



Project II: Manufacturing & Product Distribution

VLBA II – System Architectures

Magdeburg Research and Competence Cluster
Workgroup Business Informatics I

Faculty of Computer Science
Department of Technical and Business Information Systems

Otto-von-Guericke-University Magdeburg

Kavyashree Byalya Nanjegowda

Vachana Visweswaraiah

Chethan Chinnabhandara Nagaraj

Medini Rachapudi

Vinuthna Batchu



AGENDA

- Introduction
- Detailed Tasks
- Our Approach
- Tools and Technologies Used
- Data Exploration and Cleaning
- Clustering Based on Geographical Location
- Forecasting Product Sales and Materials Required
- Interpretation and Implications of Results



INTRODUCTION

- MWW serves multiple markets with different demands.
- For the future production planning and sales strategy, the profits of the manufactured goods in the past and for the future need to be analyzed.
- Due to potential regional deviations, the forecasts are expected to be presented specifically for each operated market.
- Moreover, the board values rich visualizations of the results.



DETAILED TASKS

- Because of potential regional differences, the existing customer markets need to be clustered based on their geographical location with the help of a machine learning algorithm. Use visualization technologies of the Google Cloud for the presentation of the clusters



DETAILED TASKS

- Based on the identified clusters, select one sensible market and use data analytics and machine learning models to forecast product sales and required materials to assist management in decision-making for production planning, procurement, and marketing. Use MLflow to orchestrate the Data Science lifecycle
- Interpret and summarize the findings and implications for the business in the form of a dashboard and/or report.

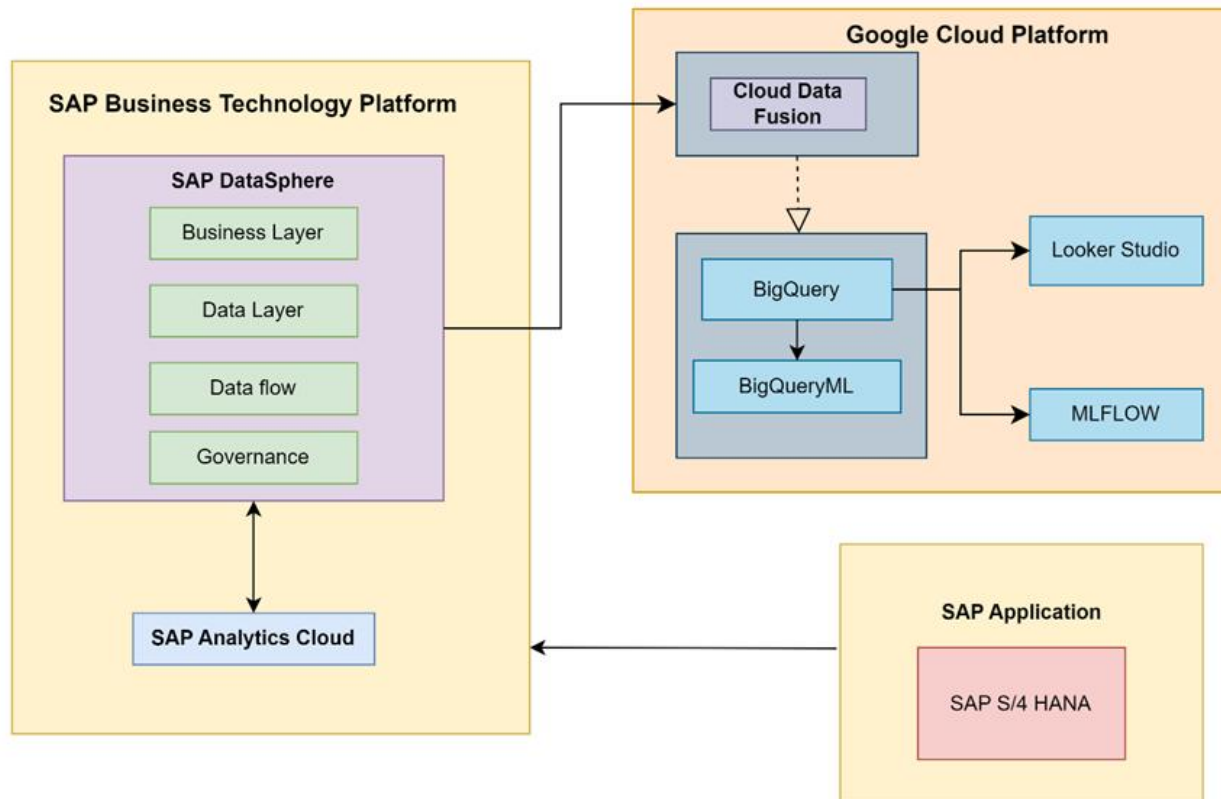


TOOLS AND TECHNOLOGIES USED

- Google Cloud BigQuery
- BigQuery
- BigQuery ML
- MLflow
- Google Cloud Looker Studio

