mlangles

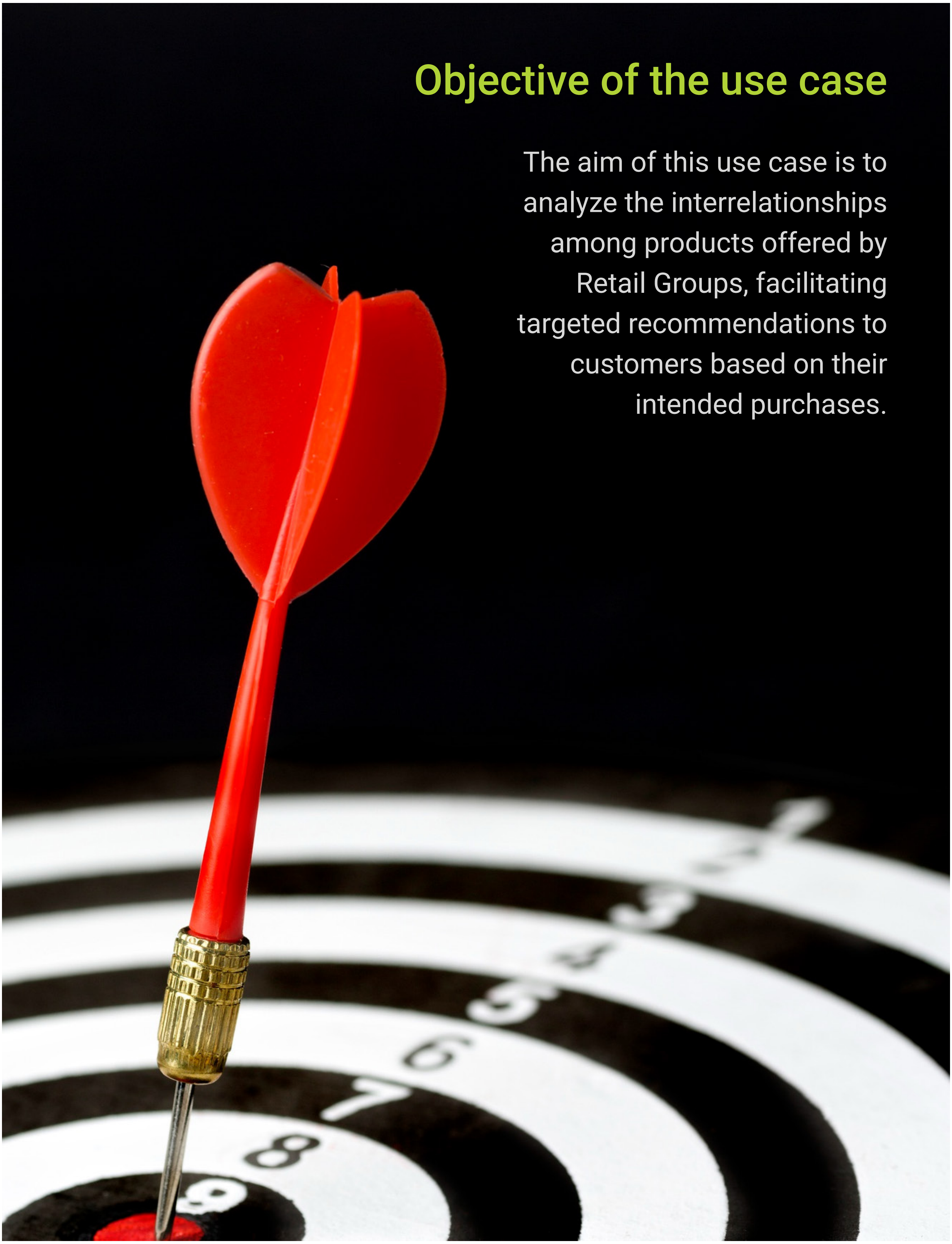
CloudAngles

mlangles Predictive AI

Retail Fashion Recommendations

Objective of the use case

The aim of this use case is to analyze the interrelationships among products offered by Retail Groups, facilitating targeted recommendations to customers based on their intended purchases.



Explanation of the use case

Fashion retail brands and businesses have online market value of 42.80 billion USD and millions of stores worldwide. The online store offers shoppers an extensive selection of products to browse through. But with too many choices, customers might not quickly find what interests them or what they are looking for, and ultimately, they might not make a purchase. To enhance the shopping experience, product recommendations are key. More importantly, helping customers make the right choices also has a positive implication for sustainability, as it reduces returns and thereby minimizes emissions from transportation.

The dataset consists of three files:

