Pngme Risk Score Model Explainability

Model Release Date	2023-05-19
Model Version	`KE-9cd0f81`
Parser Version	`a-v10.0.0 / fe-v6.0.0`
Approver	Ben Fasoli

Executive Summary

Pngme's risk score model can be used by lending institutions to improve outcomes using predictive modeling methodologies based on a rich dataset of a user's financial activities. This document details Pngme's modeling framework and presents a comprehensive evaluation of predictive performance tested against observed loan outcomes.

Data Platform Overview

Pngme's core technology consists of the following capabilities:

1. Accessing SMS data from a user's phone, with their consent
2. Extracting structured financial data from SMS sent by financial institutions to a user
3. Creating data features derived from those quantitative financial data
4. Predicting loan outcome based on a user's SMS-derived data features

In Kenya, Pngme extracts and structures data sent to users by over 150 different financial institutions. Pngme's data platform makes use of 2,353 unique regular expression templates to extract structured account balances and debit/credit transactions for a user's depository accounts, as well as loan disbursements and repayments for a user's loans. The extraction also identifies discrete financial events, such as instances of previous loan repayments. These structured data are summarized into features describing the user's past financial activity, for example, the count of a user's previous loan defaults.

Pngme's Risk Score model uses these financial data to predict the probability that the user will repay or default on a future short-term loan. The Risk Score was developed and is appropriate for short-term (30d) type loan products.

Model Overview

Pngme's model predicts the probability of term loan defaults. We train an eXtreme Gradient Boosted Trees Classifier (XGBoost) model to predict the probability of term loan defaults using 37 different financial features calculated for each of 47,887 term loans issued by multiple Kenyan lenders. We define a bad outcome as an instance of a loan default notification from a lender, and a good outcome as an instance of a loan (fully) repaid notification from a lender.

Model Performance

Pngme continuously tests model performance during development and after deployment to evaluate model results and reliability. The Risk Score predicts the probability of term loan default which is evaluated against observations of true loan outcomes. We use cross validation with time series resampling to quantify prediction uncertainty using out-of-sample loan outcomes. For each performance statistic, we report the arithmetic mean and 95% confidence interval of cross validation results.

Gini	0.43 ± 0.13
AUC-ROC	0.71 ± 0.06
Log-loss	0.55 ± 0.07

Data Platform

Data Sources

Financial data is collected from SMS messages stored on users' mobile phones using Pngme's Android SDK. Pngme uses 2,353 unique regular expression templates to extract structured account balances, transactions, and behavioral labels related to a user's depository and loan accounts. In Kenya, Pngme supports over 150 different financial institutions.

Model Training Dataset

Pngme's model is trained using 47,887 observed loan outcomes for 19,706 users. Specifically, Pngme observes loan originations and loan terminations (repaids or defaults) in the SMS history of 19,706 users, and trains against these observed outcomes. The term loans for these users were issued by the following lending institutions:

Institution name	**Term loan count**
fairkash	15377
mshwari	4548
hustlerfund	4190
zashloan	4154
easycash	3531
creditmoja	2226
flashpesa	2006
lendplus	1802
zillions	1702
kashbeanke	1588
grolatech	894
instapesa	884
directcash	720

cloudloan	586
linkcash	500
mykes	264
credithela	226
kcb	200
timiza	198
cashnow	173
okash	162
hikash	152
kcbmpesa	148
creditkes	128
kcb_mpesa	126
loanpesa	114
gogopesa	105
other (<100)	1183

Feature Definitions

The following features are used in the current production model. See the [Feature Selection](#) section for a discussion of how these features were selected among the 95+ feature candidates considered for this model.

average_end_of_day_loan_balance_0_30

The time-average of the user's balances, for all balances observed in SMS from known lenders, across all known lenders. The average is calculated over a period of zero to 30 days prior to the prediction date.

average_end_of_day_loan_balance_31_90

Same as average_end_of_day_loan_balance_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

count_loan_missed_payment_events_0_30

The number of SMS received indicating that the user missed a loan payment, for a loan from any known lender. The count is summed over a period of zero to 30 days prior to the prediction date.

