



Klarna Credit Model

Klarna Case - 02/2022

Index

1- Introduction

2- Data

3- Solutions

4- Architecture

5- Model

6- Results

Introduction



Introduction

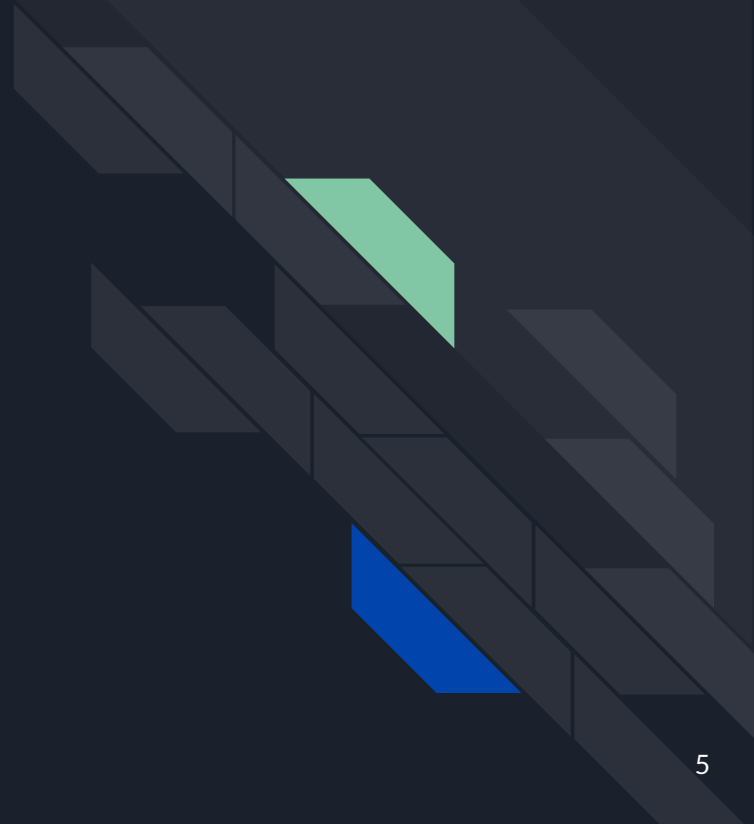
Klarna is a Swedish fintech with operations throughout Europe, US and Australia. The company offers online payment solutions for storefronts and direct clients, as well as credit products to their customers.

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



Klarna.

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 accounts
 - 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
 - 25 numerical columns
 - 16 categorical columns



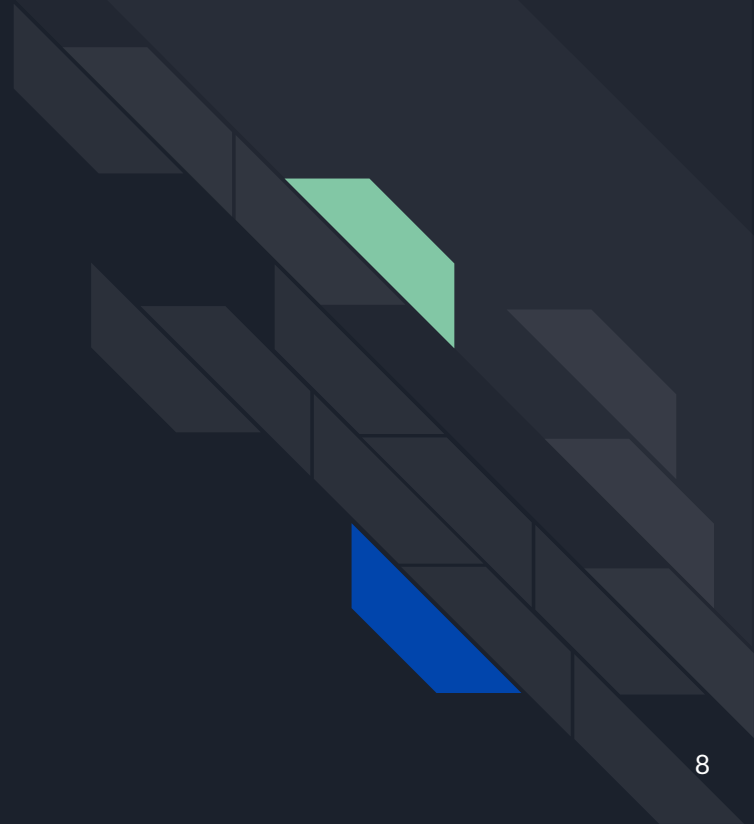
Features

- 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:

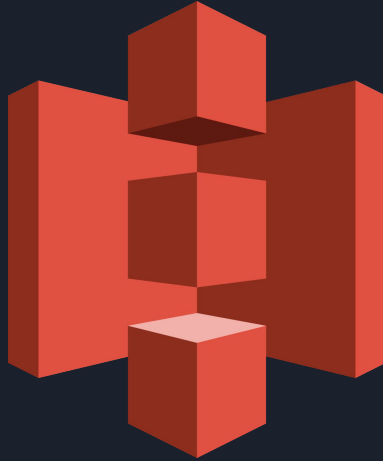
- Classification model for default probability estimation
- Airflow platform with custom operators for automating model (re)training and batch predictions using SageMaker
- KServe endpoint for online prediction
- Kubernetes environment for Airflow containers orchestration
- CI/CD pipeline built on CodePipeline 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 were stored in S3 buckets.

Online data analysis 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

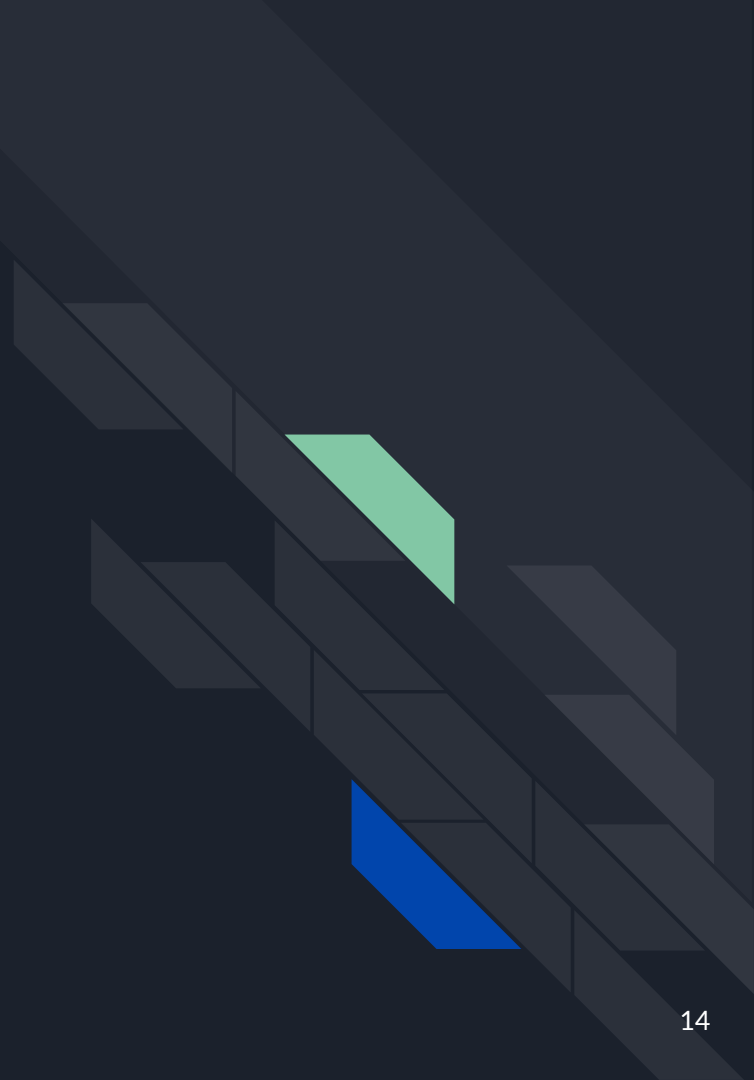


CI/CD pipeline via CodePipeline for deploying changes on the project resources upon pull requests on the project repository.

