BSFI CASE STUDY

Author Shreyas Patil

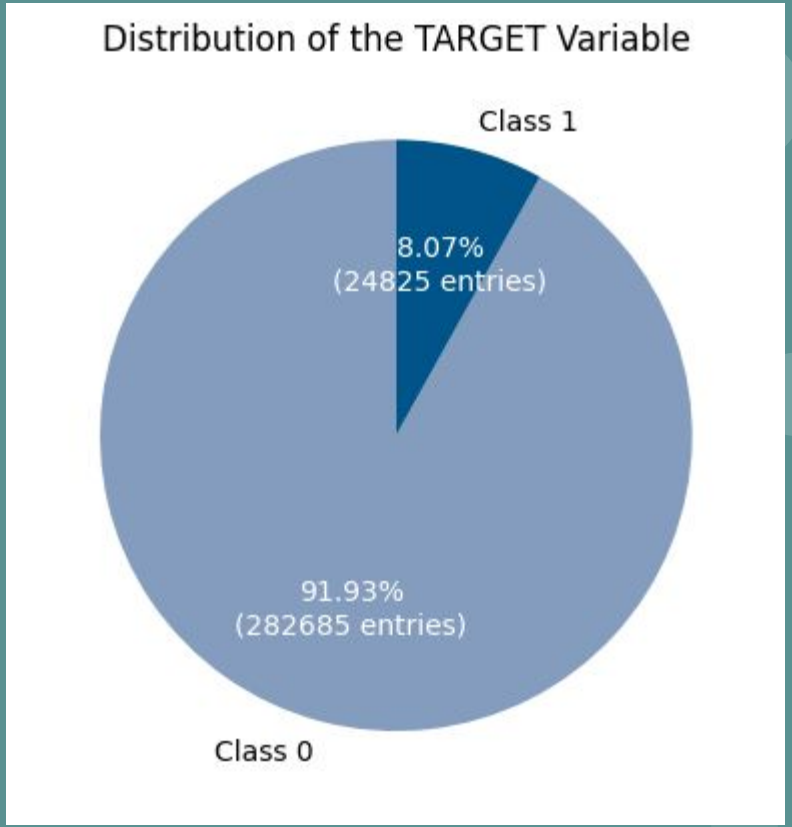
Problem Statement

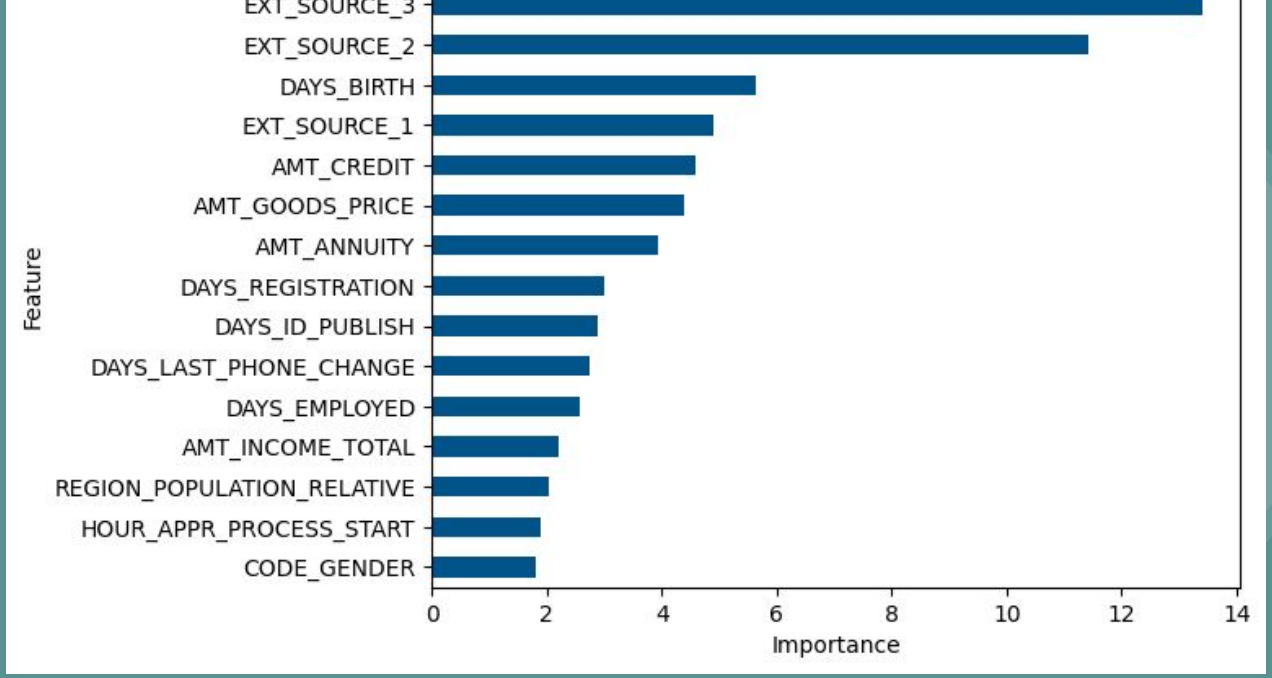
- Home Credit in deciding which loan applications should be disbursed, and which should be rejected, based on the applicant’s past behaviour and application information.

