Forecasting Sales for Online Retailer

Team Name: Data Wizards

Group Members:

Shachi Hardi (sh4615)

Shenova Davis (ssd2184)

Ritayan Patra (rp3247)

COLUMBIA ENGINEERING

The Fu Foundation School
of Engineering and Applied Science

Introduction

Goal: Develop a time series forecasting model to predict future sales for a retail company.

Why?

- Helps in better demand planning to meet customer needs.
- Optimizes inventory management to reduce overstock and stockouts.
- Supports strategic pricing and promotions to maximize revenue.

Techniques Used: Time series methods like ARIMA, Prophet, and LSTM to capture complex sales patterns.

Business Impact: The project aims to provide actionable insights for improving profitability, customer satisfaction, and operational efficiency.

Data Description

Online Retail Dataset

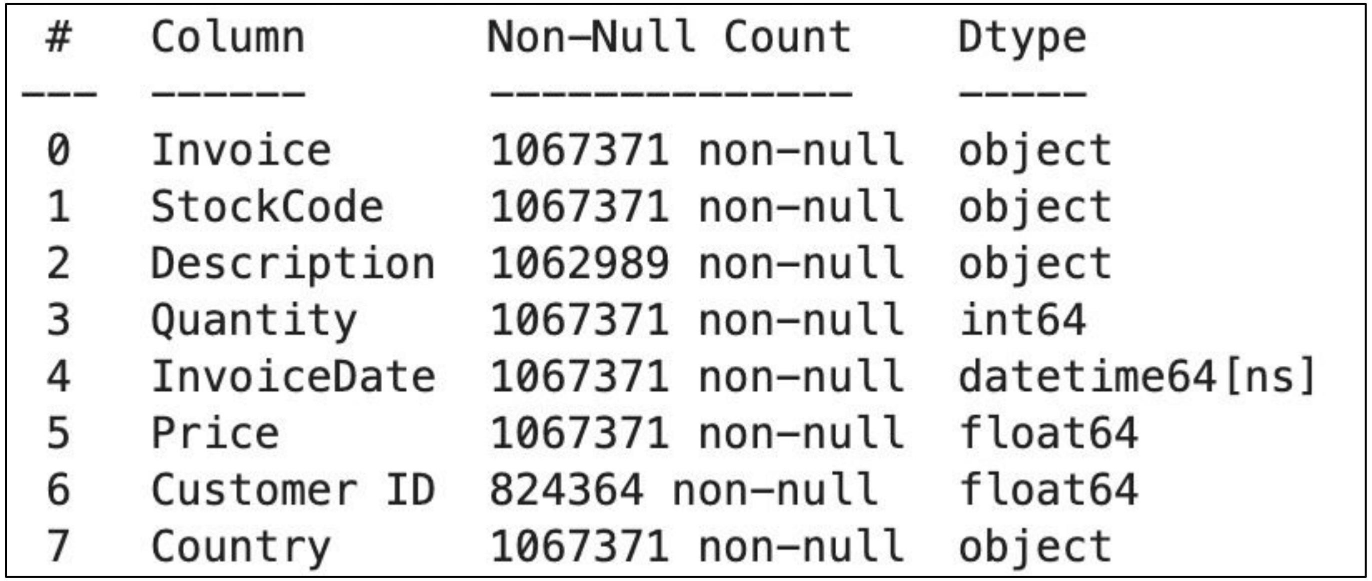
Source: Link

No. of Columns: 8

No. of Rows: 1M +

The important columns/features which are essential for data analysis are -

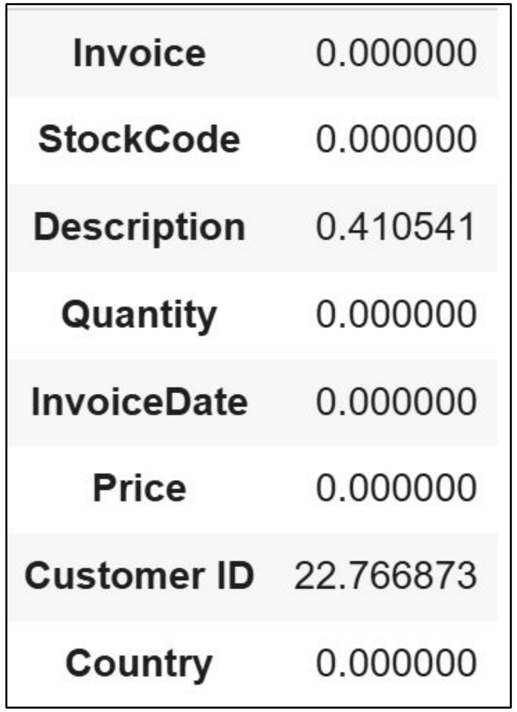
- Description
- InvoiceDate
- Price
- Country
- Customer ID



```
 #   Column        Non-Null Count    Dtype
---  ------        --------------    -----
 0   Invoice       1067371 non-null  object
 1   StockCode     1067371 non-null  object
 2   Description   1062989 non-null  object
 3   Quantity      1067371 non-null  int64
 4   InvoiceDate   1067371 non-null  datetime64[ns]
 5   Price         1067371 non-null  float64
 6   Customer ID   824364 non-null   float64
 7   Country       1067371 non-null  object
```

Data Preprocessing

- Missing Customer IDs: 238,625 missing values.
  - 236,122 records are valid but lack a Customer ID → Impute as "Unknown Customer ID".
  - The rest contain illogical values and need to be removed.
- Invalid Records to Drop:
  - 768 records: Price = 0 & Quantity < 0.
  - 981 records: Price = 0 & Quantity > 0.
  - 5 records: Price < 0.
  - 749 records: Price > 0 & Quantity < 0.
- Cancelled Orders:
  - Invoice IDs starting with 'C' represent cancellations → These have negative quantities and should be separated from actual sales.
- A ‘**Sales**’ column is generated by multiplying ‘**Quantity**’ and ‘**Price**’, which will be used for model building and prediction.



Invoice	0.000000
StockCode	0.000000
Description	0.410541
Quantity	0.000000
InvoiceDate	0.000000
Price	0.000000
Customer ID	22.766873
Country	0.000000

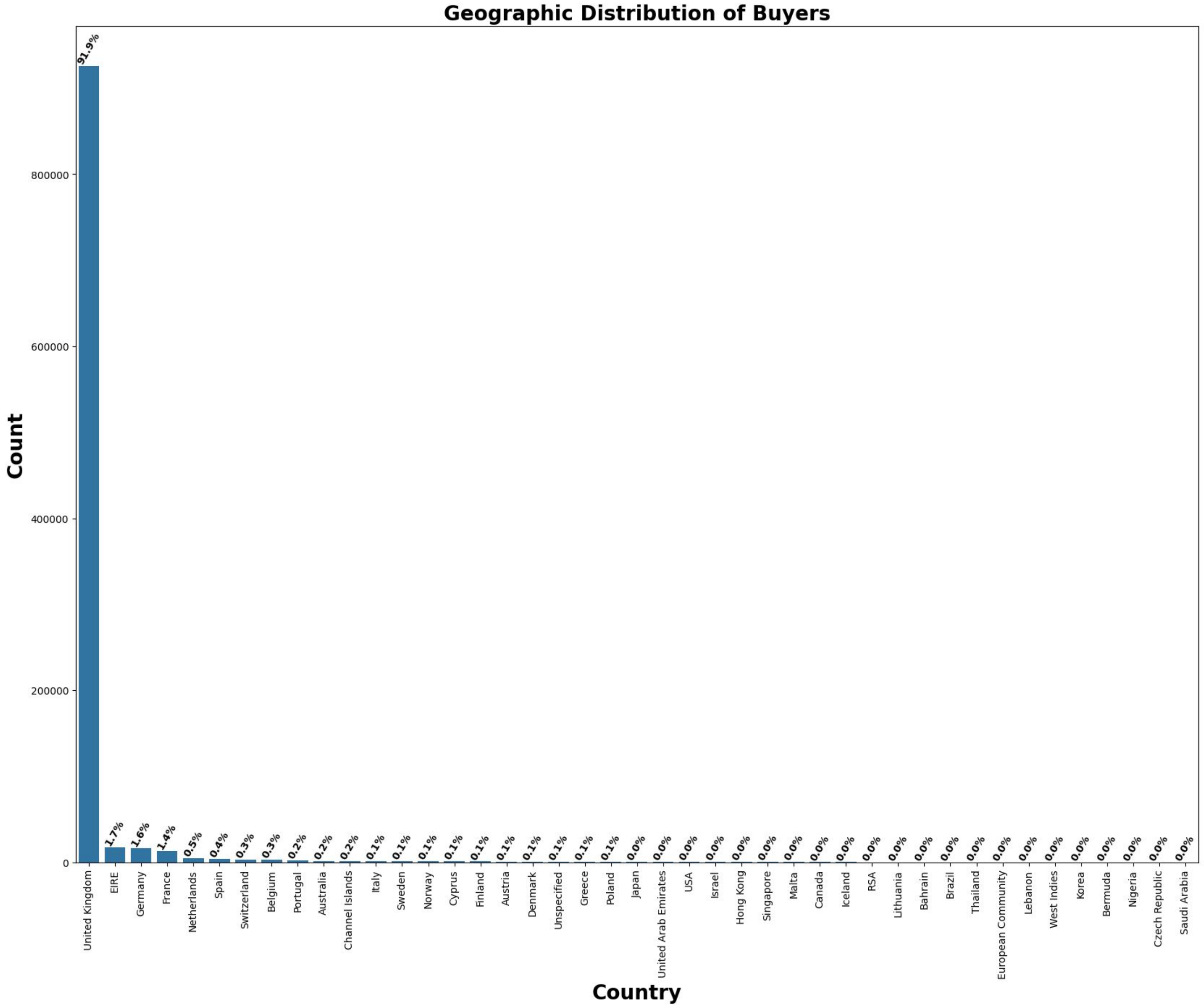
**Percentage of Missing Values*

EDA

The analysis of geographic distribution reveals that the majority of buyers (91.9%) are from the United Kingdom.

A smaller portion of customers come from Ireland (EIRE), Germany, France, and the Netherlands, with each contributing less than 2%. Buyers from other countries make up an even smaller fraction of the total sales.

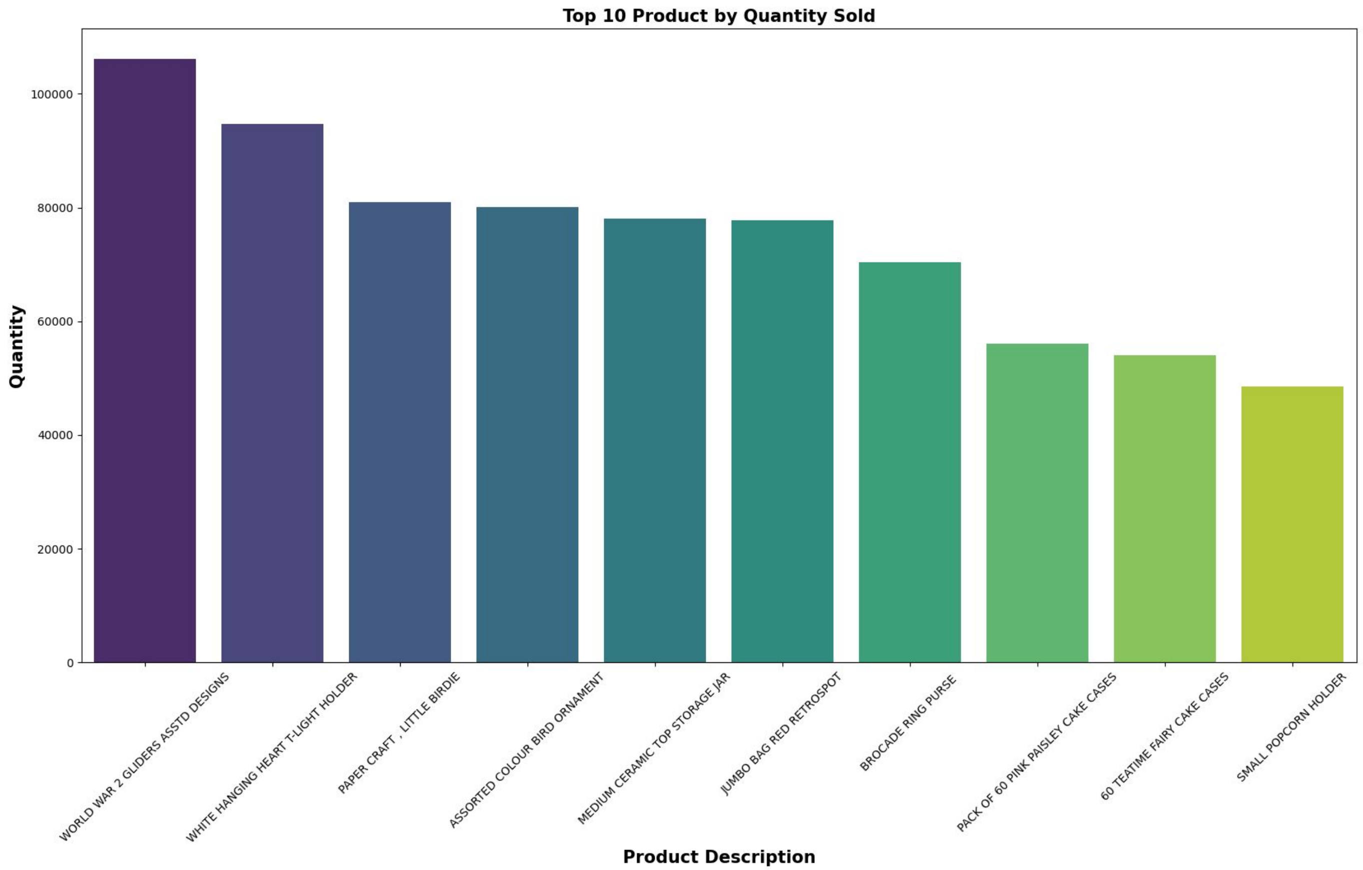
For Deliverable 1, the focus is on transactions originating from the United Kingdom, as it represents the largest customer base in the dataset.