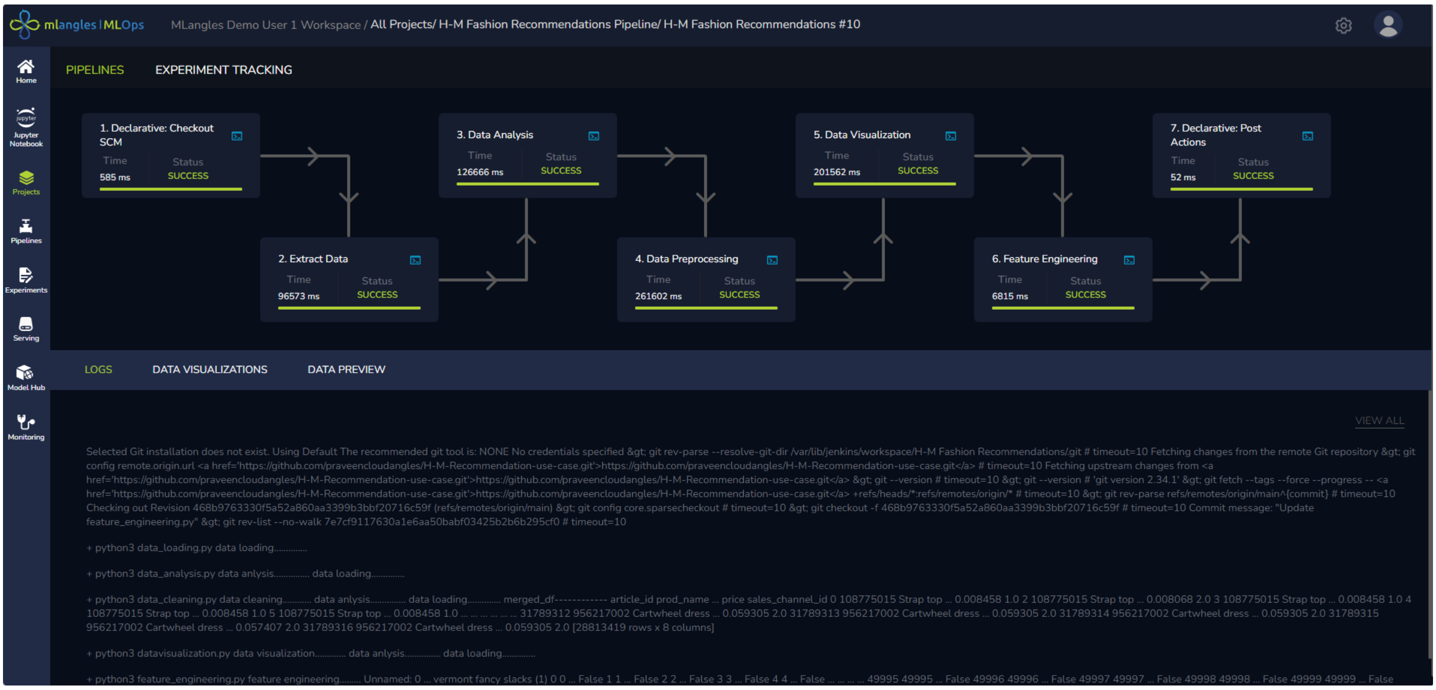
articles.csv: detailed metadata for each article id available for purchase such as product code, product name, product type number, product type name, product group name, colour group code, colour group name.

customers.csv: metadata for each customer id in dataset such as club member status, fashion news frequency, age, postal code

transactions_train.csv : the training data, consisting of the purchases each customer for each date, as well as additional information such as customer_id, article_id, price, sales_channel_id.

Working of the use case

Step 1: Data Engineering and Pipeline Creation



Data Extraction: The data is obtained from a Kaggle competition, and the files are in CSV format.

Data Analysis: The following steps were performed to ensure the data integrity.

- Identify Missing Values: Check for crucial data gaps to maintain accuracy.
- Detect Duplicate Entries: Remove redundant data for consistency.

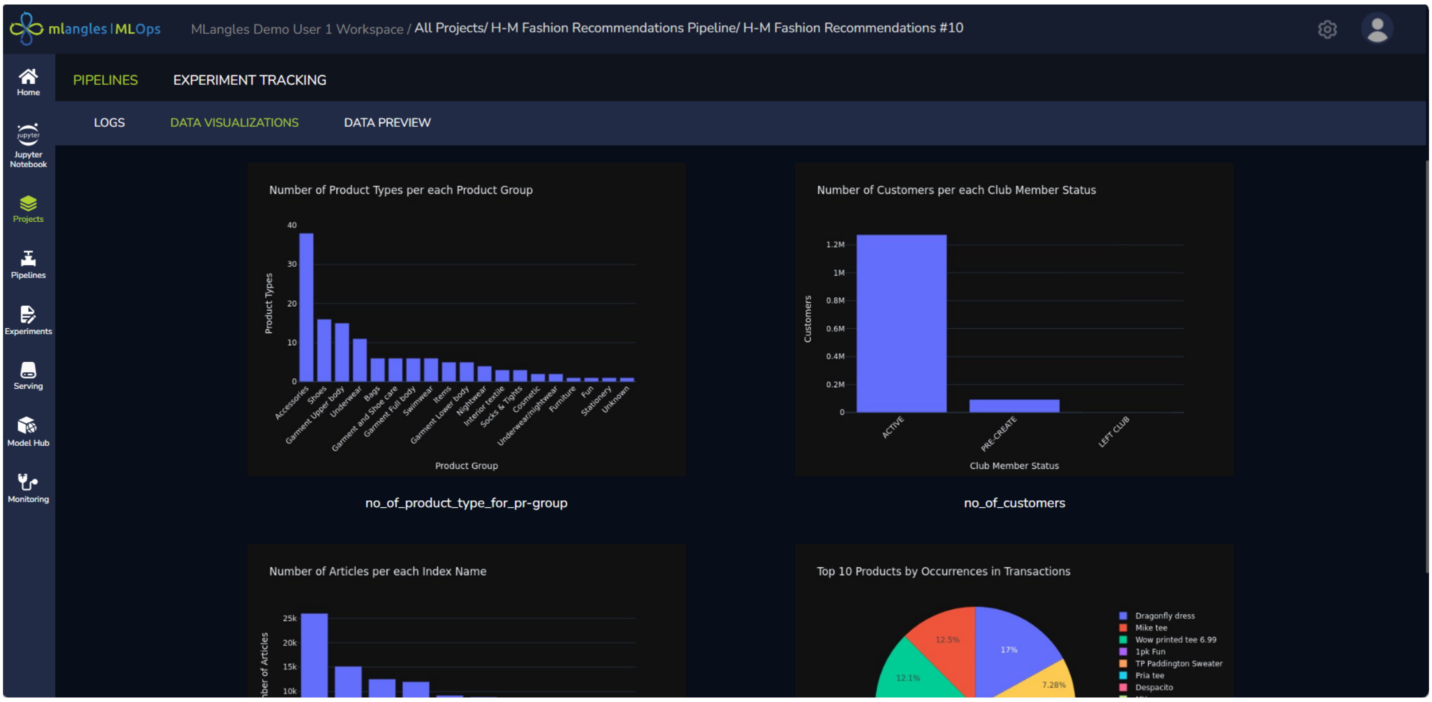
Data Preprocessing: The data in the three files are merged based on the customer ID and article ID so that the resulting data frame consists of the transaction history of the customers.

