Capstone 2 Walmart sales prediction

Yan Gao

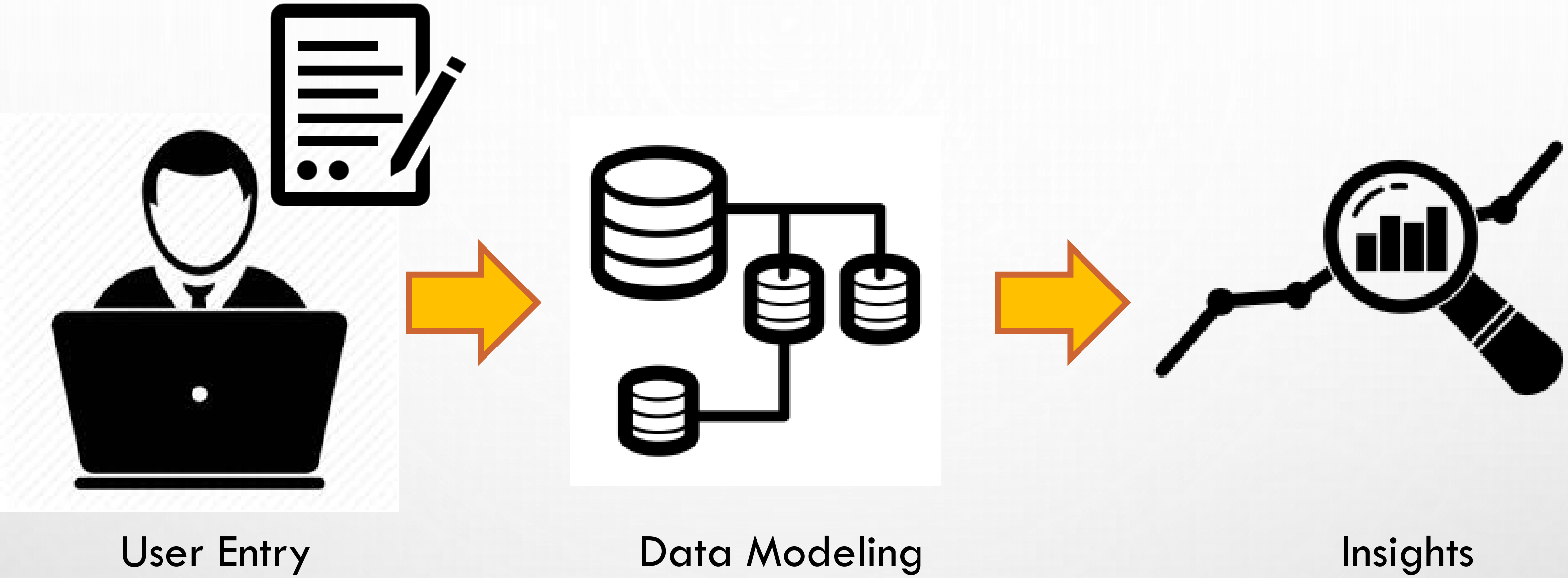
Problem Statement

Utilizing advanced machine learning techniques to forecast Walmart sales trend at individual store and department level with consideration of seasonality, economic indicators and promotion events

Benefits:

- Use as a guidance to set sales target and properly arrange personnel schedule
- Supply chain to properly manage the inventory and allocate their resources
- help with strategic planning to maximize profitability and revenue

Goal



- Develop a robust model to predict weekly sales across various Walmart stores and departments
- Create a predictive framework that is both accurate and scalable, ultimately aiding Walmart in strategic decision-making

Data Sources

kaggle

Recruitment Prediction Competition

Walmart Recruiting - Store Sales Forecasting

Use historical markdown data to predict store sales

Dataset Description

- Weekly Sales

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

- Features

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

- Time span between 2010-02-05 and 2012-10-26 on weekly basis
- Three dataset:
  - Weekly Sales
  - Features(promotion events, economic info, temperature, holidays)
  - Store info (type, size)
- Missing values in Markdowns:50%
- Missing values in CPI and unemployment:7%
- Total number of time series:3331
- Total number of time series with time gap: 695, ~21% of all the time series

Candidate Features

Variable	Description
Temperature	Local temperature
Fuel Price	Fuel Price
CPI	consumer price index, indicator of inflatation
Unemployment	unemployment rate
Year	
Month	
Season	
Week	
day	

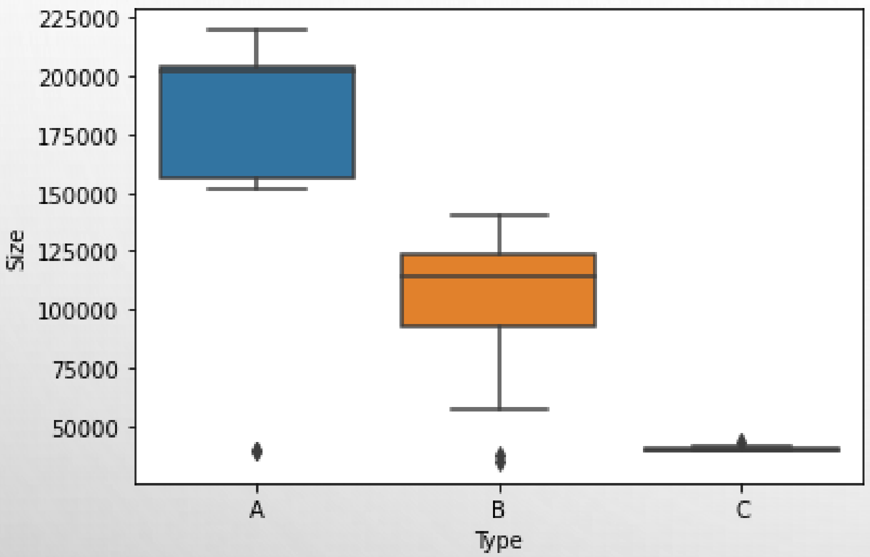
Variable	Description
Holiday Name	Birthday of Martin Luther King Jr Christmas Day Columbus Day Independence Day Labor Day Memorial Day New Year s Day Superbowl Thanksgiving Day Veterans Day Washington s Birthday non Holiday

