Klarna
Credit Model

Klarna Case - 02/2022

Index

1- Introduction

2- Data

3- Solutions

4- Architecture

5- Model

6- Results

Introduction

Introduction

Klarna is a swedish fintech with operations throughout Europe, US and Australia. The company offers online payment solutions for storefronts and direct clients, as well as credit products to theirs customers.

With the intent of diminishing the company financial exposure while at the same time allowing for the expansion the company’s credit products, this project sought the creation of a new ML model for providing a more assertive default forecast.

Data

Raw Data

- Partial data provided for the project
- 99.976 accounts (89.976 for training and 10.000 for validation)
    - Default = 0: 88.688 accoutns
    - Default = 1: 1.288 accounts
- Very unbalanced dataset
    - 1.4% of accounts in default
- 43 columns:
    - 1 ID column + 1 target column
    - 41 data columns
        - 37 numerical colums
        - 4 categorical columns

Features

- Categorical variables encoded using LabelEncoder
- Features selected using a Lasso regression for calculating L1 correlation between each feature and target
    - 5 features dropped
- In-depth analysis of data correlation in appendix

Dropped Features:

- account_incoming_debt_vs_paid_0_24m
- num_arch_written_off_0_12m
- status_max_archived_0_24_months'
- account_worst_status_3_6m'
- status_last_archived_0_24m

Solutions

Solutions

Developed:

- Classification model for default probability estimation
- Airflow platform with custom operators for automating model (re)training and batch predictions using SageMaker
- KServe endpoint for online prediction
- Kubernetes environment for Airflow containers orchestration
- CI/CD pipeline built on CodePipeline for deploying the project resources
- Versioning of the developed solutions using a Github repository

Not developed:

- Redundant environments for Development and Production
- Automatic cloud resources management via Terraform
- Automatic cluster scaling via EKS
- Automatic scaling and authentication of the online prediction endpoints
- Model versioning via DVC or MLFlow

S3 & Athena

Object storage service provided by AWS for serving files across different applications.

The provided data and all resources generated throughout the case where stored in S3 buckets.

Online data analsysis for data stored in S3. Tables were created for loading the provided data in order to simulate a real DataLake.

Airflow & Kubernetes

Workflow management platform used for integrating the AWS resources used in the project.

Provided workflow automation for extracting data from the DataLake (Athena), (re)training and versioning the model and providing batch predictions. All of that was done through the use of DAGs.

Kubernetes environment for orchestration of the projects containers of both Airflow and KServe.

It was deployed on top of AWS EC2.

SageMaker & KServe

Machine Learning platform on AWS for exploratory data analysis and model training.

Both features were used in the model development process.

Model serving framework running on top of Kubernetes. It was used in the project to provide API endpoints for serving online predictions.

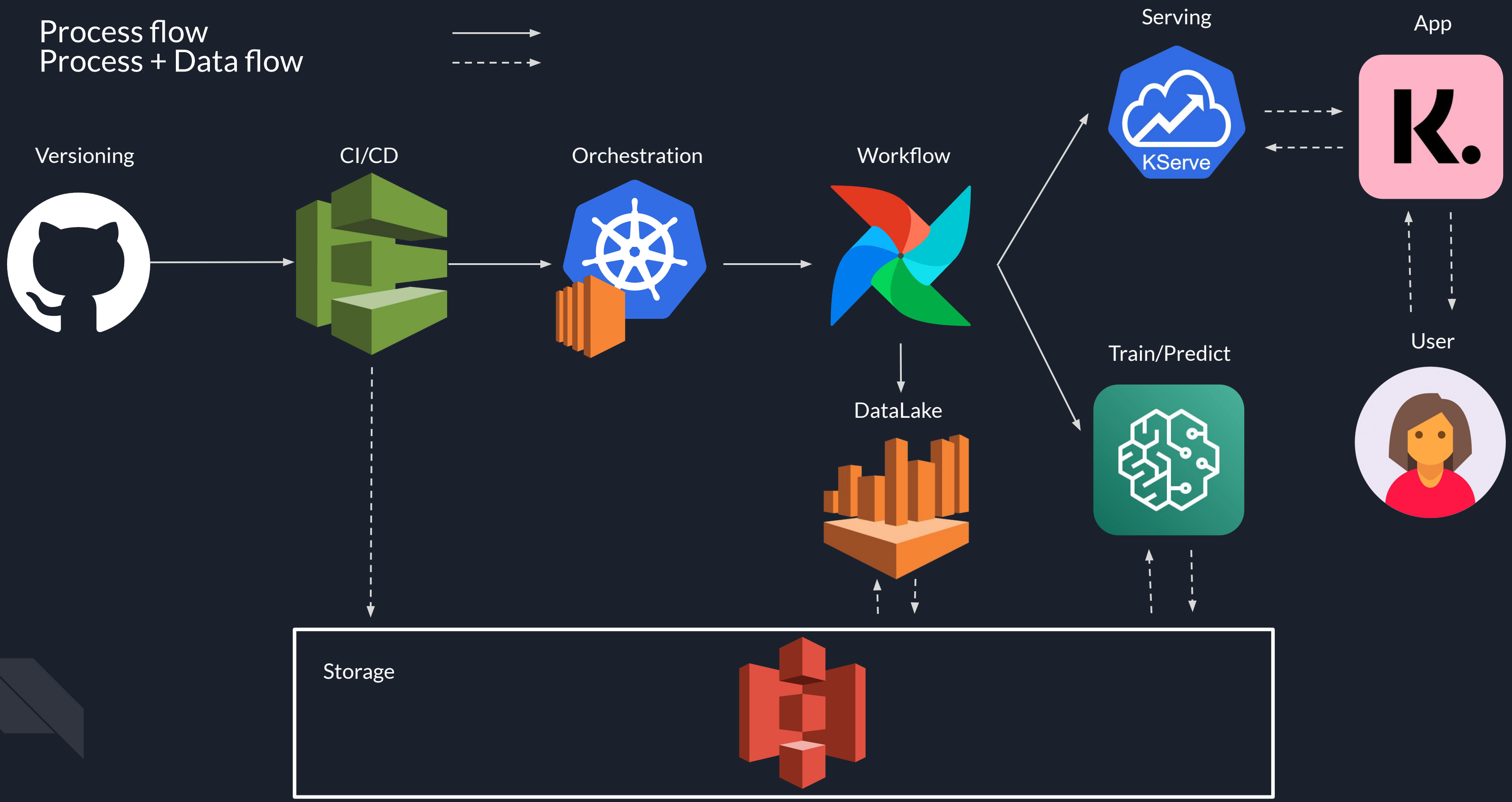
CI/CD & Versioning

Versioning of the project resources done using a GitHub repository

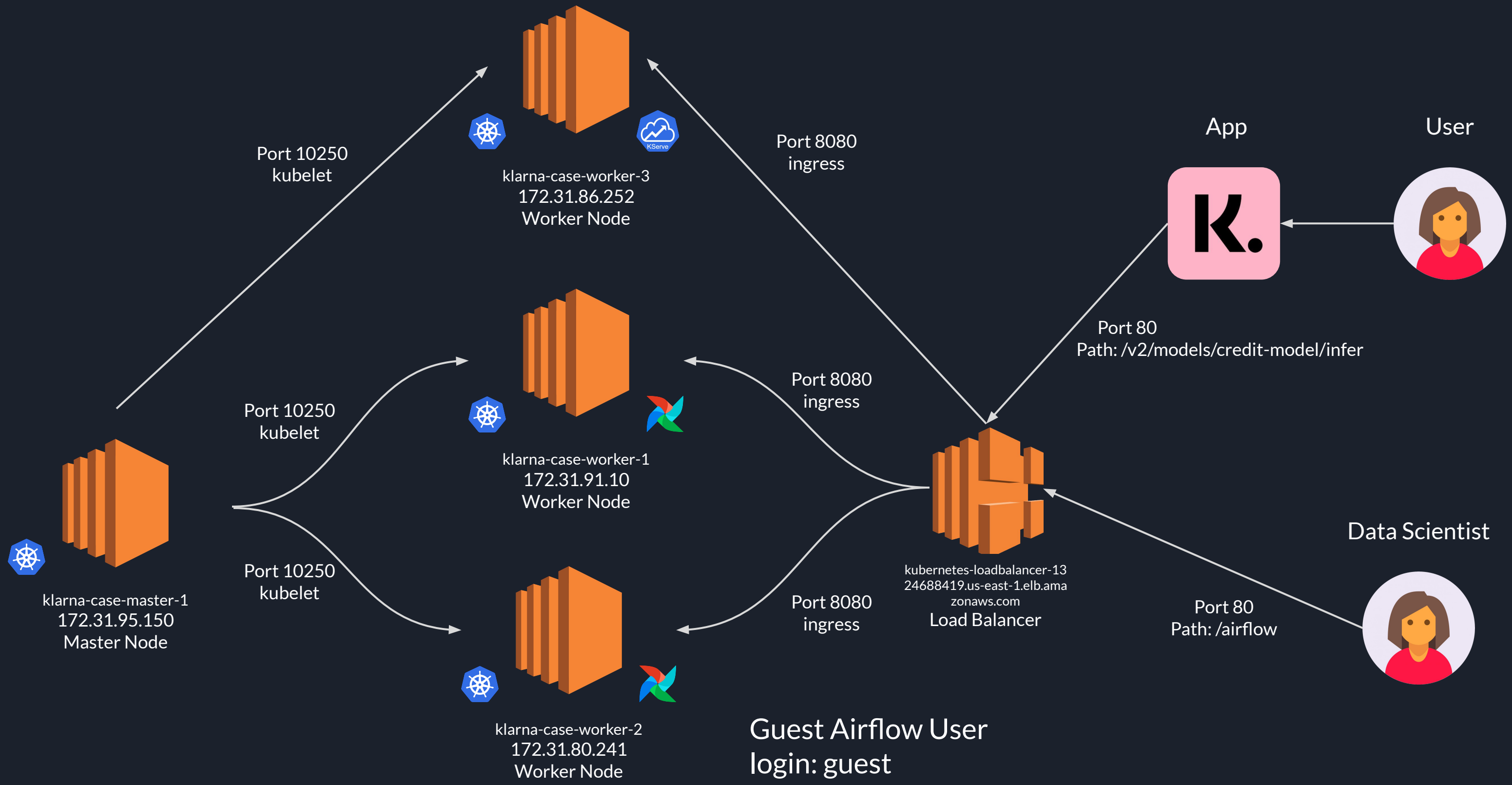
CI/CD pipeline via CodePipeline for deploying changes on the project resources upon pull requests on the project repository.

Architecture

Workflow Architecture



Systems Architecture



Guest Airflow User
login: guest
pass: guest1234

Model

Model

Selected Model: CatBoost (Tree regression model with gradient boosting)

Challenger models:

- LogisticRegression (LR)
- CatBoostClassifier (CBC)
- RandomForestClassififier (RFC)
- XGBoostClassifier (XGB)

Train-test split: 70/30

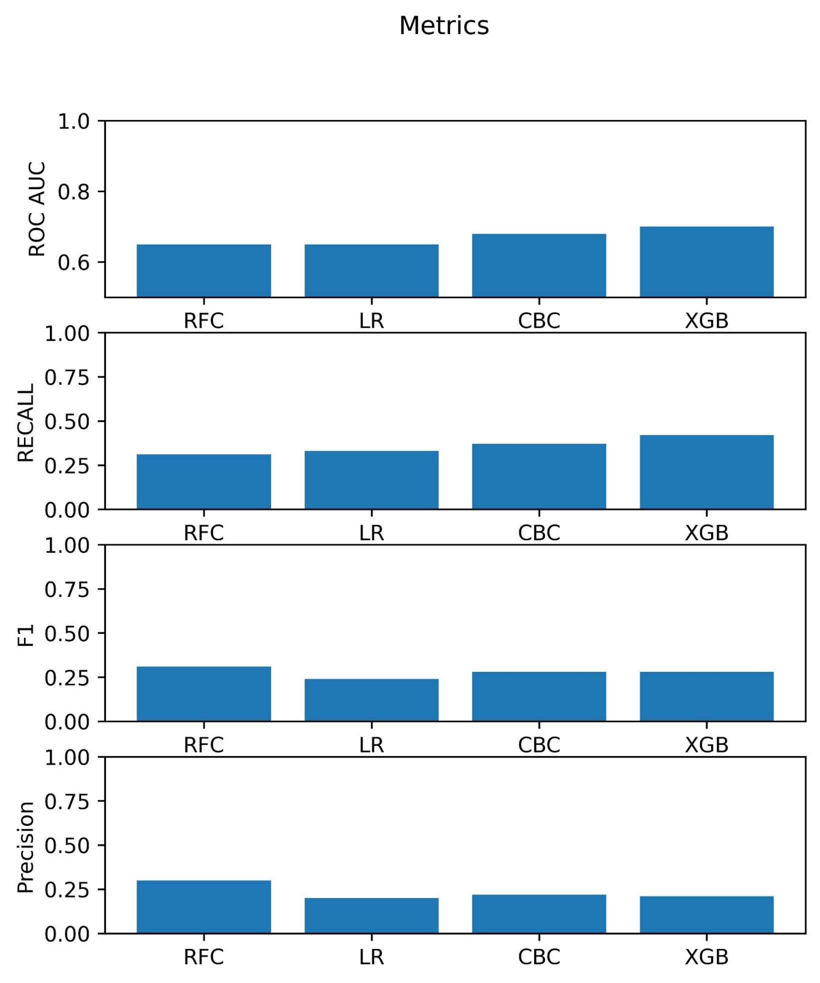
Hyperparameters tuning performed in all models

