Empowering Parents:
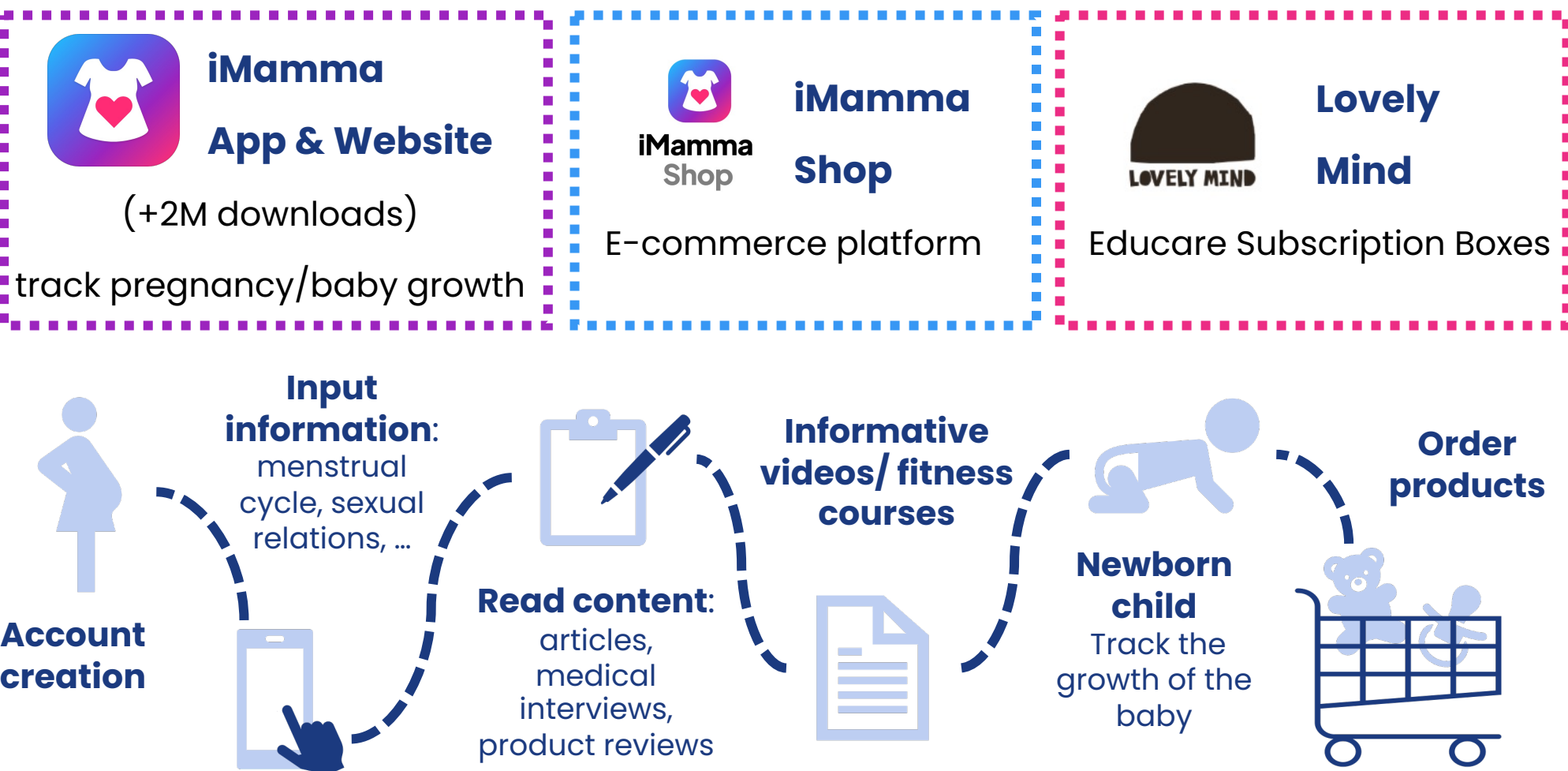## Personalized Recommendations For Italy's Top Pregnancy App

MBAn Students: **Jan** Reig Torra & **Gavin** Findlay
Faculty Advisor: **Thodoris** Lykouris
Angelini Stakeholders: **Francesco** Tortora, **Artem** Ugarov & **Lida** Sorino

Angelini Consumer

MIT MANAGEMENT BUSINESS ANALYTICS

The Company

Angelini Industries is a multinational conglomerate headquartered in Italy.