EDA

The top selling products by the online retail store are →

- WORLD WAR 2 GLIDERS ASSTD DESIGNS
- WHITE HANGING HEART T-LIGHT HOLDER
- PAPER CRAFT , LITTLE BIRDIE
- ASSORTED COLOUR BIRD ORNAMENT
- MEDIUM CERAMIC TOP STORAGE JAR
- JUMBO BAG RED RETROSPOT
- BROCADE RING PURSE
- PACK OF 60 PINK PAISLEY CAKE CASES
- 60 TEATIME FAIRY CAKE CASES
- SMALL POPCORN HOLDER

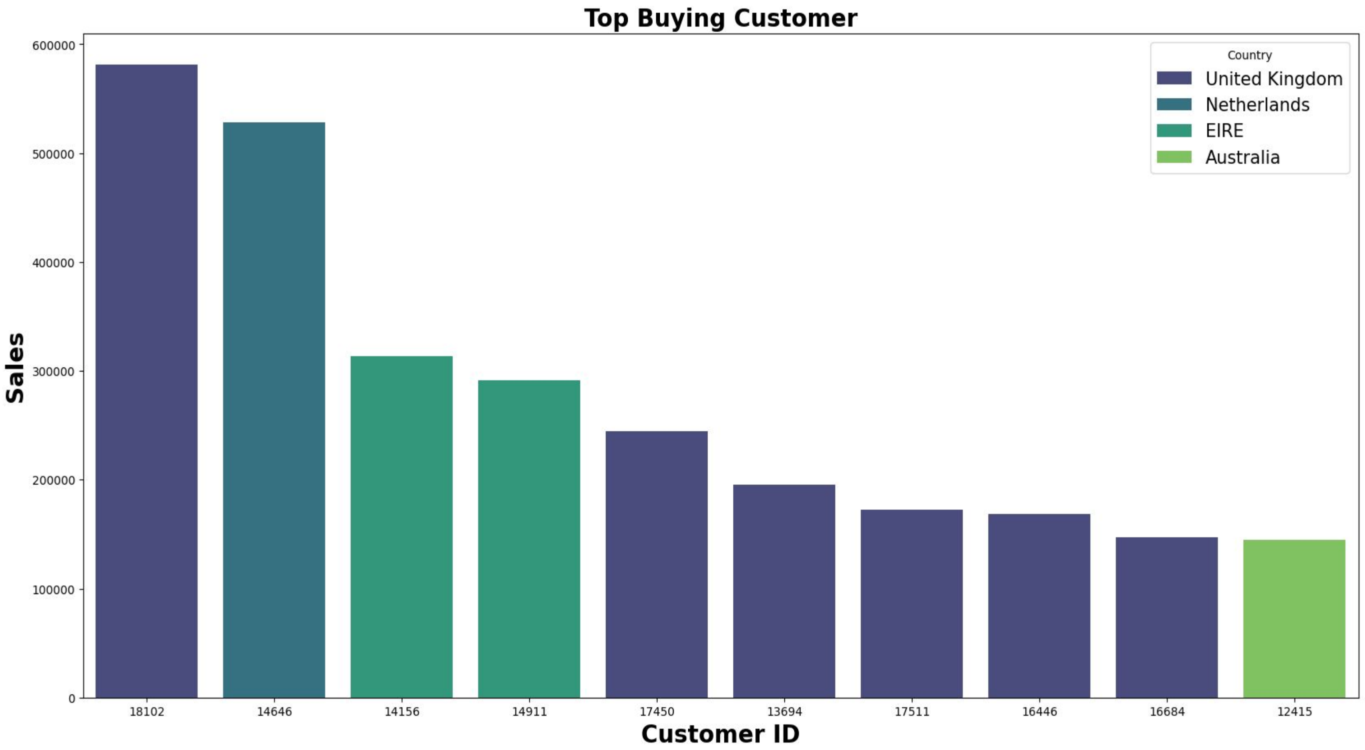


EDA

The top 10 buying customers are from countries United Kingdom, Netherlands, EIRE (Ireland) and Australia. The customer IDs of such customers are →

- 18102
- 14646
- 14156
- 14911
- 17459
- 13694
- 17511
- 16446
- 16684
- 12415

These customers are have bought significant amount of products from this online store.



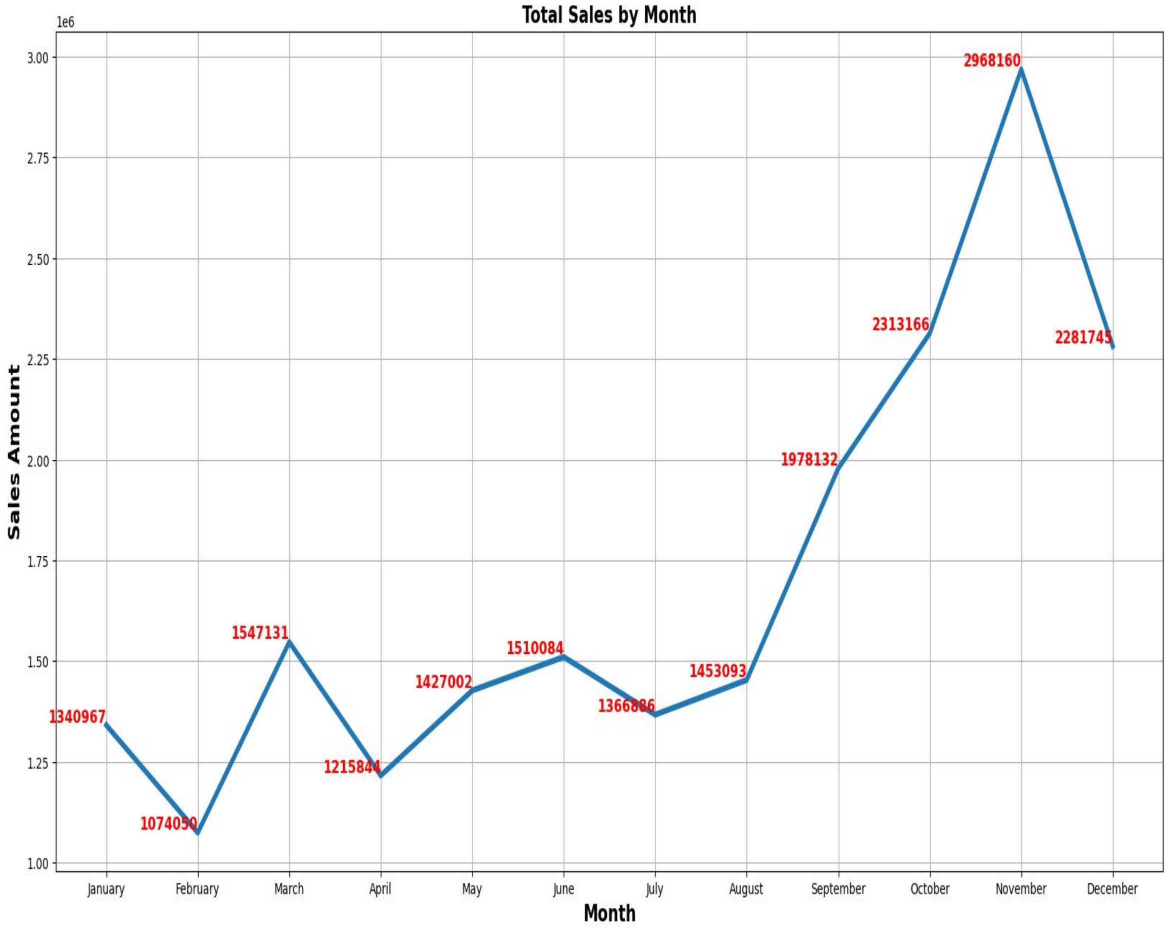
EDA

Sales are relatively low at the start of the year but gradually increase as the months progress.

By mid-year, around July, there's a noticeable rise, which becomes more pronounced from August onward.

The highest sales occur in November, driven by major shopping events like Black Friday and the holiday season.

This trend highlights a clear seasonal pattern, with consumer spending peaking toward the end of the year.

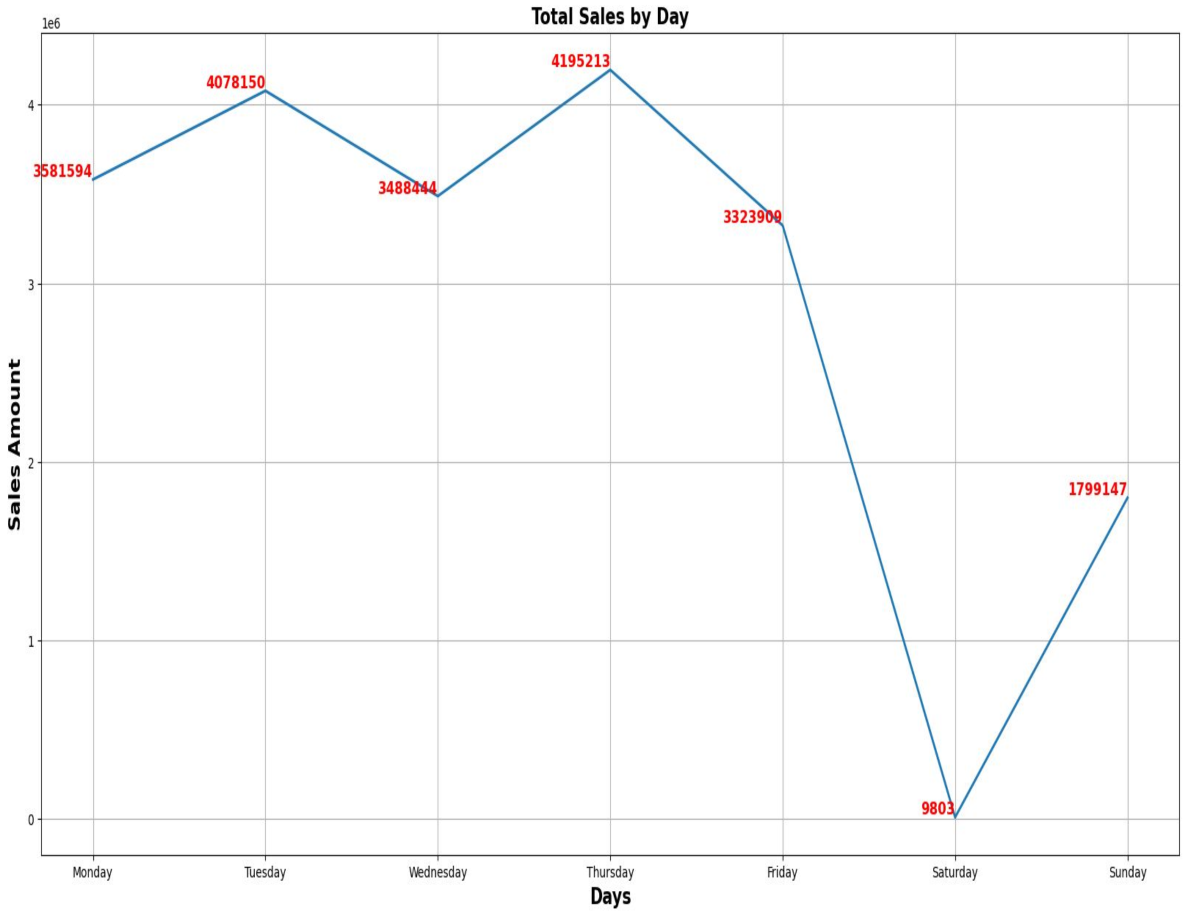


EDA

The "Total Sales by Day" graph reveals a mid-week peak in sales, rising from Monday (3.58M) to a high on Thursday (4.20M) before declining on Friday (3.32M).

