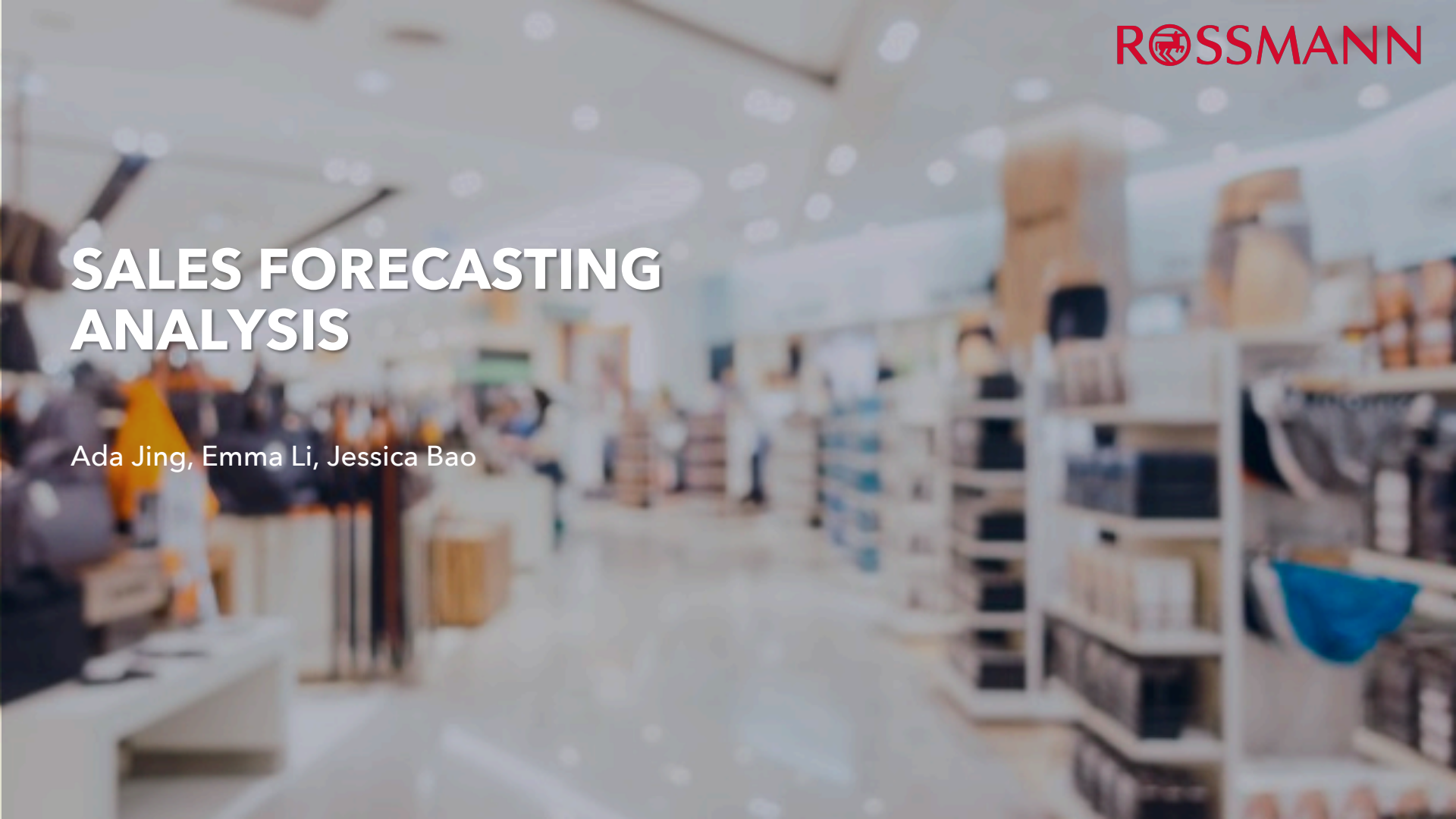


SALES FORECASTING ANALYSIS

Ada Jing, Emma Li, Jessica Bao

WHAT'S OUR GOAL?

SIMPLY PUT, OUR **OVERARCHING GOAL** IS TO **PREDICT ROSSMANN STORE FUTURE SALES DATA** BASED ON AVAILABLE INFORMATION **IN THE PRESENT TIME FRAME**

OBJECTIVES

EXPLORE

Understands factors impact sales, including data related to store activities (i.e. promotions) and timing (i.e. holidays), and decipher the ones with highest impact on model performance.

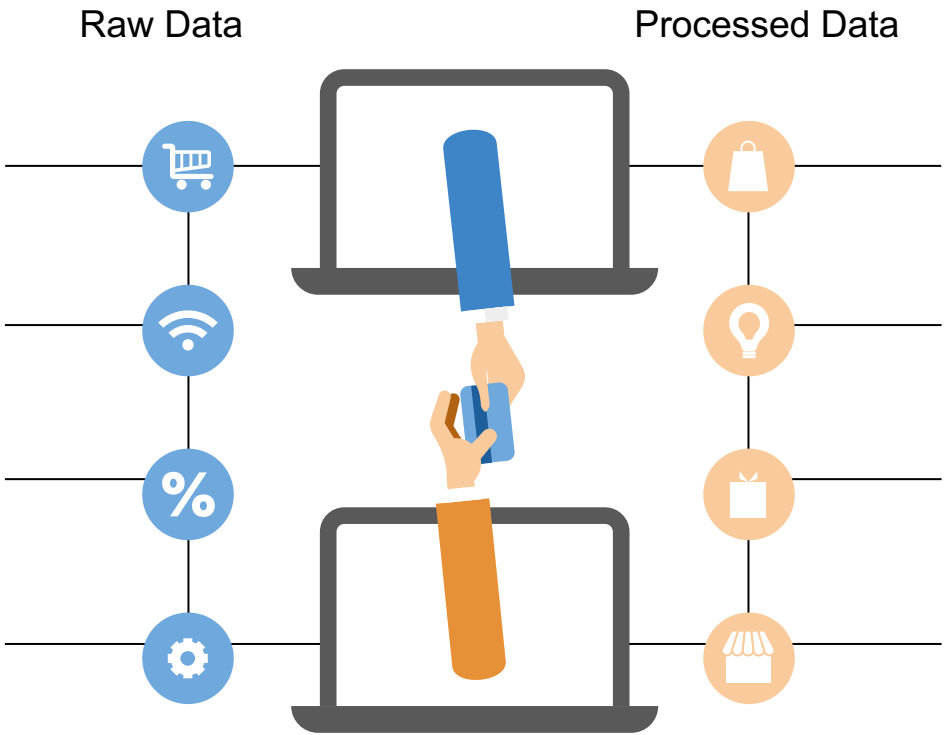
ANALYZE

Experiment with various analysis technique to identify the solution balances prediction accuracy with model complexity.

