Forecasting Walmart Store Sales

Abhirag Nagpure

Indiana University

Abstract. Accurate forecasting is a highly fallible albeit important component in multiple aspects for an organization. Publicly traded organizations can set reasonable projections for quarters with accurate product demand and sales forecasts, and can also hire or fire staff accordingly. This project aims to forecast sales figures for multiple Walmart stores. Taking into account external factors as provided in the data, I start out originally with the hypotheses: extreme temperatures negatively impact the sales, holiday weeks see higher sales than usual, and unemployment is also inversely correlated with sales. After a series of experiments with both linear regressors and complex ARIMAX models, I fail to find concrete evidence to reject or fail to reject all the hypotheses.

1 Introduction

Forecasting essentially is a technique to estimate the value of a variable in the future - using its past historical values. A common example is predicting sales from past sales data - analyzing seasonal characteristics or any trends associated with the data. Another common example is predicting stock prices. Accurate forecasting is a challenging and complex task - assuming that historical trends are stable and continuous and will continue to be so can lead to an erroneous forecasting model. Another challenge is the presence of exogenous variables that are not provided in the data-set - there may be many more variables affecting the target variable than those provided. As aptly said by Box [2], "all models are wrong, but some are useful". Hence the task of forecasting involves finding a model that is most useful, to say the least.

2 Data

2.1 Source

The data is obtained from a Kaggle competition: Walmart Recruiting - Store Sales Forecasting [6]. The data is available across 45 Walmart stores located in different regions, with each store containing multiple departments. To increase the problem complexity, there are some holiday markdown events that are included - which are known to affect the sales - but modeling that impact is a big task.

2.2 Description and Pre-processing

The data is spread across 3 CSV files:

1. train.csv - This file contains the actual sales figures for each department of each store across 143 time-steps: weekly observations recorded from February 5, 2010 to November 1, 2012. The features and their descriptions are included in Table 1.

Table 1. Features and their description included in train.csv

Feature	Description
Store	the store number
Dept	the department number
Date	the date on which the observation is recorded
Weekly_Sales	the actual sales figures for the corresponding department and store
IsHoliday	Whether that the week is a special holiday week

2. stores.csv
   This file contains anonymised information about the stores - indicating their type and size. The features and their descriptions are included in Table 2.

Table 2. Features and their description included in stores.csv

Feature	Description
Store	the store number
Type	the type of the store (one of A, B or C)
Size	the size of the store

3. features.csv
   This file contains additional information about the stores and regional activities. The features and their descriptions are included in Table 5.

I employed the following pre-processing steps:

1. Merge all the different data files into a single data frame object using inner joins.
2. Subset out a part of the data with Store number as 1 and Dept number as 1.
3. Since more than half of the time-steps have missing values for the markdown variables (MarkDown1 - MarkDown5), remove these from the data.
4. Keep a hold-out set for forecasting from time-step 130 on-wards.

Table 3. Features and their description included in features.csv

Feature	Description
Store	the store number
Date	the date on which the observation is recorded
Temperature	the average temperature in the region
Fuel_Price	the average cost of fuel in the region
MarkDown (1-5)	anonymised data related to promotional markdowns
CPI	the consumer price index
Unemployment	the unemployment rate
IsHoliday	Whether that the week is a special holiday week

3 Preliminary Analysis

3.1 Time-Series Plots

Plotting the criterion variable (weekly_sales) and the predictor variables (temperature, unemployment, CPI and fuel_price) is essential to draw some preliminary ideas about the underlying data distribution.

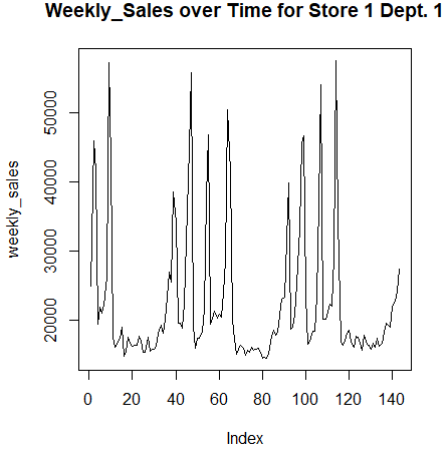

Fig. 1. Plot of weekly sales figures over time

The spikes as depicted in Fig. 1 seemed at first to indicate seasonality, but upon conducting the Dickey-Fuller test [4], a p-value of 0.01 was obtained - meaning the the time-series is indeed stationary. As an alternative, a KPSS test

[7] was also conducted, and the stationarity was confirmed.

