PRIMARY Connection score 24’

**Date of creation:** 7/25/2024

**Scoring analyst:** Zhalgas Zhienbekov

# Introduction

A glance at the model concerns the connection rate probability for primary phone numbers. It was aimly collected only primary phone numbers from call list for the last 7 month including the historical and future date as out of time. Due to population size differences within the considering period, the overall GINI varies for almost 5%. There are 8 winning predictors that passed all feature selection filters.

The name of File : model\_connect\_score\_24\_1\_final.pkl

DWH\_Source: OWNER\_EXT.T\_ZZ\_CONNECTION\_SCORE\_P\_ACTUAL (Actual Data)

Hadoop Source: bdp\_score\_stg. monthly\_connection\_score24\_p\_hdfs

## Document version control

|  |  |  |  |
| --- | --- | --- | --- |
| Document version | Date of issue | Author | Modification details |
| 1 | 7/25/2024 | Zhalgas Zhienbekov | Modified filling |

## Basic information

|  |  |
| --- | --- |
| Country | Kazakhstan |
| Area | CRM |
| Type | Connection |
| Process | Call connection |

## Development reason

Request of CRM to create connection model.

## Initiation document

Initiation document as an attachment.

Attach initiation document as an object (icon).

# Data sample

## Basic sample information

|  |  |
| --- | --- |
| Product type | None product |
| Client type | All clients |
| Other specifications | Model is built using on phone numbers who has contacted with bank (data from First Credit Bureau,..). |

## Data sample SQL script

SQL script generating data for the sample as an attachment.

CRM Datamart

## Target

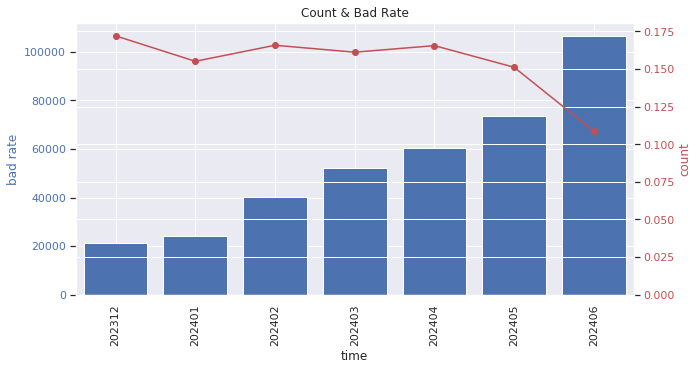
### Definition

Phone numbers that interacted with bank within the 6 month and aggregated predictors for the last 6 month before call attempt

### Target Analysis

Description of construction of sample. The sample size is 378K rows and was randomly selected calls with Call\_result = 1 and 0 in relationship 1:1 and dtime\_call\_start stratified

Below graph show true bad rate



## Development sample

### Time period

Time period of the development sample

January 1st 2023 – May 31st 2024

### Sampling description

How the development sample was divided into Training, Validation and Test samples

|  |  |  |
| --- | --- | --- |
| Sample name | Sample share | Sample description |
| Training | 60% | Randomly chosen majority of phone numbers, contacted with bank. |
| Validation | 20% | Randomly chosen part of phone numbers, used for performance and stability control, to avoid overfitting. |
| Test | 20% | For further performance evaluation of the chosen model. |