Final model calibrated using Platt calibration

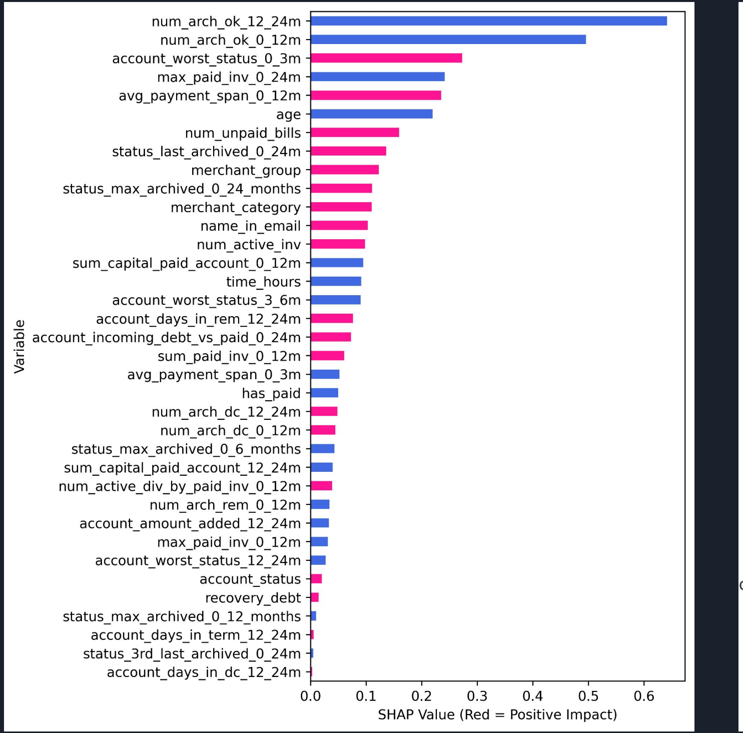
Results

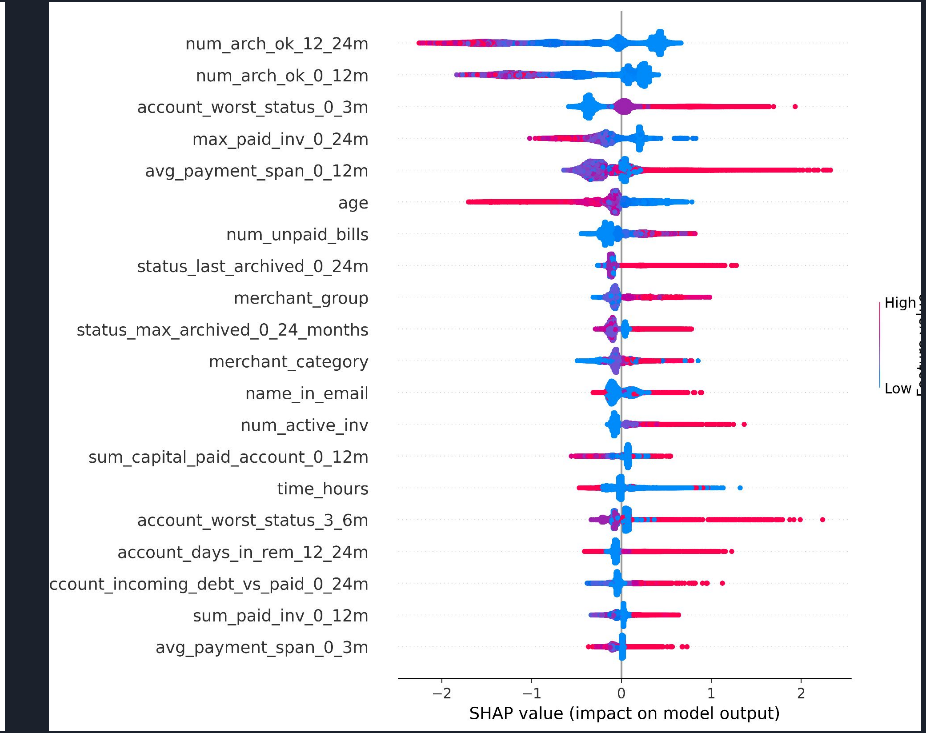
Metrics

Model	ROC AUC	F1	Recall	Precision
XGB	0.70	0.28	0.42	0.21
CBC	0.68	0.28	0.37	0.22
RFC	0.65	0.31	0.31	0.30
LR	0.65	0.24	0.33	0.20



Feature Importance

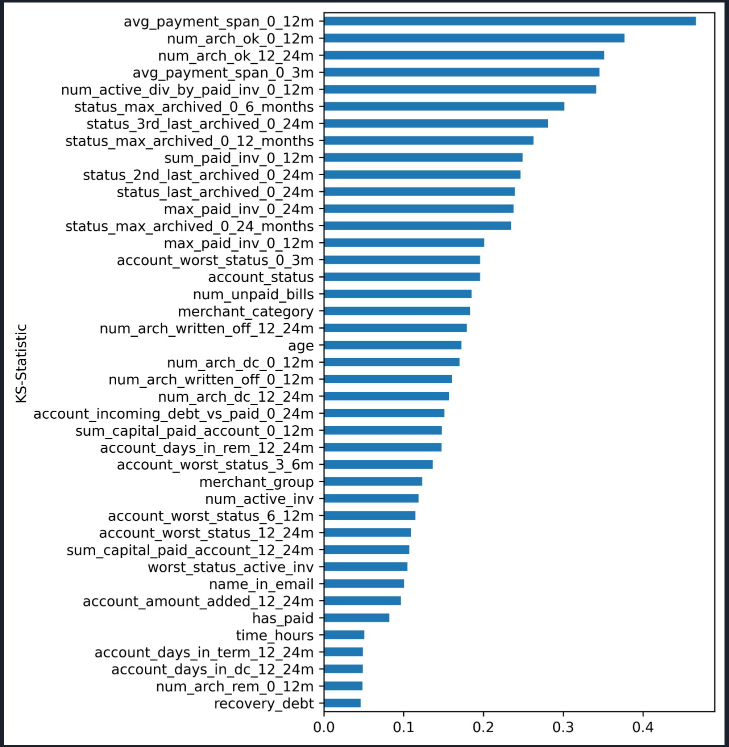




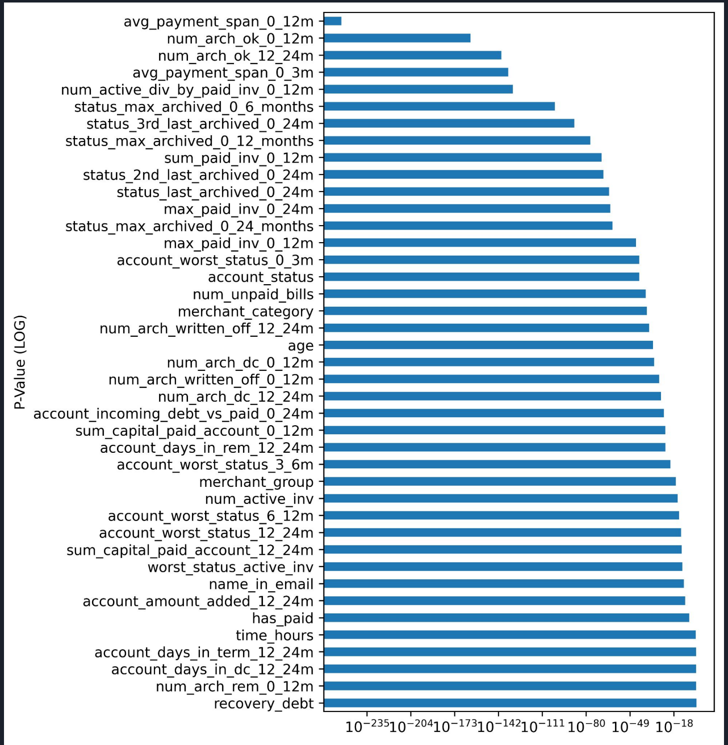
Thanks!

Appendix

KS-Test

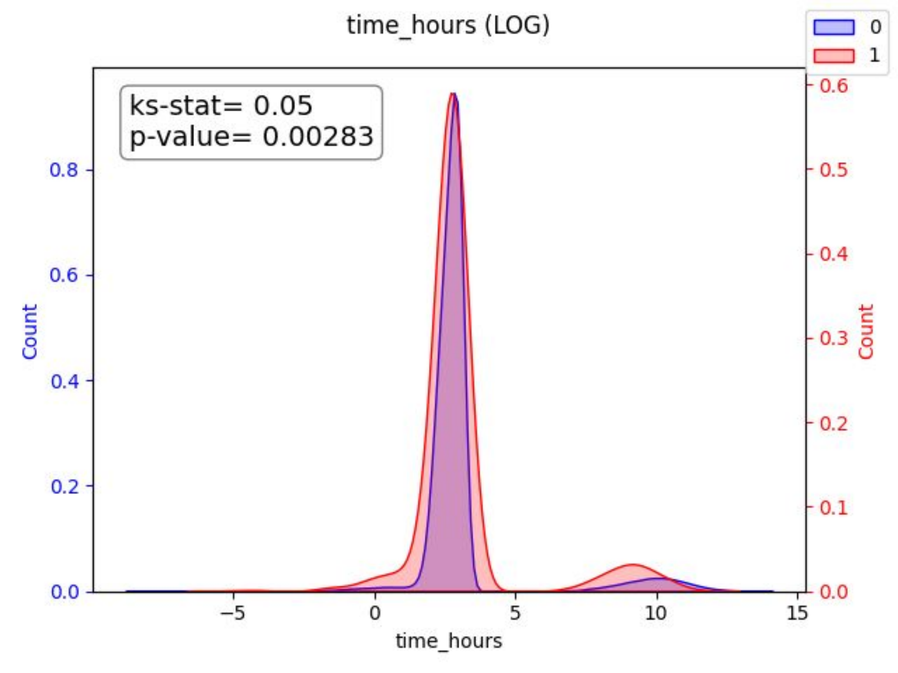


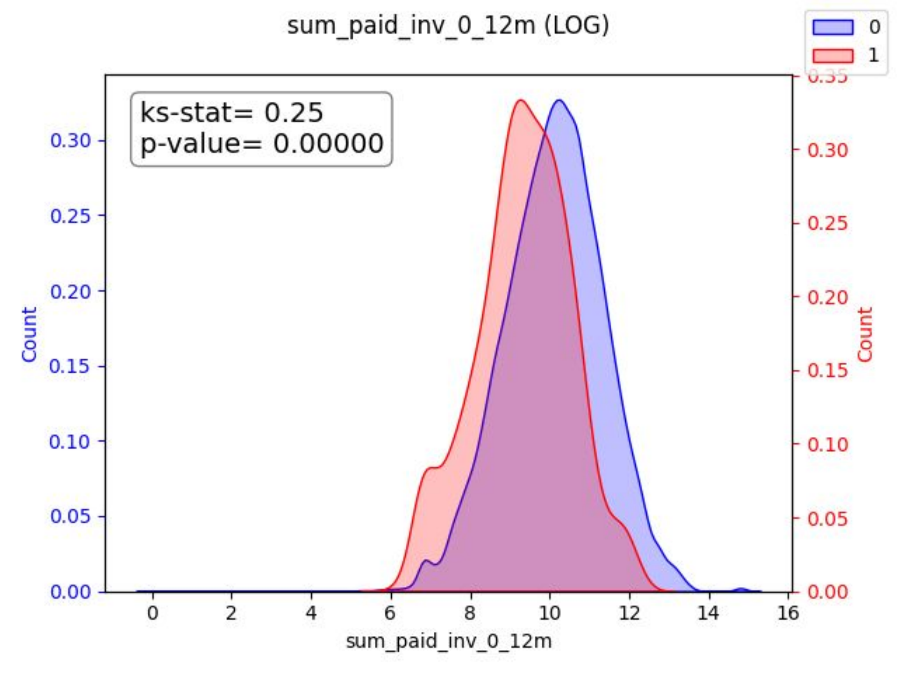
avg_payment_span_0_12m
num_arch_ok_0_12m
num_arch_ok_12_24m
avg_payment_span_0_3m
num_active_div_by_paid_inv_0_12m
status_max_archived_0_6_months
status_3rd_last_archived_0_24m
status_max_archived_0_12_months
sum_paid_inv_0_12m
status_2nd_last_archived_0_24m
status_last_archived_0_24m
max_paid_inv_0_24m
status_max_archived_0_24_months
max_paid_inv_0_12m
account_worst_status_0_3m
account_status
num_unpaid_bills
merchant_category
num_arch_written_off_12_24m
age
num_arch_dc_0_12m
num_arch_written_off_0_12m
num_arch_dc_12_24m
account_incoming_debt_vs_paid_0_24m
sum_capital_paid_account_0_12m
account_days_in_rem_12_24m
account_worst_status_3_6m
merchant_group
num_active_inv
account_worst_status_6_12m
account_worst_status_12_24m
sum_capital_paid_account_12_24m
worst_status_active_inv
name_in_email
account_amount_added_12_24m
has_paid
time_hours
account_days_in_term_12_24m
account_days_in_dc_12_24m
num_arch_rem_0_12m
recovery_debt
KS-Statistic
0.0
0.1
0.2
0.3
0.4