A sharp drop occurs on Saturday (9.8K), suggesting minimal activity, followed by a partial rebound on Sunday (1.8M).

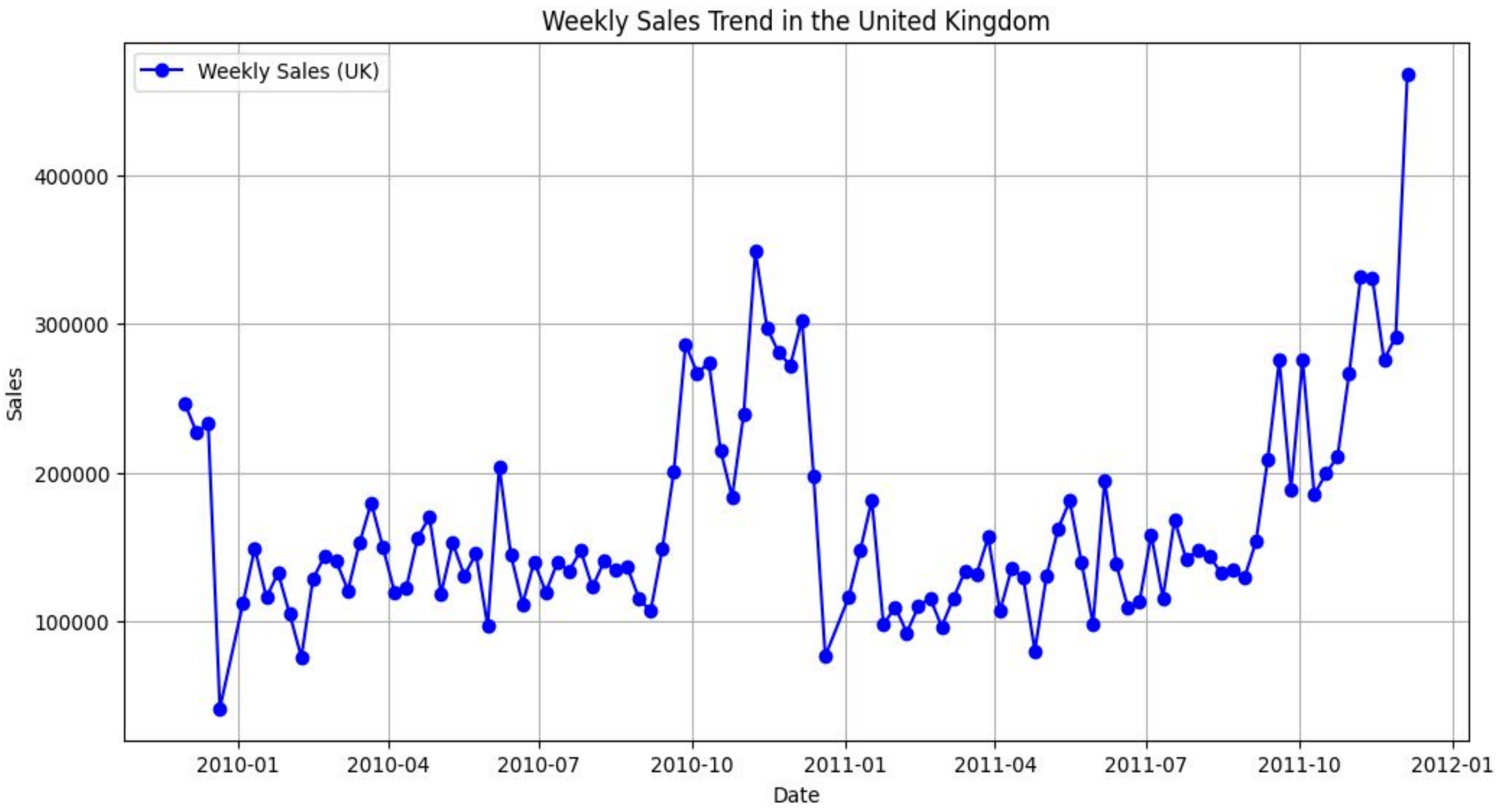
This pattern indicates strong sales during the week with a significant weekend dip.



Weekly Data Plot

The chart displays weekly sales trends from 2010 to 2011 in the UK.

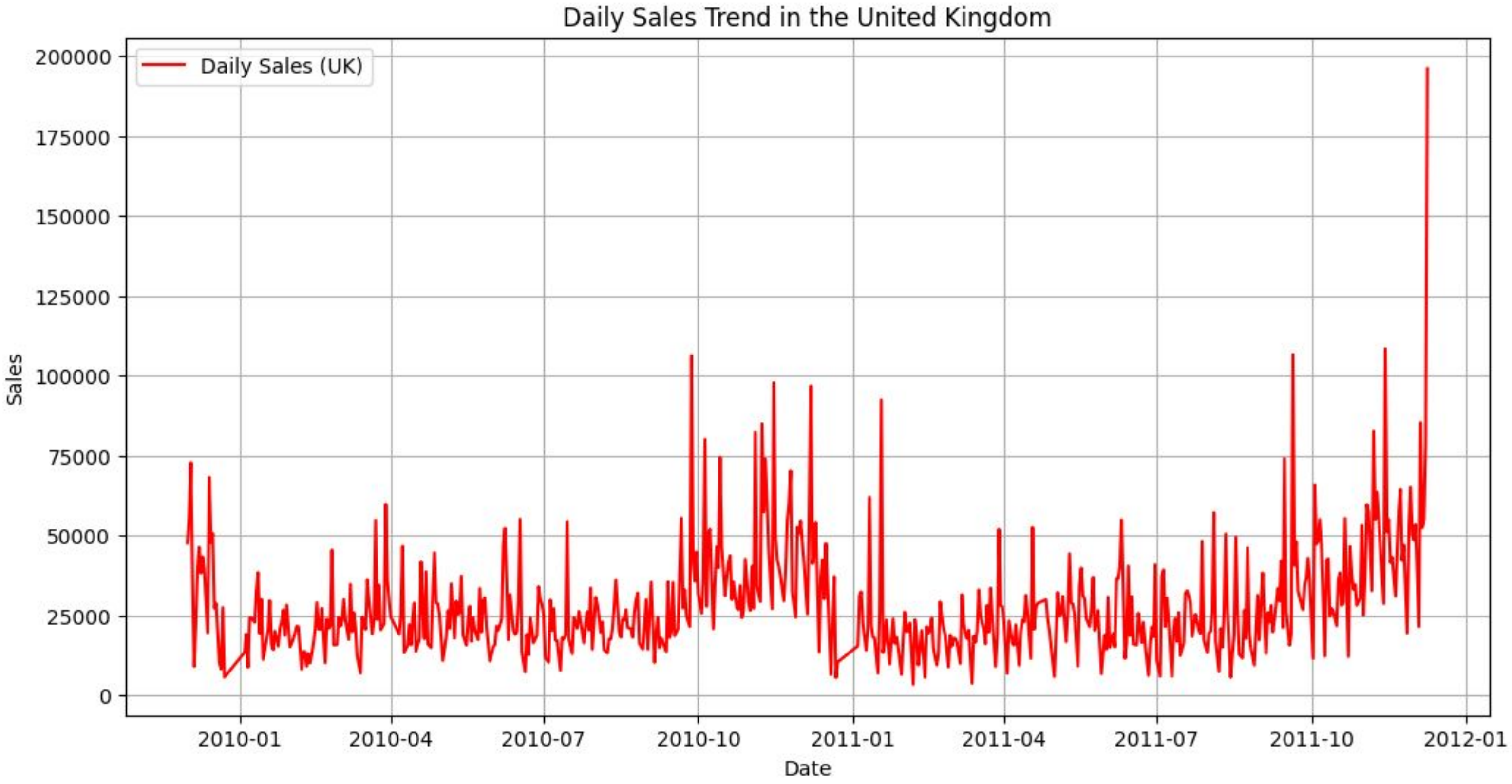
Sales fluctuate with seasonal variations, showing a strong upward trend in late 2011, likely due to holiday demand.



Daily Data Plot

The chart illustrates the daily sales trend of an online retail store in the United Kingdom. While sales exhibit frequent spikes throughout the year, the overall trend remains relatively low for most months.

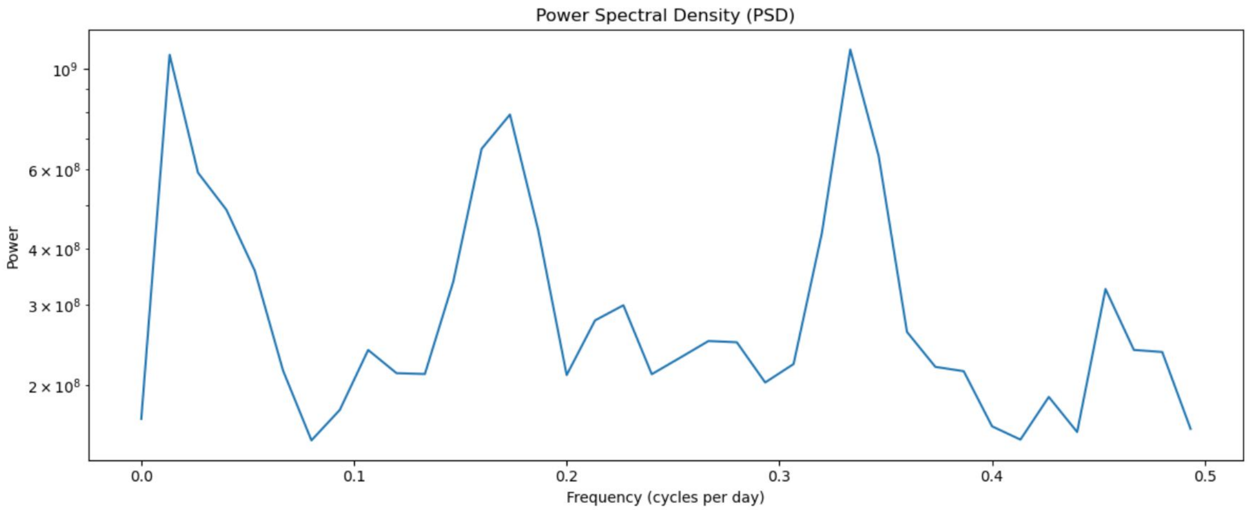
However, a noticeable increase occurs in October, November, and December, with sales reaching their highest point during this period.

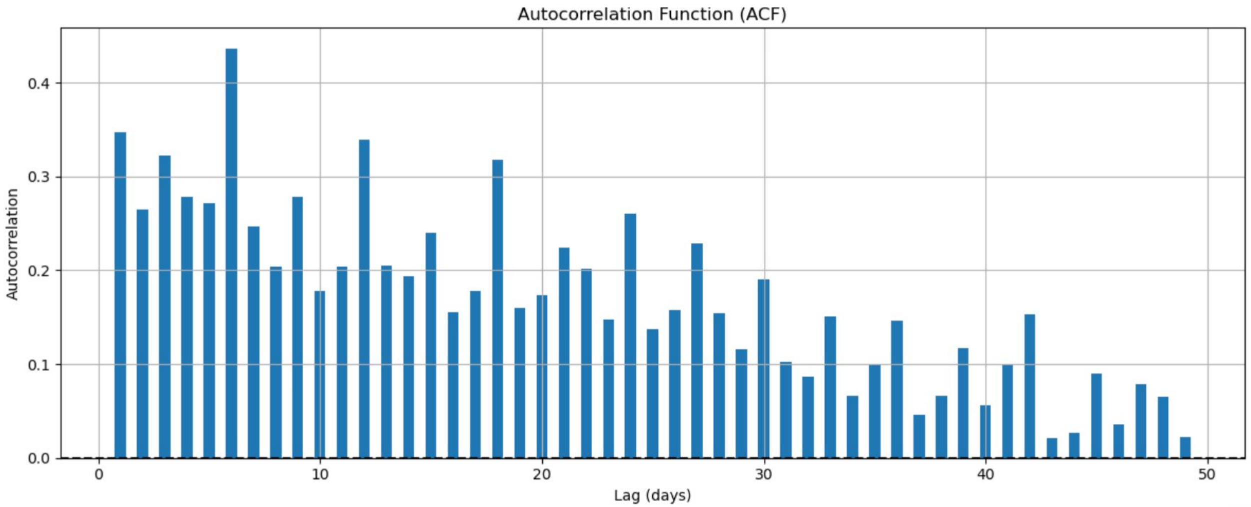


Pre Modelling

Power Spectral Density and Autocorrelation

- The Power Spectral Density (PSD) plot reveals dominant frequencies, indicating weekly and monthly seasonal trends in sales.
- The Autocorrelation Function (ACF) plot shows high autocorrelation at small lags, indicating short-term persistence, and spikes at 7, 14, and 21 days highlight strong weekly seasonality.
- As lag increases, autocorrelation weakens. Overall, the PSD and ACF analysis suggest that sales data follows weekly cycles, and forecasting models should account for this seasonality.