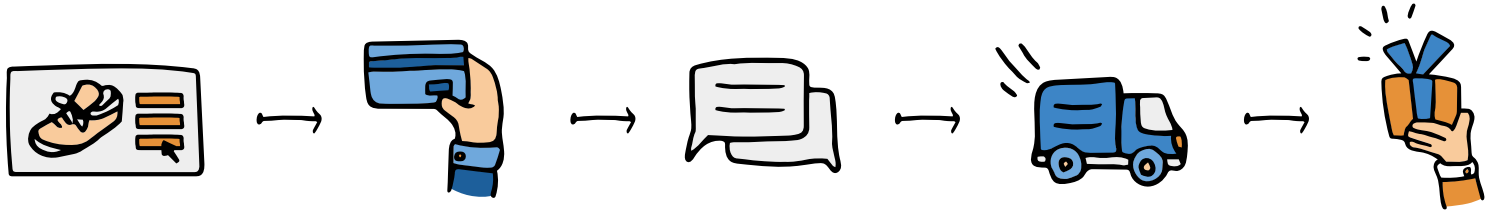
INFER

Deploy the winning model and explore future opportunities

DATA OVERVIEW



DATA EXPLORATION



1

Exploration

Examine data to determine if there are missing values

2

Decomposition

Decompose the data to detrend

3

ACF

Examine the overall ACF to determine if differencing is needed

4

Spectrum Analysis

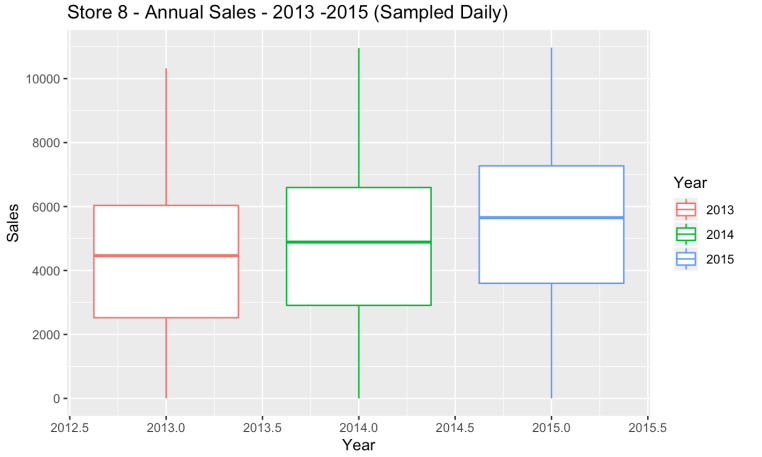
Run a spectral analysis in order to discover underlying periodicities

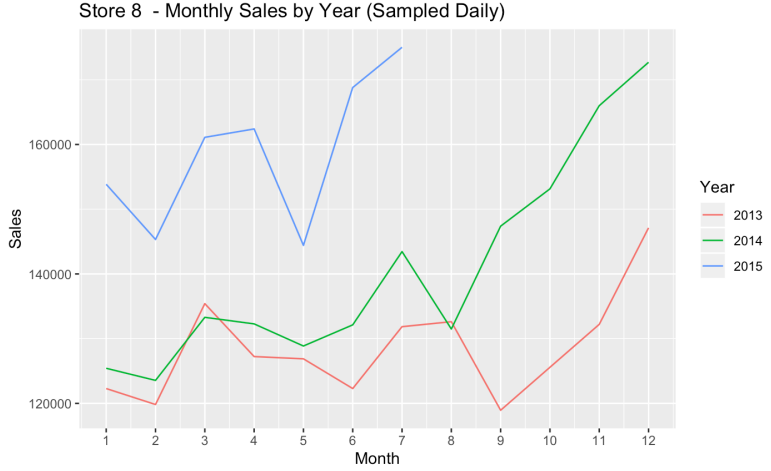
5

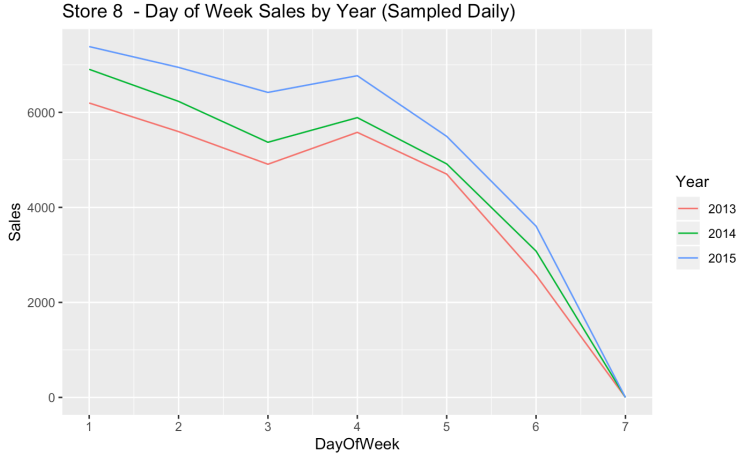
Train/Test Split

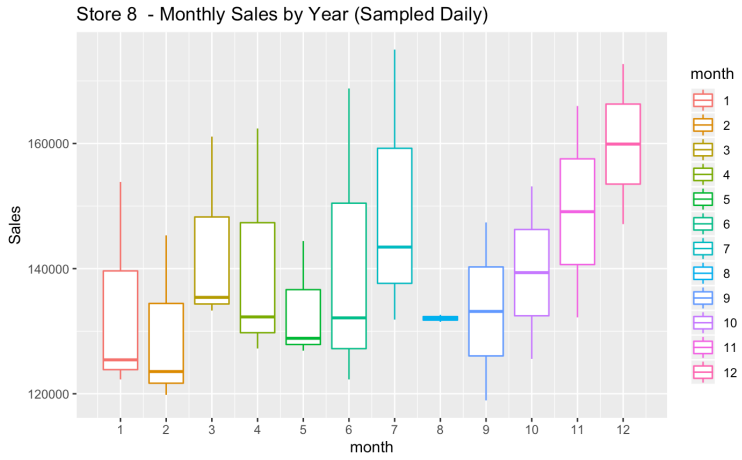
Randomly select one store and reserved 3 weeks Sales for testing

