UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Weather Driven Sales Prediction
Regression Analysis

Walmart

Group 9

Adil Alkhateeb
Linda Wong
Mahammad Ali

April 21st , 2018

Overview

- *Walmart Challenge*
- *Dataset Overview & Preparation*
- *Regression Analysis*
- *Conclusion*
- *Q & A*

Walmart Challenge

kaggle

Walmart Recruiting II: Sales in Stormy Weather
Predict how sales of weather-sensitive products are affected by snow and rain
Recruitment · 3 years ago · tabular data, regression

Jobs
485 teams

Predict the sales of 111 potentially weather-sensitive products (like umbrellas, bread, and milk) around the time of major weather events at 45 of their retail stores

20 Automated Weather Observing System (AWOS) stations covering 45 stores

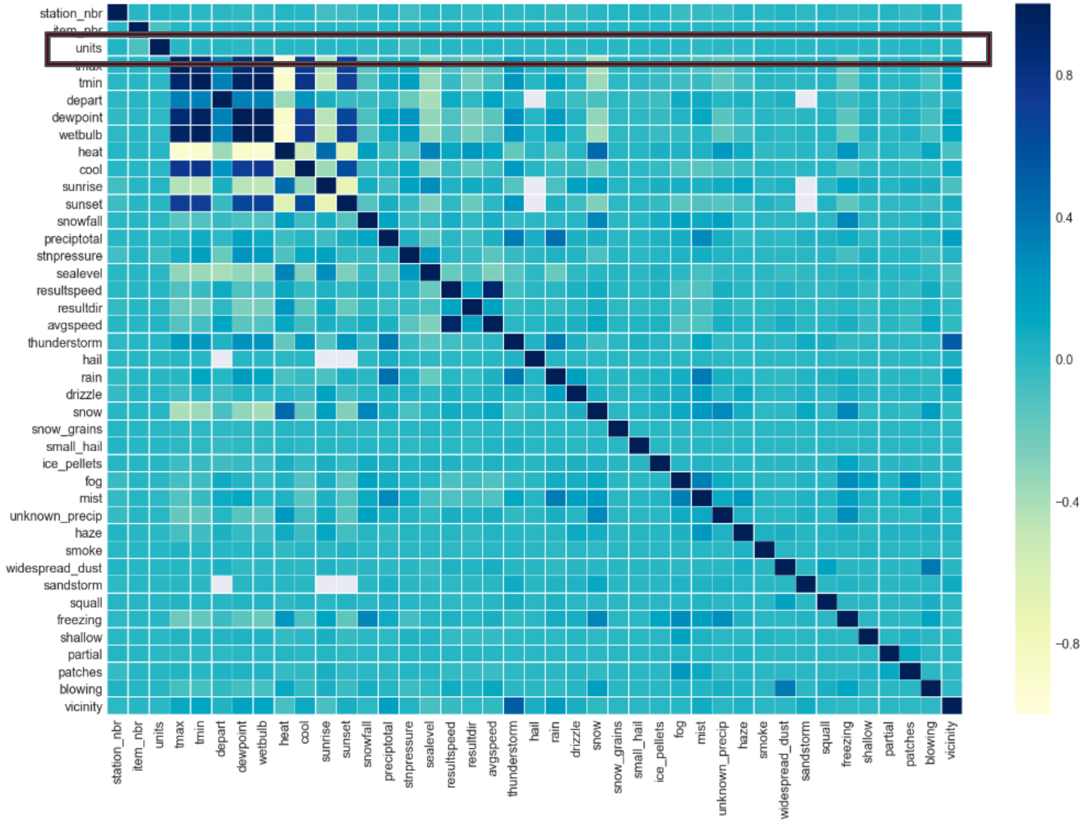
Daily weather measurements of 18 local climatological data

tmax	tmin	tavg	depart	dewpoint	wetbulb	heat	cool	sunrise
sunset	codesum	snowfall	preciptotal	stnpressure	sealevel	resultspeed	resultdir	avgspeed

