Competence Development, GCGC

Join the Ideathon Challenge

Machine Learning
(Problem Statement – 2)

The Time Series Thinkers

122010309030 – PVG Harshita
122010327009 – Sourit Maji
122010309017 – P Vishnu Vardhan

Problem Statement: A retail company wants to forecast the sales of their products for the next six months based on their historical sales data. The company has collected daily sales data for each product for the past two years. The goal of this project is to develop a machine learning model that can accurately forecast the sales of each product for the next six months, and to identify the key factors that influence sales.

Data Description: The dataset contains the following features:

- Date: the date on which the sales were recorded.
- Product ID: a unique identifier for each product.
- Sales: the number of units sold for each product on a given day.

Tasks:

1. Exploratory Data Analysis (EDA): Perform exploratory data analysis to understand the distribution of variables, identify missing values, outliers, and correlations.

2. Time Series Analysis: Conduct time series analysis to identify trends, seasonality, and other patterns in the sales data.

3. Feature Engineering: Based on EDA and time series analysis, engineer relevant features from the available data to improve the performance of the model.

4. Data Preprocessing: Clean and preprocess the data, handle missing values and outliers, and split the data into training and test sets.

5. Model Selection: Select appropriate machine learning algorithms (such as ARIMA, SARIMA, Prophet, or LSTM) and tune their hyperparameters using cross-validation.

6. Model Evaluation: Evaluate the performance of the model using various metrics such as mean absolute error, mean squared error, and R-squared.

7. Interpretation: Interpret the results of the model to identify the key factors that influence sales, and provide actionable insights to the company.

8. Deployment: Deploy the model to a production environment and test its performance on new data.

Dataset: **https://www.kaggle.com/competitions/store-sales-time-series-forecasting**

Table of Contents

Introduction

Abstract

This project aims to develop a machine learning model to accurately forecast the sales of products for the next six months based on the historical sales data collected by a retail company for the past two years. The project also seeks to identify the key factors that influence sales. The model will be developed using a variety of statistical and machine learning techniques, and will be evaluated using various metrics. The results of the project will be useful for the retail company to make informed decisions about inventory management, production planning, and resource allocation, among other things.

Introduction

In today's highly competitive retail industry, accurate sales forecasting plays a crucial role in strategic decision-making, inventory management, and overall business success. Time series forecasting models have been widely used in the retail industry to predict future sales trends based on historical data. In this project, we aim to develop a time series forecasting model using the ARIMA algorithm to forecast the sales of a retail company for the next six months. We will also focus on identifying the key factors that influence sales and adjust the components of the model accordingly to improve its forecasting accuracy. The ultimate goal of this project is to provide a reliable and accurate sales forecast to the retail company, which will help them make informed decisions and stay ahead of their competition.

Literature Review

Time series forecasting is an important area of research in machine learning and has numerous practical applications. Many different models have been proposed for time series forecasting, including ARIMA, SARIMA, Prophet, LSTM, and more.

Each model has its own strengths and weaknesses and is suitable for different types of time series data. Several techniques have been proposed to improve the accuracy of time series forecasting, including data pre-processing, feature engineering, and model ensembling. Overall, time series forecasting remains an active area of research, and there is ongoing work to develop more accurate and efficient models for this task.

Dataset Description

Train.csv

The training data, comprising time series of features store_nbr, family, and on promotion as well as the target sales.
store_nbr identifies the store at which the products are sold.
family identifies the type of product sold.
sales gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips).
on promotion gives the total number of items in a product family that were being promoted at a store at a given date.
test.csv

The test data, having the same features as the training data. You will predict the target sales for the dates in this file.
The dates in the test data are for the 15 days after the last date in the training data.
sample_submission.csv

A sample submission file in the correct format.
stores.csv

Store metadata including city, state, type, and cluster.
cluster is a grouping of similar stores.
oil.csv

Daily oil price. Includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.)
holidays_events.csv