### Sample statistics

Table showing number of contacts, take up and take up rate for each month of the development sample.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CALL\_RESULT** | | | | | **ROWS** | | | | | **CALL\_RESULT\_RATE** | | | | |
| **data\_ type** | **hoot** | **oot** | **test** | **train** | **valid** | **hoot** | **oot** | **test** | **train** | **valid** | **hoot** | **oot** | **test** | **train** | **valid** |
| **MONTH** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 202312 | **3661** |  |  |  |  | 21275 |  |  |  |  | 0.172 |  |  |  |  |
| 202401 |  |  | 748 | 2243 | 748 |  |  | 4821 | 14460 | 4820 |  |  | 0.155 | 0.155 | 0.155 |
| 202402 |  |  | 1331 | 3993 | 1332 |  |  | 8029 | 24086 | 8029 |  |  | 0.166 | 0.166 | 0.166 |
| 202403 |  |  | 1682 | 5048 | 1683 |  |  | 10438 | 31316 | 10439 |  |  | 0.161 | 0.161 | 0.161 |
| 202404 |  |  | 2006 | 6020 | 2007 |  |  | 12120 | 36361 | 12121 |  |  | 0.166 | 0.166 | 0.166 |
| 202405 |  |  | 2220 | 6660 | 2220 |  |  | 14685 | 44054 | 14684 |  |  | 0.151 | 0.151 | 0.151 |
| 202406 |  | 11588 |  |  |  |  | 106633 |  |  |  |  | 0.109 |  |  |  |

## Out-of-time sample

Future out of time is the data of call gathered for the period June 2024

Historical out of time is the data of call gathered for the period June 2024

### Time period

Time period of the out-of-time sample

December 2023 (HOOT)

June 2024 (OOT)

### Sample statistics

Table showing number of calls, conversion rate for each month of the out-of-time sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type of out-of-time | Period | Target amount | Count all sample | Conversion Rate |
| HOOT | 2023-12-01 - 2023-12-30 | 3661 | 21275 | 0.172 |
| OOT | 2024-06-01 - 2024-06-30 | 11588 | 106633 | 0.108 |

## Oversampling

Not used

# Methods

## Method used for variables grouping

### WOE transformation

Whether WOE transformation was used.

Categorical and numerical variables are used as they are in LGBM. WOE trasnfromation is applied for Predictor Power analysis in calculating IV

## Feature Selection Methods

The feature selection process for the model was conducted using the following methods:

1. **Variance Thresholding**:
   * Features with zero variance were removed. This was accomplished by excluding all features that did not exhibit any variation across the dataset.
2. **Missing Values Threshold**:
   * Features with a proportion of missing values exceeding 60% were dropped from consideration. This ensured that only features with sufficient completeness were included in the model.
3. **Predictor Power Test**:
   * Features were selected based on their Information Value (IV) metrics, with the following criteria applied:
     + IV Train values were required to be within the range (0.049, 0.8)
     + IV Test values were required to be within the range (0.05, 0.8)
     + IV Validate values were required to be within the range (0.049, 0.8)
     + IV OOT values were required to be within the range (0.05, 0.8)
   * Only features meeting these criteria were retained for further analysis.
4. **Population Stability Index (PSI) Check**:
   * The PSI was evaluated for the training, test, and validation samples. Features were selected based on the following PSI criterion:
     + Only features with a monthly PSI average less than 0.2 and greater than 0 were considered. This criterion is used to ensure the stability of features across different datasets.
5. **Autocorrelation Test**:
   * To avoid multicollinearity, features were selected such that no pair of selected features had a correlation coefficient greater than 0.7. The variable\_clustering method from the scoring module was utilized for this purpose.
6. **Feature Importance Evaluation**:
   * The final set of selected features was evaluated for their importance using the Gain parameter. This step ensured that the retained features contributed significantly to the model's predictive performance.

These methods collectively ensured a robust and effective feature selection process, optimizing the model's performance and interpretability.

## Method used for variable selection

### Method type

E.g. stepwise selection model.

LGBM feature selection model (gain and weight importance based). At first, we did PSI computation and then variable clustering based on WOE with highest correlation 0.7.