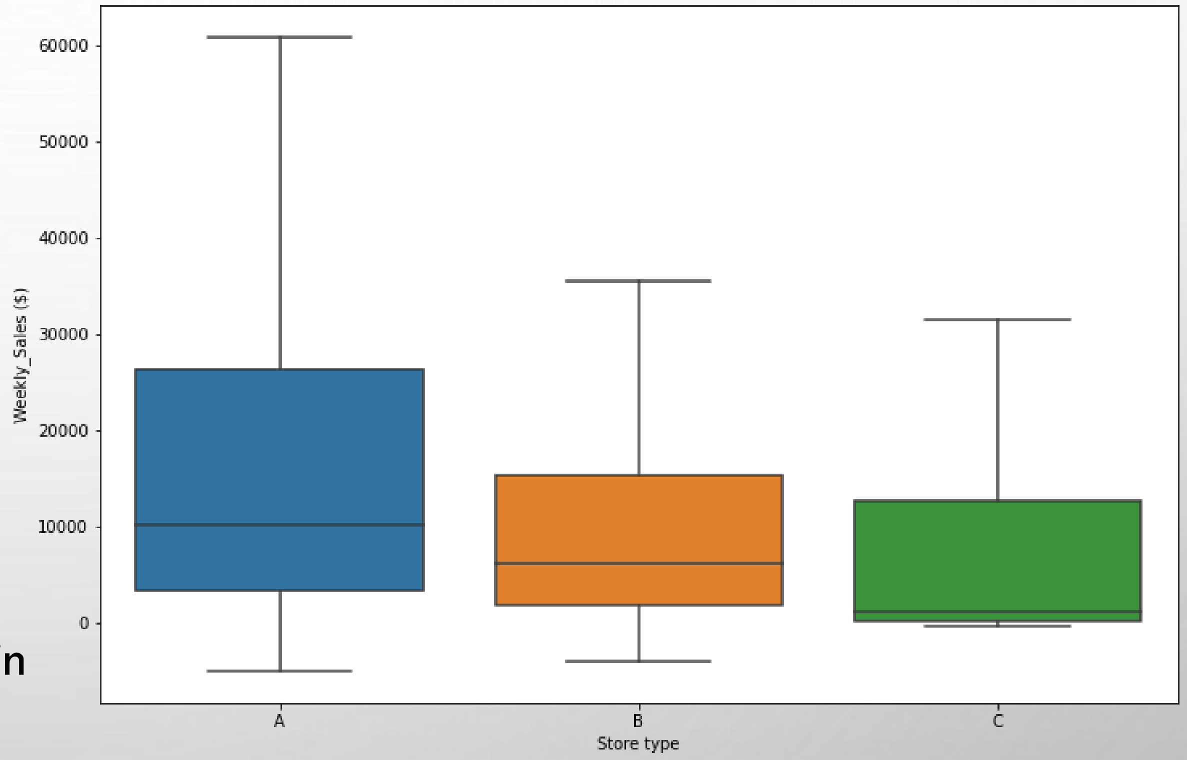
Data Wrangling

- Filling missing values for CPI, unemployment with rolling window average
- Fill markdown values before 2011-11-11 with zeros
- Remove negative weekly sales values
- Select time series in weekly_sales dataset without time gaps
- Merge three datasets (Store, Feature, Weekly_Sales)
- Create new features like year, season, week, etc and holiday names

Data Analysis

- Store Type Analysis

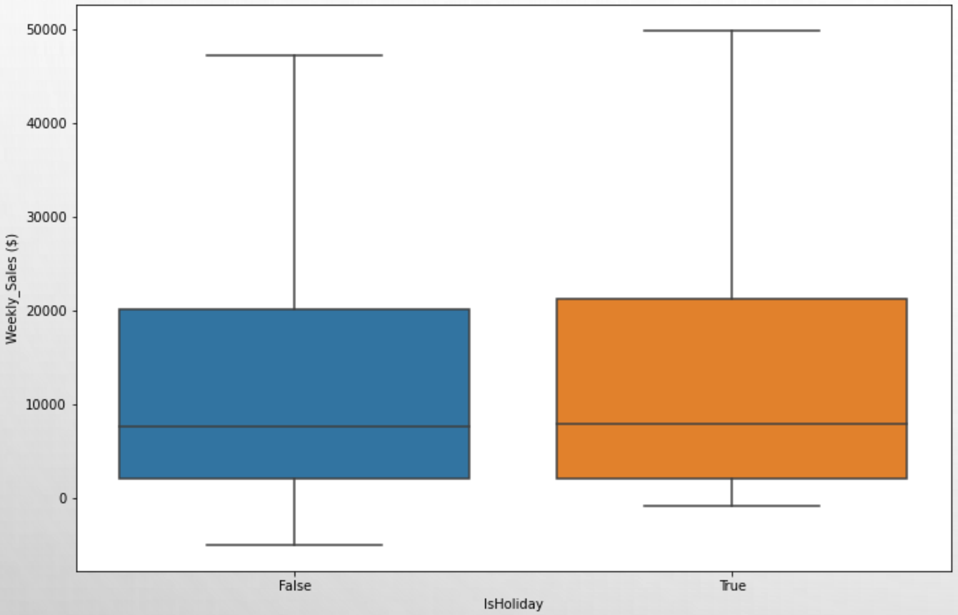


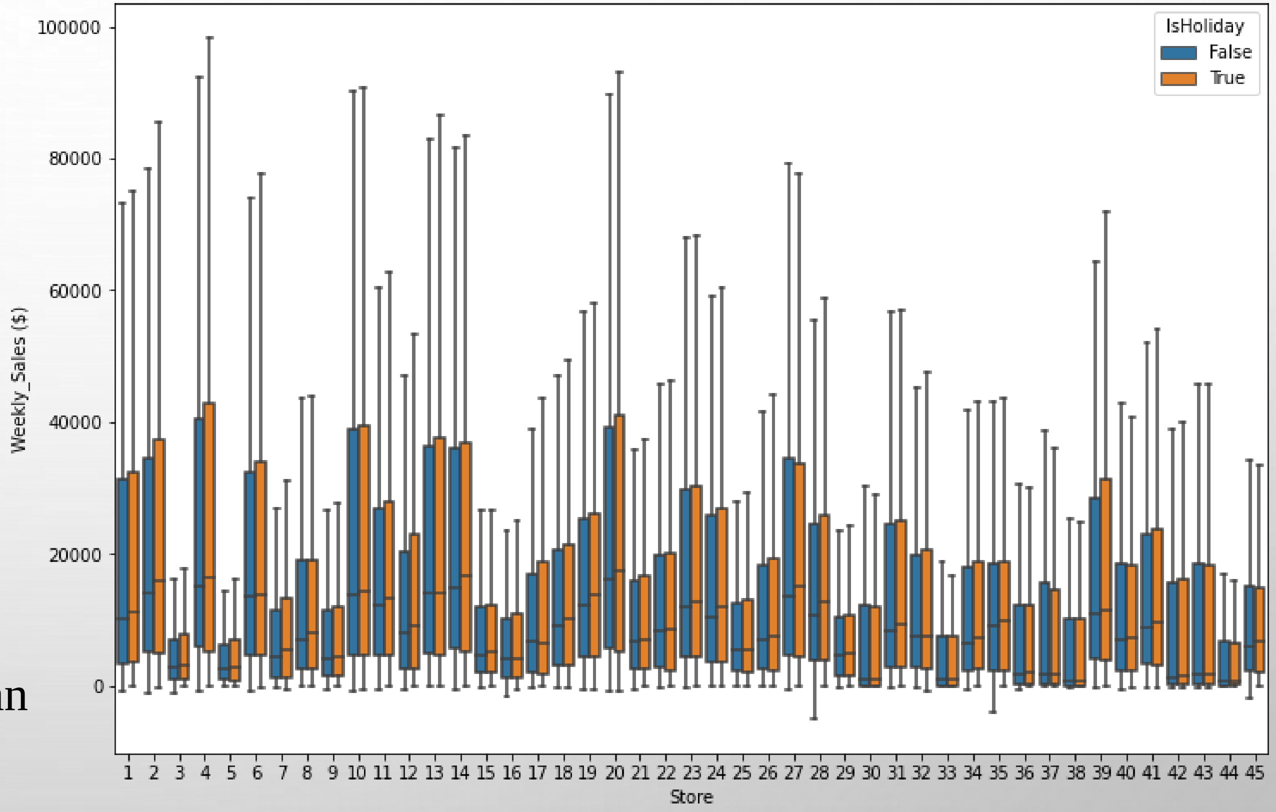


- Half of the stores are type A with largest size in average
- Type A tends to have highest average weekly sales value