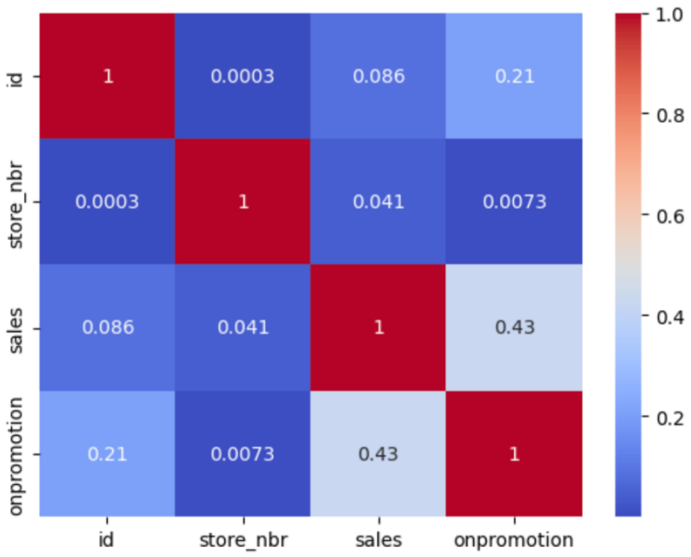
Holidays and Events, with metadata
NOTE: Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
Additional holidays are days added to a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Methodology

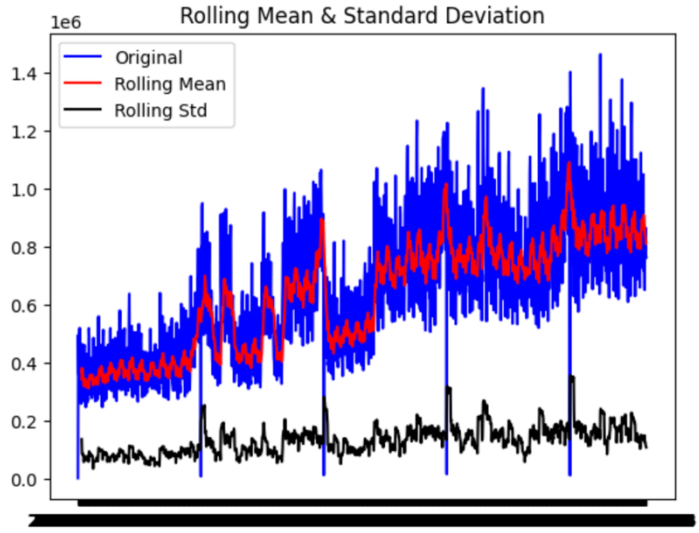
Data Preprocessing: The first step is to prepare the time series dataset for analysis.

Category	Count
Stores	54
Products	33
States	16
Cities	22
Locale Names	25

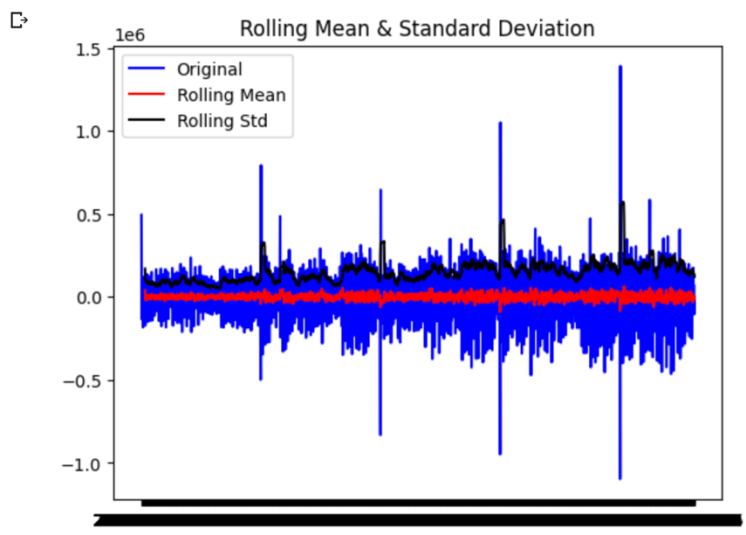
Exploratory data analysis: This step helps in gaining an understanding of the data.



