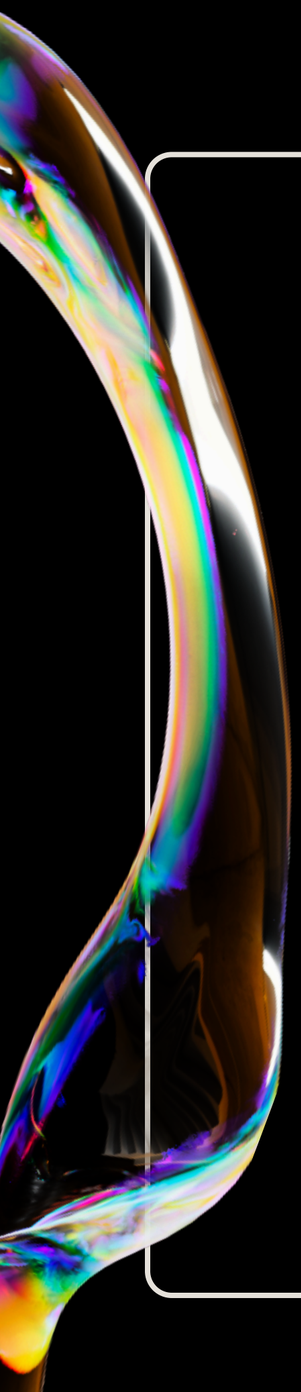
PIPEDRIVE

Presented by Jako Rostami

https://github.com/jakorostami

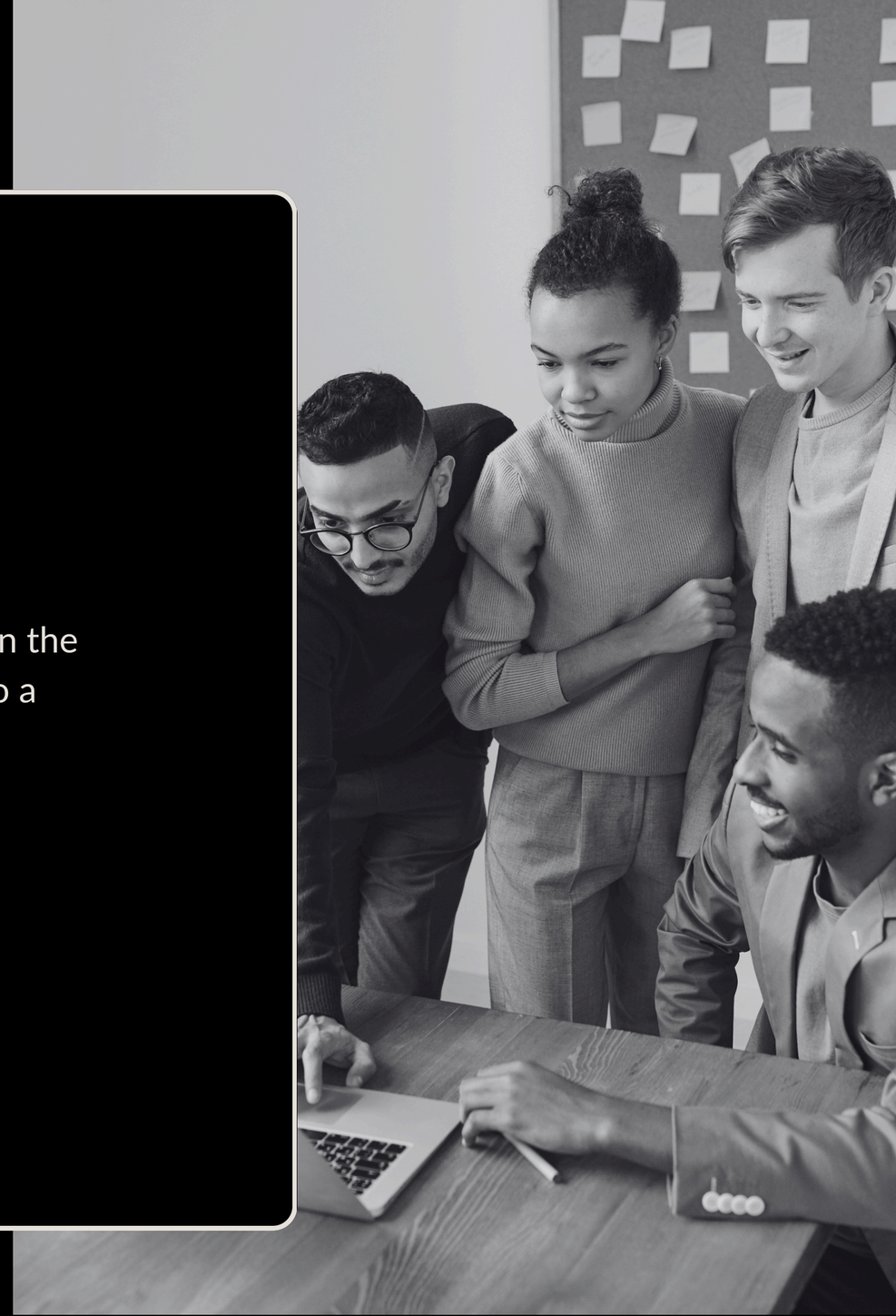
rostami.jako@gmail.com

CONTENT



PROBLEM STATEMENT

We have been given the task to create a machine learning pipeline in production that predicts the likelihood of a customer transaction within the last two months of the latest purchase. Essentially, this breaks down to a success vs. failure problem.