Data Analysis

- Holiday Effect

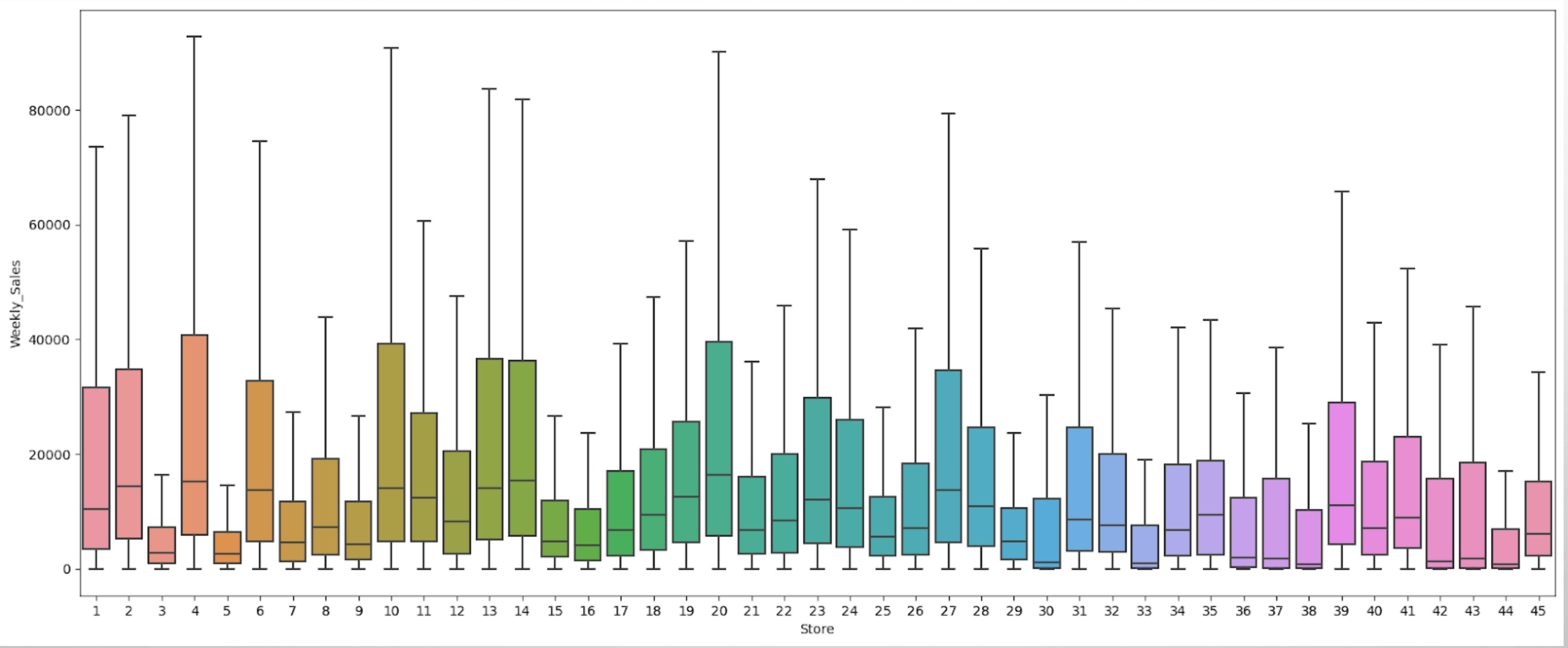




- Holiday weekly sales is always slightly higher than non-holidays

Data Analysis

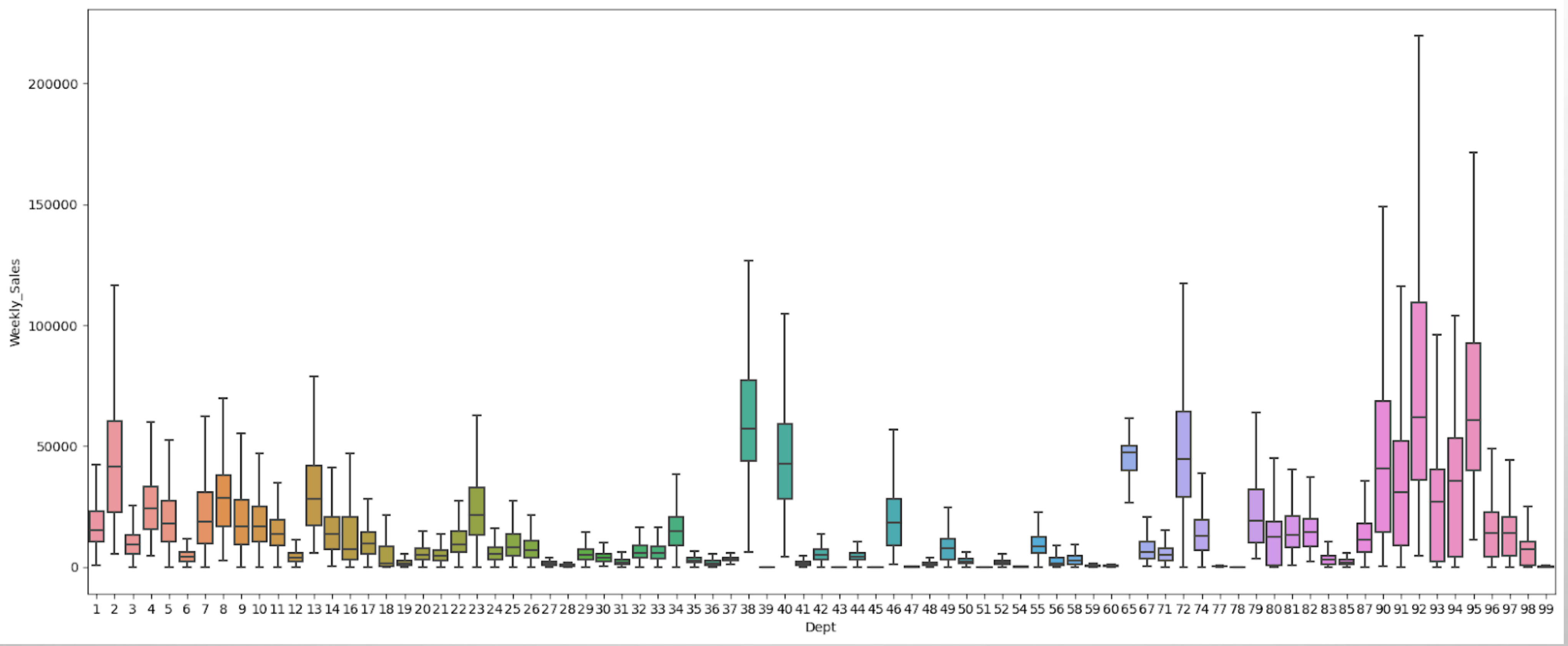
- Weekly Sales Distribution for Each Store



- Stores 20, 14, and 4 rank top 3 on average weekly sales
- Stores 30, 33, 38, and 44 exhibit the lowest average weekly sales

