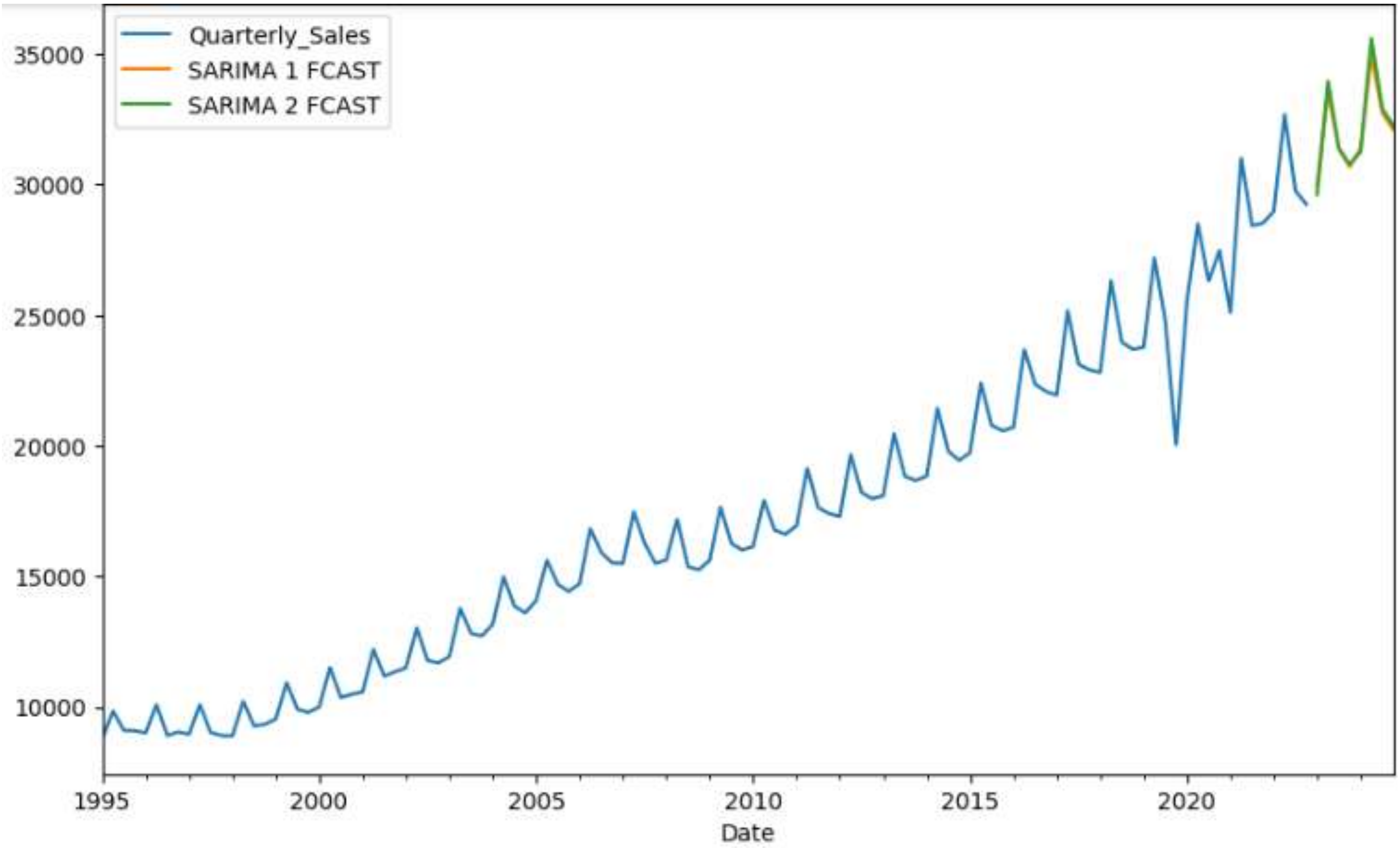
Forecasting Future Trends: A Time Series Analysis of Retail Sales in New Zealand



Project Overview

Welcome to my time series forecasting project, inspired by the practical insights gained from the book "Time Series Forecasting in Python." Here, I apply the knowledge acquired from the book to predict retail sales trends in New Zealand over the next two years.