DATA EXPLORATION

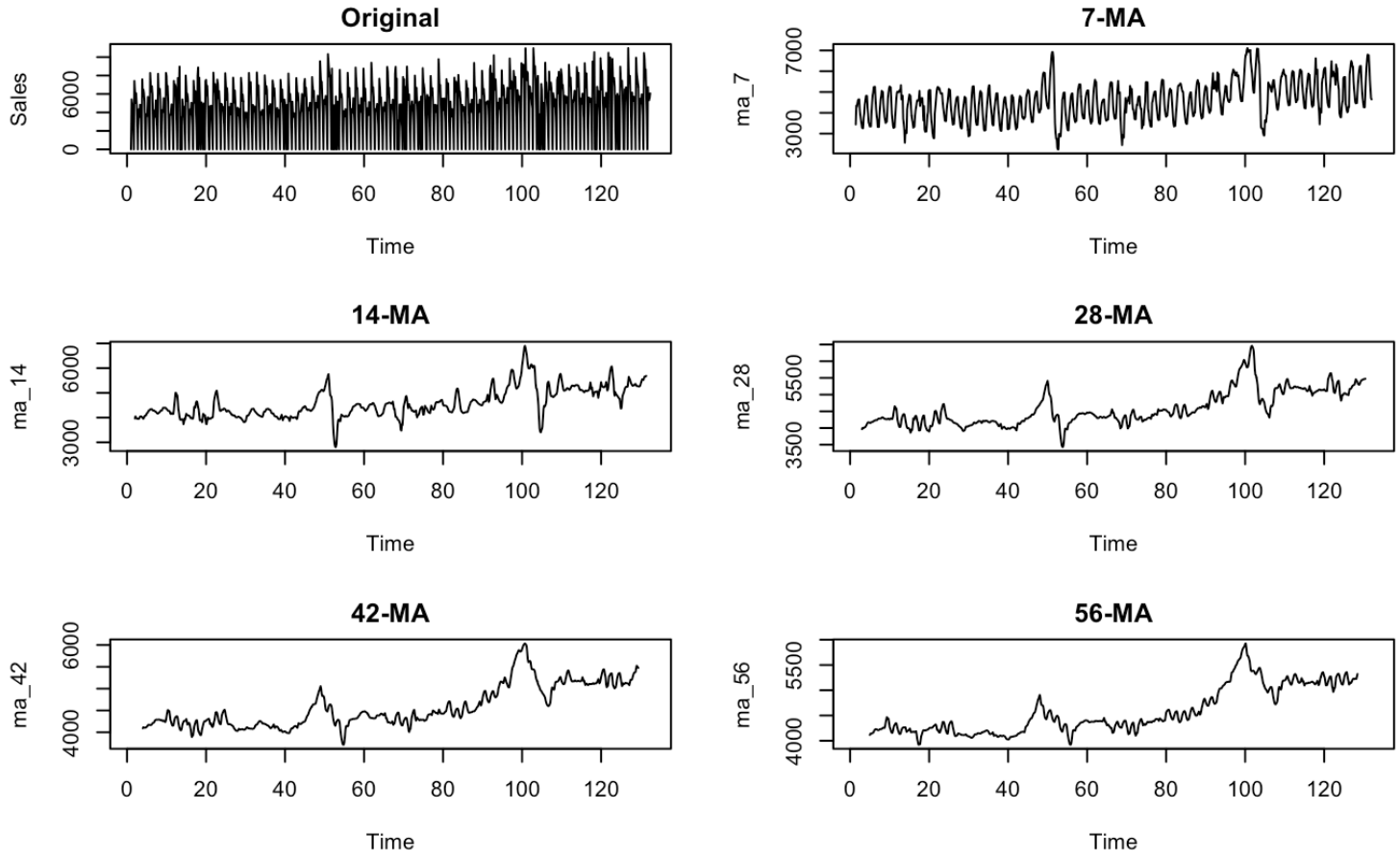




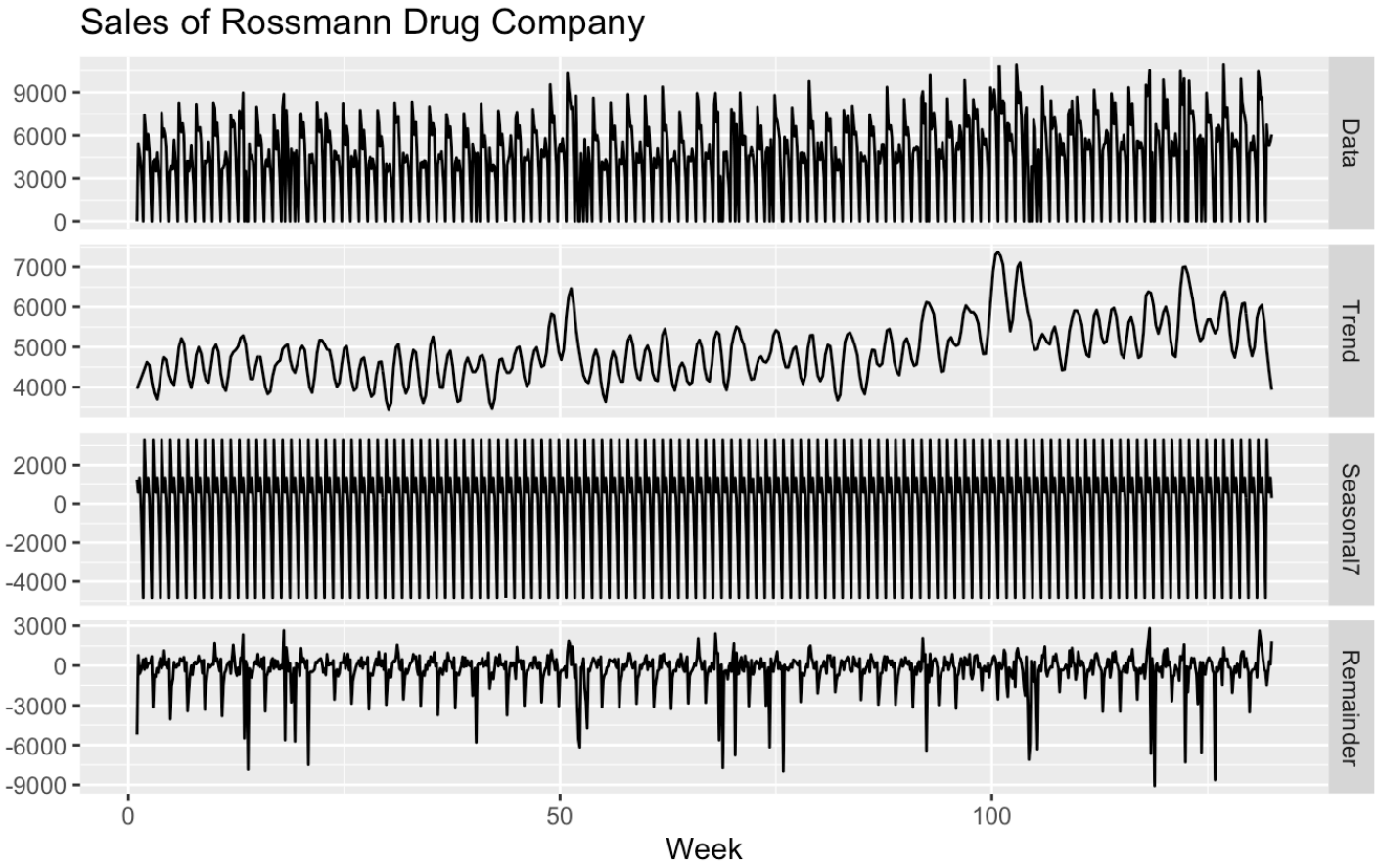




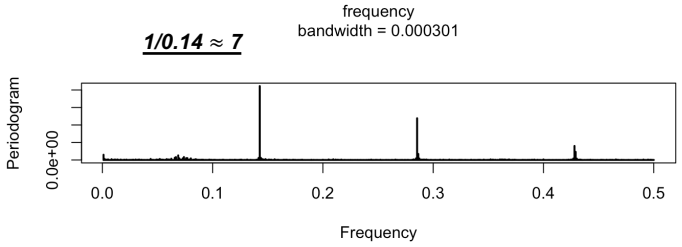
Moving Average

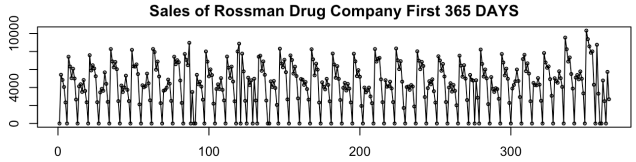


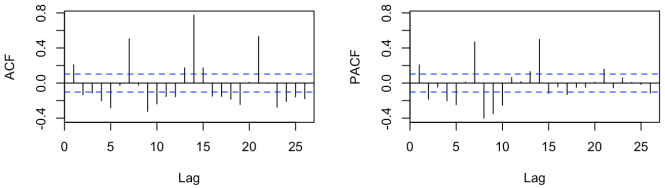
Decomposition



Stationary







Magnitude of variance

-

Box Cox Transformation

Trend

Seasonal Pattern

-

Differencing

High Auto Correlation across lags

-

Differencing

KPSS test:

P <0.05

Reject Null

-

Non Stationary

Box Cox Transformation

- Stabilize the Variance
- KPSS: p-value = 0.0214
- ADF: p-value < 0.05
  Reject!

First Order Differencing

- KPSS: p-value>0.05
- ADF: p-value <0.05
  Past!

Seasonal Differencing

- KPSS: p-value>0.05
- ADF: p-value <0.05
  Past!

Seasonal/Non Seasonal

- KPSS: p-value>0.05
- ADF: p-value <0.05
  Past!

DATA MODELING

01.

Baseline Models:

- Mean Forecast
- Naïve Forecast
- Seasonal Naïve Forecast
- Holt-Winters

02.

ARMA Models:

- Auto ARIMA
- Arima
- Seasonal ARIMA

03.

Linear Models:

- TSLM
- Linear regression with ARIMA error

04.

Advanced Models:

- VAR
- TBATS
- Neural Network

DATA MODELING

01.

Baseline Models:

- Mean Forecast
- Naïve Forecast
- Seasonal Naïve Forecast
- Holt-Winters

02.

ARMA Models:

- Auto ARIMA
- Seasonal ARIMA

03.

Linear Models:

- TSLM
- Linear regression with ARIMA error

04.

