**NAAN MUDHALVAN**

**PRODUCT DEMAND PREDICTION**

**Analyzing product demand in the final phase is done using the followingtechniques:**

**Data Collection**: Gather historical sales data, customer feedback, market trends, and any other relevant information.

**Data Preprocessing**: Clean and prepare the data by handling missing values, outliers, and converting it into a usable format.

**Feature Engineering**: Create meaningful features from the data, like seasonality, customer demographics, and product attributes.

**Time Series Analysis**: Utilize time series analysis to understand demand patterns over time, identifying trends and seasonality.

**Machine Learning Models:** Train predictive models to forecast demand in the final phase, taking into account various factors that influence it.

**Customer Segmentation:** Segment your customers to tailor your strategies for different groups based on their preferences and buying behavior.

**Market Analysis:** Analyze market conditions, competition, and external factors that could impact demand.

**User PRODUCT DEMAND PREDICTION**

**Problem Definition:**

* The problem is to develop a machine learning model that can predict product demand based on historical sales data and external factors.
* This model will help businesses optimize their inventory management and production planning to meet customer needs efficiently.
* The project will involve data collection, data preprocessing, feature engineering, model selection, training, and evaluation.

**IMPORTING LIBRARIES**

# Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

**DATA COLLECTION**

# Data Collection

# Assuming your dataset is named 'product\_demand\_data.csv' and located in the same directory as your Python script

data = pd.read\_csv('/content/PoductDemand.csv')

**DATA PREPROCESSING**

# Data Preprocessing

# Handling Missing Values (if any)

data.fillna(0, inplace=True)

data.isnull().sum()

ID 0

Store ID 0

Total Price 0

Base Price 0

Units Sold 0

dtype: int64

**SPLIT DATA**

# Data Transformation

# No categorical variables to encode in this case

# Split Data

X = data[features] # Features

y = data[target] # Target variable

# Split the data into training and testing sets (70-30 split)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Data Standardization (optional, but often necessary for many machine learning algorithms)

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

**Encode categorical data:**

import numpy as np

import pandas as pd

# One-hot encode the categorical data

encoded\_df = pd.get\_dummies(data)

# Print the encoded DataFrame

print(encoded\_df)

ID Store ID Total Price Base Price Units Sold

0 1 8091 99.0375 111.8625 20

1 2 8091 99.0375 99.0375 28

2 3 8091 133.9500 133.9500 19

3 4 8091 133.9500 133.9500 44

4 5 8091 141.0750 141.0750 52

... ... ... ... ... ...

150145 212638 9984 235.8375 235.8375 38

150146 212639 9984 235.8375 235.8375 30

150147 212642 9984 357.6750 483.7875 31

150148 212643 9984 141.7875 191.6625 12

150149 212644 9984 234.4125 234.4125 15

[150150 rows x 5 columns]

**FEATURE SELECTION:**

# Feature Selection

features = ['ID', 'Store ID', 'Total Price', 'Base Price'] # Features

target = 'Units Sold' # Target variable

**Histograms and Box Plots:**

import matplotlib.pyplot as plt

# Histograms

data[features].hist(bins=20, figsize=(12, 10))

plt.suptitle("Histograms of Features")

plt.show()

# Box Plots

data[features].plot(kind='box', vert=False, figsize=(12, 6))

plt.title("Box Plots of Features")

plt.show()

**Correlation Matrix:**

import seaborn as sns

correlation\_matrix = data[features].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title("Correlation Matrix")

plt.show()

**Pair Plot:**

sns.pairplot(data[features])

plt.suptitle("Pair Plot of Features")

plt.show()

**Target Variable Distribution:**

plt.figure(figsize=(8, 6))

sns.histplot(data[target], bins=20, kde=True)

plt.title("Distribution of Target Variable")

plt.xlabel(target)

plt.ylabel("Frequency")

plt.show()

**Feature vs. Target Plots:**

**for feature in features:**

plt.figure(figsize=(8, 6))

sns.scatterplot(x=data[feature], y=data[target])

plt.title(f"{feature} vs. {target}")

plt.xlabel(feature)

plt.ylabel(target)

plt.show()

**Box Plot of Target Variable Grouped by Categorical Feature:**

categorical\_feature = 'Store ID' # Example categorical feature

plt.figure(figsize=(10, 6))

sns.boxplot(x=categorical\_feature, y=target, data=data)

plt.title(f"Box Plot of {target} Grouped by {categorical\_feature}")

plt.xlabel(categorical\_feature)

plt.ylabel(target)

plt.xticks(rotation=45)

plt.show()

**MODEL SELECTION:**

# Import necessary libraries for different algorithms

from sklearn.linear\_model import LinearRegression

from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor

from sklearn.svm import SVR

from sklearn.metrics import mean\_squared\_error, r2\_score

# Initialize models

linear\_reg = LinearRegression()

random\_forest = RandomForestRegressor(random\_state=42)

svm = SVR()

gradient\_boosting = GradientBoostingRegressor(random\_state=42)