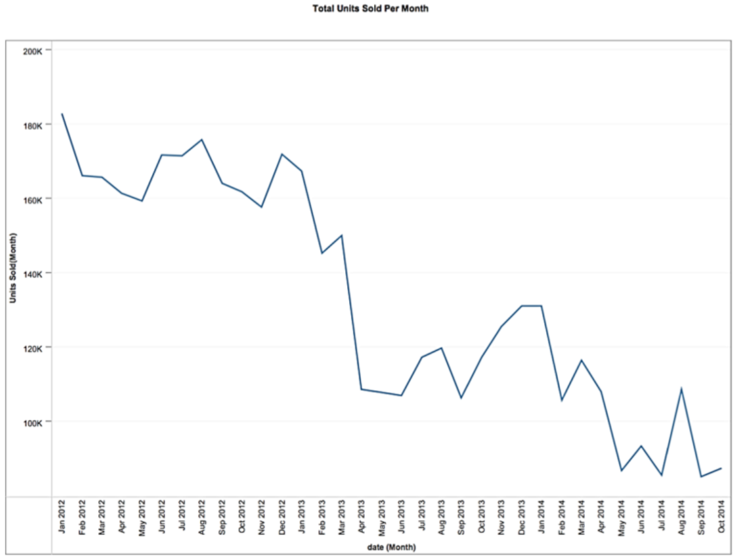
Daily products sales per store

Duration from Jan 2012 to Oct 2014

Dataset Overview



Minimal to moderate correlation between units and weather condition



Declining sales trend

station_nbr	date	tmax	tmin	tavg	depart	dewpoint	wetbulb	heat	cool	sunrise	sunset	codesum	snowfall	preciptotal	stnpressure	sealevel	resultspeed	resultdir	avgspeed
1	2012-01-01	52	31	42	M	36	40	23	0	-	-	RA FZFG BR	M	0.05	29.78	29.92	3.6	20	4.6
2	2012-01-01	48	33	41	16	37	39	24	0	716	1626	RA	0	0.07	28.82	29.91	9.1	23	11.3
3	2012-01-01	55	34	45	9	24	36	20	0	735	1720		0	0	29.77	30.47	9.9	31	10
4	2012-01-01	63	47	55	4	28	43	10	0	728	1742		0	0	29.79	30.48	8	35	8.2
6	2012-01-01	63	34	49	0	31	43	16	0	727	1742		0	0	29.95	30.47	14	36	13.8
7	2012-01-01	50	33	42	M	26	35	23	0	-	-		0	0	29.15	30.54	10.3	32	10.2

Top 3 selling items 45, 5, and 9

Products masked into item numbers only to maintain their anonymity and reduce potential prediction bias

Walmart

Data Preparation

Missing Data Filling by Interpolation

Using the surrounding days within the same station

```
for i in range(stations.size):
    weather.loc[weather.station_nbr == stations[i]] = weather.loc[weather.station_nbr == stations[i]]\
    .interpolate(method='time',limit_direction = "both")
```

date	station_nbr	tmax	tmin	depart	dewpoint
2012-05-30	20	91.0	68.0	NaN	63.0
2012-05-31	20	NaN	NaN	NaN	NaN
2012-06-01	20	87.0	58.0	NaN	50.0

Encoding weather phenomena flags into 32 binary features

Expanded the weather features significantly from 18 to 49 feature

codesum
RA FZFG BR
RA
NaN
NaN
NaN

	rain	freezing_rain	fog	mist
0	1	0	1	1
1	1	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0

Items - ~~Stores~~ - Stations linking

The final combined clean DataFrame had a total of 2,038,737 rows and 41 columns with 18,367 records per item

Multiple Regression Analysis

Multiple Regression Analysis

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

Walmart

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

Backwards

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

Walmart

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted

Backwards

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted

Evaluation Criteria: **R2_Adj**

Folds	item	selection	R2_Adj	MSE
fold1	45	Forward	0.560251715	13478.00381
fold2	45	Forward	0.565887125	14433.33723
fold3	45	Forward	0.563067961	13811.41438
fold4	45	Forward	0.56039181	14609.54234
fold5	45	Forward	0.566636608	13935.26545

