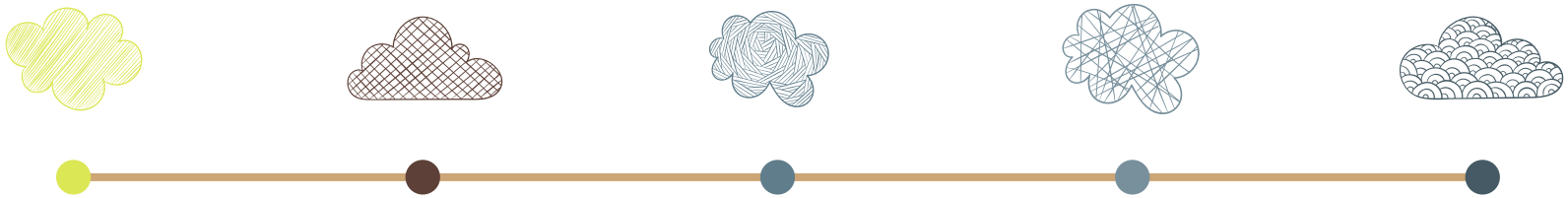
Credit Card Fraud Detection

Group 4 - MSiA Cloud Engineering

Yiyang(Jade) Cao, Yuwen(Maria) Meng,
Ziyan(Cheryl) Liu

Slide Link: https://docs.google.com/presentation/d/1s8pWavL2X3jiGc2KmfDSNCHJ0e7_R3tI2ZZnvNvHQ74/edit?usp=sharing

Presentation Agenda



01
Business Problem

02
Data Introduction
&
EDA

03
ML OPs Flow

04
Cloud Architecture
&
Web App
&
Cost Estimate

05
Next Step

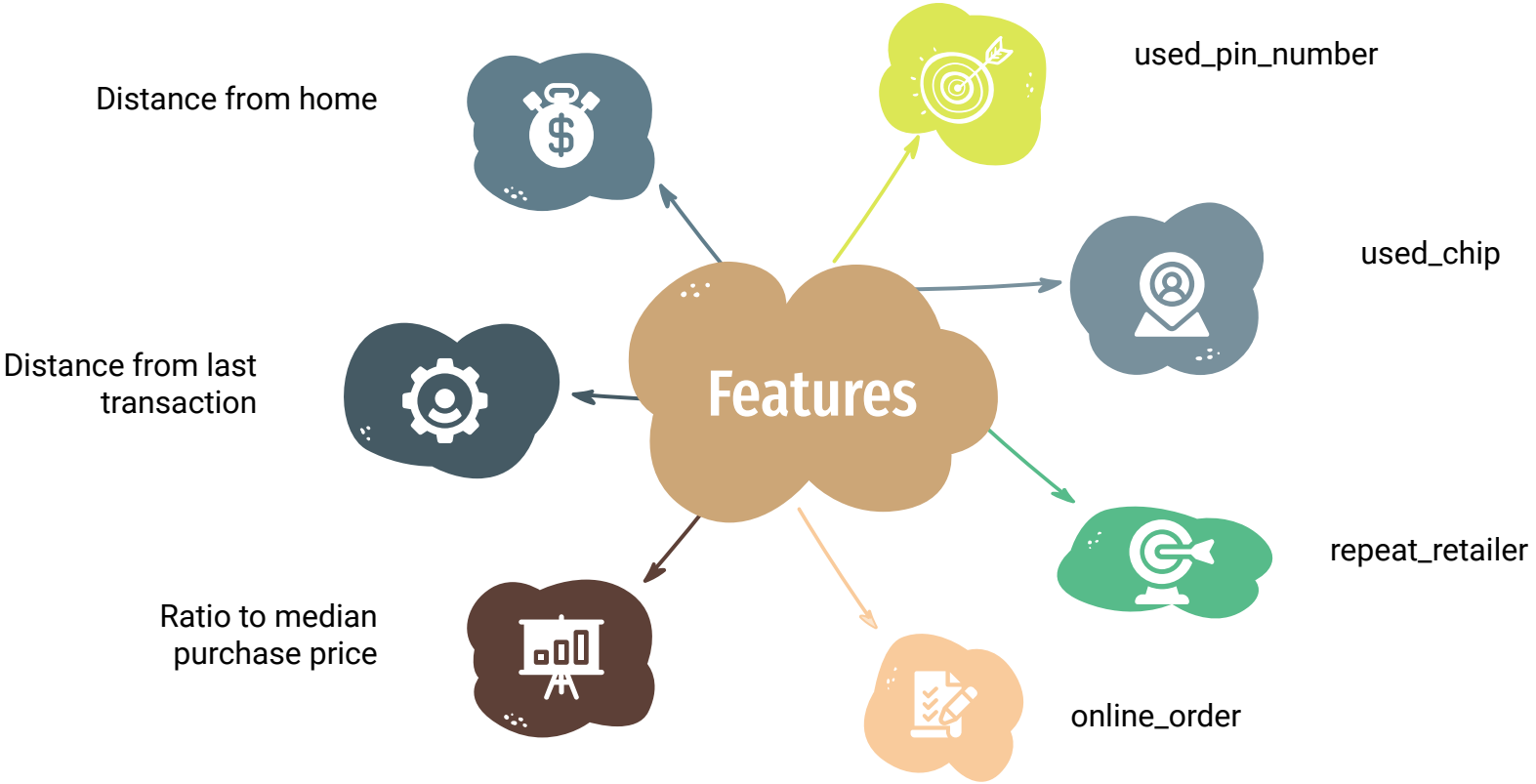
Business Problem - Credit Card Fraud Detection