Power Spectral Density (PSD)


Autocorrelation Function (ACF)

Modelling

Train-Test Split

To evaluate the forecasting model, we divided the dataset into:

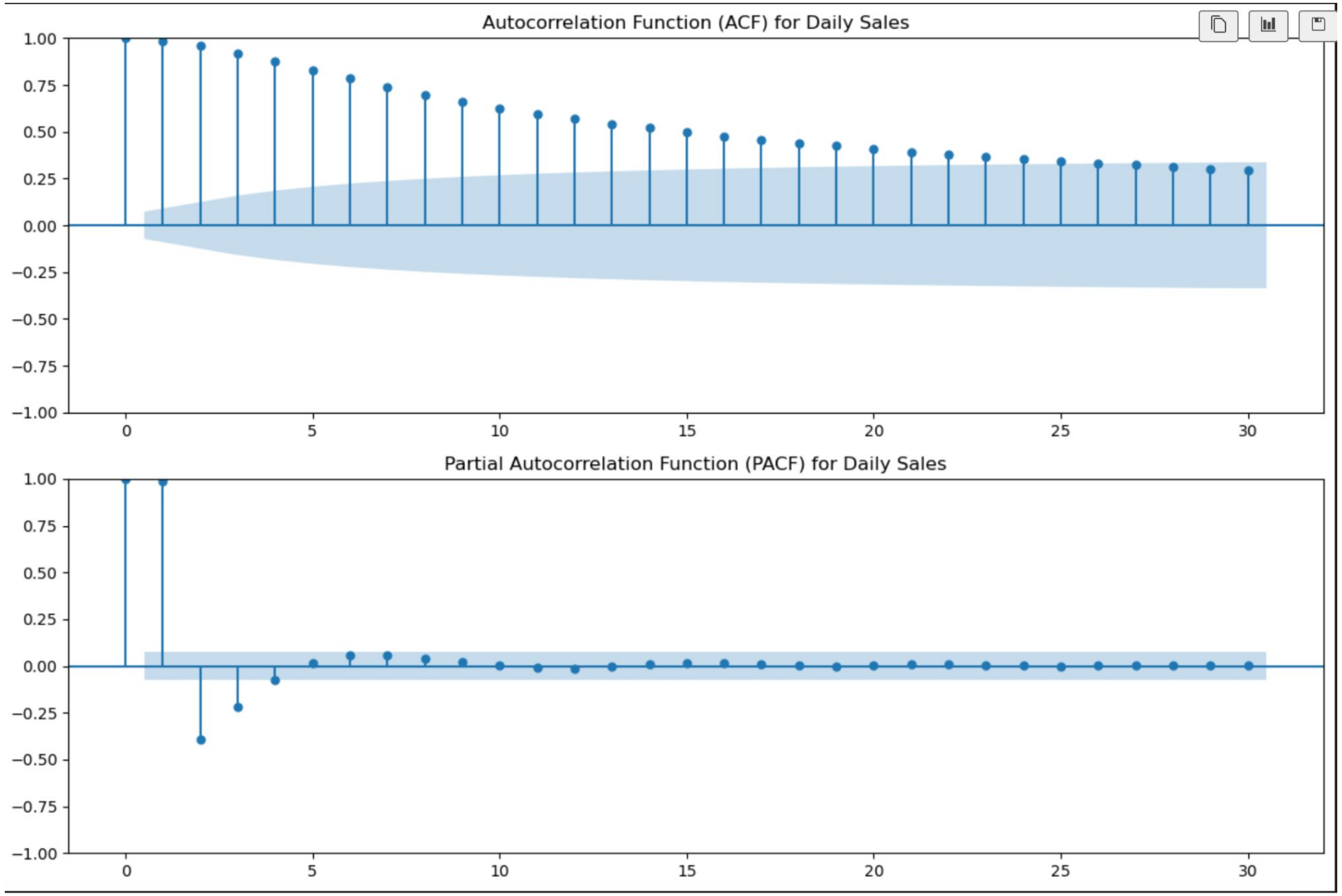
- Training Set (80%): Used to train the model.
- Testing Set (20%): Used for final evaluation.

The following slides display the use of models like ARIMA, SARIMA, SARIMAX, FbProphet and LSTM for the purpose of predicting daily sales (price x quantity).

Models have been evaluated through the use of MAPE (Mean Absolute Percentage Error) on the test data.

Arima Model

MAPE: 27.68%



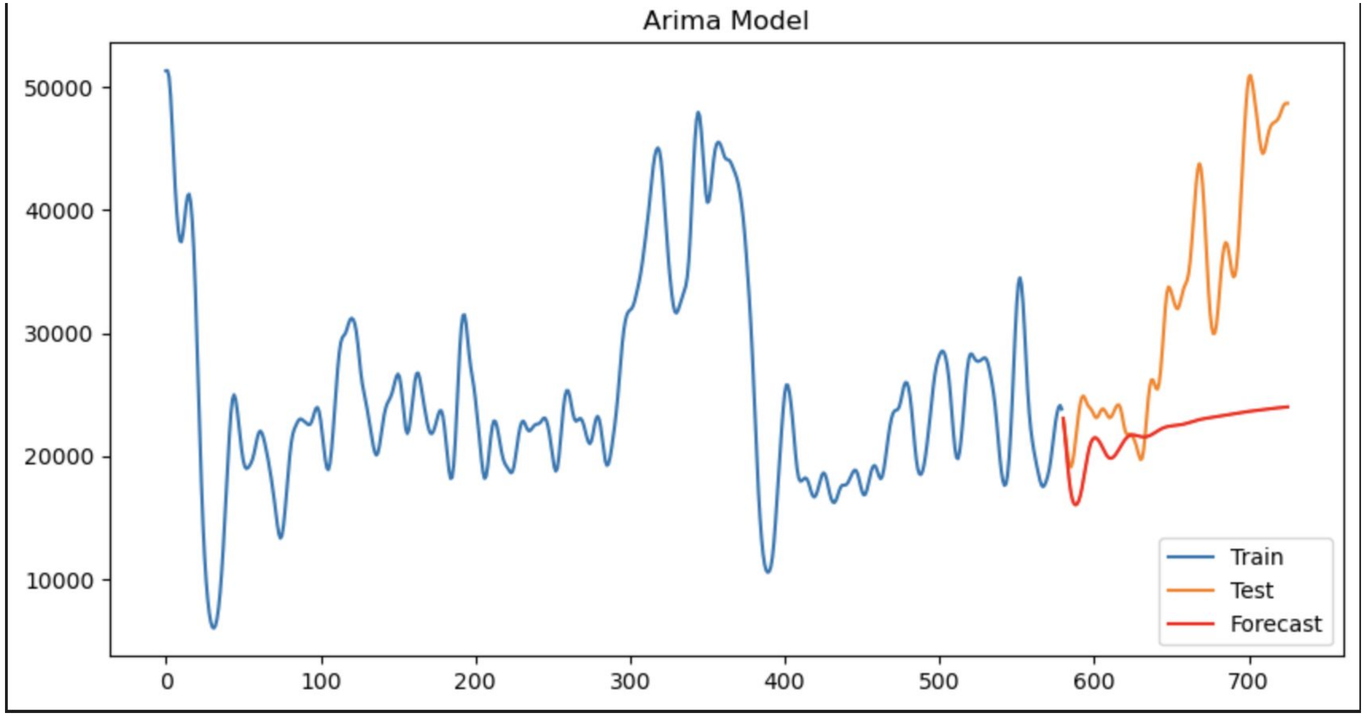
The ACF plot shows a declining pattern, indicating non-stationarity, requiring differencing (d=1).

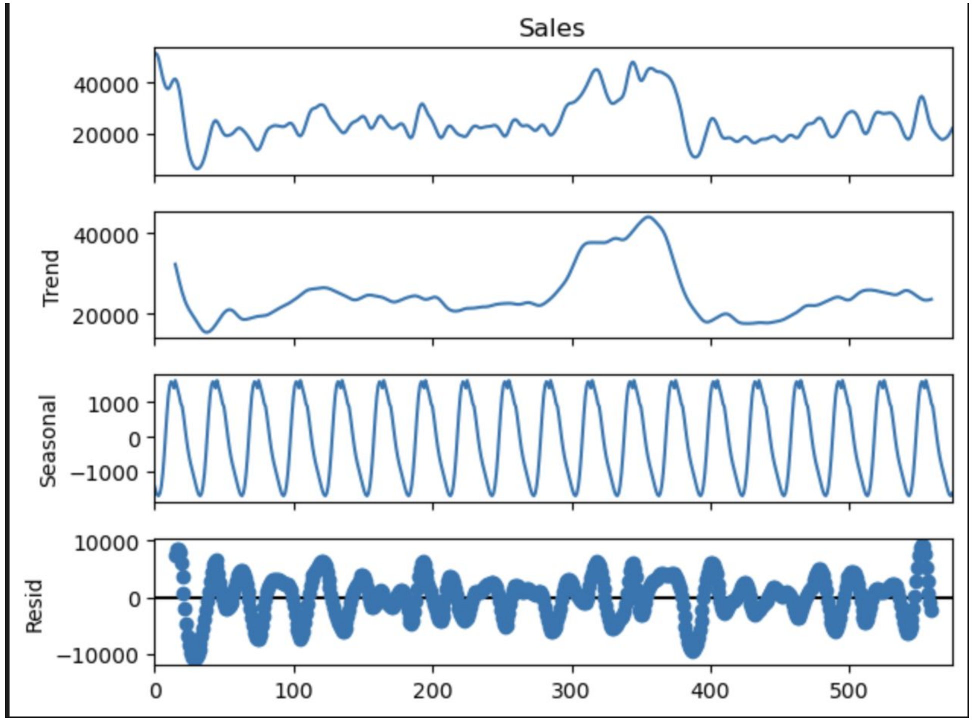
The PACF plot cuts off at lag 1, suggesting an AR(1) process.

Based on this, the ARIMA (1,1,0) model is suitable, with one autoregressive term, one differencing step, and no moving average component.