count_loan_missed_payment_events_31_90

Same as count_loan_missed_payment_events_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

count_loan_opened_events_0_30

The number of SMS received indicating that the user either had a loan approved and/or had a loan disbursed, from any known lender. The count is summed over a period of zero to 30 days prior to the prediction date.

count_loan_opened_events_31_90

Same as count_loan_opened_events_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

count_insufficient_funds_events_0_30

The number of SMS received indicating that the user had insufficient funds to conduct a transaction. The count is summed over a period of zero to 30 days prior to the prediction date.

count_insufficient_funds_events_31_90

Same as count_insufficient_funds_events_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

count_of_institutions_0_30

The number of distinct financial institutions from which the user received one or more financial SMS during a period of zero to 30 days prior to the prediction date.

count_of_institutions_0_30

Same as count_of_institutions_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

count_transactions_depository_0_30

The number of SMS received indicating that the user conducted a depository transaction. The count is summed over a period of zero to 30 days prior to the prediction date.

count_transactions_depository_31_90

Same as count_transactions_depository_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

daily_average_of_stacked_loan_alerts_0_30

The daily average number of lenders with whom a user has received a loan-related financial alert over a period of zero to 90 days prior to the prediction date. Loan related alerts include LoanDefaulted, LoanMissedPayment, LoanRepaid, LoanApproved, LoanDisbursed, LoanRepayment, and LoanRepaymentReminder. The daily average number of unique lenders indicates user activity with multiple lenders on the same days.

daily_average_of_stacked_loan_alerts_31_90

Same as daily_average_of_stacked_loan_alerts_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

difference_count_of_loans_opened_to_loans_repaid_0_30

The difference between a) the number of SMS received indicating that the user was approved for a loan, or had a loan disbursed (of any loan to any lender) and b) the number of SMS received indicating that the user made a loan repayment (of any loan to any lender). The counts for a) and b) are summed over a period of zero to 30 days prior to the prediction date.

difference_count_of_loans_opened_to_loans_repaid_31_90

Same as difference_count_of_loans_opened_to_loans_repaid_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

difference_count_of_loans_opened_to_loans_delinquent_0_30

The difference between a) the number of SMS received indicating that the user was approved for a loan, or had a loan disbursed (of any loan to any lender) and b) the number of SMS received indicating that the user is past due or defaulted on a loan (of any loan to any lender). The counts for a) and b) are summed over a period of zero to 30 days prior to the prediction date.

difference_count_of_loans_opened_to_loans_delinquent_31_90

Same as difference_count_of_loans_opened_to_loans_delinquent_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

median_end_of_day_depository_balance_0_30

The time-median of the user’s total balance summed across depository accounts at the end of each day. Balances are observed in SMS from known financial institutions. The average is calculated over a period of zero to 30 days prior to the prediction date.

median_end_of_day_depository_balance_31_90

Same as median_end_of_day_depository_balance_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

min_end_of_day_depository_balance_0_30

The minimum end-of-day (EOD) depository total balances, across all institutions where balances are observed in SMS. Total balances means the sum of EOD balances across all institutions on a daily basis. End-of-day (EOD) means the most recent notification of account balance backward looking from the end of each calendar day. The minimum is taken on the EOD observations over a period of zero to 30 days prior to the prediction date.

min_end_of_day_depository_balance_31_90

Same as min_end_of_day_depository_balance_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

net_cash_flow_0_30

Difference between credit and debit transactions across all depository accounts. Transactions are observed in SMS from known financial institutions. The difference is calculated over a period of zero to 30 days prior to the prediction date.

net_cash_flow_31_90

Same as net_cash_flow_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

stdev_end_of_day_day_depository_balance_0_30

The standard deviation of end-of-day (EOD) depository total balances, across all institutions where balances are observed in SMS. Total balances means the sum of EOD balances across all institutions on a daily basis. End-of-day (EOD) means the most recent notification of account balance backward looking from the end of each calendar day. The standard deviation is calculated on the EOD observations over a period of zero to 30 days prior to the prediction date.