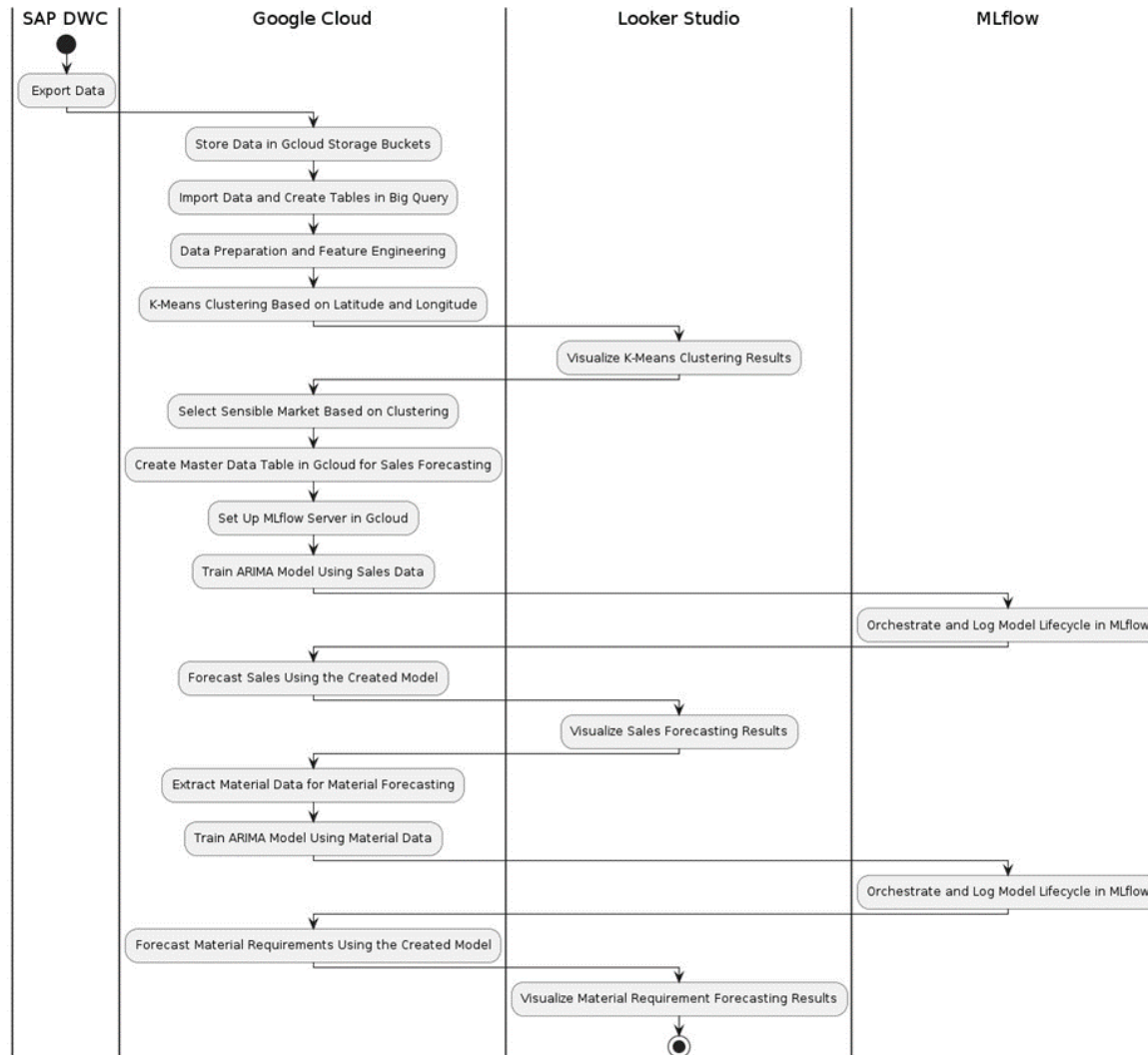


ARCHITECTURE





OUR APPROACH





DATA EXPLORATION AND CLEANING

Given data

- MPD_CustomerAttr: location of customers
- MPD_CustomerText: customer description
- MPD_MaterialAttr: information regarding materials
- MPD_MaterialOrders: past material orders
- MPD_OrderItems: past orders
- MPD_Orders: past orders with customer information
- MPD_ProductAttr: product information
- MPD_ProductMat: required materials for manufacturing
- MPD_ProductText: product names



DATA EXPLORATION AND CLEANING

Data Exploration

- Conducted thorough exploration to understand dataset structure, relationships, and key characteristics.

Data Cleaning

- Imputed mean value for erroneous data in the table MPD-ProductMat for the Amount_Required column
- Corrected city name in the table 'MPD-CustomerAttr' to ensure accurate location-based clustering



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Data Preparation

- Extended the "**MPD_CustomerAttr**" table by adding latitude and longitude columns using the Google Maps Geocoding API for precise geographical analysis.
- The obtained coordinates will be utilized to perform clustering, grouping customers based on geographical proximity for better market segmentation.