One of its divisions, **Angelini Consumer**, wants to build the "Lonely Planet Guide for Young Families", through its entities:

- **iMamma App & Website** (+2M downloads) track pregnancy/baby growth
- **iMamma Shop** E-commerce platform
- **Lovely Mind** Educare Subscription Boxes

Account creation → **Input information**: menstrual cycle, sexual relations, … → **Read content**: articles, medical interviews, product reviews → **Informative videos/ fitness courses** → **Newborn child** Track the growth of the baby → **Order products**

Problem Statement

Angelini possesses a valuable, **untouched data collection**. The primary business goals involve

- Uniting raw data sets and extracting business insights/KPIs
- Leveraging data to convert content consumers into purchasing customers

Objectives

- Pipeline for preprocessing & cleaning data
- Unite data from 3 sources & illustrate relations
- Identify KPIs to measure success and visualize them
- Understand the customer journey: When/why do users become customers?
- Improve customer experience by providing product recommendations

Main Deliverables

- **Dashboard**
- **Sales Forecasts**
- **Customer Segmentation & Panel Data Model**
- **Recommendation System**

Dataset

USER DATABASE
Information input by user, from 2016 onwards
- **Demographics**
- **Tracking Inputs – Pregnancy / Children Section**

WEB ANALYTICS DATA
User behavior in the App and Website from **Piwik**.
- **Click** data, **events**, and **visits**
- Available from May 2023
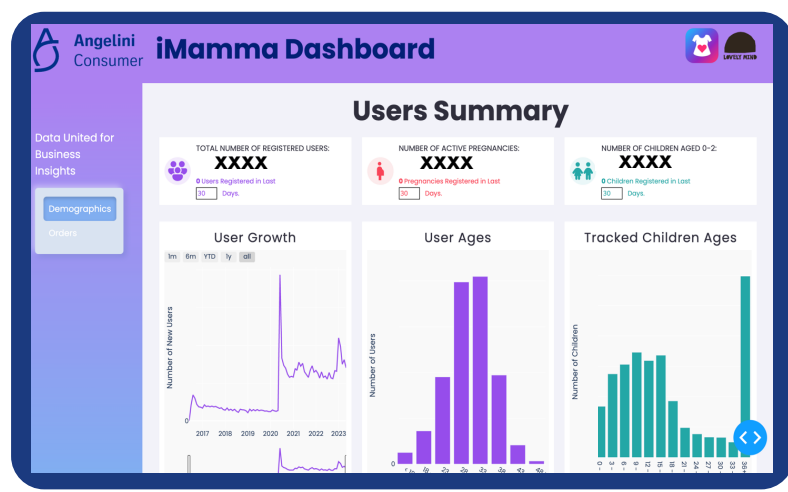
ORDERS & TRANSACTIONS DATA
E-commerce data from **Salesforce**
- **Orders** & **Transactions** with iMamma & Lovely Mind
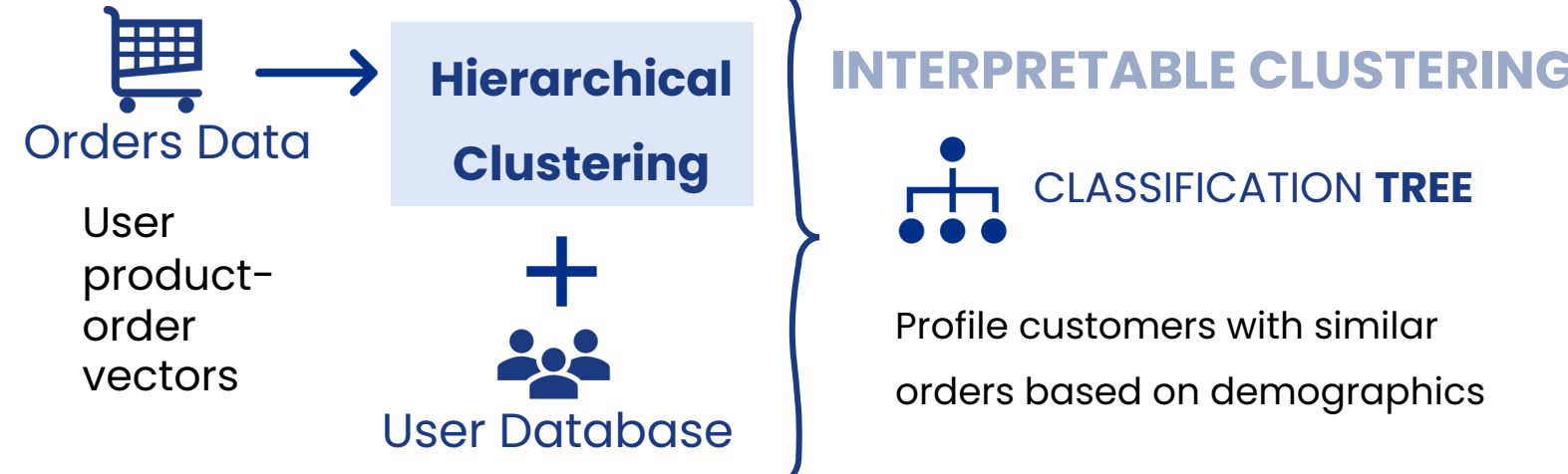- Products added to **Babylist**