stdev_end_of_day_day_depository_balance_31_90

Same as stdev_end_of_day_day_depository_balance_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

sum_of_airtime_credits_0_30

Total of credit transactions across all airtime accounts. Transactions are observed in SMS from known airtime providers. The sum is calculated over a period of zero to 30 days prior to the prediction date.

sum_of_airtime_credits_0_31_90

Same as sum_of_airtime_credits_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

sum_of_credits_0_30

Total of credit transactions across all depository accounts. Transactions are observed in SMS from known financial institutions. The sum is calculated over a period of zero to 30 days prior to the prediction date.

sum_of_credits_0_31_90

Same as sum_of_credits_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

sum_of_debits_0_30

Total of debit transactions across all depository accounts. Transactions are observed in SMS from known financial institutions. The sum is calculated over a period of zero to 30 days prior to the prediction date.

sum_of_debits_0_31_90

Same as sum_of_debits_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

sum_of_loan_repayments_0_30

The sum of repayment amounts (credit transactions) for all loans from known lenders, where an repayment amount is observed in the SMS indicating the repayment. The summation is computed for all repayment events occurring over a period of zero to 30 days prior to the prediction date.

sum_of_loan_repayments_31_90

Same as sum_of_loan_repayments_0_30 except calculated over a period of 31 to 90 days prior to the prediction date.

slope_end_of_day_depository_balance_0_90

Rate of change of daily total balance held in all depository accounts. Total balances means the sum of end-of-day balances across all institutions on a daily basis. End-of-day means the most recent notification of account balance backward looking from the end of each calendar day. The linear slope is calculated in time using an ordinary least squares regression for the period of zero to 30 days prior to the prediction date.

slope_end_of_day_loan_balance_0_90

Rate of change of daily term loan debt totaled across all open loans. Total balances means the sum of end-of-day balances across all open loans on a daily basis. End-of-day means the most recent loan balance notification backward looking from the end of each calendar day. A linear slope is calculated using an ordinary least squares regression to represent the time rate of change over the period of zero to 30 days prior to the prediction date.

sum_of_depository_balances_latest

The sum of the latest balances for depository accounts, across all institutions where balances are observed in SMS. Latest means the most recent notification of account balance backward looking from the prediction date.

Statistics for Training Featureset

The features and distributions of feature values used to train Pngme's model are as follows.

Feature name	**Unique**	**P1**	**P50**	**P99**	**Missing**
average_end_of_day_loan_balance_0_30	38,766	0	9,921	222,285	0.1%
average_end_of_day_loan_balance_31_90	35,003	0	4,888	198,033	15.6%
count_loan_missed_payment_events_0_30	266	0	3	69	0.2%
count_loan_missed_payment_events_31_90	7,973	0	2	98	16.2%
count_loan_opened_events_0_30	166	0	7	51	0.2%
count_loan_opened_events_31_90	7,846	0	3	53	16.2%
count_insufficient_funds_events_0_30	213	0	6	53	0.2%
count_insufficient_funds_events_31_90	7,942	0	8	96	16.2%
count_of_institutions_0_30	45	3	14	32	0.0%
count_of_institutions_31_90	42	0	9	31	0.0%
count_transactions_depository_0_30	1,365	5	226	1,077	0.0%
count_transactions_depository_31_90	1,893	0	240	1,616	0.0%
daily_average_of_stacked_loan_alerts_0_30	1,349	1	2	6	0.0%
daily_average_of_stacked_loan_alerts_31_90	2,357	0	1	5	0.0%
difference_count_of_loans_opened_to_loans_repaid_0_30	301	-74	-10	19	0.2%
difference_count_of_loans_opened_to_loans_repaid_31_90	8,020	-125	-13	11	16.2%
difference_count_of_loans_opened_to_loans_delinquent_0_30	325	-55	2	40	0.2%
difference_count_of_loans_opened_to_loans_del	8,021	-75	0	31	16.2%