### Input parameters

Input hyperparameters of LightGBM model

**Hyperparamter space:**

{'colsample\_bytree': 0.7500000000000002, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.06, 'max\_depth': 4, 'metric': 'auc', 'min\_child\_weight': 60, 'min\_data\_in\_leaf': 450, 'num\_boost\_round': 100000, 'num\_leaves': 52, 'objective': 'binary', 'reg\_alpha': 0.42705975987053457, 'reg\_lambda': 0.3287598336004255, 'seed': 1234, 'subsample': 0.5, 'verbose': 1}

Actual gini:

0.6805149929344425

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

{'colsample\_bytree': 0.6500000000000001, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.08, 'max\_depth': 4, 'metric': 'auc', 'min\_child\_weight': 70, 'min\_data\_in\_leaf': 50, 'num\_boost\_round': 100000, 'num\_leaves': 60, 'objective': 'binary', 'reg\_alpha': 0.34195383454203165, 'reg\_lambda': 0.2541497287923511, 'seed': 1234, 'subsample': 0.7000000000000002, 'verbose': 1}

Actual gini:

0.6802091797396885

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

{'colsample\_bytree': 0.8500000000000003, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.08, 'max\_depth': 4, 'metric': 'auc', 'min\_child\_weight': 60, 'min\_data\_in\_leaf': 350, 'num\_boost\_round': 100000, 'num\_leaves': 6, 'objective': 'binary', 'reg\_alpha': 0.14380645510559387, 'reg\_lambda': 0.28696058054987783, 'seed': 1234, 'subsample': 0.7500000000000002, 'verbose': 1}

Actual gini:

0.6802313619693121

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

{'colsample\_bytree': 0.55, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.04, 'max\_depth': 4, 'metric': 'auc', 'min\_child\_weight': 20, 'min\_data\_in\_leaf': 200, 'num\_boost\_round': 100000, 'num\_leaves': 26, 'objective': 'binary', 'reg\_alpha': 0.22824794318474584, 'reg\_lambda': 0.9643024565893034, 'seed': 1234, 'subsample': 0.6500000000000001, 'verbose': 1}

Actual gini:

0.6804475089013264

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

{'colsample\_bytree': 0.8000000000000003, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.04, 'max\_depth': 3, 'metric': 'auc', 'min\_child\_weight': 80, 'min\_data\_in\_leaf': 200, 'num\_boost\_round': 100000, 'num\_leaves': 54, 'objective': 'binary', 'reg\_alpha': 0.6174373794572313, 'reg\_lambda': 0.22767205578632388, 'seed': 1234, 'subsample': 0.8000000000000003, 'verbose': 1}

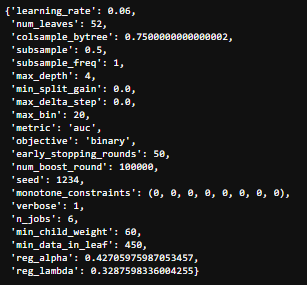
Actual gini:

0.6800832550225333

**Best combination of parameters is:**

{'colsample\_bytree': 0.7500000000000002, 'early\_stopping\_rounds': 50, 'learning\_rate': 0.06, 'max\_depth': 4, 'metric': 'auc', 'min\_child\_weight': 60, 'min\_data\_in\_leaf': 450, 'num\_boost\_round': 100000, 'num\_leaves': 52, 'objective': 'binary', 'reg\_alpha': 0.42705975987053457, 'reg\_lambda': 0.3287598336004255, 'seed': 1234, 'subsample': 0.5, 'verbose': 1}

Insert parameter setting, e.g. as a screenshot:.