Walmart

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted

Backwards

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted

Evaluation Criteria: **R2_Adj**

Folds	item	selection	R2_Adj	MSE
fold1	45	Forward	0.560251715	13478.00381
fold2	45	Forward	0.565887125	14433.33723
fold3	45	Forward	0.563067961	13811.41438
fold4	45	Forward	0.56039181	14609.54234
fold5	45	Forward	0.566636608	13935.26545

Over Fitting

Walmart

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted	**MSE**

Backwards

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted	**MSE**

Walmart

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted	**MSE**

Backwards

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted	**MSE**

Evaluation Criteria:

			R2_Adj		MSE		MSE Improvement	
Folds	*item*	*selection*	*R2_Adj*	*MSE*	*R2_Adj*	*MSE*	*MSE Delta*	*%*
fold1	45	Forward	0.560251715	13478.00381	0.450041386	6377.37534	-7100.63	-53%
fold2	45	Forward	0.565887125	14433.33723	0.449557781	6731.300762	-7702.04	-53%
fold3	45	Forward	0.563067961	13811.41438	0.449110138	6450.697522	-7360.72	-53%
fold4	45	Forward	0.56039181	14609.54234	0.457753509	7176.632181	-7432.91	-51%
fold5	45	Forward	0.566636608	13935.26545	0.464221146	7389.851892	-6545.41	-47%