As a further step, I attempted to decompose the time series into seasonal and trend components using both the decompose [9] and the stl [3] functions in the stats package in R - both yielding the same error - stating that "the series is not periodic or has less than two periods".

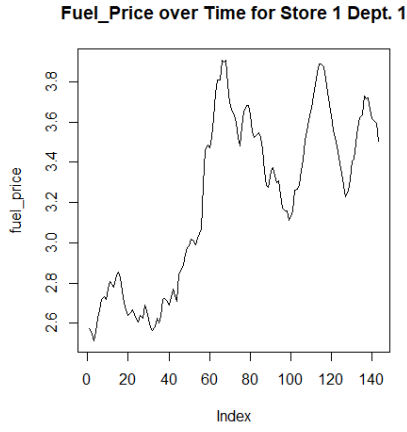

Fig. 2. Plot of fuel_price over time

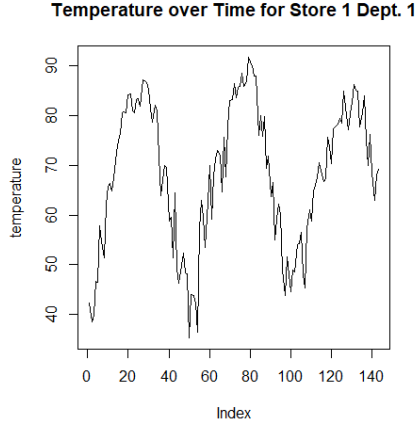

Fig. 3. Plot of temperature over time

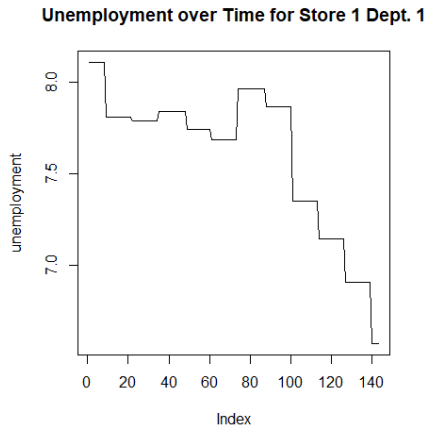

Fig. 4. Plot of unemployment over time

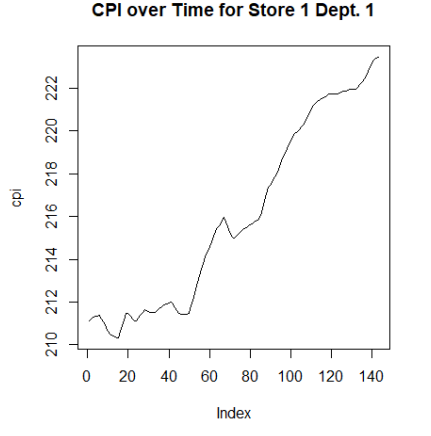

Fig. 5. Plot of CPI (consumer price index) over time

The same tests as mentioned for the criterion variable were carried out for the temperature predictor variable also - with the sames results, despite some reoccurring spiked curves as depicted in Fig. 3.

3.2 Auto Covariance plots

The ACF and PACF plots are helpful in determining the order of the MA and AR processes to feed those into the more complex ARIMA or ARIMAX models.

