



Enter

Exit



Today's checklist

-  Project introduction
-  EDA
-  Time series prediction
-  Data preprocessing
-  Model building
-  Best model selection
-  Final prediction and conclusion

- Project introduction
- EDA
- Time Series Prediction
- Data preprocessing
- Model building
- Best model selection
- Final prediction and conclusion

Project introduction

-  Project introduction
-  EDA
-  Time series prediction
-  Data preprocessing
-  Model building
-  Best model selection
-  Final prediction and conclusion

 Data period: 2010/02/05 – 2012/10/26

 Data set shape: 421,570, 16

 No. of stores: 45

 No. of dept: 81

Discount

Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Type	Size
0	1	2010-02-05	24924.50	False	42.31	2.57	NaN	NaN	NaN	NaN	NaN	211.10	8.11	A	151315
1	1	2010-02-05	50605.27	False	42.31	2.57	NaN	NaN	NaN	NaN	NaN	211.10	8.11	A	151315
2	1	2010-02-05	13740.12	False	42.31	2.57	NaN	NaN	NaN	NaN	NaN	211.10	8.11	A	151315
3	1	2010-02-05	39954.04	False	42.31	2.57	NaN	NaN	NaN	NaN	NaN	211.10	8.11	A	151315
4	1	2010-02-05	32229.38	False	42.31	2.57	NaN	NaN	NaN	NaN	NaN	211.10	8.11	A	151315

Our objective:

Predict each store's department Weekly_Sales (y) during 2012/11/2 – 2013/7/26

Eg. 45 stores x 81 departments x 39 weeks

Project introduction

EDA

Time series prediction

Data preprocessing

Model building

Best model selection

Final prediction and conclusion

Sales overview

- The sales trend by year was very similar with no obvious up/down trend between years
- Peak is seen at Xmas (max: \$81M) while trough is at the beginning of the year (min: \$40M), so the gap is a double
- However, 2011 Christmas shows a more significant drop vs 2010
- 4 Holidays noted by Walmart, and we see Easter impact is significant too



- 🔍 Project introduction
- 🔍 EDA
- 🔍 Time series prediction
- 🔍 Data preprocessing
- 🔍 Model building
- 🔍 Best model selection
- 🔍 Final prediction and conclusion

Sales overview / Factors correlation

- **Size and Dept are more positively corr. to weekly sales**
- **For the other variables, most are weakly correlated, so the focus of the exploratory analysis would be on size and dept**
- **Store type is missing here as it is not numerical categorized, will look into it in EDA**

Correlation Matrix

	Store	Dept	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday_y	Size	Year	Month	Week	Day
Store	1	0.024	-0.085	-0.00055	-0.05	0.065	-0.21	0.21	-0.00055	-0.18	0.003	0.001	0.001	-1.5e-05
Dept	0.024	1	0.15	0.00092	0.0044	0.0036	-0.0075	0.0078	0.00092	-0.003	0.0037	0.0009	0.00088	-0.00068
Weekly_Sales	-0.085	0.15	1	0.013	-0.0023	-0.00012	-0.021	-0.026	0.013	0.24	-0.01	0.028	0.028	-0.0062
IsHoliday_x	-0.00055	0.00092	0.013	1	-0.16	-0.078	-0.0019	0.01	1	0.00059	-0.057	0.12	0.13	0.045
Temperature	-0.05	0.0044	-0.0023	-0.16	1	0.14	0.18	0.097	-0.16	-0.058	0.066	0.24	0.24	0.027
Fuel_Price	0.065	0.0036	-0.00012	-0.078	0.14	1	-0.16	-0.034	-0.078	0.0034	0.78	-0.041	-0.031	0.028
CPI	-0.21	-0.0075	-0.021	-0.0019	0.18	-0.16	1	-0.3	-0.0019	-0.0033	0.075	0.0053	0.0063	0.0027
Unemployment	0.21	0.0078	-0.026	0.01	0.097	-0.034	-0.3	1	0.01	-0.068	-0.24	-0.012	-0.015	-0.0038
IsHoliday_y	-0.00055	0.00092	0.013	1	-0.16	-0.078	-0.0019	0.01	1	0.00059	-0.057	0.12	0.13	0.045
Size	-0.18	-0.003	0.24	0.00059	-0.058	0.0034	-0.0033	-0.068	0.00059	1	-0.005	-0.0012	-0.0013	-0.00039
Year	0.003	0.0037	-0.01	-0.057	0.066	0.78	0.075	-0.24	-0.057	-0.005	1	-0.19	-0.18	0.0058
Month	0.001	0.0009	0.028	0.12	0.24	-0.041	0.0053	-0.012	0.12	-0.0012	-0.19	1	1	0.016
Week	0.001	0.00088	0.028	0.13	0.24	-0.031	0.0063	-0.015	0.13	-0.0013	-0.18	1	1	0.1
Day	-1.5e-05	-0.00068	-0.0062	0.045	0.027	0.028	0.0027	-0.0038	0.045	-0.00039	0.0058	0.016	0.1	1