This project focuses on applying theoretical concepts from the book *"Time Series Forecasting in Python"* to real-world scenarios. By implementing Python-based techniques recommended in the book, I aim to actively utilize and demonstrate the acquired skills. The specific focus is on forecasting retail sales, utilizing Stats NZ data and following the step-by-step approach outlined in the book written by Marco Peixeiro.

Book: Peixeiro, Marco. *Time Series Forecasting in Python.* Manning Publications, 2022.

Project Structure

The project is structured with the following sections:

Import Libraries: Set up necessary tools for analysis.

Load the Data: Acquire the quarterly retail sales dataset.

Data Preprocessing: Clean and prepare the data.

Inspect and Visualize: Explore and visualize dataset characteristics.

Identifying Seasonal Patterns: Analyze for recurring patterns, especially seasonality.

Baseline Model: Establish a baseline for performance reference.

Predictions and MAPE: Generate initial predictions and calculate MAPE.

SARIMA Model: Implement SARIMA with ADF test for stationarity.

Train and Test Sets: Divide data for model validation.

Optimize SARIMA Function: Fine-tune parameters for optimal performance.

Run Optimized Function and Residual Analysis: Apply optimized SARIMA function and analyze residuals.

Predictions and MAPE: Generate SARIMA predictions and calculate MAPE.

Forecast into the Future: Extend SARIMA predictions for future trends.

LSTM RNN Model: Implement LSTM Recurrent Neural Network.

Scale the Data and Data Windowing: Scale data and apply windowing for LSTM.

Building and Fitting the Model: Construct and train the LSTM model.

Predictions and MAPE: Generate LSTM predictions and calculate MAPE.

Forecast into the Future: Extend LSTM predictions for future trends.

Conclusions: Summarize findings and insights.

MAPE Comparison and Final Predictions: Compare MAPE values between SARIMA and LSTM models and present final predictions.

Documentation

Questions from the Datasheets for Datasets paper, v7.

Sections:

- Motivation and Composition
- Collection process
- Preprocessing/cleaning
- Uses and Distribution
- Maintenance
- Conclusion

Dataset Composition

1- For what purpose was the dataset created?

2 - Who created the dataset?

The dataset, managed by Stats NZ, New Zealand's official data agency, tracks quarterly retail sales in the country. The numbers, expressed in millions and at current prices, cover 114 data points from 1995 to 2023.

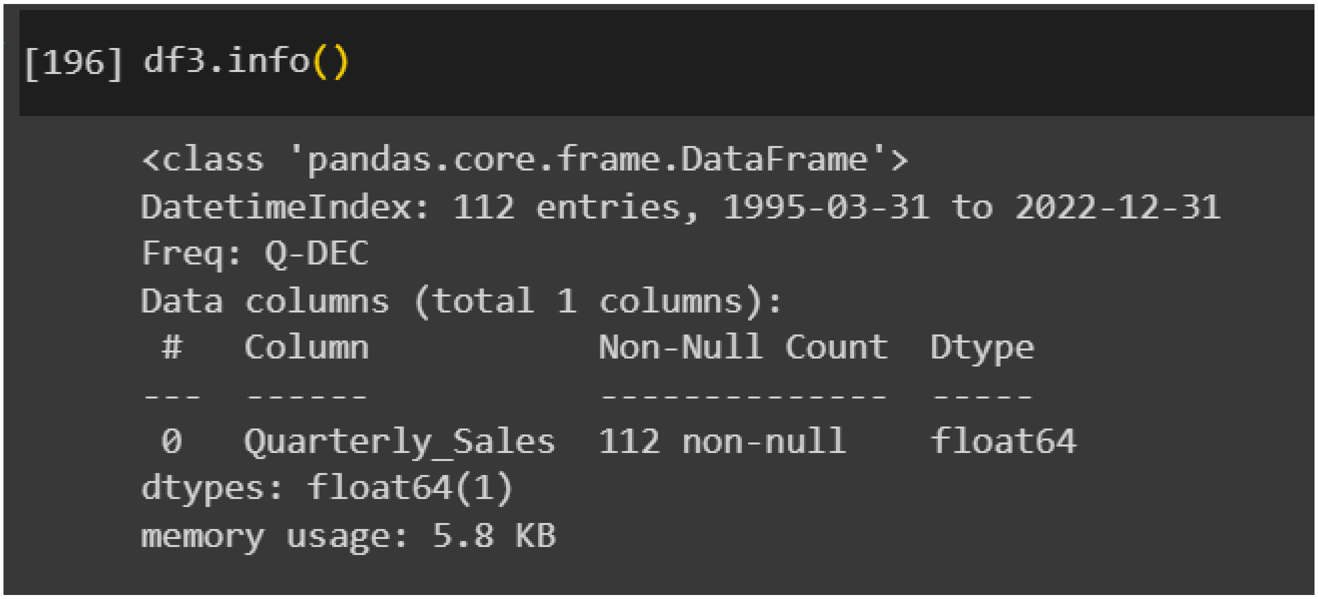
3 - What do the instances that comprise the dataset represent?