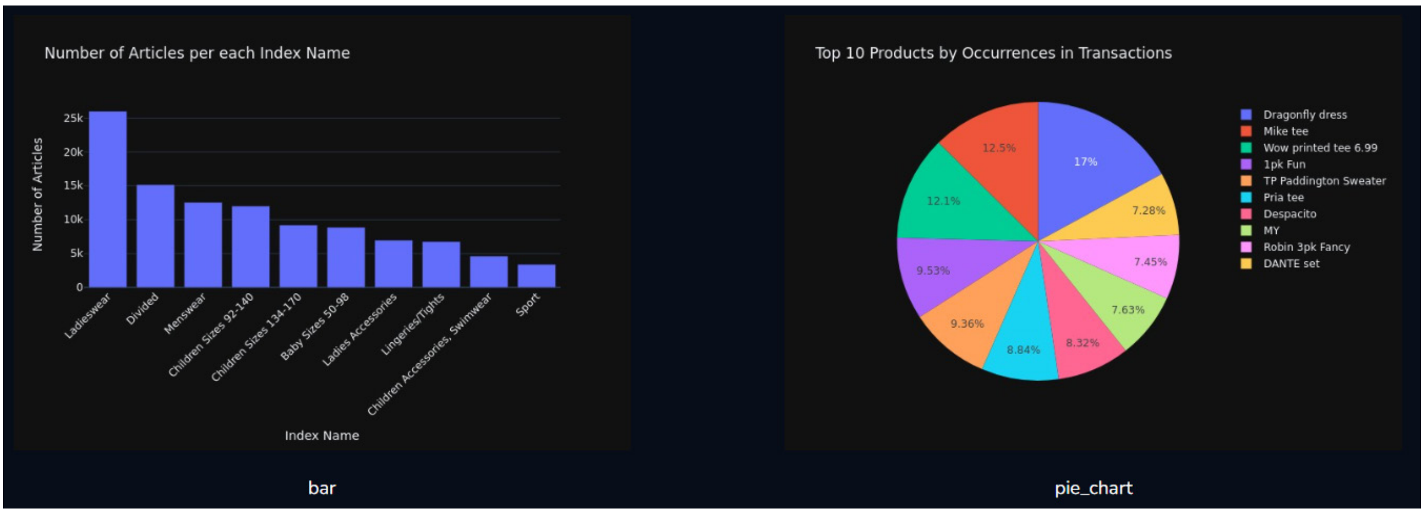
Data Visualization: This step involves representing data graphically to reveal patterns and insights. Here are brief explanations of the visualizations used in this use case.

Histogram: Histograms are graphical representations of the distribution of data. They consist of a series of adjacent rectangles or bars, where the area of each bar represents the frequency or relative frequency of data within a specific range or "bin." Typically, the horizontal axis represents the range of values, divided into intervals, while the vertical axis represents the frequency or relative frequency of occurrences within each interval. Histograms are commonly used in statistics to visualize the shape, central tendency, and variability of data distributions. They are particularly useful for identifying patterns, outliers, and underlying trends in datasets.

Pie charts: A pie chart is a circular statistical graphic divided into slices to illustrate numerical proportions. Each slice represents a proportion of the whole, with the size of each slice corresponding to the magnitude of the proportion it represents. Pie charts are effective for displaying the relative sizes of various categories or parts of a whole. They are commonly used to visualize percentages, proportions, or distributions in data sets, making it easy to understand the composition of a data set briefly. However, they are less effective for comparing individual values or showing trends over time, especially when there are many categories or the differences between categories are small.

The histograms below describe number of product types per each product group and number of articles per each index name whereas the pie chart effectively represents the top 10 products bought by the customers by which it can be noted that dragonfly dress is the most bought at around 17%





Feature Engineering: This step encompasses various tasks aimed at enhancing the quality and relevance of features used in machine learning models. The merged data frame is converted into a tabular format which contains the various possible products as the columns and true/false depending on whether the user bought those products in a transaction which is indicated by the rows.

Below is the image of the cleansed dataset after the data engineering steps.

20 DEN 2P TIGHTS	3P SNEAKER SOCKS	7P BASIC SHAFTLESS	ALEX TRS (J)	BAMA	BARABOOM (1)	BASIC SWEATPANTS	BECKA HOODIE	BIRD TEE	BOWIE	BOX 4P TIGHTS	BRIT BABY TEE	BRITTANY LS	CHARLIE SKIRT	CALISTA (1)	CAT TEE.	CHARLIE TOP	CHARLOTTE BRAZILIAN AZA.LOW 2P	CH SF
false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	true	false	false	false	false	false	false	false	false	false	false	false	false	false	true	fal
false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	false	false	false	false	false	false	false	false	false	false	true	false	false	false	false	fal
false	false	false	false	true	false	false	true	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	false	true	false	false	false	false	false	false	false	true	false	false	false	false	false	fal
true	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	fal
false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	false	true	false	fal

Data Versioning:

- Various processed data versions can be generated through different transformations applied to the same raw dataset, such as deleting columns or applying various transformations on specific columns.
- Throughout the data pipeline, diverse transformations can be executed at each iteration. Consequently, the resulting data at the pipeline's end is systematically versioned.
- Given that each version of the final data is distinct, models trained on these different versions will exhibit varying behaviors.

Step 2: Experiment Tracking - Modelling

After preparing the cleansed data, the next step involves training the model using this cleaned dataset. Given that this is an association rules problem, the most common models that are suitable for the task include the FPGrowth, Hmine, ECLAT and Apriori.