Advanced Models:

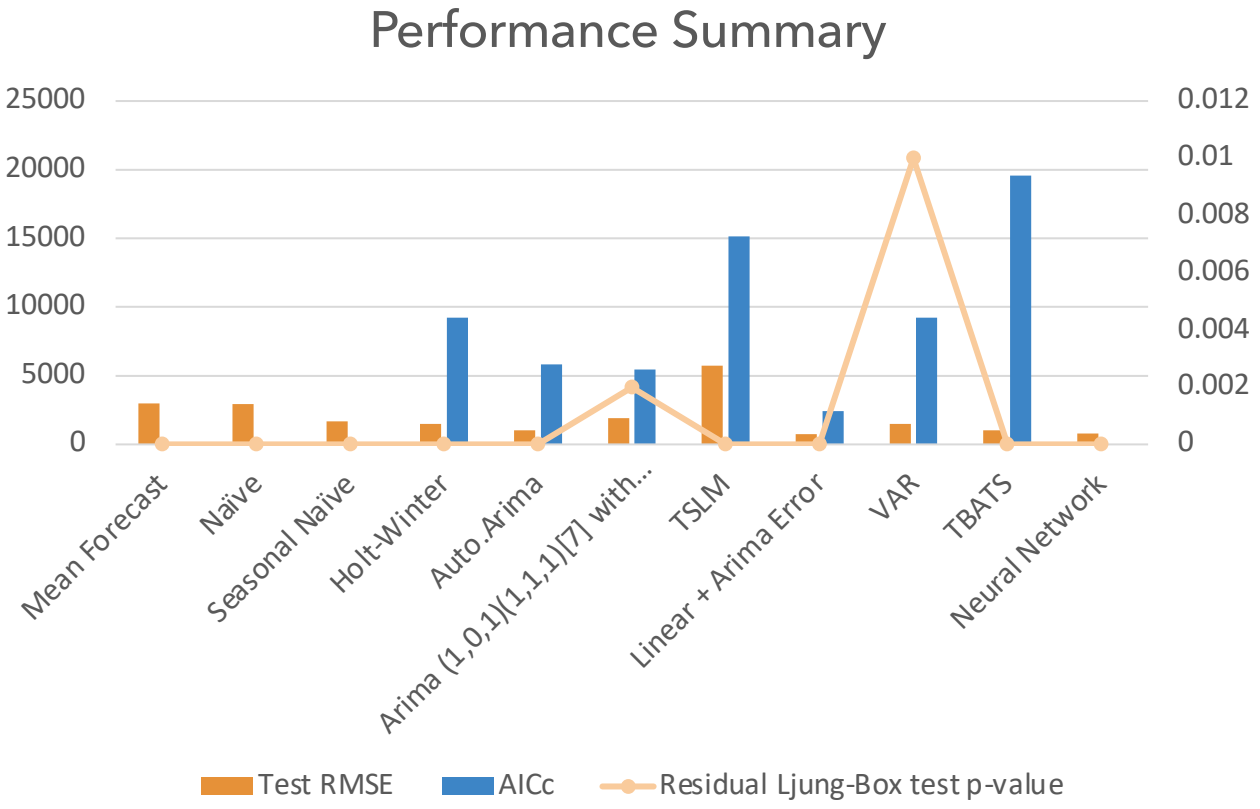
- VAR
- TBATS
- Neural Network

RMSE
How accurate is our model?

AIC
How efficient is our model?

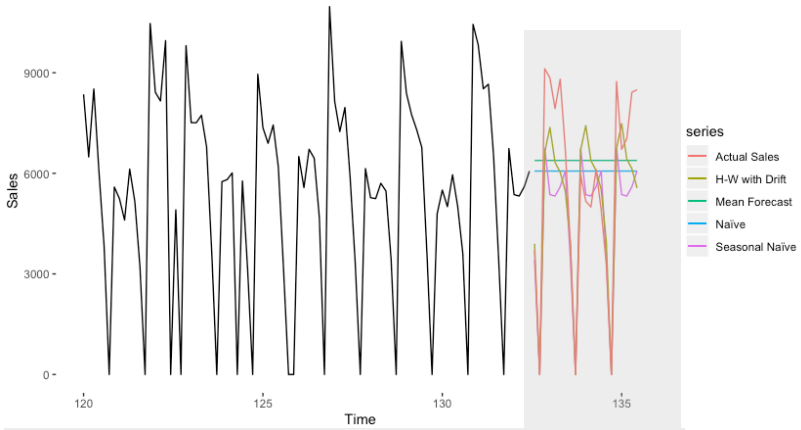
Ljung-Box Test
How comprehensive is our model?

DATA MODELING

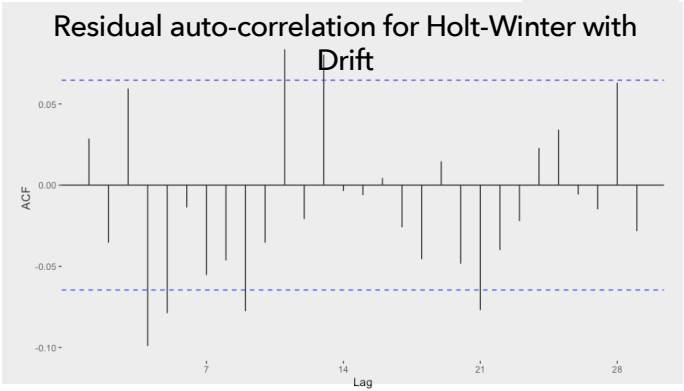


MODEL SUMMARY: BASELINE

Baseline forecast vs. actual



Residual auto-correlation for Holt-Winter with Drift



RMSE for Baseline Models

2992
Mean Forecast

2931
Naïve Forecast

1696
Seasonal Naïve Forecast

1500
Holt-Winter Forecast

Seasonal information increased model accuracy

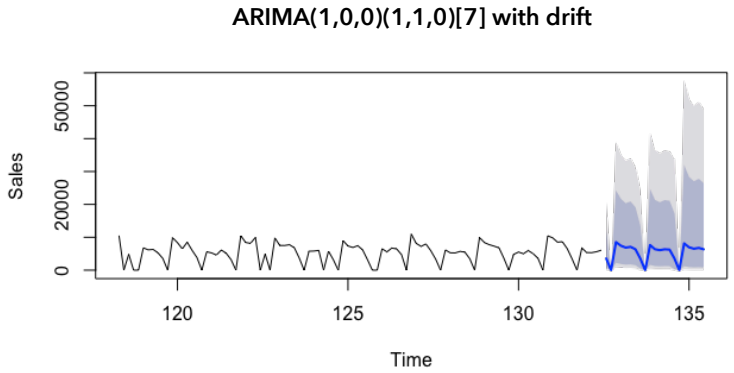
- Comparing the four baseline models, we see by incorporating seasonal information, our models showed significant improvement in performance

Opportunity to further explore remaining information in residuals

- However, all baseline models still showed residuals with strong autocorrelation, indicating opportunity to further utilize the available data.

MODEL SUMMARY: ARIMA

Arima model outperformed the baseline models in minimizing error



1022
RMSE

RMSE for Baseline Models

2992
Mean Forecast

2931
Naïve Forecast

1696
Seasonal Naïve Forecast

1500
Holt-Winter Forecast

MODEL SUMMARY: EACF/ARIMA

ARMA(1,4)
First Order Differencing

AR/MA
0 1 2 3 4
0 x x o o x
1 x x x o o
2 x x o o o
3 x x x o o
4 x x o x o

2560
RMSE
17069
AICc

ARMA(3,4)
Seasonal Differencing

AR/MA
0 1 2 3 4
0 x x x x x
1 x x x x o
2 x x x x x
3 o x x x o
4 x x x x o

1758
RMSE
16286
AICc

sARIMA(0,0,0)(0,11)[7]
BoxCox

1747
RMSE
5476
AICc

sARIMA(0,1,1)(1,11)[7]
BoxCox

2199
RMSE
5487
AICc

sARIMA(1,0,1)(1,11)[7]
BoxCox

1712
RMSE
5475
AICc

sARIMA(0,1,1)(1,0,1)[7]
BoxCox

1785
RMSE
5528
AICc

MODEL SUMMARY: EACF/ARIMA

ARMA(1,4)
First Order Differencing

AR/MA

	0	1	2	3	4
0	x	x	o	o	x
1	x	x	x	o	o
2	x	x	o	o	o
3	x	x	x	o	o
4	x	x	o	x	o

2559.93
RMSE
17069.47
AICc

ARMA(3,4)
Seasonal Differencing

AR/MA

	0	1	2	3	4
0	x	x	x	x	x
1	x	x	x	x	o
2	x	x	x	x	x
3	o	x	x	x	o
4	x	x	x	x	o

1758.43
RMSE
16286.52
AICc

sARIMA(0,0,0)(0,11)[7]
BoxCox

1747
RMSE
5476
AICc

sARIMA(0,1,1)(1,11)[7]
BoxCox

2199
RMSE
5487
AICc

sARIMA(1,0,1)(1,11)[7]
BoxCox

1712
RMSE
5475
AICc

sARIMA(0,1,1)(1,0,1)[7]
BoxCox

1785
RMSE
5528
AICc

ARIMA MODELS

ARIMA(1,0,0)(1,1,0)[7] with drift – Auto Arima

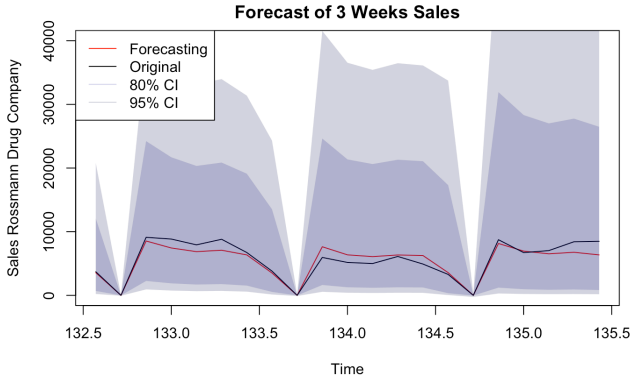
5805	1022	<0.01
AICc	RMSE	Ljung-Box test p-value