Versioning of the project resources done using a GitHub repository

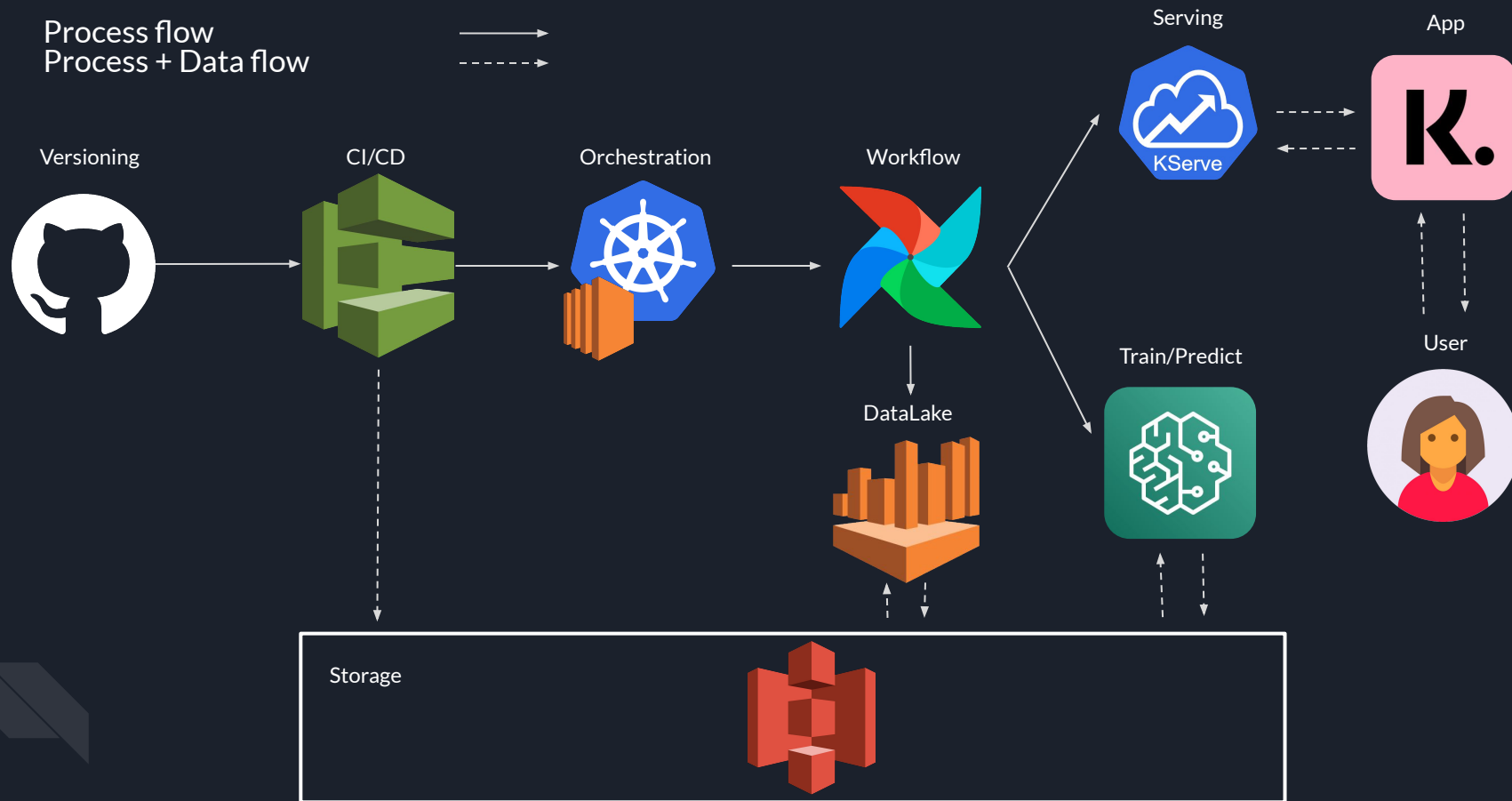


Architecture

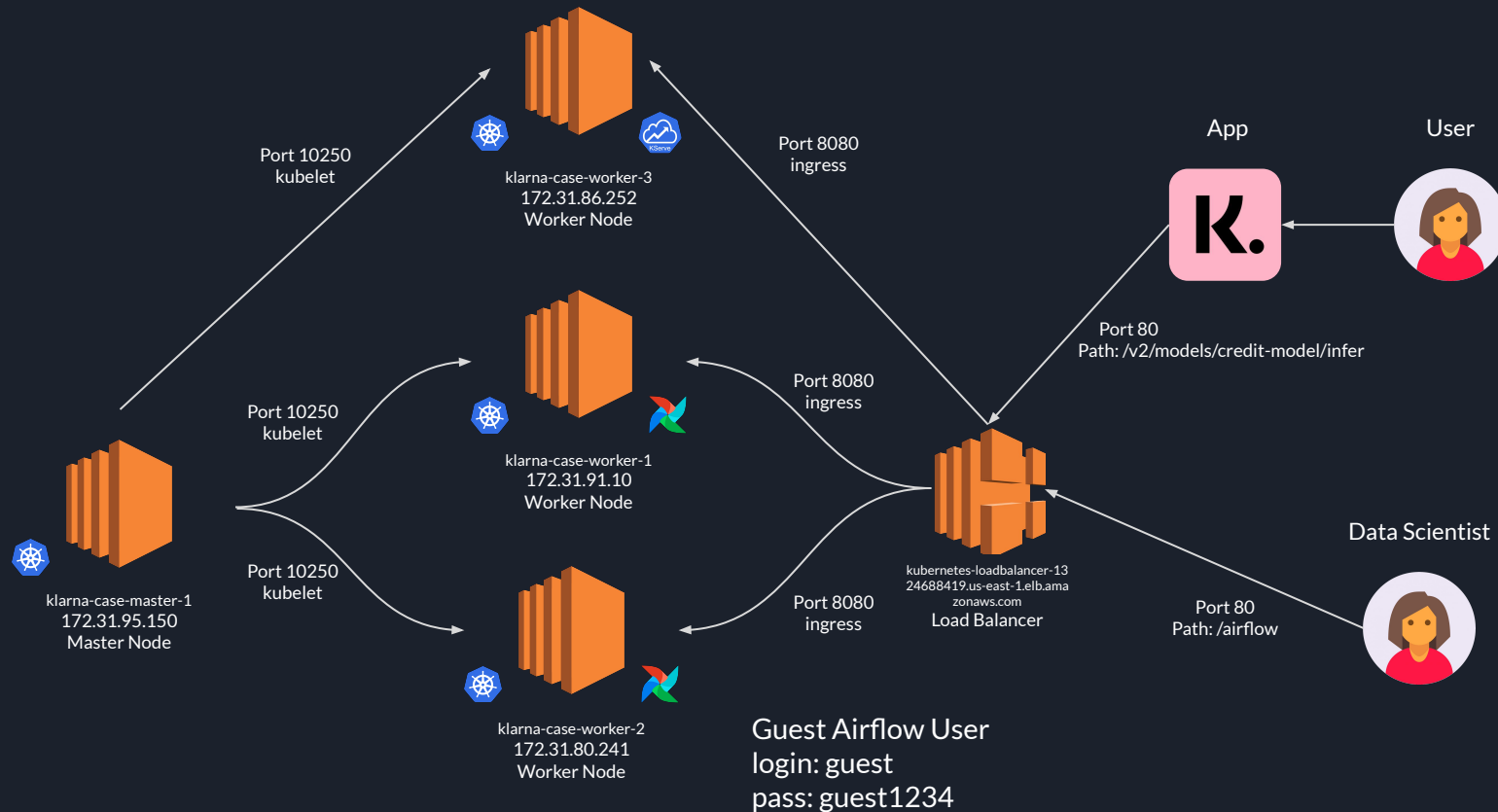


Workflow Architecture

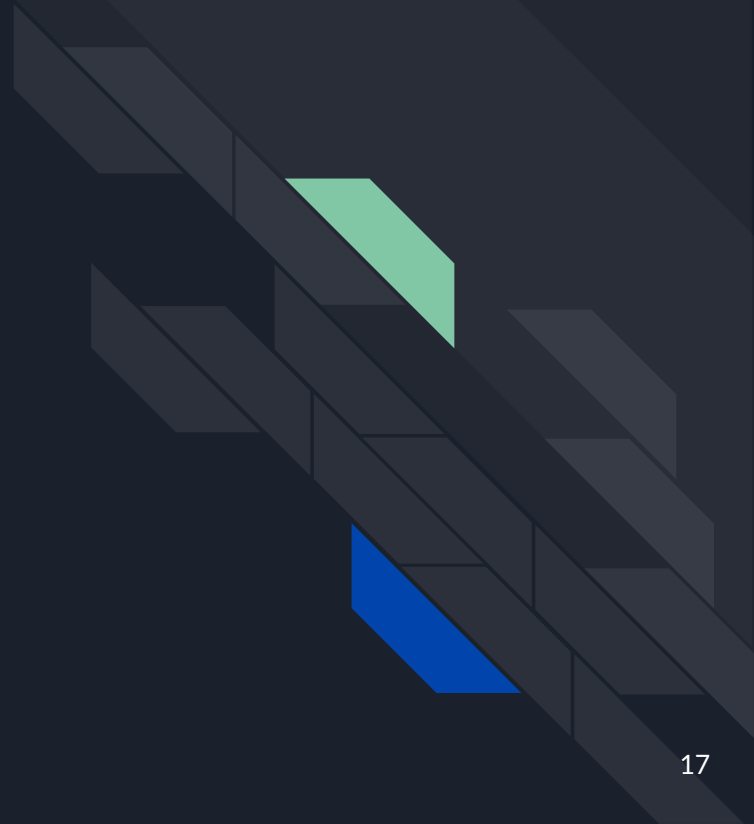
Process flow
Process + Data flow



Systems Architecture



Model





Model

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

Challenger models:

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

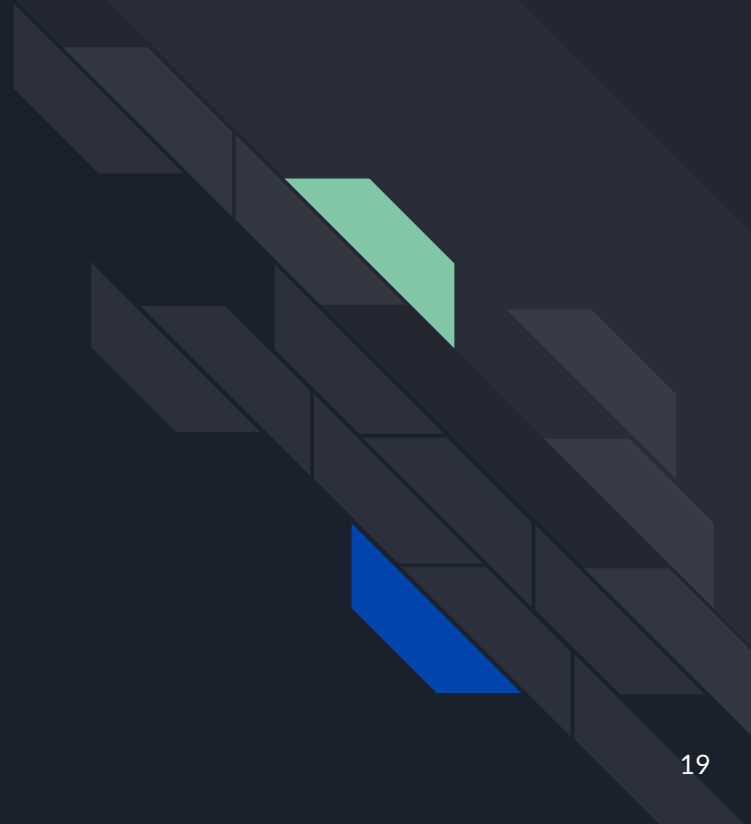
Train-test split: 70/30

Hyperparameters tuning performed in all models

Final model calibrated using Platt calibration

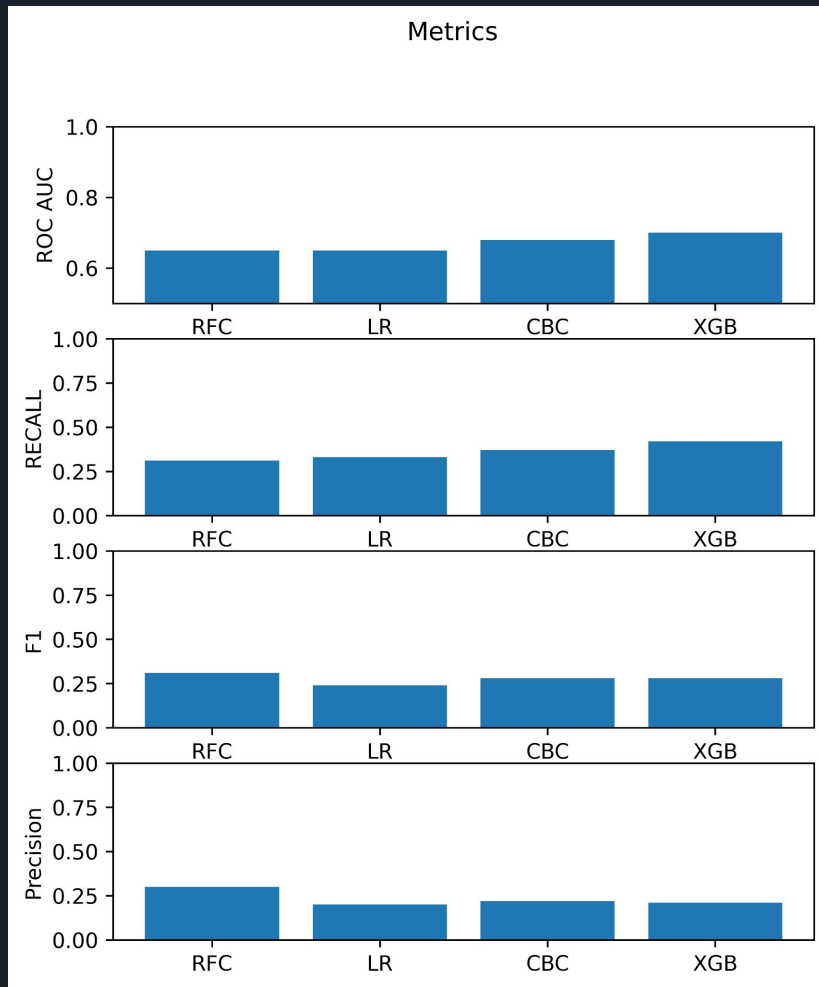
dmlc
XGBoost

Results

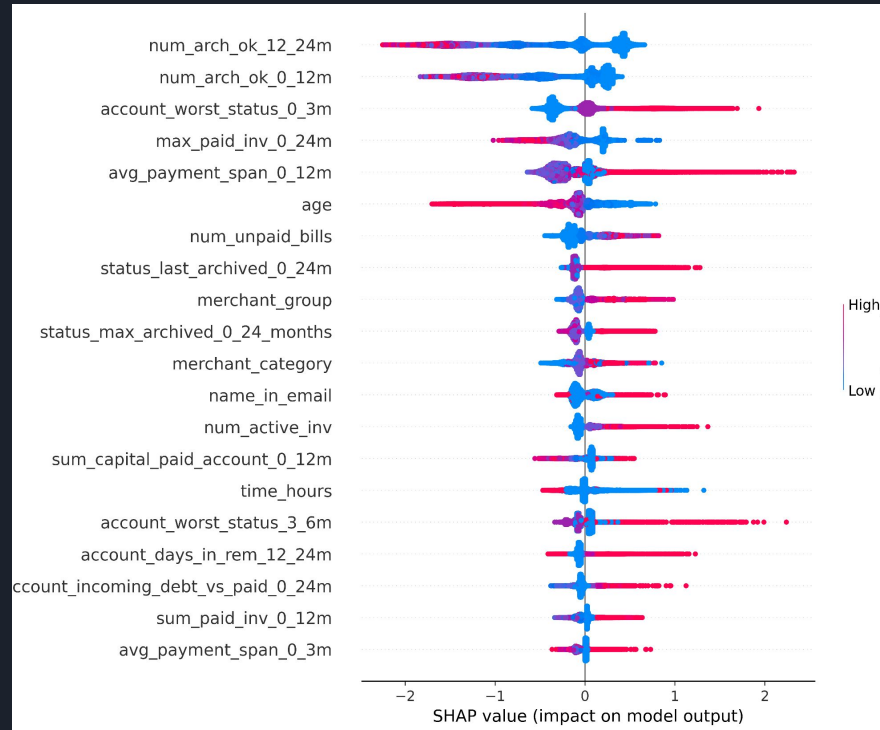
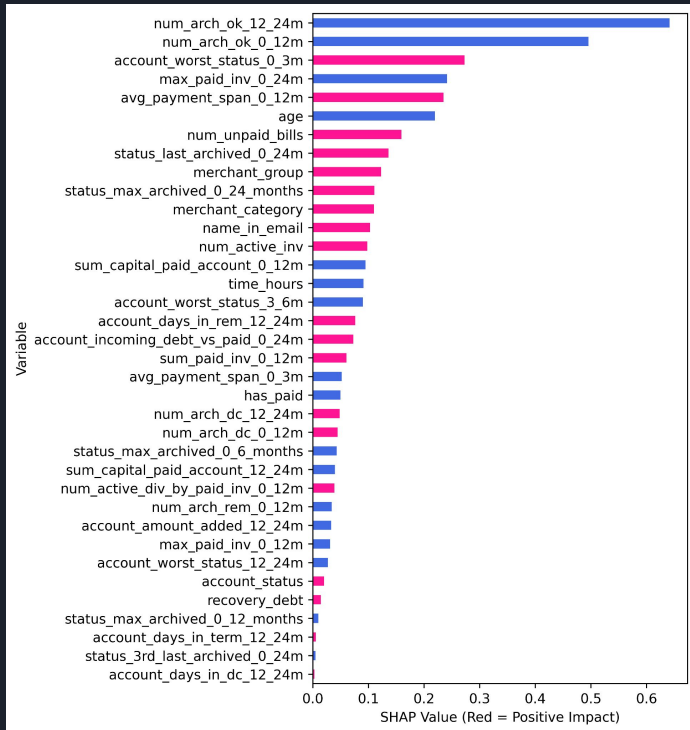