- ML models -> detect credit card fraud
- More than **5 million** records are being stolen on a daily basis (Data Breach Index)
    - Both for Card-Present and Card-not Present type of payments

Pipeline
Training different model versions

Inference
Streamlit App for fraudulent transaction prediction

Data Introduction

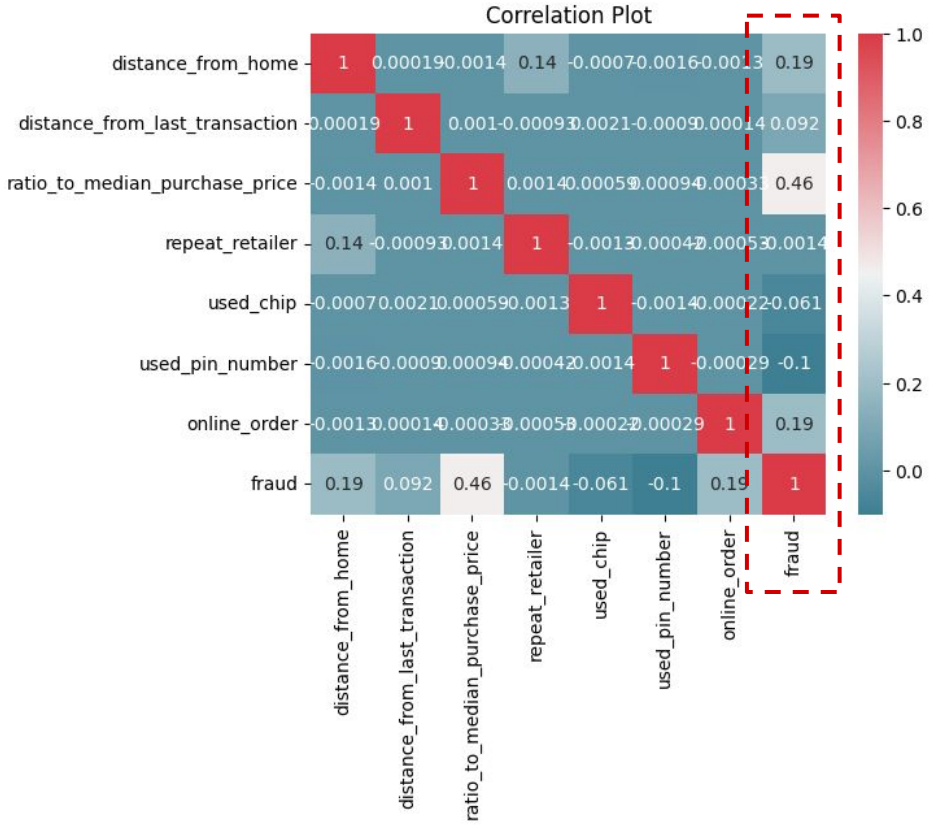


EDA (Heatmap)

Findings:

- Only *ratio_to_median_purchase_price* has a relatively higher correlation with fraud.

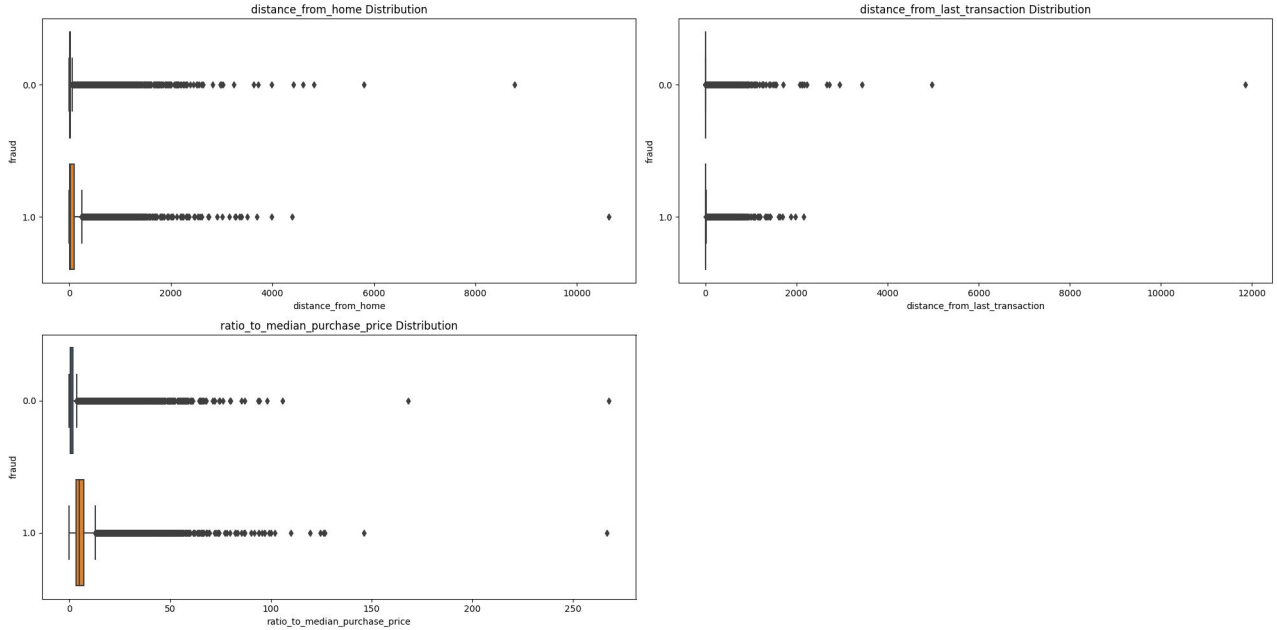
- There's no trivial correlation between each variable for our data. Multicollinearity would not be an issue that needs to solve.