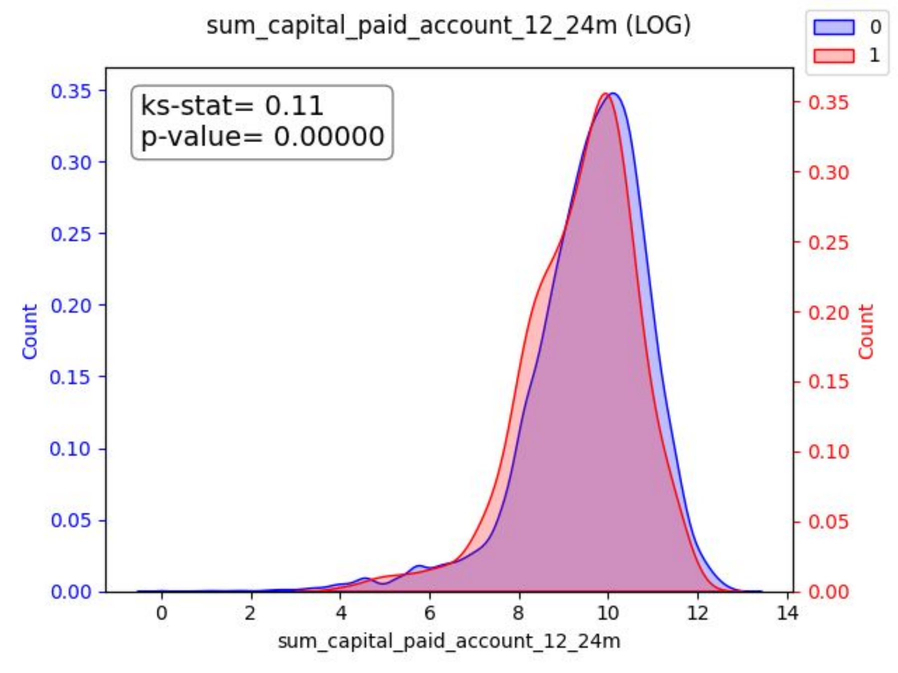
avg_payment_span_0_12m
num_arch_ok_0_12m
num_arch_ok_12_24m
avg_payment_span_0_3m
num_active_div_by_paid_inv_0_12m
status_max_archived_0_6_months
status_3rd_last_archived_0_24m
status_max_archived_0_12_months
sum_paid_inv_0_12m
status_2nd_last_archived_0_24m
status_last_archived_0_24m
max_paid_inv_0_24m
status_max_archived_0_24_months
max_paid_inv_0_12m
account_worst_status_0_3m
account_status
num_unpaid_bills
merchant_category
num_arch_written_off_12_24m
age
num_arch_dc_0_12m
num_arch_written_off_0_12m
num_arch_dc_12_24m
account_incoming_debt_vs_paid_0_24m
sum_capital_paid_account_0_12m
account_days_in_rem_12_24m
account_worst_status_3_6m
merchant_group
num_active_inv
account_worst_status_6_12m
account_worst_status_12_24m
sum_capital_paid_account_12_24m
worst_status_active_inv
name_in_email
account_amount_added_12_24m
has_paid
time_hours
account_days_in_term_12_24m
account_days_in_dc_12_24m
num_arch_rem_0_12m
recovery_debt
P-Value (LOG)
10^{-235}
10^{-204}
10^{-173}
10^{-142}
10^{-111}
10^{-80}
10^{-49}
10^{-18}