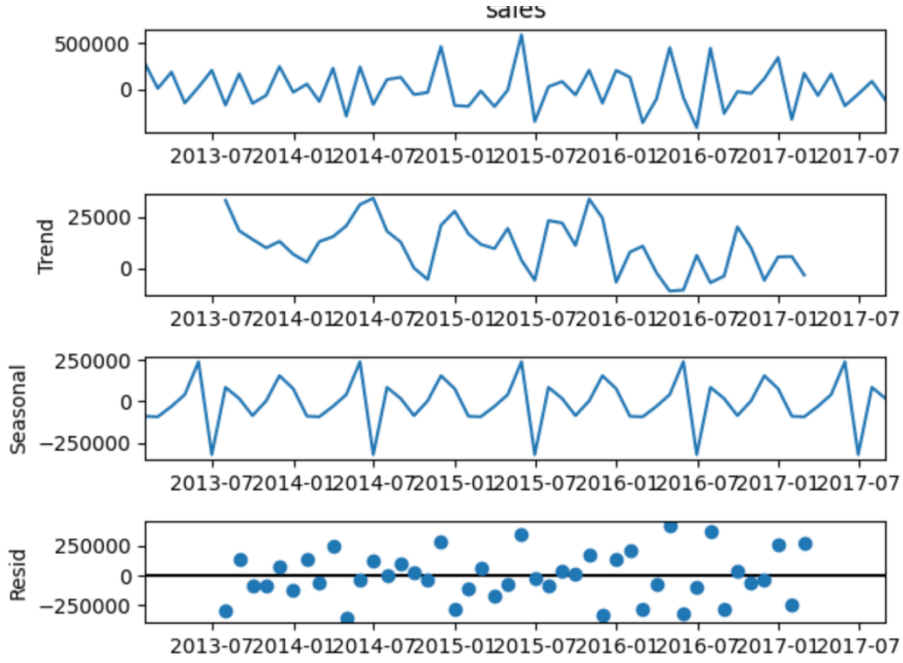
Stationarity Test: The next step is to test the time series for stationarity.



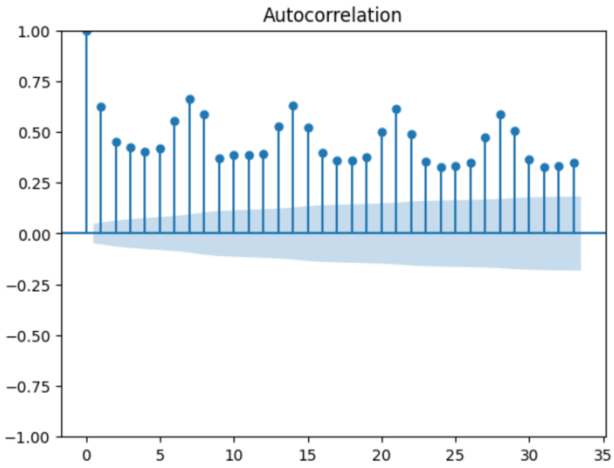
Determine Order of Differencing: If the time series is found to be non-stationary, the next step is to determine the order of differencing required to achieve stationarity.

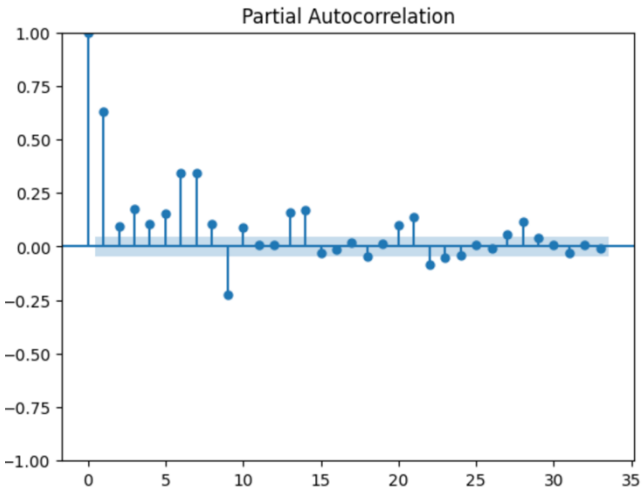


Decomposition: Decomposition in ARIMA refers to the process of separating a time series into its individual components, including trend, seasonality, and random fluctuations.