Arima Model

MAPE: 27.68%



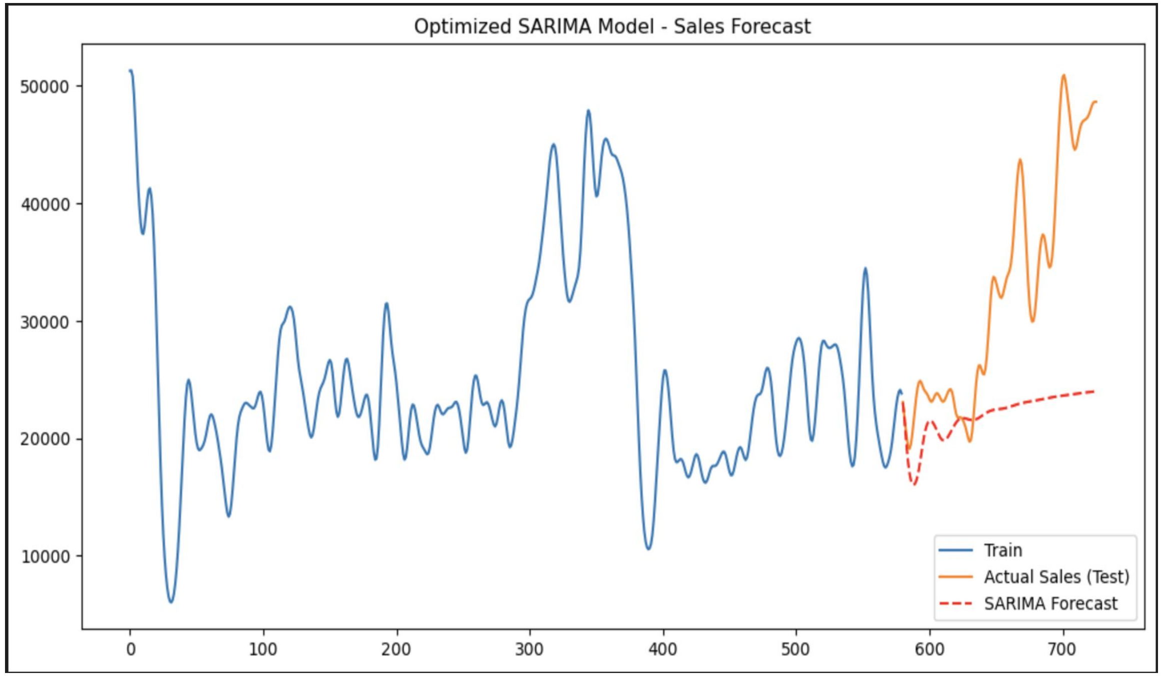


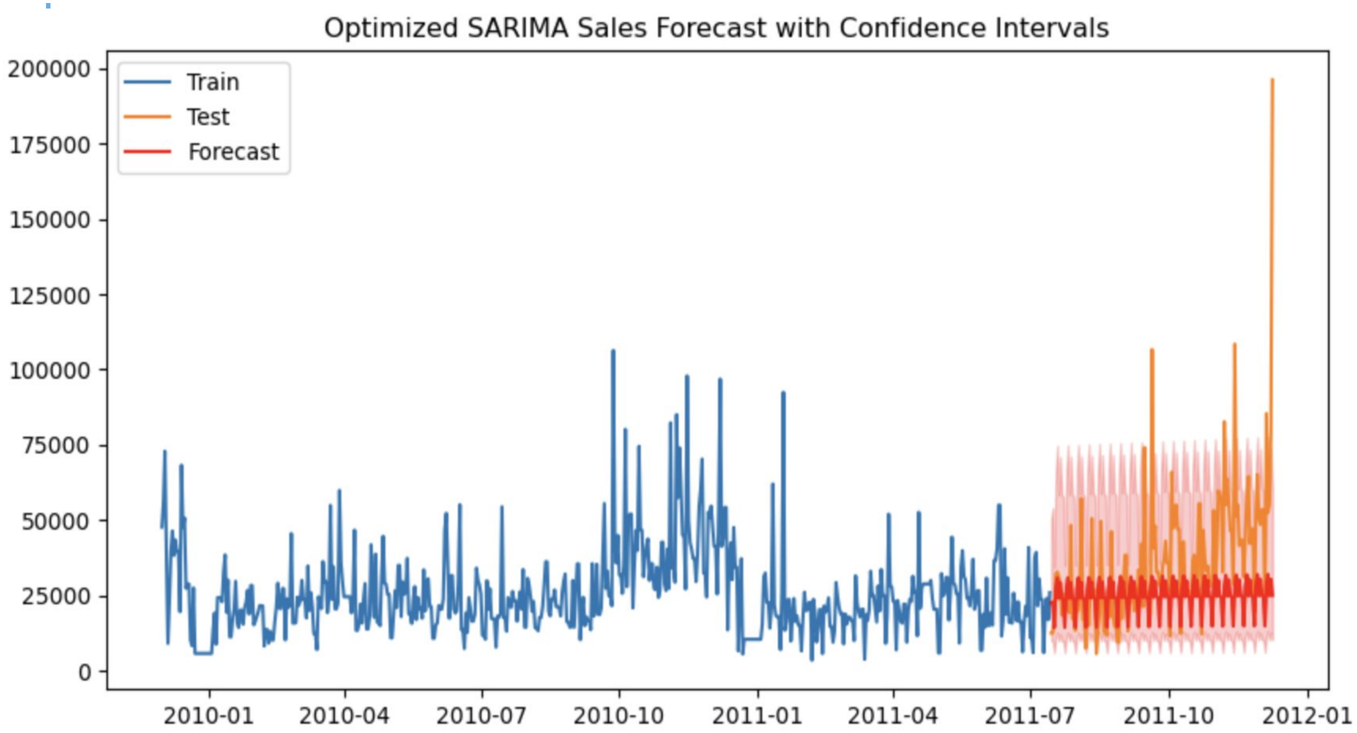
∗Seasonal Decomposition Plot

Optimized SARIMA and SARIMAX Model

MAPE: 39.29%

Accounting for Holidays and Seasonality

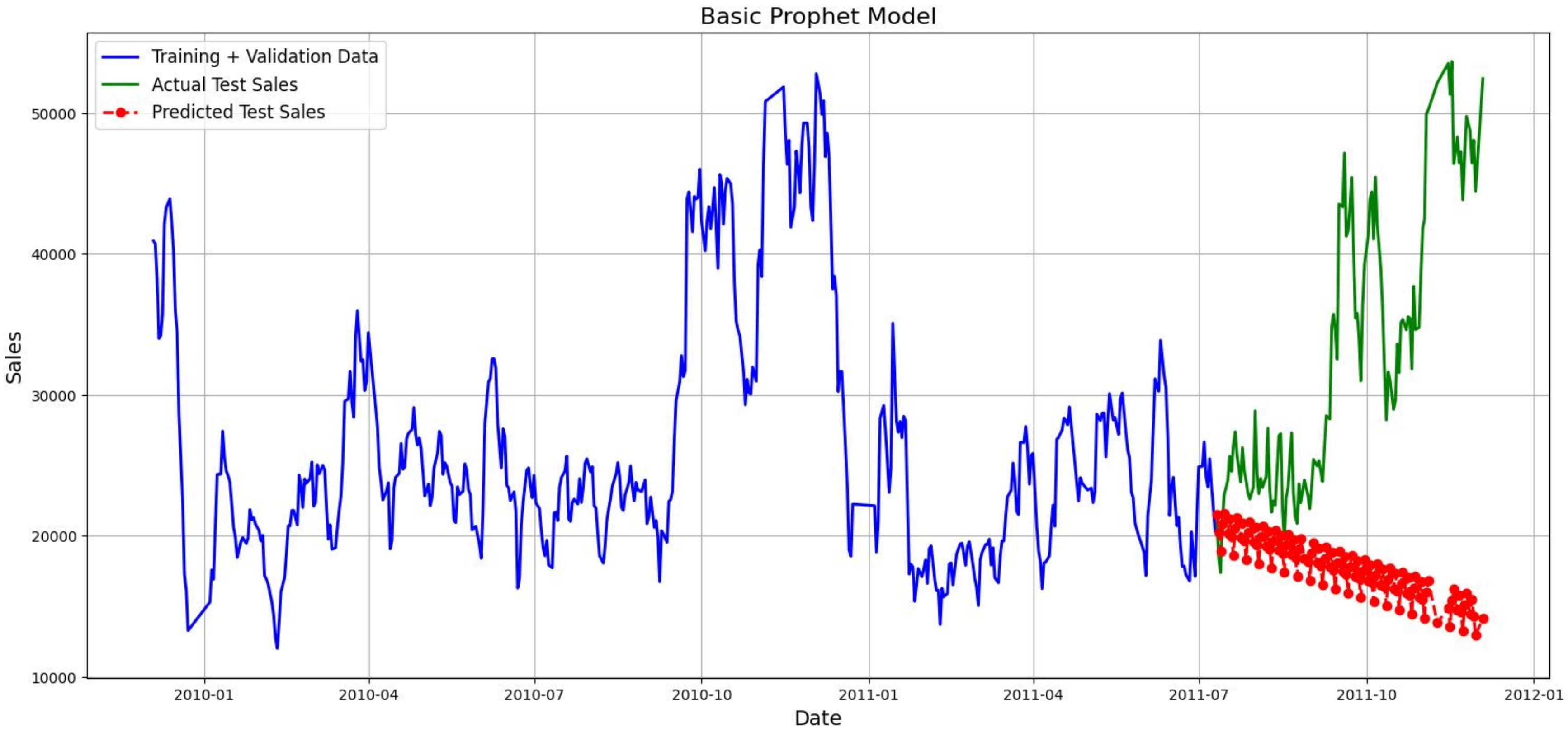




FbProphet

MAPE: 44.71%

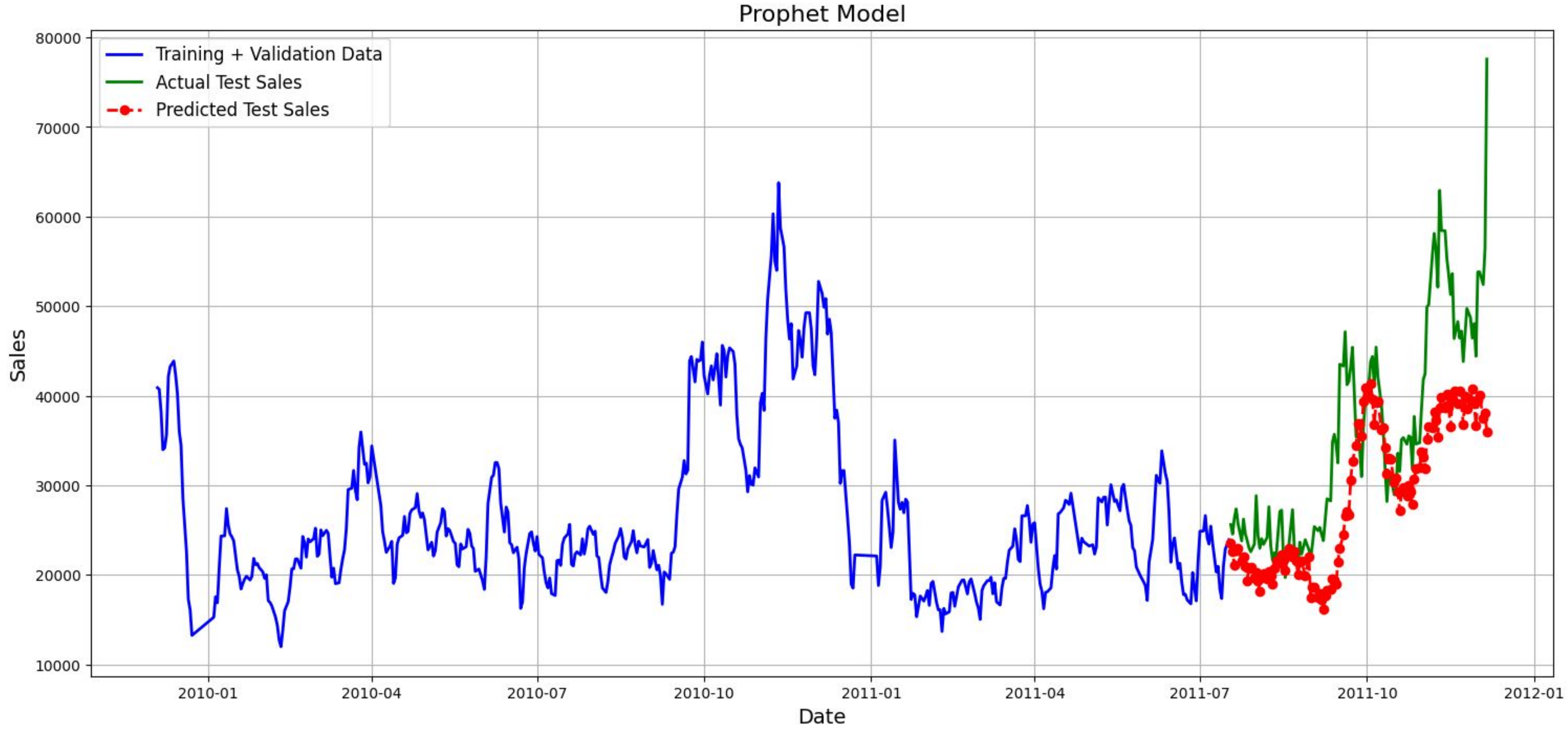
This model is the base model with default parameters.



