequences.

2 Input and Output

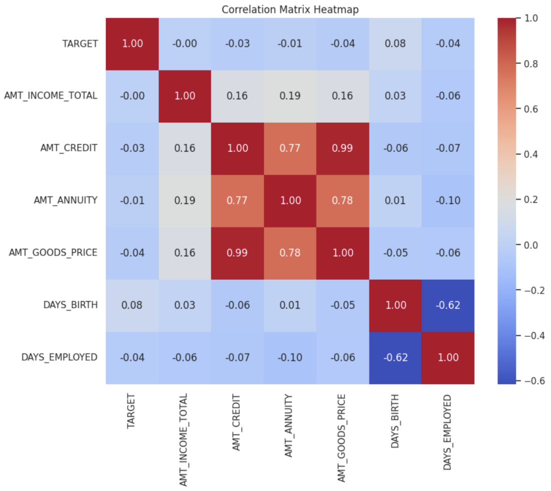
The data used by these ADS consists of various features related to loan applicants, including demographic information, financial data, and credit history. These features are collected during the loan application process and are stored in a structured format for analysis. The dataset includes information such as the applicant's age, gender, income, employment status, credit history, property ownership, and other relevant factors that can influence credit risk assessment.

The dataset under analysis comprises a diverse array of features, categorized into three primary data types: integer (int64), floating-point (float64), and object (string or categorical). Integral identifiers like SK_ID_CURR and the target variable TARGET are formatted as integers, facilitating straightforward numerical operations. Financial metrics including AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY are expressed as floating-point numbers, allowing for precision in representing monetary values. Descriptive attributes such as NAME_CONTRACT_TYPE, CODE_GENDER, and OCCUPATION_TYPE are stored as objects, indicating textual or categorical data.

The dataset shows a significant range in the presence of missing values across various features. Essential identifying information and most categorical features such as SK_ID_CURR, TARGET, and NAME_CONTRACT_TYPE have no missing values, indicating completeness in the core application data. However, certain features display a substantial number of null entries, notably in areas related to the applicant's assets and environment. For instance, OWN_CAR_AGE and various attributes detailing apartment specifics (like APARTMENTS_AVG, BASEMENTAREA_AVG) and the related MODE and

MEDI features exhibit over 150,000 missing values, which could impede the model's learning process where these features are crucial. The EXT_SOURCE features, which are external data sources, also have notable gaps (EXT_SOURCE_1 has 173,378 nulls, and EXT_SOURCE_3 has 60,965 nulls), potentially affecting the predictive power of external credit scores. Furthermore, the OBS_30_CNT_SOCIAL_CIRCLE and its variants, which likely relate to the applicant's social background, show around 1,021 missing entries, pointing to occasional gaps in social data collection.

Below are all correlations between features:



	TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED
TARGET	1.00	-0.00	-0.03	-0.01	-0.04	0.08	-0.04
AMT_INCOME_TOTAL	-0.00	1.00	0.16	0.19	0.16	0.03	-0.06
AMT_CREDIT	-0.03	0.16	1.00	0.77	0.99	-0.06	-0.07
AMT_ANNUITY	-0.01	0.19	0.77	1.00	0.78	0.01	-0.10
AMT_GOODS_PRICE	-0.04	0.16	0.99	0.78	1.00	-0.05	-0.06
DAYS_BIRTH	0.08	0.03	-0.06	0.01	-0.05	1.00	-0.62
DAYS_EMPLOYED	-0.04	-0.06	-0.07	-0.10	-0.06	-0.62	1.00

1. **TARGET vs. Other Features**: The correlations with TARGET are generally very weak, with none exceeding an absolute value of 0.1, indicating a very weak linear relationship.

2. **AMT_INCOME_TOTAL vs. Other Features**: There are weak positive correlations with AMT_ANNUITY and AMT_GOODS_PRICE, and a very weak positive correlation with AMT_CREDIT. Negligible correlations exist with DAYS_BIRTH and DAYS_EMPLOYED.