Tech Stack

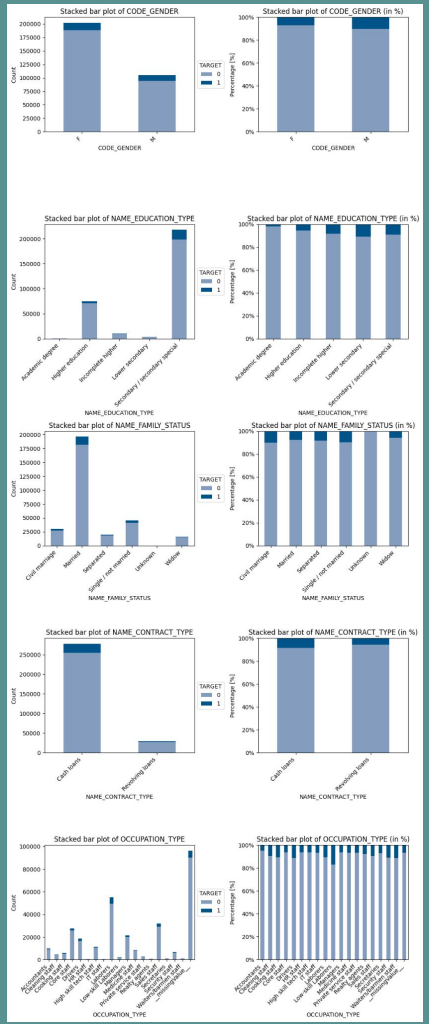
- Python
- Excel
- SQL

- The pie chart illustrates a clear imbalance in the distribution of the dependent variable TARGET: The critical class 1 appears significantly less frequently than class 0.