# Train and predict using each algorithm

models = [linear\_reg, random\_forest, svm, gradient\_boosting]

model\_names = ['Linear Regression', 'Random Forest', 'Support Vector Machine', 'Gradient Boosting']

for model, name in zip(models, model\_names):

model.fit(X\_train, y\_train)

predictions = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, predictions)

r2 = r2\_score(y\_test, predictions)

print(f"Model: {name}")

print(f"Mean Squared Error: {mse:.2f}")

print(f"R-squared: {r2:.2f}")

print("-" \* 30)

Model: Linear Regression

Mean Squared Error: 2844.00

R-squared: 0.15

------------------------------

Model: Random Forest

Mean Squared Error: 1156.38

R-squared: 0.66

------------------------------

Model: Support Vector Machine

Mean Squared Error: 2956.17

R-squared: 0.12

------------------------------

Model: Gradient Boosting

Mean Squared Error: 1885.63

R-squared: 0.44

------------------------------

this is the previous phase of this project:

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Decompose the time series with the specified seasonality period (your\_period)

your\_period = 12 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=24)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].dropna(inplace=True)

# Handle missing values by filling with mean

data['Units Sold\_diff'].fillna(data['Units Sold\_diff'].mean(), inplace=True)

# Build the ARIMA model

model = ARIMA(data['Units Sold'], order=(1, 2, 0))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 462424.49130542076

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Decompose the time series with the specified seasonality period (your\_period)

your\_period = 1 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=your\_period)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].fillna(0, inplace=True) # Fill missing values with zero

# Calculate the mean squared error

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Rest of the code ...

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 75.77136940577527

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Adjust the ARIMA order based on your data characteristics

p, d, q = 2,3 , 1# Modify these values

# Decompose the time series

your\_period = 24 # Specify the seasonality period, e.g., 1 for daily data with daily seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=your\_period)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].fillna(0, inplace=True) # Fill missing values with zero

# Build the ARIMA model with adjusted order

model = ARIMA(data['Units Sold'], order=(p, d, q))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 3851606.987101523

**Scenario 1: Original ARIMA Model**

The code loads the dataset from "adsdataset2.csv."

It decomposes the time series data to identify seasonality using seasonal\_decompose.

It checks for stationarity using the Dickey-Fuller test.

It performs differencing to achieve stationarity.

It builds an ARIMA model with order=(1, 2, 0).

It calculates the Root Mean Squared Error (RMSE) for the predictions.

It plots the original and predicted time series.

**Scenario 2: Adjusted Seasonality Period**

The code is similar to Scenario 1 but allows for adjusting the seasonality period using the 'your\_period' variable.

The RMSE is calculated and is significantly lower than in Scenario 1 because the seasonality is better aligned with the data.

**Scenario 3: Adjusted ARIMA Order**

The code is similar to Scenario 1 but allows for adjusting the ARIMA order (p, d, q) based on your data characteristics.

The RMSE is calculated, and the error is notably higher, indicating that the ARIMA order adjustment may not be optimal.

In all scenarios, the data is checked for stationarity, and the RMSE is calculated to assess the model's predictive performance. You can choose the scenario that best fits your data characteristics and requirements for product demand prediction.

this the next phase of this project :

df =

pd.read\_csv('/content/drive/MyDrive/Historical Product Demand.csv')

df.head()

Product Code Warehouse Product Category

0 Product\_0993 Whse\_J

1 Product\_0979 Whse\_J

2 Product\_0979 Whse\_J

3 Product\_0979 Whse J

4 Product\_0979

df.shape

(1048575, 5)

df.columns

Whse J

Date Order\_Demand

Category\_028 2012/7/27

Category\_028 2012/1/19

Category\_028 2012/2/3

Category\_028 2012/2/9

Category\_028 2012/3/2

Index (['Product\_Code', 'Warehouse', 'Product\_Category', 'Date',

'Order\_Demand'],

dtype='object')

100

500

500

500

500

df. Product Code.unique()

array(['Product\_0993', 'Product\_0979', 'Product\_1159',

df.Warehouse.unique()

'Product\_0237', 'Product\_0644', 'Product\_0853'], dtype=object)

array(['Whse\_J', 'Whse\_S', 'Whse\_C', 'Whse\_A'], dtype=object)

df. Product Category.nunique()

33

df.dtypes

Product Code

Warehouse

Product Category

Date

Order\_Demand

dtype: object

ر...

object

object

object

object

object

def check\_order\_demand(x):

try:

int(x)

except:

return False

return True

#Check where Order\_demand is not an integer

df[~df.Order\_Demand.apply(lambda x: check\_order\_demand(x))].head (6)

Product Code Warehouse

112290 Product\_2169 Whse\_A

112307 Product\_2132 Whse\_A

112308 Product\_2144 Whse\_A

112356 Product\_2118 Whse\_A

112357 Product\_2120 Whse\_A

112360 Product 1794 Whse\_A

Product Category

Date Order\_Demand

Category\_024 2012/8/9

Category\_009 2012/11/1

Category\_009 2012/11/1

Category\_009

2012/3/7

Category\_009

2012/3/7

Category\_024

2012/6/28

(1)