Metrics

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



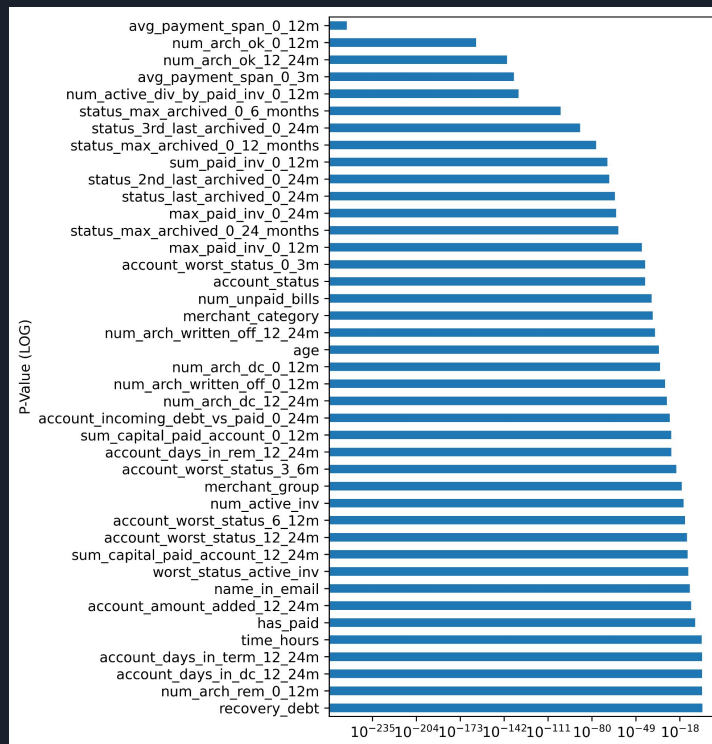
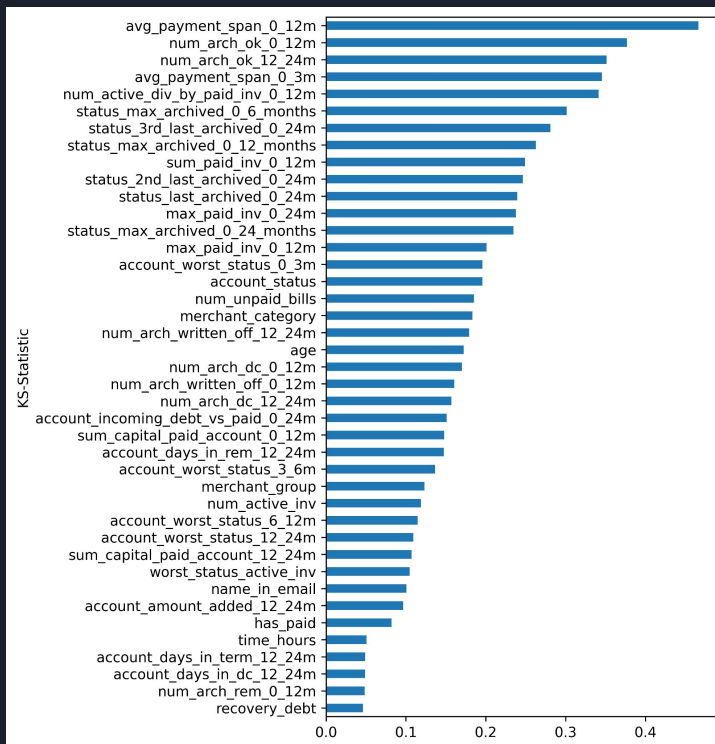
Feature Importance



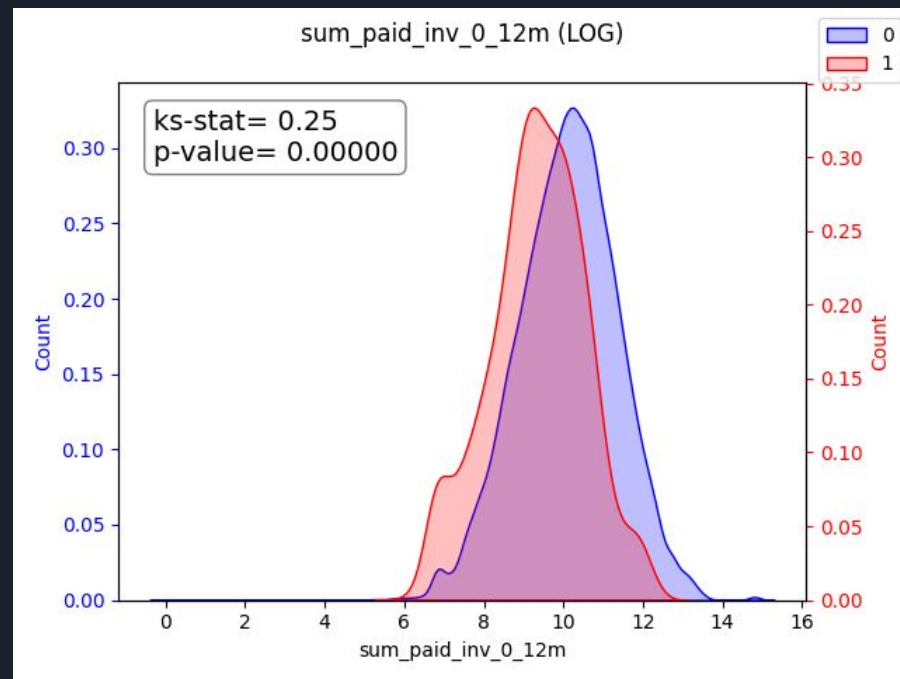
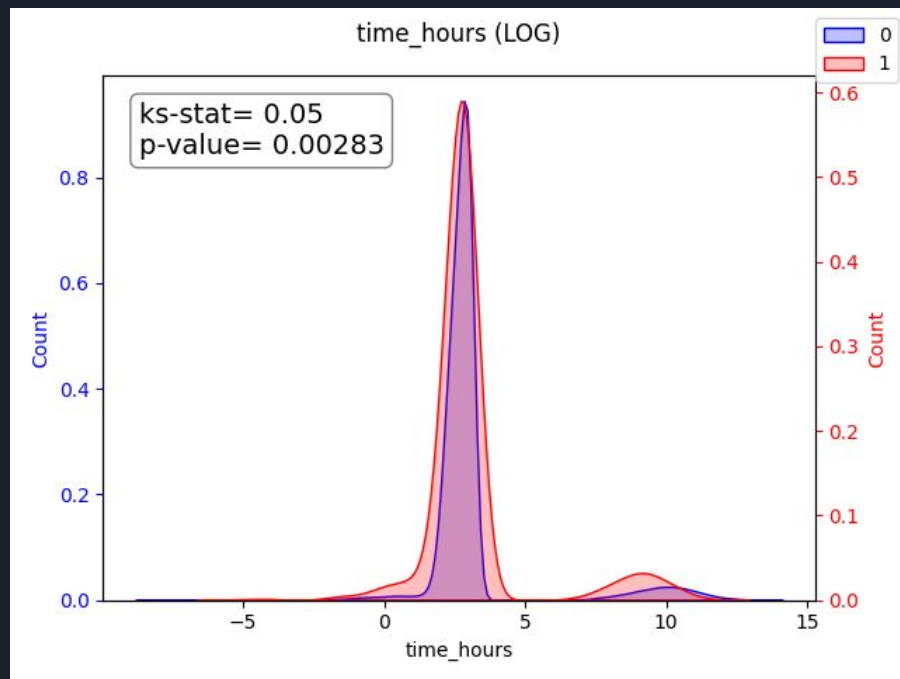
Thanks!

Appendix

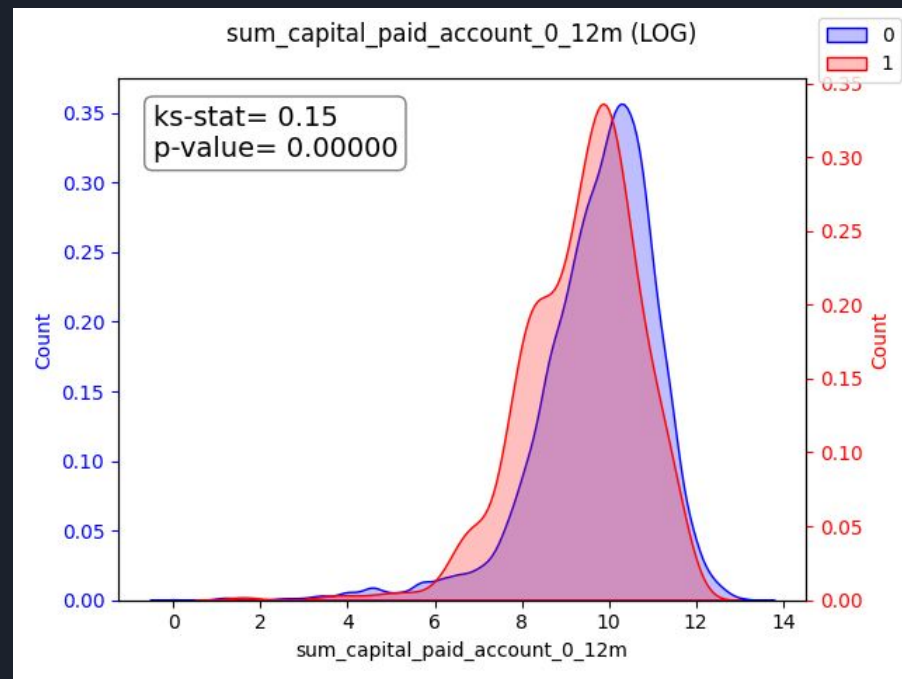
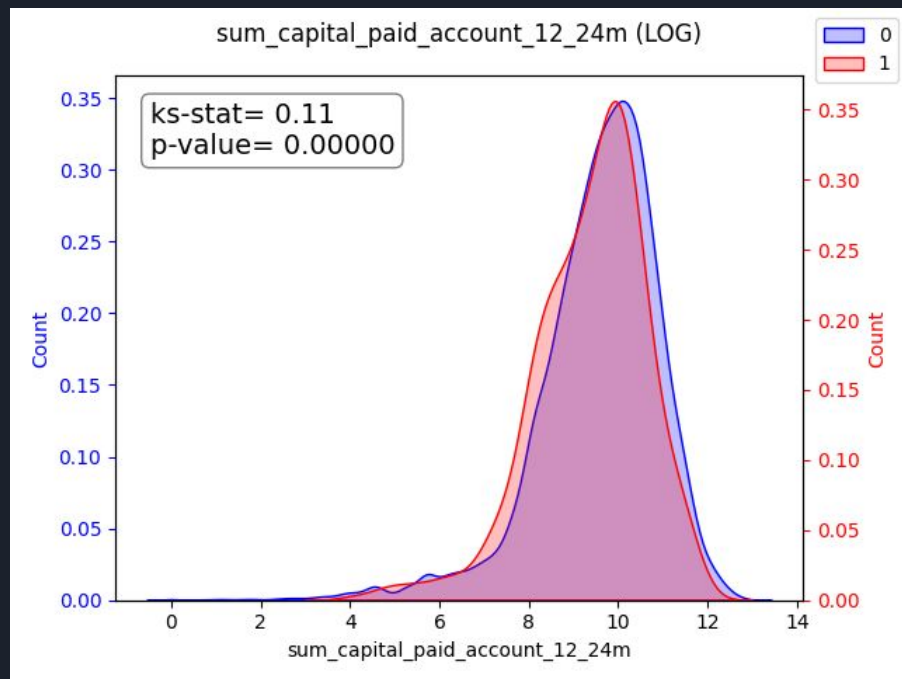
KS-Test



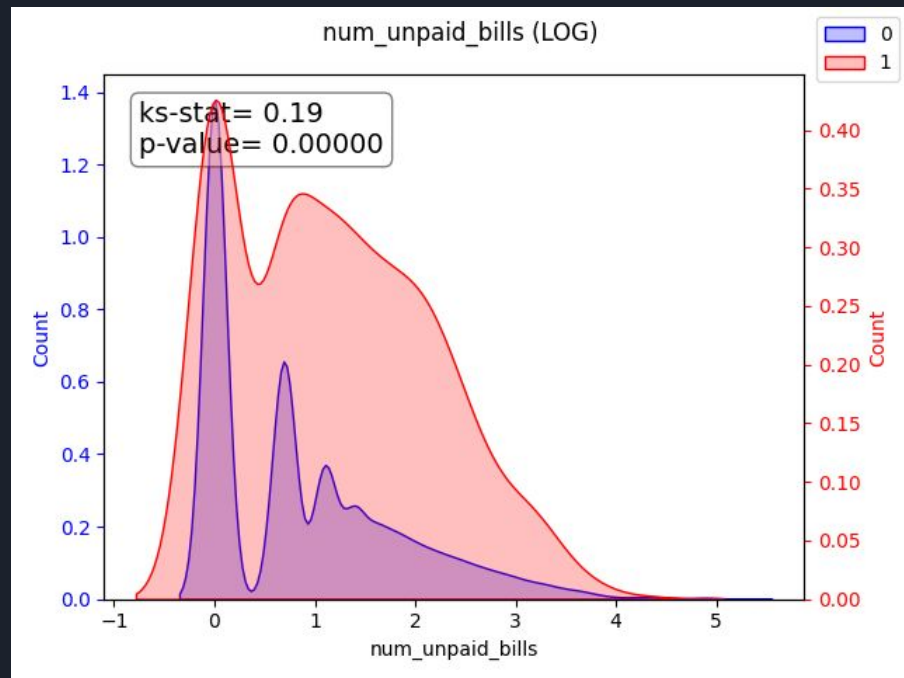
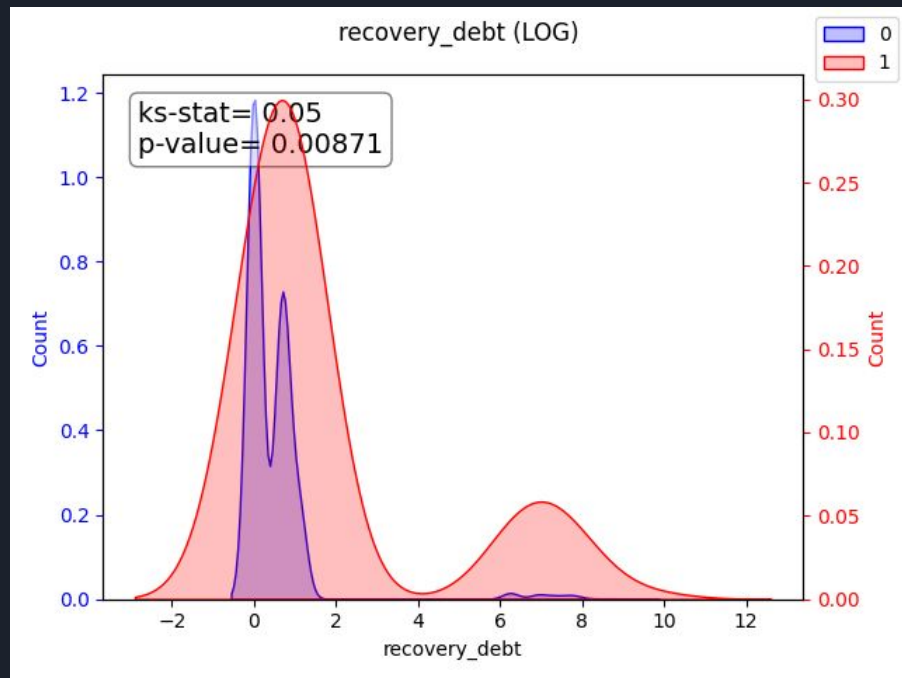
Distributions - Numerical