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Data Preparation

```
for city in cities:
    geocode_result = gmaps.geocode(city)
    if geocode_result:
        location = geocode_result[0]['geometry']['location']
        results.append({
            'City': city.split(',')[0],
            'Country': city.split(',')[1].strip(),
            'Latitude': location['lat'],
            'Longitude': location['lng']
        })
df = pd.DataFrame(results)
print(df)

df.to_csv('MPD_CustomersAttr_location_Coordinates.csv', index=False)
```



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Data Preparation

```
CREATE OR REPLACE TABLE `vlba-2024-mpd-group-5.MPD_DATA_5.MPD_CustAtt_extended` AS
SELECT
    cust.CustomerId,
    cust.City,
    cust.SalesOrg,
    cust.Country,
    cust.Currency,
    loc.Latitude,
    loc.Longitude,
    ctxt.CustDescr
FROM
    `vlba-2024-mpd-group-5.MPD_DATA_5.MPD-CustomerAttr` AS cust
LEFT JOIN
    `vlba-2024-mpd-group-5.MPD_DATA_5.MPD_CustomersAttr_location_Coordinates` AS loc
ON
    cust.City = loc.City
LEFT JOIN
    `vlba-2024-mpd-group-5.MPD_DATA_5.MPD-CustomerText` AS ctxt
ON
    cust.CustomerId = ctxt.CustomerID;
```



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Clustering

- Algorithm Selection:

Employed K-Means clustering to segment customer markets based on geographical locations using latitude and longitude coordinates, chosen for its simplicity, efficiency, scalability, and interpretability.



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Clustering

- Implementation:

Utilized BigQuery ML to execute K-Means clustering with $k=4$, determined through iterative testing to group customers into four distinct, strategically significant geographical clusters.



CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Clustering

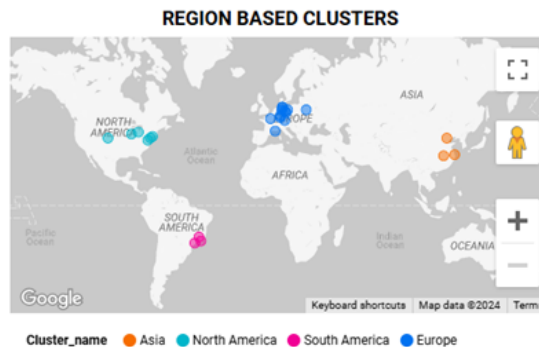
```
1 CREATE OR REPLACE MODEL `vlba-2024-mpd-group-5.MPD_DATA_5.Cust_Markt_clustering`  
2 OPTIONS(  
3   model_type='kmeans',  
4   num_clusters=4,  
5   standardize_features=true  
6 ) AS  
7 SELECT  
8   SAFE_CAST(Latitude AS FLOAT64) AS Latitude,  
9   SAFE_CAST(Longitude AS FLOAT64) AS Longitude  
10 FROM  
11   `vlba-2024-mpd-group-5.MPD_DATA_5.MPD_CustAtt_extended`;  
12
```



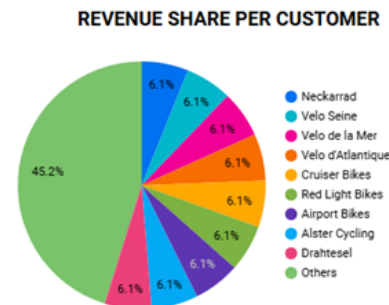
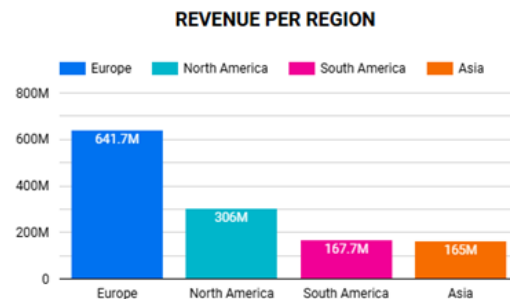
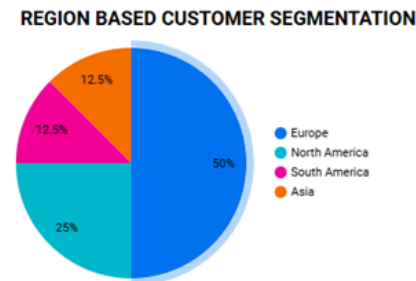

CLUSTERING BASED ON GEOGRAPHICAL LOCATION

Visualization in Looker Studio

Geographical Customer Segmentation



CUSTOMER COUNT
Record Count
24





FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Sensible Market Selection

- Data-driven approach that considers two metrics:
 - Customer Reach
 - Revenue Potential
- Custom queries to calculate the metrics