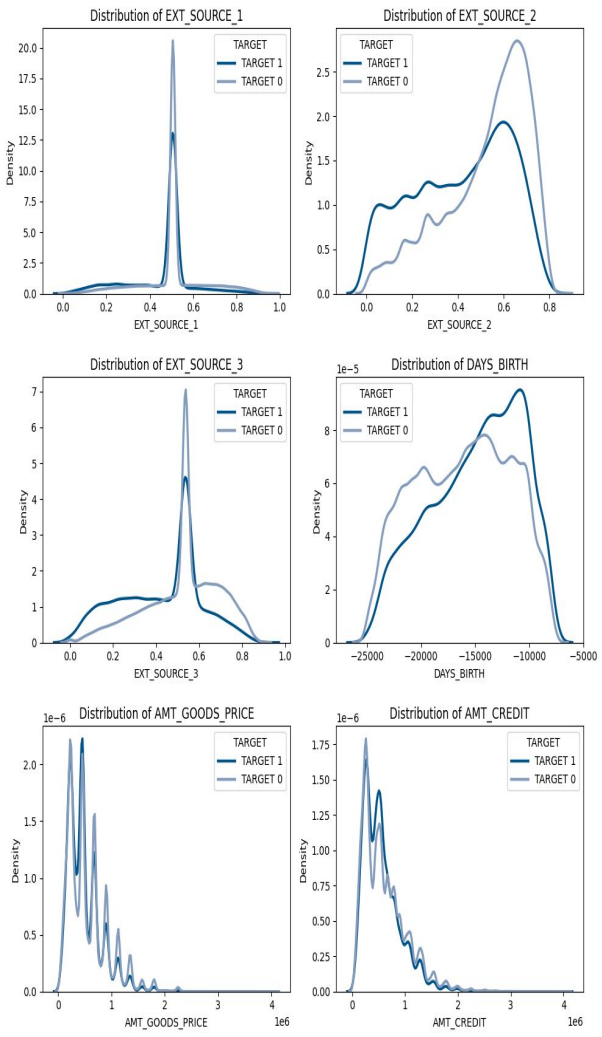




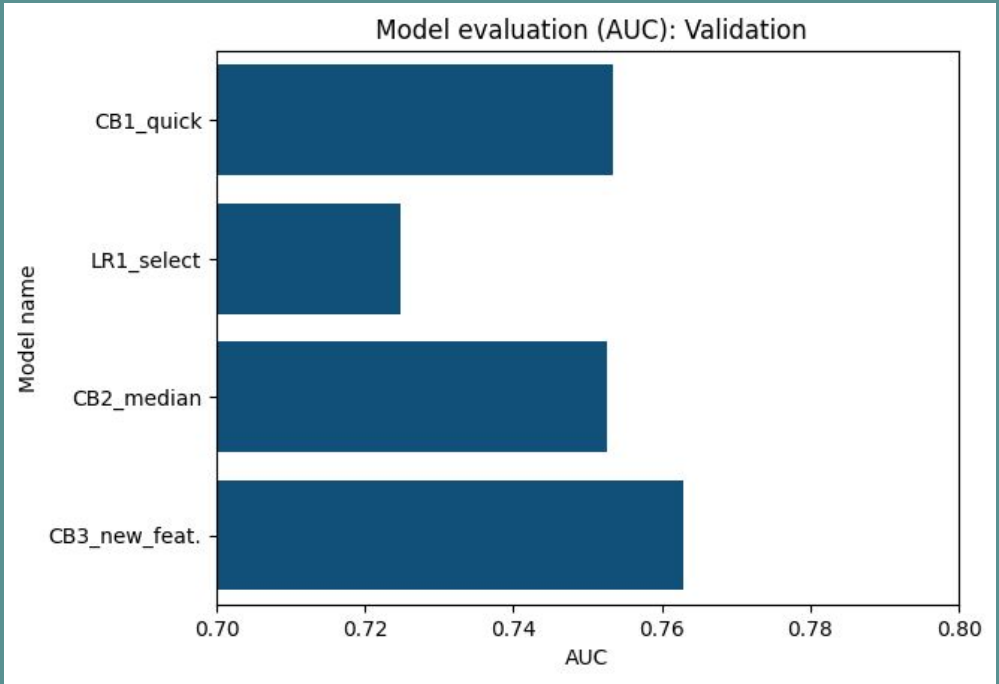
1. Influential Predictors:
   - External information sources (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) and key financial variables are identified as primary influencers.
   - Borrower age emerges as a significant factor shaping the predictive strength of the model.
2. Model Performance:
   - Our CatBoost model is evaluated against the validation set using the AUC metric.
   - The model's ability to differentiate between approved and rejected applicants is effectively measured, providing insights into its overall performance.