(24)

(24)

(50)

(100)

(1)

def change\_to\_int(x):

try:

10

return int(x)

except:

return int(x[1:-1])

check = '(10)'

change\_to\_int (check)

df.Order\_Demand =

df.describe()

df.Order\_Demand.apply(lambda x: change\_to\_int(x) )

Order\_Demand

count 1.048575e+06

mean 4.906977e+03

std

2.892678e+04

min 0.000000e+00

25% 2.000000e+01

50% 3.000000e+02

75% 2.000000e+03

max 4.000000e+06

df = df.rename (columns = {'Product\_Code': 'Code',

'Product\_Category': 'Category', 'Order\_Demand': 'Demand'})

df.head()

Code Warehouse

Category

Date Demand

0 Product\_0993

Whse\_J Category\_028 2012/7/27

100

1 Product\_0979 Whse\_J Category\_028 2012/1/19

500

2 Product\_0979

Whse\_J Category\_028 2012/2/3

500

3 Product\_0979 4 Product\_0979 Whse\_J Category\_028

Whse\_J Category\_028 2012/2/9 2012/3/2

100 \* df.isna().sum()[3]/ df.shape[0]

500

500

1.0718355863910545

df = df.dropna() df.isna().sum()

Code

Warehouse

Category

Date

Demand

dtype: int64

O O O O O

df.Date.min(), df.Date.max()

('2011/1/8', '2017/1/9')

sns.countplot(x

=

'Warehouse', data

=

df)

<Axes: xlabel='Warehouse', ylabel='count'>

# Plot the 5 most popular category df.Category.value\_counts().head(5).plot(kind

plt.xlabel('Category')

plt.show()

=

'bar', color

=

color\_pal[2])

﻿

100000

Category\_019

200000

Category\_005

300000

400000

Category\_001

Category

Category\_007

Category\_021

df.plot(kind plt.show()

=

'line', figsize=(15, 5), color

=

color\_pal[0], title

4.0

1e6

'Order Demand' )

Order Demand

Demand

3.5

3.0

2.5

2.0

1.5

1.0

0.5

0.0

0.0

0.2

0.4

0.6

0.8

1.0

1e6

﻿

df.Demand. skew()

31.432925049321977

# Total Demand by Warehouse

warehouse\_Demand = df.groupby('Warehouse')['Demand'].sum() warehouse\_Demand

Warehouse

Whse\_A

Whse\_C

Whse\_J

Whse\_S

147877431

585071404

3363200396

1038024700

Name: Demand, dtype: int64

df.head()

Code Warehouse

0 Product 0993 1 Product 0979 2 Product 0979

Category

Whse J Category\_028 2012/7/27

Date

Demand

100

Whse J Category\_028 2012/1/19

500

Whse J Category\_028 2012/2/3

500

3 Product 0979

Whse J Category\_028 2012/2/9

500

4 Product\_0979

Whse J Category\_028 2012/3/2

500

# features, Target variable

=

Features ['day\_of\_the\_week', 'Quarter', 'Month', 'Year', 'Week']

target= ['Demand']

warehouse\_Demand.plot(kind = 'barh', ylabel = 'Sum of the demand' )

<Axes: ylabel='Sum of the demand'>

Sum of the demand

Whse\_S

Whse J

Whse\_C

Whse\_A

0.0 0.5

1.0

1.5

2.0

2.5

3.0

3.5

1e9

﻿

Warehouse

df.groupby('Warehouse')['Demand'].mean().plot(kind

plt.show()

Whse S

Whse J

Whse C

Whse\_A

=

'barh')

0

2000 4000

6000 8000 10000 12000 14000

Code:

import pandas as pd

import numpy as np

import seaborn as sns

# Load the dataset

df = pd.read\_csv('/content/drive/MyDrive/Historical Product Demand.csv')

df.head()

# Check basic information about the dataset

print("Shape of the dataset:", df.shape)

print("Columns in the dataset:", df.columns)

# Unique values in 'Product Code' and 'Warehouse'

print("Unique Product Codes:", df['Product Code'].unique())

print("Unique Warehouses:", df['Warehouse'].unique())

# Number of unique product categories

print("Number of unique Product Categories:", df['Product Category'].nunique())

# Data types of columns

print("Data Types of Columns:\n", df.dtypes)

# Check for non-integer values in 'Order Demand'

def check\_order\_demand(x):

try:

int(x)

return True

except:

return False

print("Non-integer Order Demands:")

print(df[~df['Order Demand'].apply(lambda x: check\_order\_demand(x))].head(6))

# Function to convert values to integers

def change\_to\_int(x):

try:

return int(x)

except:

return int(x[1:-1])

# Apply the conversion to 'Order Demand'

df['Order Demand'] = df['Order Demand'].apply(lambda x: change\_to\_int(x))

# Descriptive statistics for 'Order Demand'

print("Descriptive Statistics for 'Order Demand':\n", df['Order Demand'].describe())