The dataset used in the project contains a single target variable, the quarterly sales (in millions at current prices) of the Retail Trade category.

The Retail Trade category includes a diverse range of businesses, they are:

- Supermarket and grocery stores
- Specialized food retailing (excluding liquor)
- Liquor retailing
- Non-store and commission based retailing
- Department stores
- Furniture, floor coverings, houseware and textile goods retailing
- Hardware, building and garden supplies
- Recreational goods retailing
- Clothing, footwear and personal accessory retailing
- Electrical and electronic goods retailing
- Pharmaceutical and other store based retailing
- Motor vehicle and parts retailing
- Fuel retailing
- Accommodation
- Food and beverage services

4 - How many instances are there in total?
The dataset contains 114 entries, capturing data up to the first and second quarters of 2023. However, for the project's simplicity and to work exclusively with full-year data, the analysis focuses on the 112 data points available up to the fourth quarter of 2022.



```
[196] df3.info()

      <class 'pandas.core.frame.DataFrame'>
      DatetimeIndex: 112 entries, 1995-03-31 to 2022-12-31
      Freq: Q-DEC
      Data columns (total 1 columns):
       #   Column           Non-Null Count  Dtype
      ---  ------           --------------  -----
       0   Quarterly_Sales  112 non-null    float64
      dtypes: float64(1)
      memory usage: 5.8 KB
```

5 - Does the dataset contain all possible instances?

The dataset contains all possible instances, totaling 112 data points, representing the quarterly sales and their corresponding dates.

6 - What data does each instance consist of?

Each instance includes quarterly sales in millions and their corresponding dates.

7 - Is there a label or target associated with each instance?

In this scenario, the target variable is represented by the quarterly sales.

8 - Is any information missing from individual instances?

There is no information missing.

9 - Are relationships between individual instances made explicit?

As mentioned earlier, the variable in this scenario is the target—quarterly sales. Prior to undertaking the project, it is presumed that a relationship exists between past values of quarterly sales and future values.

10 - Are there recommended data splits?

Following the guidance from the book, data split involves reserving the last full year of the series (2022) for the test set, with the remainder used as the training set.

11 - Are there any errors, sources of noise, or redundancies in the dataset?

There are no errors in the dataset.

12 - Does the dataset contain data that might be considered confidential?

No, the dataset does not contain confidential information. The figures represent the total sales in the entire industry, and there is no reference to individual businesses.

13 - Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No, the dataset does not contain offensive data.

14 - Does the dataset relate to people?

No, the dataset is related to quarterly retail sales and does not involve information about individuals.

Data Collection Process

15 - How was the data associated with each instance acquired?

16 - What mechanisms or procedures were used to collect the data?

The data was downloaded from Stats NZ Tatauranga Aotearoa webpage. Stats NZ Tatauranga Aotearoa is New Zealand's official data agency.

17 - Over what timeframe was the data collected?
The data was downloaded on October 2023.

Data Cleaning

18 - Was any cleaning/preprocessing of the data done?

As shown in the notebook, on section DATA PREPROCESSING, the tasks done were:

- Define a DataFrame with relevant data.
- Rename columns.
- Convert the date column into date time object.
- Set the date time object frequency.
- Set the date time object as the index.

19 - Was the “raw” data saved in addition to the cleaned data?

Yes, using the method .copy()

Uses and Distribution

20 - Is there a repository that links to any or all papers or systems that use the dataset? Link: https://www.stats.govt.nz

21 - Has the dataset been used for any tasks already?