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Sensible Market Selection

- Customer Reach

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS
Row	Region ▼	TotalCustomers ▼			
1	South America	3			
2	Asia	3			
3	Europe	12			
4	North America	6			



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Sensible Market Selection

- Revenue Potential

JOB INFORMATION		RESULTS	CHART	JSON
Row //	Region ▼ //	TotalRevenue ▼ //		
1	South America	167685012.9099...		
2	Asia	165033634.1300...		
3	Europe	641657746.1000...		
4	North America	306039037.1		



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Sensible Market Selection

- Based on the results of our analysis, we selected the **Europe** cluster as it had the highest number of customers and generated the highest revenue.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

MLflow Orchestration

- Set Up MLflow Server with SQLite backend store
- Logging and Tracking of
 - Model
 - Model Parameters
 - Metrics
 - Forecasted Values



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

MLflow Orchestration

The screenshot displays the MLflow Experiments interface in a web browser. The URL bar shows "34.141.32.30:5000/#/experiments/4". The interface has a dark blue header with the MLflow logo and version "1.30.1". Below the header, there are tabs for "Experiments" and "Models". The "Experiments" tab is active, showing a list of experiments on the left and details for "MPD_Analysis_Experiment" on the right. The experiment details include a description, experiment ID (4), and a table of runs. The table shows one matching run with the following details:

Created	Duration	Run Name	User	Source	Version	Models	metrics.mean	materials_mean	materials_mean	Countries	Time Series	Training End Ver	Desc
1 day ago	1.3min	Sales_Foreca...	root	collab_te...	-	siklearn	95.45	-45.05	0.239	[DE, 'FR]	Year	2020	Forec...



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

MLflow Orchestration

The screenshot displays the MLflow web interface for an experiment named 'Sales_Forecasting_Workflow'. The interface includes a navigation bar with 'Experiments' and 'Models' tabs. The main content area shows experiment details: Run ID (db7f377cf748485c97d6a32cc23e1e42), Date (2024-06-30 15:25:28), Source (colab_kernel_launcher.py), User (root), Duration (1.3min), and Status (FINISHED). A sidebar on the left lists the experiment's lifecycle stages: Description, Parameters (6), Metrics (12), Tags (1), and Artifacts. The 'Artifacts' section is expanded, showing a file tree with 'data', 'forecast', 'models', and 'training_evaluation'. The 'models' folder is selected, displaying the 'MLflow Model' details. This section includes a 'Model schema' table (currently empty with a note to see MLflow docs) and a 'Load the model' code snippet. The code snippet shows how to load the model using the MLflow Python API. Below the code, there are links to documentation for batch or real-time scoring using the Pyfunc model flavor.

mlflow 1.30.1 Experiments Models GitHub Docs

MPD_Analysis_Experiment > Sales_Forecasting_Workflow

Sales_Forecasting_Workflow

Run ID: db7f377cf748485c97d6a32cc23e1e42 Date: 2024-06-30 15:25:28 Source: colab_kernel_launcher.py

User: root Duration: 1.3min Status: FINISHED

Lifecycle Stage: active

- > Description Edit
- > Parameters (6)
- > Metrics (12)
- > Tags (1)
- > Artifacts

Full Paths: /mpd-bucket-group-5/4/db7f377cf748485c97d6a32cc23e1e42/artifacts/models

MLflow Model

The code snippets below demonstrate how to load the logged model. You can also register it to the model registry to version control

Model schema

Input and output schema for your model. [Learn more](#)

Name	Type
No schema. See MLflow docs for how to include input and output schema with your model.	

Load the model

```
import mlflow
logged_model = "runs:/db7f377cf748485c97d6a32cc23e1e42/models"

# Load model
loaded_model = mlflow.sklearn.load_model(logged_model)
```

See the documents below to learn how to customize this model and deploy it for batch or real-time scoring using the pyfunc model flavor.

- [API reference for the mlflow.pyfunc module](#)
- [Creating custom Pyfunc models](#)



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Data Preparation

- Sales Data Extraction and Joining:
 - Extracted relevant sales data from Google BigQuery.
 - Performed join operations on CustomerAttr, Orders, OrderItems, and ProductText tables based on user-specified countries.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Data Preparation

- Materials Data Preparation:
 - Joined the sales data (sales_table) with the ProductMat table.
 - Created a new attribute total_materials_required by multiplying Lot_size and Amount_Required.
- Utilized MLflow for centralized management and reproducibility of prepared data artifacts.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Model Training and Evaluation

- Used ARIMA PLUS for product and materials forecast
- Chosen for its ability to capture and forecast time-series patterns like seasonality and trends in sales and materials consumption data.
- Used Google BigQuery ML for scalable and efficient data processing, ensuring accurate and timely forecasts.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Model Training and Evaluation