inquent_31_90					
median_end_of_day_depository_balance_0_30	25,793	0	761	51,049	0.4%
median_end_of_day_depository_balance_31_90	28,563	0	669	47,313	20.2%
min_end_of_day_depository_balance_0_30	7,786	-52,276	0	4,329	0.4%
min_end_of_day_depository_balance_31_90	14,118	-53,313	0	5,465	20.2%
net_cash_flow_0_30	41,809	-3,675,329	-41,738	324,269	0.2%
net_cash_flow_31_90	40,738	-6,244,753	-66,286	692,354	15.7%
stdev_end_of_day_depository_balance_0_30	39,701	0	3,093	106,979	0.4%
stdev_end_of_day_depository_balance_31_90	40,187	0	4,114	131,979	20.2%
sum_of_airtime_credits_0_30	13,757	10	1,130	9,175	11.4%
sum_of_airtime_credits_31_90	29,460	20	2,378	17,496	45.9%
sum_of_credits_0_30	36,471	292	137,262	3,924,942	0.2%
sum_of_credits_31_90	35,255	0	178,340	6,757,526	15.7%
sum_of_debits_0_30	41,693	1,310	207,392	6,164,133	0.2%
sum_of_debits_31_90	40,143	0	287,690	11,097,810	15.7%
sum_of_loan_repayments_0_30	27,288	0	5,400	317,551	0.1%
sum_of_loan_repayments_31_90	25,031	0	2,209	410,913	14.5%
slope_end_of_day_depository_balance_0_90	37,653	-1,826	3	1,930	0.4%
slope_end_of_day_loan_balance_0_90	38,680	-1,033	108	6,192	1.3%
sum_of_depository_balances_latest	32,818	-16,142	1,563	153,183	0.4%

Model Theoretical Framework

Pngme uses an eXtreme Gradient Boosted Trees Classifier (XGBoost) model, which produces more accurate predictions while relying on fewer assumptions than traditional statistical models. For example, logistic regression models assume the data generating process is additively separable before applying the logistic link function and generally use hypothesis tests (e.g. t-tests) which often misrepresent the distributions of feature values. In contrast, XGBoost does not rely on these strong distributional assumptions.

Feature Selection

We calculate 95 different financial features summarizing various dimensions of a user's behavior. Of these 95 features, we follow an iterative feature selection process to identify features with the best predictive power and the least distributional drift in time. Our strategy is based on two metrics which summarize the impact of each feature on the resulting model:

1. **SH**apley **A**dditive ex**P**lanations (SHAP) values determine how strongly a feature influences the final model by iteratively randomizing feature values and quantifying the resulting change to model predictions. High SHAP values indicate important features.
2. **P**opulation **S**tability **I**ndex (PSI) discretizes feature distributions and iteratively tests for changes to feature distributions over time. We compute fixed decile bins and use PSI to quantify changes to the relative fraction of feature values falling within each bin over time. Low PSI values indicate features with stable distributions over time.

Combining the SHAP and PSI values ensures Pngme's model can produce robust predictions for future lending decisions.

Training and Evaluation Strategy

We use cross validation with time series resampling to quantify prediction uncertainty using out-of-sample loan outcomes. Cross validation allows us to evaluate prediction uncertainty and guard against overfitting to the training data by iteratively training and evaluating an ensemble of similar models, each trained on a subset of the available training data.

We split loan outcomes to form 11 folds, or distinct subsets of the total training population. We assign the oldest 50% of the training set to the first fold then evenly divide the remaining 50% of observations into 10 additional folds ordered in time, each containing 5% of the training set.

Folds from the time series splits are used to iteratively train an ensemble of models. The first model of the ensemble is trained on the first fold and evaluated by predicting outcomes in the second fold. Each subsequent iteration grows the size of the training set by including one additional fold and evaluates the resulting model against the following fold, such that exactly one model in the ensemble has predicted every outcome in the second 50% of the training set. For example, the model trained on the first 65% of the training set is evaluated against the following 5% of observed outcomes. We can estimate a representative value and confidence interval for reported performance statistics using the distributions given by the cross validation ensemble.