EDA (Numerical Features)

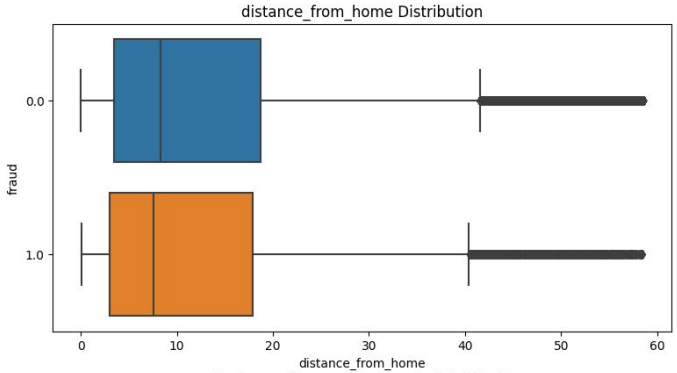
Boxplot of each numerical features

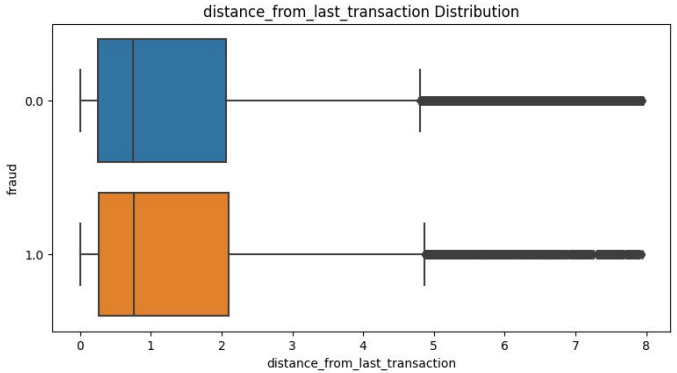
- There are some outliers in features (distance from home or transaction, purchase price).
- Drop samples with values exceeds 1.5*IQR

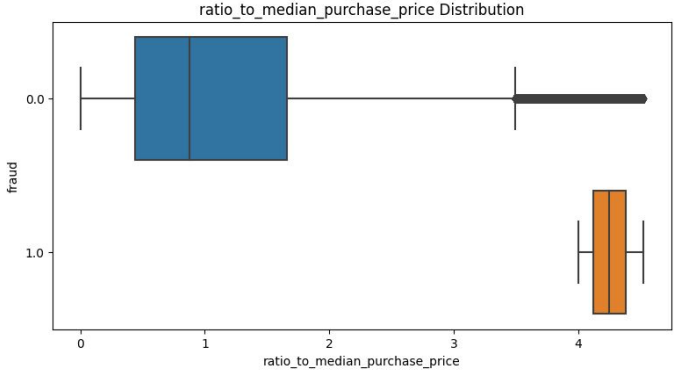


EDA (outliers removed)