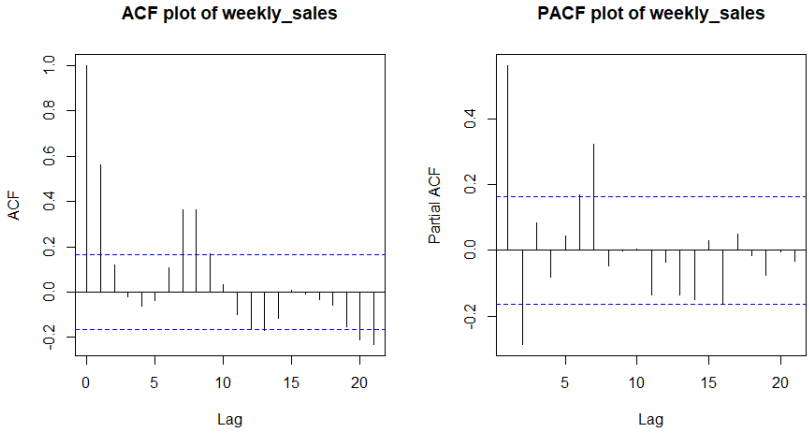

Fig. 6. ACF and PACF Plots of weekly_sales

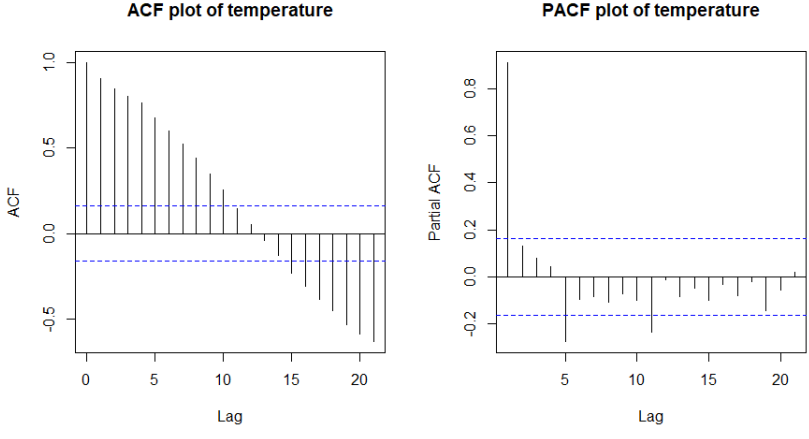

Fig. 7. ACF and PACF Plots of temperature

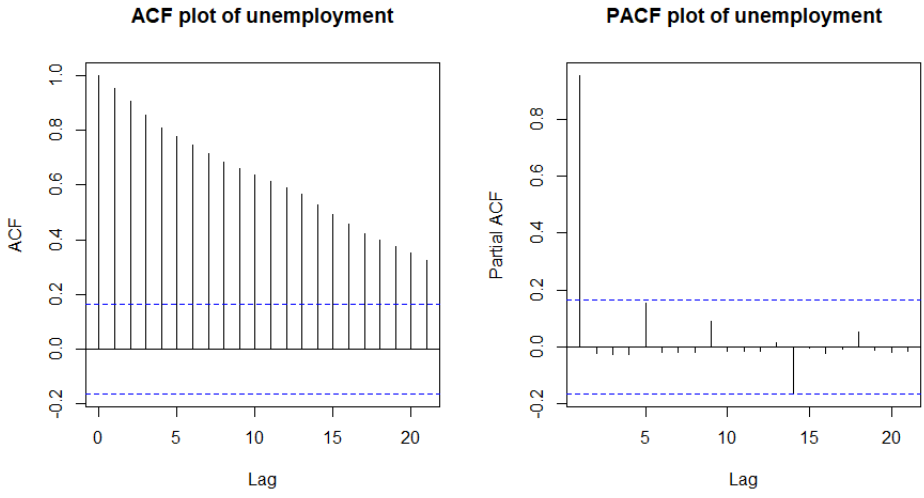

Fig. 8. ACF and PACF Plots of unemployment

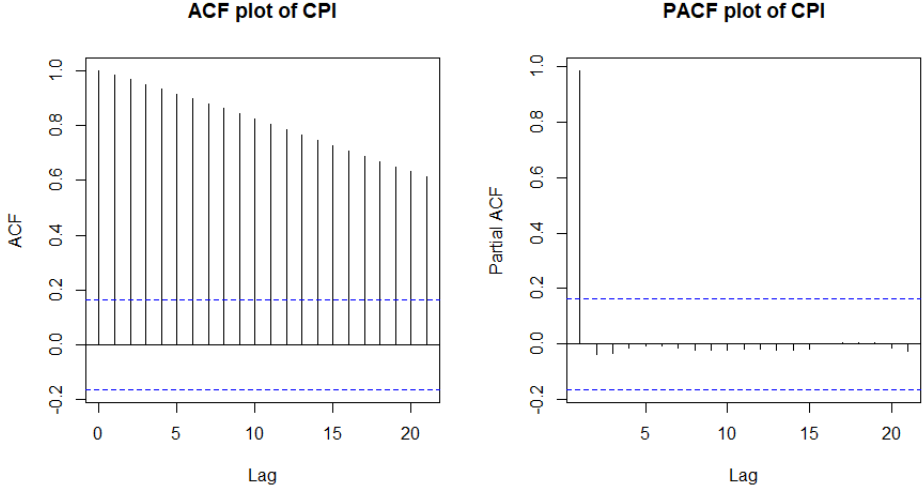

Fig. 9. ACF and PACF Plots of CPI

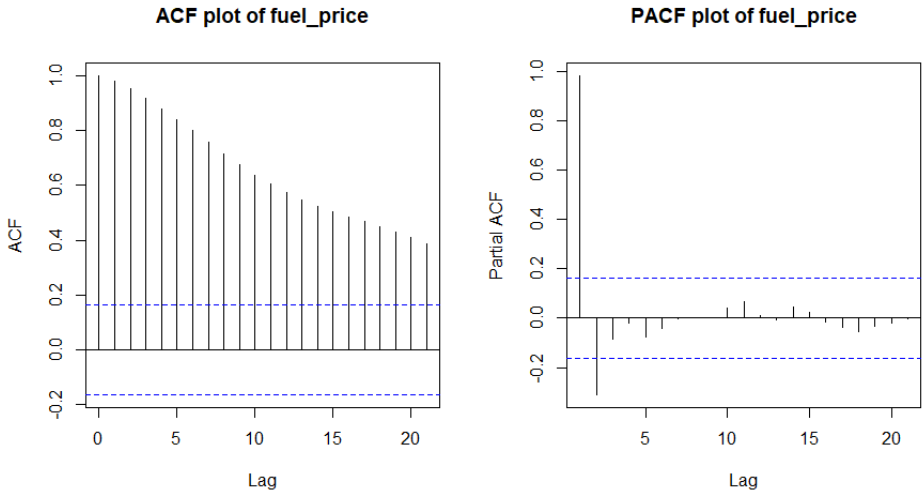

Fig. 10. ACF and PACF Plots of fuel_price

Fig. 6 indicates that the weekly_sales series has significant correlation at lags 1, 7 and 8.

Fig. 7, Fig. 8 and Fig. 9 indicate that the temperature, unemployment and CPI series have significant correlation at lag 1. Differencing these series by 1 seems to be apt. Fig. 10 indicates the fuel_price series has a significant correlation at lag 2.

3.3 Cross Covariance plots

Analysing cross-covariance plots is important since correlation does not imply causation and can imply spurious relationships among variables.

As evident by Fig. 12, Fig. 13, and Fig. 14, there does not seem to be any significant correlation among the corresponding variables - except for temperature as indicated by Fig. 11.