22 - Is there anything about the composition of the dataset or the way it was collected and cleaned that might impact future uses?

23 - Are there tasks for which the dataset should not be used?

24 - Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

25 - How will the dataset will be distributed (e.g., zip file, website, GitHub)?

26 - How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

27 - Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

28 - If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset was created and is managed and maintained by Stats NZ (https://www.stats.govt.nz).

To access the dataset, visit the Stats NZ webpage and navigate to the 'Retail Trade (ANZSIC06) - RTT' group. Look for the 'Sales and stocks by industry, in current and constant prices (SAFC)' table, updated quarterly in March, June, September, and December

Conclusion

Models performance

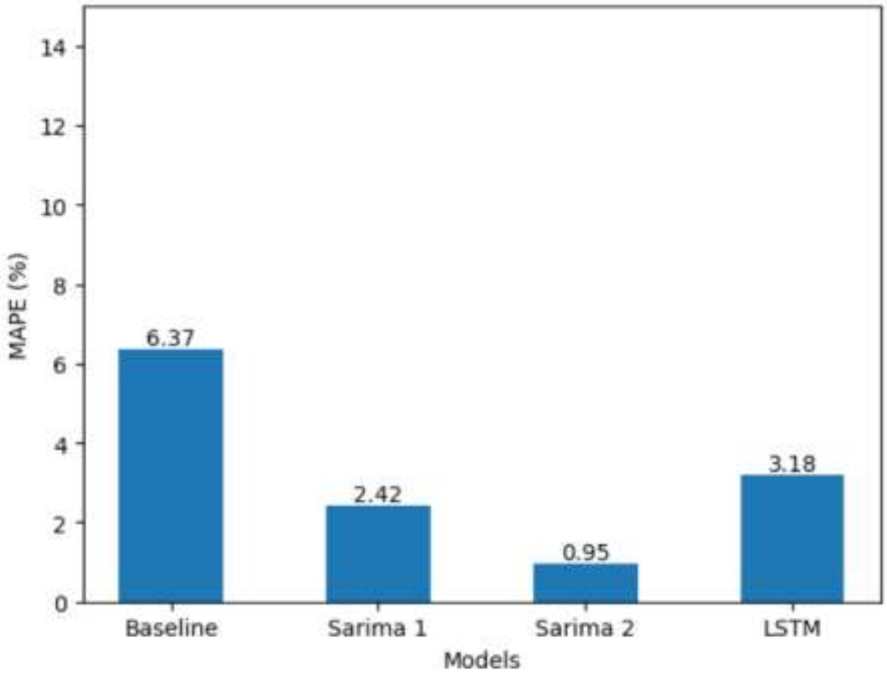
The metric used to evaluate the performance was Mean Absolute Percentage Error **(**MAPE**).**