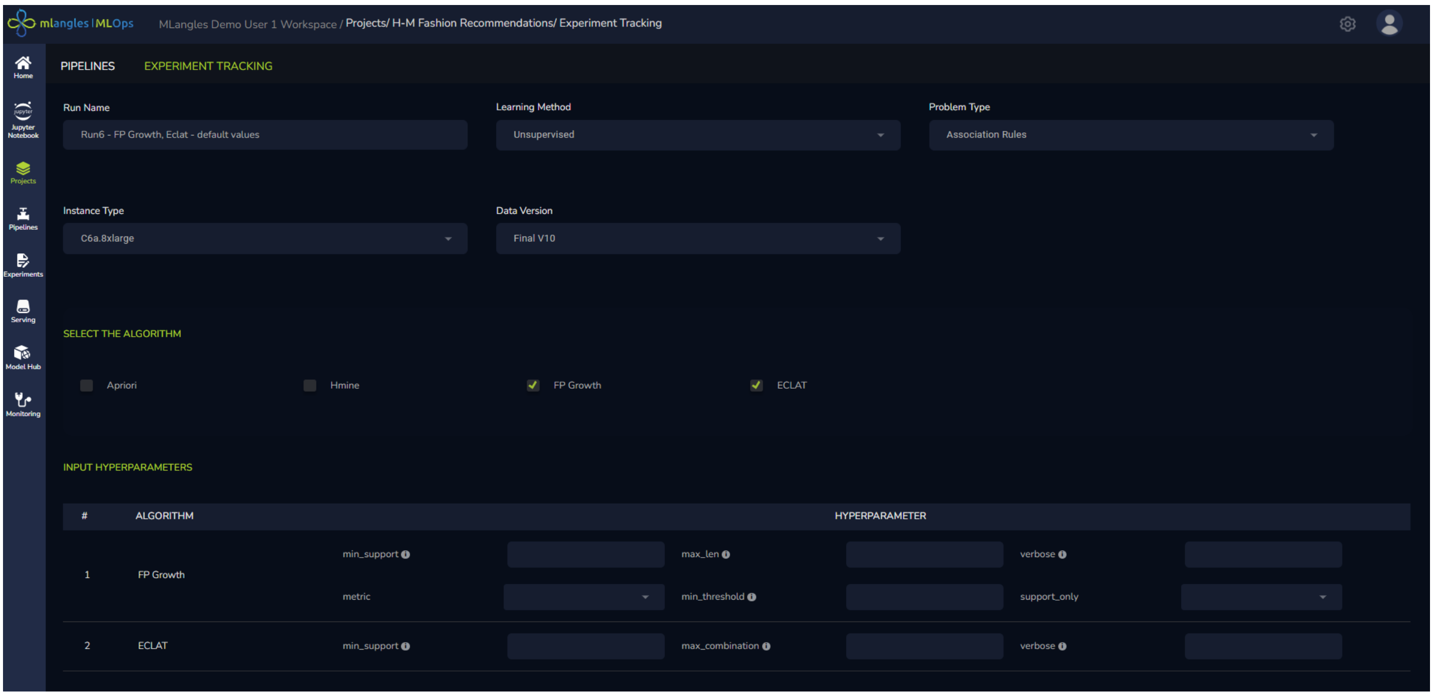
FP-Growth: The FP-Growth algorithm is a popular method for frequent itemset mining in transactional databases. It efficiently discovers frequent itemsets without generating candidate itemsets explicitly. FP-Growth constructs a compact data structure called the FP-tree, which represents the transactions and their itemsets in a compressed form. By recursively mining the FP-tree and its conditional FP-trees, it identifies frequent itemsets with high efficiency.

Hmine: It is a method used for mining high-utility itemsets from transaction databases. H-Mine considers the utility or profit associated with each item in a transaction. By optimizing utility-based measures, H-Mine efficiently discovers itemsets that maximize the overall utility, making it particularly useful in domains where item values vary and transactions involve multiple items with different utilities.

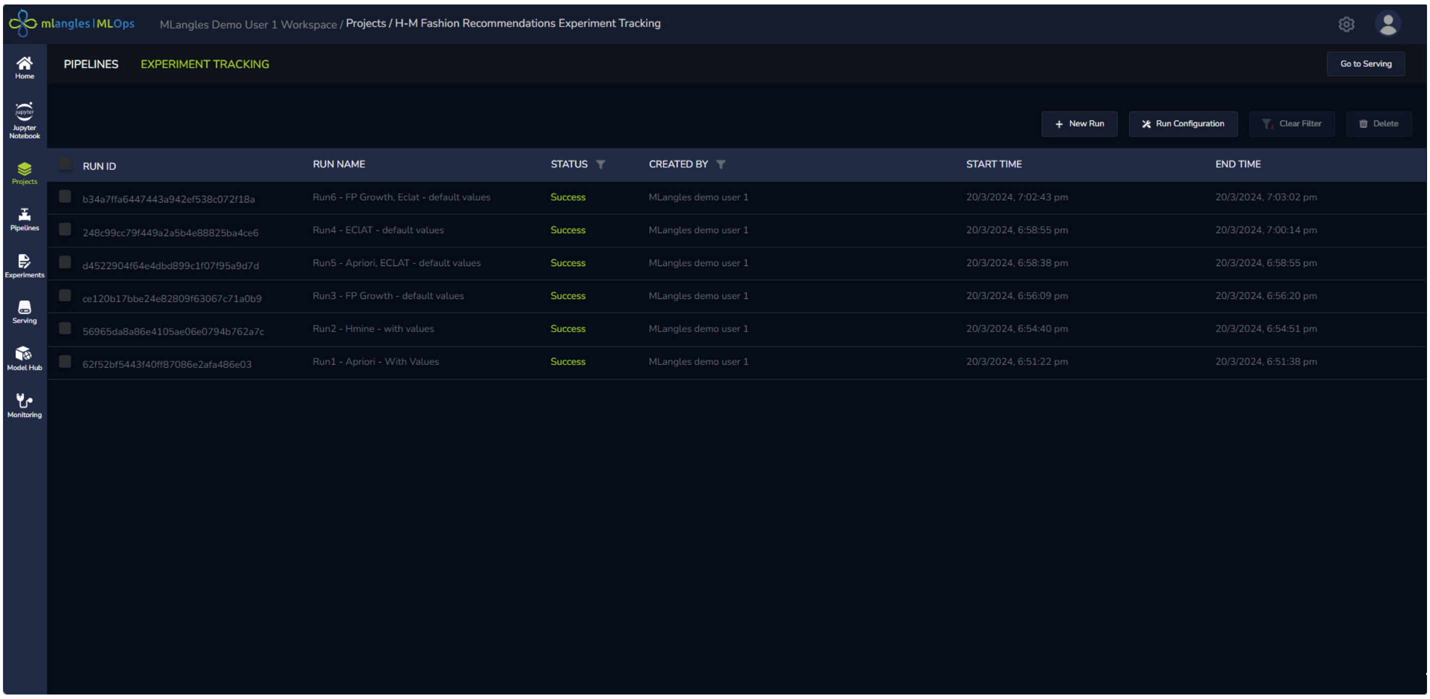
ECLAT (Equivalence Class Clustering and Bottom-Up Lattice Traversal): It is an association rule mining algorithm used to discover frequent itemsets in transaction databases. Unlike the Apriori algorithm, ECLAT does not generate candidate itemsets explicitly. Instead, it utilizes a depth-first search approach and vertical data format to efficiently explore the lattice structure of itemsets. By exploiting equivalence classes and intersecting transactions, ECLAT identifies frequent itemsets and their corresponding support counts.