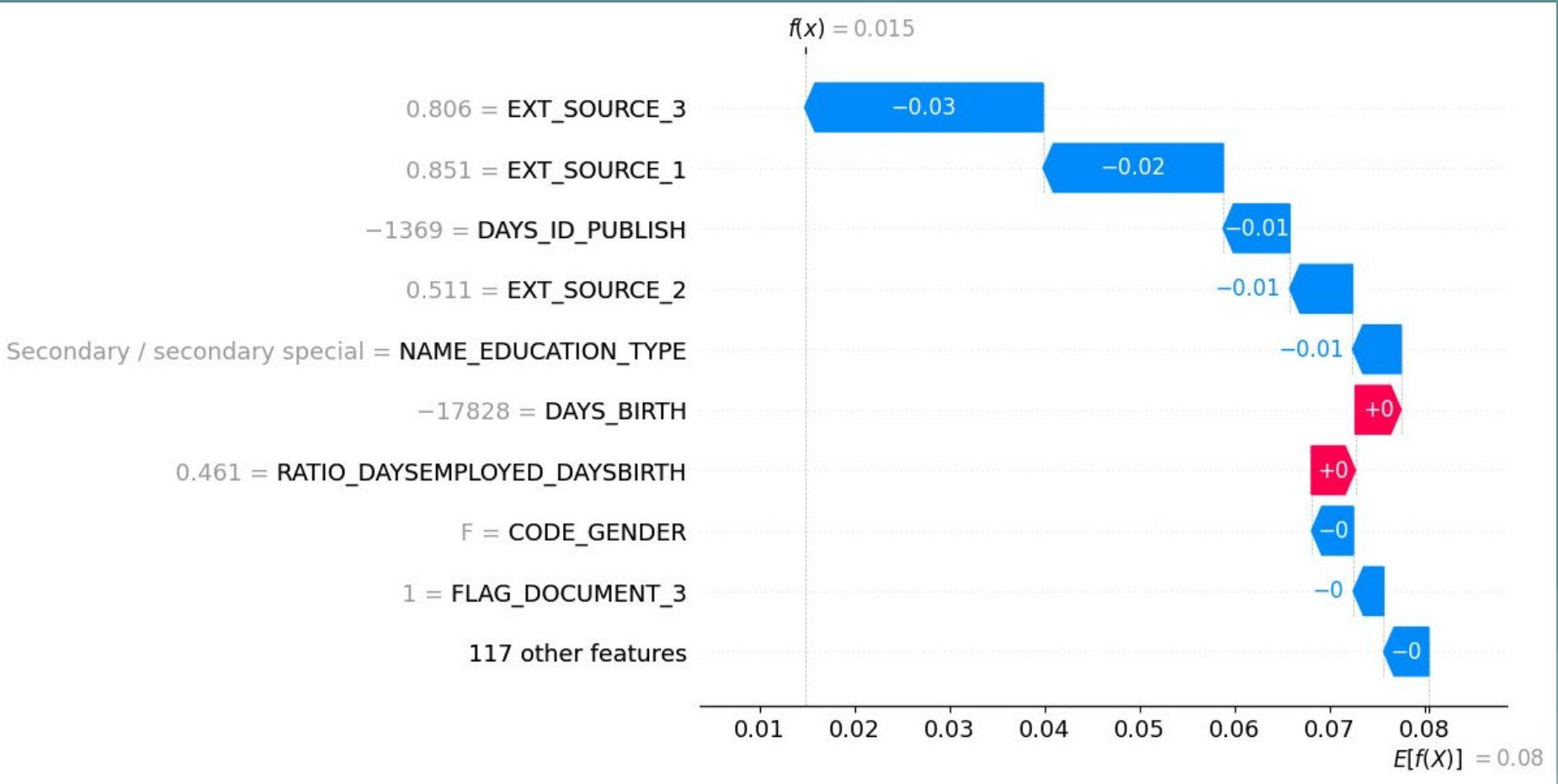
- Higher education (NAME_EDUCATION_TYPE) and advanced occupational (OCCUPATION_TYPE) correlate with lower default rates, whereas lesser education and basic occupations are linked to higher defaults.
- Male applicants (CODE_GENDER) show a higher tendency to default than females.
- Loan type matters; cash loans exhibit a higher default risk than revolving loans (NAME_CONTRACT_TYPE).
- Family status (NAME_FAMILY_STATUS) shows no significant variation in default rates.
- Rare categories like XNA in CODE_GENDER and UNKNOWN in NAME_FAMILY_STATUS may be omitted from models to streamline the feature set.