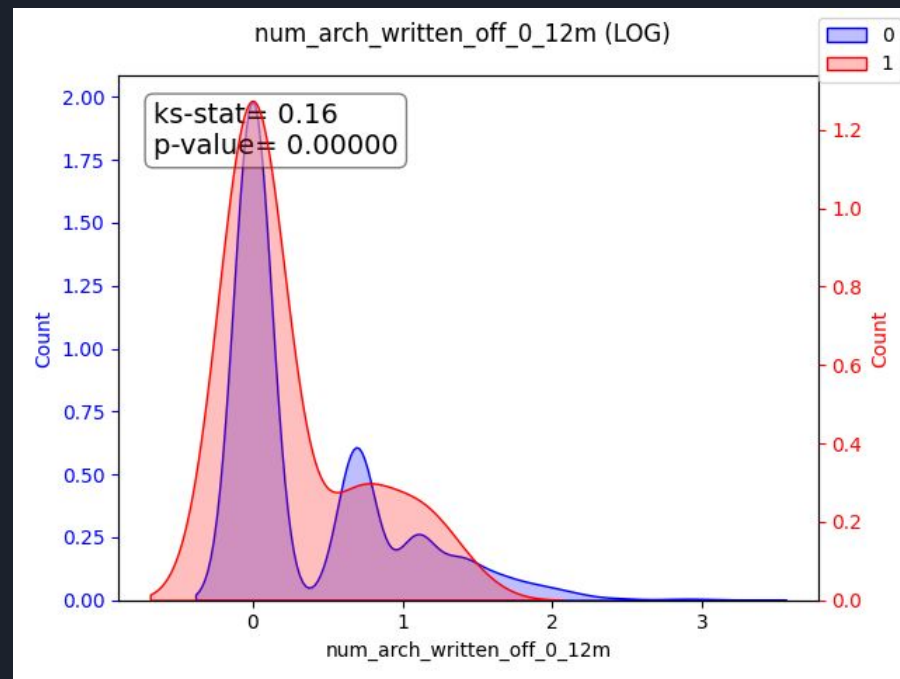
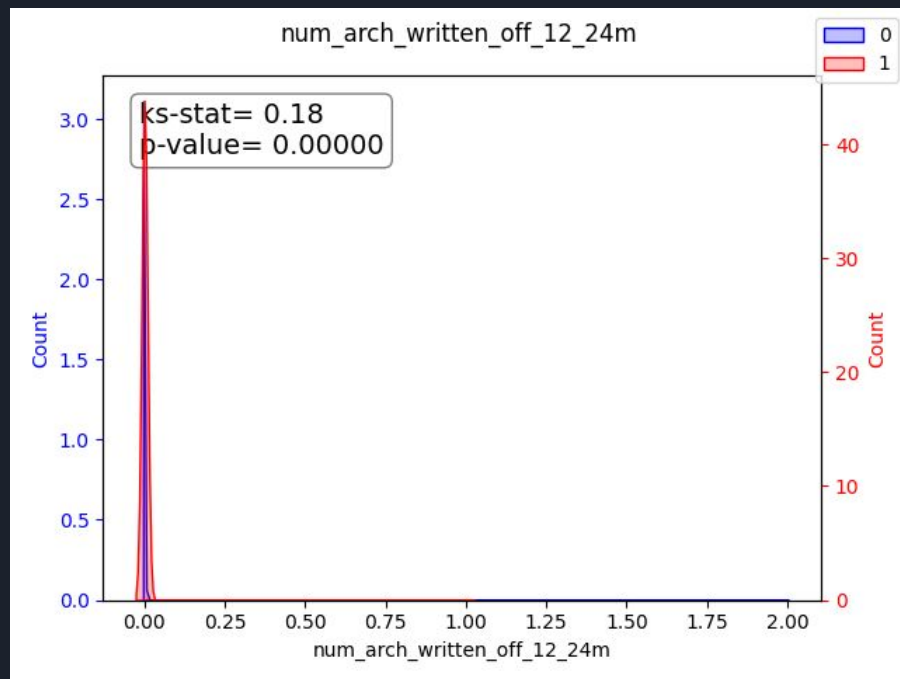
Distributions - Numerical



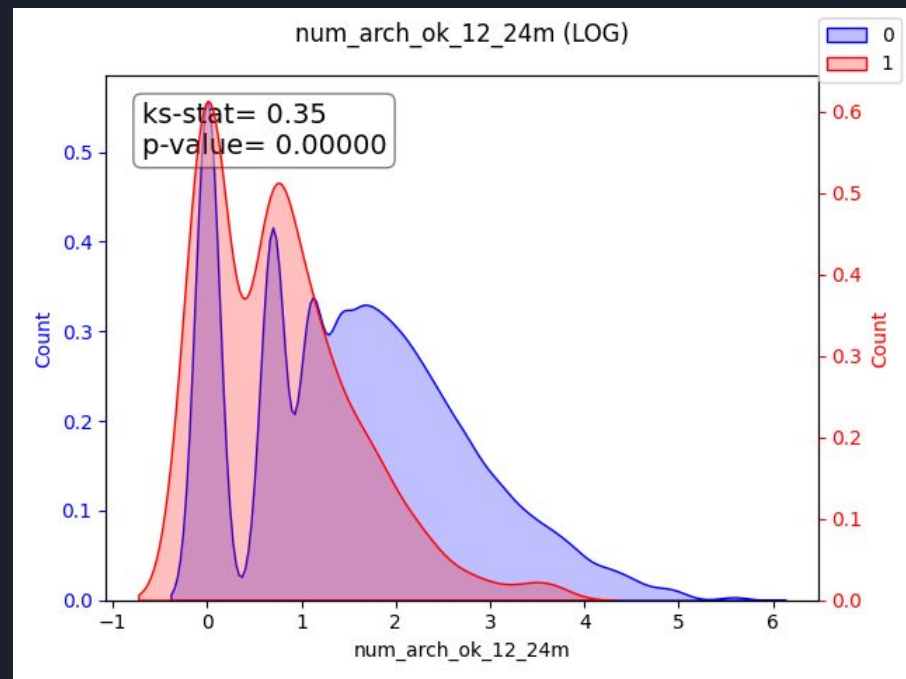
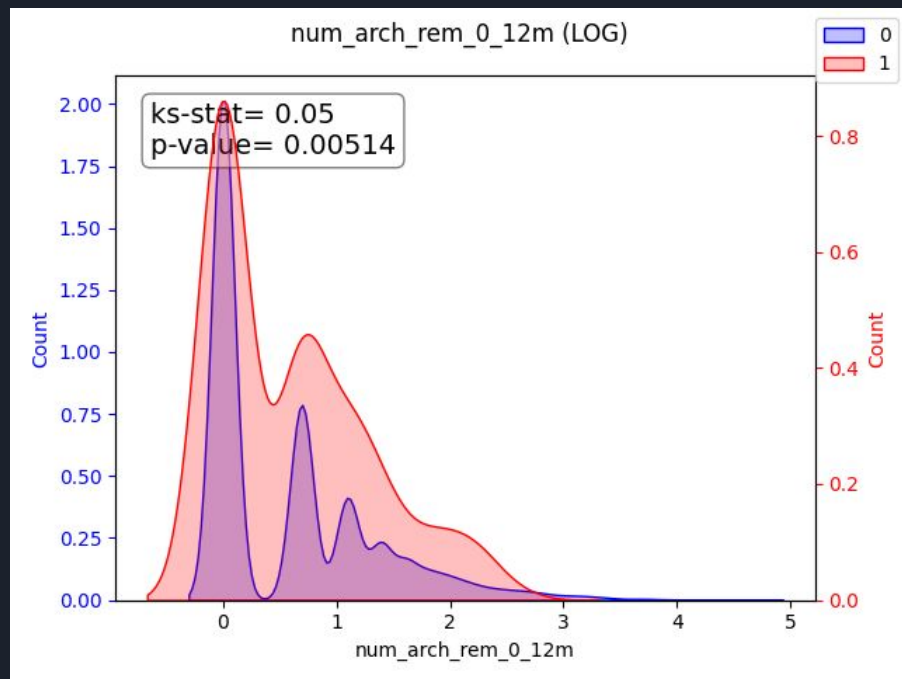
Distributions - Numerical



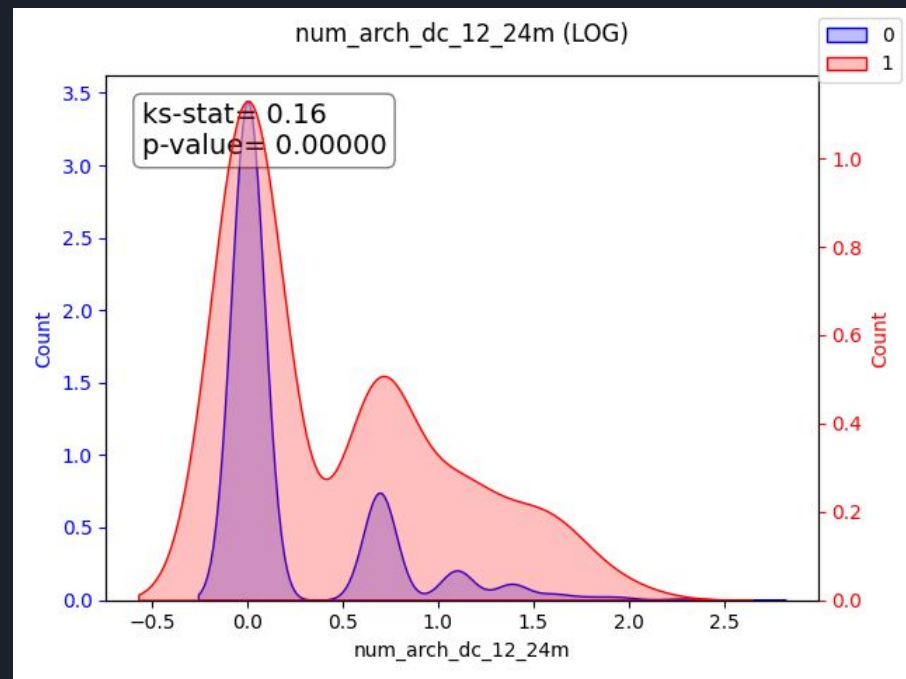
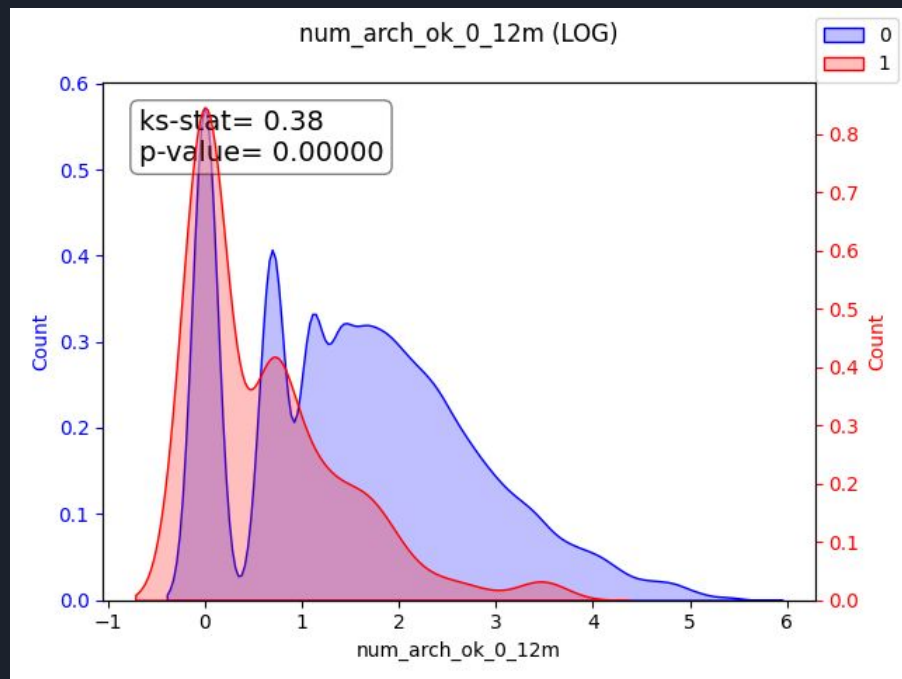
Distributions - Numerical