## Software used

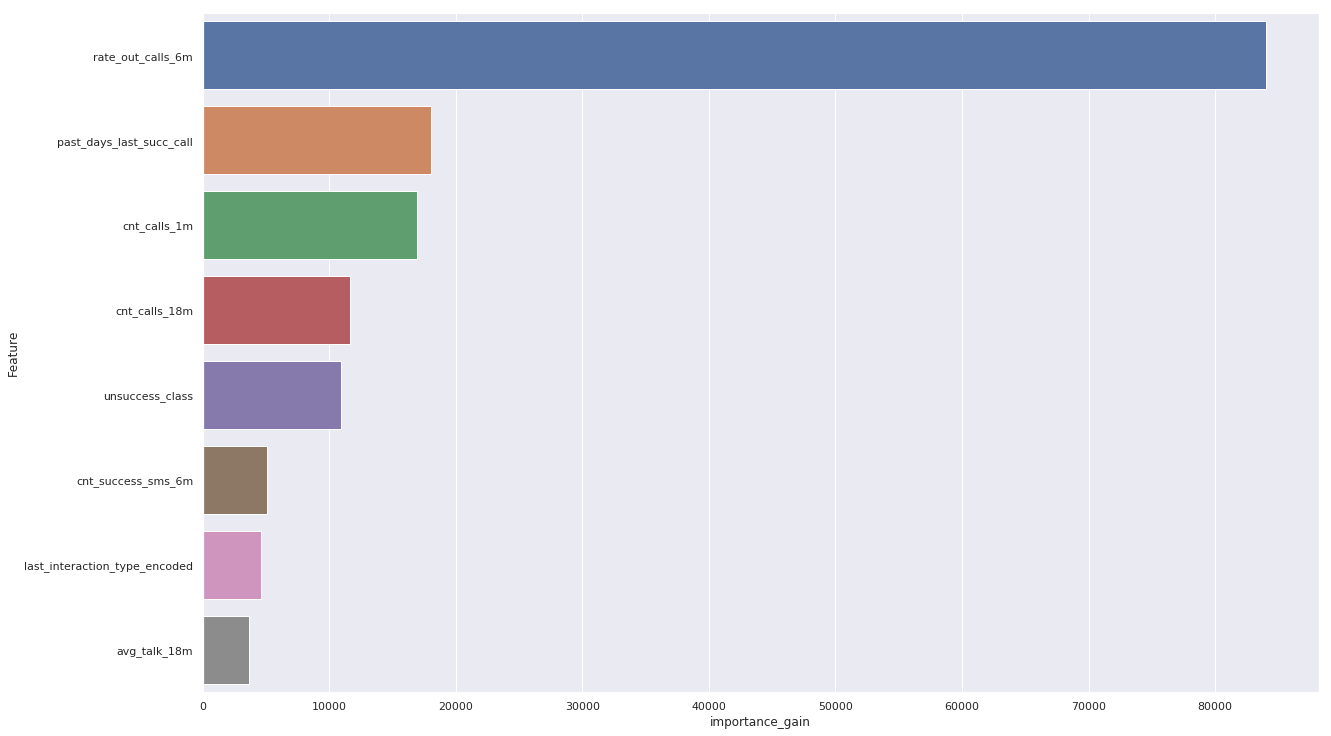
Software used for the scorecard development (e.g. SAS Enterprise miner, FICO Model Builder, R…)

Home Credit Python Scoring Workflow v0.11.0

GradientBoosting\_0\_8\_3, Python\_3.7.3

# Final model

## LGBM Feature Importance



* Estimated coefficient for each decision trees

## Predictors

Definitions of scorecard predictors + description of each predictor’s categories and their business interpretation

|  |  |
| --- | --- |
| **Predictors** | **Definition** |
| avg\_talk\_18m | The average duration of calls made over the past 18 months. This predictor reflects the average call length, providing insight into communication habits over a longer period. |
| cnt\_calls\_18m | The total number of calls made in the past 18 months. This indicates overall calling activity during this period. |
| cnt\_calls\_1m | The total number of calls made in the past month. This predictor gives a recent snapshot of calling behavior. |
| cnt\_success\_sms\_6m | The total number of successful SMS messages sent in the past 6 months. This reflects the user's text messaging activity. |
| last\_interaction\_type\_encoded | An encoded value representing the type of the last interaction (e.g., call, SMS, email). This helps to categorize the most recent interaction. |
| past\_days\_last\_succ\_call | The number of days since the last successful call. This metric indicates how recently the user made a successful call. |
| rate\_out\_calls\_6m | The ratio of outgoing calls to the total number of calls in the past 6 months. This shows the proportion of calls initiated by the user. |
| unsuccess\_class | A classification of unsuccessful interactions. This predictor helps to identify and categorize unsuccessful communication attempts also no attempts since it consist zero calls as 0 |