- Models are trained on the previously prepared data for both Sales Forecast and Material Forecast

```
def train_model_sales(countries, start_year, end_year):  
    countries_str = ', '.join(f'"{country}"' for country in countries)  
    bq_query = f"""  
    CREATE OR REPLACE MODEL `vlba-2024-mpd-group-5.MPD_DATA_5.MPD_Product_Sales_Forecast1`  
    OPTIONS(  
        model_type='ARIMA_PLUS',  
        auto_arima=True,  
        time_series_data_col='total_sales',  
        time_series_timestamp_col='Year',  
        time_series_id_col='ProdDescr'  
    ) AS  
    WITH sales_data AS (  
        SELECT  
            DATE_TRUNC(Date, YEAR) AS Year,  
            SUM(SalesQuantity) AS total_sales,  
            ProdDescr  
        FROM  
            `vlba-2024-mpd-group-5.MPD_DATA_5.TASK2_MPD_DATA_FOR_SALES_FORECAST`
```



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Model Training and Evaluation

- Models are trained on the previously prepared data for both Sales Forecast and Material Forecast

```
def train_model_mat(countries, start_year, end_year):  
  
    countries_str = ', '.join(f'{{country}}' for country in countries)  
  
    bq_query1 = f"""CREATE OR REPLACE MODEL `vlba-2024-mpd-group-5.MPD_DATA_5.Material_Forecast`  
OPTIONS(  
    model_type='ARIMA_PLUS',  
    auto_arma=True,  
    time_series_data_col='total_materials_required',  
    time_series_timestamp_col='Year',  
    time_series_id_col='MatId'  
) AS  
WITH material_data AS (  
    SELECT  
        DATE_TRUNC(Date, YEAR) AS Year,  
        SUM(Total_materials_Required) AS total_materials_required,  
        MatId  
    FROM  
        `vlba-2024-mpd-group-5.MPD_DATA_5.TASK2_MPD_DATA_FOR_MATERIAL_FORECAST`
```



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Model Training and Evaluation

Used the in-built ML.EVALUATE() function to assess model performance

```
#3: Model Evaluation
def evaluate_model_sales():
    eval_query1 = """
    Select
    *
    from
        ML.EVALUATE(MODEL `vlba-2024-mpd-group-5.MPD_DATA_5.MPD_Product_Sales_Forecast1`);
    """
    eval_job = client.query(eval_query1)
    eval_result = eval_job.result()
    eval_df = eval_result.to_dataframe()
    eval_df.to_csv('/tmp/sales_training_evaluation.csv', index=False)
    mlflow.log_artifact('/tmp/sales_training_evaluation.csv', artifact_path="training_evaluation")
    return eval_df

def evaluate_model_mat():
    eval_query = """
    Select
    *
    from
        ML.EVALUATE(MODEL `vlba-2024-mpd-group-5.MPD_DATA_5.Material_Forecast`);
    """
    eval_job = client.query(eval_query)
    eval_result = eval_job.result()
    eval_df = eval_result.to_dataframe()
    eval_df.to_csv('/tmp/material_training_evaluation.csv', index=False)
    mlflow.log_artifact('/tmp/material_training_evaluation.csv', artifact_path="training_evaluation")
    return eval_df
```



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Deployment and Monitoring

- BigQuery ML's ARIMA_PLUS model automatically handles hyperparameter tuning.
- Once the ARIMA_PLUS model is trained using BigQuery ML, it is ready for use in forecasting without additional deployment steps.
- Streamlines the process, making it more efficient and user-friendly.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Deployment and Monitoring

- After training, evaluating the models and validating the forecast results, we registered and logged them to the MLflow server.
- Monitoring with MLflow: Log performance metrics and hyperparameters of ARIMA PLUS models for product sales and materials.

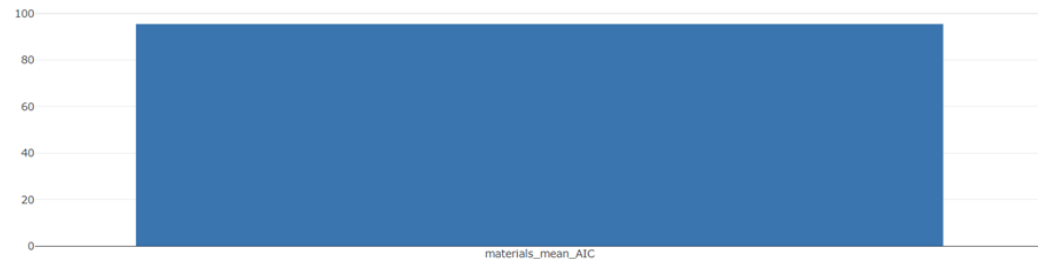


FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Deployment and Monitoring

✓ Metrics (12)

Name	Value
materials_mean_AIC 📈	95.45
materials_mean_log_likelihood 📈	-45.05
materials_mean_non_seasonal_d 📈	0.239
materials_mean_non_seasonal_p 📈	0.408
materials_mean_non_seasonal_q 📈	0.408
materials_mean_variance 📈	8094198.1
sales_mean_AIC 📈	111.1
sales_mean_log_likelihood 📈	-52.6
sales_mean_non_seasonal_d 📈	0.103
sales_mean_non_seasonal_p 📈	0.655
sales_mean_non_seasonal_q 📈	0.31
sales_mean_variance 📈	4074.3



Metric	Latest	Min	Max
materials_mean_AIC	95.45 (step=0)	95.45 (step=0)	95.45 (step=0)



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

Implemented Flexible Forecasting Approaches:

- **Forecast by Period:** Predict values for a user-defined number of years for both Product Sales and Materials.
- **Forecast for a Particular Year:** Predict values for a user-specified year for both Product Sales and Materials.