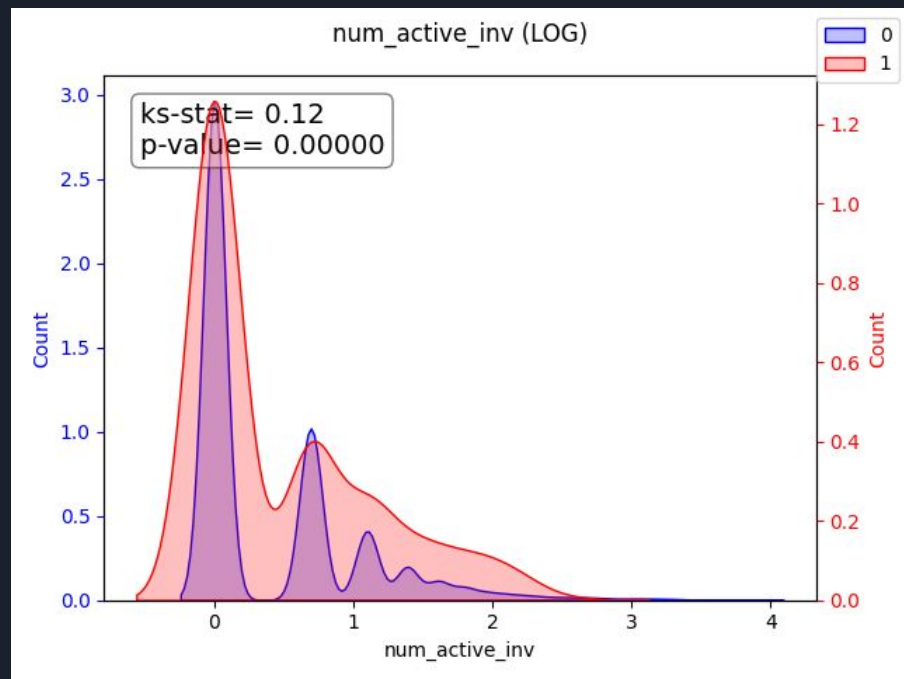
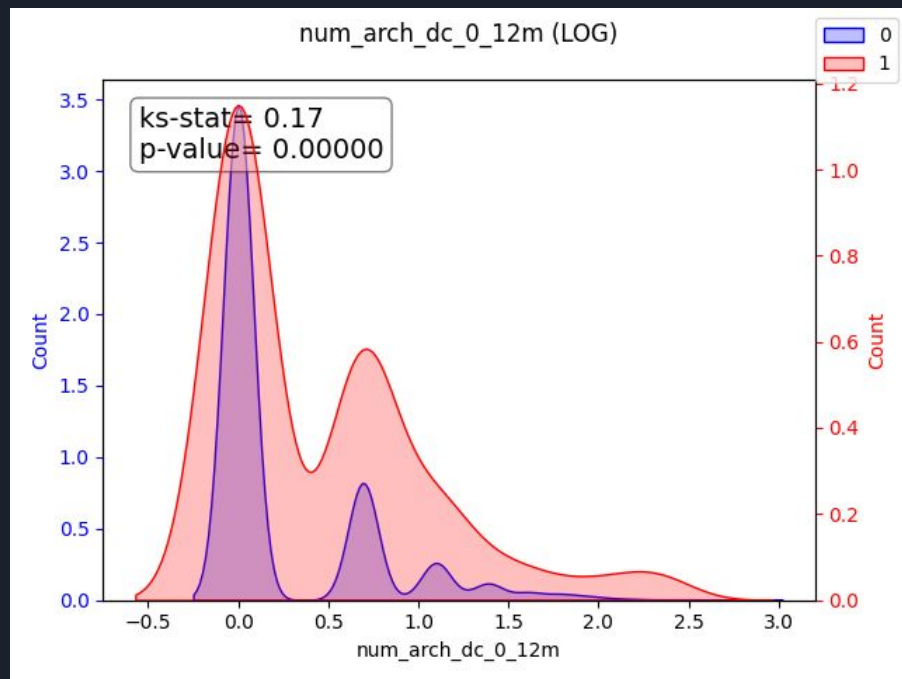
Distributions - Numerical



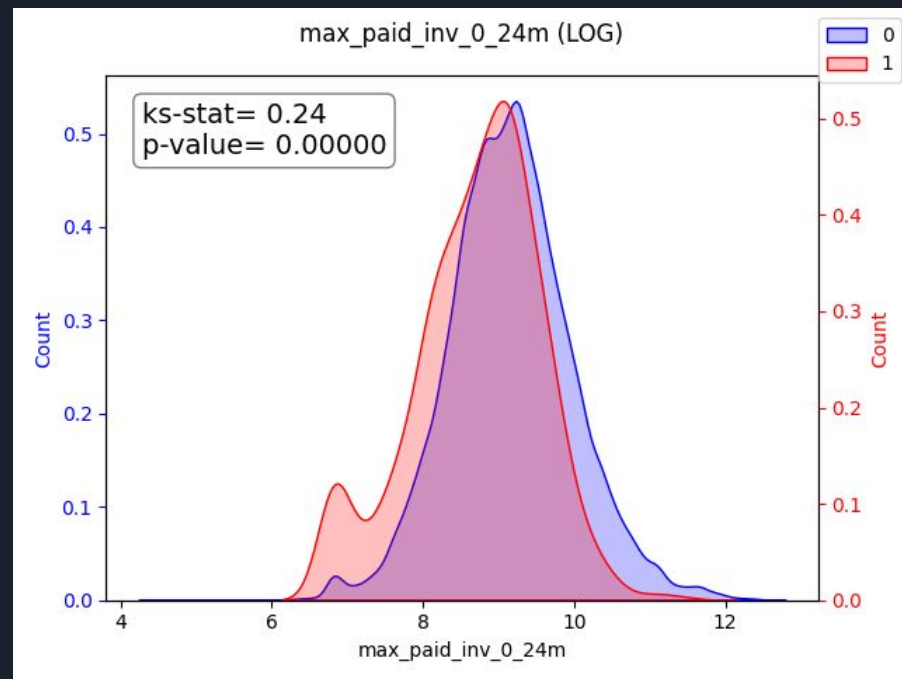
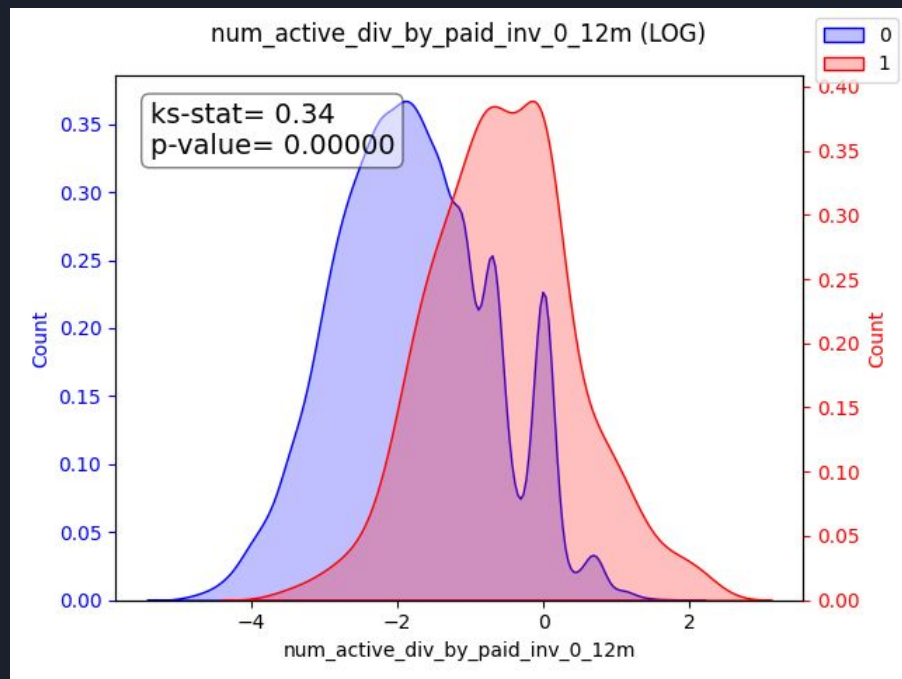
Distributions - Numerical



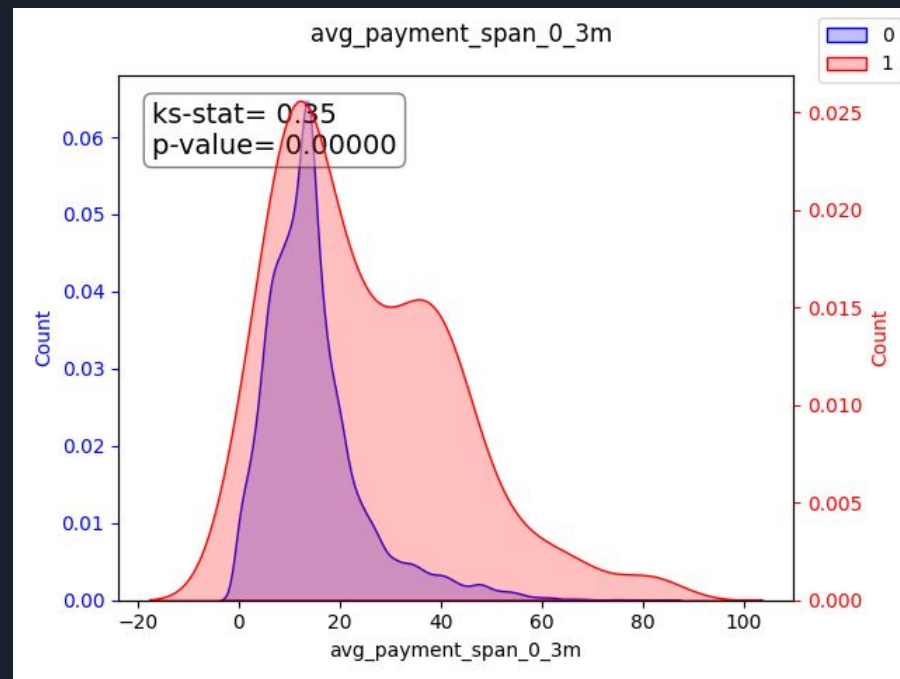
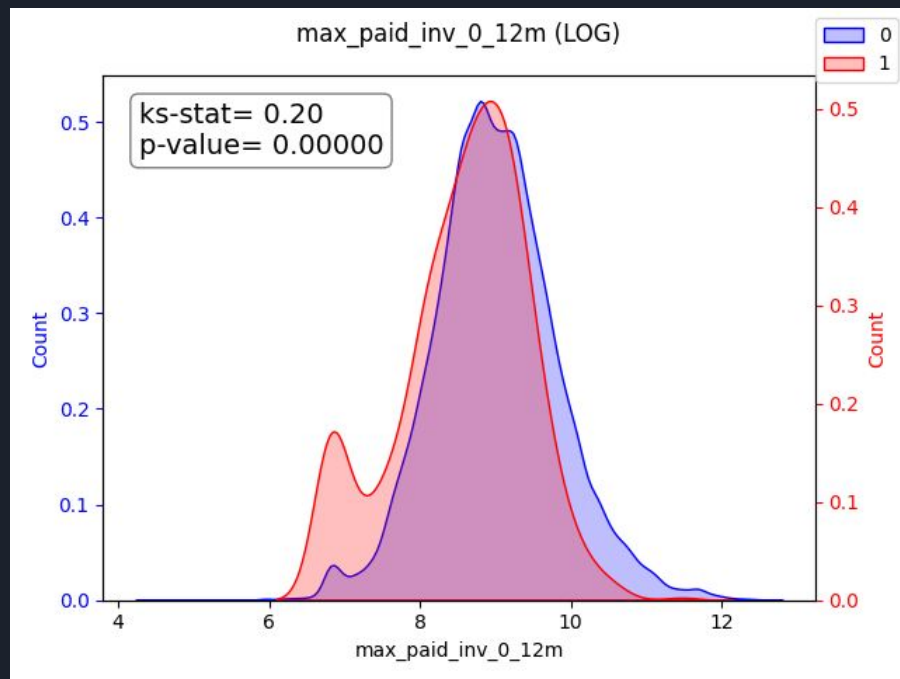
Distributions - Numerical