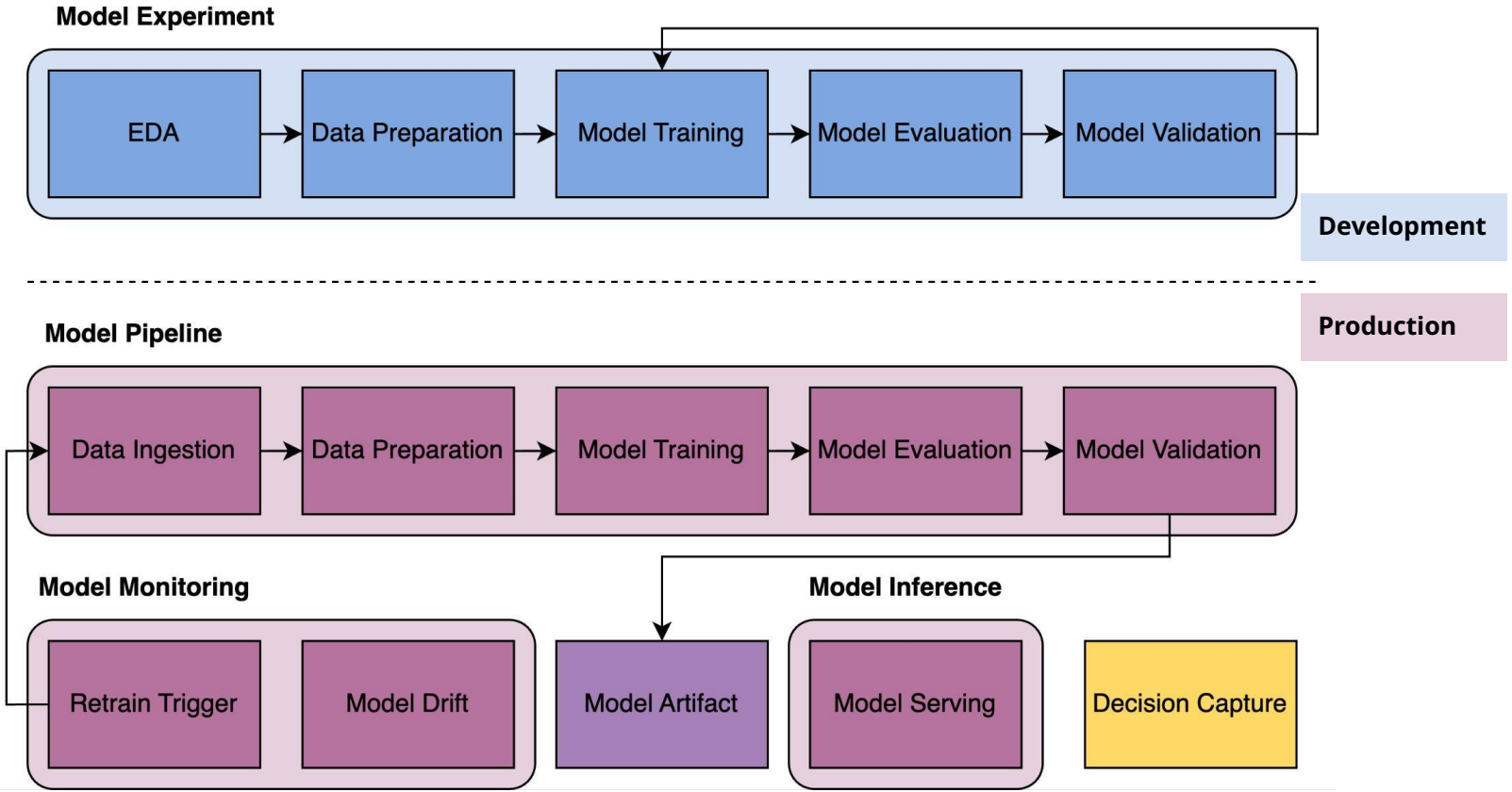
- Boxplot of each numeric features after outlier removal