Multiple Regression Analysis

Forward

Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted | **MSE**

Backwards

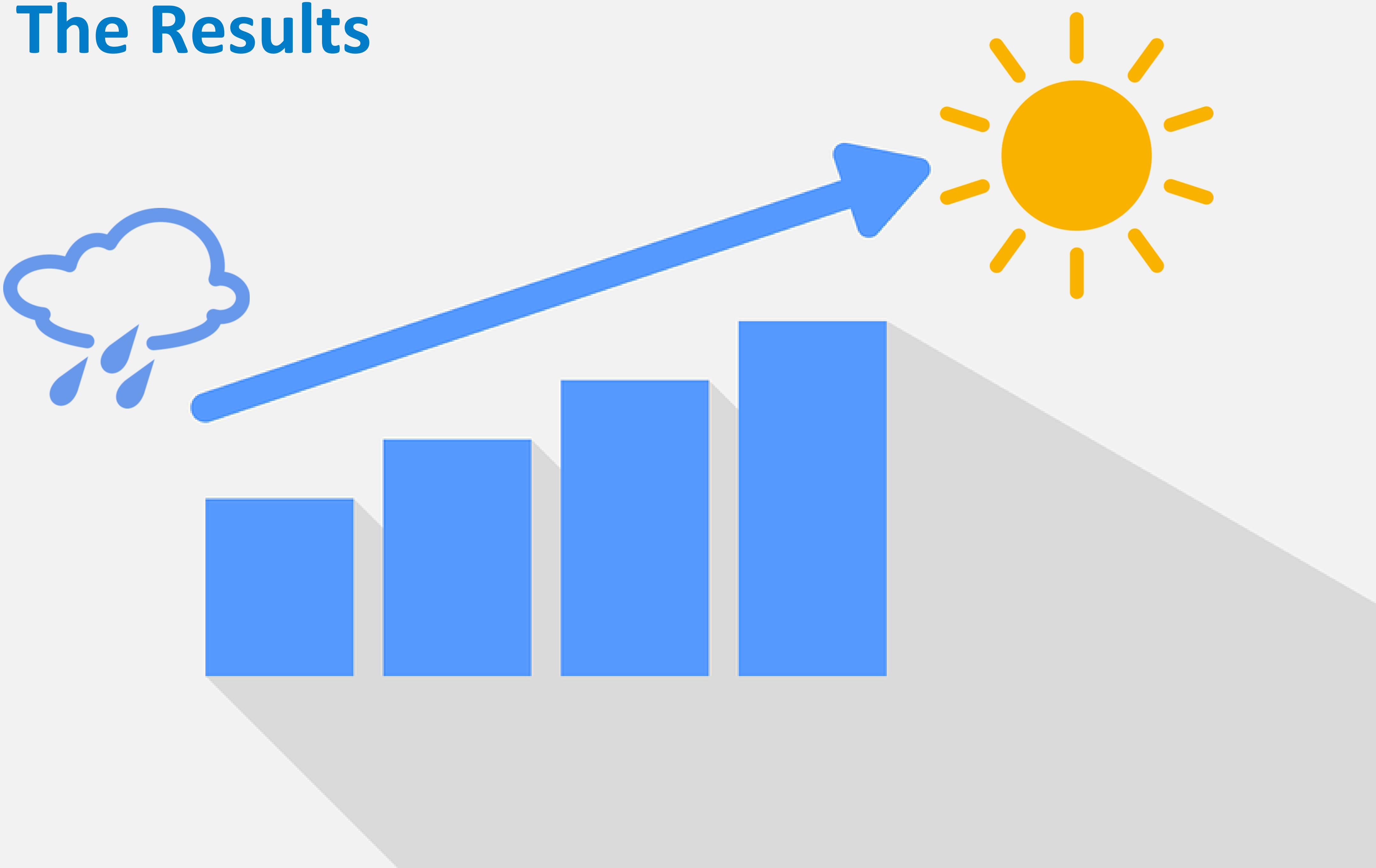
Train / Test
80 : 20

Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

R^2 Adjusted | **MSE**

Evaluation Criteria:			**R2_Adj**		**MSE**		**MSE Improvement**	
Folds	*item*	*selection*	*R2_Adj*	*MSE*	*R2_Adj*	*MSE*	*MSE Delta*	*%*
fold1	45	Forward	0.560251715	13478.00381	0.450041386	6377.37534	-7100.63	-53%
fold2	45	Forward	0.565887125	14433.33723	0.449557781	6731.300762	-7702.04	-53%
fold3	45	Forward	0.563067961	13811.41438	0.449110138	6450.697522	-7360.72	-53%
fold4	45	Forward	0.56039181	14609.54234	0.457753509	7176.632181	-7432.91	-51%
fold5	45	Forward	0.566636608	13935.26545	0.464221146	7389.851892	-6545.41	-47%

The Results



Optimum Prediction Models

item_nbr	model	selection	rsquared_adj	MSE
1	<statsmodels.regression.linear_model.Regressio...	backward	0.048489	0.130139
2	<statsmodels.regression.linear_model.Regressio...	backward	0.061097	0.918617
3	<statsmodels.regression.linear_model.Regressio...	backward	0.089020	0.075939
4	<statsmodels.regression.linear_model.Regressio...	backward	0.007371	0.026946
5	<statsmodels.regression.linear_model.Regressio...	forward	0.176935	3611.746326

Runtime
1 hr 40 mins

*Optimum models **serialized** and **saved** using 'Pickle' package for **immediate** prediction*

Item	Quantity Sold	Selection
45	1,005,111	Forward
9	916,615	Forward
5	846,662	Forward
44	577,193	Backward
16	226,772	Backward

__Forward__ selection was better in predicting __high__ sales items

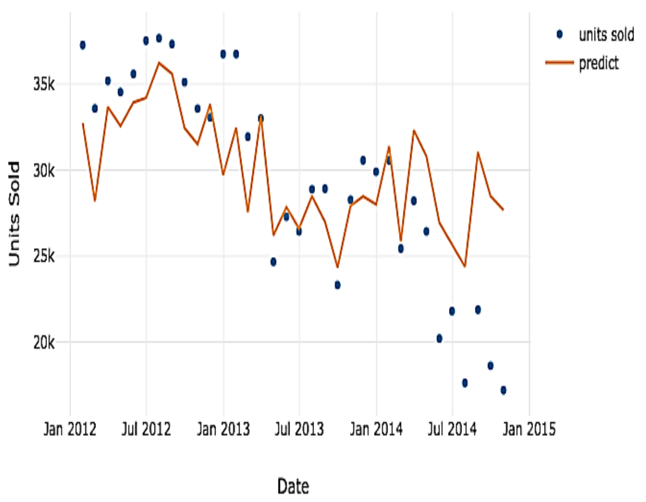
__Backward__ elimination was better in predicting __low__ sales items