# Predictors characteristics

## Marginal contribution

Marginal contribution is computed on full sample. This marginal contribution was calculated by subtracting the gini of the model with the full predictors and the one with the listed predictor taken out from the model.

|  |  |
| --- | --- |
| **Predictors** | **Marginal contribution LGBM** |
| unsuccess\_class | 0.0369411 |
| cnt\_calls\_18m | 0.230748 |
| cnt\_success\_sms\_6m | 0.503698 |
| t\_interaction\_type\_encoded | 0.673323 |
| past\_days\_last\_succ\_call | 1.51589 |
| cnt\_calls\_1m | 1.69687 |
| rate\_out\_calls\_6m | 4.30638 |

# Correlations

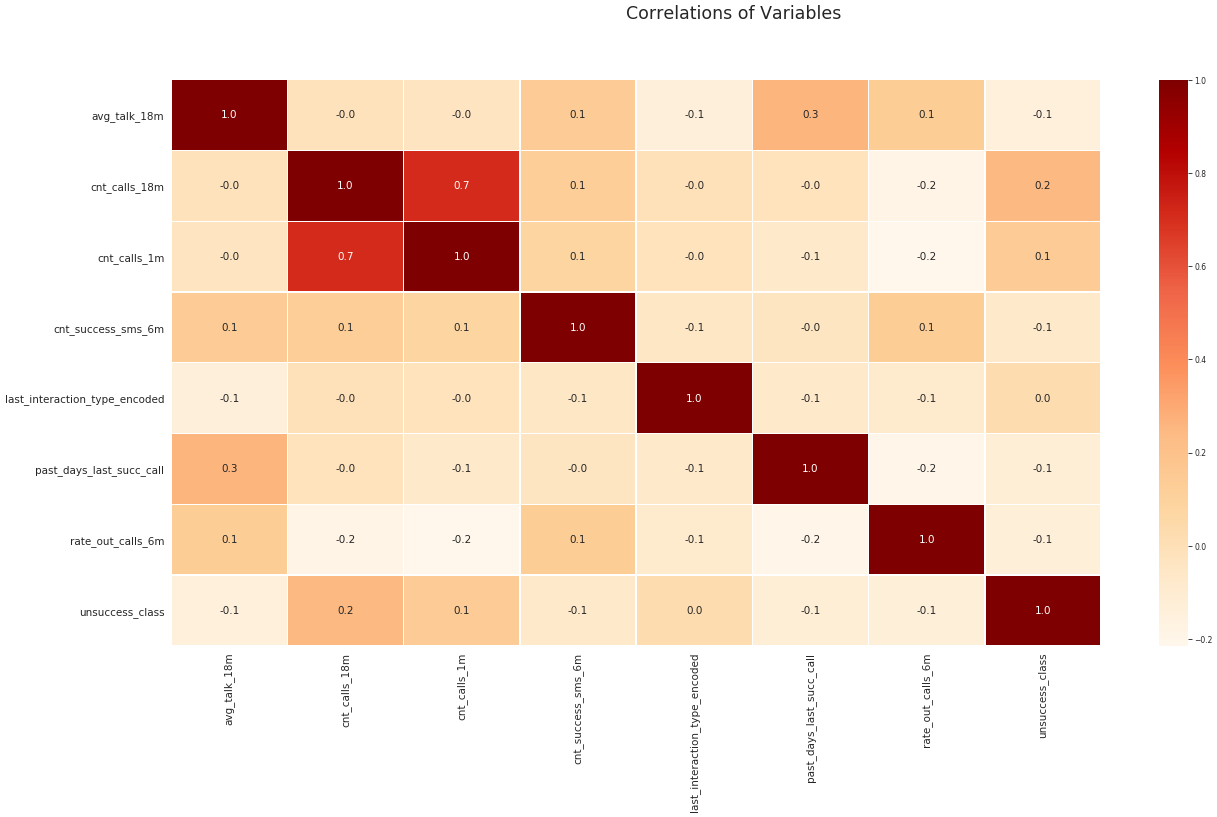
## Highest correlation

A black text on a white background

Description automatically generated

## Correlation matrix of raw data

These data are used in the model. Autocorrelation of predictors



Correlation of predictors to the target

A screenshot of a phone

Description automatically generated

# Model evaluation

To evaluate stability of the model. A model is stable if ROC and cumulative lift curves are almost the same on training and validation samples.