3. **AMT_CREDIT vs. Other Features**: There's a strong positive correlation with AMT_GOODS_PRICE and a weak positive correlation with AMT_ANNUITY. Negligible correlations exist with DAYS_BIRTH and DAYS_EMPLOYED.

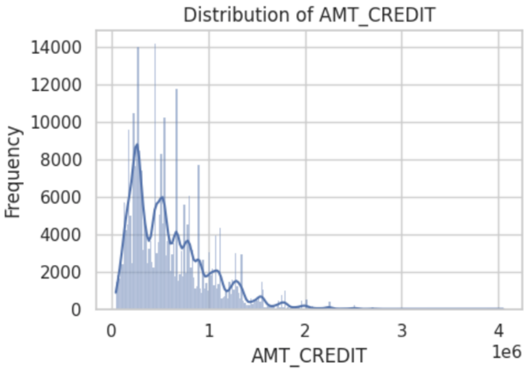
4. **AMT_ANNUITY vs. Other Features**: Moderate positive correlation with AMT_CREDIT and weak positive correlations with AMT_GOODS_PRICE and AMT_INCOME_TOTAL. Negligible correlations exist with DAYS_BIRTH and DAYS_EMPLOYED.

5. **AMT_GOODS_PRICE vs. Other Features**: Strong positive correlation with AMT_CREDIT and weak positive correlations with AMT_ANNUITY and AMT_INCOME_TOTAL. Negligible correlations exist with DAYS_BIRTH and DAYS_EMPLOYED.

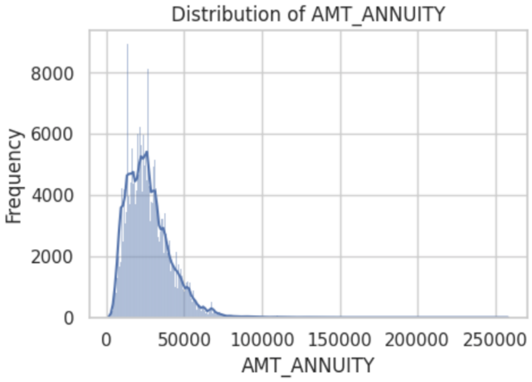
6. **DAYS_BIRTH vs. Other Features**: There's a very weak negative correlation with TARGET, and negligible correlations with all other features.

7. **DAYS_EMPLOYED vs. Other Features**: Very weak negative correlation with TARGET, moderate negative correlation with DAYS_BIRTH, and negligible correlations with other features.

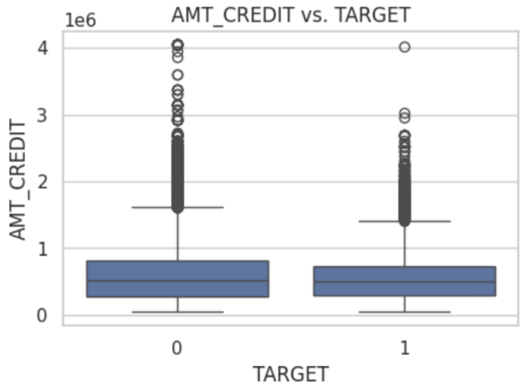
The following analysis provides insights into the distribution of key financial and demographic variables of loan applicants within the dataset.



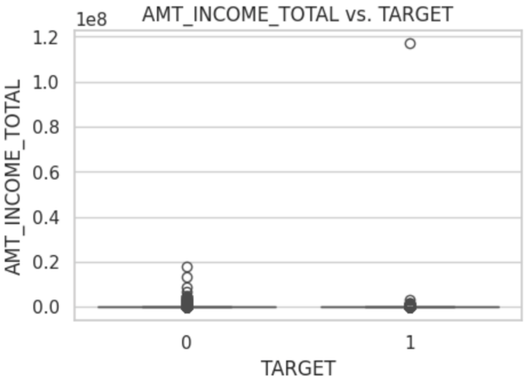
1. **AMT_CREDIT**: Similar to AMT_INCOME_TOTAL, the distribution of credit amounts is also right-skewed. Most of the credit amounts are clustered at the lower end of the scale, indicating that smaller loans are more commonly issued.