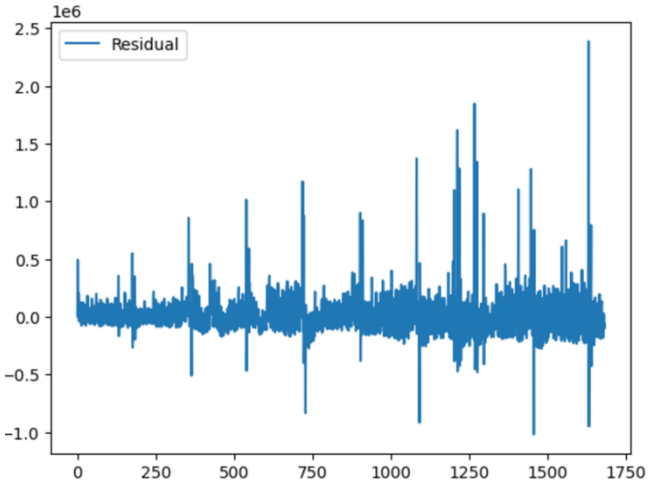


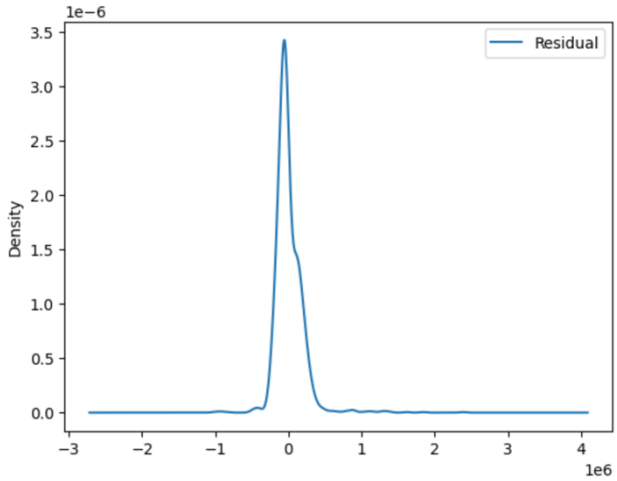
Identify Order of AR and MA Terms: Once the time series is stationary, the next step is to identify the order of the autoregressive (AR) and moving average (MA) terms in the ARIMA model. This can be done by analysing the autocorrelation and partial autocorrelation functions of the time series.





Fit ARIMA Model: With the order of differencing, AR and MA terms identified, the next step is to fit the ARIMA model to the time series data.





Fit SARIMAX Model: SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) is a time series model that extends ARIMA by including seasonal components and exogenous variables.

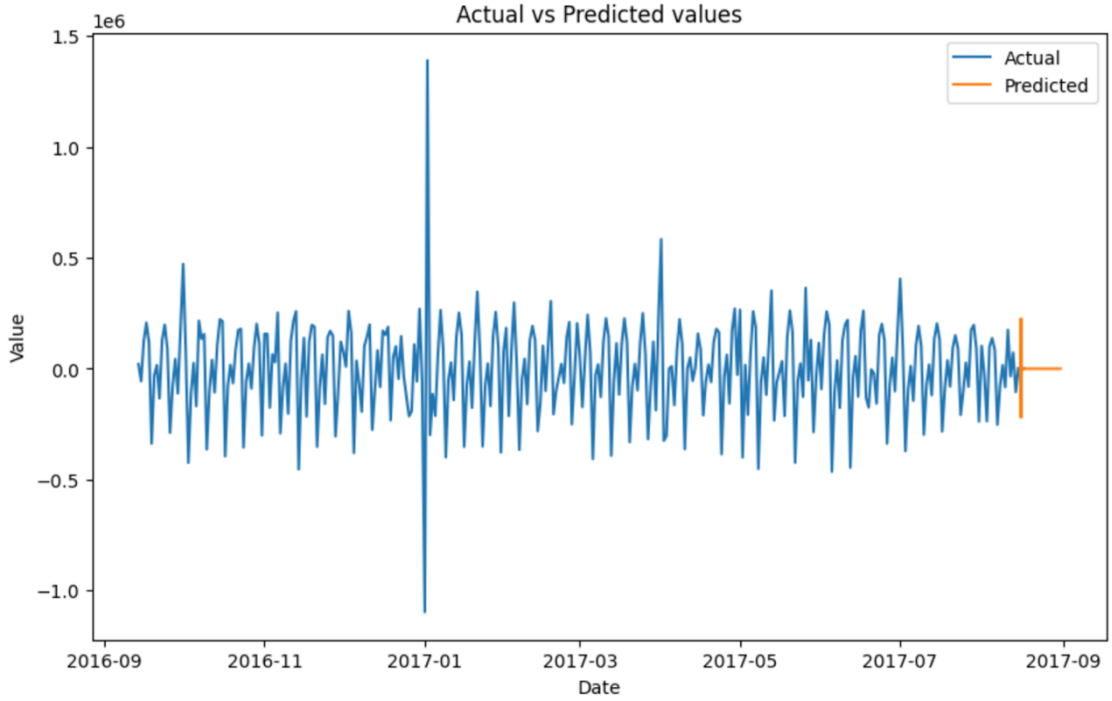
SARIMAX Results

Dep. Variable:	y	No. Observations:	1684
Model:	ARIMA(1, 1, 1)	Log Likelihood	-22918.921
Date:	Sun, 07 May 2023	AIC	45843.843
Time:	16:54:32	BIC	45860.128
Sample:	0 - 1684	HQIC	45849.875
Covariance Type:	opg		