Data Analysis

- Weekly Sales Distribution for Each Department



– Department 92 show highest average weekly sales

Data Analysis

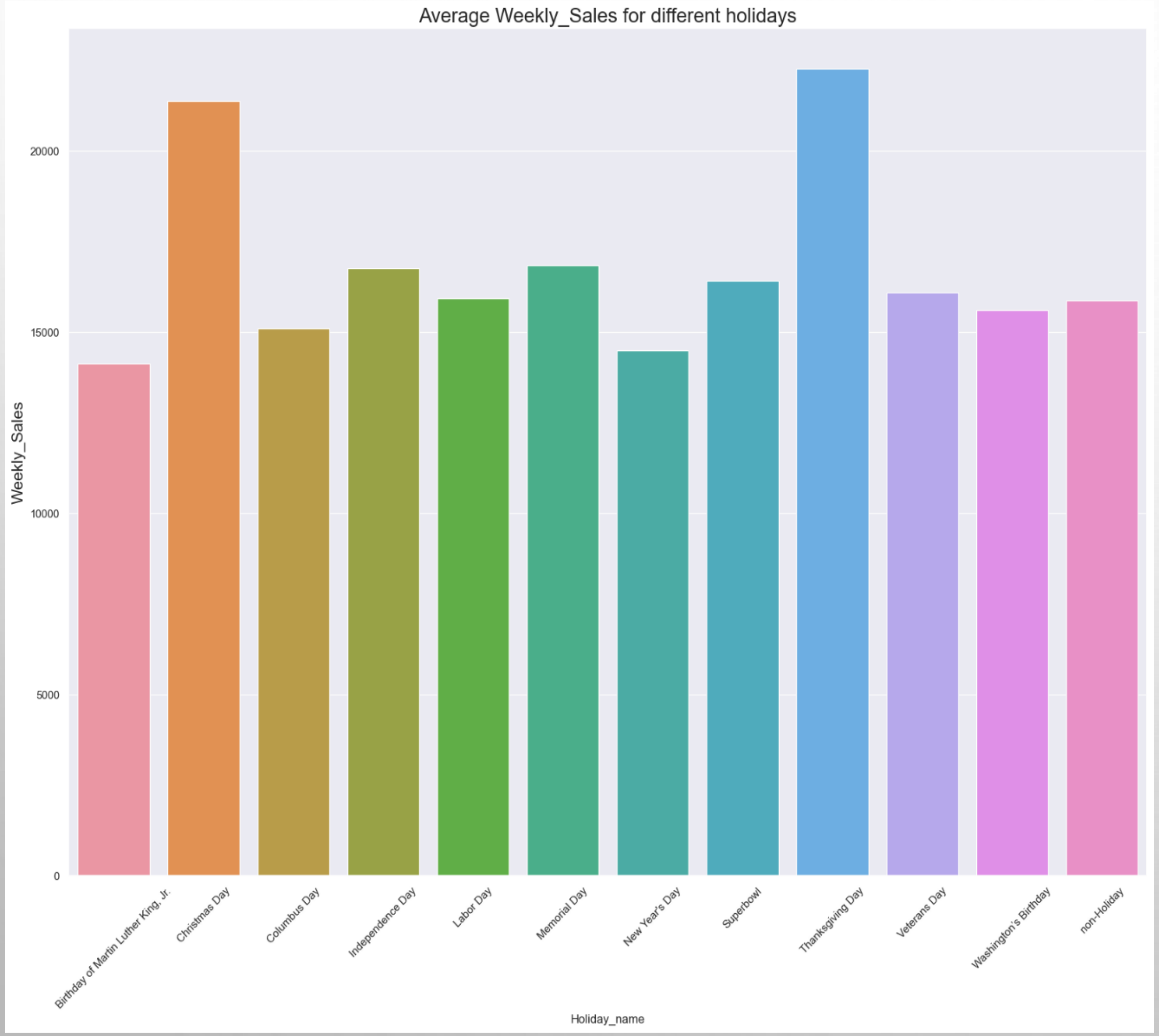
- Periodic Trend of Weekly Sales

– The end of the year consistently brings a surge in sales irrespective of the year, this could be a critical window for revenue generation

– The overall sales trend remains largely stable from year to year, allowing for predictable business
planning and resource allocation

Data Analysis

- Holiday Effect on Averaged Weekly Sales



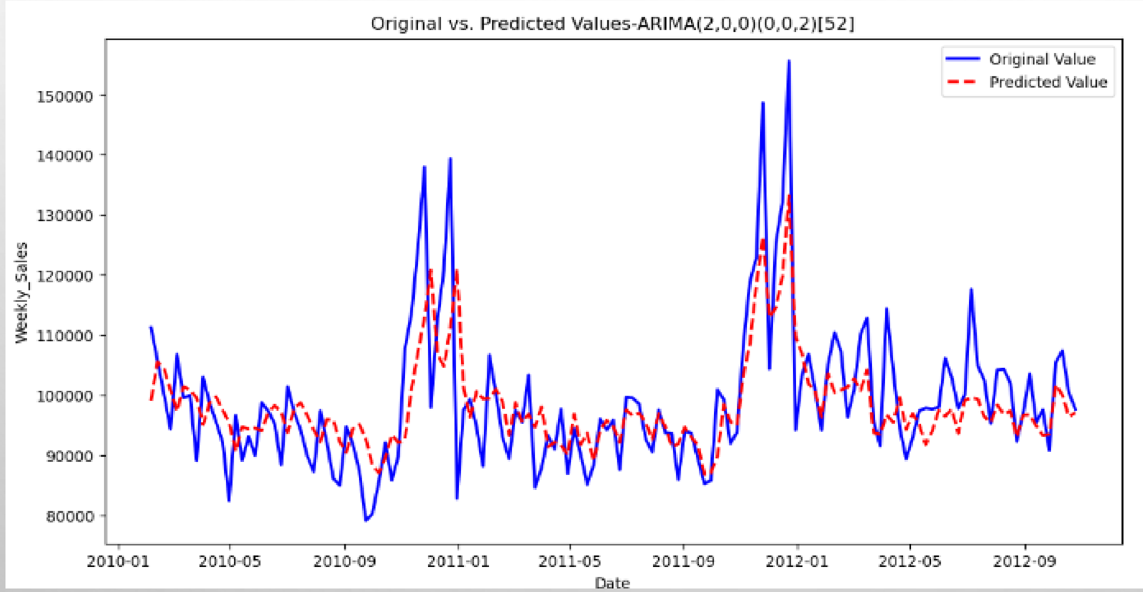
– The peak in average weekly sales consistently occurs during the weeks of Thanksgiving and Christmas

– These periods offer a unique set of opportunities for revenue growth and customer acquisition, allowing us to allocate resources more efficiently and capitalize on these peak sales window

Baseline Model

- **ARIMA algorithm**

detailed model description	MAPE for training	MAPE for testing
ARIMA(1,1,1)	0.089	0.079
ARIMA(1,0,4)_GridSearch	0.065	0.05268
AUTO_ARIMA((2,0,0)(0,0,2)[52]	0.063	0.04824