ARIMA(1,0,1)(1,1,1)[7] with drift

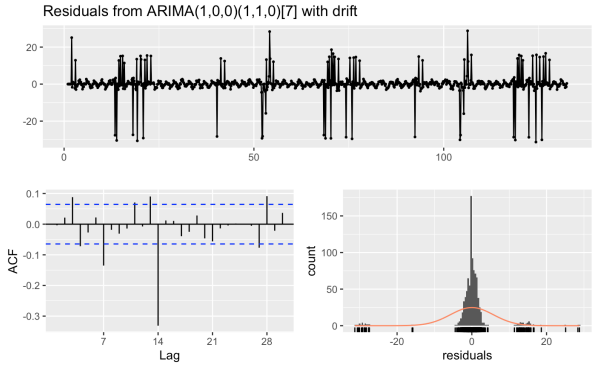
5474	1889	<0.01
AICc	RMSE	Ljung-Box test p-value

- Simpler Model
- Better Validation Result
- Both didn’t generate white noise

ARIMA(1,0,0)(1,1,0)[7] with drift



Residuals from model still showed auto-correlation



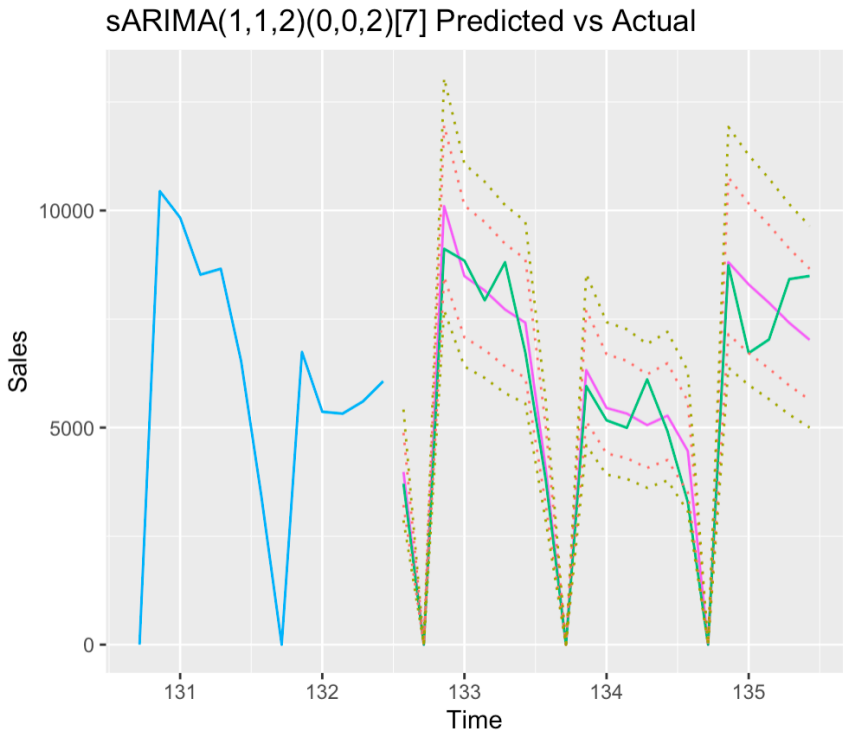
MODEL SUMMARY: LINEAR MODEL WITH ARIMA ERROR

Date	Sales	Customers	Open	Promo	StateHoliday
0.15	1.00	0.97	0.77	0.69	-0.25
SchoolHoliday	year	month	week	day	DayOfWeek
0.11	0.13	0.04	0.04	-0.03	-0.71

Variables chosen based on correlation to sales (magnitude > 0.5)

- Open: as expected, there is a strong correlation between sales and whether or not a store is open
- Promo: we expect that promotions have a positive correlation with a store's sales
- DayOfWeek: negative correlation aligns with Sundays (7) being closed

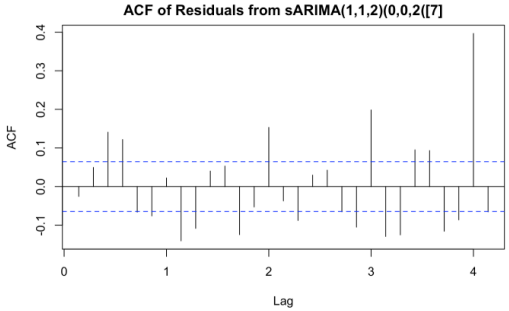
MODEL SUMMARY: LINEAR MODEL WITH ARIMA ERROR



764
RMSE

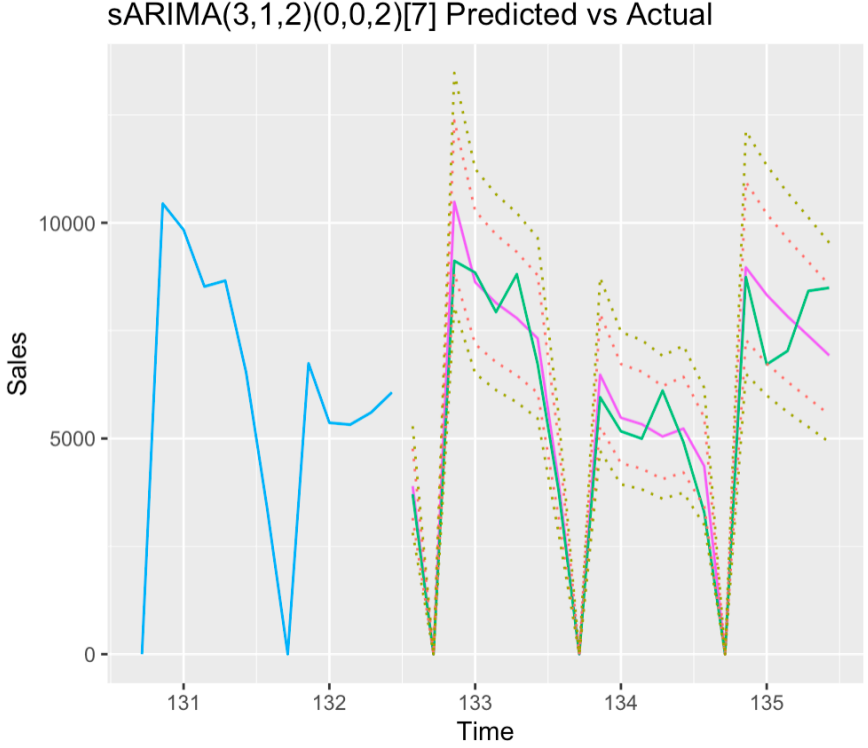
2400
AICc

<0.01
Ljung-Box test p-value




```
Series: sales
Regression with ARIMA(1,1,2)(0,0,2)[7] errors
Box Cox transformation: lambda= 0.2045416

Coefficients:
          ar1      ma1      ma2     sma1    sma2     Open   Promo  DayOfWeek
      -0.9029  -0.0455  -0.9161  0.3345  0.4026  26.9716  2.3035    -0.3542
s.e.   0.0512   0.0421   0.0408  0.0393  0.0293   0.1489  0.0647     0.0297

sigma^2 estimated as 0.7824:  log likelihood=-1191.09
AIC=2400.19   AICc=2400.38   BIC=2443.6

Training set error measures:
                    ME     RMSE      MAE MPE MAPE      MASE      ACF1
Training set 58.94127 740.8819 526.195 NaN  Inf 0.2935999 0.1634032
```