Apriori: It is a classic approach in data mining for discovering frequent itemsets within transactional databases. It operates based on the "apriori principle," which states that if an itemset is frequent, then all of its subsets must also be frequent. The algorithm works iteratively, starting with finding individual items' frequencies, then generating candidate itemsets of larger sizes, and finally pruning those that do not meet a minimum support threshold. This process continues until no new frequent itemsets can be found.

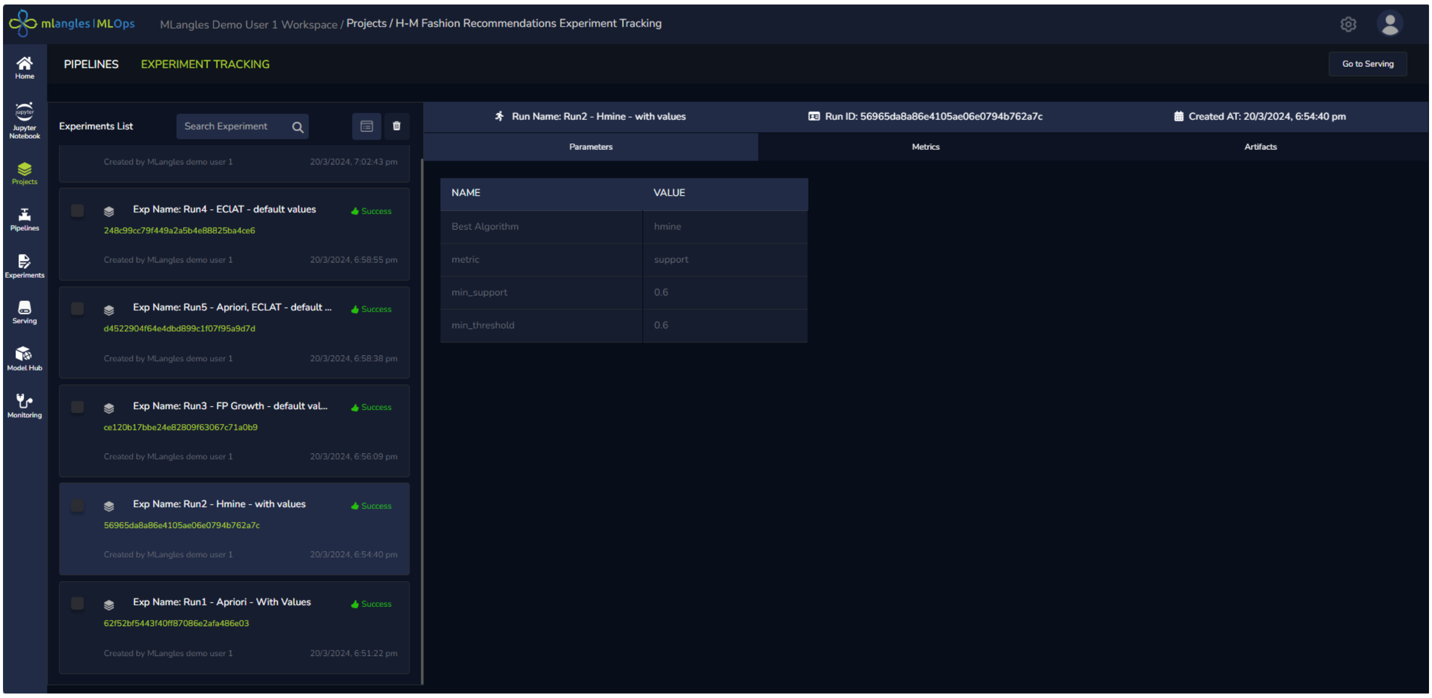
Runs can be created by selecting the appropriate data version. Each of the data version would correspond to a successful data pipeline.



Once the run is created, various details can be extracted, including parameters utilized during training, and metrics and artifacts. These insights provide valuable information about the model's performance and behavior, aiding in understanding its effectiveness and potential areas for improvement.