Results are saved as CSV files and logged as artifacts in MLflow.



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

User Inputs

Enter countries (comma-separated): DE, FR

Enter training start year: 2010

Enter training end year: 2020

Enter forecast period (in years): 7

Enter a specific year for forecast: 2026



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

Sales Product Forecast for specific period

ProdDescr	forecast_timestamp	forecast_value
Air Pump	2021-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2022-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2023-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2024-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2025-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2026-01-01 00:00:00+00:00	1893.3000000000002
Air Pump	2027-01-01 00:00:00+00:00	1893.3000000000002



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

Materials Forecast for a specific period

Full Path:gs://mpd-bucket-group-5/4/bdd1b82317ab4716a94f22c09945130f/artifacts/forecast/materials_forecast_for_period_7_years.csv

Size: 85.04KB

Previewing the first 500 rows

ProdDescr	MatId	Year	total_materials_required
Air Pump	PLAS1000	2021-01-01 00:00:00+00:00	184437.0
Air Pump	PLAS1000	2022-01-01 00:00:00+00:00	184437.0
Air Pump	PLAS1000	2023-01-01 00:00:00+00:00	184437.0
Air Pump	PLAS1000	2024-01-01 00:00:00+00:00	184437.0
Air Pump	PLAS1000	2025-01-01 00:00:00+00:00	184437.0



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

Sales Product Forecast for a specific year

ProdDescr	forecast_timestamp	forecast_value
Air Pump	2026-01-01 00:00:00+00:00	1893.3000000000002
Alu Scoot	2026-01-01 00:00:00+00:00	1534.4
Deluxe Touring Bike-Black	2026-01-01 00:00:00+00:00	601.8000000000001
Deluxe Touring Bike-Red	2026-01-01 00:00:00+00:00	409.9662039841683
Deluxe Touring Bike-Silver	2026-01-01 00:00:00+00:00	1140.9
E-Bike Tailwind	2026-01-01 00:00:00+00:00	336.44545455982467



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

Forecasting

Materials Forecast for a Specific year

ProdDescr	MatId	Year	total_materials_required
Air Pump	PLAS1000	2026-01-01 00:00:00+00:00	184437.0
Air Pump	VENT1000	2026-01-01 00:00:00+00:00	221812.5
Alu Scoot	BATT1000	2026-01-01 00:00:00+00:00	428.2343659026982
Alu Scoot	SCLE1000	2026-01-01 00:00:00+00:00	290.3
Alu Scoot	SCRA1000	2026-01-01 00:00:00+00:00	580.6
Alu Scoot	SSEA1000	2026-01-01 00:00:00+00:00	320.33708557449404



FORECASTING PRODUCT SALES AND MATERIALS REQUIRED

MLflow Logged Artifacts

▼ Artifacts

- ▼ data
 - materials_table.csv
 - sales_table.csv
- ▼ forecast
 - materials_forecast_for_period_7_years.csv
 - materials_forecast_for_year_2026.csv
 - sales_forecast_for_period_7_years.csv
 - sales_forecast_for_year_2026.csv
- ▼ models
 - ▶ metadata
 - MLmodel
 - conda.yaml
 - model.pkl
 - python_env.yaml
 - requirements.txt
- ▼ training_evaluation
 - material_training_evaluation.csv
 - sales_training_evaluation.csv