MODEL SUMMARY: LINEAR MODEL WITH ARIMA ERROR



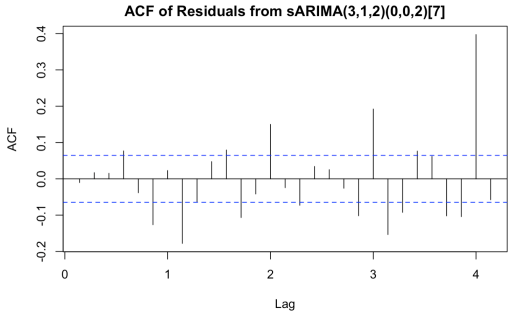
790
RMSE

2373
AICc

<0.01
Ljung-Box test p-value

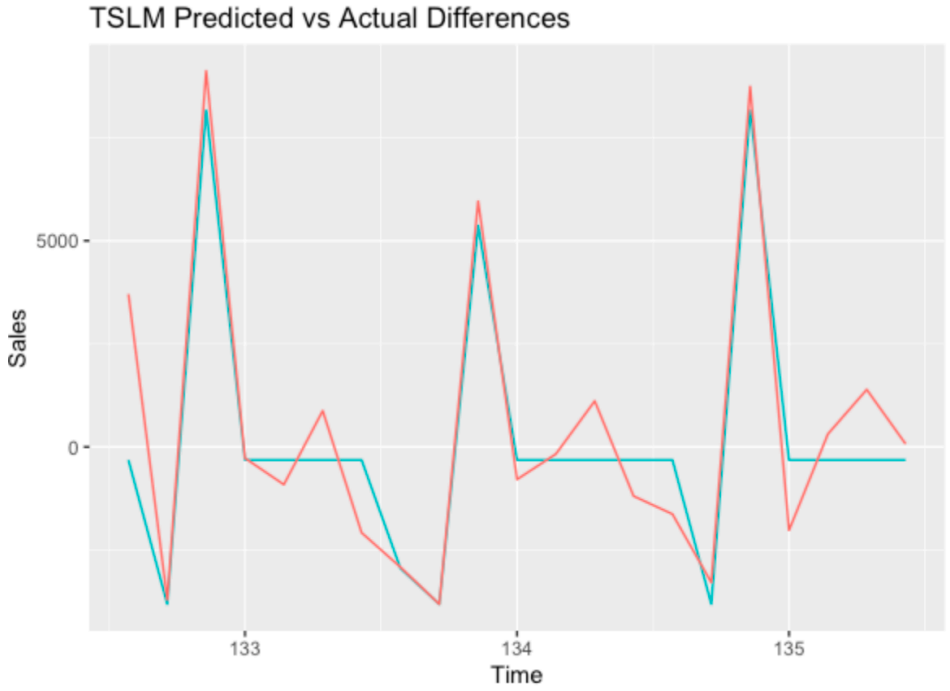
AR/MA

	0	1	2	3	4	5
0	x	x	x	x	x	o
1	x	x	o	o	x	o
2	x	x	x	o	o	x
3	x	x	o	o	o	o
4	x	x	o	o	o	x
5	x	x	o	x	x	x




```
Series: sales
Regression with ARIMA(3,1,2)(0,0,2)[7] errors
Box Cox transformation: lambda= 0.2045416

Coefficients:
          ar1     ar2     ar3      ma1     ma2     sma1    sma2     Open   Promo  DayOfWeek
      0.2607  0.0412  0.1844  -1.2677  0.2771  0.3095  0.3982  26.9737  2.2855    -0.3769
s.e.  0.0949  0.0352  0.0345   0.0939  0.0924  0.0396  0.0288   0.1341  0.0642     0.0255

sigma^2 estimated as 0.7582:  log likelihood=-1175.57
AIC=2373.14   AICc=2373.43   BIC=2426.21

Training set error measures:
                     ME     RMSE      MAE MPE MAPE      MASE      ACF1
Training set 60.93221 719.4383 511.3967 NaN  Inf 0.2853429 0.1843593
```

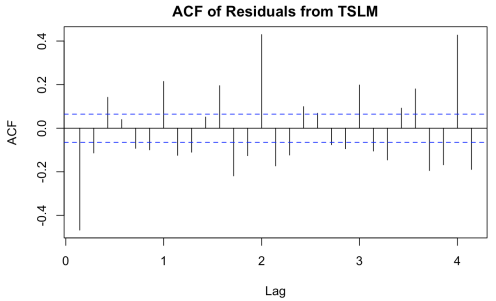
MODEL SUMMARY: TIME SERIES LINEAR REGRESSION



5748
RMSE

15162
AICc

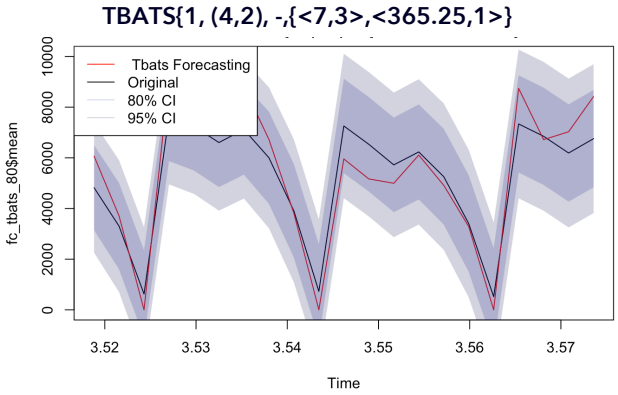
<0.01
Ljung-Box test p-value

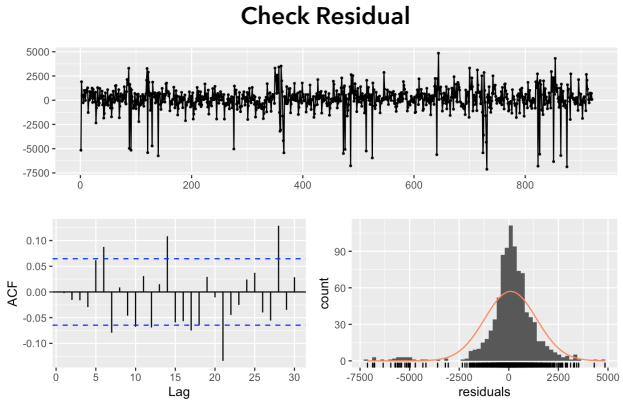



```
Call:
tslm(formula = sales_diff1 ~ predictors_diff1, lambda = "auto")

Residuals:
    Min      1Q  Median      3Q     Max
-4752.6  -597.0   -21.0   651.7  3343.7

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -29.50      30.06  -0.981    0.327
predictors_diff1open         2608.70      73.83  35.332   <2e-16 ***
predictors_diff1promo        1978.67      93.40  21.185   <2e-16 ***
predictors_diff1dayofweek    -234.91      19.44 -12.082   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 911.9 on 916 degrees of freedom
Multiple R-squared:  0.8688,    Adjusted R-squared:  0.8684
F-statistic:  2022 on 3 and 916 DF,  p-value: < 2.2e-16
```