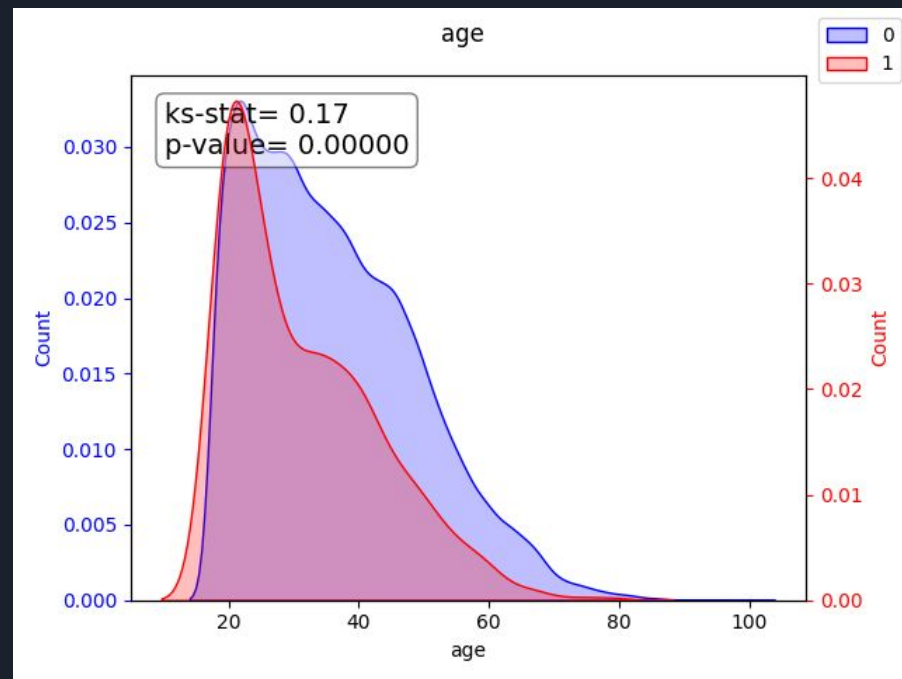
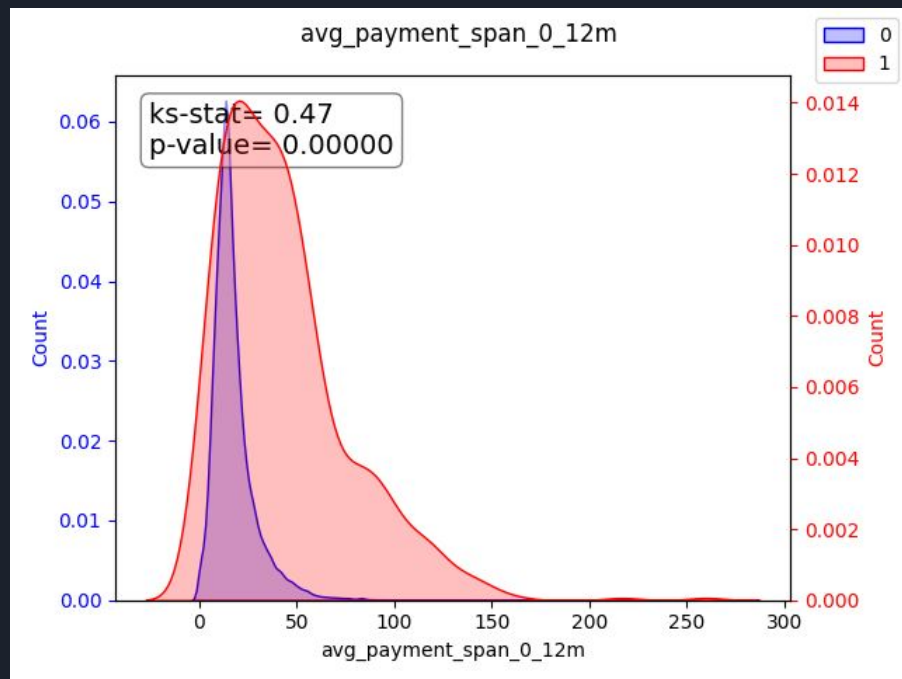
Distributions - Numerical



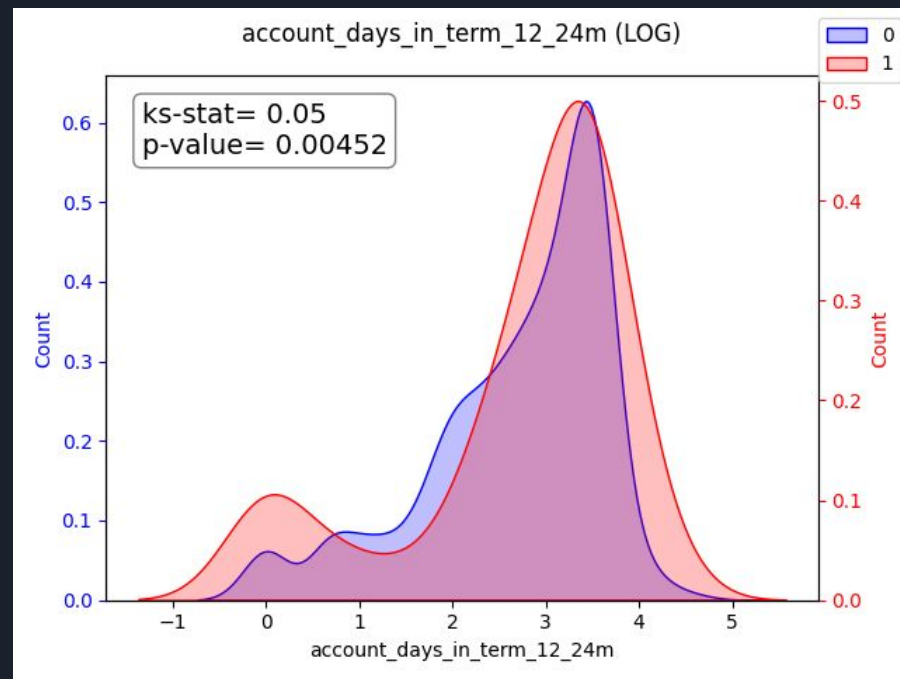
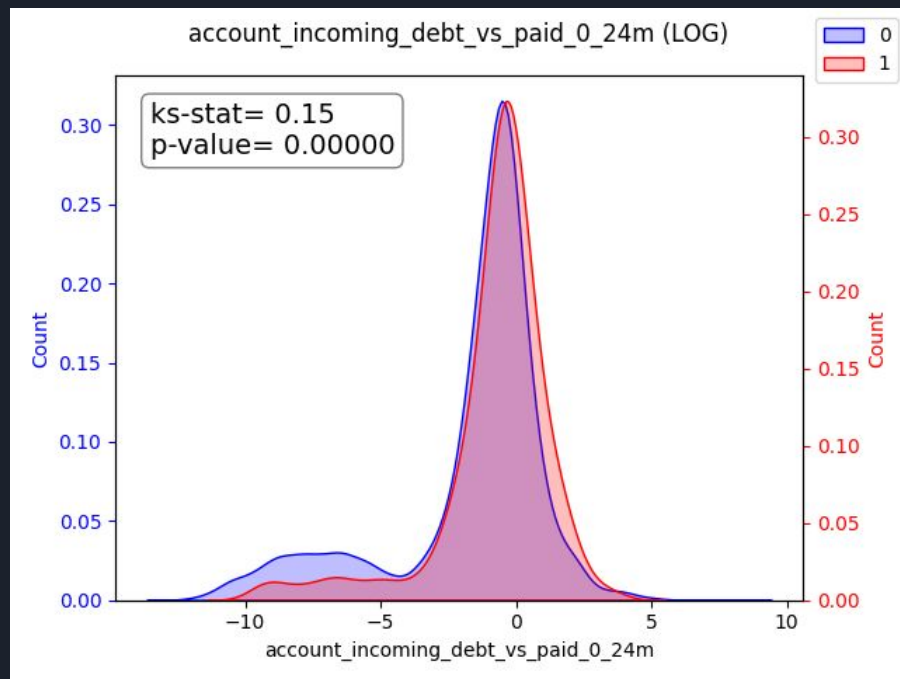
Distributions - Numerical



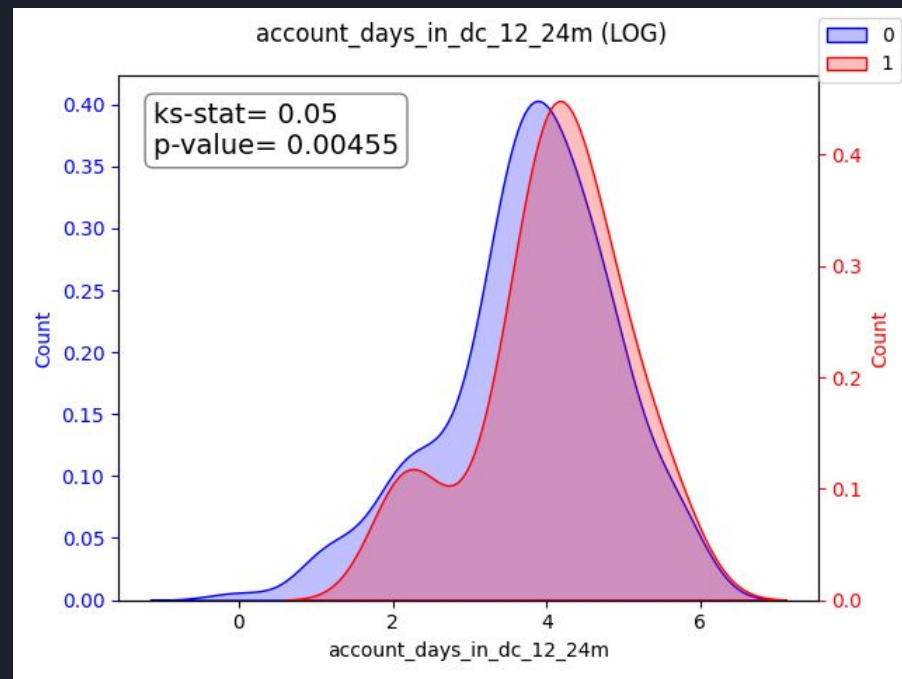
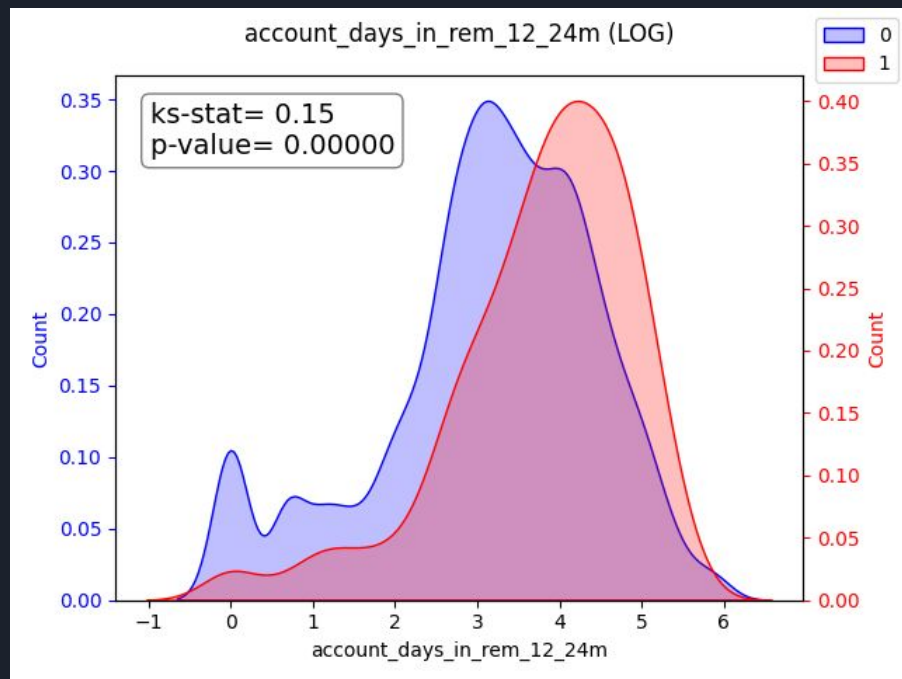
Distributions - Numerical