## Performance (samples comparison)

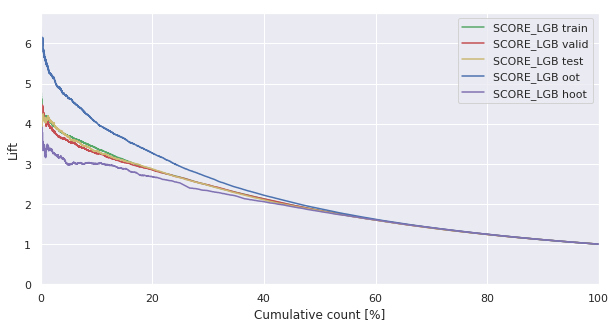
Performance comparison of training and validation samples (Gini, cumulative lift 10%, K-S statistics)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **Gini Mean** | **Gini std** | **CI 5%** | **CI 95%** |
| train | 0.689108 | 0.002309 | 0.685822 | 0.692868 |
| valid | 0.684762 | 0.003984 | 0.677644 | 0.690931 |
| test | 0.682667 | 0.003769 | 0.677133 | 0.688631 |
| oot | 0.715200 | 0.003222 | 0.710116 | 0.719840 |
| hoot | 0.660796 | 0.006409 | 0.650527 | 0.670409 |

## Cumulative Lift curves (samples comparison)

Comparison of cumulative lift curves on training and validation samples

Insert chart/picture. Example:



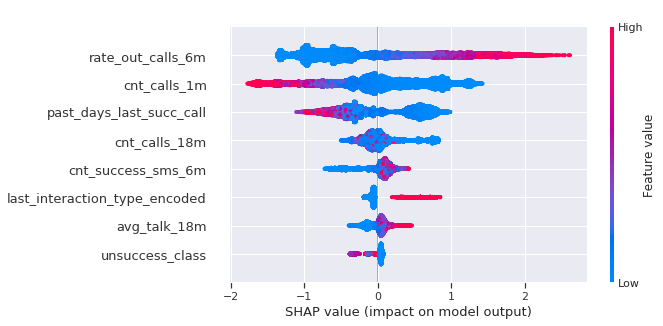
## ROC curves (samples comparison)