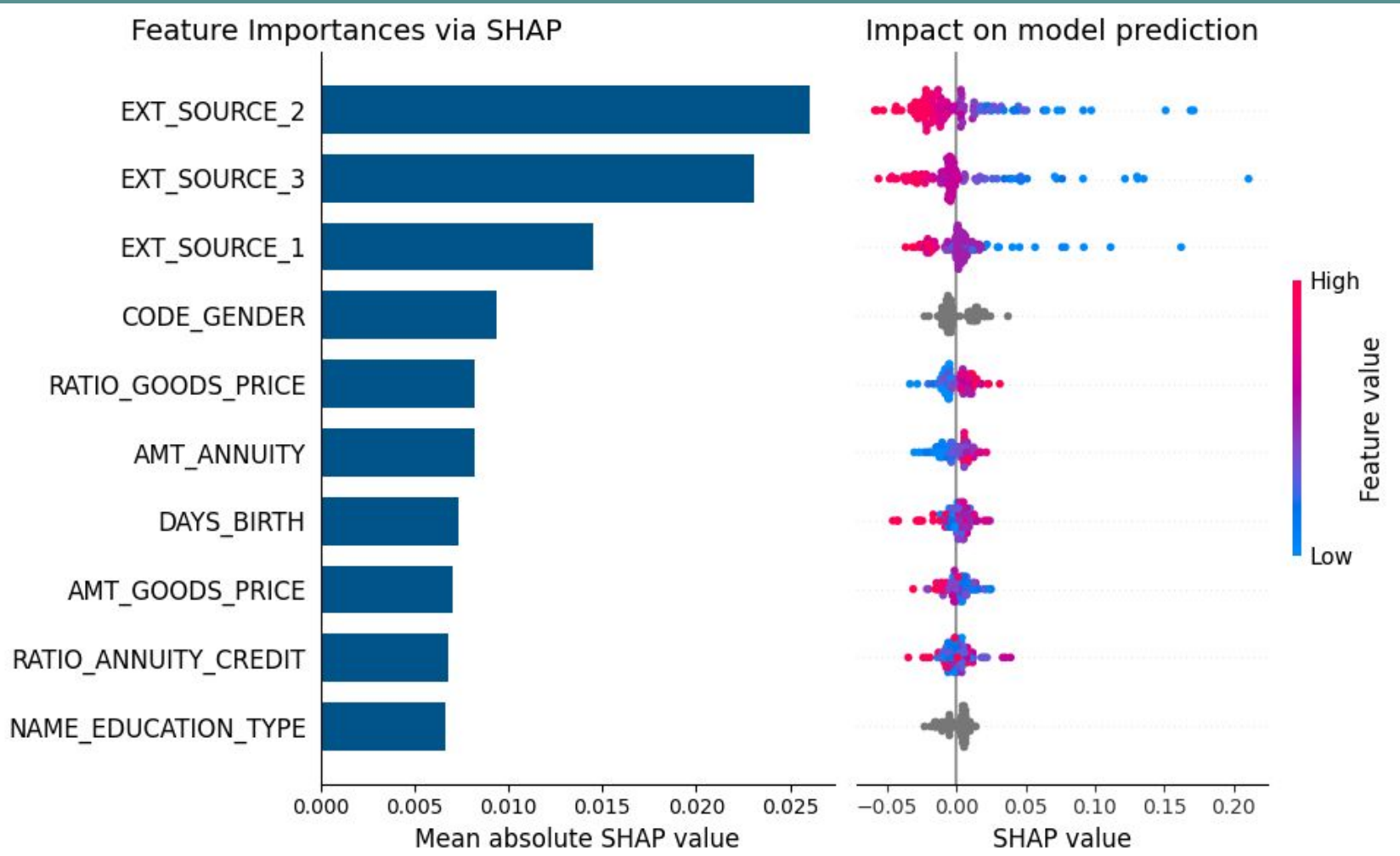


- EXT_SOURCE features, which likely represent external credit scores, show a noticeable negative correlation with default risk. Higher values of EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 are associated with lower probabilities of defaulting.
- DAYS_BIRTH, representing the age of the applicant in days, exhibits that younger applicants are more prone to defaulting. There is a trend suggesting that the risk of default decreases with age.
- AMT_GOODS_PRICE, indicating the value of the goods for which the loan is taken, suggests that lower-priced goods correlate with a higher likelihood of default.
- Similarly, AMT_CREDIT, the total credit amount borrowed, is observed to have a relationship with default rates, with smaller loans having a tendency towards higher default frequencies.