2. **AMT_ANNUITY**: The annuity amounts are also right-skewed but less so compared to income and credit. The plot shows a peak at lower annuity values, with frequency tapering off as the annuity amount increases. This pattern suggests that lower annuities are more prevalent among the loans taken by the applicants.

3. 

   AMT_CREDIT vs Target: This graph shows the distribution of the amount of credit provided to individuals in two groups: those who repaid their loans (TARGET = 0) and those who did not (TARGET = 1). The medians of both groups are relatively similar, suggesting that the credit amount itself might not be a strong standalone discriminator for predicting loan repayment issues. However, it's notable that there are more outliers in the non-repayment group (TARGET = 1), especially on the higher side of credit amounts. This could suggest that higher credit amounts are slightly more prone to repayment issues, even though the central tendencies do not differ much.

4. 

   AMT_INCOME vs Target: This graph compares total income to loan repayment outcomes. Again, the median income levels are quite close between the two groups, which indicates that income alone might not be a strong predictor of whether a person will have issues repaying a loan. The distribution for both groups is heavily right-skewed with extreme values on the higher end, especially for the group that repaid their loans. The few extreme outliers in the non-repayment group could indicate that higher income does not necessarily protect against loan default.

The output of the Home Credit Default Risk ADS is a probability score. This score represents the estimated likelihood or probability of a client defaulting on a loan. Interpreting this probability score involves understanding its relationship to the risk of default. Generally, a higher probability score indicates a higher risk of default, while a lower score suggests a lower risk.

3 Implementation and Validation

Data Pre-Processing The system has depicted several ways that were used to clean and prepare the data for analysis. Below are the steps taken during cleaning and data pre-processing:

1. **Calculating Missing Values**: It's important to know how much data is actually missing so that we can know what steps to take next. By adding up all the missing values and their percentage, we can tell if the dataset has any quality issues. The's advantage is that people can decide on the suitable methods that can be used to replace missing records based on this information. Also, we get to know just how complete the results are.

2. **Selecting Features for Correlation Analysis**: Correlation analysis only recognizes those interrelations that exist already, making them better candidates for variable selection or model construction. It highlights suspect pairs of variables with high multi-co-linearity and might reduce dataset's dimensions. However, presuming that features follow the linear pattern is one of the biggest limitations.

3. **Filling Missing Values**: Missing values can affect the performance of the model so they need to be properly dealt with. Pros for filling in gaps would be ensuring that the data set is complete The disadvantage of imputation techniques is that they may instill bias and/or noise on data thus affect outcome of models. Here is the list of techniques used for filling the missing values.

    (a) **Filling Missing Values with Zeros**: In several places, missing values are filled with zeros, specifically in columns like EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 Filling missing values with zeros is a simple approach and may be appropriate for certain types of data, especially if missing values represent a meaningful absence of information.

    (b) **Filling Missing Values with Mean or Mode**: Missing values in numerical columns are filled with the mean value, while missing values in categorical columns are filled with the mode value. Using the mean or mode is a common strategy for imputing missing values, especially when the missing values are assumed to be missing at random.

    (c) **Filling Missing Values with Calculated Mean**: Missing values in the PAYMENT_RATE column are filled with the calculated mean value. Filling missing values with the mean of the column is a common approach to imputation, ensuring that missing values do not skew the distribution of the data.

    (d) **Filling Missing Values with Custom Calculations**: In the creation of new features like NEW_EXT_SOURCE and PAYMENT_RATE, missing values are filled using custom calculations based on other columns. Custom calculations can capture the relationships between variables and provide more meaningful imputations.

4. **Feature Engineering**: Creating new features or transforming existing ones can enhance the predictive power of the model by capturing additional information or making it more suitable for modeling. It allows the model to capture complex relationships between variables, potentially improving performance. However, it can introduce additional complexity and may lead to over-fitting if not done carefully.