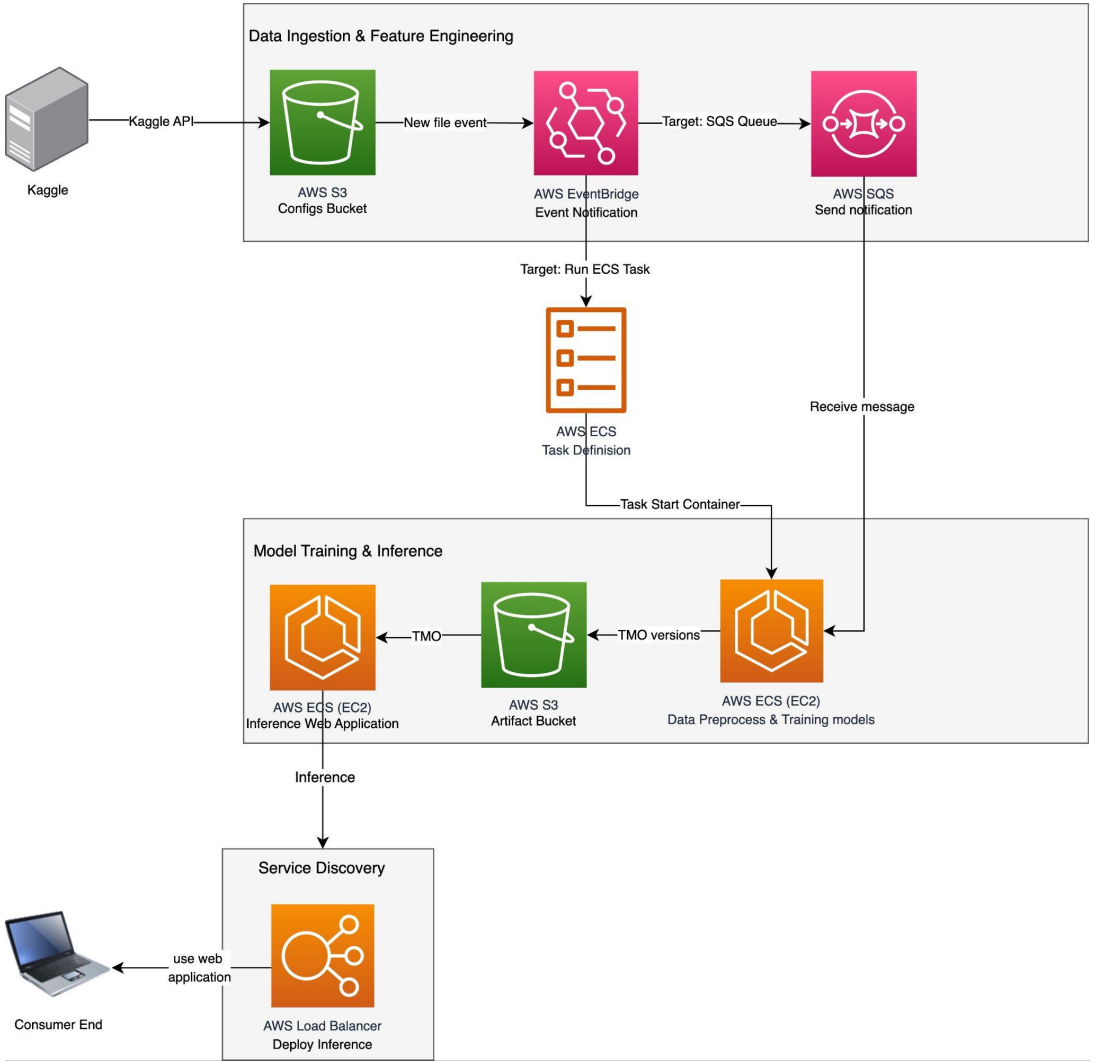




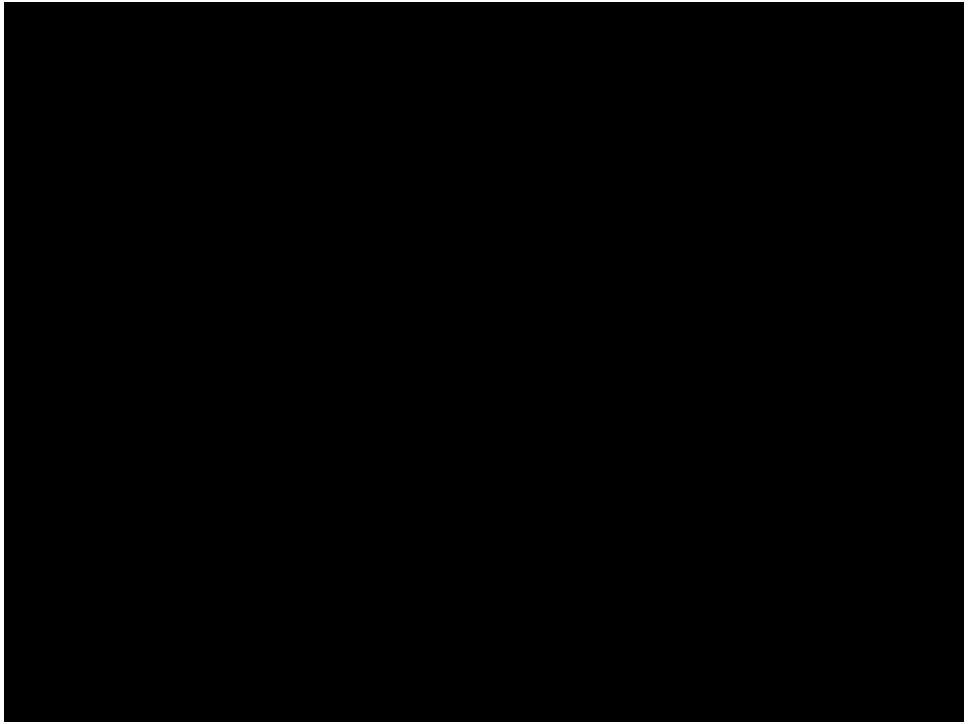
ML OPs Flow


Model Experiment
EDA
Data Preparation
Model Training
Model Evaluation
Model Validation
Development
Production
Model Pipeline
Data Ingestion
Data Preparation
Model Training
Model Evaluation
Model Validation
Model Monitoring
Model Inference
Retrain Trigger
Model Drift
Model Artifact
Model Serving
Decision Capture

Cloud Architecture



Inference app



Streamlit demo

- 2 model versions
- Input features using sliding value and select value