Project introduction



EDA



Time series prediction



Data preprocessing



Model building



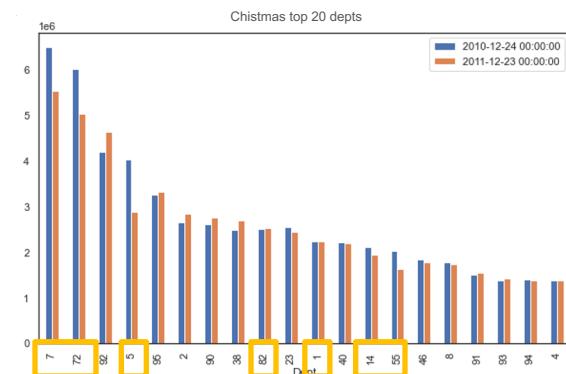
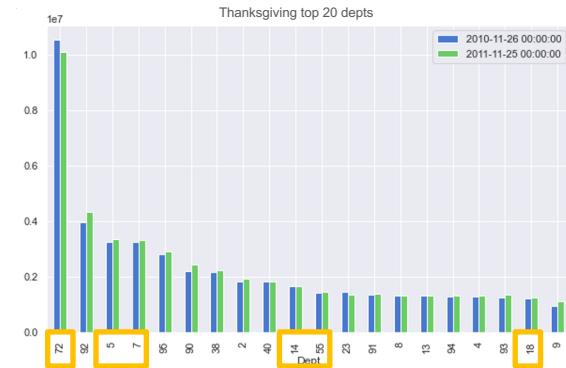
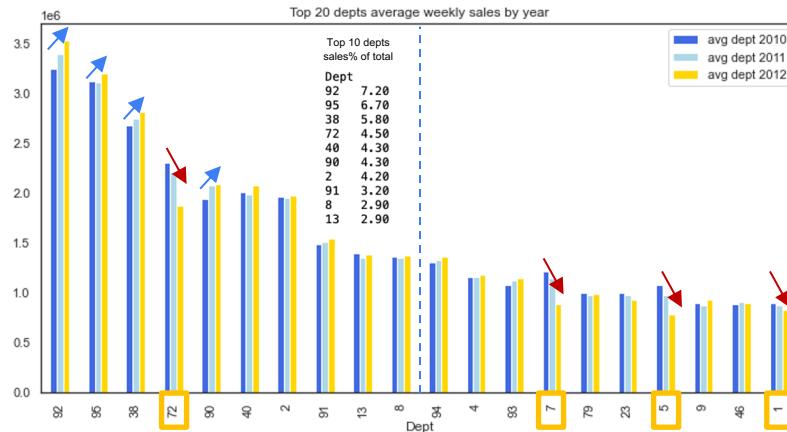
Best model selection



Final prediction and conclusion

Sales overview / Factors correlation / Core depts

- Total 81 depts, top 20 depts account for 70% total sales**
- Esp. dept starting with '9': 4 depts accounting for 20%**
- Performing depts: 92, 95, 38, 90**
- Declining depts: 72, 7, 5, 1**
- Thanksgiving and Christmas product mix is different**
- Festive depts: 72, 7, 5, 14, 82, 55, 1, 18**



Sales overview / Factors correlation / Core depts / Store types

 Project introduction

 EDA

 Time series prediction

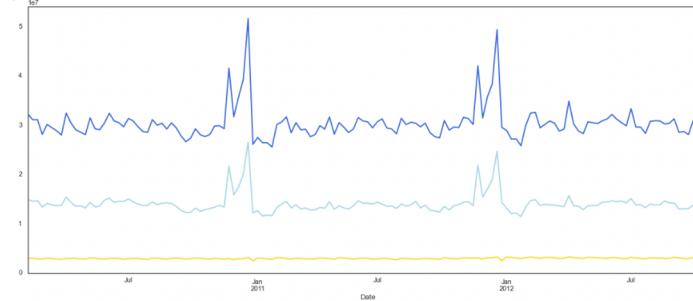
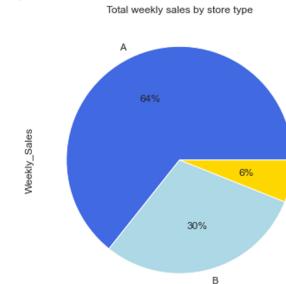
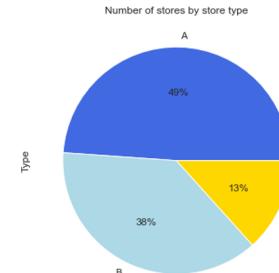
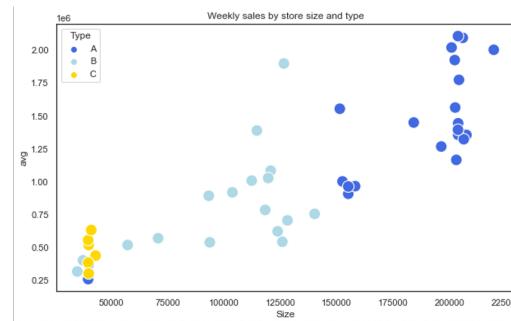
 Data preprocessing

 Model building

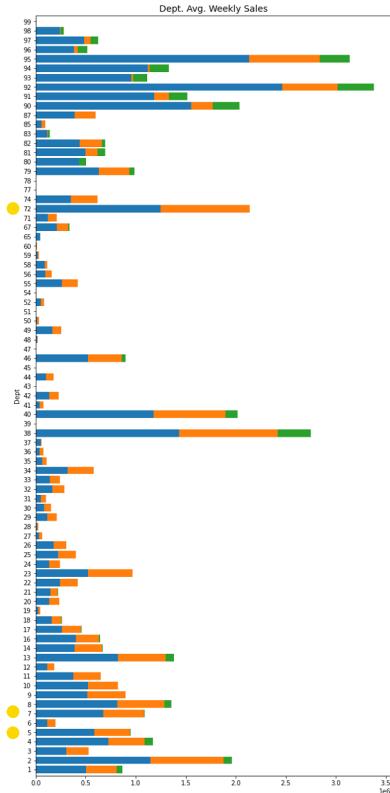
 Best model selection

 Final prediction and conclusion

- **Store types are characterized according to size and income**
- **Larger size, higher sales, smaller size, lower sales**
- **However, in terms of sales per size unit by type: negative corr.**
- **e.g. Type C has the highest sales/unit size, implying highest sales efficiency**
- **Type C is immune to holidays impact because festive depts are rarely carried**