Creating folds from time series splits enforces that the test set always occurs after the training set. In other words, no model is allowed to learn from future outcomes. We use time series splits instead of the more common k-fold cross validation, where the training set is randomly sampled without replacement into distinct folds, to enable our model evaluation framework to test for drift in predictions over time.

Hyperparameter Optimization

XGBoost models use a set of hyperparameters which configure how the model learns from the training data. We perform an exhaustive grid search over combinations of hyperparameter values and use cross validation to identify the set of hyperparameters which give the highest ROC-AUC evaluated against out-of-sample loan outcomes while guarding against overfitting to the training data. The hyperparameters governing how the model learned from the training set are as follows.

Number of unique trees (n_estimators)	500
Max tree depth (max_depth)	8
Tree weight adjustment rate (learning_rate)	0.01
Iterative resampling ratio (subsample)	0.5

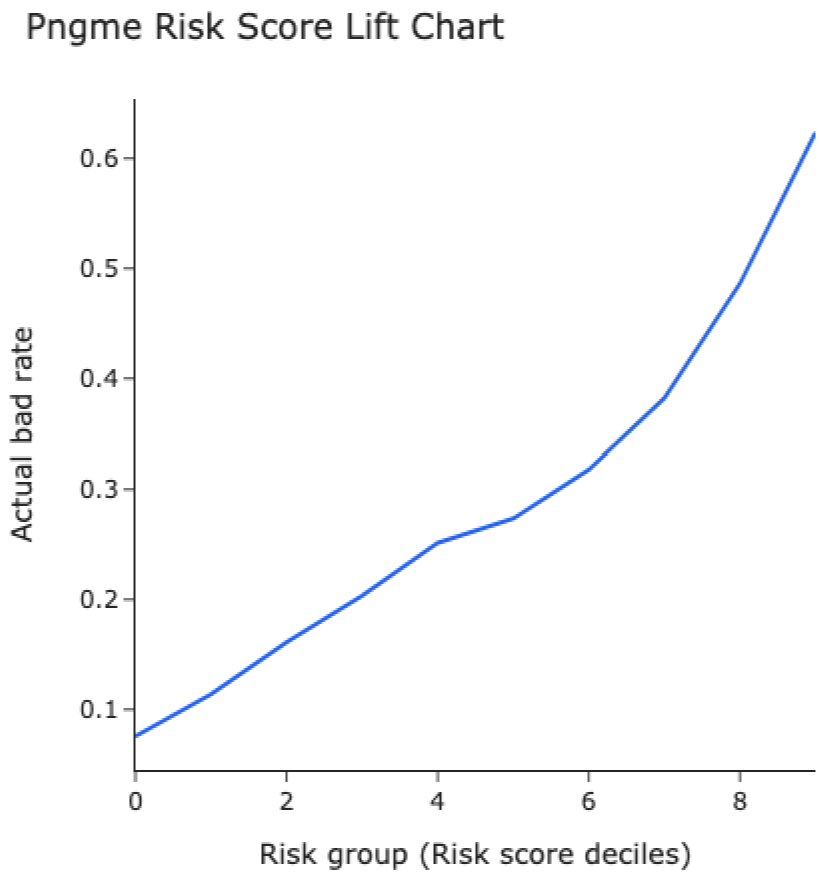
Model Performance

We use cross validation with time series resampling to quantify prediction uncertainty using out-of-sample loan outcomes. For each performance statistic, we report the arithmetic mean and 95% confidence interval of cross validation results.