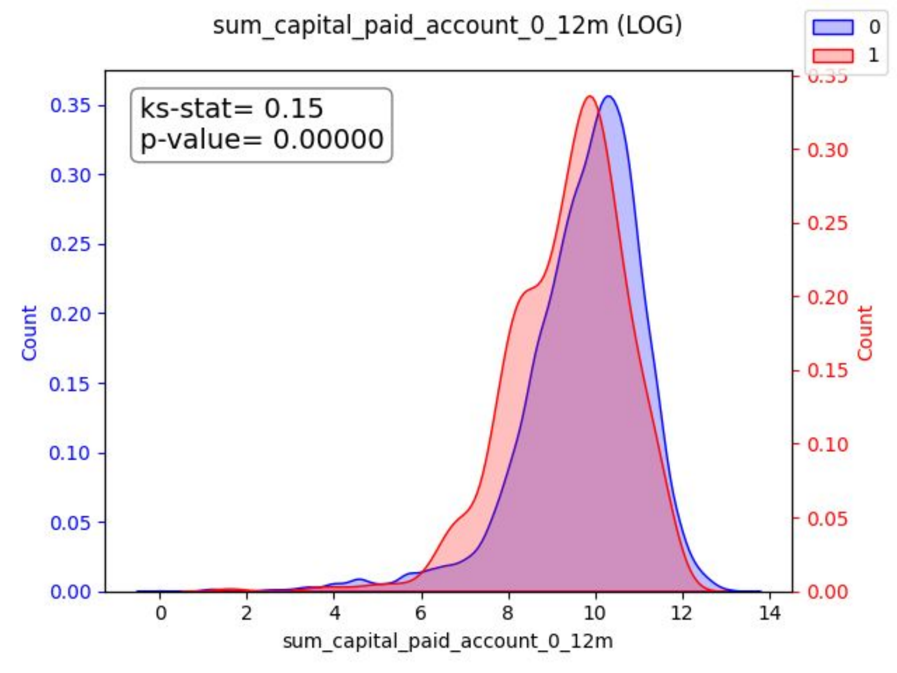
Distributions - Numerical



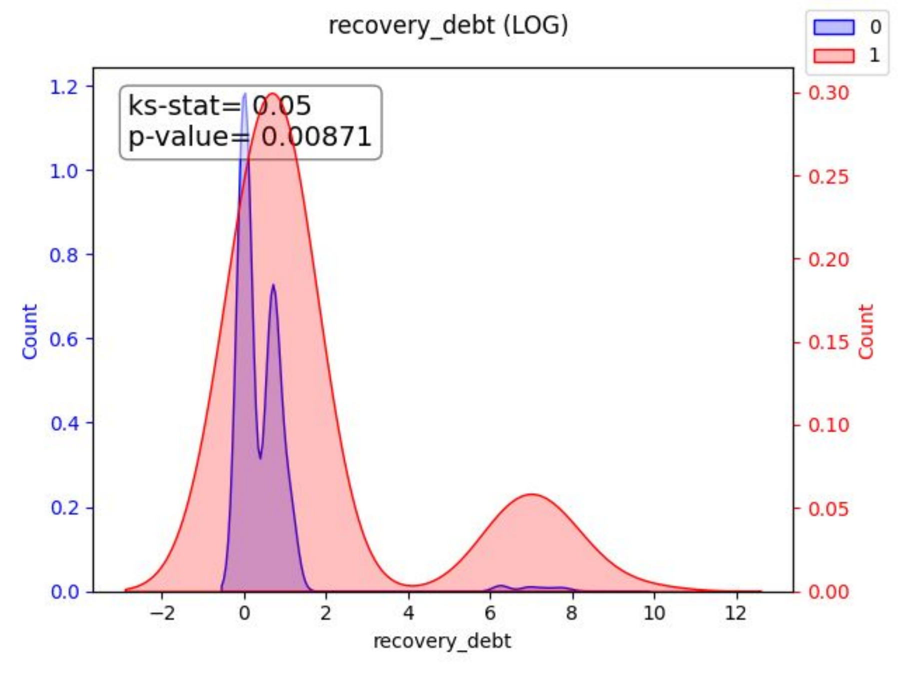


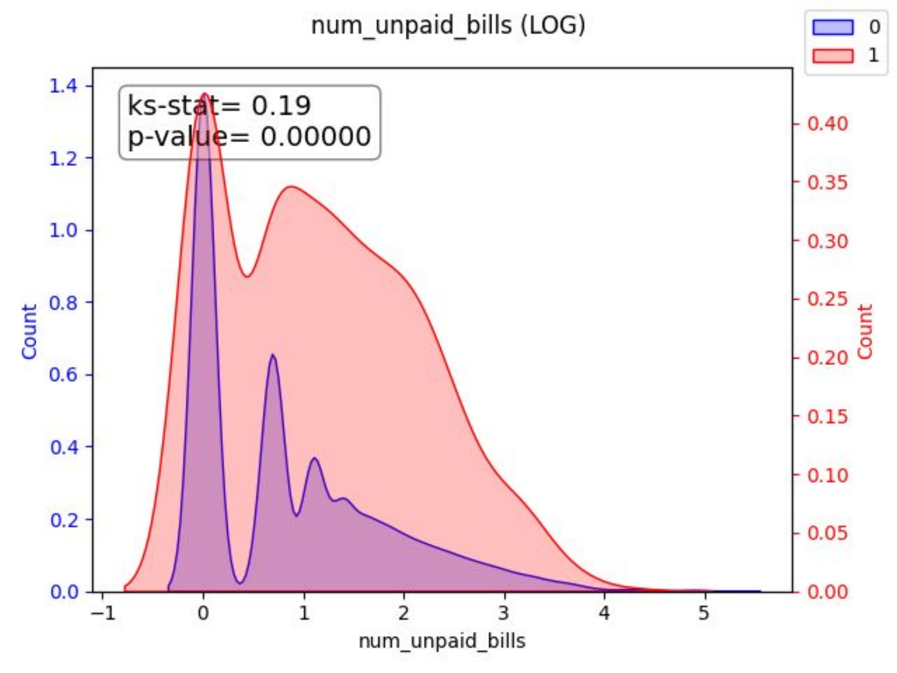
Distributions - Numerical



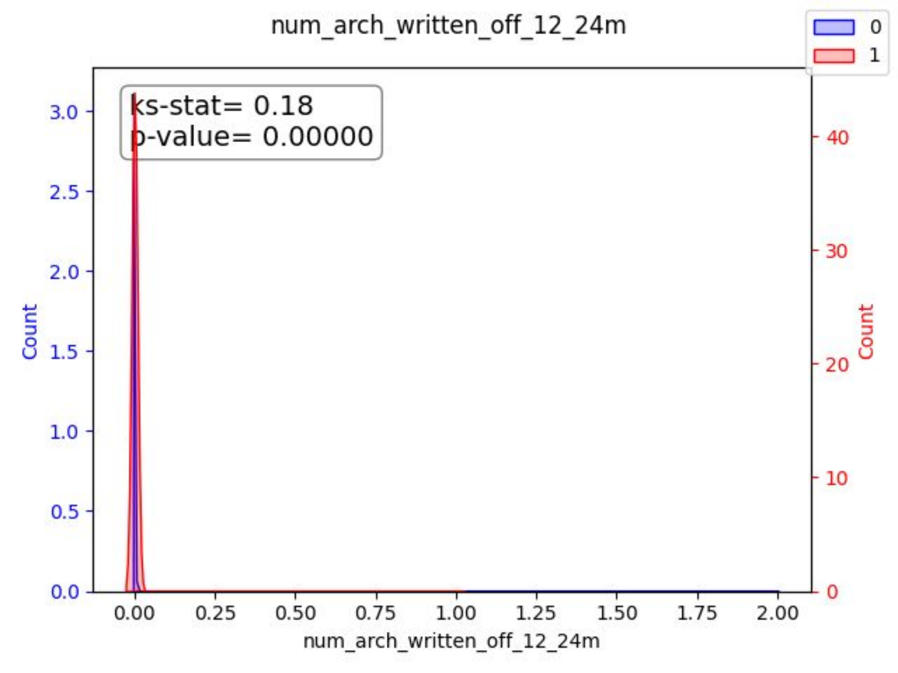


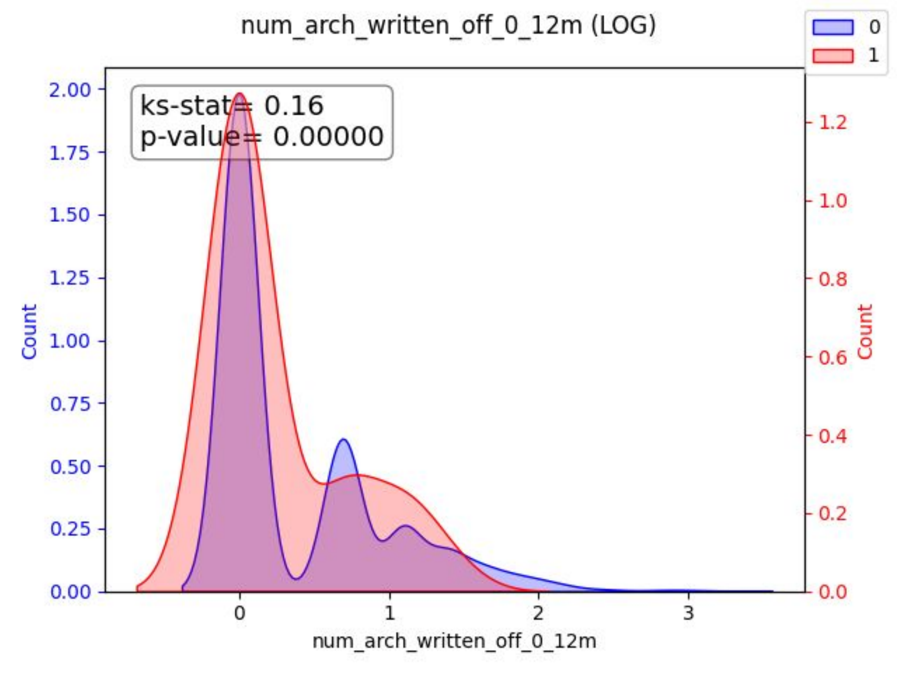
Distributions - Numerical



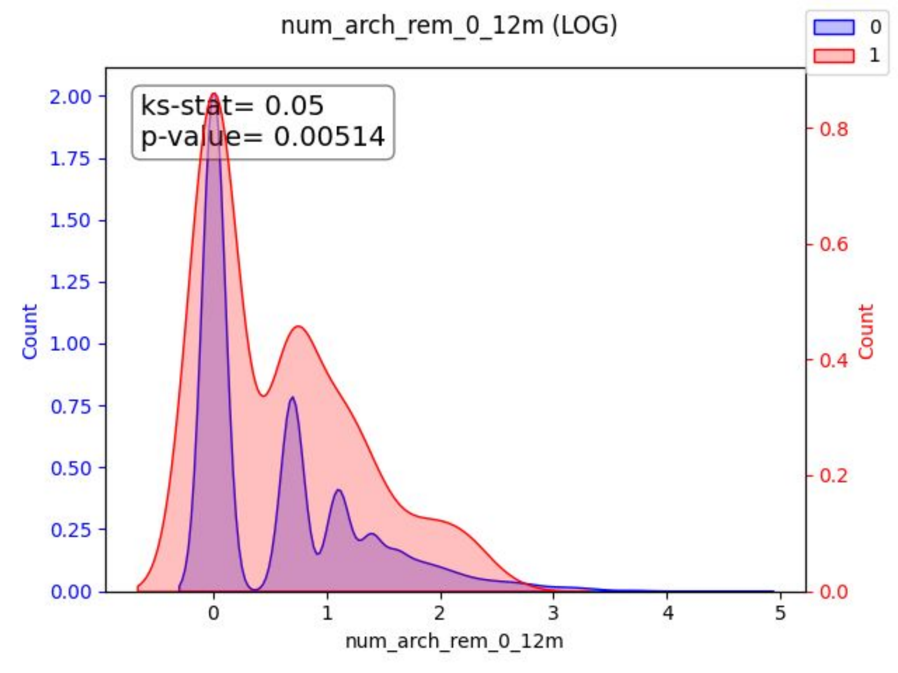


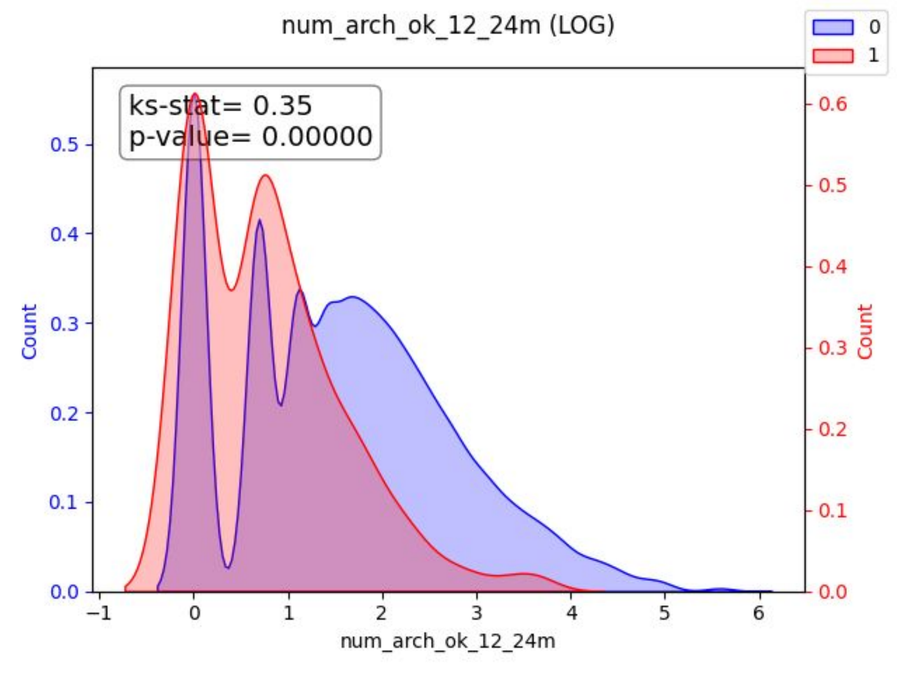
Distributions - Numerical



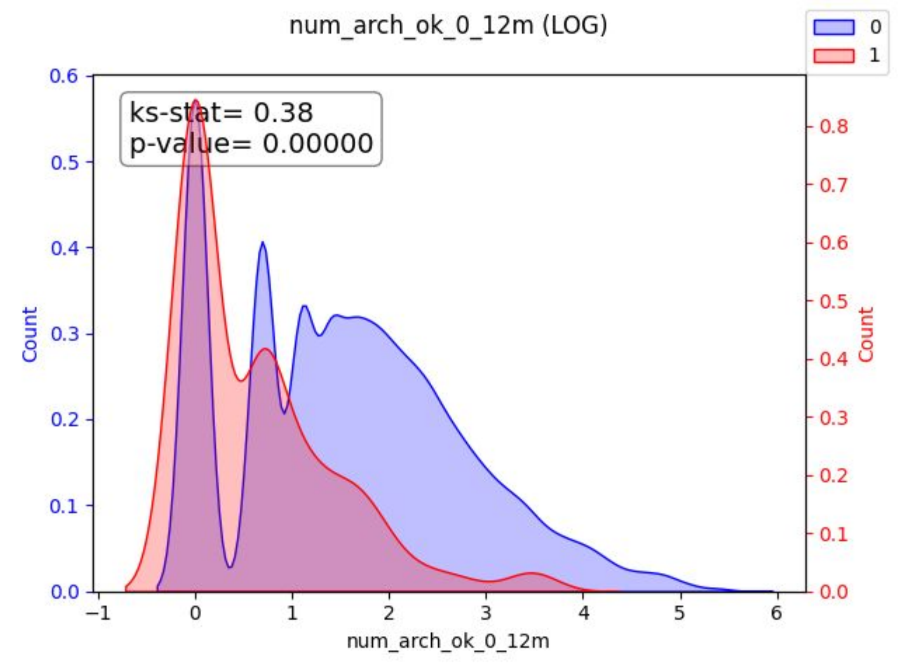


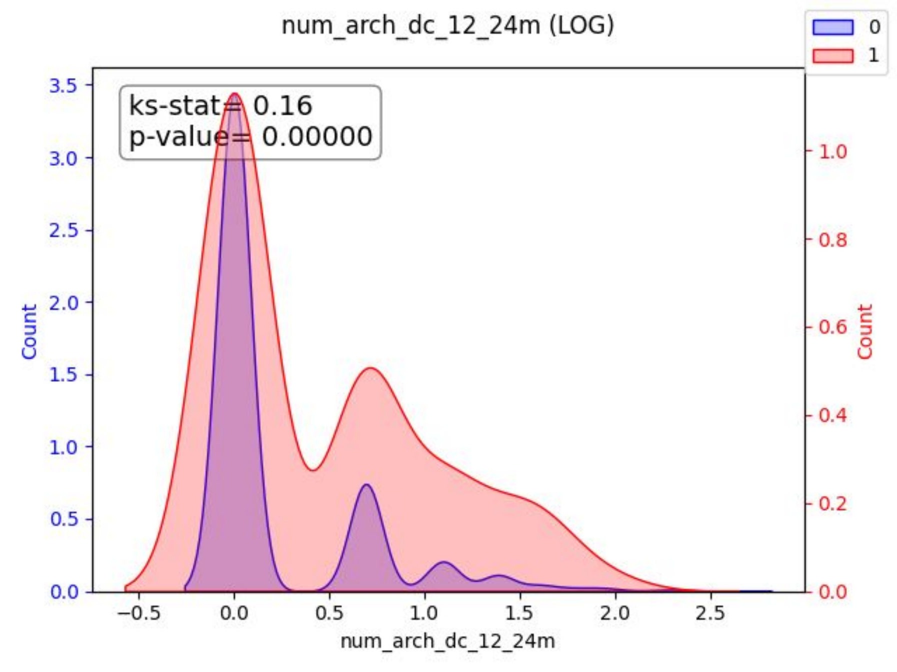