Type	A	B	C
count	22.00	17.00	6.00
mean	177247.73	101190.71	40541.67
mean	1376673.47	822994.96	472614.83
Sales/size unit	7.77	8.13	11.66



 Project introduction

 EDA

 Time series prediction

 Data preprocessing

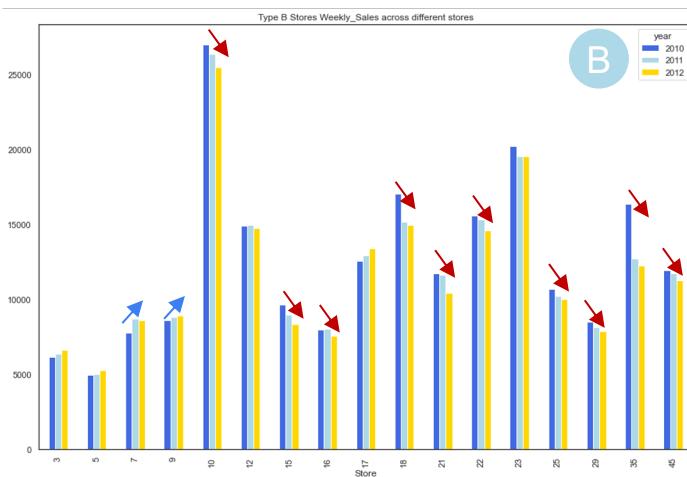
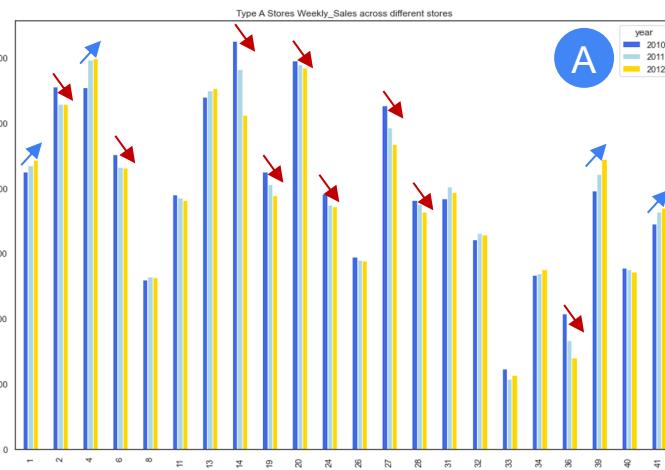
 Model building

 Best model selection

 Final prediction and conclusion

Sales overview / Factors correlation / Core depts / Store types / Stores performance

- 7 stores show growing trend (esp. 4, 39)
- 19 stores show declining trend (esp. 14, 27, 36, 35, 10)
- Need to look into the driving depts for the highlighted stores