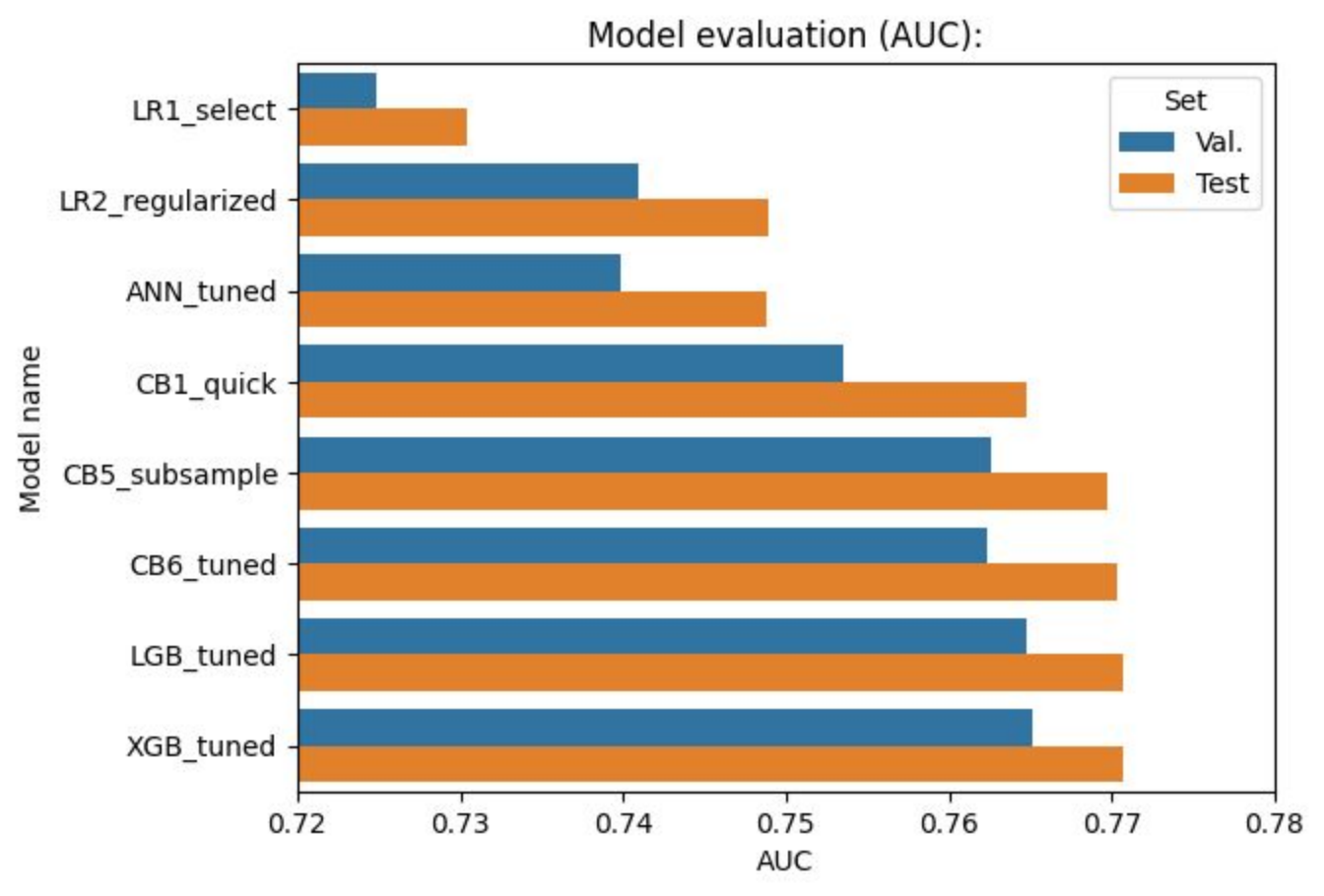


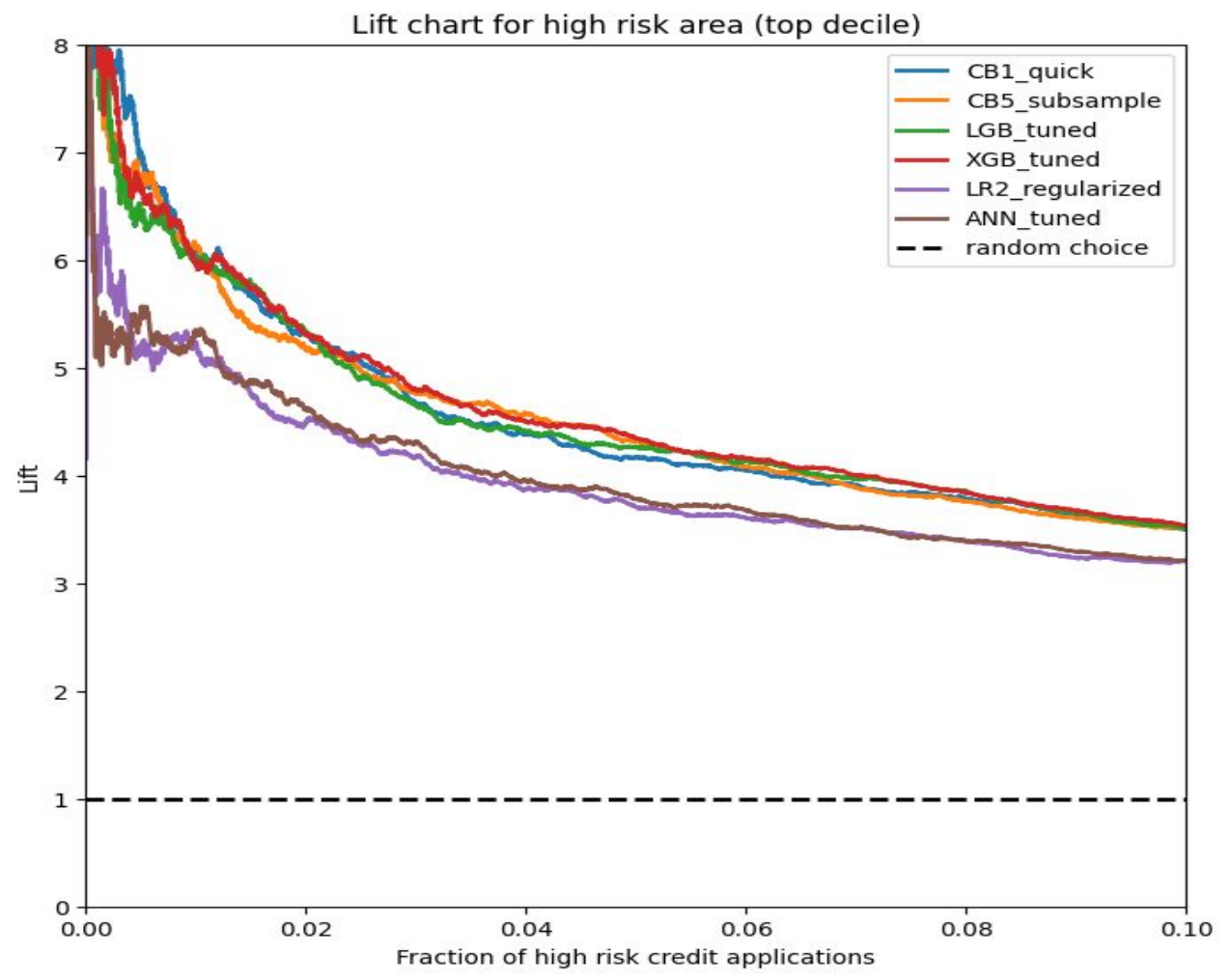
- The new features have noticeably improved the performance of our latest CatBoost model. Feature engineering can be the key to improving model performance. In a sense, our most important variables ext_source_* are also the result of (external) feature engineering.