- Output: fraud alert and probability
- Use ECS cluster

Cost Estimate

Link: https://calculator.aws/#/estimate?id=82de0239d303747f8718a8216d1d3bee4aee5068

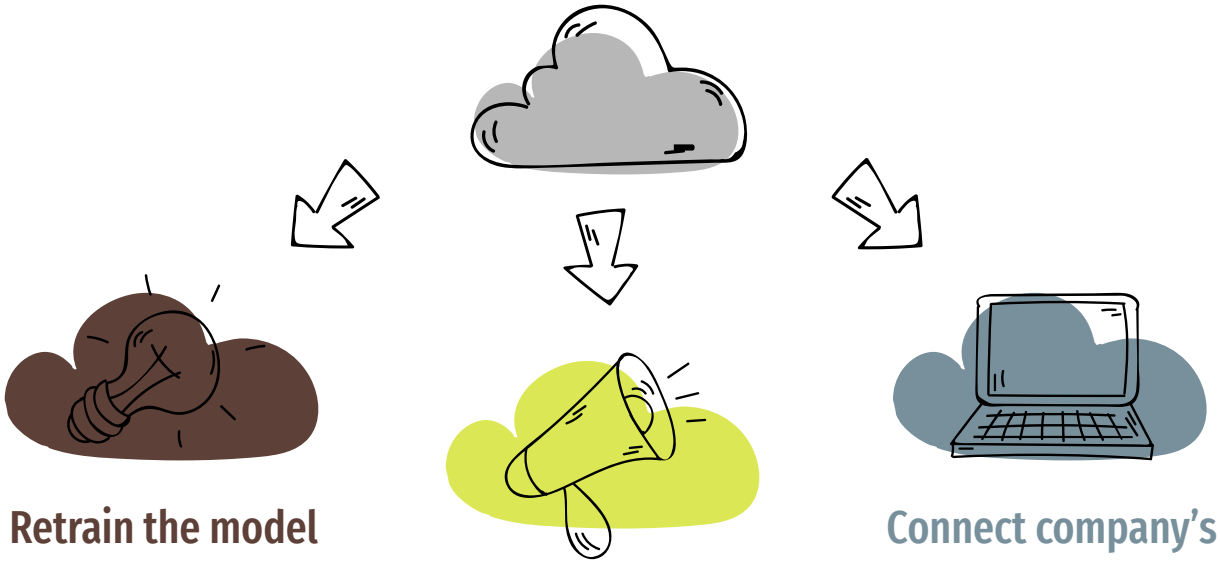
Model pipeline and inference

Distribute incoming traffic across multiple instances

Store artifacts and models

Hold Event bridge messages for switching config and experiments for ECS task

Next steps



Retrain the model on the whole dataset

Prediction summary about model metrics

Connect company's database rather than kaggle public data

Thank You!