METHODOLOGY

The Dashboard

- **Customized dashboard** for measuring business performance
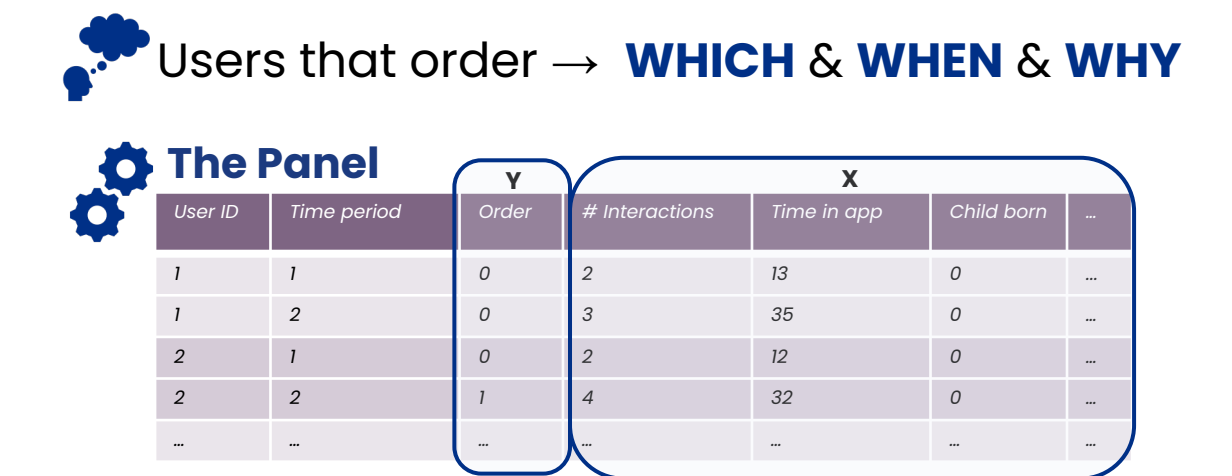- Linked to **Data Lake** for live updates

Understanding the Costumer Journey

CUSTOMER SEGMENTATION AND CLUSTERING

Orders Data → **Hierarchical Clustering** + User Database

User product-order vectors

INTERPRETABLE CLUSTERING

CLASSIFICATION TREE

Profile customers with similar orders based on demographics

PANEL DATA BINARY CLASSIFICATION

Users that order → **WHICH** & **WHEN** & **WHY**

The Panel

User ID	Time period	Y	X			
		Order	# Interactions	Time in app	Child born	…
1	1	0	2	13	0	…
1	2	0	3	35	0	…
2	1	0	2	12	0	…
2	2	1	4	32	0	…
…	…	…	…	…	…	…