Distributions - Numerical



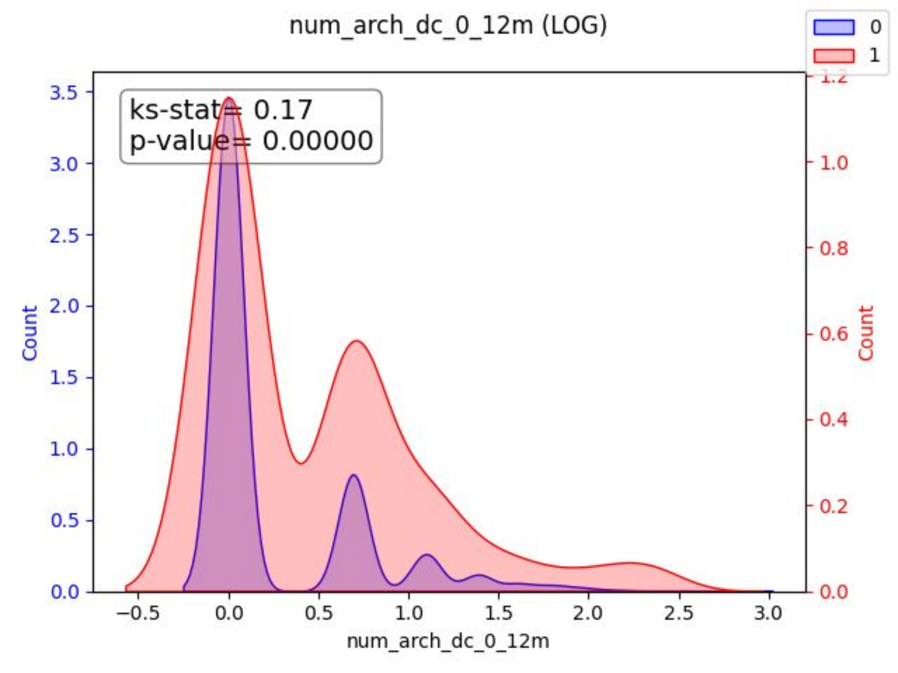


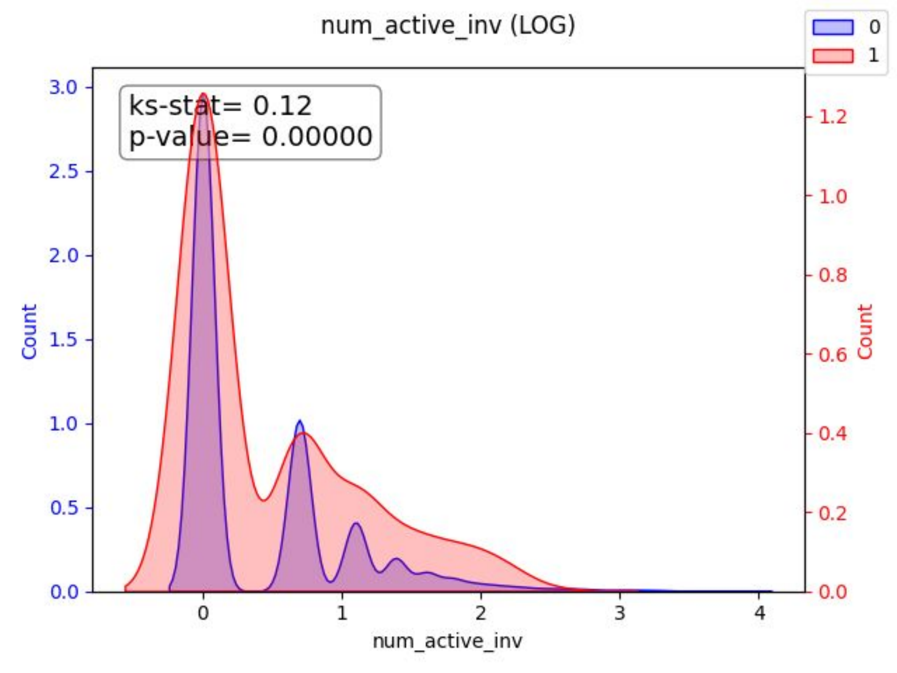
Distributions - Numerical



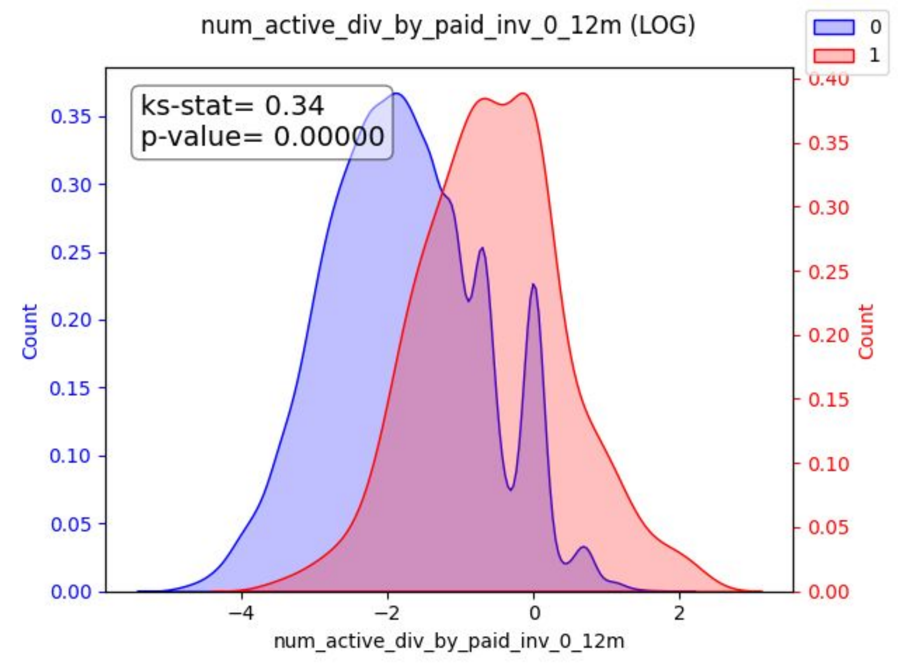


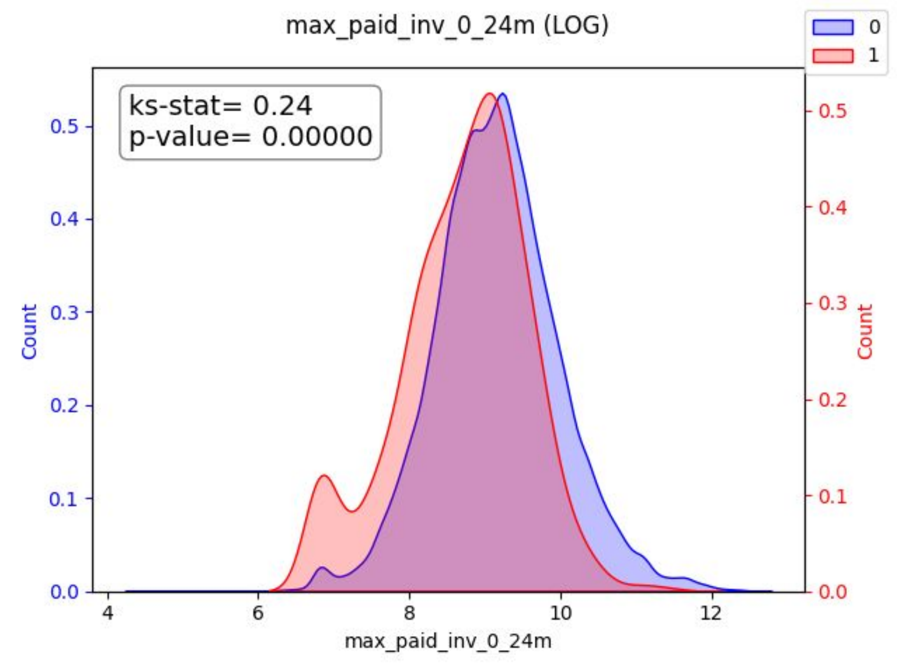
Distributions - Numerical



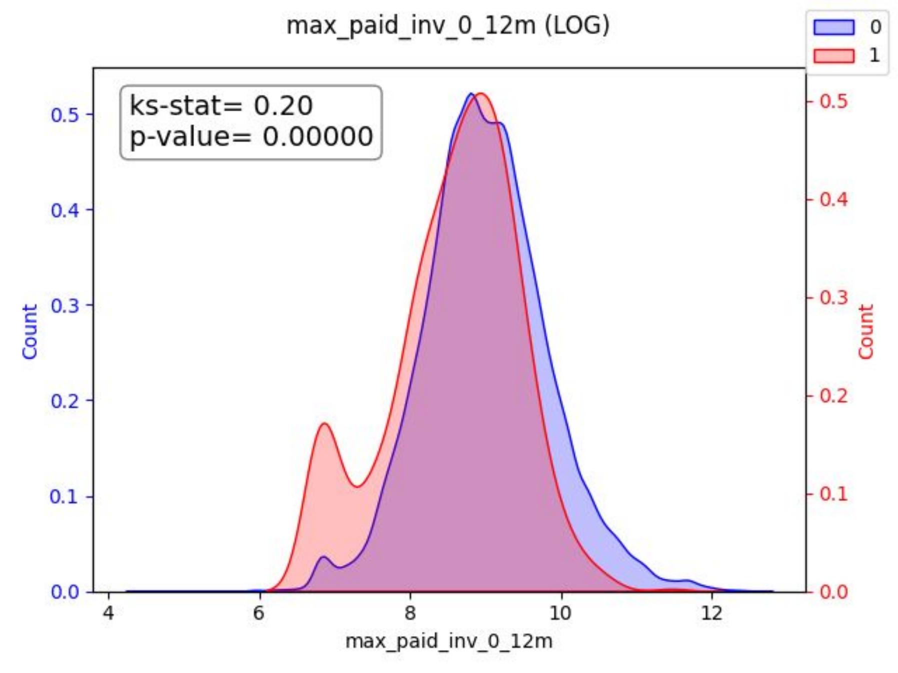


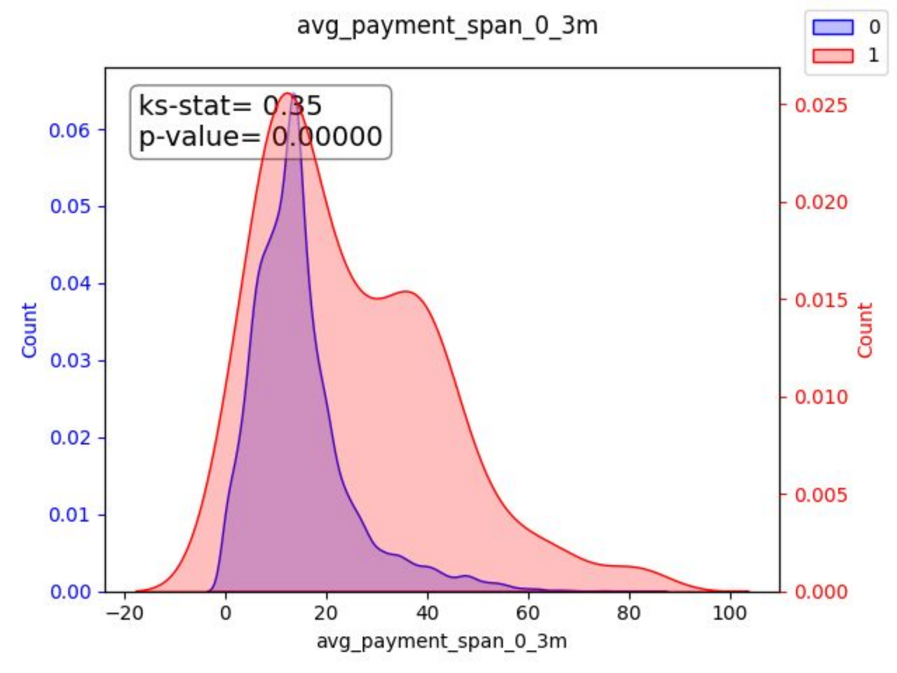
Distributions - Numerical



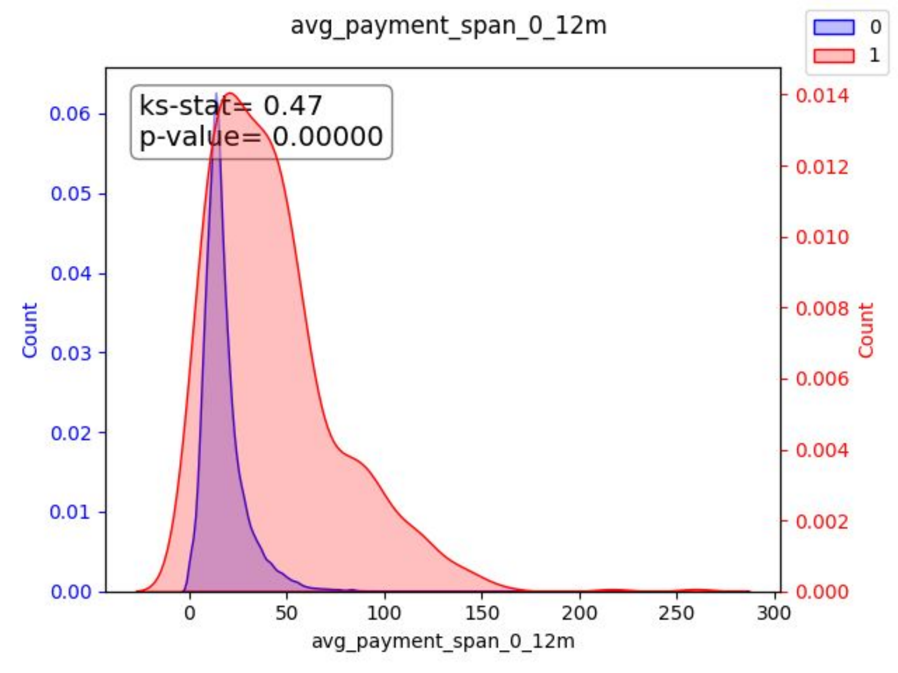


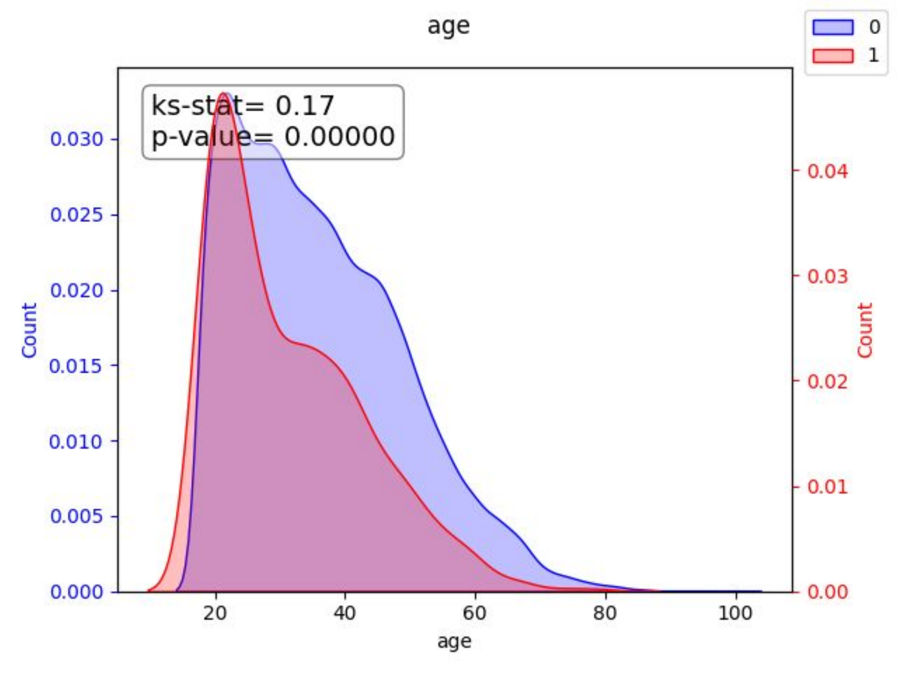