Comparison of ROC curves on training and validation samples

A graph of a number of numbers

Description automatically generated

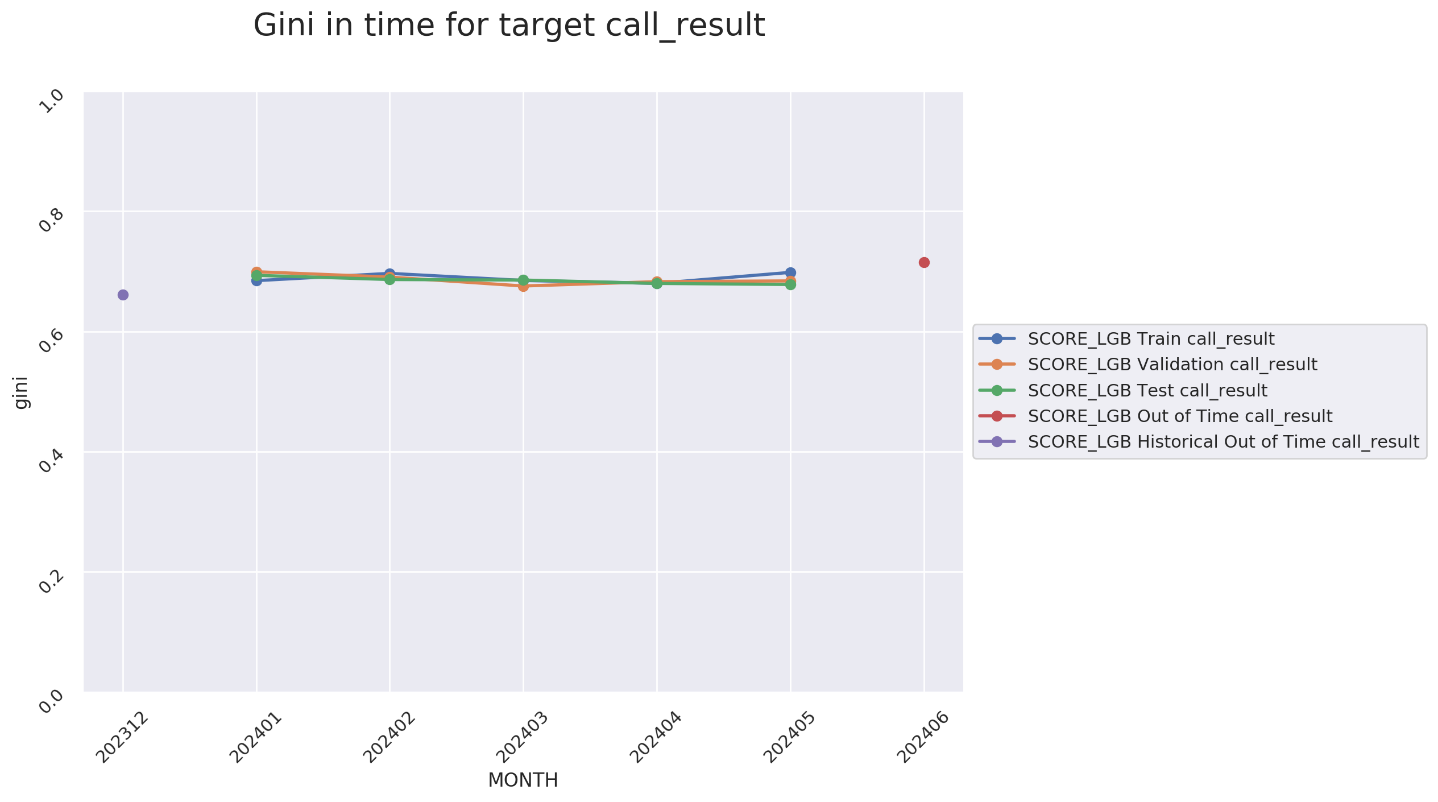
## SHAP Values



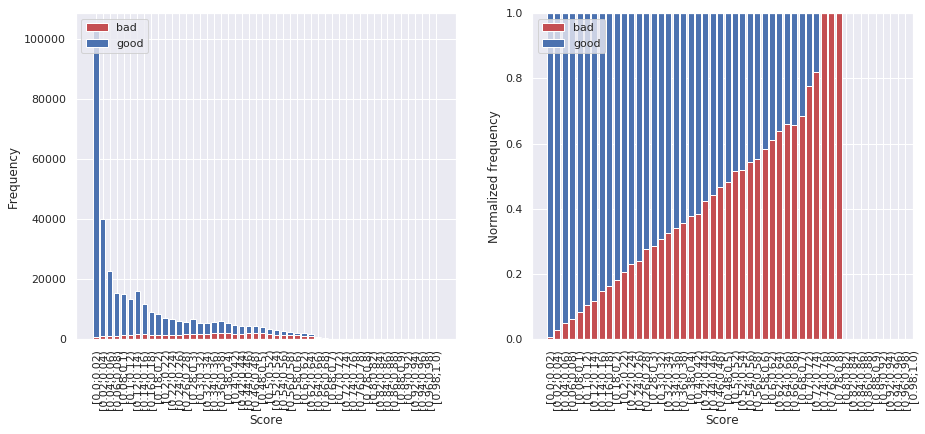
A graph with blue bars

Description automatically generated

## GINI in time



## Score distribution



A close-up of a graph

Description automatically generated