Gini	0.43 ± 0.13
AUC-ROC	0.71 ± 0.06
Log-loss	0.55 ± 0.07

Lift Chart

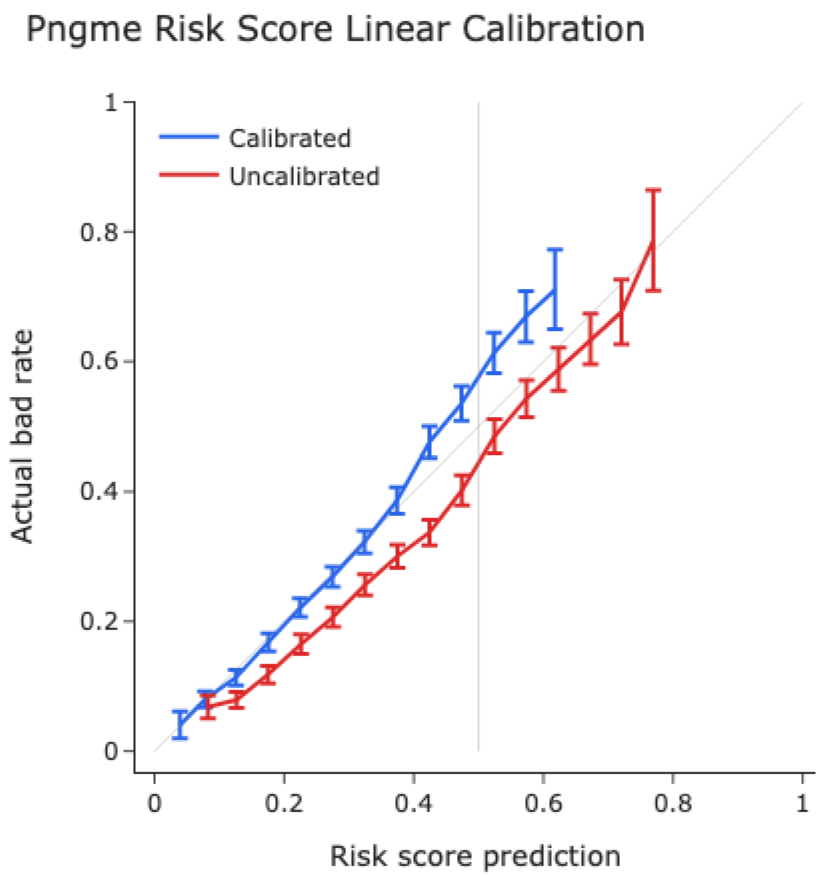
The lift chart visualizes the effectiveness of Pngme's predictive model relative to performance without using the predictive model. We discretize predictions into deciles and compare each risk group with the aggregate observed term loan outcomes within each group. For example, the 0-1 group represents the actual bad-rates observed in the cross-validation holdout loan outcomes for the 10% of lowest predicted outcomes. Conversely, the 9-10 group represents the observed bad rates for the 10% of highest predicted outcomes.