SARIMA 1 **MAPE = %2.42**

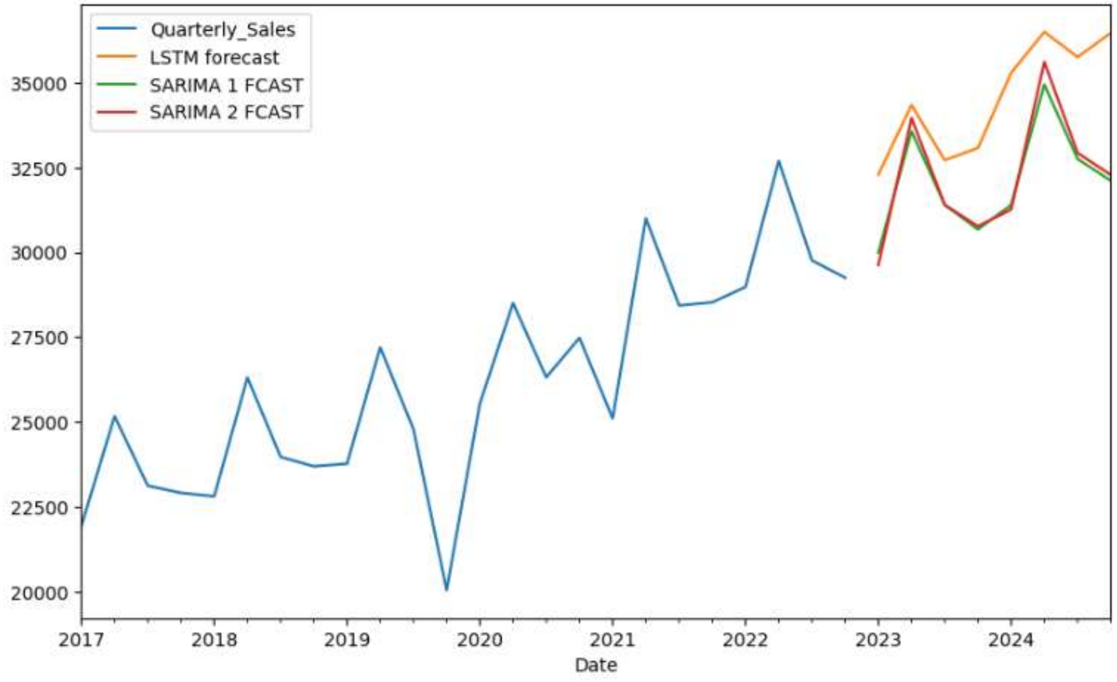
SARIMA 2 **MAPE = %0.95**

LSTM **MAPE = %2.5 – %5.5**

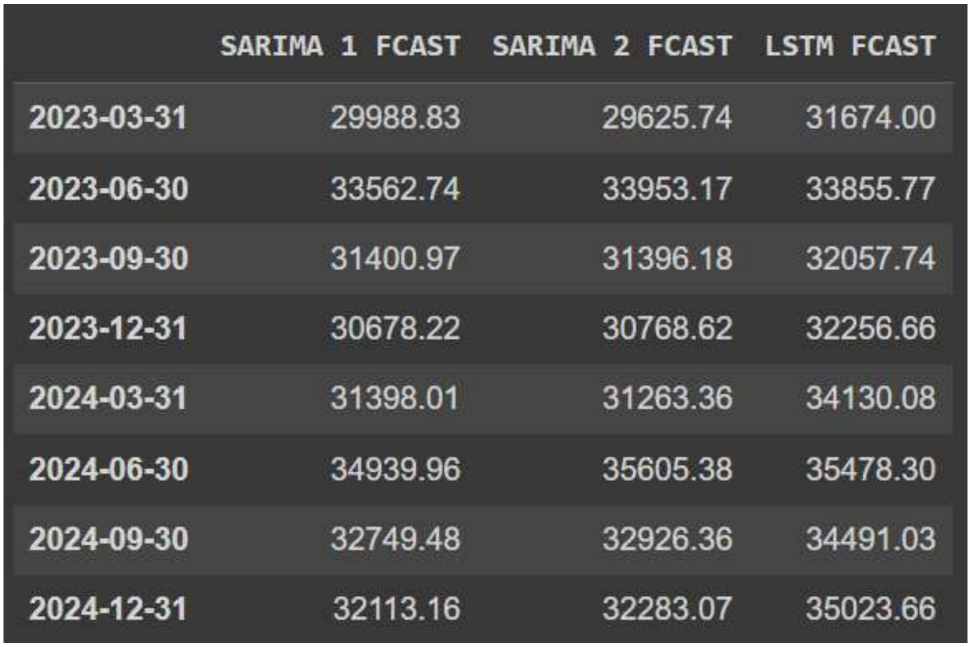
Models performance



Forecast plot



Forecast chart



	SARIMA 1 FCAST	SARIMA 2 FCAST	LSTM FCAST
2023-03-31	29988.83	29625.74	31674.00
2023-06-30	33562.74	33953.17	33855.77
2023-09-30	31400.97	31396.18	32057.74
2023-12-31	30678.22	30768.62	32256.66
2024-03-31	31398.01	31263.36	34130.08
2024-06-30	34939.96	35605.38	35478.30
2024-09-30	32749.48	32926.36	34491.03
2024-12-31	32113.16	32283.07	35023.66

Summary:

The results showed that SARIMA 2 was the most accurate, followed closely by SARIMA 1. Even though LSTM faced challenges due to our smaller dataset, it still provided useful predictions.

SARIMA models performed better here, but including LSTM allowed us to see how it stacks up in a scenario where it's not ideal. This comparison helps us understand which models might work best based on the data we have. Further tweaking and exploring different models could improve predictions, especially with more extensive datasets.