Recommendation System

- Limited sales data (**less than 1%** of users have ordered, & most only ordered once)
- **Content-Based Filtering**
- **IDEA**: Infer user demand from app behavior (articles read, products viewed/added to babylists) by linking iMamma content to products

STEP 1
Product Categorization
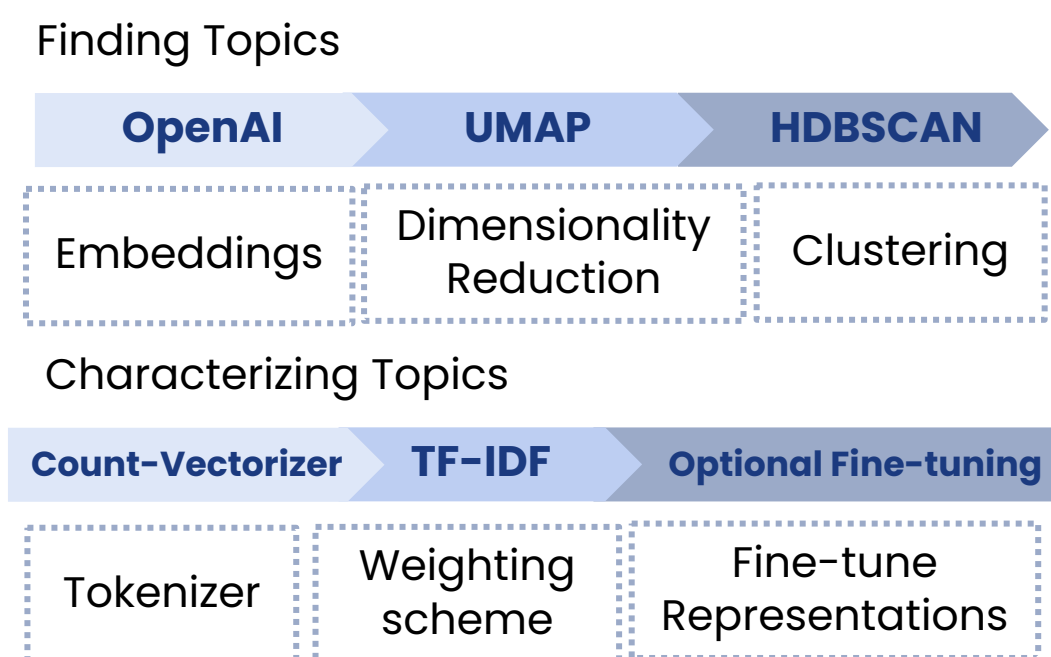
Construct categories of **product substitutes**

Classify each category:
1. Repeat-buy products (e.g. Pampers)
2. Repeat-buy within category (e.g. Newborn Rompers)
3. Non-repeat buy (e.g. Stroller)

