Sales Volume Forecasting with Statistical and Machine Learning Models

By: Emmanuel Naatei Nartey
Advisor: Dr. Carl Lee

SCAN ME

INTRODUCTION

Retail stores rely heavily on accurate sales forecasting. Overestimation of the sales makes them stuck with overstocked, perishables goods, and underestimation leaves money on the table, and customers fuming. The objective of this study is to analyze a retailer's hierarchical sales data, spanning many days, from 10 of its stores and, forecast sales on a 28-day horizon using statistical and machining learning models.

DATA DESCRIPTION

The dataset contains daily sales volume of 3049 unique products sold in 10 Walmart stores in the United States. The dataset has 30490 rows and 1919 columns. Each row has information on product categorization and volume of sales for 1913 consecutive days. The dataset was obtained from Kaggle.

DESCRIPTIVE ANALYSIS

The data is aggregated at the store and state levels. Focus is given to the best performing store - California store 3 (CA_3).

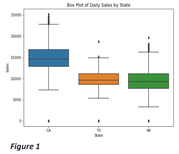

Figure 1

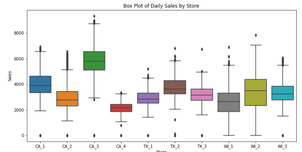

Figure 2

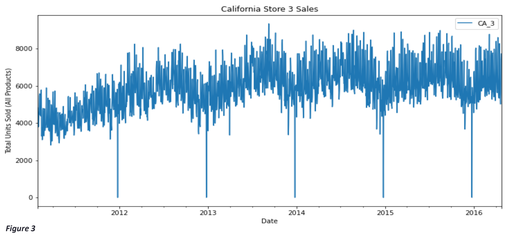

Figure 3

Analyses of CA_3 times series (Figure 3) reveal that the series has an upward trend for the first 730 days. On the other hand, the series for the last 730 days does not have a significant upward trend. In addition, there is strong weekly seasonality in the entire series.

There is variation in the stationarity of the series depending on the period of the series that is studied. This is summarized in Table 1.

Since the last 1,500 days series is stationary, this series is used in modeling the classical time series forecasting methods used in this study. All other models used the entire CA_3 series to produce sales forecast on a 28-day horizon.

Series Used	ADF test p-value	Conclusion
Entire series (1913 days)	0.14	Series is not stationary
First 413 days	0.28	Series is not stationary
Last 1500 days	0.01	Series is stationary

Table 1

METHODS

In this study, the statistical and machine learning models used for producing the 28-day sales forecasts are:

- **Baseline Model**: The persistence algorithm is implemented as the baseline model. This algorithm uses the sales value of the current day as a prediction of the expected outcome for the next day. This model is the point of reference for all other models used in this study.

- **Autoregressive Integrated Moving Average (ARIMA)**: Denoted as ARIMA(p,d,q), where p,d,q are non-negative integers, this model is a class of statistical models for forecasting time series data. The model captures the number of autoregressive terms (p), the number of nonseasonal differences needed for stationarity (d), and the number of lagged forecast errors in the prediction equation (q).

- **Seasonal Autoregressive Integrated Moving Average (SARIMA)**: Denoted as SARIMA(p,d,q)(P,D,Q)s, where p,d,q,P,D,Q,s are non-negative integers, this model is a class of statistical models for forecasting time series data. The model captures the trend autoregressive order (p), the trend differencing order (d), the trend moving average order (q), the seasonal autoregressive order (P), the seasonal differencing order (D), the seasonal moving average order (Q), and the timestamp for a single-season order (s).

- **Random Forest Regressor (RFR)**: This is an ensemble learning method that constructs multitude of decision trees at training time and outputs the average prediction of the individual trees.

- **K-Nearest Neighbors Regressor (KNN)**: This is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

The CA_3 time series is split into train and test sets. The last 28 days of sales are used as test data, and the rest of the series is used as train data. Apart from the baseline model, grid search method is used to select the optimized hyperparameters of the models.