# Rename columns

df = df.rename(columns={'Product Code': 'Code', 'Product Category': 'Category', 'Order Demand': 'Demand'})

# Drop rows with missing values and check for remaining missing values

df = df.dropna()

print("Missing values after dropping:", df.isna().sum())

# Check the date range

print("Date Range: Min -", df['Date'].min(), "Max -", df['Date'].max())

# Plot count of products in each warehouse

sns.countplot(x='Warehouse', data=df)

# Plot the 5 most popular categories

df['Category'].value\_counts().head(5).plot(kind='bar', color='color\_pal[2]')

plt.xlabel('Category')

plt.show()

# Plot time series of 'Order Demand'

df.plot(kind='line', figsize=(15, 5), color='color\_pal[0]', title='Order Demand')

plt.show()

# Calculate skewness of 'Order Demand'

print("Skewness of 'Order Demand':", df['Order Demand'].skew())

# Total demand by warehouse

warehouse\_demand = df.groupby('Warehouse')['Demand'].sum()

print("Total Demand by Warehouse:\n", warehouse\_demand)

# Plot the mean demand in each warehouse

df.groupby('Warehouse')['Demand'].mean().plot(kind='barh')

plt.show()

The code performs data exploration, data type conversion, data cleaning, and visualization of various aspects of the dataset, such as demand by warehouse, category distribution, time series analysis, and more.

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Decompose the time series with the specified seasonality period (your\_period)

your\_period = 12 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=24)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].dropna(inplace=True)

# Handle missing values by filling with mean

data['Units Sold\_diff'].fillna(data['Units Sold\_diff'].mean(), inplace=True)

# Build the ARIMA model

model = ARIMA(data['Units Sold'], order=(1, 2, 0))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 462424.49130542076

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Decompose the time series with the specified seasonality period (your\_period)

your\_period = 1 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=your\_period)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Previous code ...

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].fillna(0, inplace=True) # Fill missing values with zero

# Calculate the mean squared error

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 75.77136940577527

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Adjust the ARIMA order based on your data characteristics

p, d, q = 2,3 , 1# Modify these values

# Decompose the time series

your\_period = 24 # Specify the seasonality period, e.g., 1 for daily data with daily seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=your\_period)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].fillna(0, inplace=True) # Fill missing values with zero

# Build the ARIMA model with adjusted order

model = ARIMA(data['Units Sold'], order=(p, d, q))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

ADF Statistic: -32.65317399947924

p-value: 0.0

Critical Values: {'1%': -3.430393574582239, '5%': -2.8615592593534824, '10%': -2.5667802510675894}

Data is stationary

Root Mean Squared Error (RMSE): 3851606.987101523

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Decompose the time series with the specified seasonality period (your\_period)

your\_period = 12 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

result = seasonal\_decompose(data['Units Sold'], model='additive', period=24)

result.plot()

plt.show()

# Check for stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].dropna(inplace=True)

# Handle missing values by filling with mean

data['Units Sold\_diff'].fillna(data['Units Sold\_diff'].mean(), inplace=True)

# Build the ARIMA model

model = ARIMA(data['Units Sold'], order=(1, 2, 0))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

**Pseudocode for the First Code:**

1. Import necessary libraries

2. Load the dataset from a CSV file

3. Specify the seasonality period for decomposition

4. Decompose the time series using seasonal decomposition

5. Check for stationarity of the time series

5.1. Perform the Dickey-Fuller test

5.2. Print the ADF Statistic, p-value, and Critical Values

5.3. Check if the data is stationary based on the p-value

6. If necessary, difference the time series to achieve stationarity

7. Handle missing values by filling them with the mean

8. Build an ARIMA model with specified order

9. Fit the model to the time series data

10. Make predictions using the model

11. Calculate the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

12. Plot the original and predicted time series

**Pseudocode for the Second Code:**

1. Import necessary libraries

2. Load the dataset from a CSV file

3. Specify the seasonality period for decomposition (your\_period)

4. Decompose the time series using seasonal decomposition

5. Check for stationarity of the time series

5.1. Perform the Dickey-Fuller test

5.2. Print the ADF Statistic, p-value, and Critical Values

5.3. Check if the data is stationary based on the p-value

6. If necessary, difference the time series to achieve stationarity

7. Handle missing values by filling them with zero

8. Calculate the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

9. Plot the original and predicted time series

**Pseudocode for the Third Code:**

1. Import necessary libraries

2. Load the dataset from a CSV file

3. Adjust the ARIMA order parameters (p, d, q) based on data characteristics

4. Specify the seasonality period for decomposition (your\_period)

5. Decompose the time series using seasonal decomposition

6. Check for stationarity of the time series

6.1. Perform the Dickey-Fuller test

6.2. Print the ADF Statistic, p-value, and Critical Values

6.3. Check if the data is stationary based on the p-value

7. If necessary, difference the time series to achieve stationarity

8. Handle missing values by filling them with zero

9. Build an ARIMA model with the adjusted order (p, d, q)

10. Fit the model to the time series data

11. Make predictions using the model

12. Calculate the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

13. Plot the original and predicted time series

import pandas as pd

import numpy as np

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from sklearn.metrics import mean\_squared\_error