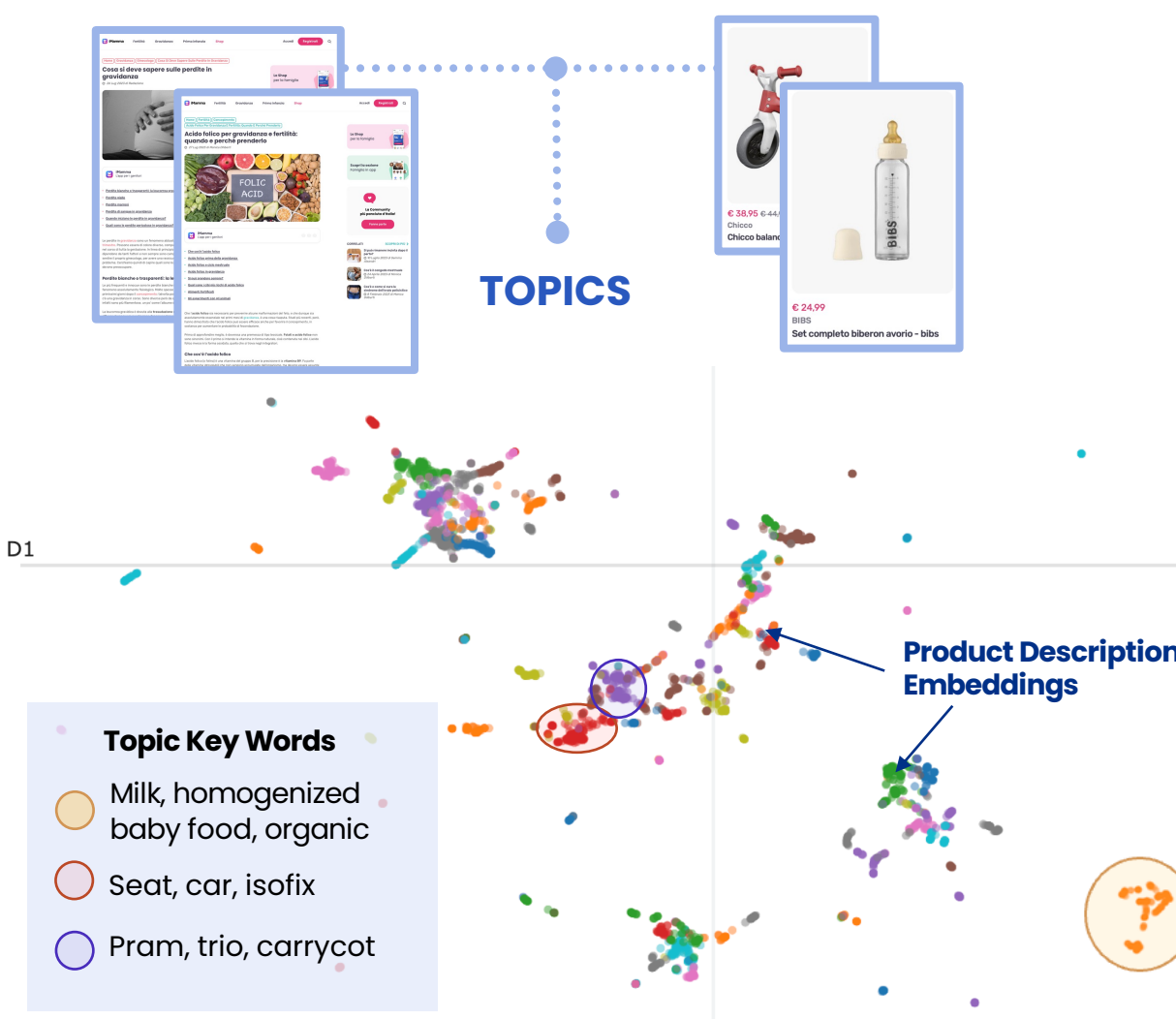<u>Method</u>: **LLMs** and manual brute force

STEP 2
Topic Modeling / Feature Extraction

NLP techniques on **product descriptions to uncover TOPICS**

OUTPUT → Distribution of topics per product

EMBEDDING-BASED APPROACH

Finding Topics

OpenAI → **UMAP** → **HDBSCAN**

Embeddings | Dimensionality Reduction | Clustering

Characterizing Topics

Count-Vectorizer → **TF-IDF** → **Optional Fine-tuning**

Tokenizer | Weighting scheme | Fine-tune Representations

NEXT → Relate Content to Topics via Embeddings

TOPICS

D1

Product Description Embeddings

Topic Key Words
- Milk, homogenized baby food, organic
- Seat, car, isofix
- Pram, trio, carrycot

STEP 3
Demand estimation

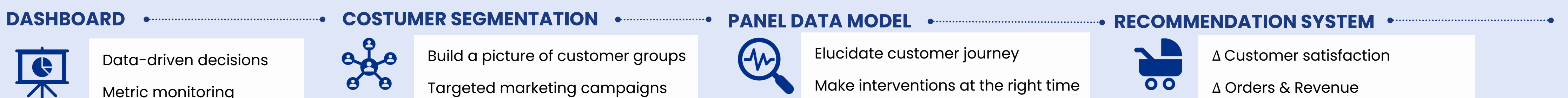Compute product-feature vectors and user-feature vectors