FbProphet

MAPE: 17.73%

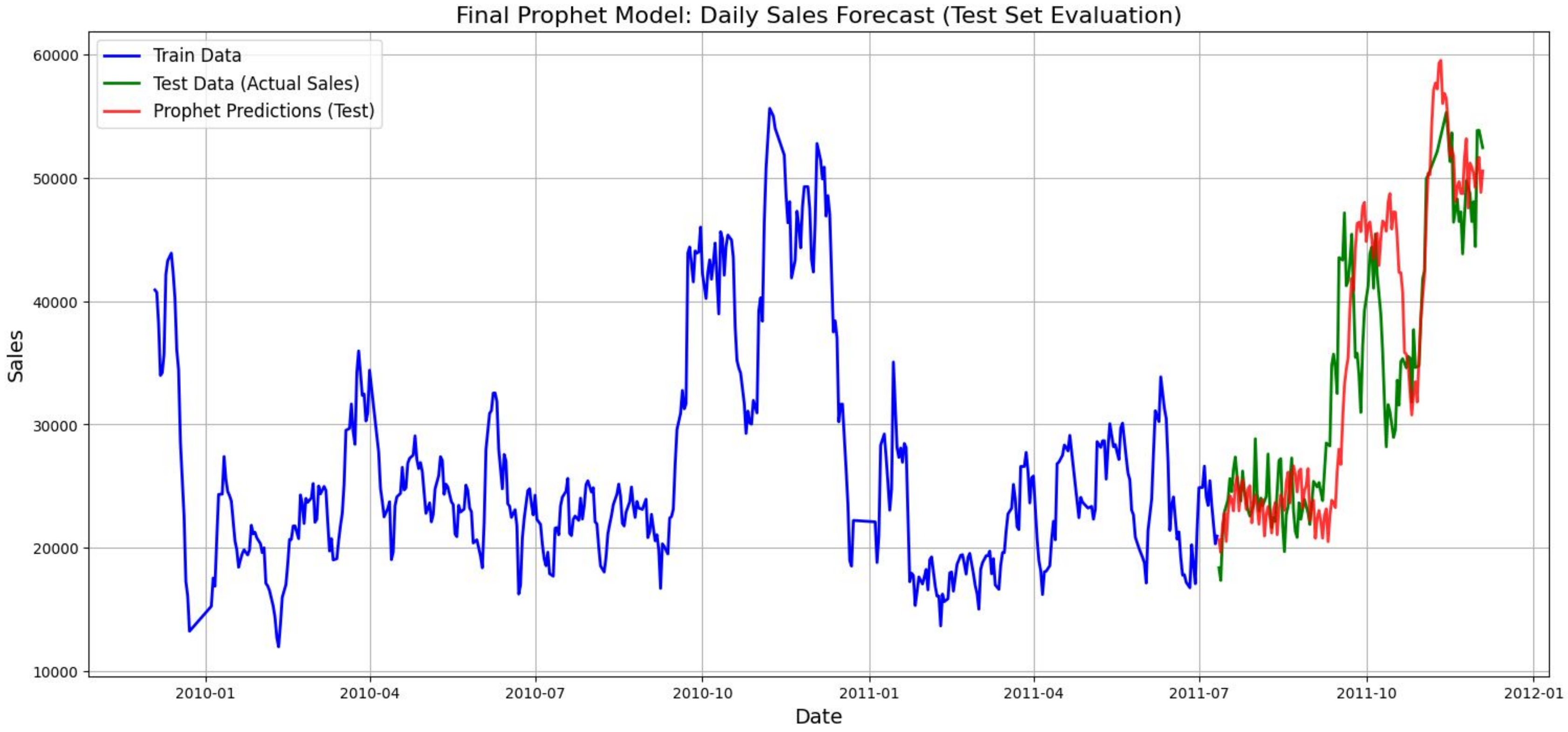
This model includes external regressors and log transformation on Sales.



FbProphet

MAPE: 13.06%

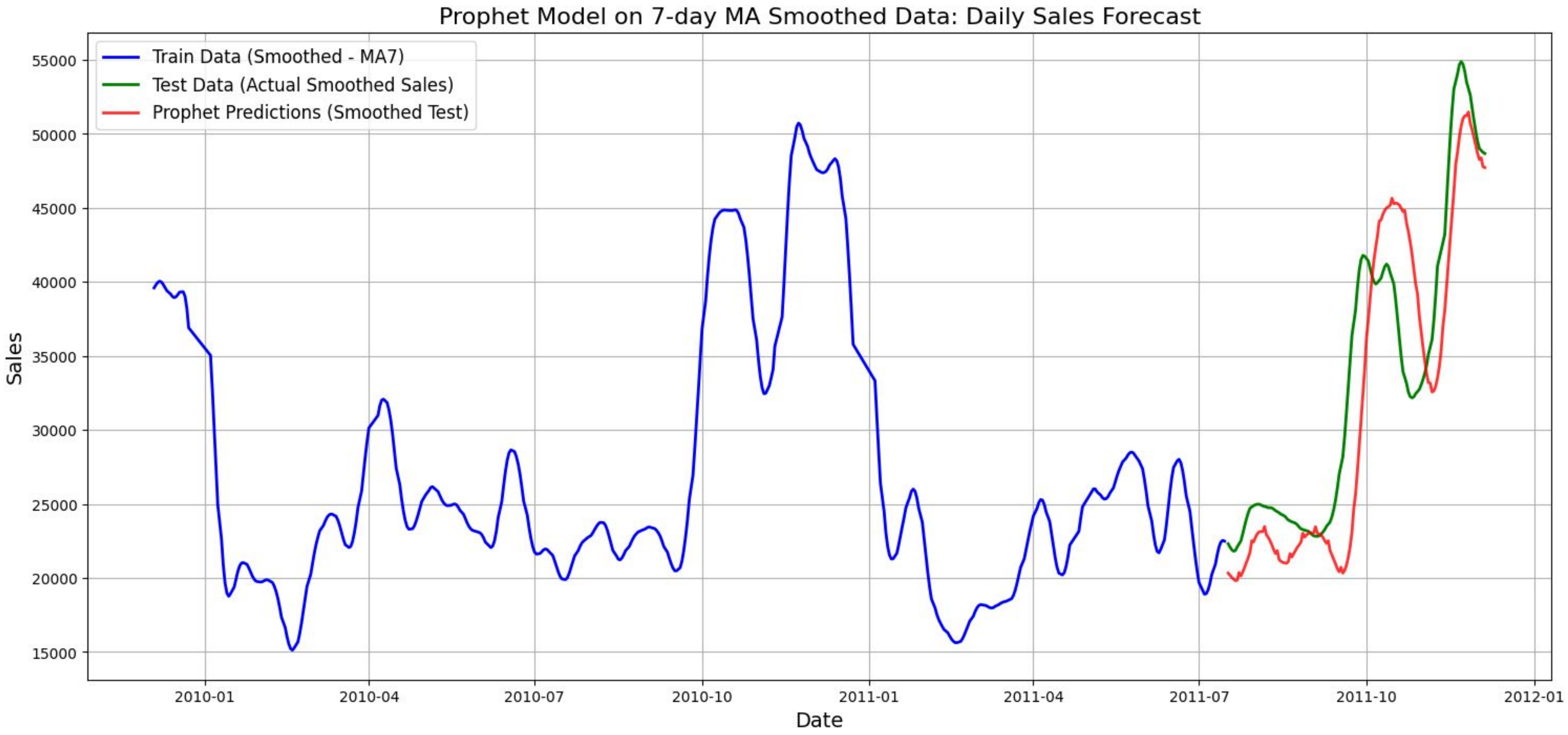
In this model, outliers are removed. Additionally, growth is set to logistic.



FbProphet

MAPE: 12.29%

In this model, Prophet is fit on smoothed data. Data is smoothed using Moving Average with window 7.



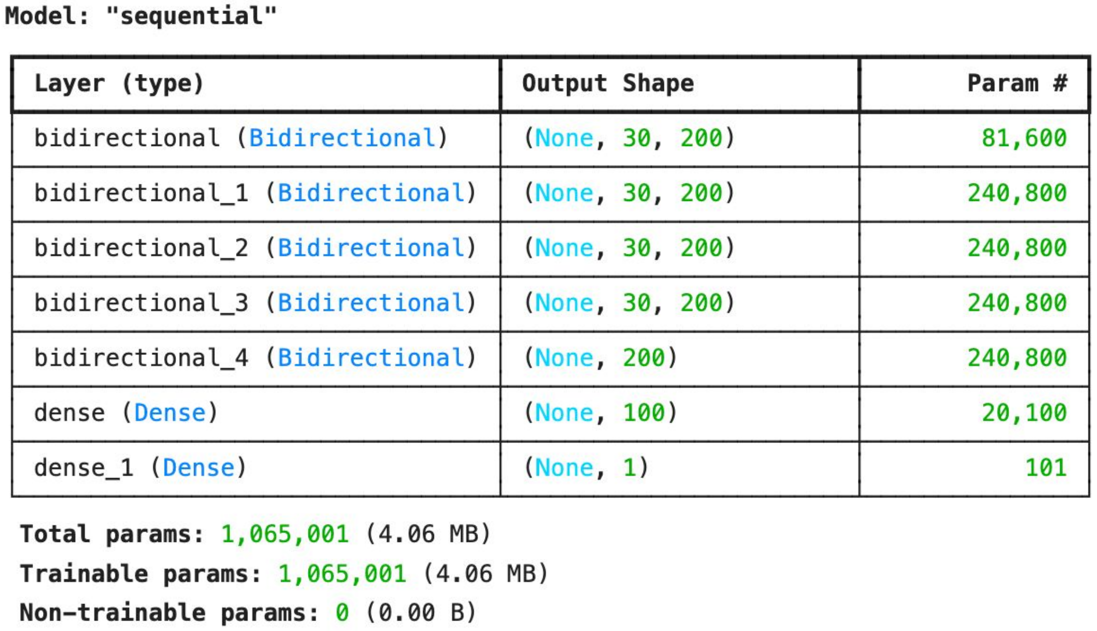
Bidirectional LSTM

A Bidirectional LSTM processes data in both forward and backward directions, capturing context from both past and future states. It is effective in NLP, speech recognition, and time-series analysis.

Key Features:

- Context Awareness – Utilizes information from both directions for better predictions.
- Enhanced Accuracy – Captures dependencies missed by unidirectional models.

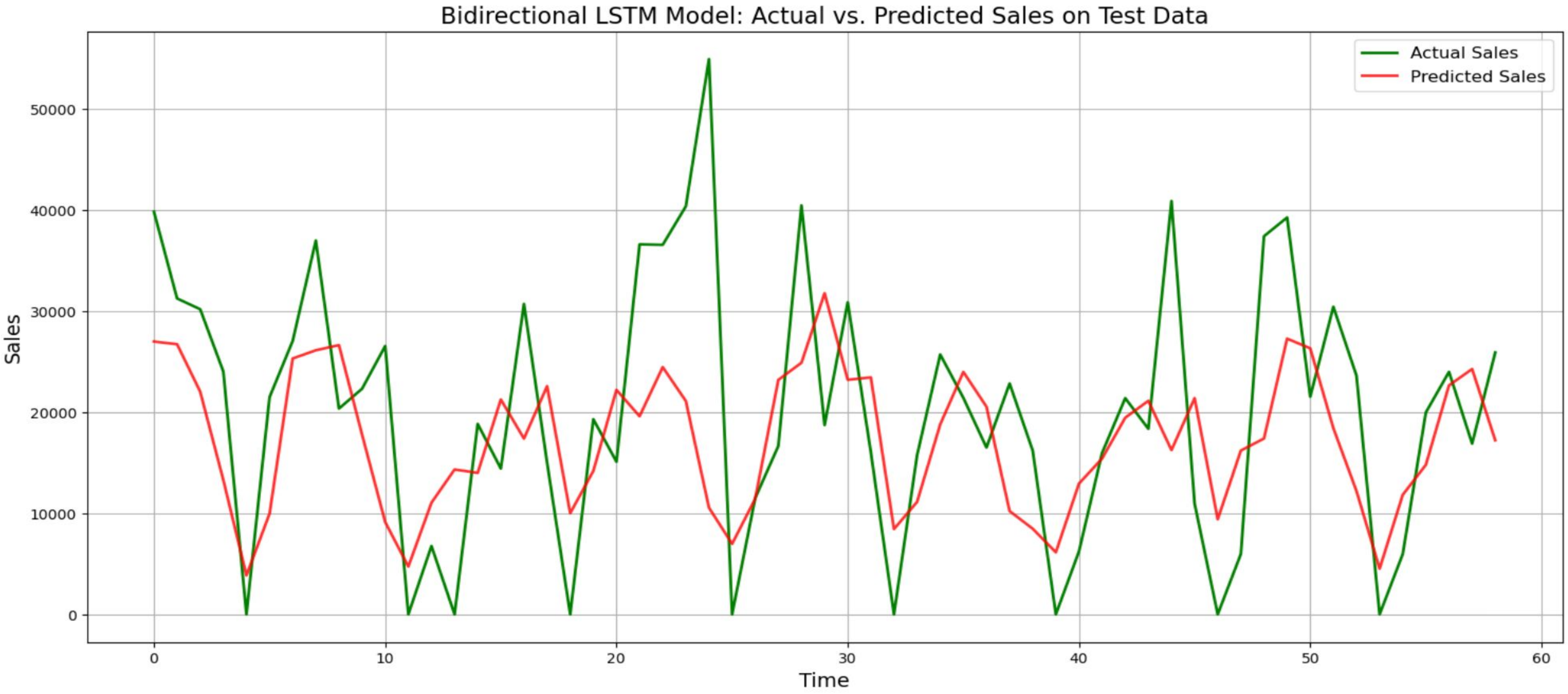
Experiments with different layers, Dropout, and Batch Normalization led to a model with a MAPE of ~39%, serving as a baseline.



```
Model: "sequential"
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ bidirectional (Bidirectional)   │ (None, 30, 200)        │        81,600 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ bidirectional_1 (Bidirectional) │ (None, 30, 200)        │       240,800 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ bidirectional_2 (Bidirectional) │ (None, 30, 200)        │       240,800 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ bidirectional_3 (Bidirectional) │ (None, 30, 200)        │       240,800 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ bidirectional_4 (Bidirectional) │ (None, 200)            │       240,800 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                   │ (None, 100)            │        20,100 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 1)              │           101 │
└─────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 1,065,001 (4.06 MB)
 Trainable params: 1,065,001 (4.06 MB)
 Non-trainable params: 0 (0.00 B)
```