DATA

E-Commerce Sales Data

5 months of sales data showing each transaction id and each associated customer with it. The granularity is on the hour-minute-second level with daily marks.

Products and Categories

The data shows us values per product, category, gender, state, city, the device type, customer login type, and delivery type. Meaning we can see a hierarchical structure to the provided data.

Purchase != Successful

The target of interest is whether the transaction was successful or not. This means that customers can make a purchase but that does not mean it is successful. Speculation tells us that customer lack funds, incorrect payment details, or technical issues.

ANALYTICAL APPROACH

Rapid Data Analysis

Given the nature of the data, the time available, and the task at hand - we have provided a rapid analysis that focuses on the data quality and analysis part.

Advanced Price Engineering

Following a scientific and Data Science approach allows Pipedrive to always provide high data quality. E.g, we found that the Amount in US$ and individual prices come with errors so we did statistical imputation by separating the dirty data from the clean.

Predictive Success

With the previous steps done, we have set the foundation for creating the prediction service for consumer usage.

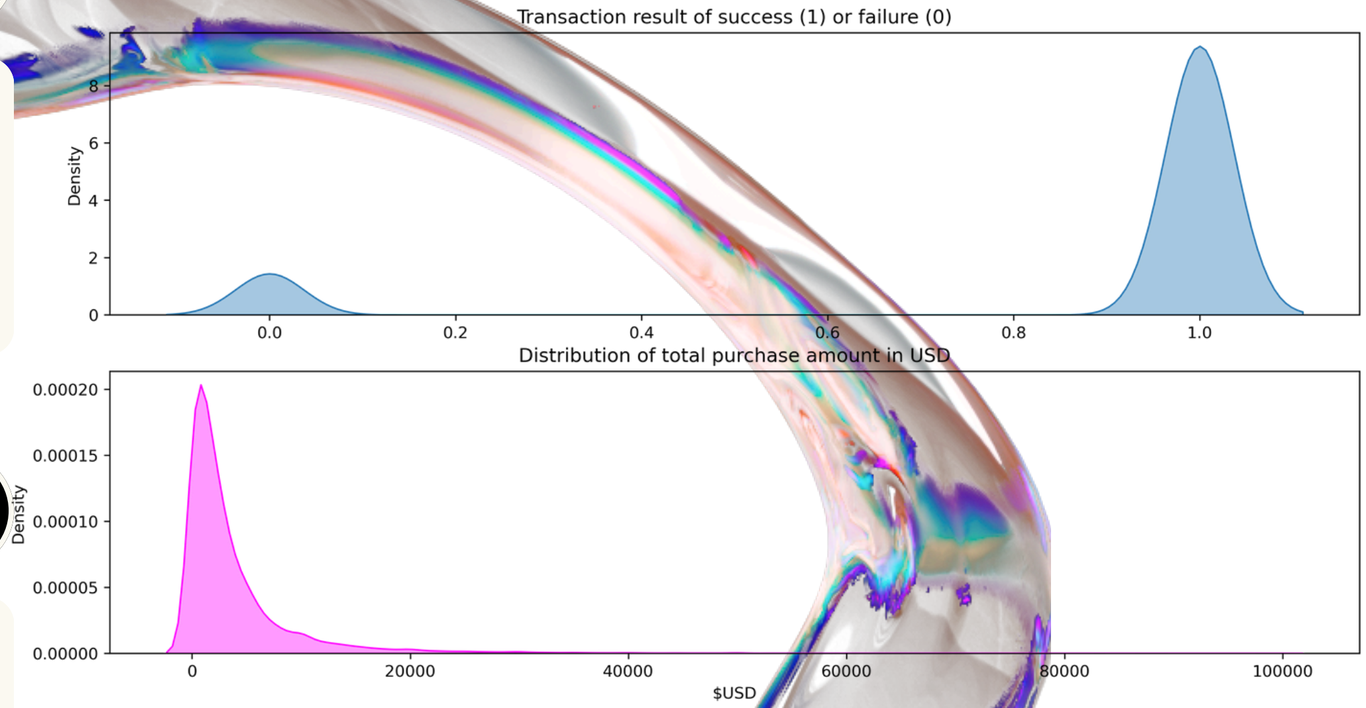
FINDINGS

IMBALANCED DATASET

We identify an imbalanced dataset with about 87% transactions successful - so we ignored methods to balance the dataset because of time and that it has no realistic statistical representation.