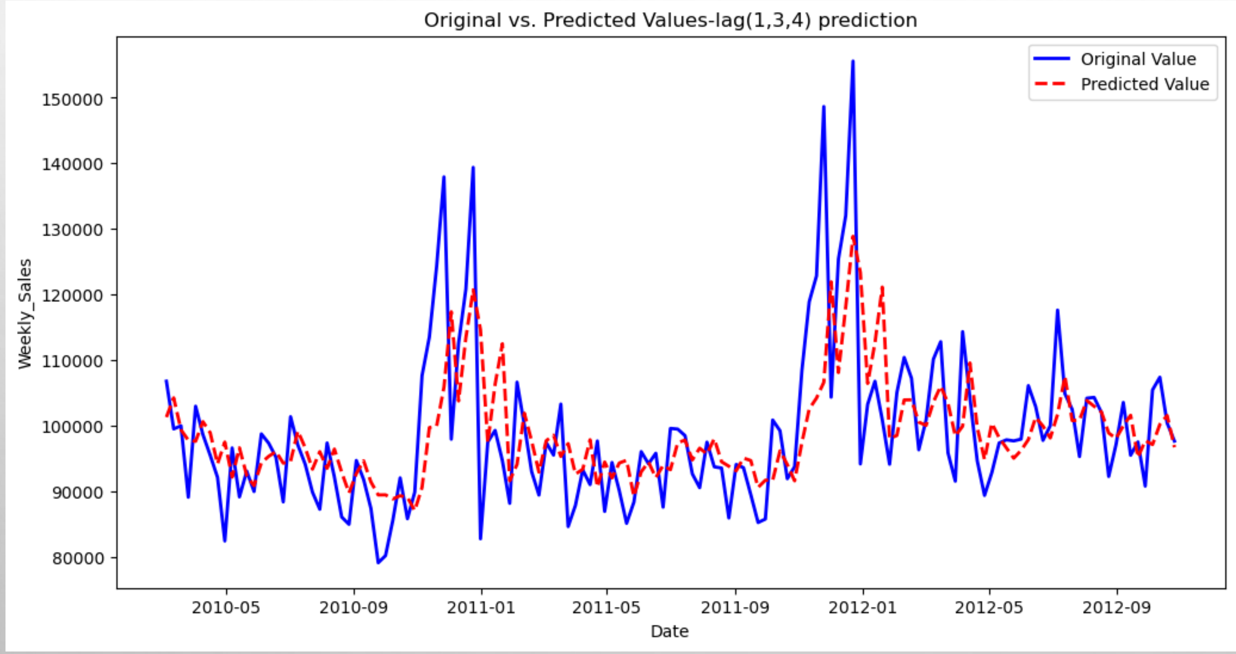


A representative time series was selected as the initial basis for selecting algorithms

Baseline Model

- **Linear regression**

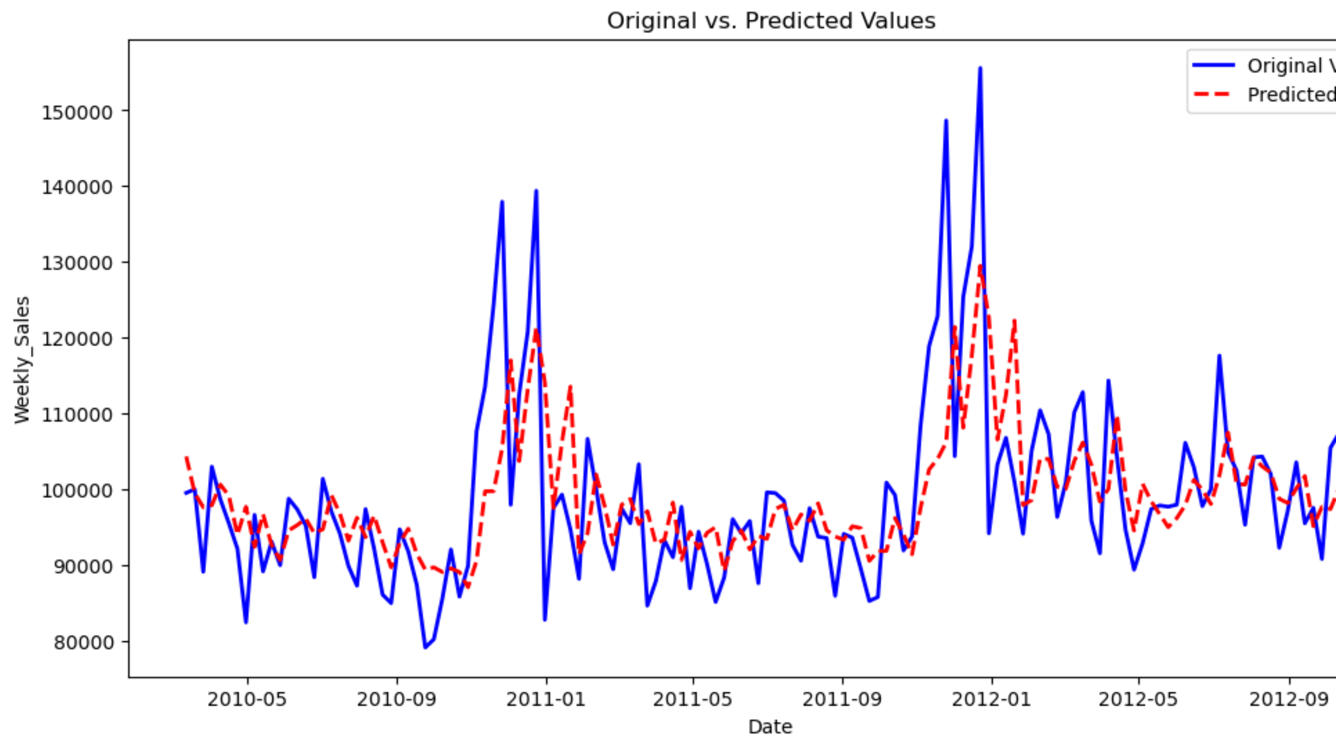
detailed model description	R^2 for training	R^2 for testing	MAPE for training	MAPE for testing
linear regression with lag_1	0.27	-0.1228	0.08	0.0503
linear regression with lag_1-5	0.42	0.0879	0.07	0.0495
linear regression with lag_1-7	0.43	0.1927	0.07	0.0475
linear regression with lag_(1,3,4)	0.37	0.2078	0.073	0.0444