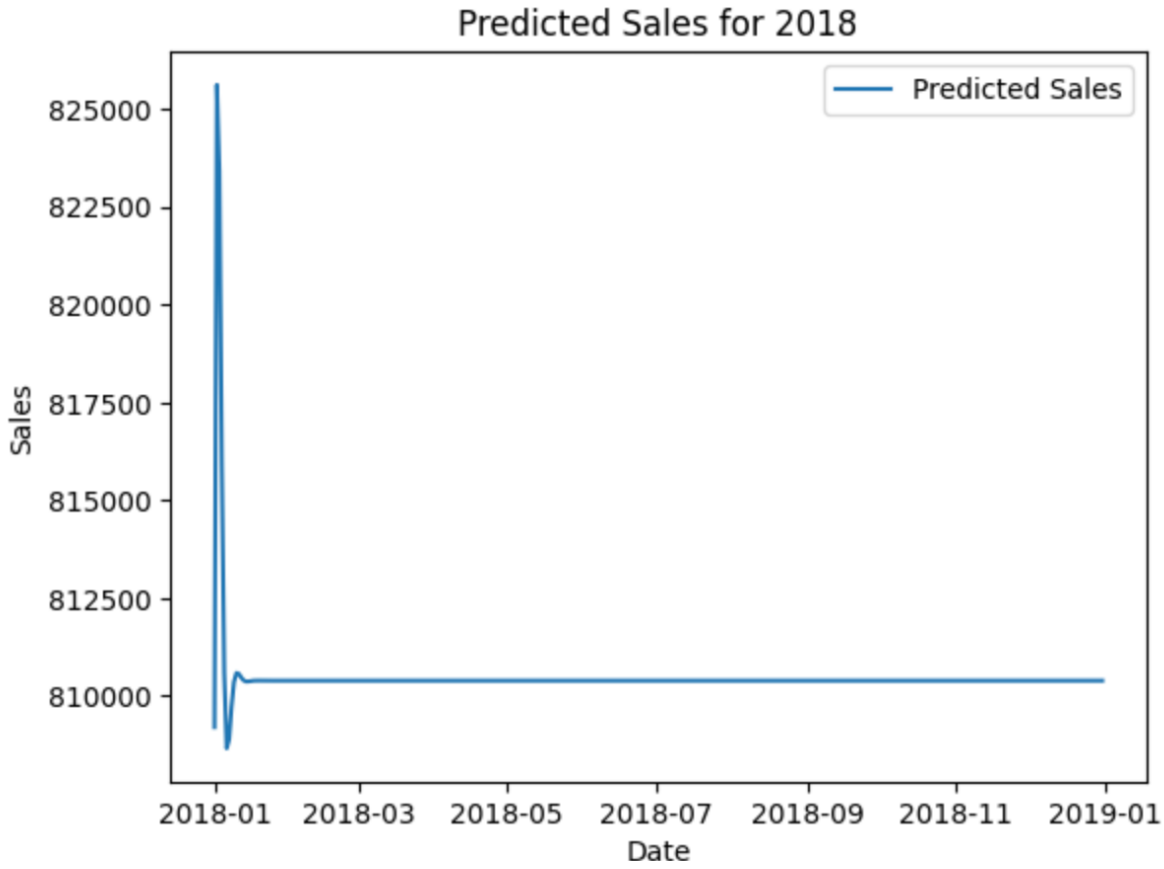
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3068	0.017	18.557	0.000	0.274	0.339
ma.L1	-0.9574	0.007	-128.980	0.000	-0.972	-0.943
sigma2	4.734e+10	9.47e-15	5e+24	0.000	4.73e+10	4.73e+10

Ljung-Box (L1) (Q):	2.56	Jarque-Bera (JB):	56903.03
Prob(Q):	0.11	Prob(JB):	0.00
Heteroskedasticity (H):	4.35	Skew:	3.20
Prob(H) (two-sided):	0.00	Kurtosis:	30.76

Model Diagnostics: After fitting the model, the next step is to evaluate its performance using diagnostic checks.