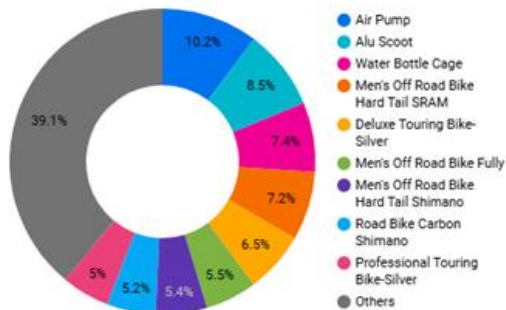


INTERPRETATION AND IMPLICATIONS OF RESULTS

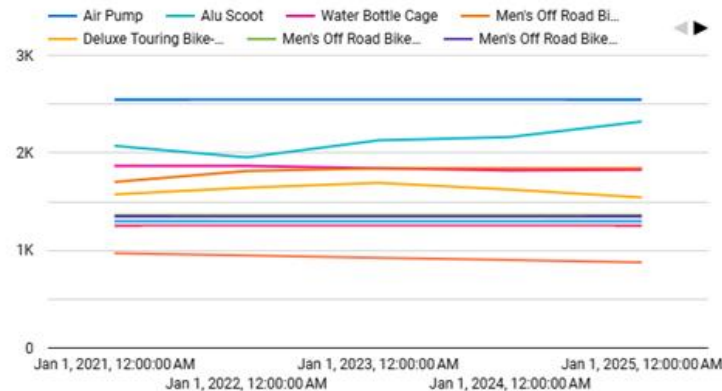
Product Sales Forecast

PRODUCT SALES FORECAST REPORT

PRODUCT SALES FOR 5 YEARS



5 YEAR SALES FORECAST PER PRODUCT



SOLD UNITS PER PRODUCT

forecast_value
124,836.24

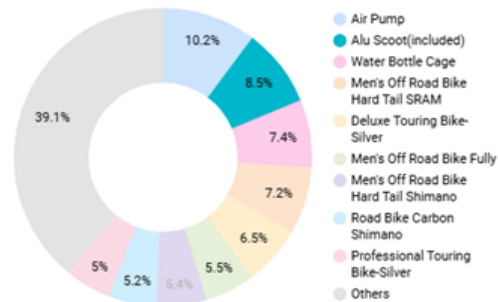


INTERPRETATION AND IMPLICATIONS OF RESULTS

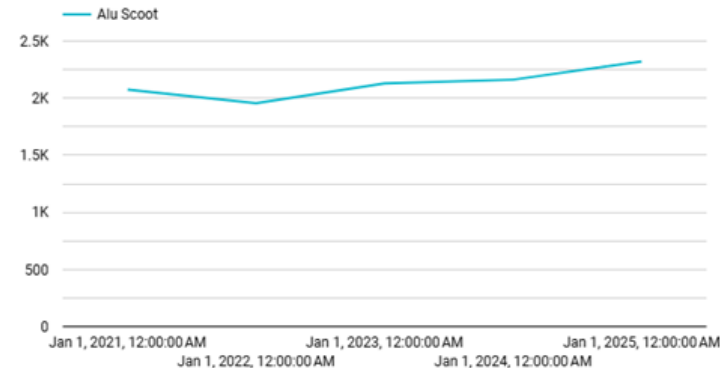
Product Sales Forecast

PRODUCT SALES FORECAST REPORT

PRODUCT SALES FOR 5 YEARS



5 YEAR SALES FORECAST PER PRODUCT



SOLD UNITS PER PRODUCT

forecast_value

10,638.9



INTERPRETATION AND IMPLICATIONS OF RESULTS

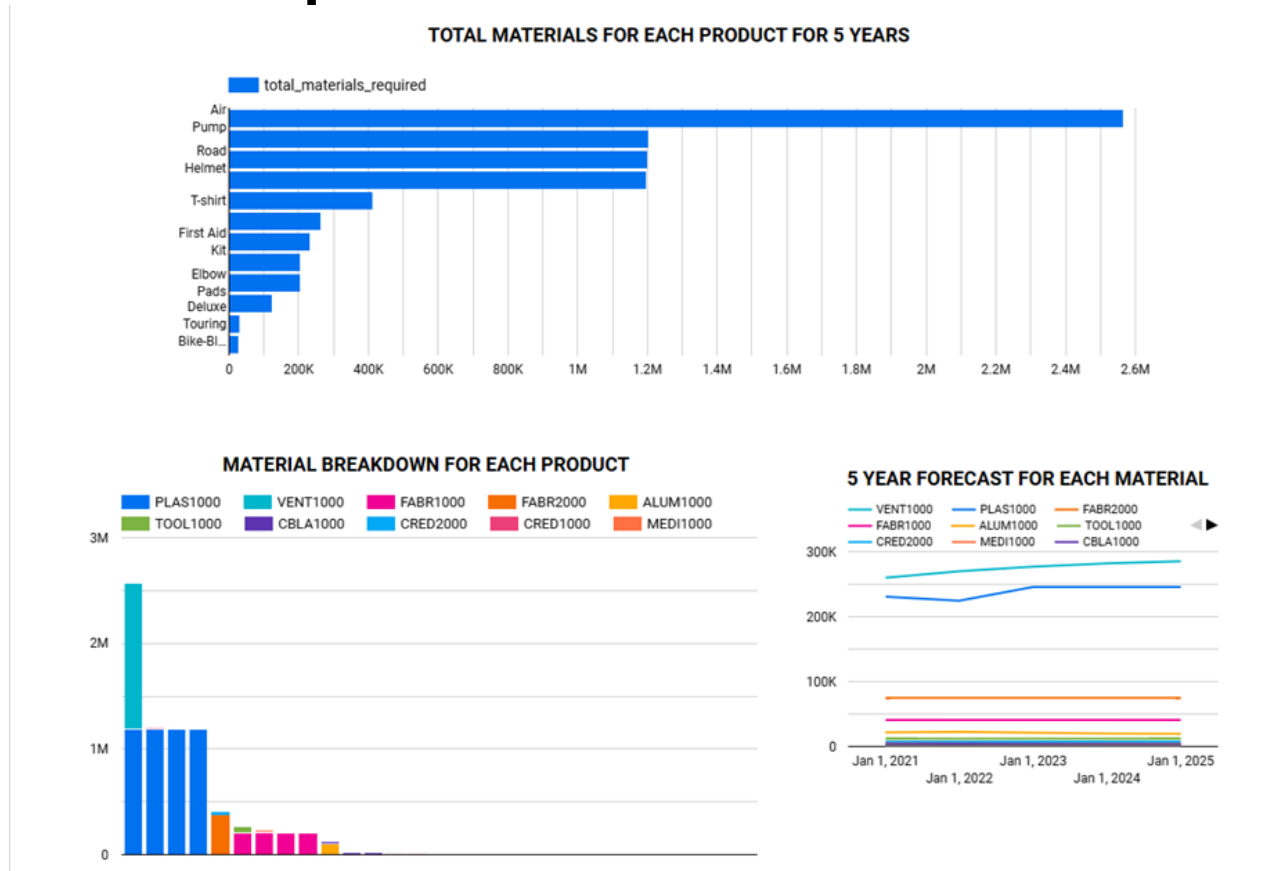
Business Implications

- Product Strategy - With 39.1% of sales in "Others," diversify and innovate continuously.
- Market Demand - Focus on growing products like "Air Pump" and "Alu Scoot"