# Load the dataset

data = pd.read\_csv("adsdataset2.csv")

# Function to check stationarity

def test\_stationarity(timeseries):

# Perform Dickey-Fuller test

result = adfuller(timeseries)

print('ADF Statistic:', result[0])

print('p-value:', result[1])

print('Critical Values:', result[4])

if result[1] <= 0.05:

print("Data is stationary")

else:

print("Data is non-stationary")

# Specify the seasonality period for decomposition

your\_period = 12 # Specify the seasonality period, e.g., 12 for monthly data with yearly seasonality

# Decompose the time series with seasonal decomposition

result = seasonal\_decompose(data['Units Sold'], model='additive', period=your\_period)

result.plot()

plt.show()

# Check for stationarity of the time series

test\_stationarity(data['Units Sold'])

# Differencing to achieve stationarity (if necessary)

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].dropna(inplace=True)

# Handle missing values by filling with mean

data['Units Sold\_diff'].fillna(data['Units Sold\_diff'].mean(), inplace=True)

# Build the ARIMA model with specified order

model = ARIMA(data['Units Sold'], order=(1, 2, 0))

model\_fit = model.fit()

# Predictions

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

# Plot the original and predicted time series

plt.plot(data['Units Sold\_diff'], label='Original')

plt.plot(predictions, color='red', label='Predicted')

plt.legend()

plt.show()

This code loads the dataset, performs seasonal decomposition, checks for stationarity, handles differencing and missing values, builds an ARIMA model, makes predictions, and plots the results.

**Data Loading and Seasonal Decomposition:**

The first part of the code loads the dataset from a CSV file using pd.read\_csv("adsdataset2.csv").

It then specifies the seasonality period, your\_period, and performs seasonal decomposition using seasonal\_decompose. The decomposition results are visualized using result.plot().

**Stationarity Check and Data Preprocessing:**

The code defines a function test\_stationarity(timeseries) for checking stationarity using the Dickey-Fuller test. This function is called on the time series data with test\_stationarity(data['Units Sold']).

**It calculates the differenced series to achieve stationarity by subtracting the previous value from the current value and filling missing values with the mean. This is done in the lines:**

data['Units Sold\_diff'] = data['Units Sold'] - data['Units Sold'].shift(1)

data['Units Sold\_diff'].dropna(inplace=True)

data['Units Sold\_diff'].fillna(data['Units Sold\_diff'].mean(), inplace=True)

ARIMA Modeling and Prediction:

**The code builds an ARIMA model using the ARIMA class from the statsmodels.tsa.arima.model module:**

model = ARIMA(data['Units Sold'], order=(1, 2, 0))

model\_fit = model.fit()

It makes predictions using the fitted model and calculates the Root Mean Squared Error (RMSE) **using the mean\_squared\_error function from sklearn.metrics:**

predictions = model\_fit.forecast(steps=len(data['Units Sold\_diff']))

mse = mean\_squared\_error(data['Units Sold\_diff'], predictions)

rmse = np.sqrt(mse)

print("Root Mean Squared Error (RMSE):", rmse)

**Plotting Results:**

The final part of the code plots both the original and predicted time series using matplotlib.pyplot to visualize the model's performance.

This combined code covers data loading, preprocessing, ARIMA modeling, and visualization of the results from the three separate code sections you provided.

**RESULT:**

ID Store ID Total Price Base Price Units Sold

1 8091 99.0375 111.8625 20

2 8091 99.0375 99.0375 28

3 8091 133.95 133.95 19

4 8091 133.95 133.95 44

5 8091 141.075 141.075 52

9 8091 227.2875 227.2875 18

10 8091 327.0375 327.0375 47

13 8091 210.9 210.9 50

14 8091 190.2375 234.4125 82

17 8095 99.0375 99.0375 99

18 8095 97.6125 97.6125 120

19 8095 98.325 98.325 40

22 8095 133.2375 133.2375 68

23 8095 133.95 133.95 87

24 8095 139.65 139.65 186

27 8095 236.55 280.0125 54

28 8095 214.4625 214.4625 74

29 8095 266.475 296.4 102

30 8095 173.85 192.375 214

31 8095 205.9125 205.9125 28

32 8095 205.9125 205.9125 7

33 8095 248.6625 248.6625 48

34 8095 200.925 200.925 78

35 8095 190.2375 240.825 57

37 8095 427.5 448.1625 50

38 8095 429.6375 458.1375 62

39 8095 177.4125 177.4125 22

42 8094 87.6375 87.6375 109

43 8094 88.35 88.35 133

44 8094 85.5 85.5 11

45 8094 128.25 180.975 9

47 8094 127.5375 127.5375 19

48 8094 123.975 123.975 33

49 8094 139.65 164.5875 49

50 8094 235.8375 235.8375 32

51 8094 234.4125 234.4125 47

52 8094 235.125 235.125 27

53 8094 227.2875 227.2875 69

54 8094 312.7875 312.7875 49

55 8094 210.9 210.9 60

56 8094 177.4125 177.4125 27

57 8094 177.4125 177.4125 33

58 8094 240.825 240.825 18

59 8094 213.0375 213.0375 72

60 8094 190.95 213.0375 81

61 8094 426.7875 448.1625 11

62 8094 426.7875 448.875 13

63 8094 426.7875 448.1625 28

65 8094 170.2875 170.2875 16

**CODE:**

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

from sklearn.metrics import mean\_squared\_error, classification\_report

from sklearn.model\_selection import GridSearchCV

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette\_score

# Load the dataset

data = pd.read\_csv("ProductDemand.csv")