Distributions - Numerical



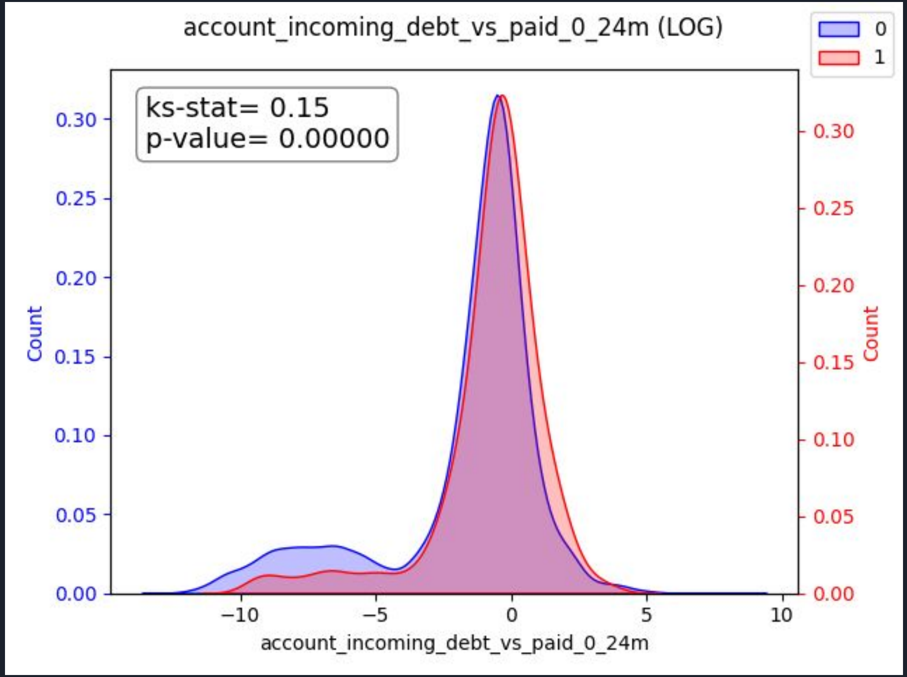


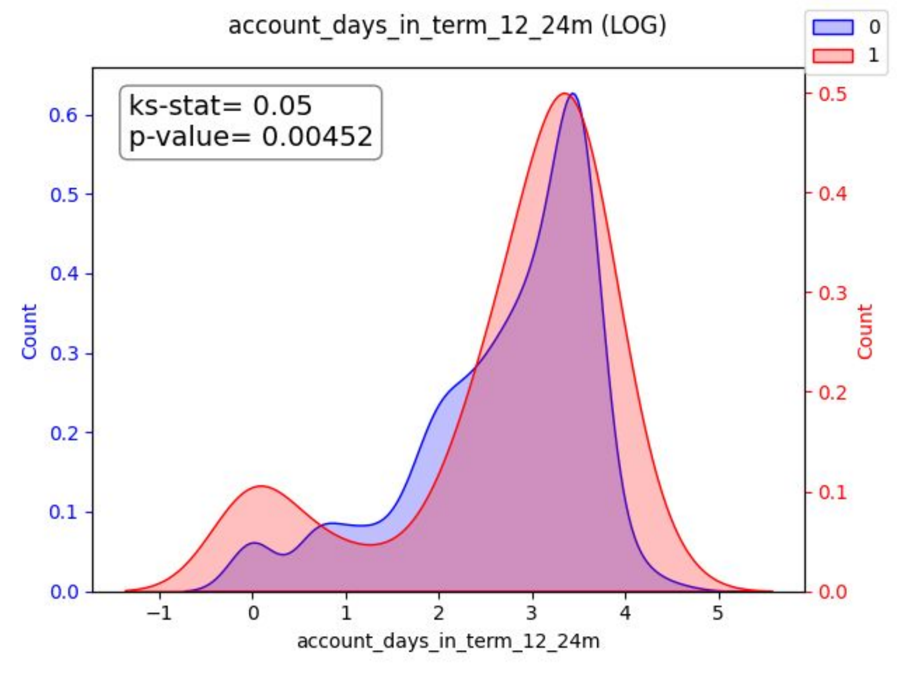
Distributions - Numerical



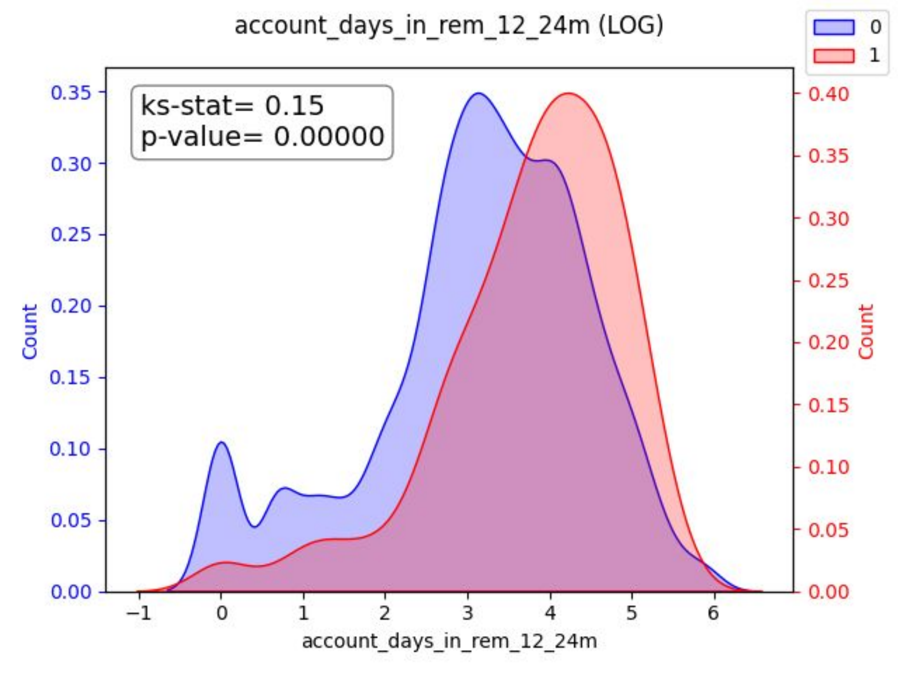


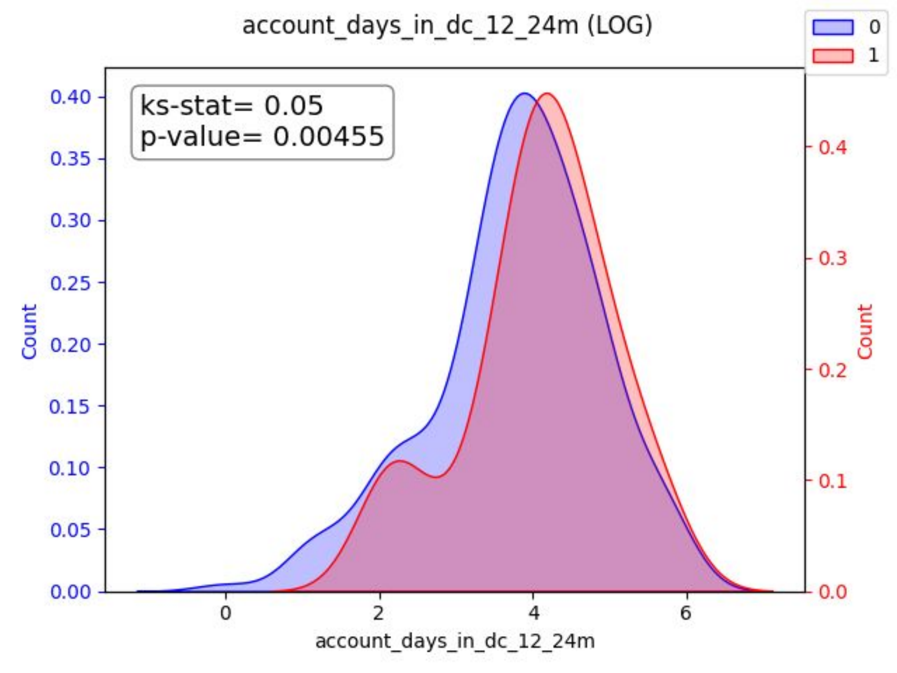
Distributions - Numerical



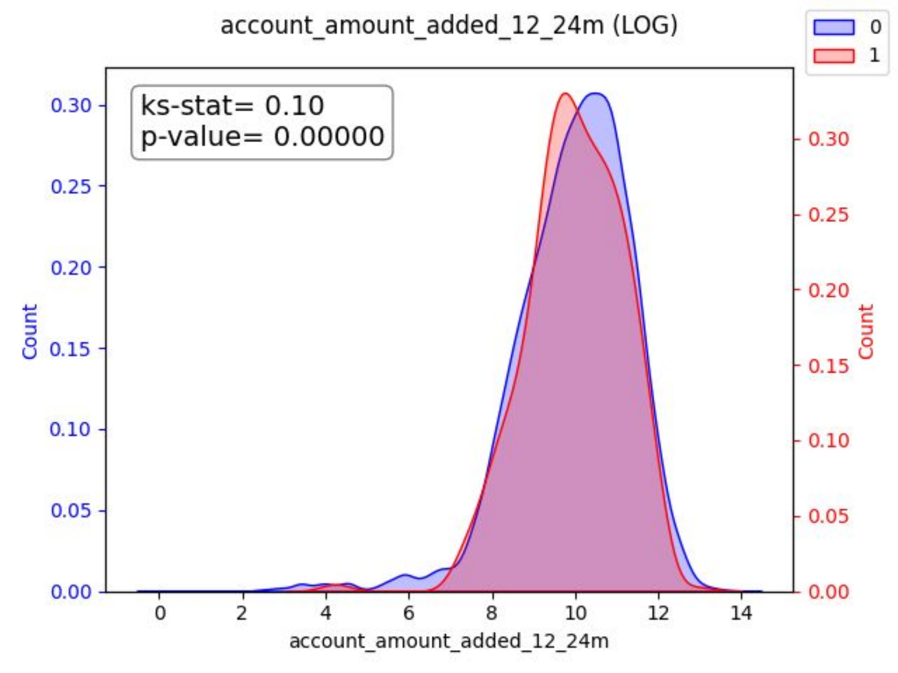


Distributions - Numerical

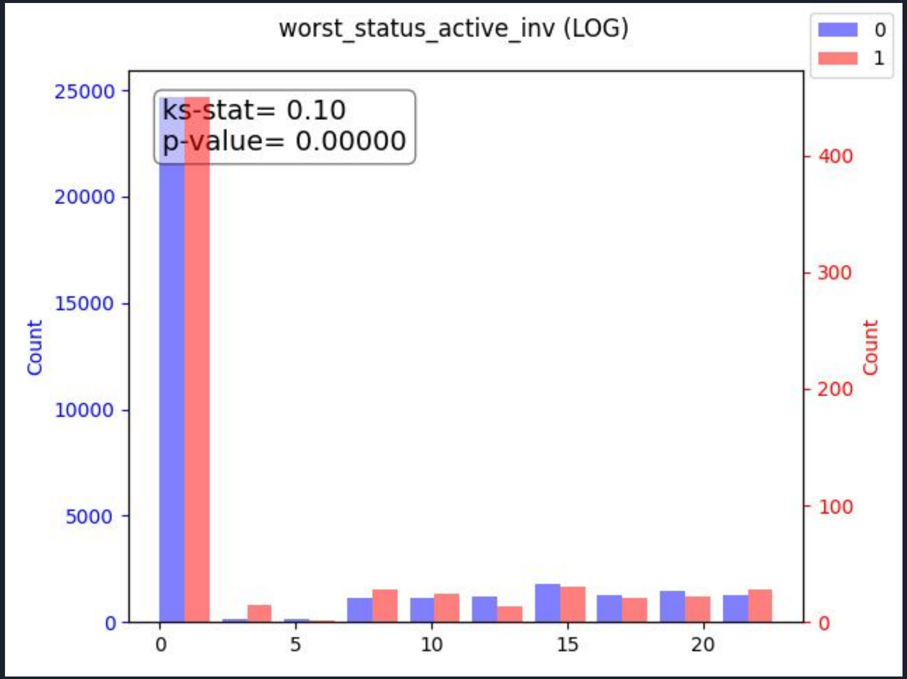


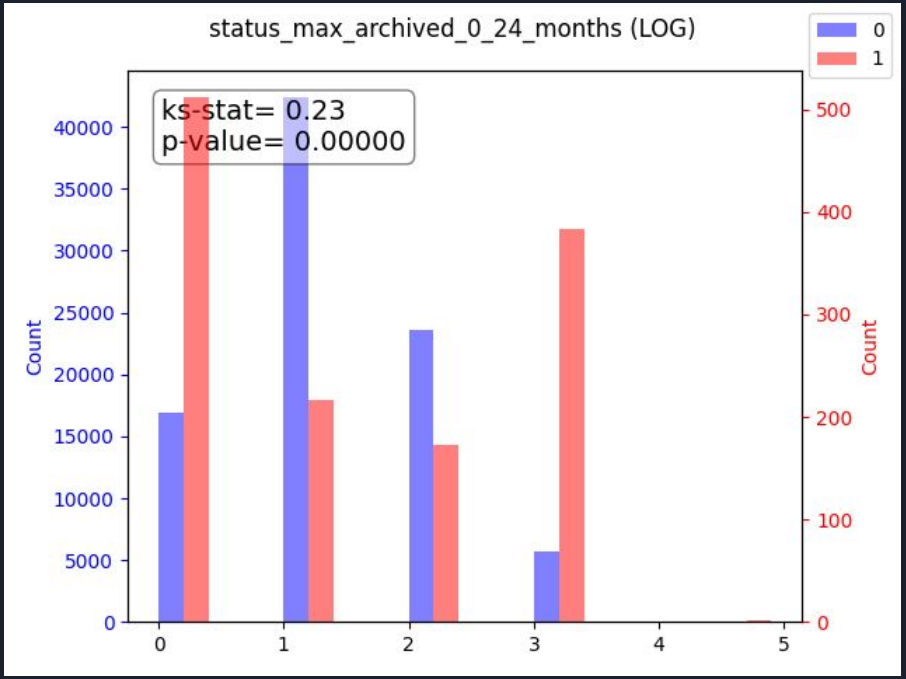


Distributions - Numerical

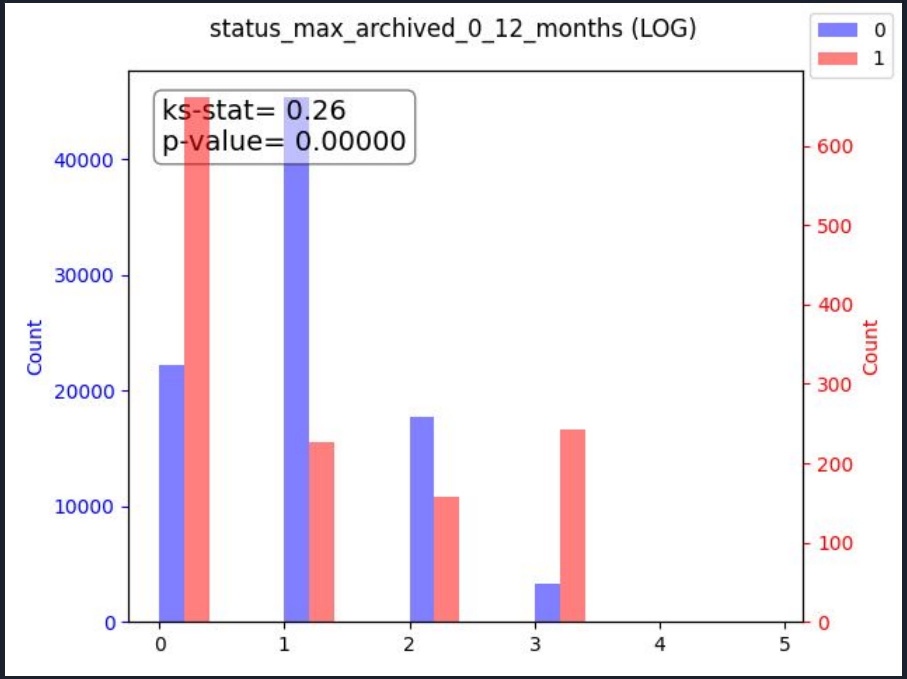