Project introduction



EDA



Time series prediction



Data preprocessing



Model building

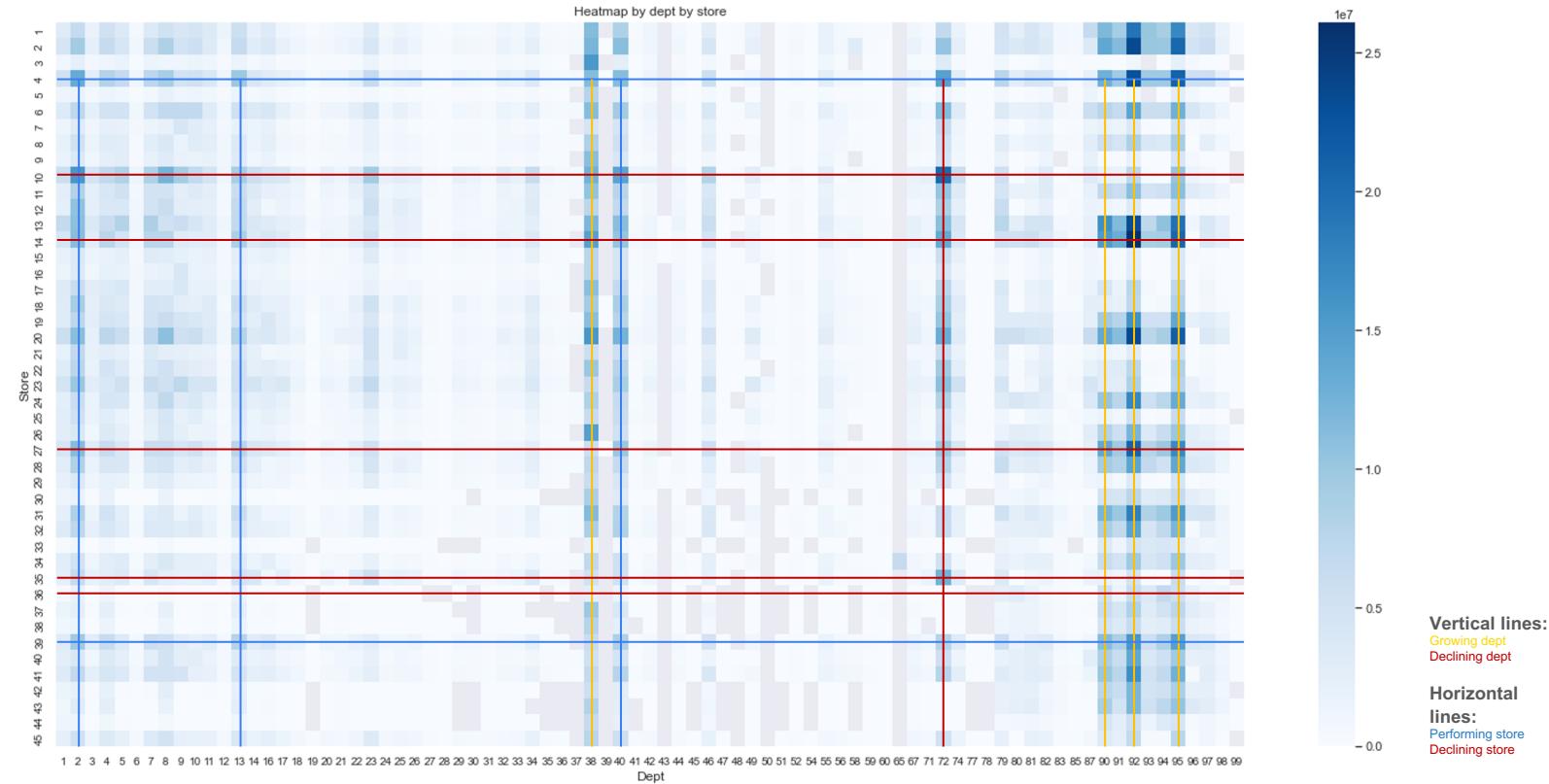


Best model selection



Final prediction and conclusion

Sales overview / Factors correlation / Core depts / Store types / Stores performance





Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection



Final prediction and conclusion



- Easter chance (growth index: 1.1 vs average weekly) > budget allocation shift from Super Bowl & Labor Day
- Festive depts (72, 7, 5) are declining and dragged down 2011 Christmas
- Chances seen for growing depts 92, 95, 90
- Type C stores have the highest sales efficiency, potential store expansion store type
- Type C stores do not carry festive depts much, consider seasonal listing by selecting few top items to create festive corner
- All stores should expand depts 92, 95, 90 shelf space by taking space from under-performing depts, so further range review by store with heatmap is needed
 - Eg. Expand 92, 95, 90 at store 10 where these depts are very weak