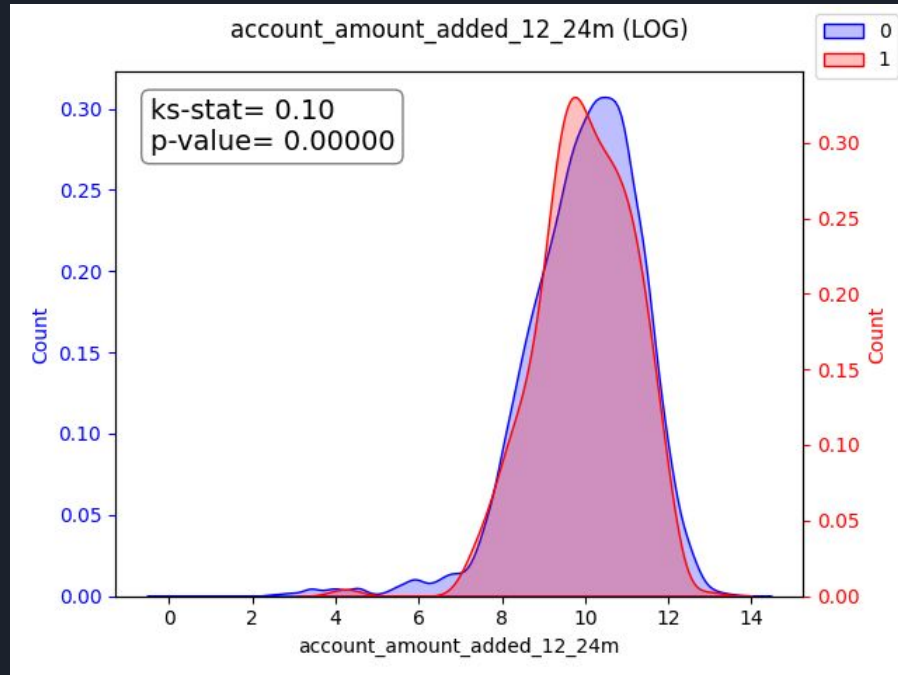
Distributions - Numerical



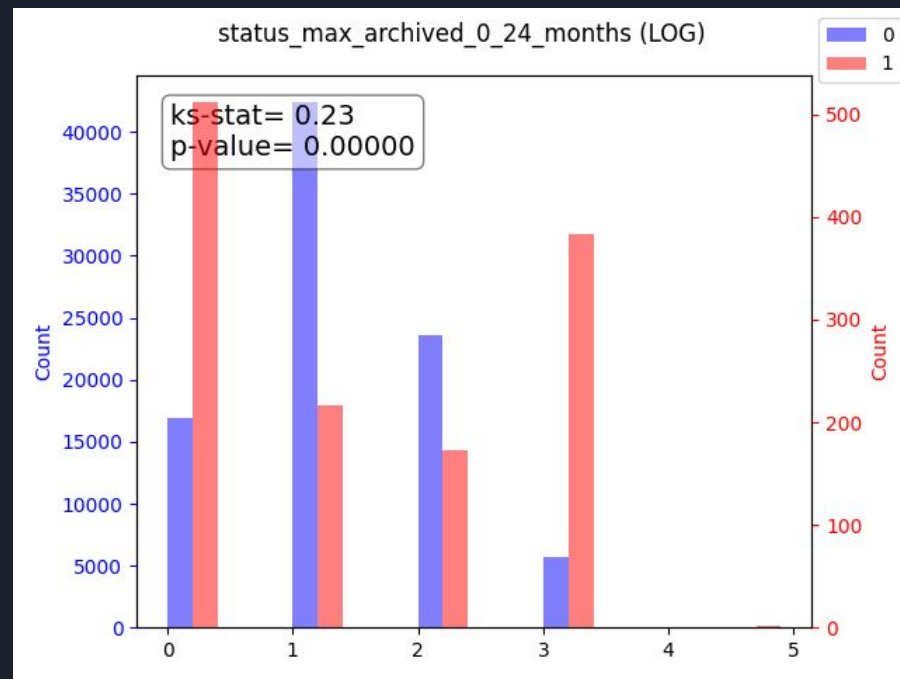
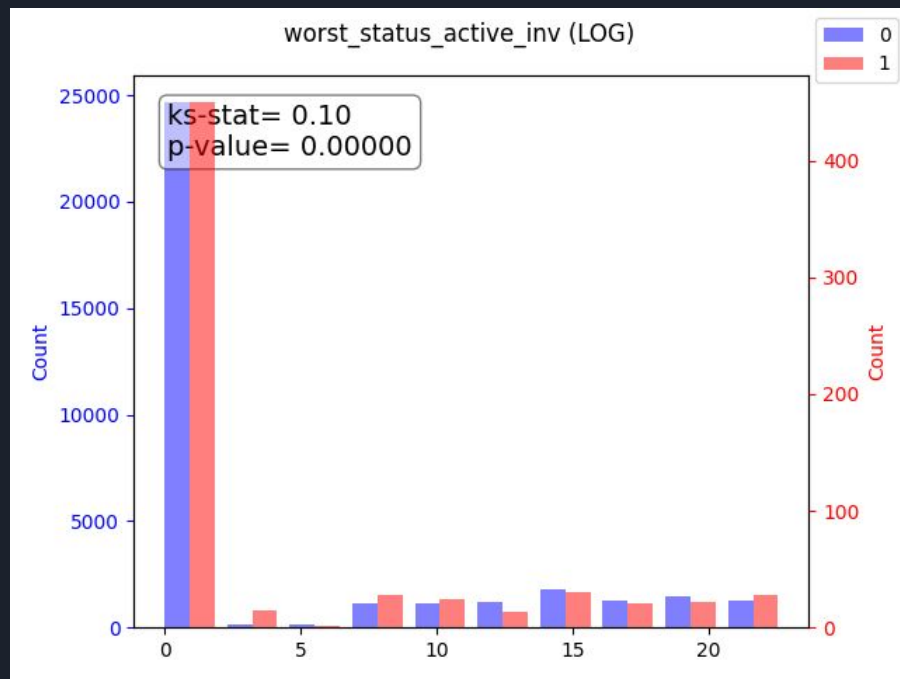
Distributions - Numerical



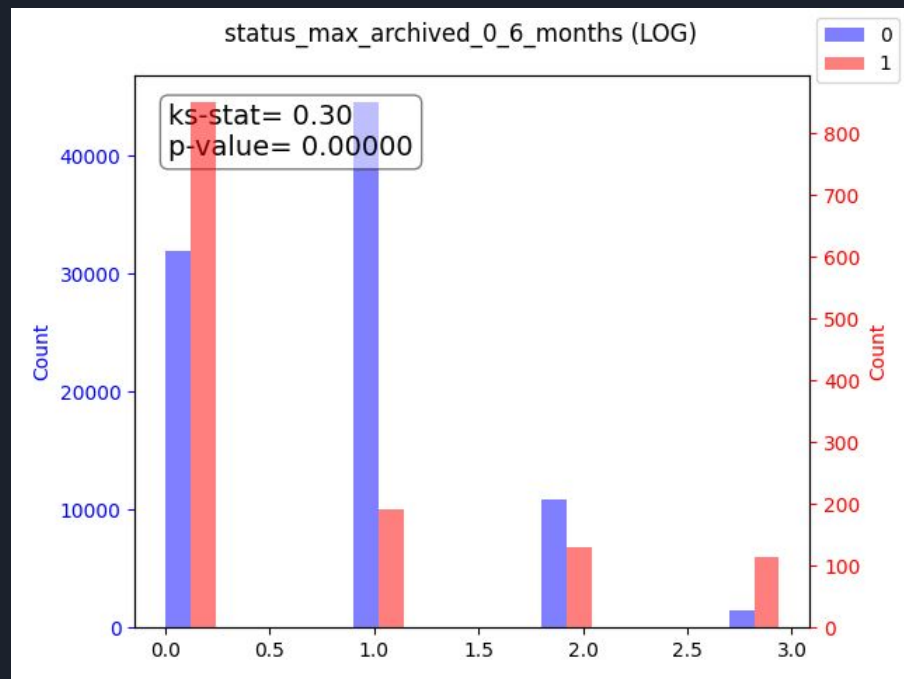
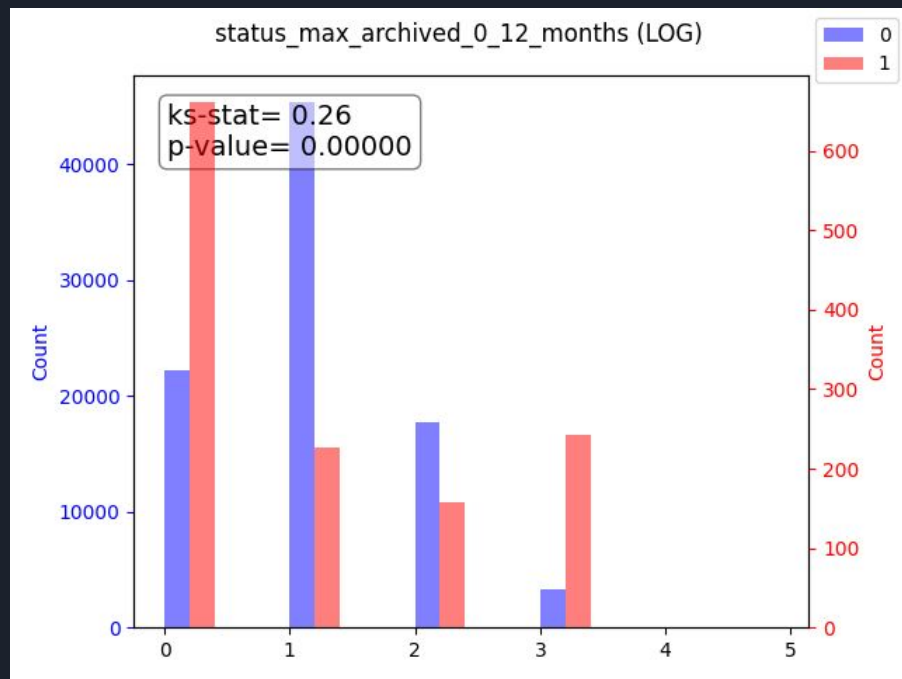
Distributions - Numerical



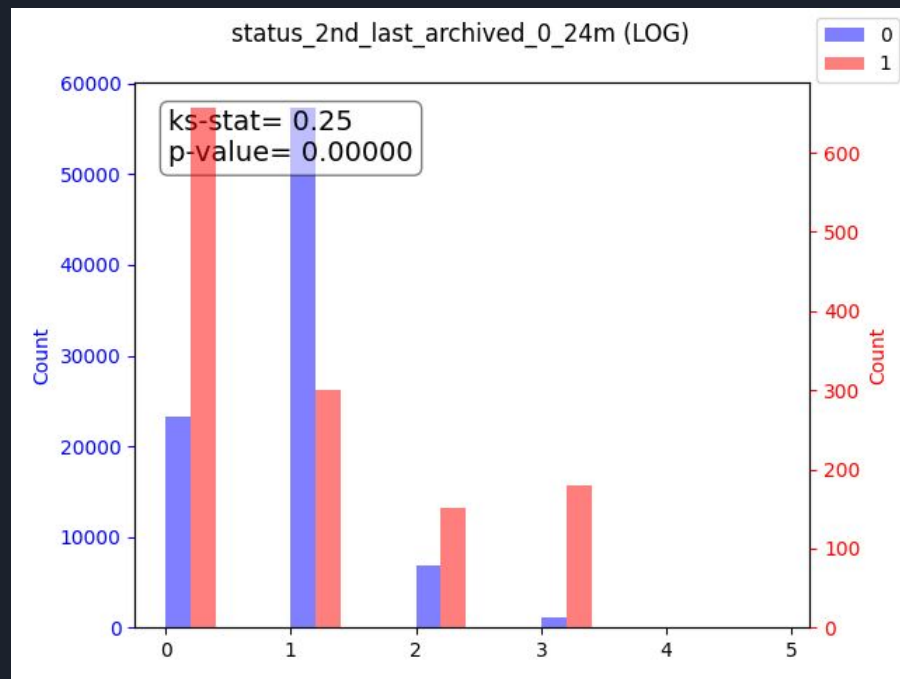
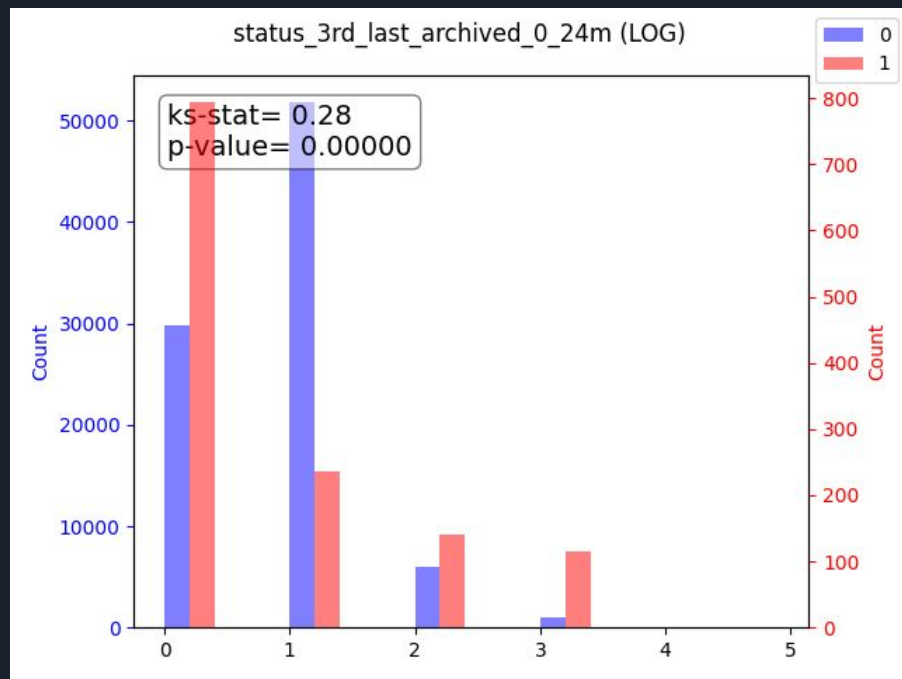
Distributions - Categorical



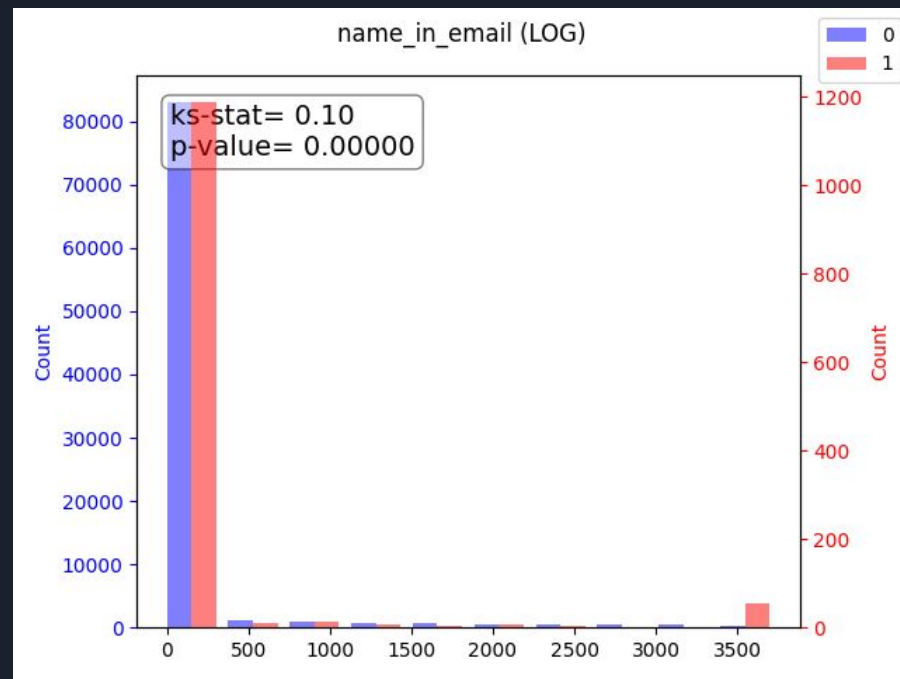
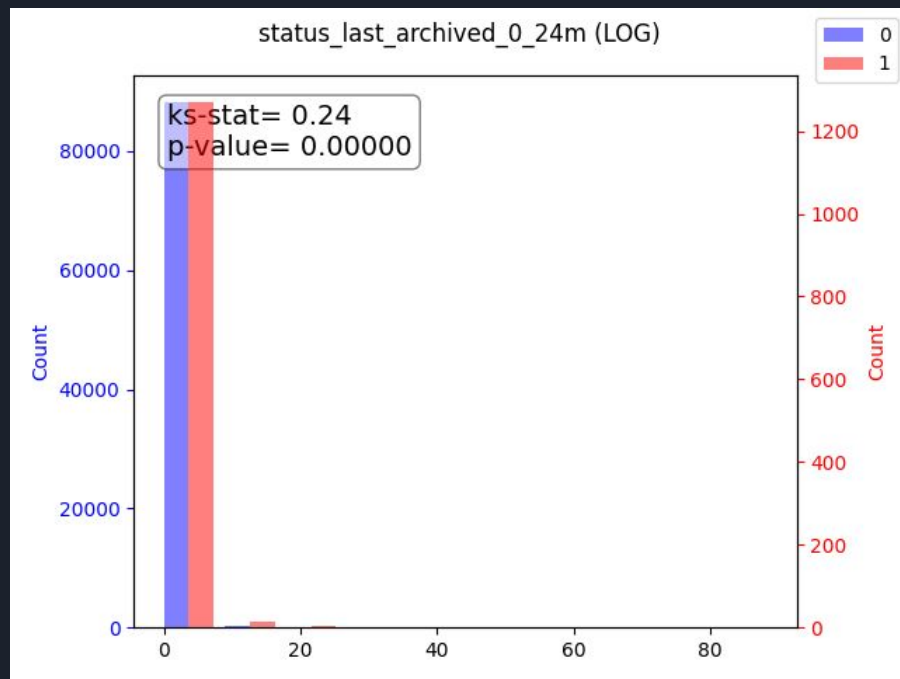
Distributions - Categorical