f(x) = 0.015
0.806 = EXT_SOURCE_3
0.851 = EXT_SOURCE_1
−1369 = DAYS_ID_PUBLISH
0.511 = EXT_SOURCE_2
Secondary / secondary special = NAME_EDUCATION_TYPE
−17828 = DAYS_BIRTH
0.461 = RATIO_DAYSEMPLOYED_DAYSBIRTH
F = CODE_GENDER
1 = FLAG_DOCUMENT_3
117 other features
−0.03
−0.02
−0.01
−0.01
−0.01
+0
+0
−0
−0
−0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
E[f(X)] = 0.08


Feature Importances via SHAP
EXT_SOURCE_2
EXT_SOURCE_3
EXT_SOURCE_1
CODE_GENDER
RATIO_GOODS_PRICE
AMT_ANNUITY
DAYS_BIRTH
AMT_GOODS_PRICE
RATIO_ANNUITY_CREDIT
NAME_EDUCATION_TYPE
0.000
0.005
0.010
0.015
0.020
0.025
Mean absolute SHAP value
Impact on model prediction
−0.05
0.00
0.05
0.10
0.15
0.20
SHAP value
High
Feature value
Low

- As previously observed in the CatBoost internal feature importance plot in Section 1, SHAP also identifies key features such as EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_BIRTH as crucial determinants for the predictions of our CatBoost model. This underscores their significance in the classification task we are examining.
- SHAP Feature Importance Plot:
  - Provides a global overview of feature impact on model predictions.
  - Ranks features by mean absolute SHAP scores, indicating their general influence across all predictions.
- SHAP Summary Plot:
  - Offers a detailed perspective on feature impact distribution across all dataset instances.
  - Unmasks relationships between feature values and their impact on model predictions, emphasizing variations from one instance to another.


Model evaluation (AUC):
Set
Val.
Test
LR1_select
LR2_regularized
ANN_tuned
CB1_quick
CB5_subsample
CB6_tuned
LGB_tuned
XGB_tuned
Model name
0.72
0.73
0.74
0.75
0.76
0.77
0.78
AUC


Lift chart for high risk area (top decile)
CB1_quick
CB5_subsample
LGB_tuned
XGB_tuned
LR2_regularized
ANN_tuned
random choice
Lift
Fraction of high risk credit applications
0
1
2
3
4
5
6
7
8
0.00
0.02
0.04
0.06
0.08
0.10

Test Data Performance:

- Surprisingly, all models exhibit improved performance on the test data, potentially attributed to differences in samples and TARGET distribution, as discussed in Subsection 1.3.

Model Evaluation and Comparison:

- XGB_tuned demonstrates the best prediction quality, with faster hyperparameter tuning compared to LGB_tuned.
- Top-performing models, including CB5_subsample (untuned), LGB_tuned, and XGB_tuned, benefit from effective feature engineering, outperforming the base model CB1_quick.

Cumulative Lift Chart Analysis:

- The cumulative lift chart illustrates the model's ability to enhance the "hit rate" and influence credit defaults.
- Notably, focusing on the 6% of applications with the highest default probability, with a lift score of 4.2, could impact 25% of all credit defaults.

Model Comparison for Top Decile:

- No clear winner emerges among gradient tree boosting models in the top decile, indicating similar lift scores.
- The baseline model CB1_quick is considered for defining the top risk area, and the focus shifts to creating a list of the riskiest loan applications, comparing predicted and true values using test data.