Time Series Prediction



Project introduction



EDA



Time series prediction



Data preprocessing



Model building

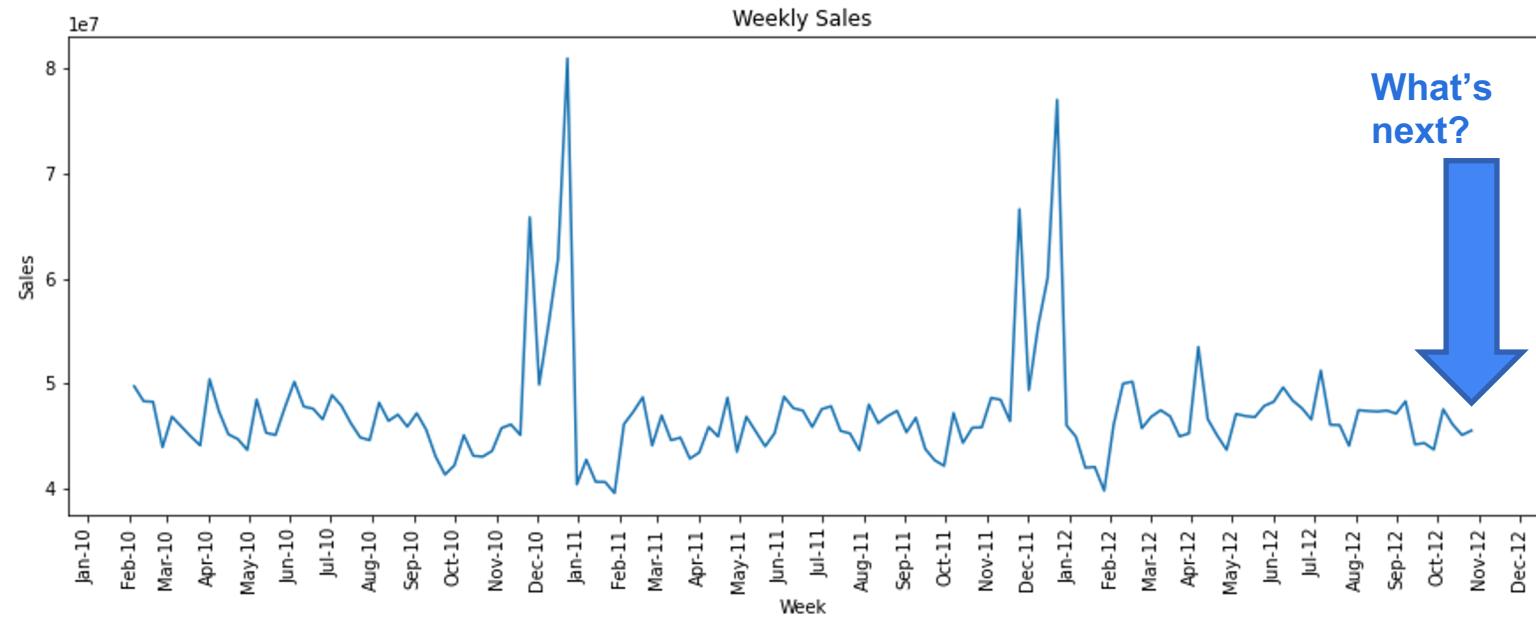


Best model selection



Final prediction and conclusion

Time Series



Time Series Prediction



Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection



Final prediction and conclusion

Time Series/Stationarity Check





Project introduction



EDA



Time series prediction



Data preprocessing



Model building



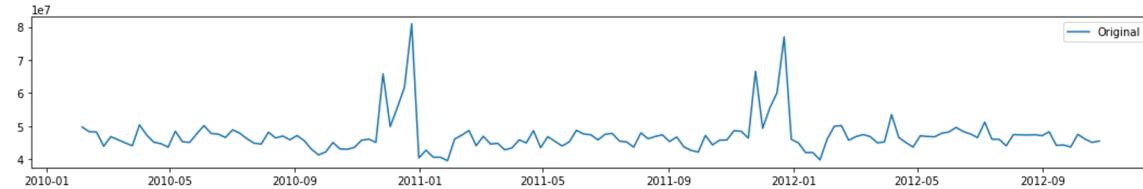
Best model selection



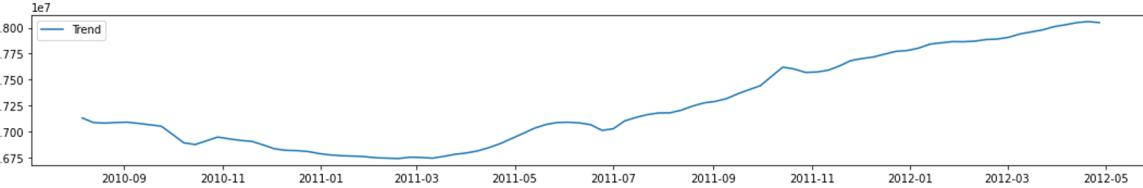
Final prediction and conclusion

Time Series/Stationarity Check/Decomposition

Time Series

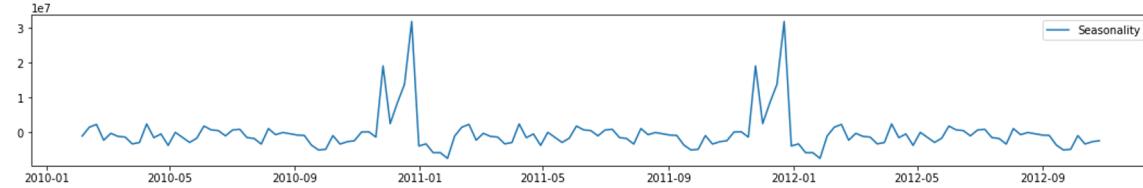


Trend

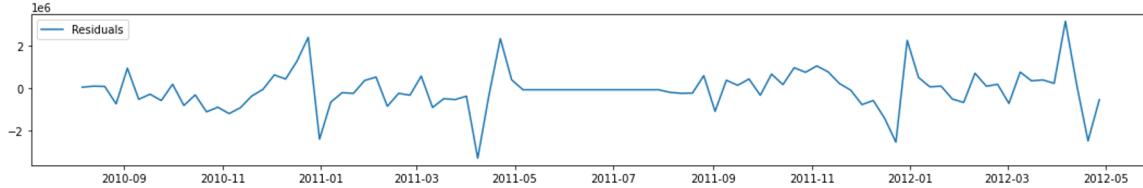


Seasonality

SARIMA model



Residuals



 Project introduction EDA Time series prediction Data preprocessing Model building Best model selection Final prediction and conclusion

Time Series/Stationarity Check/Decomposition/Model Evaluation

SARIMAX Results

```
=====
Dep. Variable: Weekly_Sales No. Observations: 143
Model: SARIMAX(1, 0, 1)x(1, 0, 1, 52) Log Likelihood -354.325
Date: Mon, 21 Feb 2022 AIC 718.650
Time: 02:27:34 BIC 733.464
Sample: 02-05-2010 HQIC 724.670
          - 10-26-2012
Covariance Type: opg
=====
```

	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.9978	0.006	170.329	0.000	0.986	1.009
ma.L1	-0.8931	0.047	-19.009	0.000	-0.985	-0.801
ar.S.L52	0.9976	0.004	268.608	0.000	0.990	1.005
ma.S.L52	-0.7203	0.199	-3.616	0.000	-1.111	-0.330
sigma2	2.3687	0.396	5.987	0.000	1.593	3.144

<=====

Ljung-Box (L1) (Q):	1.89	Jarque-Bera (JB):	252.02
Prob(Q):	0.17	Prob(JB):	0.00
Heteroskedasticity (H):	0.97	Skew:	1.23
Prob(H) (two-sided):	0.90	Kurtosis:	9.02