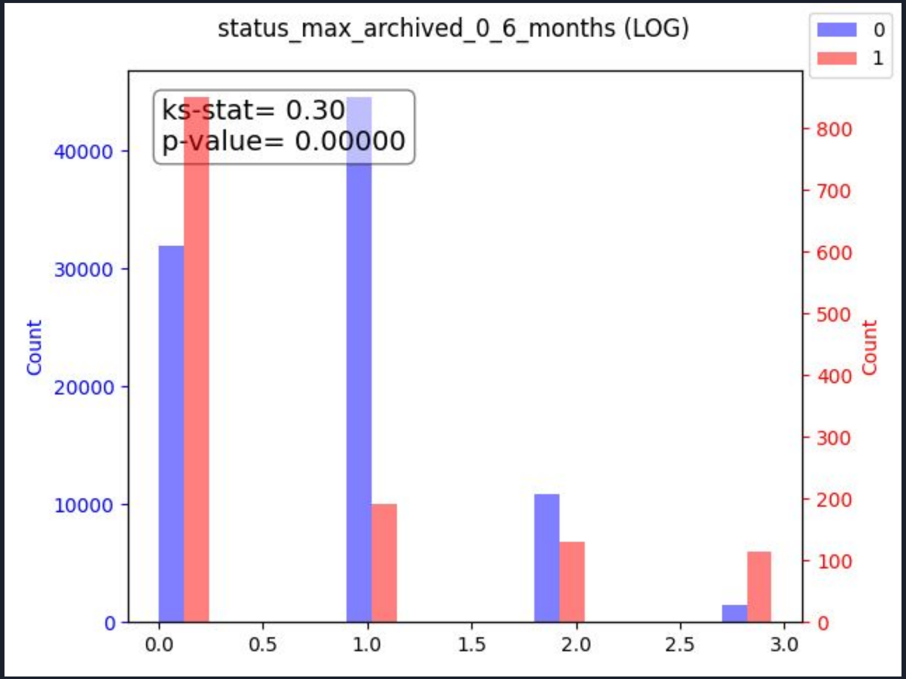
Distributions - Categorical



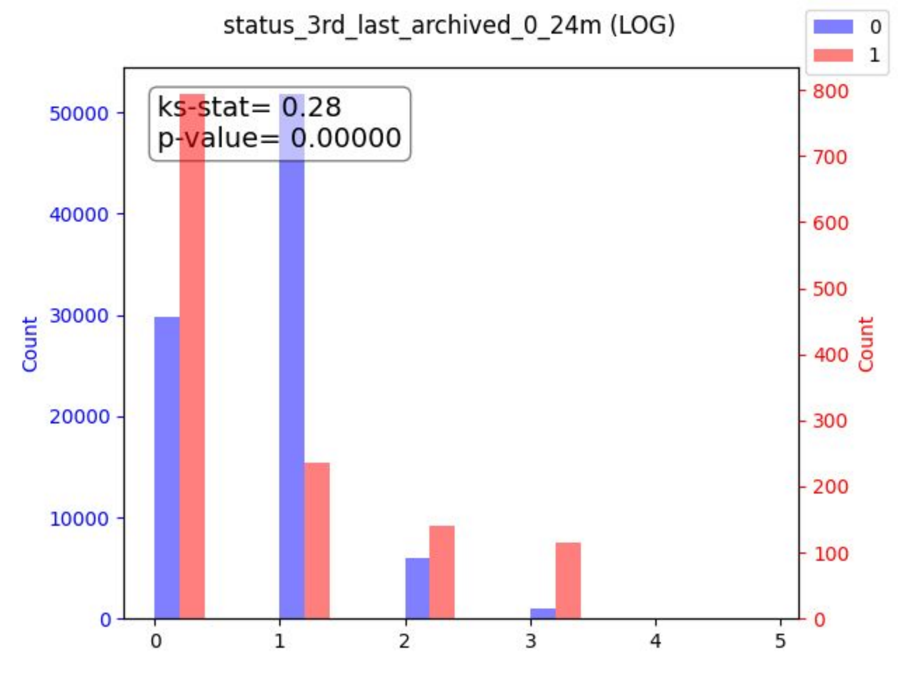


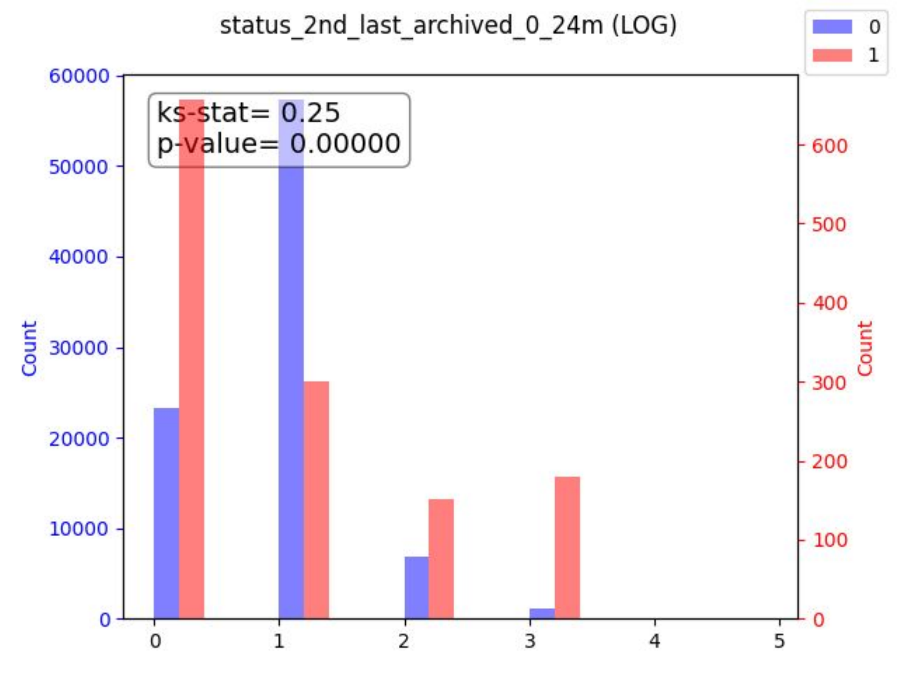
Distributions - Categorical



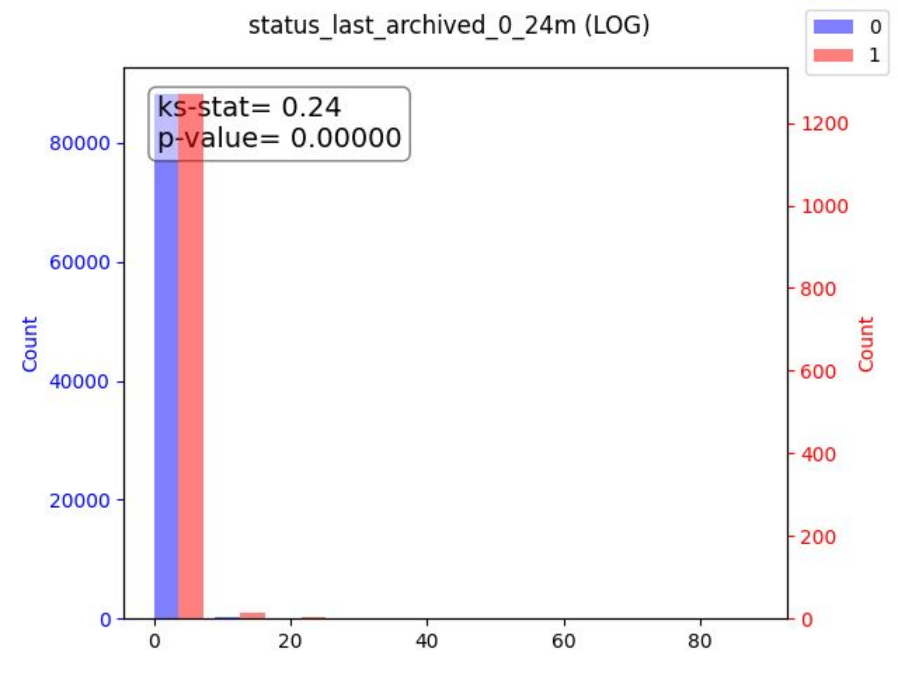


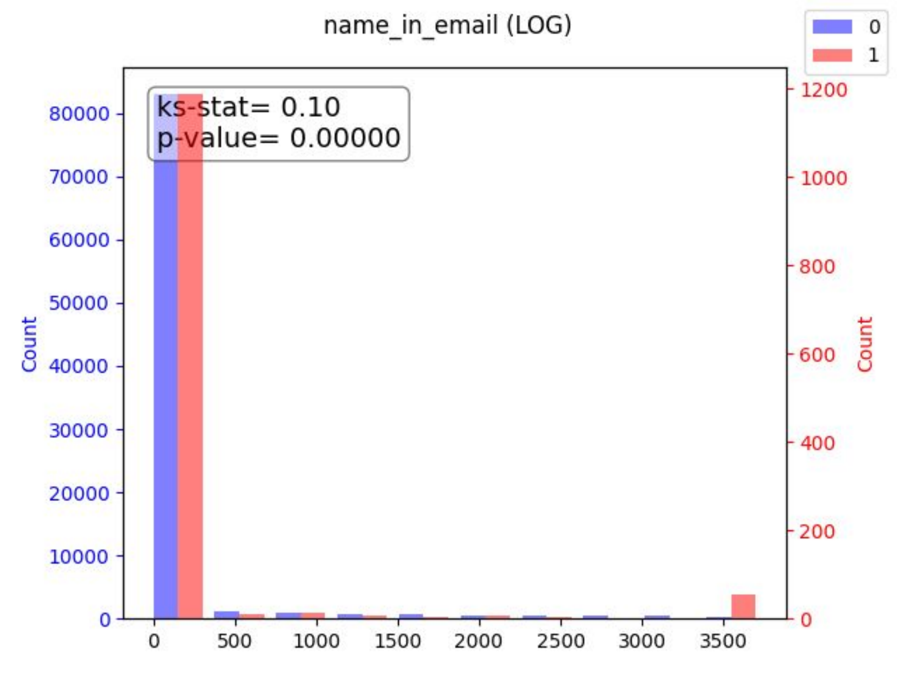
Distributions - Categorical



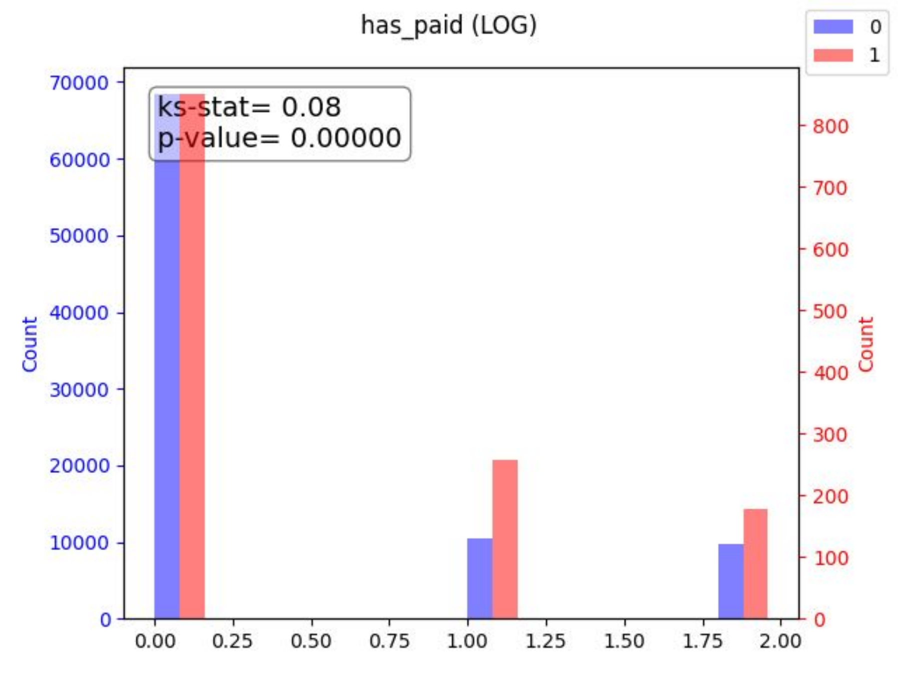


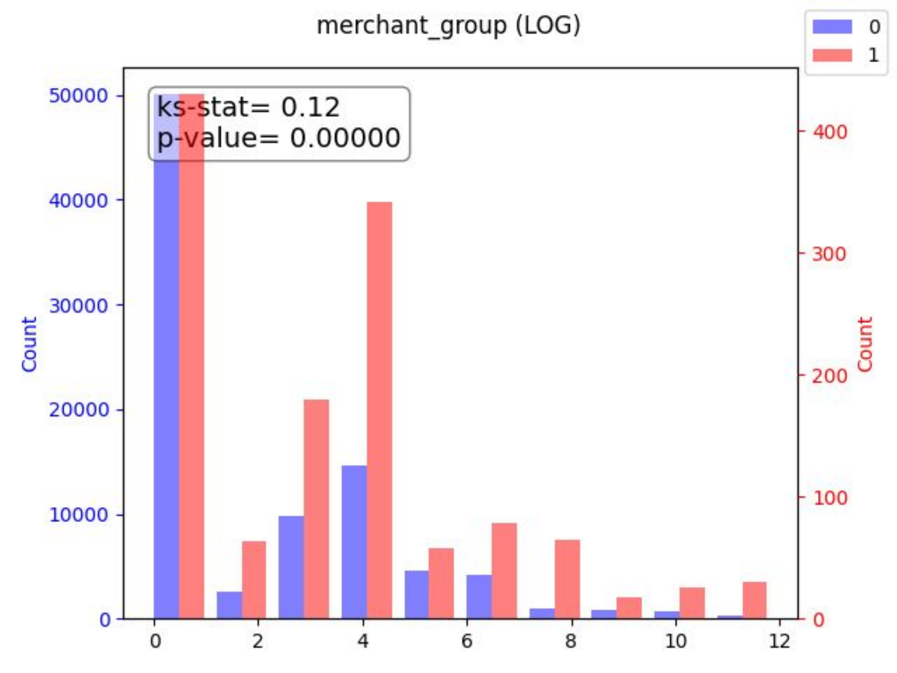
Distributions - Categorical



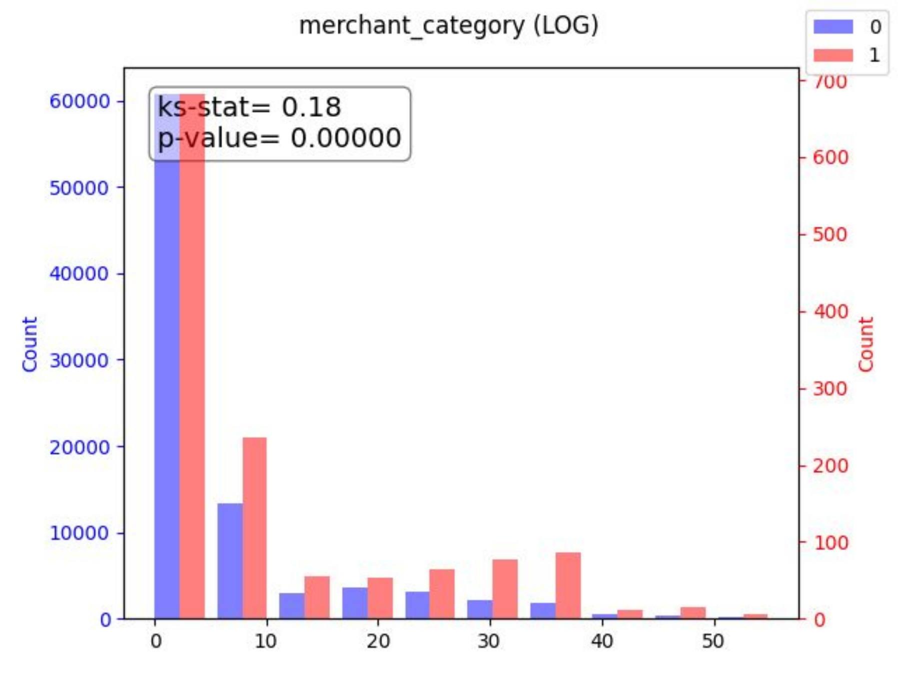