Baseline Model

- **Linear regression with exogenous features**

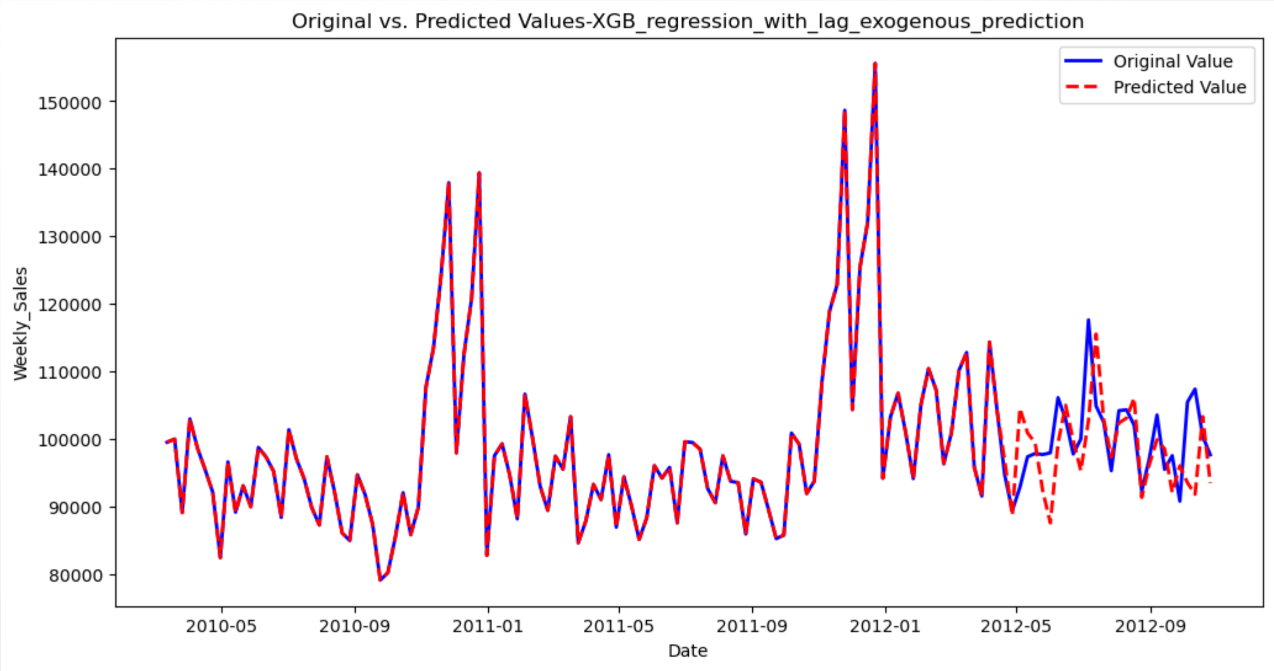
detailed model description	R^2 for testing	MAPE / mean MAPE for training	Variance of MAPE	MAPE / mean MAPE for testing	Variance of MAPE
linear regression with exogenous features	-0.74	0.06		0.066	
linear regression with exogenous features_5-fold validation	-1349	0.05	0.007	1.1	1.89
linear regression with exogenous_walk_froward validation	-0.74	0.06		0.066	0.04
linear regression with exogenous-grid search	-1.36	0.074		0.067	
linear regression with exogenous and lagged values	-0.75	0.06		0.067	
linear regression with exogenous and lagged values-grid search	-0.08	-0.06		0.04	



Baseline Model

- **<u>Tree based regression</u>**

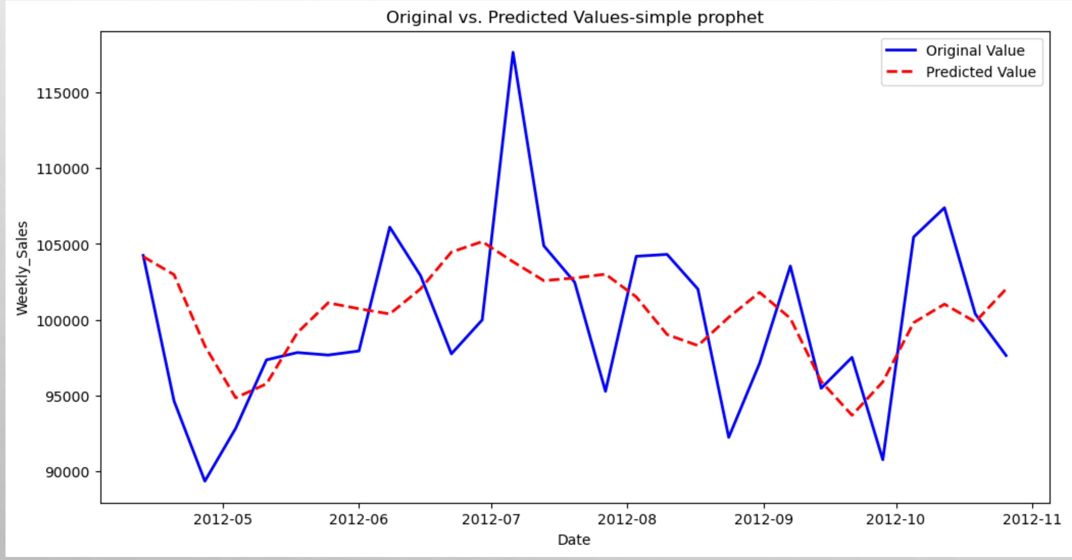
detailed model description	R^2 for testing	MAPE / mean MAPE for training	MAPE/ mean MAPE for testing	Variance of MAPE
Random forest regression with exogenous features	-2.2	0.04	0.08	
Random forest regression with walk_froward_validation	-2.2	0.04	0.08	0.07
Random forest regression with _lagged_exogenous_features	-3.73	0.03	0.1	
Random forest regression with lagged_exogenous_features_walk_froward_validation	-3.73	0.03	0.1	0.08
Simple XG_Boost		0	0.049	
XG_Boost with grid search		0	0.094	
XG_Boost with selected features		0	0.069	

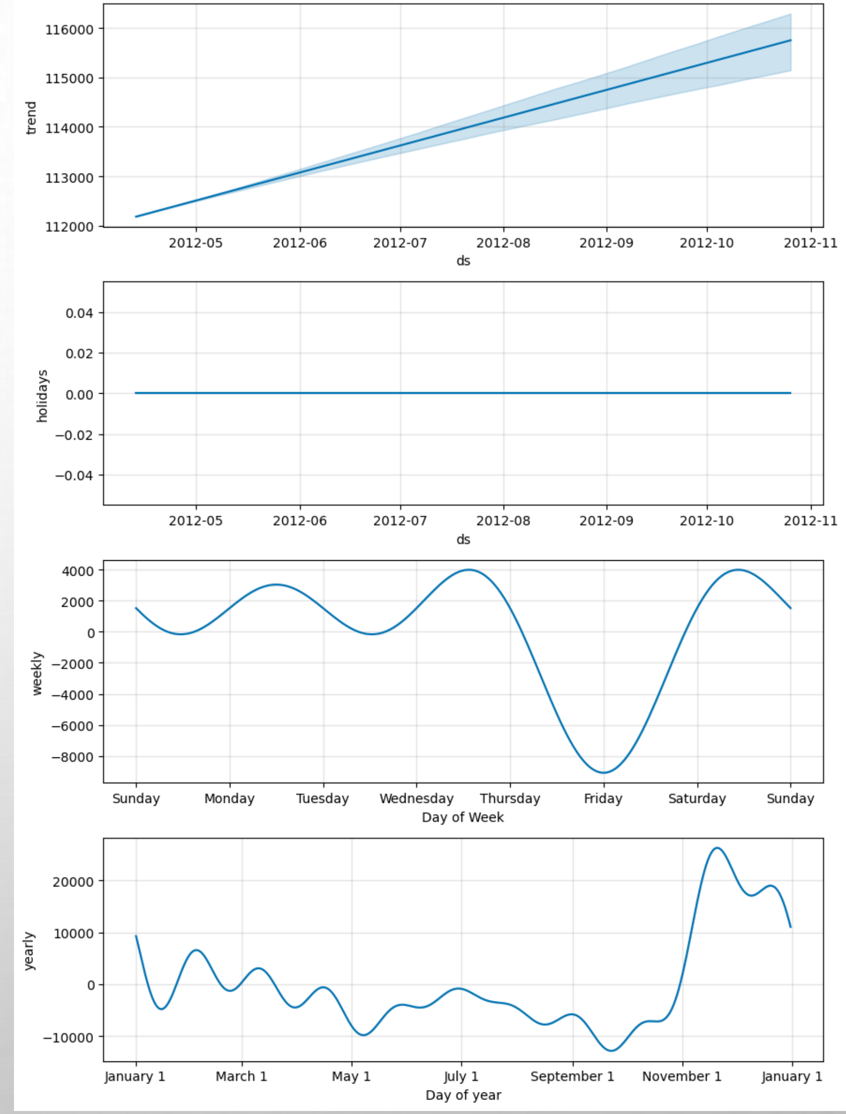


Baseline Model

- **Prophet model**

detailed model description	R^2 for testing	MAPE for test
simple_prophet	0.16	0.04
added_holiday_prophet	0.17	0.042
added_exgeneous_holiday_prophet	-0.73	0.06
lagged_exgeneous_holiday_prophet_grid_search	0.14	0.036