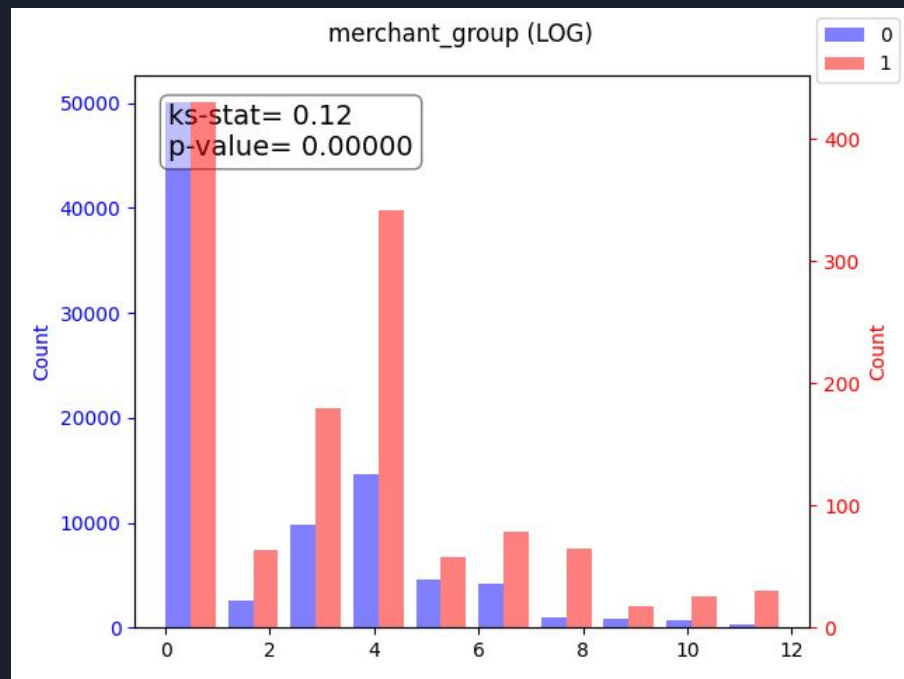
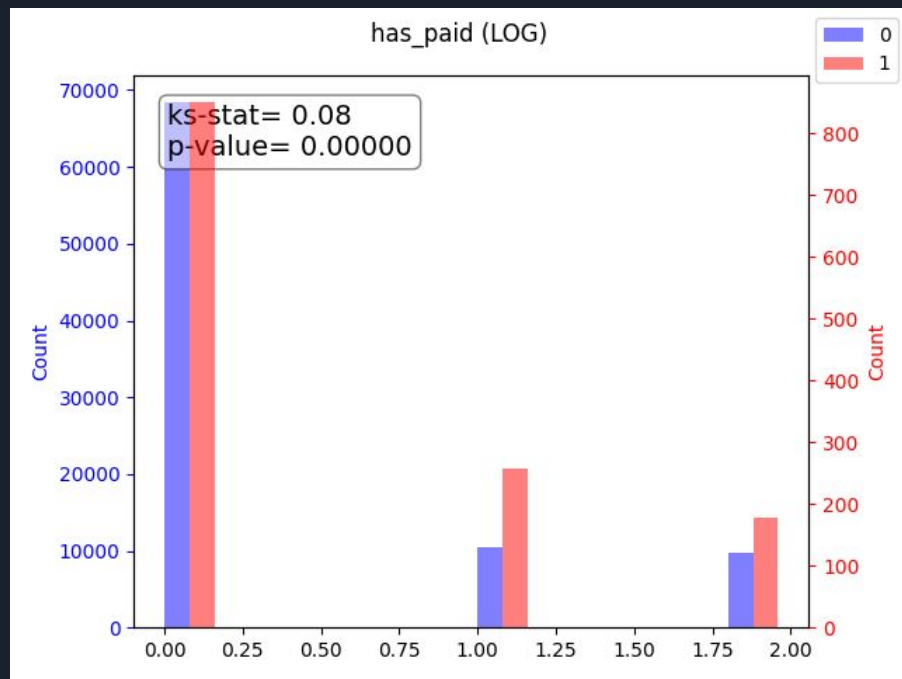
Distributions - Categorical



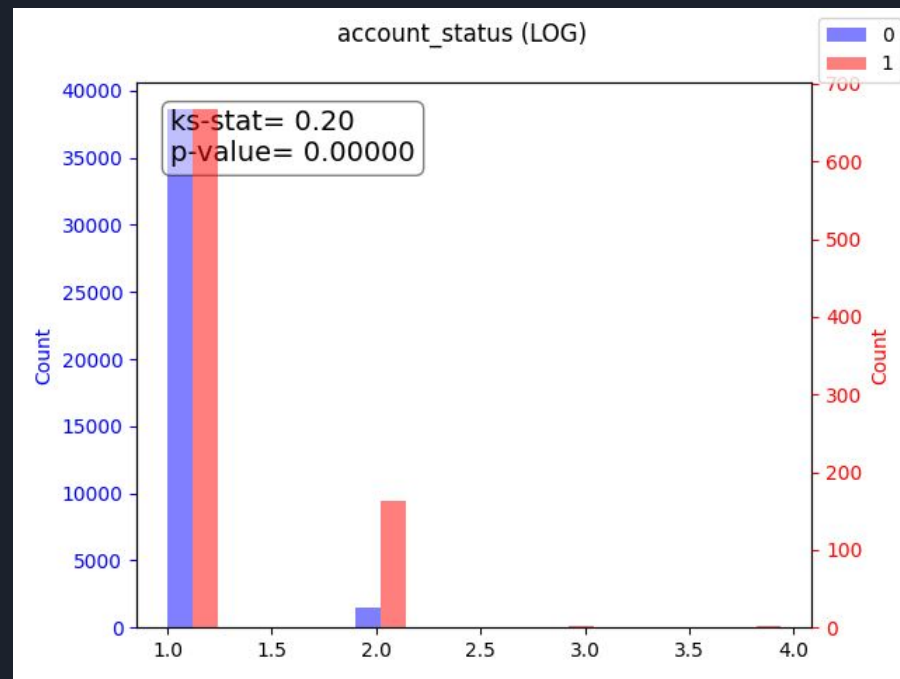
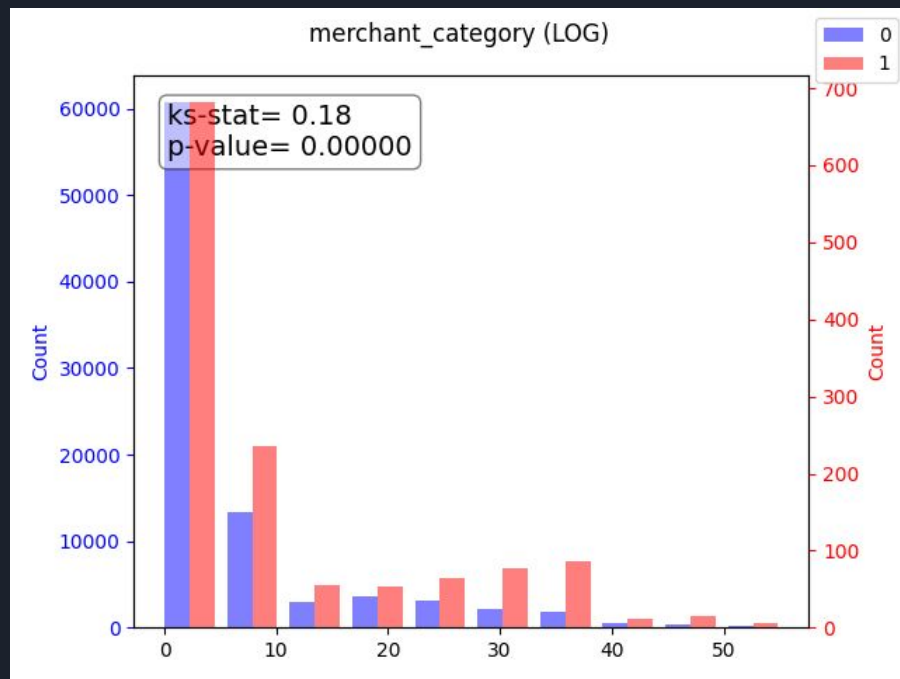
Distributions - Categorical



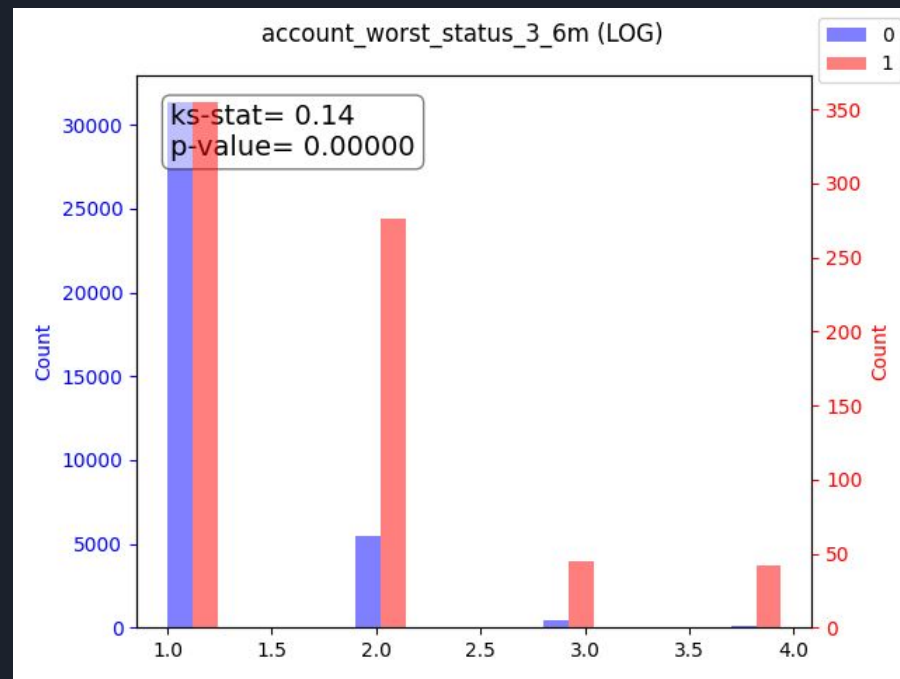
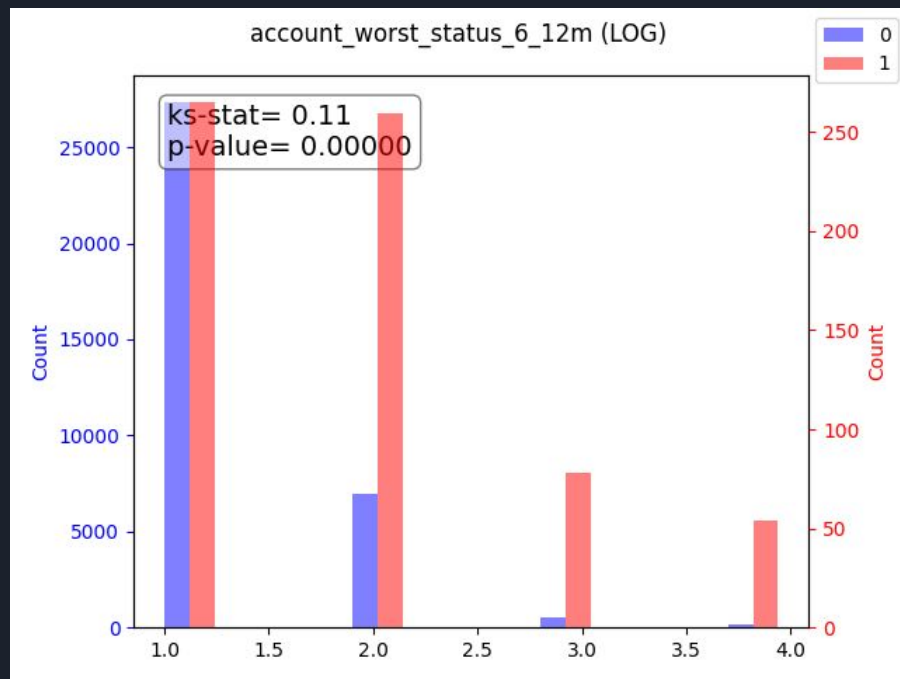
Distributions - Categorical



Distributions - Categorical



Distributions - Categorical



Distributions - Categorical