<=====

Time Series Prediction

Project introduction

EDA

Time series prediction

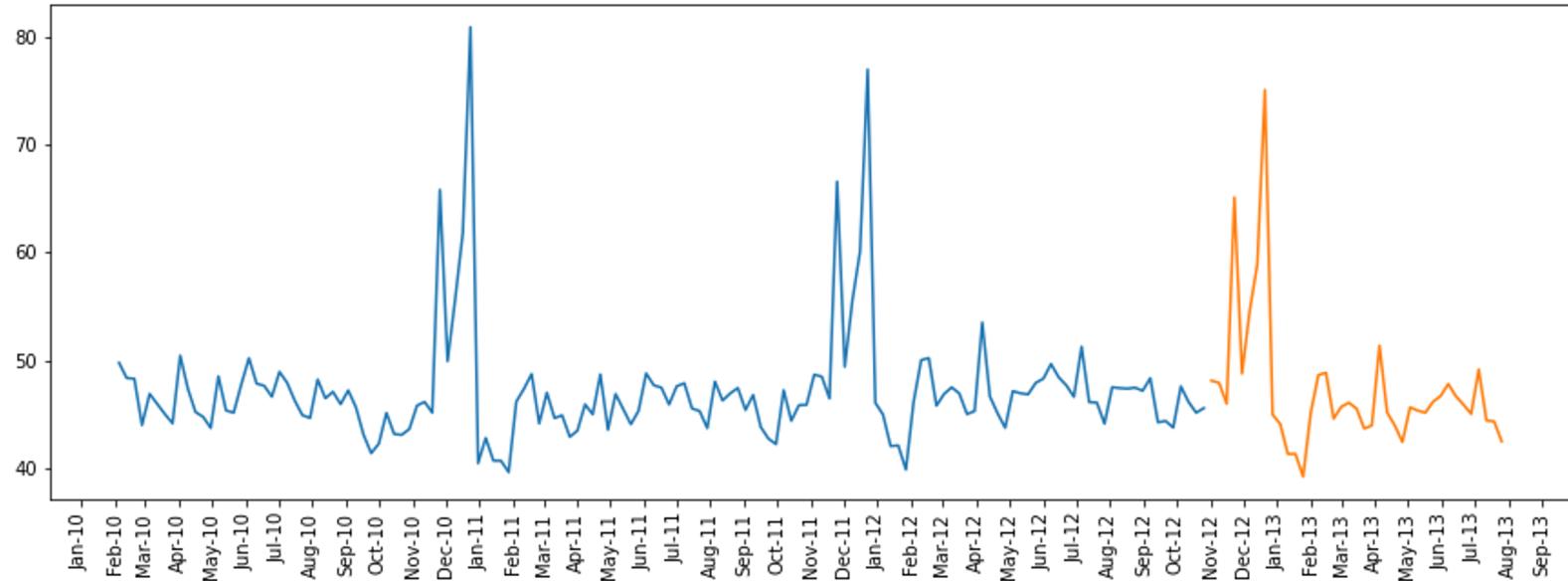
Data preprocessing

Model building

Best model selection

Final prediction and conclusion

Time Series/Stationarity Check/Decomposition/Model Evaluation/Prediction



Data Processing

 Project introduction

 EDA

 Time series prediction

 Data preprocessing

 Model building

 Best model selection

 Final prediction and conclusion

Data Cleaning/Data for Training

	Store	Dept	Weekly_Sales	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3
count	421570.000000	421570.000000	421570.000000	421570.000000	421570.000000	150681.000000	111246.000000	137091.000000
mean	22.200546	44.260317	15981.58123	60.090059	3.361027	7246.420196	3334.28621	1439.41384
std	12.785297	30.492054	22711.83519	18.447931	0.458515	8291.221345	9475.557325	9623.013290
min	1.000000	1.000000	-4988.940000	-2.060000	2.472000	0.270000	-265.760000	-29.100000
25%	11.000000	18.000000	2079.650000	46.680000	2.933000	2240.270000	41.600000	5.080000
50%	22.000000	37.000000	7612.030000	62.090000	3.452000	5347.450000	192.000000	24.600000
75%	33.000000	74.000000	20205.852500	74.280000	3.738000	9210.900000	1926.940000	103.990000
max	45.000000	99.000000	693099.360000	100.140000	4.468000	88646.760000	104519.540000	141630.610000
			MarkDown4	MarkDown5	CPI	Unemployment	Size	
			count	134967.000000	151432.000000	421570.000000	421570.000000	421570.000000
			mean	3383.168256	4628.975079	171.201947	7.960289	136727.915739
			std	6292.384031	5962.887455	39.159276	1.863296	60980.583328
			min	0.220000	135.160000	126.064000	3.879000	34875.000000
			25%	504.220000	1878.440000	132.022667	6.891000	93638.000000
			50%	1481.310000	3359.450000	182.318780	7.866000	140167.000000
			75%	3595.040000	5563.800000	212.416993	8.572000	202505.000000
			max	67474.850000	108519.280000	227.232807	14.313000	219622.000000

Data for training

Data Processing

Project introduction

EDA

Time series prediction

Data preprocessing

Model building

Best model selection

Final prediction and conclusion

Data Cleaning /Data for Training /Data for Forecasting

	Store	Dept	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
count	115064.000000	115064.000000	115064.000000	115064.000000	114913.000000	86437.600000	105235.000000	102176.000000	115064.000000
mean	22.238207	44.339524	18.796040	3.581546	7681.16439	3734.0729	2403.088666	3356.219071	3922.31189
std	12.809930	30.656410	9.543562	0.239442	10691.60716	8323.4014	13767.939313	7570.501545	19445.10745
min	1.000000	1.000000	-18.922222	2.872000	-2781.450000	-35.740000	-179.260000	0.220000	-185.170000
25%	11.000000	18.000000	11.894444	3.431000	1966.460000	180.350000	15.100000	155.460000	1309.300000
50%	22.000000	37.000000	20.000000	3.606000	4842.290000	742.590000	78.260000	840.940000	2390.430000
75%	33.000000	74.000000	27.055556	3.766000	9439.140000	2735.670000	272.580000	3096.920000	4227.270000
max	45.000000	99.000000	37.344444	4.125000	103184.980000	71074.170000	149483.310000	65344.640000	771448.100000

	CPI	Unemployment	Size
count	76902.000000	76902.000000	115064.000000
mean	176.961347	6.868733	136497.688921
std	41.239967	1.583427	61106.926438
min	131.236226	3.684000	34875.000000
25%	138.402033	5.771000	93638.000000
50%	192.304445	6.806000	140167.000000
75%	223.244532	8.036000	202505.000000
max	228.976456	10.199000	219622.000000

Data for forecasting

Save money. Live better.