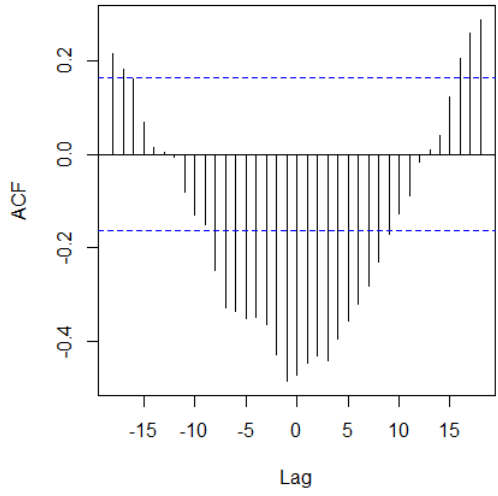

Fig. 11. CCF Plot of temperature and weekly_sales

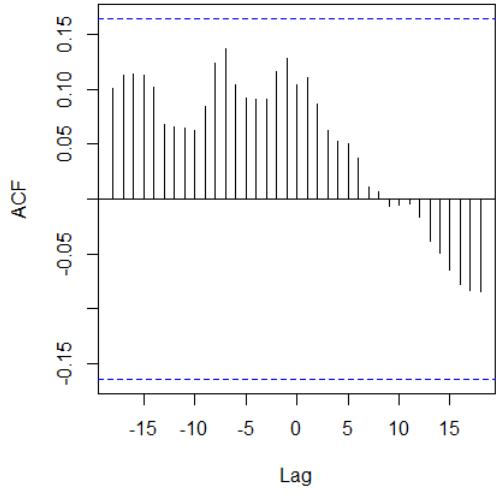

Fig. 12. CCF Plot of unemployment and weekly_sales

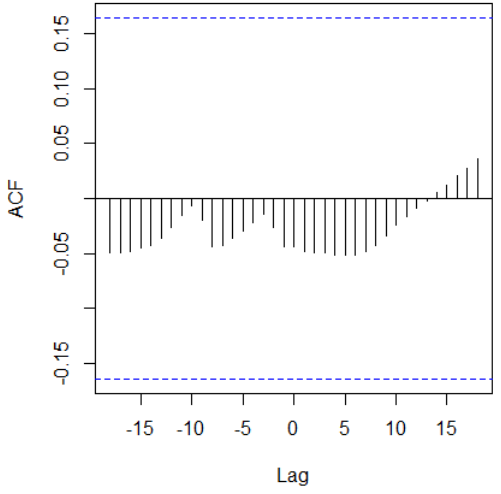

Fig. 13. CCF Plot of CPI and weekly_sales

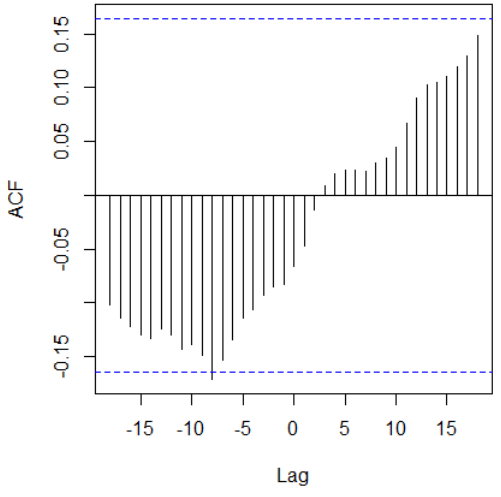

Fig. 14. CCF Plot of fuel_price and weekly_sales

4 Initial Hypotheses

I started out with three initial hypotheses, completely based on intuition and some basic assumptions:

1. Extreme temperatures negatively impact the sales
2. Holiday weeks see higher sales than usual
3. Unemployment is also inversely correlated with sales

5 Modeling

5.1 Using Linear Regressors

I attempted to fit multiple linear models to the data and evaluated them using the Akaike Information Criterion [1] and the Bayesian Information Criterion [8].

Table 4. Analysis of multiple linear regressors on the data. Note that if a variable is lagged by a factor of n, it has been represented by a adding a suffix _n

Model No.	Features	AIC	BIC
1	weekly_sales_1	2707.125	2715.704
2	weekly_sales_7	2605.533	2613.969
3	weekly_sales_8	2585.155	2593.567
4	temperature, cpi, unemployment, fuel_price	2748.738	2765.943
5	weekly_sales_1, temperature, cpi, unemployment, fuel_price	2702.872	2722.891
6	weekly_sales_1, weekly_sales_7, weekly_sales_8, temperature, cpi, unemployment, fuel_price	2541.815	2567.051
7	weekly_sales_1, weekly_sales_7, weekly_sales_8, temperature, cpi, cpi_1, unemployment, unemployment_1, fuel_price, fuel_price_1, fuel_price_2	2537.449	2573.901
8	weekly_sales_1, weekly_sales_7, cpi_1, unemployment_1, fuel_price_1, fuel_price_2	2530.348	2552.78

5.2 Using ARIMAX model

Since there are external regressors in the data, I attempted to fit multiple ARIMAX models with different configurations. Also, I utilised the results returned from the *auto.arima* function from the forecast package [5] in R.

Table 5. Analysis of multiple ARIMAX models on the data

Model Parameters	Features	AIC	BIC
ARIMAX(1,0,1)	weekly_sales_1, weekly_sales_7, weekly_sales_8, temperature, cpi, cpi_1, unemployment, unemployment_1, fuel_price, fuel_price_1, fuel_price_2	2529.435	2571.496
ARIMAX(0,0,1)	weekly_sales_1, weekly_sales_7, weekly_sales_8, temperature, cpi, cpi_1, unemployment, unemployment_1, fuel_price, fuel_price_1, fuel_price_2	2527.835	2567.092
ARIMAX(1,0,1)	weekly_sales_1, weekly_sales_7, cpi_1, unemployment_1, fuel_price_1, fuel_price_2	2529.396	2557.436
ARIMAX(0,0,2)	weekly_sales_1, weekly_sales_7, cpi_1, unemployment_1, fuel_price_1, fuel_price_2	2526.278	2554.319

6 Forecasting and Residuals Analysis

Using the best linear model (Model 8) and the best ARIMAX model (ARIMAX(0,0,2)), I analysed their residuals and predictions on the held-out forecasting data.

Residuals of a good-fit model should resemble white noise - which can be determined by the ACF plot of the residuals.