Extended Model

- **Extended model summary with auto-ML**

Summary of MAPE for selected model for each store/dept

Index	model_name	val_MAPE
(1, 1)	Ensemble	0.036588491
(1, 2)	Ensemble	0.022880428
(1, 3)	Ensemble	0.102711383
(1, 4)	Ensemble	0.034287863
(1, 5)	Ensemble	0.089661202
(1, 7)	Ensemble	0.065940652
(1, 8)	Ensemble	0.032639058
(1, 9)	Ensemble	0.091993969
(1, 10)	Ensemble	0.060693532
(1, 11)	Ensemble	0.11546191
(1, 12)	Ensemble	0.069269739
(1, 13)	Ensemble	0.024706026
(1, 14)	Ensemble	0.088031065
(1, 16)	Ensemble	0.079380506
(1, 17)	Ensemble	0.061286704
(1, 19)	Ensemble	0.167445631
(1, 20)	Ensemble	0.113268197
(1, 21)	Ensemble	0.065800082
(1, 22)	Ensemble	0.086555572
(1, 23)	Ensemble	0.086386558
(1, 24)	Ensemble	0.123138657

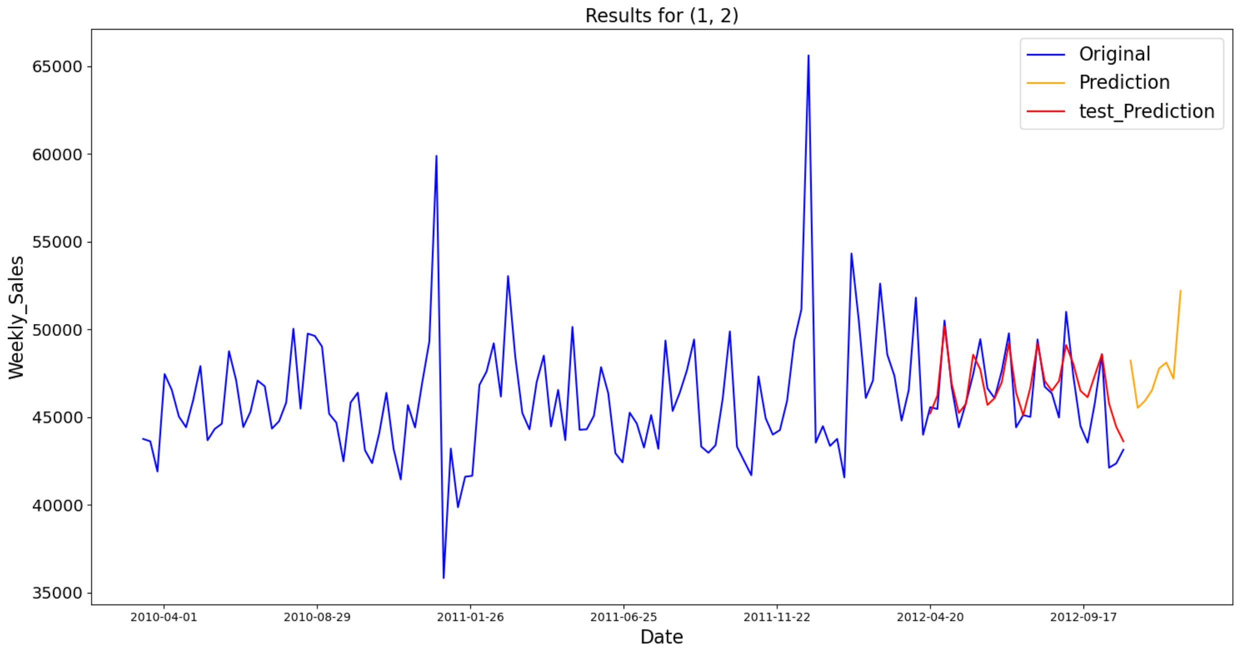
Index	model_name	val_MAPE
(1, 25)	Ensemble	0.0875075
(1, 26)	Ensemble	0.088320692
(1, 27)	Ensemble	0.123534503
(1, 28)	Ensemble	0.130092019
(1, 29)	Ensemble	0.065723621
(1, 30)	Ensemble	0.128446197
(1, 31)	Ensemble	0.190918333
(1, 32)	Ensemble	0.135011641
(1, 33)	Ensemble	0.125898524
(1, 34)	Ensemble	0.059239158
(1, 35)	Ensemble	0.139254297
(1, 36)	Ensemble	0.333464833
(1, 37)	Ensemble	0.067644359
(1, 38)	Ensemble	0.056337
(1, 40)	Ensemble	0.029778584
(1, 41)	Ensemble	0.165471755
(1, 42)	Ensemble	0.075061359
(1, 44)	Ensemble	0.075042024
(1, 46)	Ensemble	0.042453815
(1, 49)	Ensemble	0.121323123
(1, 52)	Ensemble	0.133201003
(1, 55)	Ensemble	0.091233782

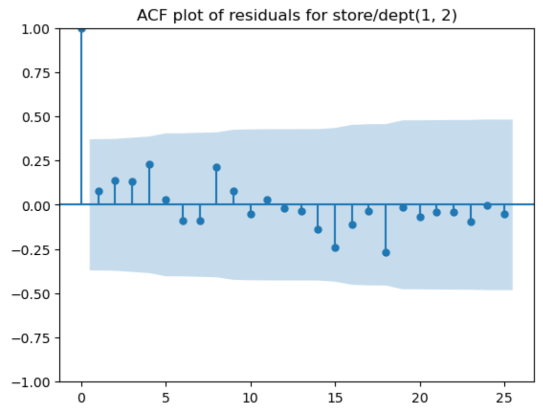
Index	model_name	val_MAPE
(1, 56)	Ensemble	0.107605313
(1, 58)	Ensemble	0.479122888
(1, 59)	Ensemble	0.159957116
(1, 60)	Ensemble	0.087710412
(1, 67)	Ensemble	0.082985515
(1, 71)	Ensemble	0.208093623
(1, 72)	Ensemble	0.080136841

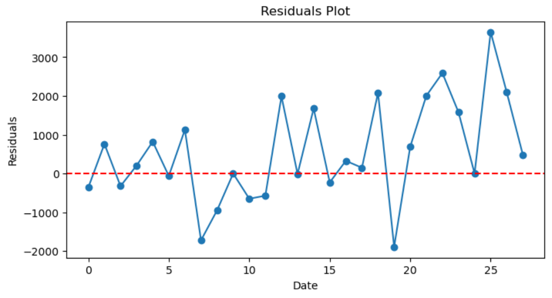
50 time series were selected to build extended model for individual time series