Data Processing



Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection



Final prediction and conclusion

Data Cleaning /Data for Training /Data for Forecasting/Null Values

	0
Store	0
Dept	0
Date	0
Weekly_Sales	0
IsHoliday	0
Temperature	0
Fuel_Price	0
MarkDown1	270889
MarkDown2	310322
MarkDown3	284479
MarkDown4	286603
MarkDown5	270138
CPI	0
Unemployment	0
Type	0
Size	0

- Markdowns contain null values
- Markdowns have no significant effect on sales
- These features are removed



Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection



Final prediction and conclusion

Data Cleaning /Data for Training /Data for Forecasting/Null Values/Label Encoding

- Store Types: Type A, Type B, Type C
- One-hot encoding
- Date: encoded into week number in the year

Data Processing

Project introduction

EDA

Time series prediction

Data preprocessing

Model building

Best model selection

Final prediction and conclusion

Data Cleaning /Data for Training /Data for Forecasting/Null Values/Label Encoding/Feature Scaling

	Store	Dept	IsHoliday	Temperature	Fuel_Price	CPI	Unemployment	Size	month	week	B	C
0	1	1	0	42.31	2.572	211.096358	8.106	151315	2	1	0	0
1	1	2	0	42.31	2.572	211.096358	8.106	151315	2	1	0	0
2	1	3	0	42.31	2.572	211.096358	8.106	151315	2	1	0	0
3	1	4	0	42.31	2.572	211.096358	8.106	151315	2	1	0	0
4	1	5	0	42.31	2.572	211.096358	8.106	151315	2	1	0	0

- Min-max scaling
- Testing/Training size: 0.25/0.75



Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection



Final prediction and conclusion

Preliminary selection

- All features are used in the preliminary round
- Eliminate not so performing models
- Further improve the models
- The best model will be used to forecast weekly sales

Model building

-  Project introduction
-  EDA
-  Time series prediction
-  Data preprocessing
-  Model building
-  Best model selection
-  Final prediction and conclusion

Preliminary selection/Results

Model	Linear Regression	Lasso	Decision Tree	Random Forest	AdaBoost	Gradient Boost	XGBoost
MAE	14600	14599	1845	1402.7	24677	24677	1901
MSE	4.76E8	4.76E8	2.07E7	1.24E7	8.06E8	8.06E8	1.44E7
RMSE	21812	21916	4555	3518	28393	28393	3799
Variance Score	0.09	0.09	0.96	0.98	-0.54	-0.54	0.97

Model building

Project introduction

EDA

Time series prediction

Data preprocessing

Model building

Best model selection

Final prediction and conclusion

Preliminary selection/Results/Model Fine-Tuning

	Model	KPIs	PCA (explained var.=0.95)	Feature Importance	HPO
 Random Forest	MAE	2599.07			1479.42
	RMSE	7314.41			3633.08
	R2 Score	0.89 ↓			0.97 -
 XGBoost	MAE	7248.34			2715.22
	RMSE	12331.89			4931.27
	R2 Score	0.70 ↓			0.95 -
 Decision Tree	MAE	3180.04			5991.13
	RMSE	10137.45			9847.82
	R2 Score	0.80 ↓			0.81 ↓