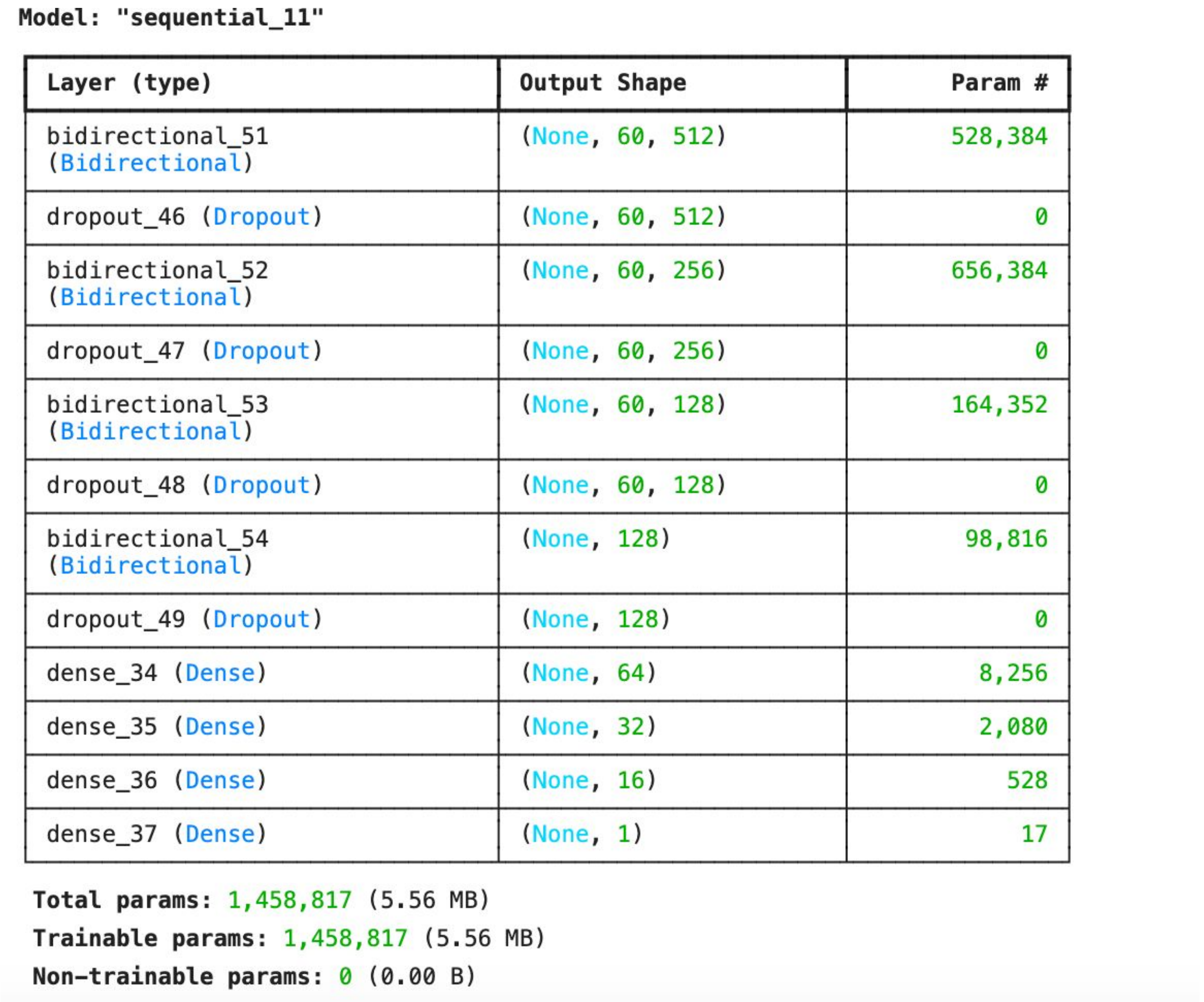
Bidirectional LSTM

MAPE: 39.81%



Improved Bidirectional LSTM

- The input to the model is transformed from daily sales data to a 7-day moving average of sales data.
- The model consists of:
  - 3 bidirectional LSTM layers
  - 3 dense (fully connected) layers
  - Activation functions are mixed and matched across layers, using ‘tanh’ and ‘relu’.
- The dense layers have a progressively decreasing number of neurons, allowing the model to approximate actual values more effectively.
- This approach improves model performance, achieving a Mean Absolute Percentage Error (MAPE) of 9.54%.

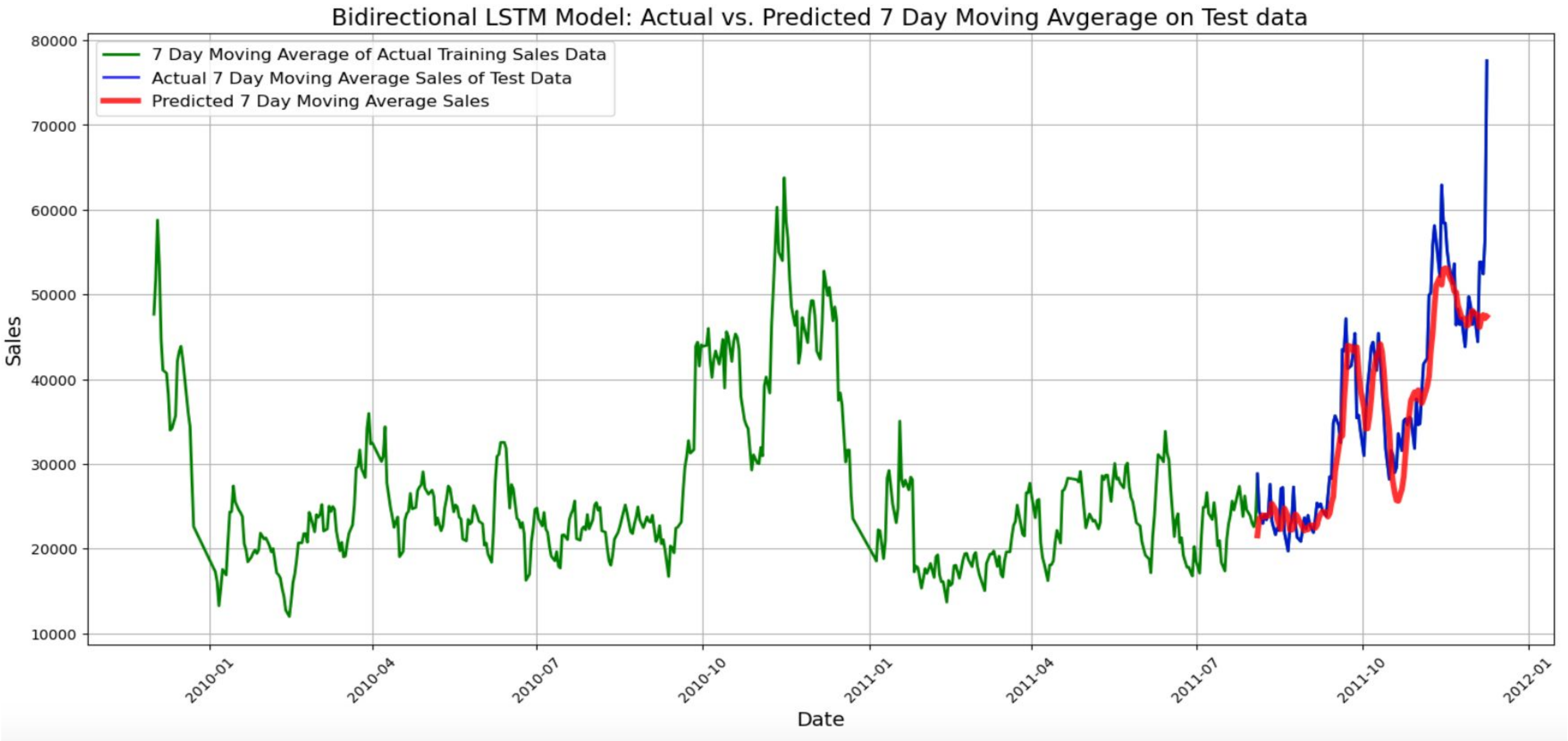


```
Model: "sequential_11"

┃ Layer (type)                    ┃ Output Shape          ┃       Param # ┃
│ bidirectional_51                │ (None, 60, 512)       │       528,384 │
│ (Bidirectional)                 │                       │               │
│ dropout_46 (Dropout)            │ (None, 60, 512)       │             0 │
│ bidirectional_52                │ (None, 60, 256)       │       656,384 │
│ (Bidirectional)                 │                       │               │
│ dropout_47 (Dropout)            │ (None, 60, 256)       │             0 │
│ bidirectional_53                │ (None, 60, 128)       │       164,352 │
│ (Bidirectional)                 │                       │               │
│ dropout_48 (Dropout)            │ (None, 60, 128)       │             0 │
│ bidirectional_54                │ (None, 128)           │        98,816 │
│ (Bidirectional)                 │                       │               │
│ dropout_49 (Dropout)            │ (None, 128)           │             0 │
│ dense_34 (Dense)                │ (None, 64)            │         8,256 │
│ dense_35 (Dense)                │ (None, 32)            │         2,080 │
│ dense_36 (Dense)                │ (None, 16)            │           528 │
│ dense_37 (Dense)                │ (None, 1)             │            17 │

 Total params: 1,458,817 (5.56 MB)
 Trainable params: 1,458,817 (5.56 MB)
 Non-trainable params: 0 (0.00 B)
```

Improved Bidirectional LSTM

MAPE: 9.54%

The daily sales data is modified to 7-day moving average for better model training.

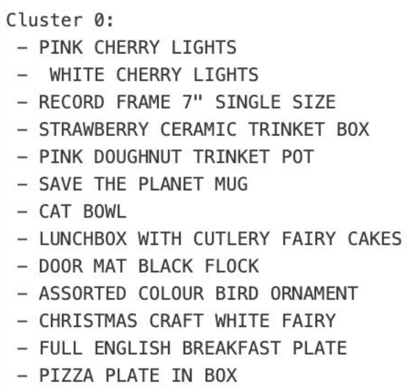


Results

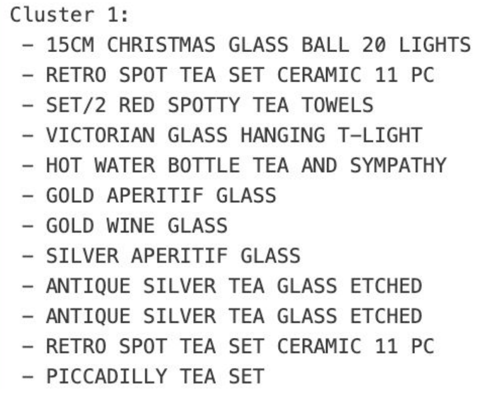
Model	MAPE (%)
ARIMA	37.89
SARIMA	27.68
SARIMAX	39.29
FbProphet	12.29
Bi-LSTM	39.81
Improved Bi-LSTM	9.54

Future Scope

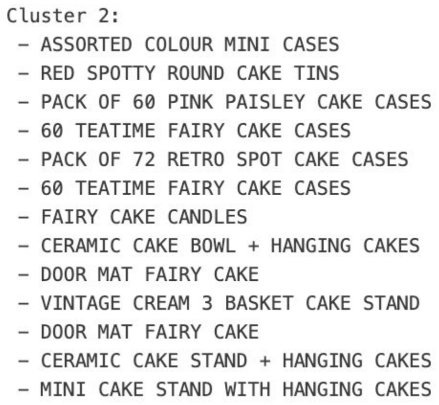
Grouping Product Categories using clustering with Product Descriptions



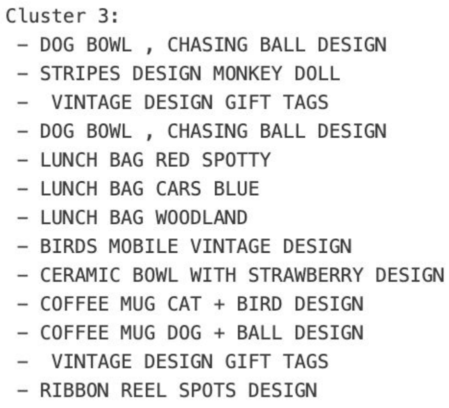
```
Cluster 0:
 - PINK CHERRY LIGHTS
 -  WHITE CHERRY LIGHTS
 - RECORD FRAME 7" SINGLE SIZE
 - STRAWBERRY CERAMIC TRINKET BOX
 - PINK DOUGHNUT TRINKET POT
 - SAVE THE PLANET MUG
 - CAT BOWL
 - LUNCHBOX WITH CUTLERY FAIRY CAKES
 - DOOR MAT BLACK FLOCK
 - ASSORTED COLOUR BIRD ORNAMENT
 - CHRISTMAS CRAFT WHITE FAIRY
 - FULL ENGLISH BREAKFAST PLATE
 - PIZZA PLATE IN BOX
```