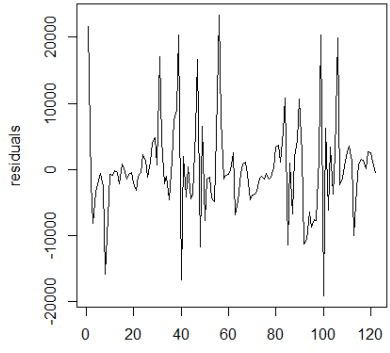

Fig. 15. Residuals of the best ARIMAX model. ACF plot of these residuals does not show any significant correlation

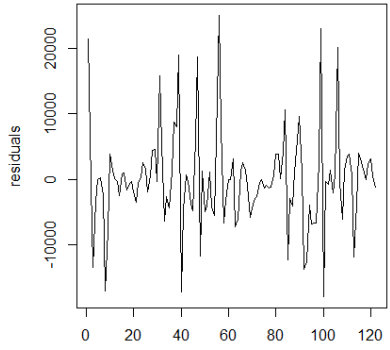

Fig. 16. Residuals of the best linear model. ACF plot of these residuals shows slightly significant correlation at lag 2

7 Conclusions

The actual values (mean-differenced) and the values forecasted by the best models are plotted and shown in Fig. 17. It is evident that neither of these models can be used in practice - essentially rendering them unsuitable for practical applications. Some possible reasons behind the failure of the models are:

1. Limited data: With only 130 time-steps to work with, training data was scarce. This possibly has impacted decomposition - the trend seems to be evident as well as the seasonality (but the decomposition models failed to work).
2. Only one department from one randomly selected store was used in this project. Generalizing a forecast model on a larger scale would involve extensive studies.

With these results, there does not seem enough evidence to reject or fail to reject the hypotheses I had started out with. Some support is obtained from cross-correlation plots but nothing definite can be concluded with these results.

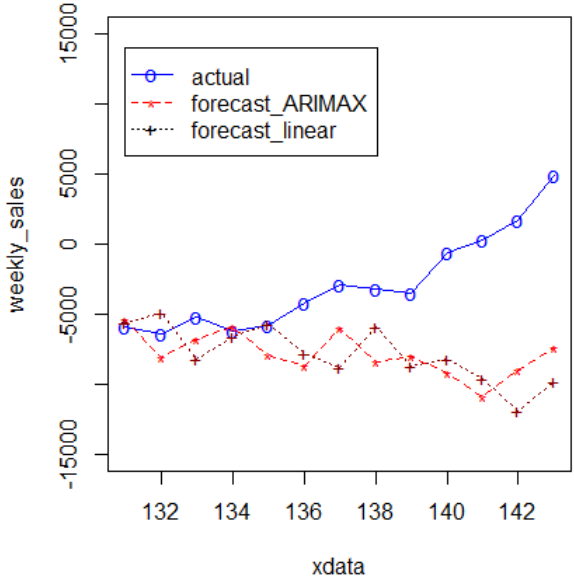

Fig. 17. Stacking up predictions against the actual weekly_sales values

References

1. Akaike, H., Petrov, B.N., Csaki, F.: Second international symposium on information theory (1973)
2. Box, G.E.P.: Science and statistics. Journal of the American Statistical Association **71**(356), 791–799 (1976). https://doi.org/10.1080/01621459.1976.10480949
3. Cleveland, R., Cleveland, W.S., McRae, J.E., Terpenning, I.J.: Stl: A seasonal-trend decomposition procedure based on loess (with discussion) (1990)
4. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association **74**(366a), 427–431 (1979). https://doi.org/10.1080/01621459.1979.10482531
5. Hyndman, R., Khandakar, Y.: Automatic time series forecasting: The forecast package for r. Journal of Statistical Software, Articles **27**(3), 1–22 (2008). https://doi.org/10.18637/jss.v027.i03, https://www.jstatsoft.org/v027/i03
6. Kaggle: Walmart recruiting - store sales forecasting (2014), https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting
7. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics **54**(1), 159 – 178 (1992). https://doi.org/https://doi.org/10.1016/0304-4076(92)90104-Y
8. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6**(2), 461–464 (03 1978). https://doi.org/10.1214/aos/1176344136, https://doi.org/10.1214/aos/1176344136
9. Wynn, H.P.: The advanced theory of statistics, vol. 3, 4th edition, kendall, sir maurice, stuart, a. and ord, j. k., high wycombe: Charles griffin, 1983. price: £37.50. pages: 780. Journal of Forecasting **4**(3), 315–315 (1985). https://doi.org/10.1002/for.3980040310