Final results had __Backwards__ models selected for __101__ items and __Forward__ models selected for __9__ items

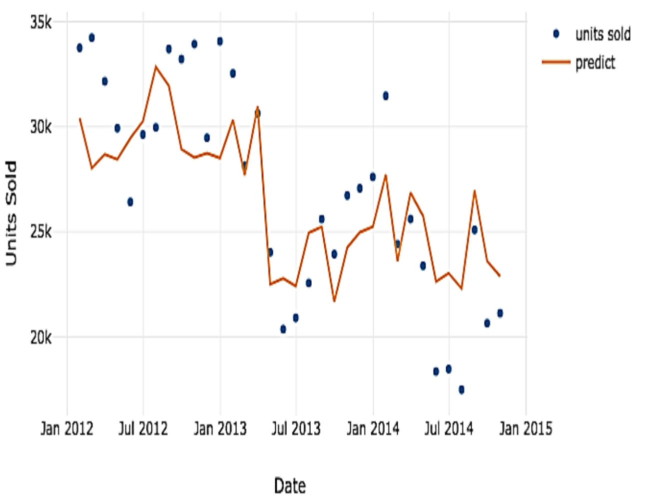
Walmart

Top Three Items Prediction Results

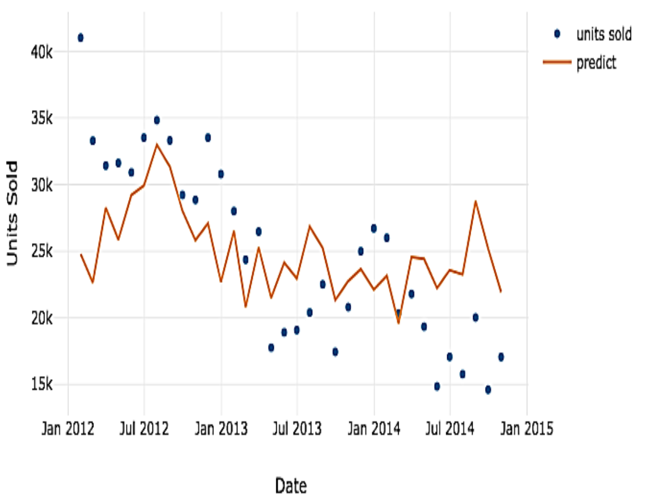
Monthly Sale Prediction for Item 45



Monthly Sale Prediction for Item 9

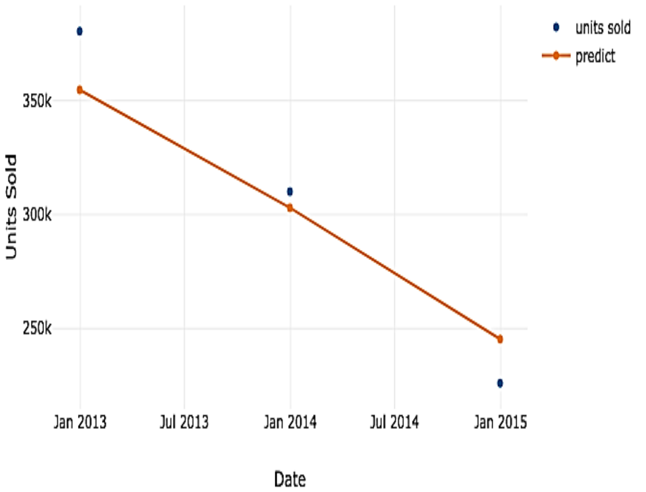


Monthly Sale Prediction for Item 5

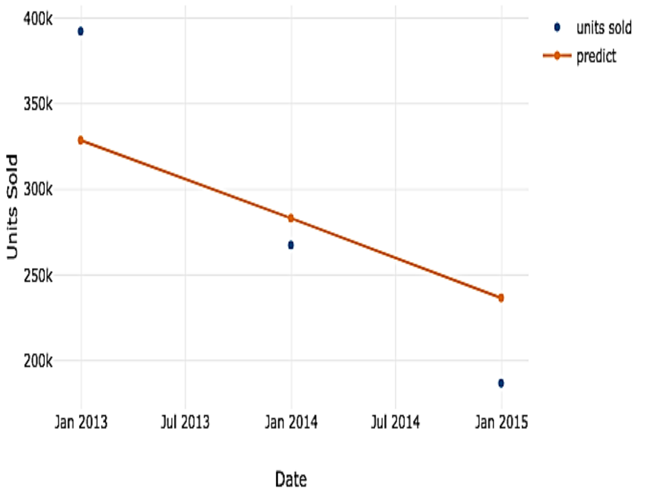


Yearly Sale Prediction for Item 45

Yearly Sale Prediction for Item 9