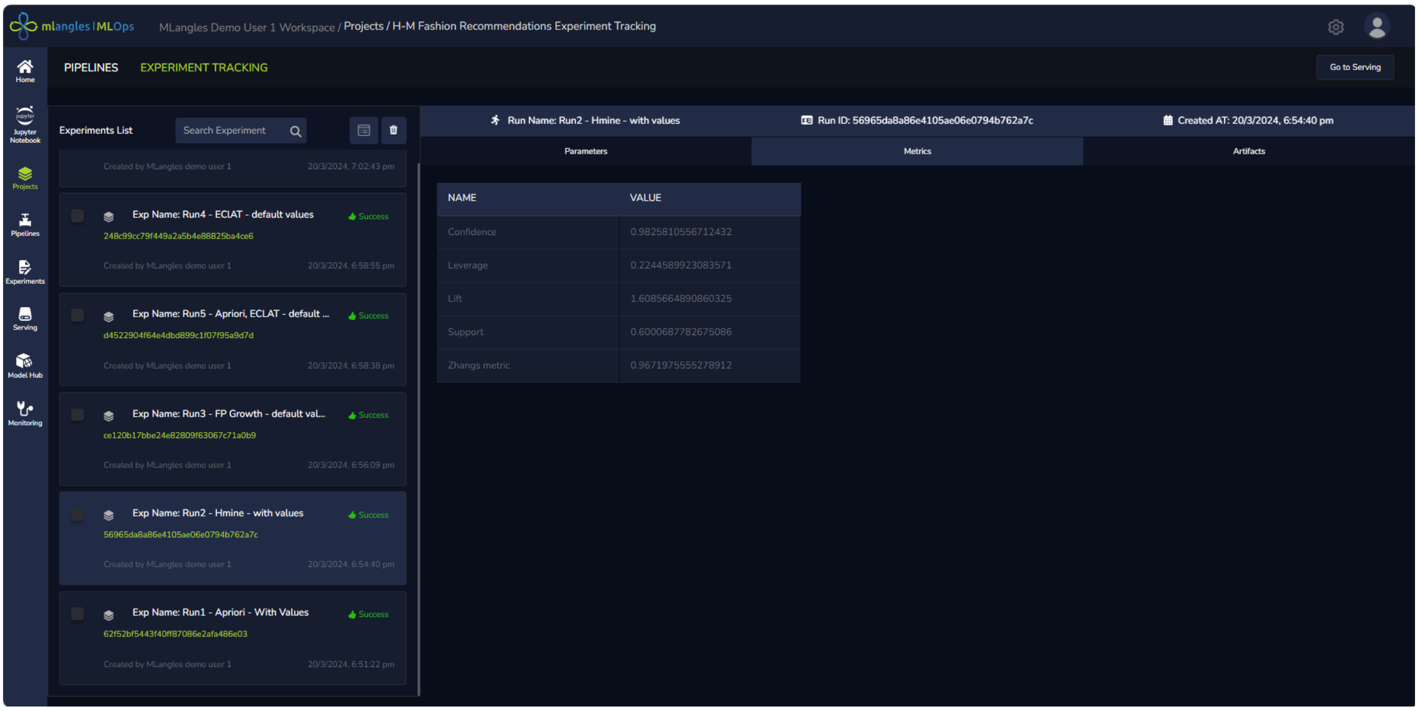


Extracting parameters used during training allows for reproducibility and transparency, ensuring that the model's settings are documented and accessible for future reference.