ADVANCED MODELS

ARIMA Models

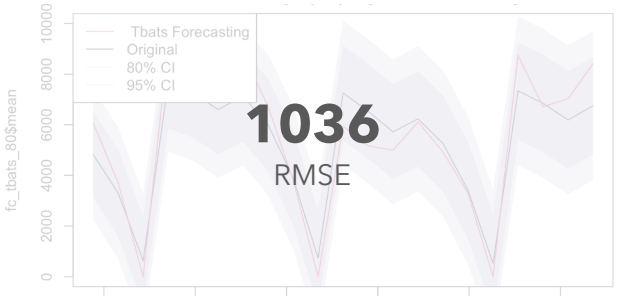




ADVANCED MODELS

ARIMA Models

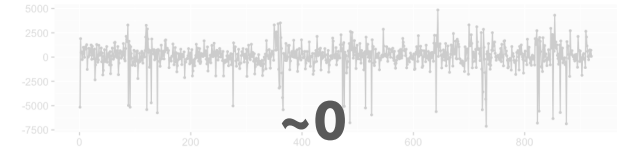
TBATS{1, (4,2), -,{<7,3>,<365.25,1>}



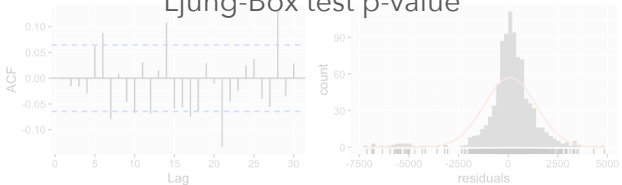
1036
RMSE

19555
AIC

Check Residual

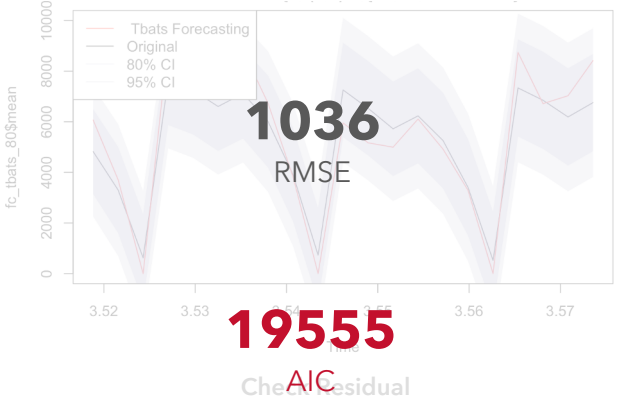


~0
Ljung-Box test p-value

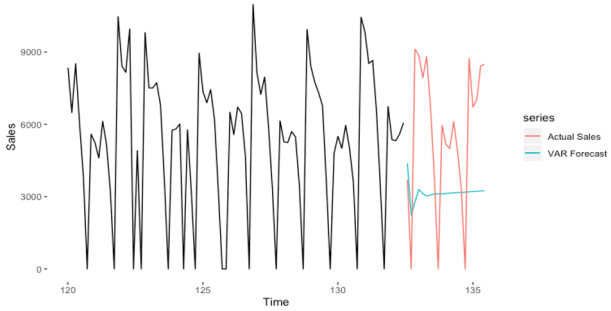


ADVANCED MODELS

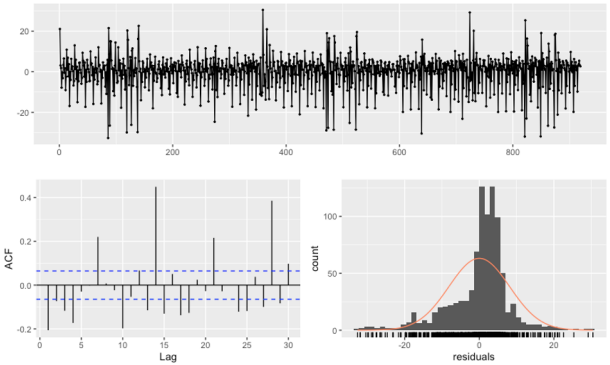
TBATS{1, (4,2), -,{<7,3>,<365.25,1>}



VAR(1)

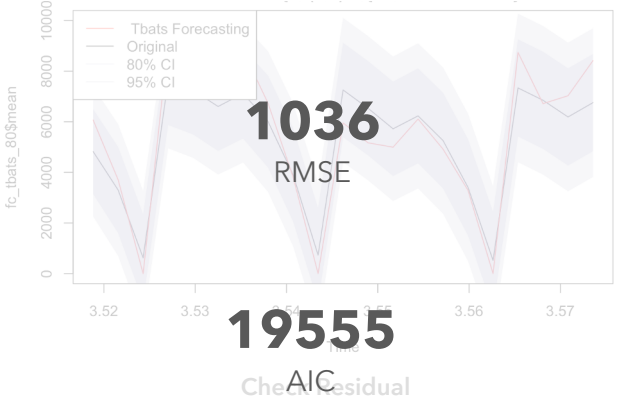


Check Residual



ADVANCED MODELS

TBATS{1, (4,2), -,{<7,3>,<365.25,1>}

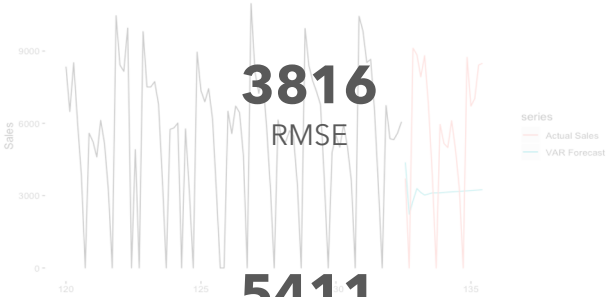


1036
RMSE

19555
AIC

~0
Ljung-Box test p-value

VAR(1)



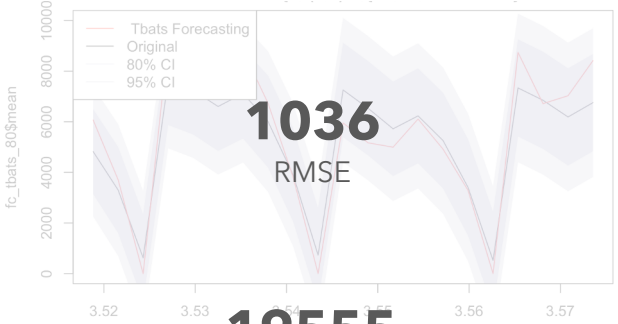
3816
RMSE

5411
AIC

0.01
Ljung-Box test p-value

ADVANCED MODELS

TBATS{1, (4,2), -,{<7,3>,<365.25,1>}

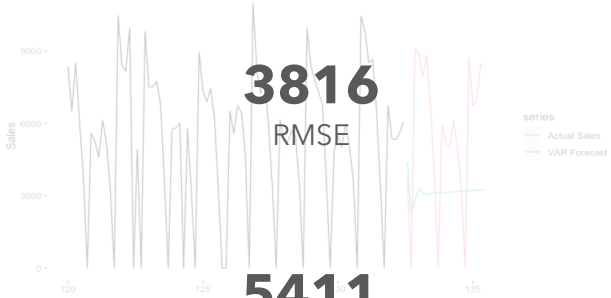


1036
RMSE

19555
AIC

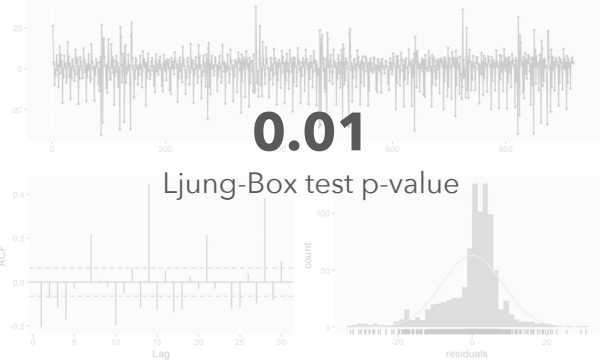
~0
Ljung-Box test p-value

VAR(1)