1) User-feature vectors

When a user:
- **Buys/adds to cart/views/adds to babylist** a product
- **Reads an article**

Update implied rating for related topics/features

2) Product-feature vectors
- Topic-affinity
- Price bracket (budget/mid-range/luxury) (from categorization)
- Target age-range

Compute demand via similarity of **user-feature** and **product-feature** vectors

BUSINESS IMPACT

DASHBOARD
- Data-driven decisions
- Metric monitoring

COSTUMER SEGMENTATION
- Build a picture of customer groups
- Targeted marketing campaigns

PANEL DATA MODEL
- Elucidate customer journey
- Make interventions at the right time

RECOMMENDATION SYSTEM
- Δ Customer satisfaction
- Δ Orders & Revenue

Results

Panel Data Model
Out-of-sample results
Accuracy 89.86%   **Recall 64.08%**

Key Insights:
- Users do not order before the child is born
- Users that track their children are more likely to order
- Users are most likely to repeat-order the following week, or 2 months later

Sales Forecasting Model   **Out-of-sample MAPE 3.48%**
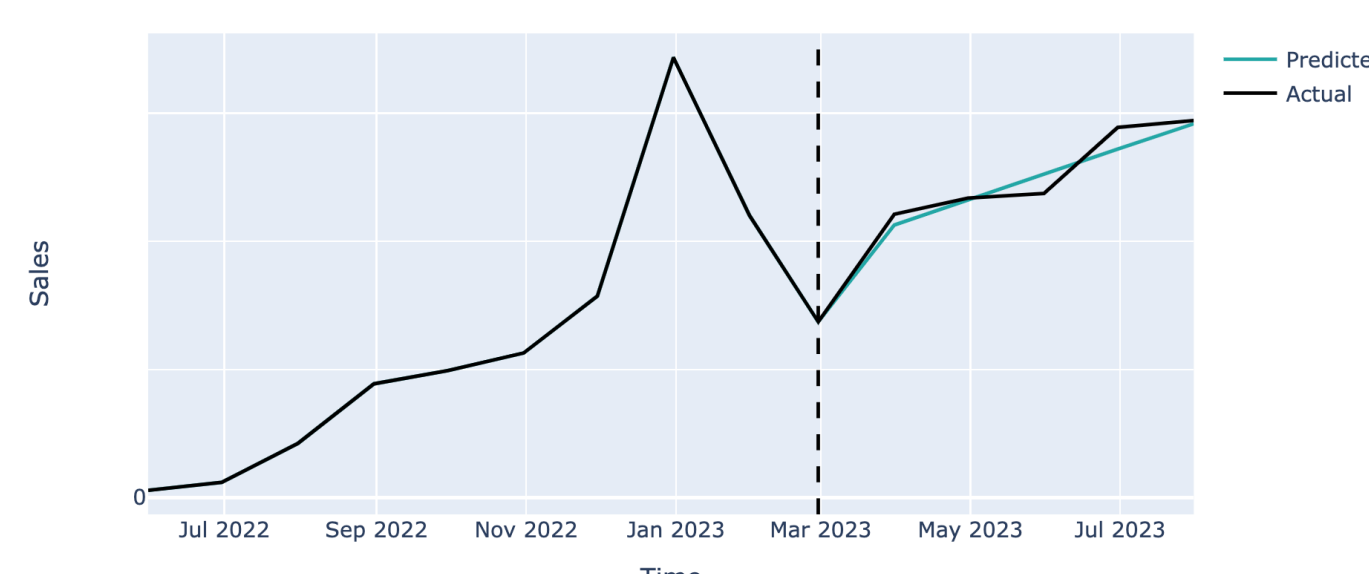
Next Steps

- Link the Dashboard with the **Data Lake**
- Incorporate app web analytics data when it becomes available
- **Implement** the personalized recommendation system, & A/B test vs "most-popular" recommendations to measure performance