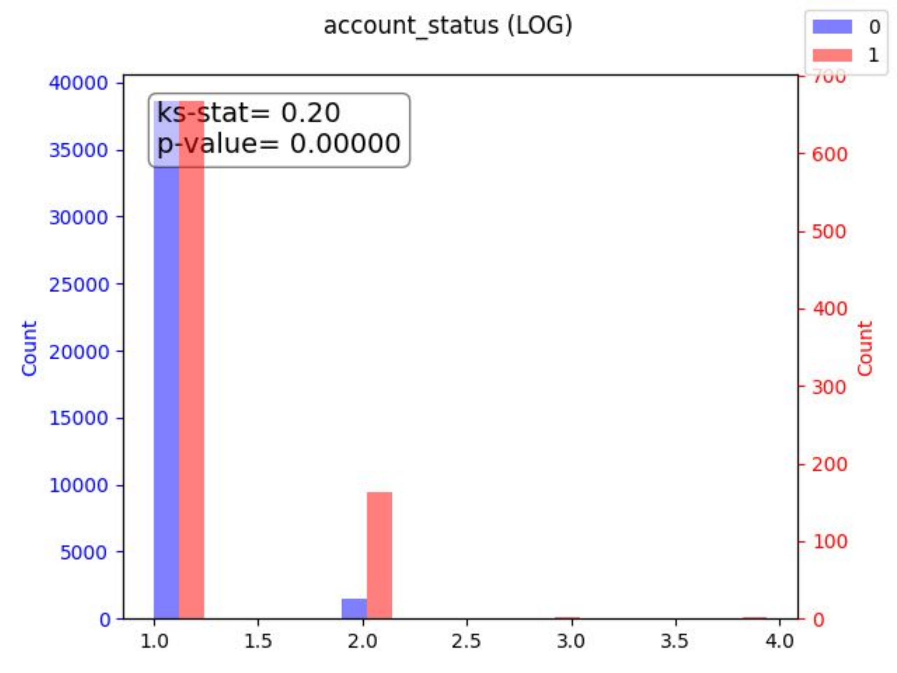
Distributions - Categorical



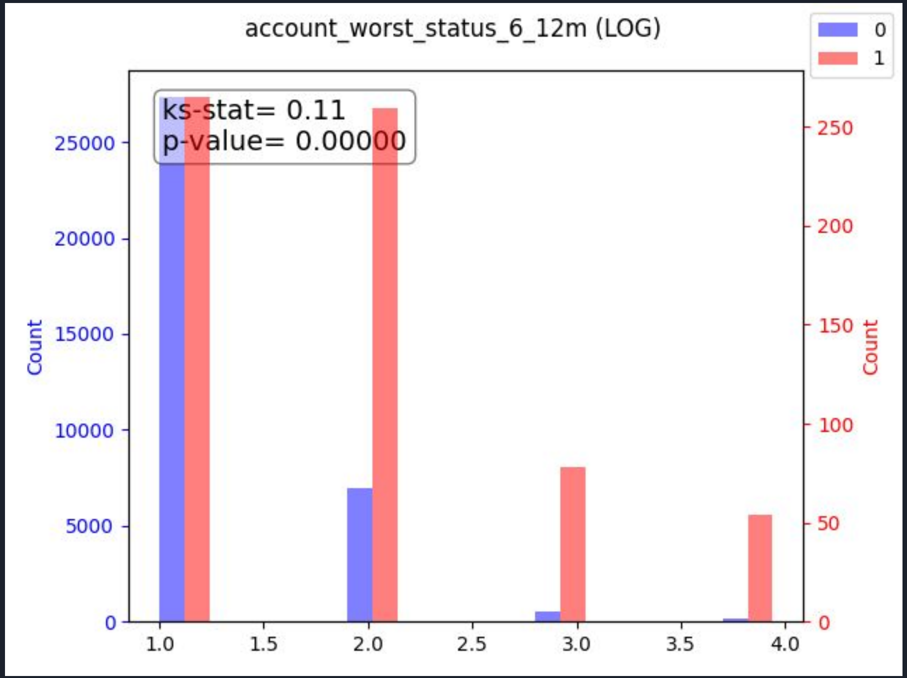


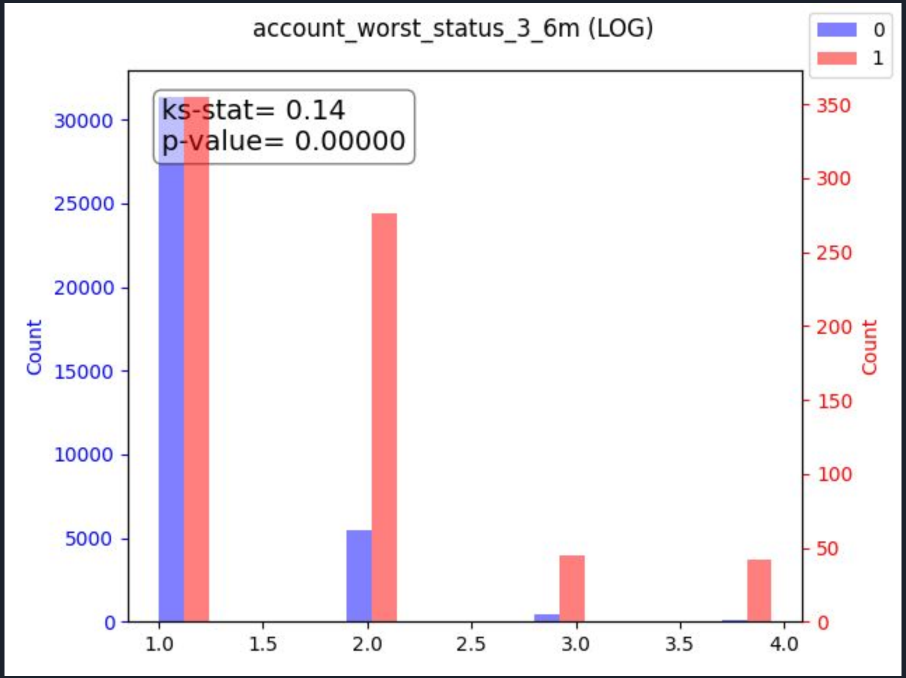
Distributions - Categorical



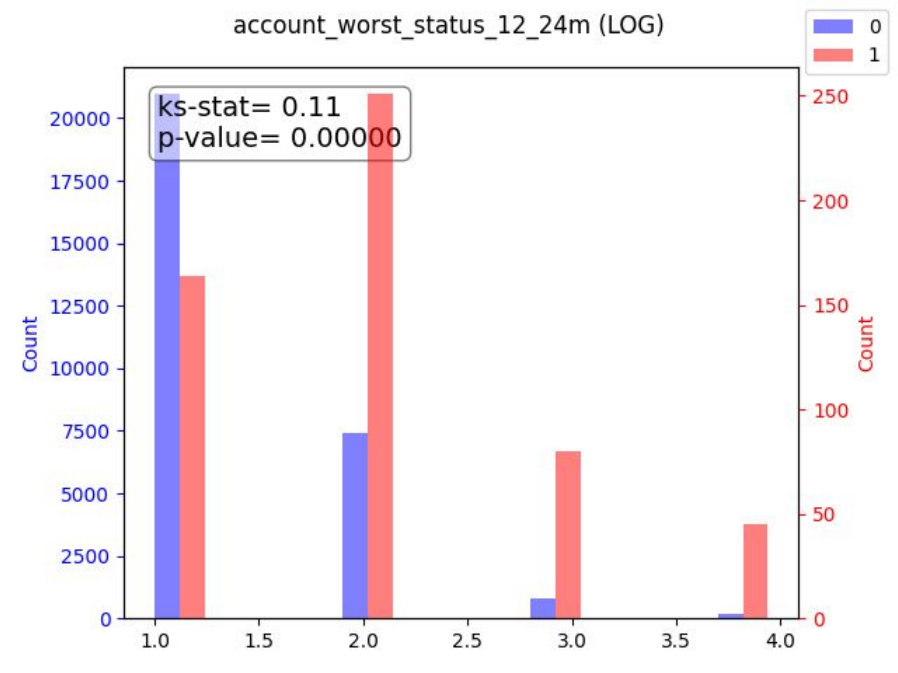


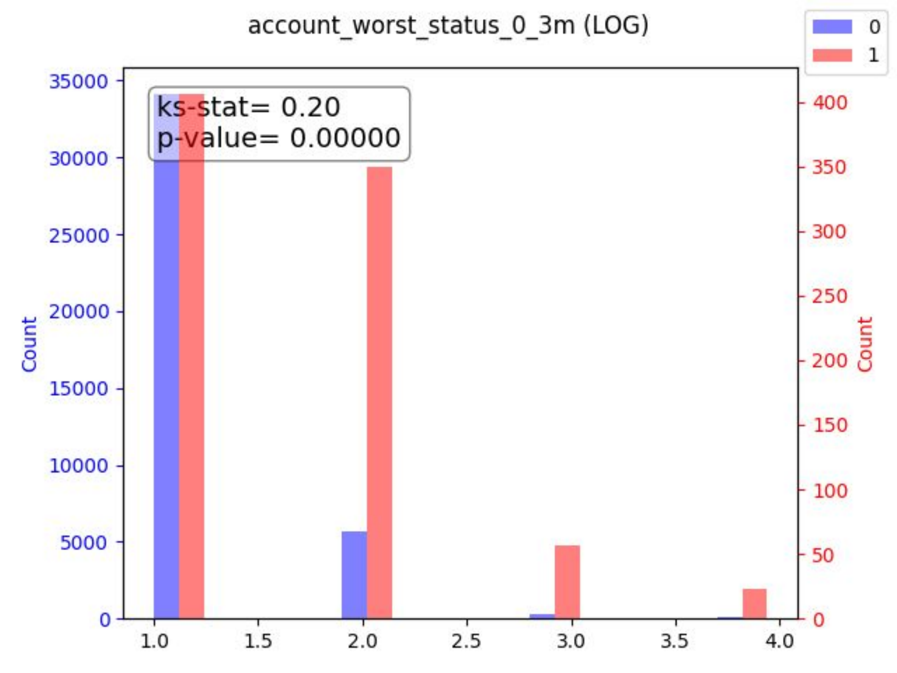
Distributions - Categorical





Distributions - Categorical


account_worst_status_12_24m (LOG)
ks-stat= 0.11
p-value= 0.00000
Count
Count


account_worst_status_0_3m (LOG)
ks-stat= 0.20
p-value= 0.00000
Count
Count