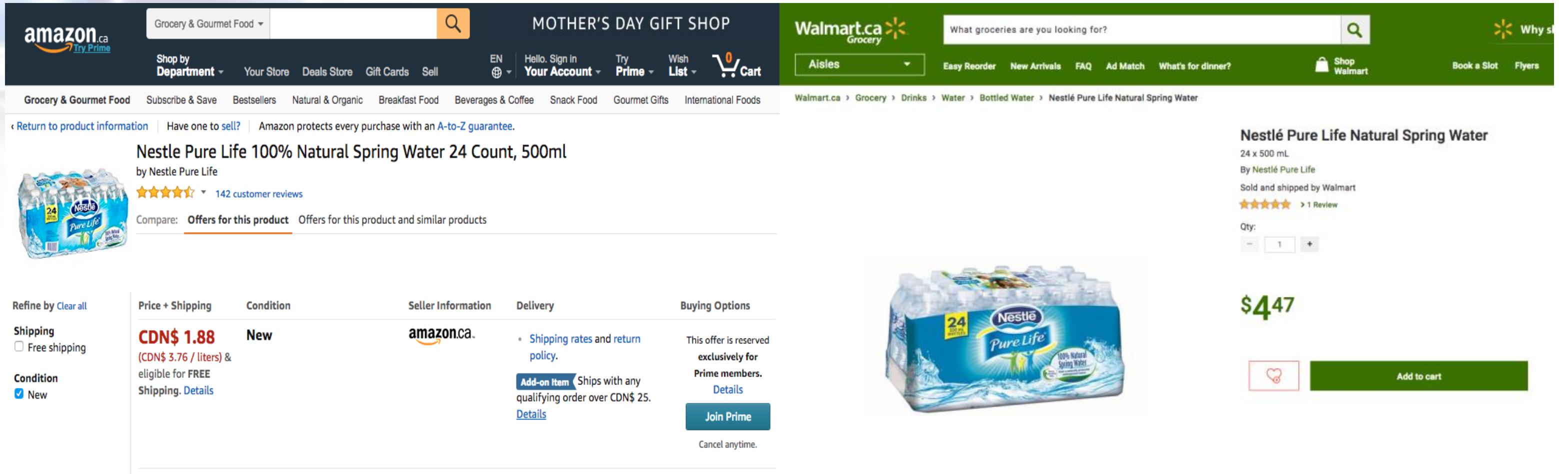
Yearly Sale Prediction for Item 5



Prediction models successfully captured the reducing sales seasonal trend and produced moderately fitted prediction models.

Conclusion

- *Weather may not be a great influencer to consumers buying behavior for basic products*
- *weather based prediction models can be improved if combined with directly related consumer buying influencers*
  - *Day of week, holidays, paycheck days, promotions*
- *Online competition*



Walmart