Calibration: Risk Score vs Actual Outcome

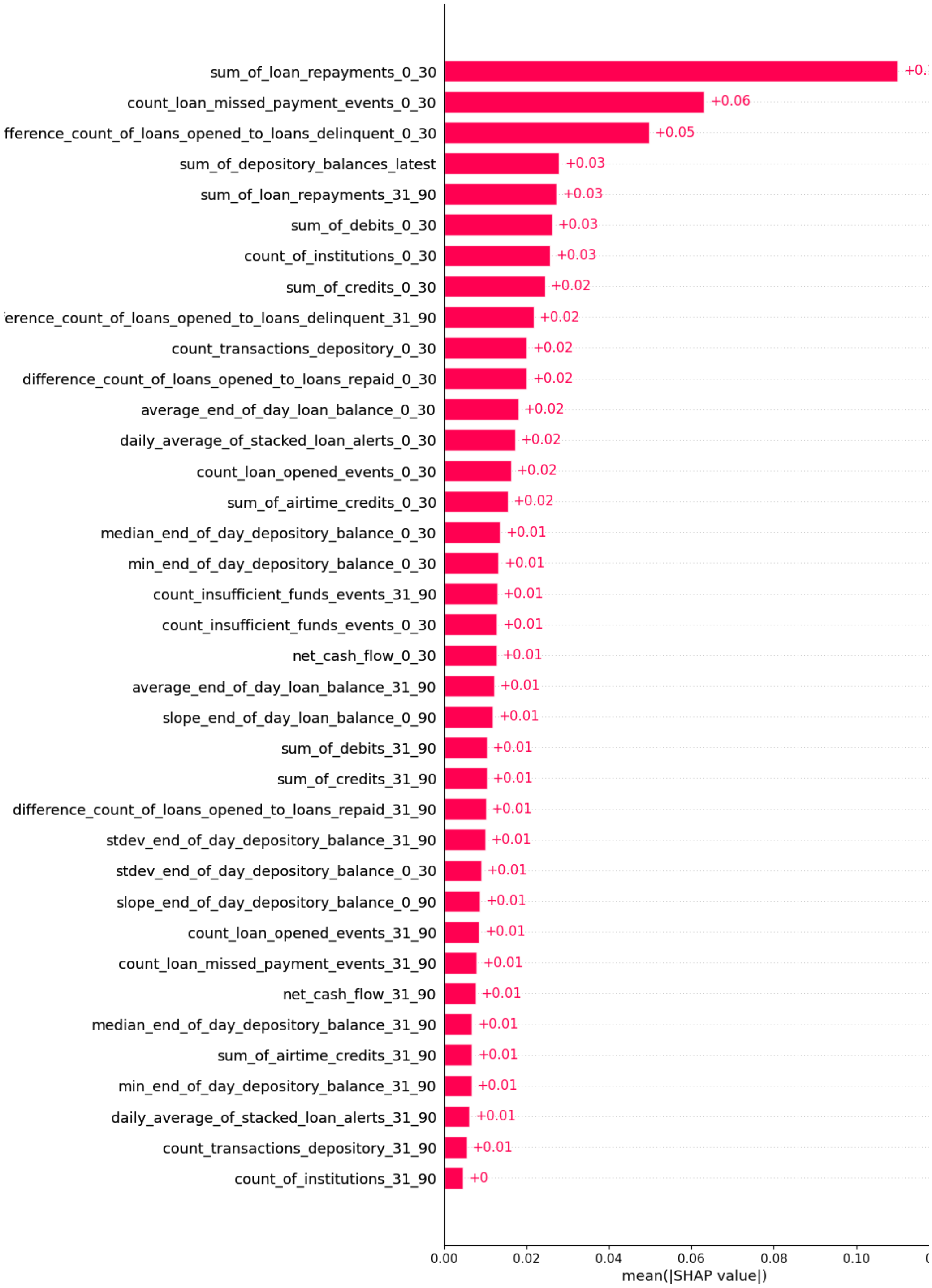
We bin Risk Score predictions from the cross validation holdout folds and calculate the fraction of true delinquent loan outcomes within each bin. We then use a linear regression to calibrate predicted probabilities to observations. Rather than fitting regression coefficients to the full range of predictions, we calibrate based on predictions where the probability of delinquency is less than 50%. This range is more representative of standard lending use cases and guards against potential non-linear bias introduced by higher predictions.

The below visualization shows the bad-rate as observed in the cross validation holdout training data versus the predicted values for those loan instances. Bad-rates are calculated between the actual predicted on bins of 0.05 increment.



Feature Importance: SHAP Values

We use **SH**apley **A**dditive ex**P**lanations (SHAP) values to determine the importance of each feature on the resulting model. These importances indicate how sensitive model predictions are to changes in feature values, rather than testing for correlation between feature values and observed outcomes. This approach to estimate feature importance for machine-learning based models is analogous to information value (IV) for logistic regression models.



Feature Conceptual Coherence: SHAP Beeswarm

The SHAP beeswarm is a visual representation of how feature values impact predicted outcomes. The beeswarm is best interpreted qualitatively by looking at both the magnitude of separation (left to right) between feature value observations and predicted outcome spread, and consistency of gradient (blue to red) between low feature values to high features values, indicating strong correlation between feature value and predicted outcomes. This relates feature values to model impact, both in magnitude and directionality.