- Sales Forecasting - Use sales trends to optimize inventory and review declining products like "Deluxe Touring Bike-Silver."



INTERPRETATION AND IMPLICATIONS OF RESULTS

Materials Required Forecast

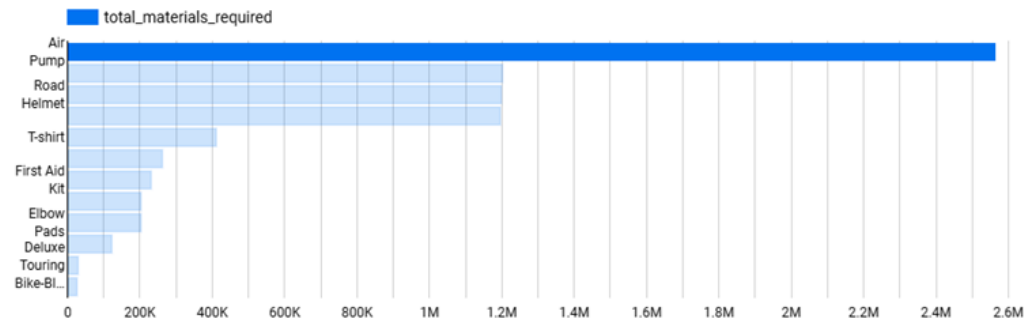




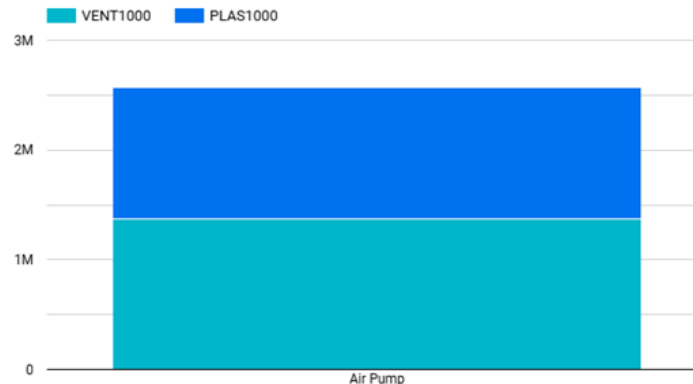
INTERPRETATION AND IMPLICATIONS OF RESULTS

Materials Required Forecast

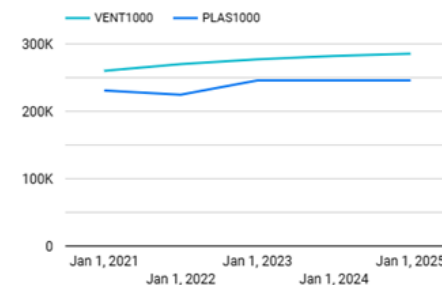
TOTAL MATERIALS FOR EACH PRODUCT FOR 5 YEARS



MATERIAL BREAKDOWN FOR EACH PRODUCT



5 YEAR FORECAST FOR EACH MATERIAL





INTERPRETATION AND IMPLICATIONS OF RESULTS

Business Implications

- Resource Allocation - Ensure a steady supply of VENT1000 and PLAS1000 for Air Pumps and Road Helmets.
- Cost Forecasting - With increased use of VENT1000, adjust budgets to accommodate rising costs.



THANK YOU