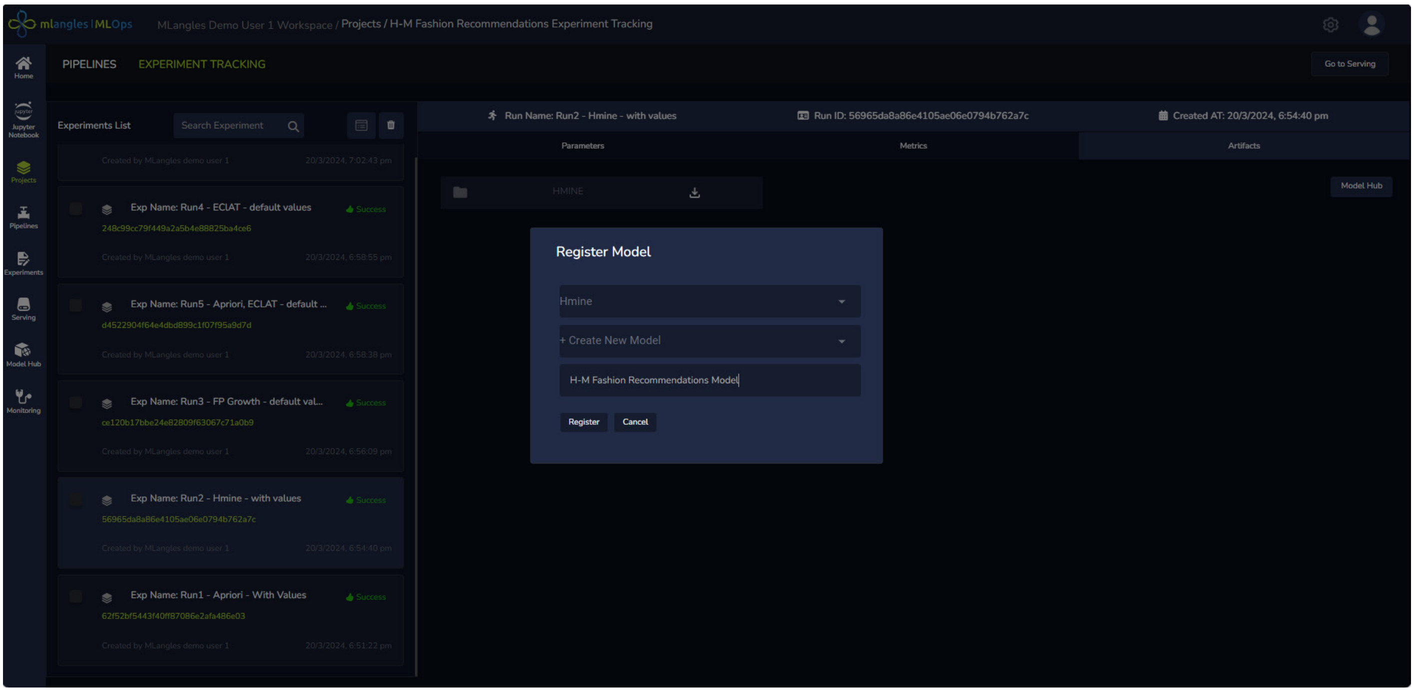


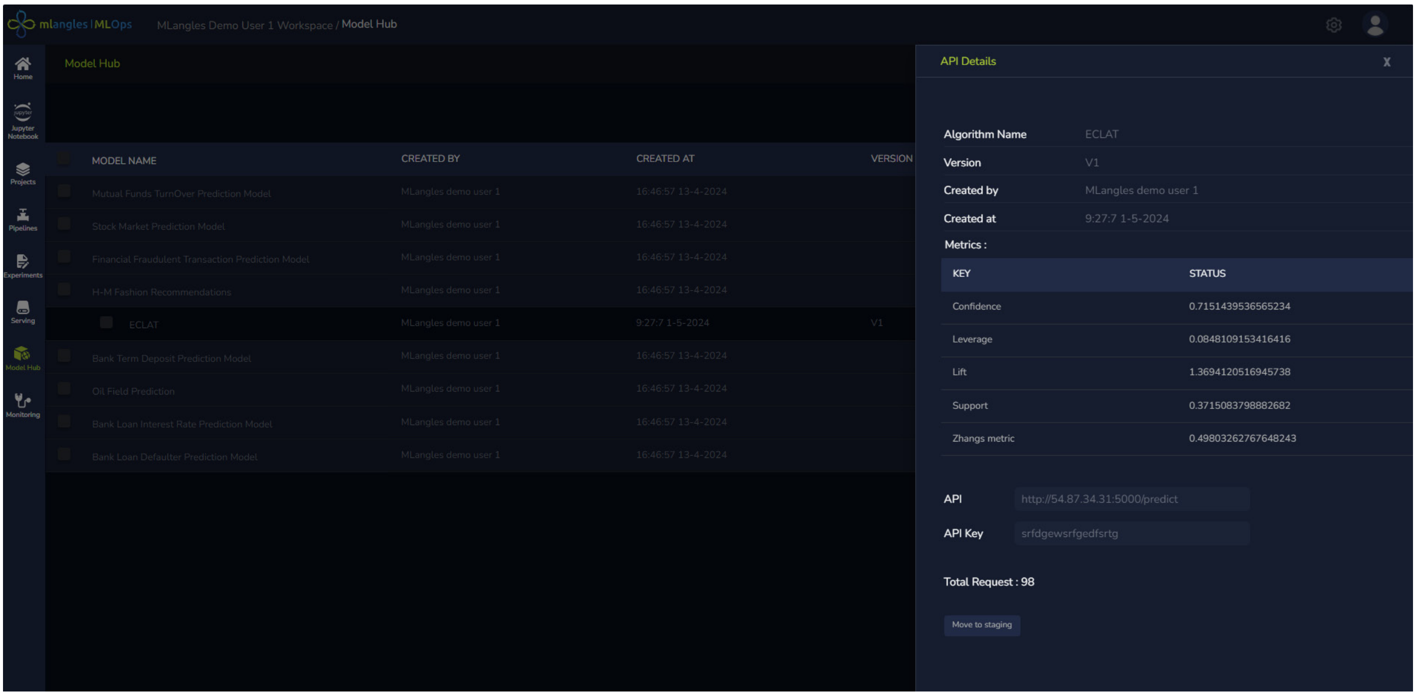
The following are common evaluation metrics used in association tasks to assess the performance of predictive models:

- Confidence measures the likelihood of occurrence of the consequent given the antecedent, providing insight into the strength of relationships between items in a dataset
- Leverage quantifies the difference between the observed frequency of co-occurrence of items in a dataset and the frequency expected under independence, indicating the significance of the association between items.
- Lift measures the degree of dependency between the antecedent and consequent of a rule, indicating how much more likely the consequent is to occur when the antecedent is present compared to its expected occurrence in the absence of the antecedent
- Support quantifies the frequency of occurrence of a specific itemset or association rule within a dataset, providing a measure of how common the association is among transactions.
- Zhang's metric is a statistical measure used to assess the significance of association rules by considering both confidence and statistical significance, providing a balanced evaluation of rule quality.



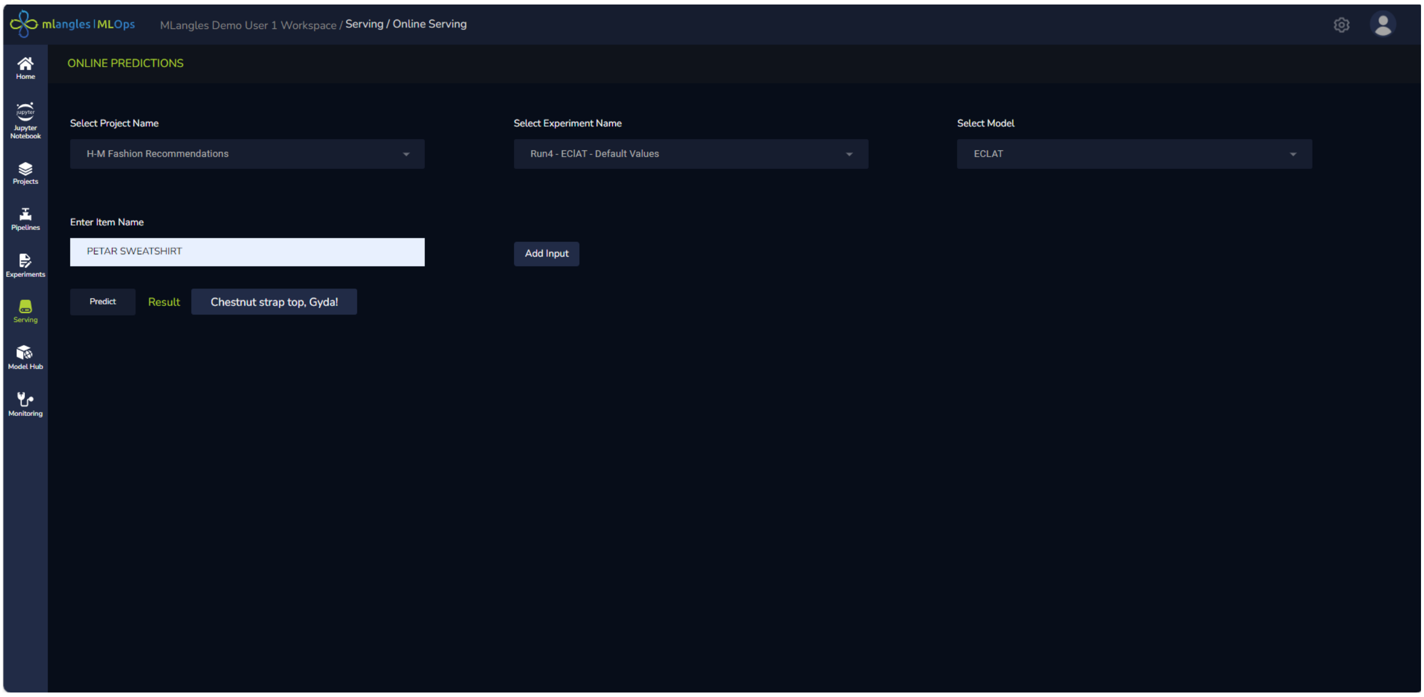
Model Hub: The best model can be pushed to a model hub where it is deployed and exposed as a REST API endpoint. This endpoint allows applications to send new data to the model for making predictions. By integrating the endpoint into applications, users can easily leverage the model's predictive capabilities in their workflows, enabling real-time decision-making based on the model's insights.