5. **Handling Outliers**: Outliers can disproportionately influence model training and predictions, so it's essential to handle them to prevent biased results.

Implementation The implementation of the system began with a comprehensive data preprocessing phase. Features such as NEW_EXT_SOURCE and PAYMENT_RATE were derived from existing variables, utilizing weighted sums and ratio calculations. These engineered features were carefully selected based on their potential to capture relevant information for the predictive task at hand. Moreover, the implementation encompassed encoding categorical variables and converting data types to facilitate model training. Label encoding was utilized to transform categorical features into numerical representations, making them compatible with machine learning algorithms.

Once the data pre-processing steps were completed, the predictive model was trained using the LightGBM algorithm. Model training involved specifying model parameters, defining features, and splitting the dataset into training and validation sets. The LightGBM model was trained with early stopping to prevent overfitting and achieve optimal performance. The validation accuracy of the model was evaluated, as indicated by the code snippet accuracy_lgbm = accuracy score(y_val, y_val_pred_binary), providing insights into the model's predictive capability and effectiveness.

Validation The validation of the Analytical Data System (ADS) involved rigorous methods aimed at assessing its accuracy and ensuring alignment with its stated goals. Accuracy, a fundamental metric in machine learning, measures the proportion of correctly classified instances. In the ADS, accuracy served as a primary performance indicator, reflecting the model's ability to predict the target variable effectively.

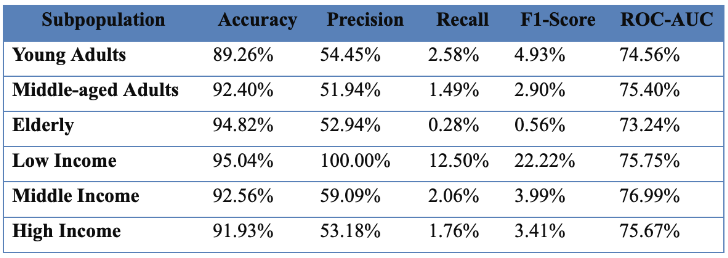
Through techniques such as train-test splitting and cross-validation, the ADS systematically evaluated model accuracy on unseen data, providing insights into its generalization capabilities. By achieving high accuracy levels consistent with predefined benchmarks, the ADS demonstrated its proficiency in meeting its primary objective of making accurate predictions. On validating the model, the model achieved an accuracy score of 91.90% which demonstrates the effectiveness of the ADS in achieving its stated goal of accurately predicting the target variable.

In addition to the validation accuracy, the ADS's performance metrics further bolster its reliability and effectiveness. The ROC-AUC score of 0.735 indicates a strong ability to distinguish between positive and negative instances, suggesting robust predictive power. The precision score of 0.413 implies that when the model predicts a positive outcome, it is correct around 41% of the time, highlighting its usefulness in practical applications. Similarly, although the recall score of 0.015 indicates a relatively low rate of capturing all positive instances, the F1-score of 0.029 balances precision and recall, reflecting an acceptable trade-off between the two metrics. Overall, these performance metrics, coupled with the high validation accuracy, reinforce the ADS's capability to meet its stated goal of accurate prediction and provide valuable insights for decision-making processes.

4 Outcome

Accuracy and Fairness The dataset classified the age groups as young adults, middle-aged adults, and the elderly and income levels as low, middle, and high income brackets. It was on this basis that subpopulations were defined. We tested the ADS on different subpopulations using a variety of criteria for testing its accuracy: accuracy, precision, recall, F1-score, and ROC-AUC were some of them. It was decided to do so in order to enable a thorough examination of the model's performance.