Final Prediction and Conclusion



Project introduction



EDA



Time series prediction



Data preprocessing



Model building



Best model selection

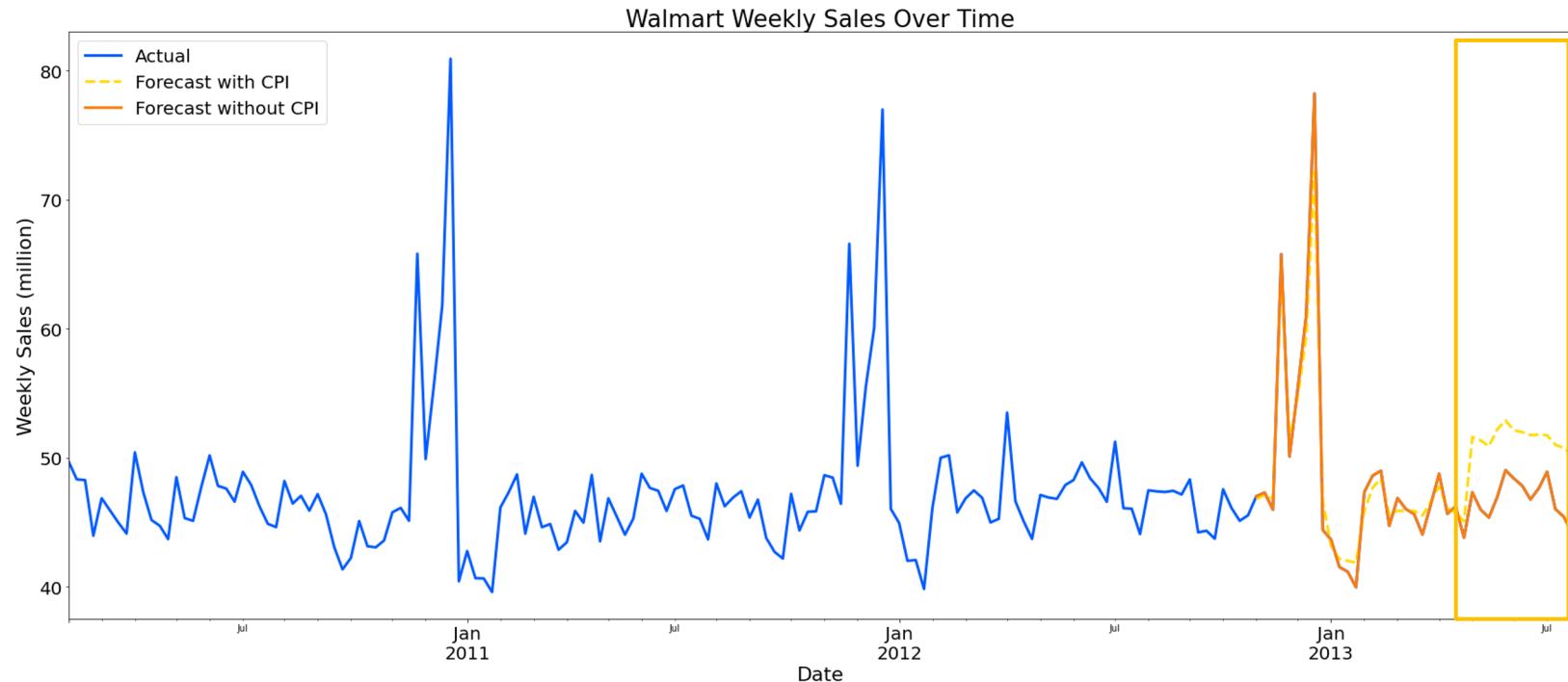


Final prediction and conclusion

Final Prediction



CPI is tricky! It is “0” in May-Jul 2013 so prediction is distorted. Thus, we removed it for the final prediction.



Final Prediction and Conclusion

Project introduction

EDA

Time series prediction

Data preprocessing

Model building

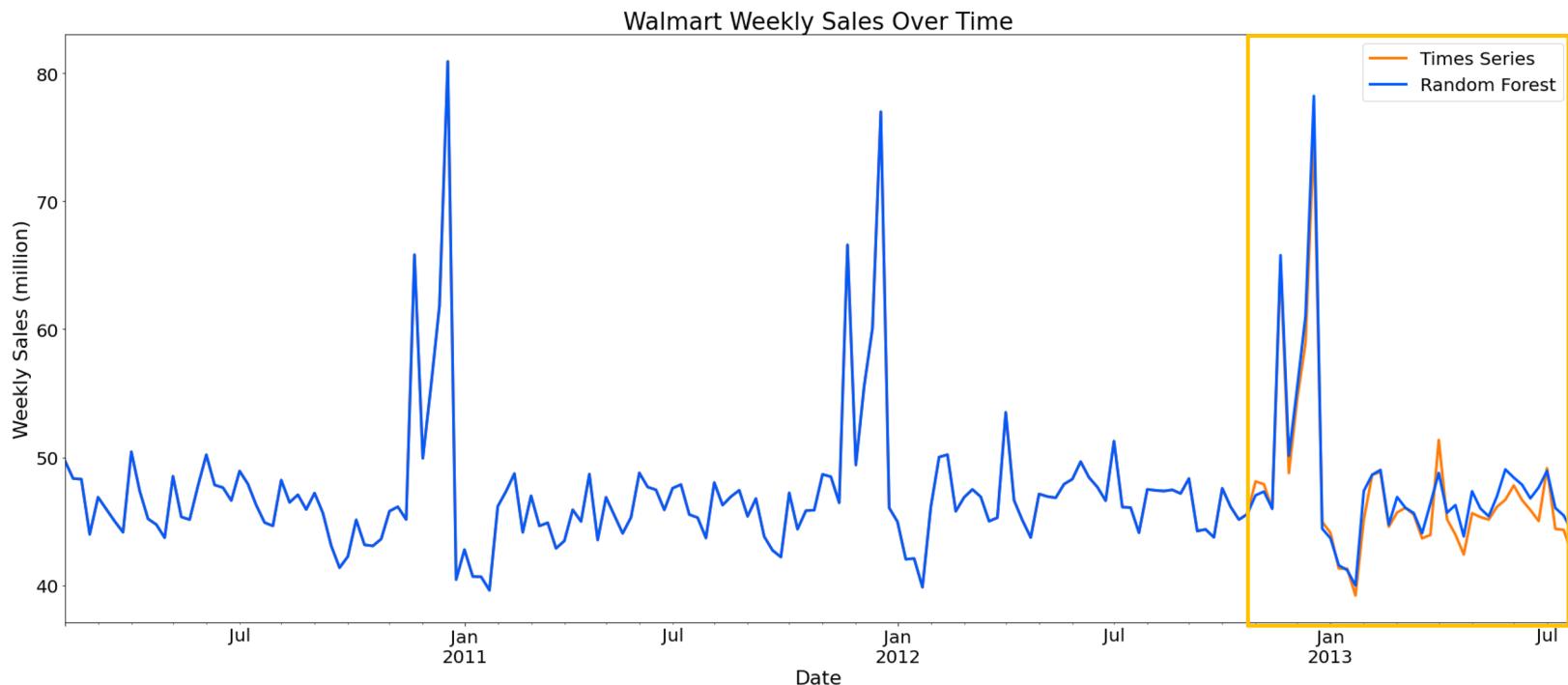
Best model selection

Final prediction and conclusion

Final Prediction / VS. Times Series Result



Similar trend is observed from SARIMA & Random Forest Result.



Final Prediction and Conclusion

-  Project introduction
-  EDA
-  Time series prediction
-  Data preprocessing
-  Model building
-  Best model selection
-  Final prediction and conclusion

Final Prediction / VS. Times Series Result / Conclusion



Threat



Festive Dept Sales Decline



Opportunities



Easter Holiday Potential



Department Extension



Type C Store Expansion