mlangles | MLOps
MLangles Demo User 1 Workspace / Projects / H-M Fashion Recommendations Experiment Tracking
PIPELINES
EXPERIMENT TRACKING
Go to Serving
Experiments List
Search Experiment
Run Name: Run2 - Hmine - with values
Run ID: 56965da8a86e4105ae06e0794b762a7c
Created AT: 20/3/2024, 6:54:40 pm
Parameters
Metrics
Artifacts
Model Hub
Exp Name: Run4 - ECLAT - default values
Exp Name: Run5 - Apriori, ECLAT - default ...
Exp Name: Run3 - FP Growth - default val...
Exp Name: Run2 - Hmine - with values
Exp Name: Run1 - Apriori - With Values
Register Model
Hmine
+ Create New Model
H-M Fashion Recommendations Model
Register
Cancel


mlangles | MLOps
MLangles Demo User 1 Workspace / Model Hub
Model Hub
MODEL NAME
CREATED BY
CREATED AT
VERSION
API Details
Algorithm Name ECLAT
Version V1
Created by MLangles demo user 1
Created at 9:27:7 1-5-2024
Metrics :
KEY STATUS
Confidence 0.7151439536565234
Leverage 0.0848109153416416
Lift 1.3694120516945738
Support 0.3715083798882682
Zhangs metric 0.49803262767648243
API http://54.87.34.31:5000/predict
API Key srfdgewsrfgedfsrtg
Total Request : 98
Move to staging

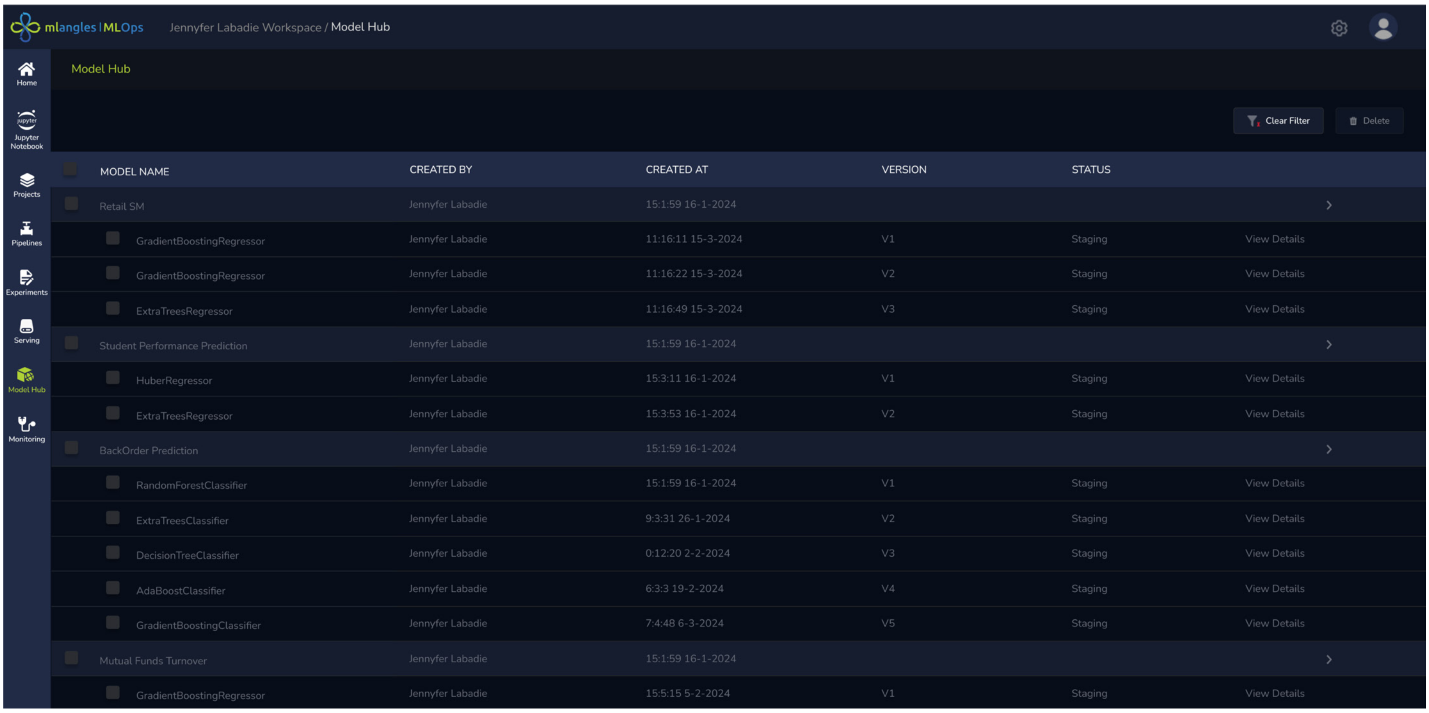
Step 3: Prediction / Serving

During model serving, a product is used as a sanity test for the model, and the set of most recommended products based on that is predicted as the output. Here, the trained ECLAT model recommends Chestnut Strap Top to be bought along with Petar Sweatshirt.



Model Hub:

- Trained models are uploaded to the model hub, whereupon deployment, a REST API endpoint is automatically generated.
- Data is transmitted to this endpoint as a request, triggering the model to execute a prediction and return the output as the response to the request.



Conclusion

In conclusion, the project of generating recommendations based on association rule mining has provided valuable insights into understanding the relationships between the products in the dataset. By leveraging techniques such as the Apriori, FP Growth algorithms and evaluating metrics like confidence, support, and lift, meaningful associations among various products have been identified. These associations enable us to make personalized recommendations to users based on their past behavior or preferences.

To setup a demo

Info.mlangles@cloudangles.com

Visit: www.mlangles.ai