DATA CONCENTRATION

Majority of the total purchase amounts are concentrated at around $4000 but ranges up to $100k.



MODEL PERFORMANCE

	Strengths	Challenges	Proposal
AdaBoost/ CatBoost/ XGBoost	Strong predictions	Requires retraining as data grows	Adaptive retraining schedule
Decision Tree/ Random Forest	Highly interpretable	Computationally heavy for larger datasets	Baseline model
Neural Net	Detecting complex patterns	Requires significant data	Retraining schedule should be sparse
Ensemble	Highest potential accuracy	Complex deployment pipeline	Use in production

DEPLOYMENT

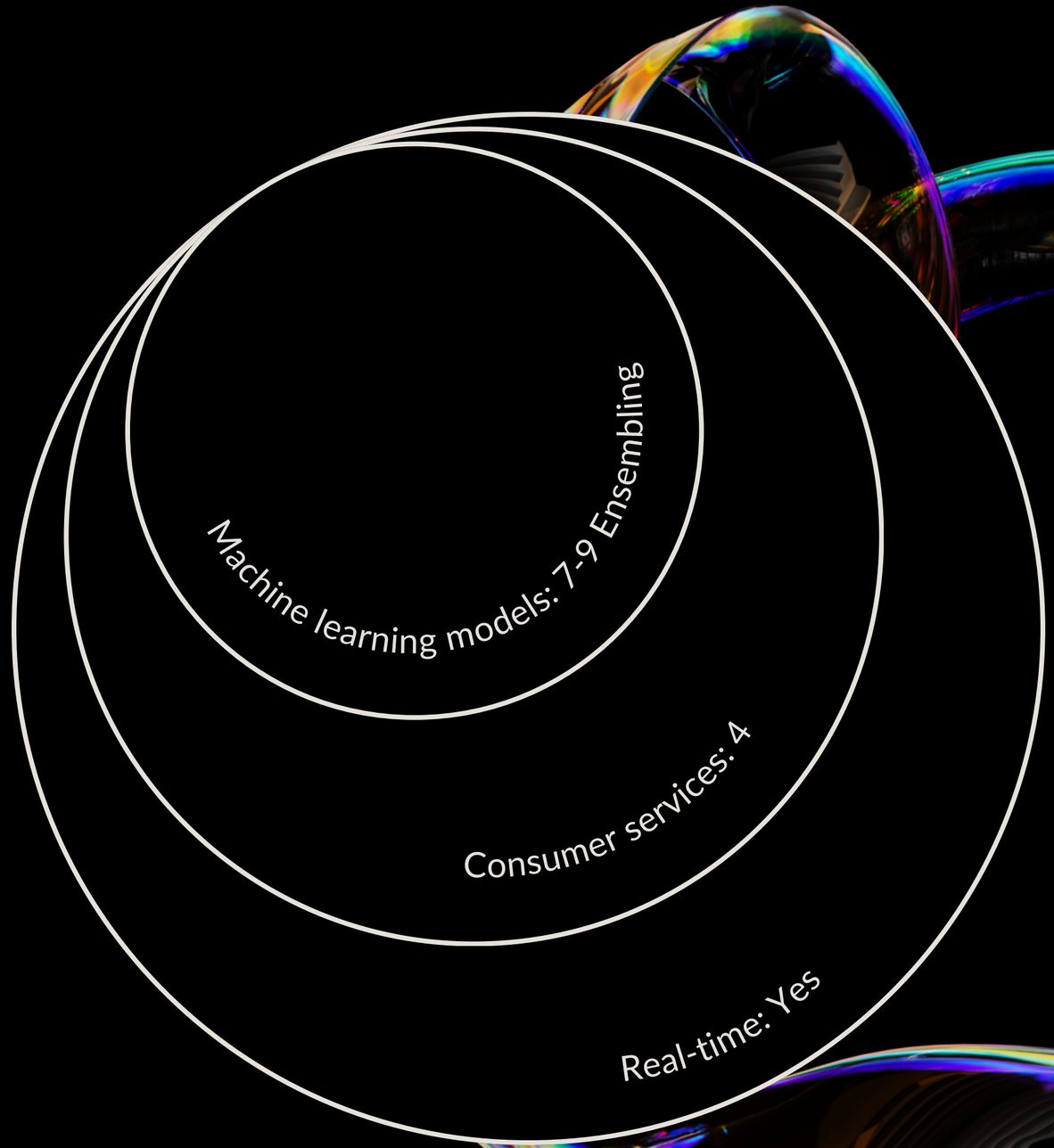
Drawing from our findings, we developed our predictive service to address data quality issues and allow the consumer to re-train in real-time!

Our layered modeling approach accounts for:

- Prediction service for consumer
- Load specific versions of your Machine learning models
- Re-train with new data in real-time
- Reload your models with new data

This strategic framework allows us to build consumer focused prediction services while accounting for the nuisance a consumer could experience.

Maybe you want to train on demand - here you go!



THANK YOU

Presented by Jako Rostami