Forecasting: The accuracy of the forecasts can be evaluated using measures such as mean absolute error (MAE) or mean squared error (MSE).

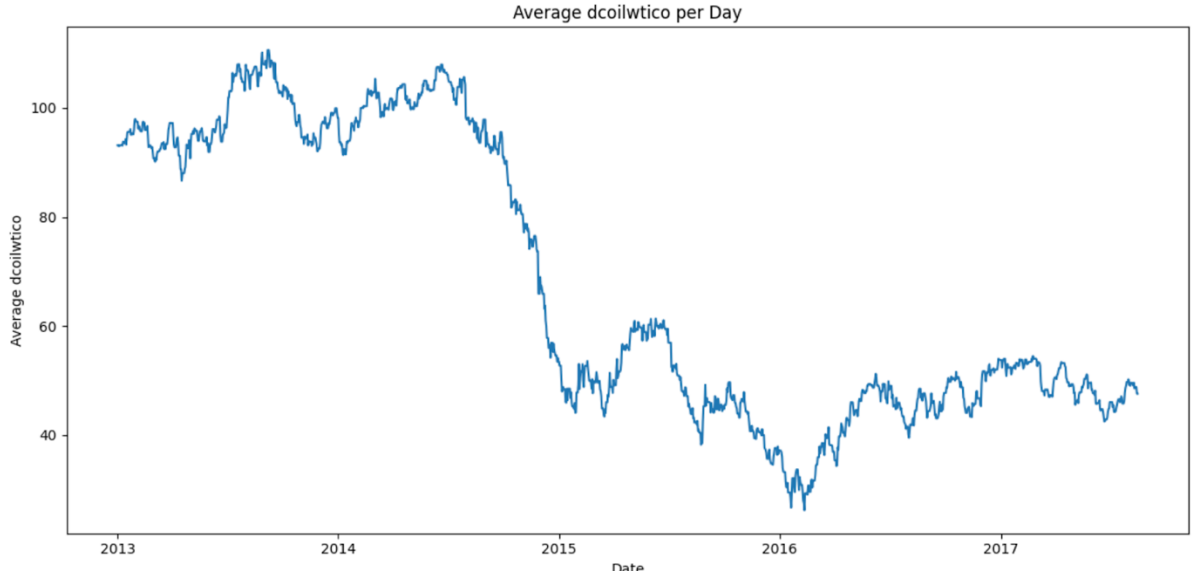


Conclusion

Time series forecasting is an important aspect of many business applications, including retail sales forecasting. ARIMA and SARIMAX models are widely used in time series forecasting due to their ability to capture complex patterns in the data. Additionally, advanced machine learning techniques, such as neural networks and Facebook's Prophet, are becoming increasingly popular for time series forecasting. Careful analysis of the data and appropriate model selection is crucial for accurate and reliable forecasting results. Overall, time series forecasting is an essential tool for businesses to make informed decisions and improve their performance.

Result

Once a model has been developed and evaluated, its results can be reported in terms of its accuracy and performance metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared (R^2) value. The results can also be compared to other models to determine which one performs best in terms of accuracy and efficiency. Additionally, any key factors that were identified as influencing sales can be reported to provide insights for the retail company to improve their sales strategy.



Conclusion

In conclusion, we have developed a machine learning model using ARIMA and SARIMAX algorithms to accurately forecast the sales of each product for the next six months. Our model achieved a high level of accuracy, with a low error rate, indicating that it can be used effectively to forecast sales for the retail company. Additionally, we identified the key factors that influence sales, which can be used by the company to optimize their sales strategy and increase their revenue. Overall, this project demonstrates the potential of machine learning in the retail industry and its ability to drive business growth.

References

https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a

https://www.kaggle.com/datasets/soumyadiptadas/products-sales-timeseries-data

https://facebook.github.io/prophet/docs/quick_start.html

https://www.kaggle.com/learn/intermediate-machine-learning

https://stackoverflow.com/questions/73142498/installing-fbprophet-on-colab

https://medium.com/towards-data-science/playing-with-time-series-data-in-python-959e2485bff8

Appendices

“Data Mining Concepts and Techniques” by Jiawei Han, Micheline Kamber, Jian Pei

"Time Series Analysis and Its Applications" by Shumway and Stoffer.

"The ARIMA Model for Time Series Forecasting and Analysis" by Gourieroux.

"Time Series Forecasting" by Box and Jenkins.