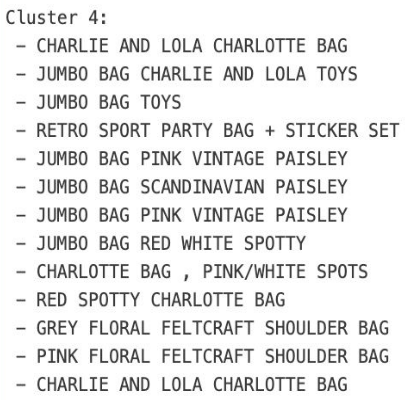
```
Cluster 1:
 - 15CM CHRISTMAS GLASS BALL 20 LIGHTS
 - RETRO SPOT TEA SET CERAMIC 11 PC
 - SET/2 RED SPOTTY TEA TOWELS
 - VICTORIAN GLASS HANGING T-LIGHT
 - HOT WATER BOTTLE TEA AND SYMPATHY
 - GOLD APERITIF GLASS
 - GOLD WINE GLASS
 - SILVER APERITIF GLASS
 - ANTIQUE SILVER TEA GLASS ETCHED
 - ANTIQUE SILVER TEA GLASS ETCHED
 - RETRO SPOT TEA SET CERAMIC 11 PC
 - PICCADILLY TEA SET
```



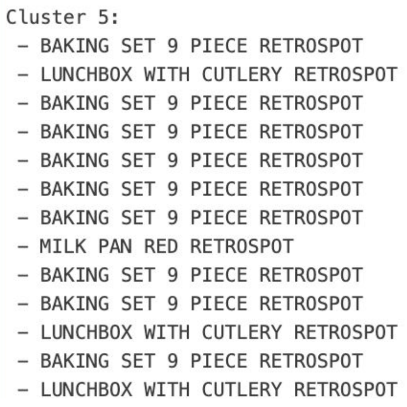
```
Cluster 2:
 - ASSORTED COLOUR MINI CASES
 - RED SPOTTY ROUND CAKE TINS
 - PACK OF 60 PINK PAISLEY CAKE CASES
 - 60 TEATIME FAIRY CAKE CASES
 - PACK OF 72 RETRO SPOT CAKE CASES
 - 60 TEATIME FAIRY CAKE CASES
 - FAIRY CAKE CANDLES
 - CERAMIC CAKE BOWL + HANGING CAKES
 - DOOR MAT FAIRY CAKE
 - VINTAGE CREAM 3 BASKET CAKE STAND
 - DOOR MAT FAIRY CAKE
 - CERAMIC CAKE STAND + HANGING CAKES
 - MINI CAKE STAND WITH HANGING CAKES
```



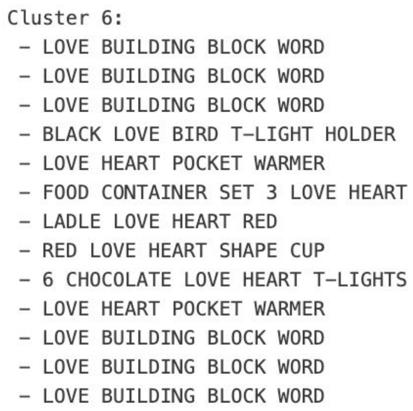
```
Cluster 3:
 - DOG BOWL , CHASING BALL DESIGN
 - STRIPES DESIGN MONKEY DOLL
 -  VINTAGE DESIGN GIFT TAGS
 - DOG BOWL , CHASING BALL DESIGN
 - LUNCH BAG RED SPOTTY
 - LUNCH BAG CARS BLUE
 - LUNCH BAG WOODLAND
 - BIRDS MOBILE VINTAGE DESIGN
 - CERAMIC BOWL WITH STRAWBERRY DESIGN
 - COFFEE MUG CAT + BIRD DESIGN
 - COFFEE MUG DOG + BALL DESIGN
 -  VINTAGE DESIGN GIFT TAGS
 - RIBBON REEL SPOTS DESIGN
```



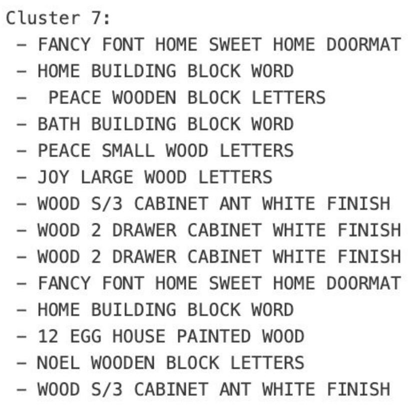
```
Cluster 4:
 - CHARLIE AND LOLA CHARLOTTE BAG
 - JUMBO BAG CHARLIE AND LOLA TOYS
 - JUMBO BAG TOYS
 - RETRO SPORT PARTY BAG + STICKER SET
 - JUMBO BAG PINK VINTAGE PAISLEY
 - JUMBO BAG SCANDINAVIAN PAISLEY
 - JUMBO BAG PINK VINTAGE PAISLEY
 - JUMBO BAG RED WHITE SPOTTY
 - CHARLOTTE BAG , PINK/WHITE SPOTS
 - RED SPOTTY CHARLOTTE BAG
 - GREY FLORAL FELTCRAFT SHOULDER BAG
 - PINK FLORAL FELTCRAFT SHOULDER BAG
 - CHARLIE AND LOLA CHARLOTTE BAG
```



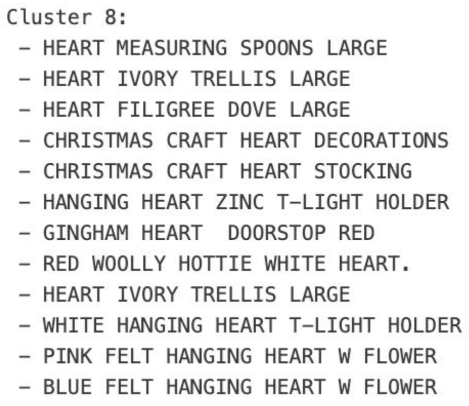
```
Cluster 5:
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - MILK PAN RED RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
 - BAKING SET 9 PIECE RETROSPOT
 - LUNCHBOX WITH CUTLERY RETROSPOT
```



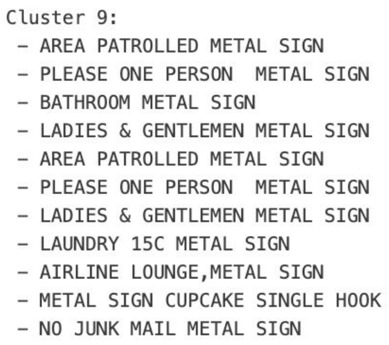
```
Cluster 6:
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - BLACK LOVE BIRD T-LIGHT HOLDER
 - LOVE HEART POCKET WARMER
 - FOOD CONTAINER SET 3 LOVE HEART
 - LADLE LOVE HEART RED
 - RED LOVE HEART SHAPE CUP
 - 6 CHOCOLATE LOVE HEART T-LIGHTS
 - LOVE HEART POCKET WARMER
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
 - LOVE BUILDING BLOCK WORD
```



```
Cluster 7:
 - FANCY FONT HOME SWEET HOME DOORMAT
 - HOME BUILDING BLOCK WORD
 -  PEACE WOODEN BLOCK LETTERS
 - BATH BUILDING BLOCK WORD
 - PEACE SMALL WOOD LETTERS
 - JOY LARGE WOOD LETTERS
 - WOOD S/3 CABINET ANT WHITE FINISH
 - WOOD 2 DRAWER CABINET WHITE FINISH
 - WOOD 2 DRAWER CABINET WHITE FINISH
 - FANCY FONT HOME SWEET HOME DOORMAT
 - HOME BUILDING BLOCK WORD
 - 12 EGG HOUSE PAINTED WOOD
 - NOEL WOODEN BLOCK LETTERS
 - WOOD S/3 CABINET ANT WHITE FINISH
```



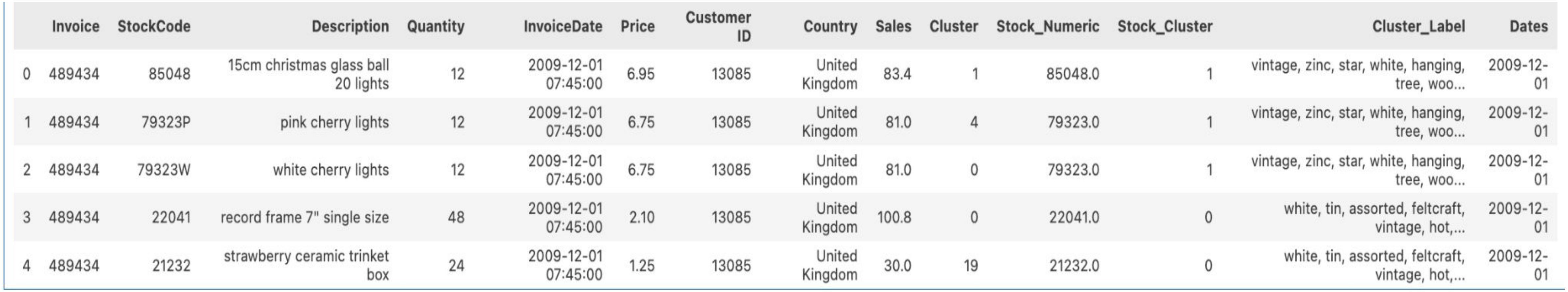
```
Cluster 8:
 - HEART MEASURING SPOONS LARGE
 - HEART IVORY TRELLIS LARGE
 - HEART FILIGREE DOVE LARGE
 - CHRISTMAS CRAFT HEART DECORATIONS
 - CHRISTMAS CRAFT HEART STOCKING
 - HANGING HEART ZINC T-LIGHT HOLDER
 - GINGHAM HEART  DOORSTOP RED
 - RED WOOLLY HOTTIE WHITE HEART.
 - HEART IVORY TRELLIS LARGE
 - WHITE HANGING HEART T-LIGHT HOLDER
 - PINK FELT HANGING HEART W FLOWER
 - BLUE FELT HANGING HEART W FLOWER
```



```
Cluster 9:
 - AREA PATROLLED METAL SIGN
 - PLEASE ONE PERSON  METAL SIGN
 - BATHROOM METAL SIGN
 - LADIES & GENTLEMEN METAL SIGN
 - AREA PATROLLED METAL SIGN
 - PLEASE ONE PERSON  METAL SIGN
 - LADIES & GENTLEMEN METAL SIGN
 - LAUNDRY 15C METAL SIGN
 - AIRLINE LOUNGE,METAL SIGN
 - METAL SIGN CUPCAKE SINGLE HOOK
 - NO JUNK MAIL METAL SIGN
```

Future Scope

Clustering by Stock Code



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Sales	Cluster	Stock_Numeric	Stock_Cluster	Cluster_Label	Dates
0	489434	85048	15cm christmas glass ball 20 lights	12	2009-12-01 07:45:00	6.95	13085	United Kingdom	83.4	1	85048.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
1	489434	79323P	pink cherry lights	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	4	79323.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
2	489434	79323W	white cherry lights	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	0	79323.0	1	vintage, zinc, star, white, hanging, tree, woo...	2009-12-01
3	489434	22041	record frame 7" single size	48	2009-12-01 07:45:00	2.10	13085	United Kingdom	100.8	0	22041.0	0	white, tin, assorted, feltcraft, vintage, hot,...	2009-12-01
4	489434	21232	strawberry ceramic trinket box	24	2009-12-01 07:45:00	1.25	13085	United Kingdom	30.0	19	21232.0	0	white, tin, assorted, feltcraft, vintage, hot,...	2009-12-01

We will apply hyperparameter tuning and cross-validation to optimize model performance and identify weaknesses.

Time series analysis on clusters will help analyze sales trends, while incorporating external factors like economic indicators, unlisted holidays, and discounts will enhance forecasting.

Additionally, we will explore alternative models such as BERTopic, XGBoost, and DBSCAN to better capture fluctuating sales trends.

Conclusion

- The project explored statistical and deep learning models to forecast daily sales in the UK.
- ARIMA and SARIMA struggled with capturing the complexity of sales patterns due to assumptions about stationarity and linearity.
  - ARIMA MAPE: 37.89%
  - SARIMA MAPE: 27.68%
- Facebook Prophet (MAPE: 12.29%) performed well, effectively handling seasonality, trend shifts, and UK holidays.
- Bi-LSTM underperformed (MAPE: 39.81%), failing to capture sales patterns effectively.
- Improved LSTM (MAPE: 9.54%) was the most accurate, demonstrating the potential of deep learning when properly tuned.
- Conclusion: The Improved LSTM provided the best predictions, but Facebook Prophet remains a strong alternative due to its flexibility and interpretability.

Thank You!