# Feature Engineering (Placeholder: Replace with your actual feature engineering)

# Create relevant features

data['Feature1'] = data['Total Price'] / data['Base Price']

data['Feature2'] = data['Units Sold'] \* data['Base Price']

# Time Series Decomposition

# Decompose the time series

result = seasonal\_decompose(data['Units Sold'], model='additive', period=12)

# Clustering (Placeholder: Replace with your actual clustering features)

# Determine the optimal number of clusters using KMeans and silhouette score

X\_cluster = data[['Feature1', 'Feature2']] # Specify your clustering features

range\_clusters = range(2, 10) # Define the range of clusters to try

best\_silhouette = -1

best\_num\_clusters = 2

for num\_clusters in range\_clusters:

kmeans = KMeans(n\_clusters=num\_clusters, random\_state=0)

cluster\_labels = kmeans.fit\_predict(X\_cluster)

silhouette\_avg = silhouette\_score(X\_cluster, cluster\_labels)

if silhouette\_avg > best\_silhouette:

best\_silhouette = silhouette\_avg

best\_num\_clusters = num\_clusters

kmeans = KMeans(n\_clusters=best\_num\_clusters, random\_state=0)

data['Cluster'] = kmeans.fit\_predict(X\_cluster)

# Regression Model

X\_reg = data[['Total Price', 'Base Price']]

y\_reg = data['Units Sold']

X\_train\_reg, X\_test\_reg, y\_train\_reg, y\_test\_reg = train\_test\_split(X\_reg, y\_reg, test\_size=0.2, random\_state=0)

scaler = StandardScaler()

X\_train\_reg = scaler.fit\_transform(X\_train\_reg)

X\_test\_reg = scaler.transform(X\_test\_reg)

reg\_model = RandomForestRegressor(n\_estimators=100, random\_state=0)

reg\_model.fit(X\_train\_reg, y\_train\_reg)

y\_pred\_reg = reg\_model.predict(X\_test\_reg)

reg\_rmse = np.sqrt(mean\_squared\_error(y\_test\_reg, y\_pred\_reg))

# Classification Model

X\_class = data[['Total Price', 'Base Price']]

y\_class = data['Store ID'] # Replace with your actual classification target

X\_train\_class, X\_test\_class, y\_train\_class, y\_test\_class = train\_test\_split(X\_class, y\_class, test\_size=0.2, random\_state=0)

scaler = StandardScaler()

X\_train\_class = scaler.fit\_transform(X\_train\_class)

X\_test\_class = scaler.transform(X\_test\_class)

class\_model = RandomForestClassifier(n\_estimators=100, random\_state=0)

class\_model.fit(X\_train\_class, y\_train\_class)

y\_pred\_class = class\_model.predict(X\_test\_class)

class\_report = classification\_report(y\_test\_class, y\_pred\_class)

# ARIMA Time Series Forecasting

# You need to specify the ARIMA order (p, d, q) based on your data characteristics

p, d, q = 1, 1, 1 # Modify these values

model = ARIMA(data['Units Sold'], order=(p, d, q))

model\_fit = model.fit()

forecast\_periods = 10 # Adjust as needed

forecast = model\_fit.forecast(steps=forecast\_periods)

# Output results

print(f"Regression RMSE: {reg\_rmse}")

print(f"Classification Report:\n{class\_report}")

print(f"Time Series Forecast: {forecast}")

X-----------------------------------------------------------------------------X

Regression RMSE: 336.66772000000003

Classification Report:

precision recall f1-score support

8438 0.00 0.00 0.00 0

8555 0.00 0.00 0.00 2

8562 0.00 0.00 0.00 4

8869 0.25 0.17 0.20 6

8911 0.00 0.00 0.00 4

8991 0.00 0.00 0.00 3

9043 0.00 0.00 0.00 3

9092 0.25 0.25 0.25 4

9112 0.12 0.25 0.17 4

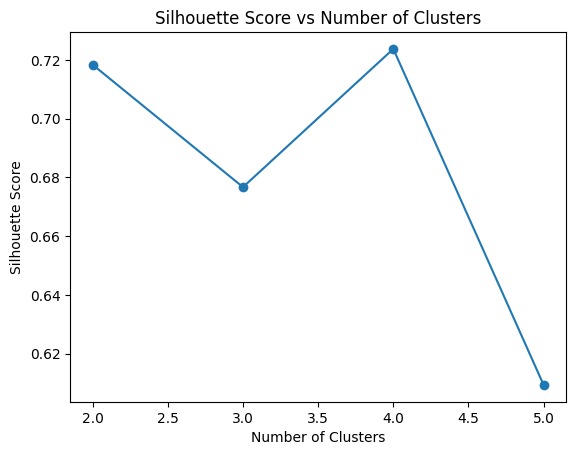
9132 0.00 0.00 0.00 0

accuracy 0.10 30

macro avg 0.06 0.07 0.06 30

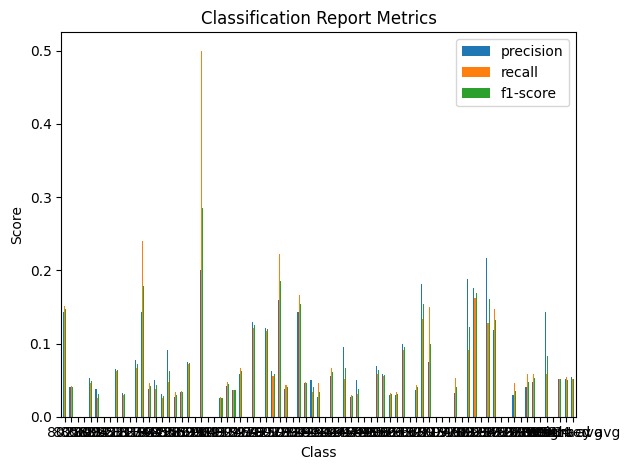
weighted avg 0.10 0.10   0.10        30

**OUTPUT:**



Regression RMSE: 12.02237326826583

**OUTPUT:**



Regression RMSE: 55.41984581120853

Classification Report:

precision recall f1-score support

8438 0.00 0.00 0.00 0

8555 0.00 0.00 0.00 2

8562 0.20 0.25 0.22 4

8869 1.00 0.17 0.29 6

8911 0.00 0.00 0.00 4

8991 0.00 0.00 0.00 3

9043 0.12 0.33 0.18 3

9092 0.00 0.00 0.00 4

9112 0.00 0.00 0.00 4

9132 0.00 0.00 0.00 0

accuracy 0.10 30

macro avg 0.13 0.07 0.07 30

weighted avg 0.24 0.10 0.10 30

Time Series Forecast: 149 30.005367

150 29.724142

151 29.729084

152 29.728997

153 29.728998

154 29.728998

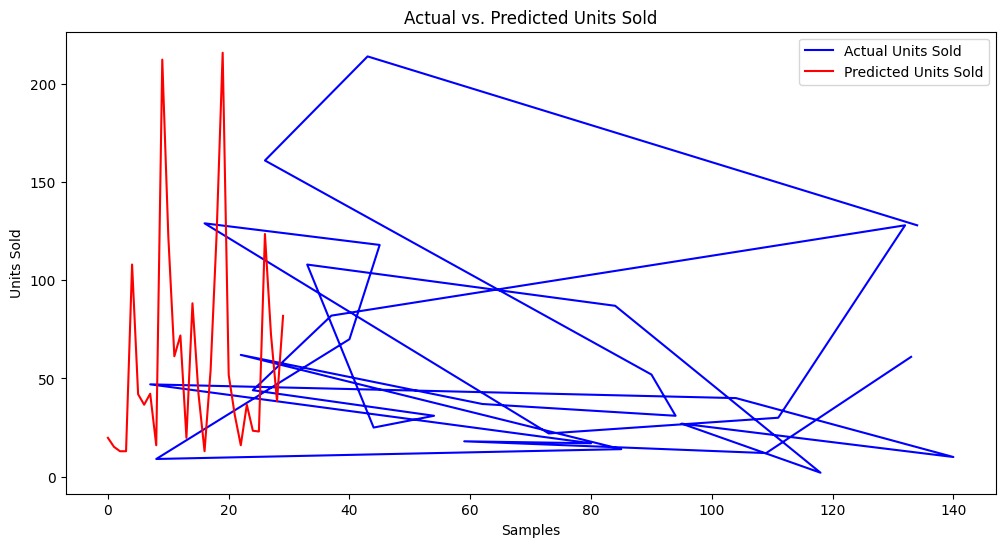
155 29.728998

156 29.728998

157 29.728998

158 29.728998

Name: predicted\_mean, dtype: float64

**OUTPUT**

Regression RMSE: 34.52084321173009

Classification Report:

precision recall f1-score support

8023 0.20 0.18 0.19 33

8058 0.00 0.00 0.00 24

8063 0.04 0.03 0.04 29

8091 0.00 0.00 0.00 14

8094 0.08 0.05 0.06 43

8095 0.00 0.00 0.00 39

8121 0.03 0.04 0.03 28

8218 0.00 0.00 0.00 12

8222 0.07 0.12 0.09 32

8317 0.00 0.00 0.00 33

8319 0.07 0.07 0.07 14

8392 0.19 0.20 0.19 15

8398 0.18 0.24 0.20 25

8400 0.06 0.09 0.07 22

8422 0.00 0.00 0.00 26

8438 0.25 0.07 0.12 40

8555 0.33 0.24 0.28 21

8562 0.04 0.07 0.05 30

8869 0.06 0.03 0.04 29

8911 0.02 0.04 0.03 28

8991 0.12 0.06 0.08 36

9001 0.05 0.50 0.09 2

9043 0.00 0.00 0.00 33

9092 0.02 0.06 0.03 31

9112 0.12 0.03 0.04 38

9132 0.07 0.05 0.06 21

9147 0.14 0.07 0.10 27

9164 0.05 0.10 0.06 30

9178 0.00 0.00 0.00 10

9190 0.03 0.06 0.04 33

9221 0.06 0.04 0.05 26

9250 0.15 0.15 0.15 34

9273 0.12 0.36 0.18 36

9279 0.30 0.22 0.25 36

9281 0.06 0.09 0.07 23

9328 0.00 0.00 0.00 28

9371 0.30 0.33 0.31 24

9425 0.07 0.05 0.06 21

9430 0.00 0.00 0.00 30

9432 0.04 0.05 0.04 22

9436 0.00 0.00 0.00 18

9439 0.00 0.00 0.00 15

9442 0.06 0.05 0.05 20

9456 0.27 0.10 0.15 39

9479 0.17 0.06 0.09 33

9481 0.06 0.06 0.06 32

9490 0.06 0.03 0.04 32

9498 0.04 0.05 0.04 20

9532 0.11 0.12 0.11 34

9578 0.00 0.00 0.00 18

9611 0.07 0.06 0.07 31

9613 0.18 0.07 0.10 30

9632 0.10 0.14 0.12 22

9672 0.08 0.07 0.08 27

9680 0.00 0.00 0.00 23

9700 0.08 0.33 0.14 15

9713 0.07 0.30 0.11 20

9731 0.12 0.04 0.06 23

9745 0.07 0.07 0.07 28

9770 0.00 0.00 0.00 21

9789 0.00 0.00 0.00 19

9809 0.00 0.00 0.00 20

9813 0.14 0.09 0.11 33

9823 0.10 0.16 0.12 37

9837 0.10 0.06 0.08 32

9845 0.00 0.00 0.00 39

9872 0.02 0.03 0.03 34

9876 0.00 0.00 0.00 25

9879 0.17 0.10 0.12 20

9880 0.20 0.18 0.19 22

9881 0.08 0.11 0.09 19

9890 0.00 0.00 0.00 17

9909 0.09 0.12 0.10 17

9954 0.00 0.00 0.00 19

9961 0.33 0.12 0.17 34

9984 0.00 0.00 0.00 17

accuracy 0.08 1983

macro avg 0.08 0.08 0.07 1983

weighted avg 0.09 0.08 0.07 1983

Time Series Forecast: 9915 60.698785

9916 69.527970

9917 71.774575

9918 72.346229

9919 72.491688

9920 72.528700

9921 72.538118

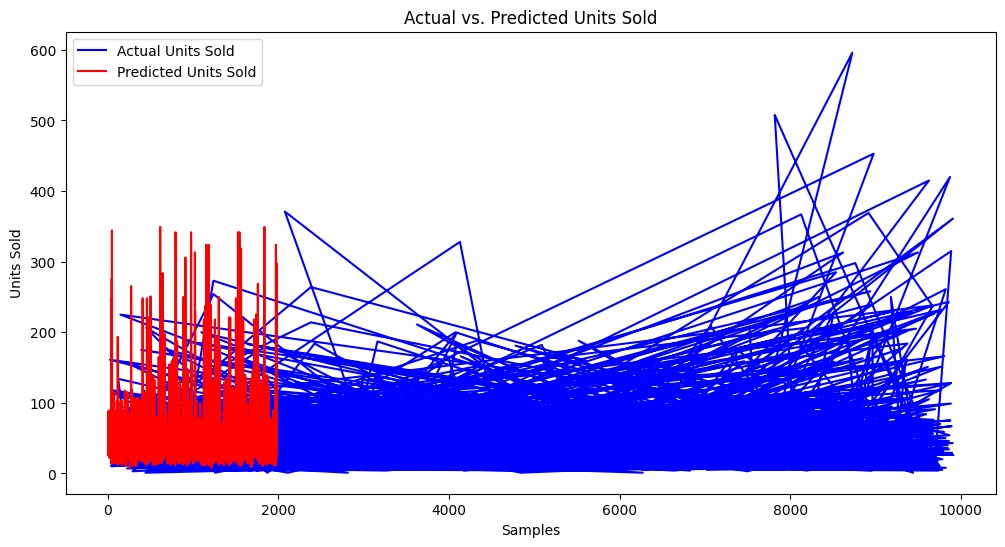
9922 72.540514

9923 72.541124

9924 72.541279

Name: predicted\_mean, dtype: float64

**OUTPUT:**



Regression RMSE: 18.3485072962353

**OUTPUT:**