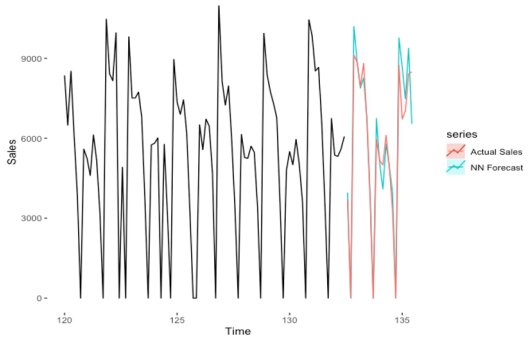
3816
RMSE

5411
AIC

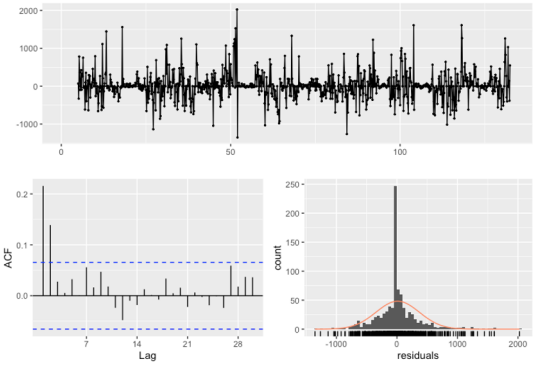


0.01
Ljung-Box test p-value

Neural Network with seasonal component

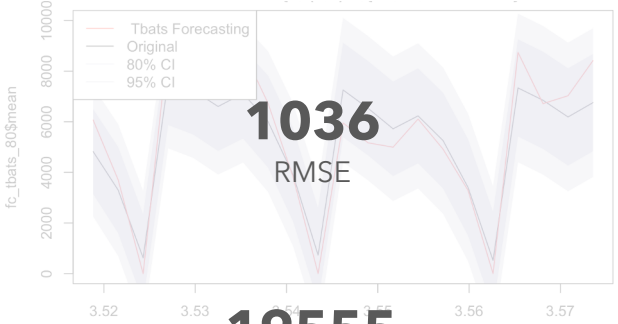


Check Residuals



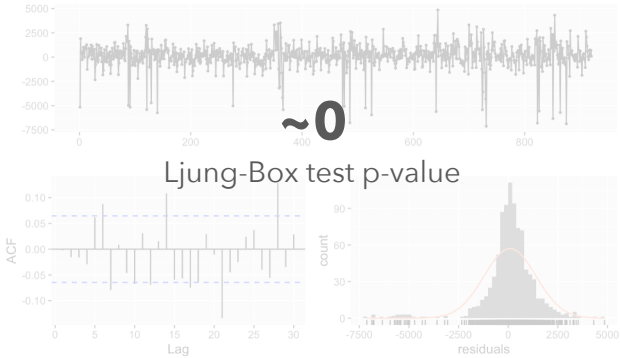
ADVANCED MODELS

TBATS{1, (4,2), -,{<7,3>,<365.25,1>}



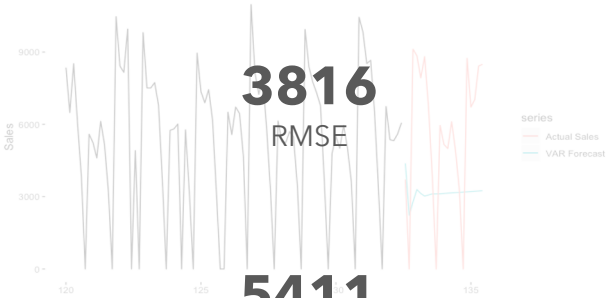
1036
RMSE

19555
AIC



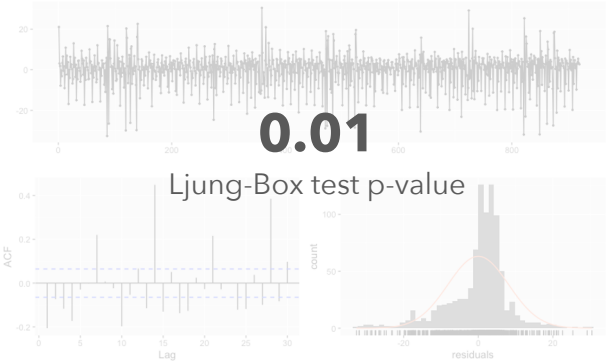
~0
Ljung-Box test p-value

VAR(1)



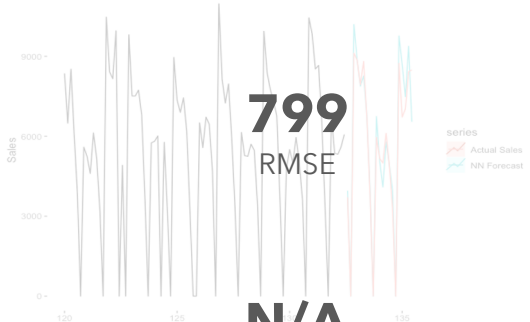
3816
RMSE

5411
AIC



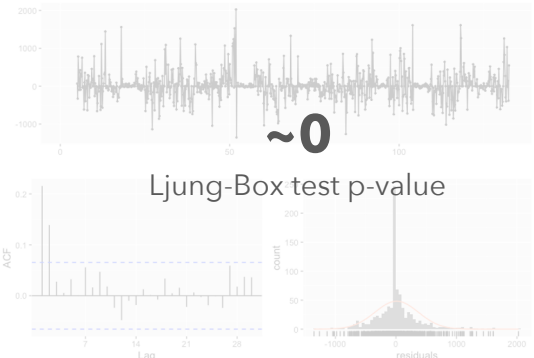
0.01
Ljung-Box test p-value

Neural Network with seasonal component



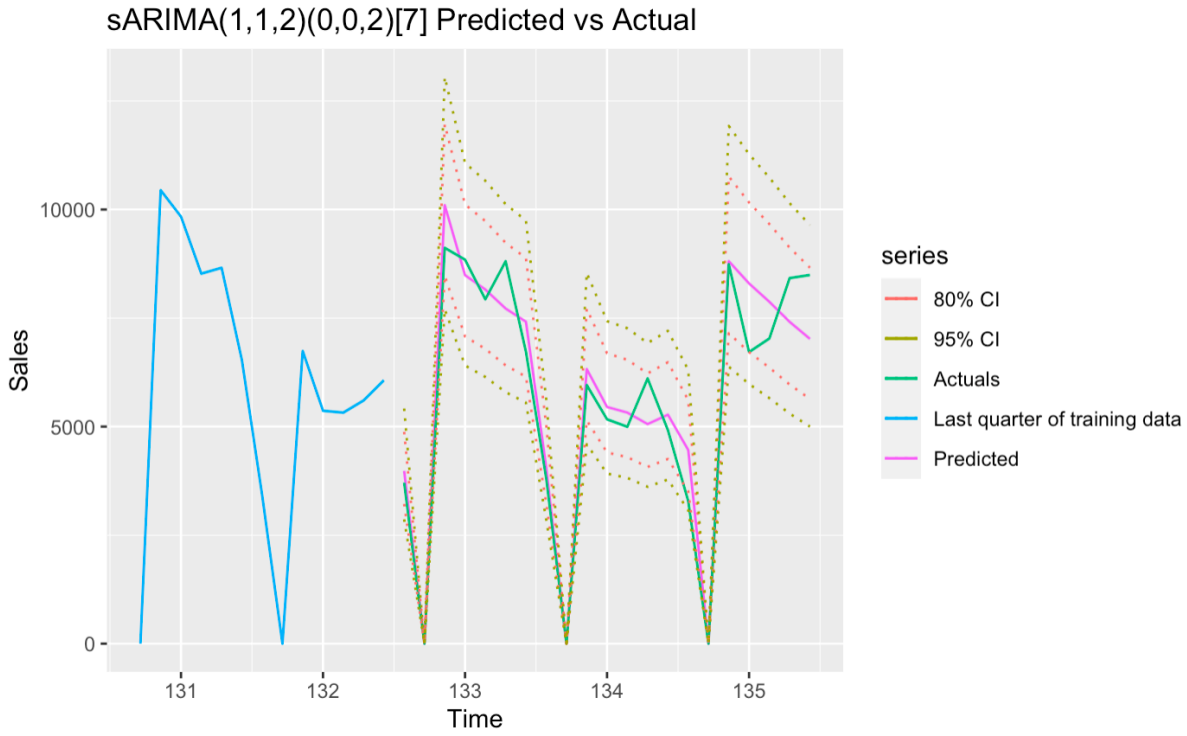
799
RMSE

N/A
AIC



~0
Ljung-Box test p-value

CONCLUSION



764
RMSE

2400
AICc

<0.01
Ljung-Box test p-value

FUTURE WORK

Feature Engineering

Scalability

Additional analysis

Thank you!

Questions?

Slide template credits to slidesgo.com and freepik.com