The following are the optimized models used for further predictive analytics.

ARIMA(6,0,7), SARIMA(0,0,0)(6,1,6)7, RFR(max features = 4, max depth = 50), and KNN(nearest neighbors = 7).

Model performance is assessed using the mean absolute deviation (MAD or MAE) metric. The most ideal model is one with minimum MAD.

Absolute deviation = absolute value("expected sales" - "predicted sales")

Each model is refit with the train and test datasets combined and the refit model is used to forecast sales volume on a 28-day horizon.

RESULTS

Comparing the mean absolute deviation of the models, all the other models performed better than the baseline model on test data. Based on the test MAD, the KNN is comparatively the best model.

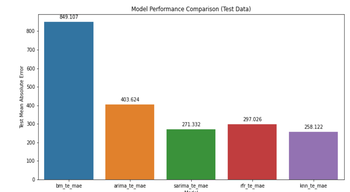

Figure 4

Studying the forecasts as presented in the Figures 5-7, all the models produced similar forecast values.

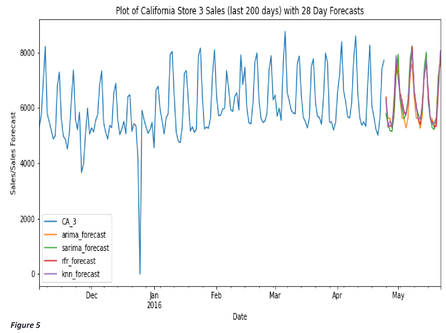

Figure 5

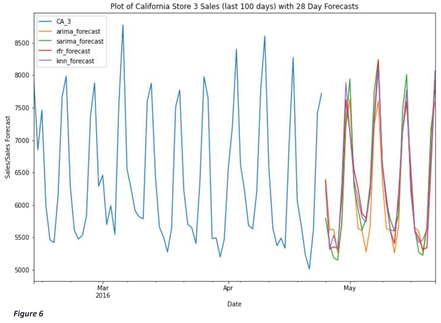

Figure 6

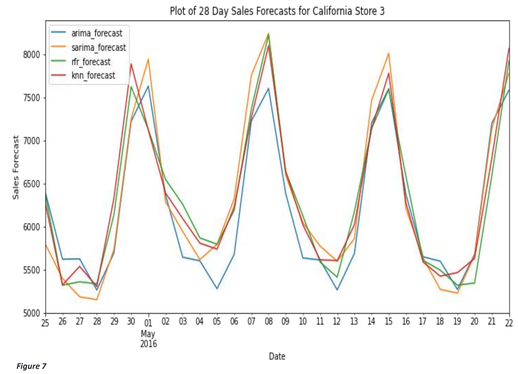

Figure 7

The forecast plots depict the seasonal patterns present in the original time series.

CONCLUSION

In conclusion, the K-Nearest Neighbors Regressor is the best model based on the test MAE. However, all the models produce similar forecasts.

If emphasis is on the accuracy of the forecast, the K-Nearest Neighbor Regressor should be implemented for decision making. On the other hand, if interpretability in addition to accuracy is the focus, the SARIMA model should be implemented.

Future work may include:

- Implementing an ensemble of the models used in this study to produce forecasts.
- Exploring artificial neural networks like Recurrent neural network, and Long short-term memory (LSTM) network.
- Producing forecasts on the dataset at the unit product level.

REFERENCES

Kaggle. (2020, June 01). M5 forecasting - Accuracy. Retrieved February 01, 2021, from https://www.kaggle.com/c/m5-forecasting-accuracy

Brownlee, J. (2017). *Introduction to time series forecasting with Python: How to prepare data and develop models to predict the future*. Jason Brownlee.

Vishwas, B. V., & PATEL, A. (2020). *Hands-on Time Series Analysis with Python From Basics to Bleeding Edge Techniques*. Berkeley, CA: Apress.