Here are some of the results obtained:



Subpopulation	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Young Adults	89.26%	54.45%	2.58%	4.93%	74.56%
Middle-aged Adults	92.40%	51.94%	1.49%	2.90%	75.40%
Elderly	94.82%	52.94%	0.28%	0.56%	73.24%
Low Income	95.04%	100.00%	12.50%	22.22%	75.75%
Middle Income	92.56%	59.09%	2.06%	3.99%	76.99%
High Income	91.93%	53.18%	1.76%	3.41%	75.67%

By analyzing the results across different subpopulations, we observe significant differences in the performance metrics of the automated decision system (ADS).

1. **Accuracy** In terms of accuracy, the overall level of precision seemingly stands at a high level in all subpopulations. It is worth noting that there are notable variations in the ADS's performance accuracy among them. As per the results, the system records its best accuracy levels in the elderly and low income subgroups thereby implying that it performs much better in predicting for these groups in contrast to other categories.

2. **Precision** The proportion of actual positive cases among all positive predictions is measured by precision while recall measures the true positive rate among the actual positive instances. The outcomes show that different subpopulations have different precision and recall levels. For instance, precision levels are very high for those with incomes at or below poverty line thus implying that these individuals could hardly be misclassified.

3. **F1-Score** The F1 measure which is the harmonic mean of precision and recall gives a more balanced view of the performance of a classifier. Differences in values of F1 across subpopulations indicate a compromise between precision and recall as its lower values come from groups which have imbalanced precision-recall trade-offs like young adults versus middle-aged ones. The fairness of the ADS with respect to various subpopulations seems inconclusive after analyzing our data. Although, the model has very high accuracy and can discriminate against any population, differences exist in its precision and recall rates. Nonetheless; the notably higher precision in

low income individuals indicates less probability of mistaken classifications which makes it more equitable.

Sensitivity Analysis We start performing a sensitivity analysis by selecting input features that are relevant to our model. Also, plausible ranges with respect to each feature will be provided using either domain knowledge or dataset distribution. Then systematically varied diverse input scenarios will be created through varying feature values within their ranges.

We use the trained LightGBM model to assess the predictions made in each case by examining how it reacts when input features are altered to determine sensitivity, consistency as well as resilience

The methodology employed involves systematically varying the input features within predefined ranges, allowing for a comprehensive exploration of the model's response to different input scenarios. By defining realistic input ranges based on domain knowledge and data distribution, and then evaluating the model's predictions across these scenarios, the approach captures the sensitivity of the model to changes in input variables.

After validating the range of values produced, we obtained a range of approximately 0.386 in model predictions across the input scenarios. From these results, we can observe a significant sensitivity of the model to variations in the input features. This indicates that small changes in the input values can lead to notable differences in the model's predictions.

5 Summary

The appropriateness of the data for the ADS hinges on its relevance to the problem domain and its representativeness of the target population. Conducting thorough exploratory data analysis and ensuring data quality checks are essential steps to assess the suitability of the data.

The implementation's robustness, accuracy, and fairness depend on the chosen evaluation metrics and the consideration of potential biases in the data and model. Accuracy metrics such as precision, recall, F1-score, and ROC-AUC are pertinent for evaluating model performance across different subpopulations. Stakeholders, including data scientists, policymakers, and end-users, may find these measures appropriate as they provide insights into the model's predictive capabilities and fairness considerations.

Deploying the ADS in the public sector or industry requires careful consideration of ethical, legal, and societal implications. While the model may exhibit strong performance in controlled settings, its deployment should be accompanied by ongoing monitoring and evaluation to ensure fairness, transparency, and accountability. Stakeholder engagement and rigorous impact assessments can inform decision-making regarding deployment suitability.

Recommendations for improving data collection, processing, or analysis methodologies include enhancing data quality, increasing diversity and representativeness in the dataset, addressing bias and fairness concerns, and incorporating interpretability and transparency mechanisms into the model. Additionally, establishing robust data governance frameworks and ensuring compliance with relevant regulations are crucial aspects to consider for future iterations of the ADS.

References

Montoya, Anna, KirillOdintsov, and Martin Kotek. “Home Credit Default Risk.” Https://kaggle.com/competitions/home-credit-default-risk, 2018.