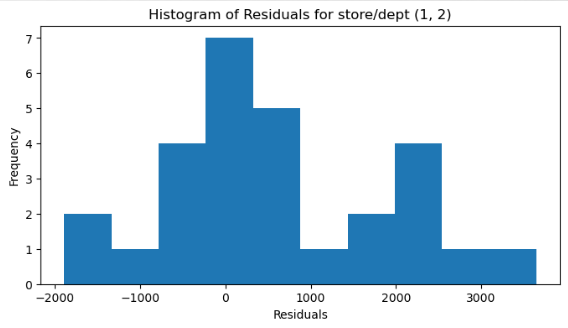
Extended Model

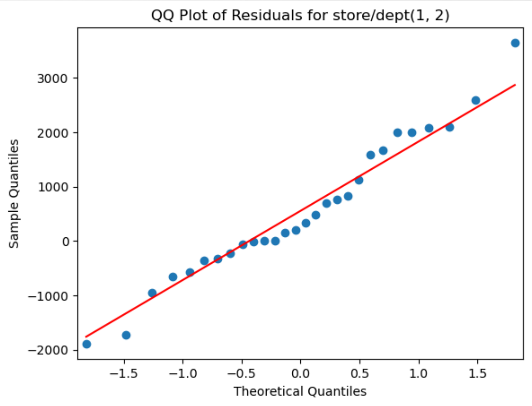
- **<u>Examples of store/dept with lowest MAPE</u>**







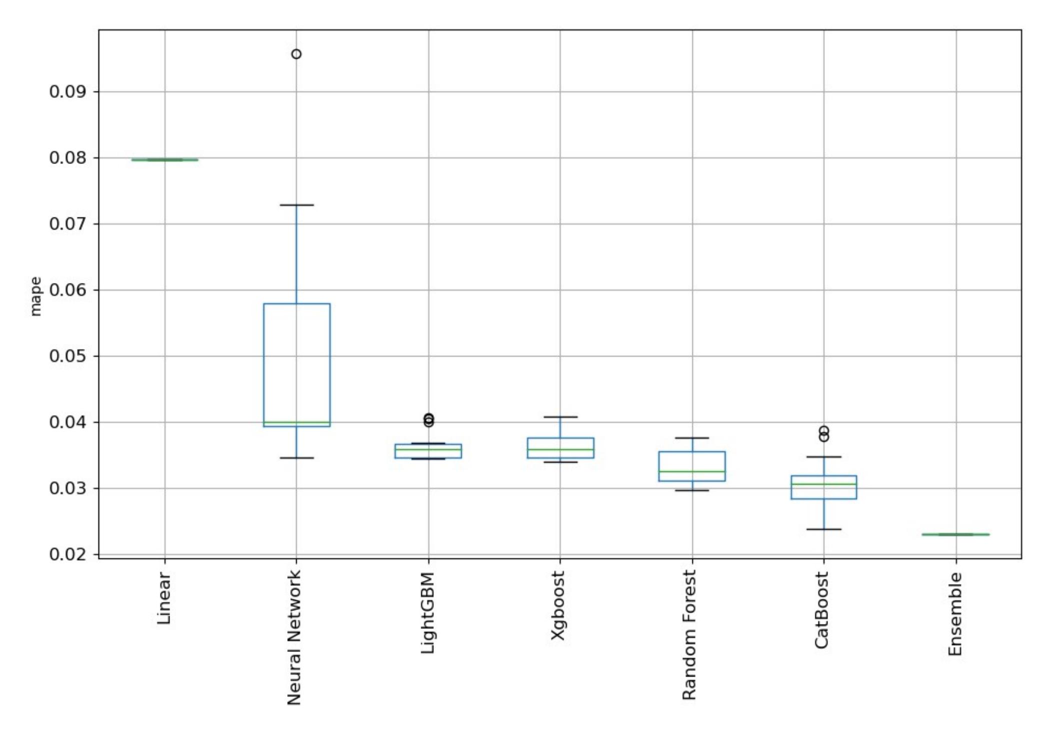




Extended Model

- **Examples of store/dept with lowest MAPE**

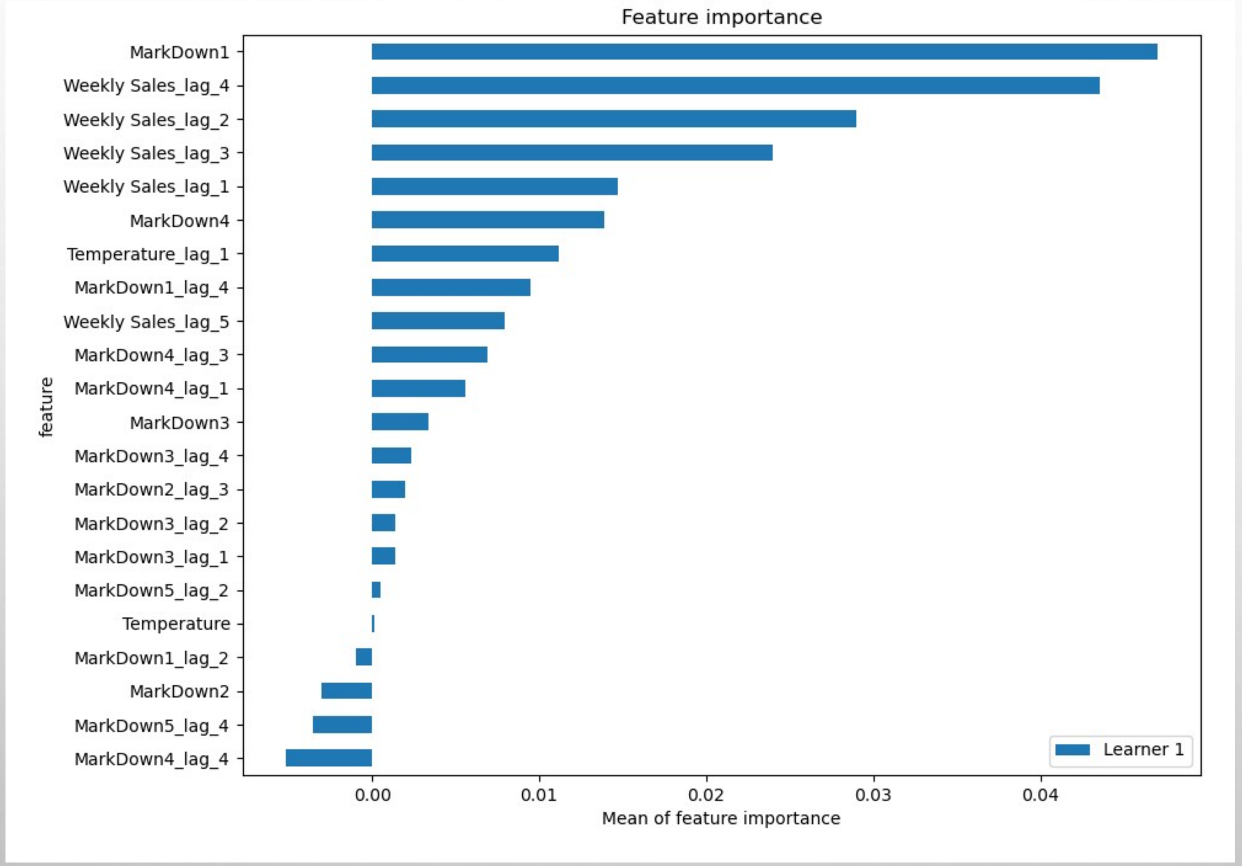
Boxplot of MAPE values for different models used in autoML



Extended Model

- **<u>Examples of store/dept with lowest MAPE</u>**

Feature importance provided by ensemble model



Summary and Recommendations

Sales Patterns & Trends:

- Clear seasonal trends were identified in the sales data, with consistent peaks during the end of the year, especially during Thanksgiving and Christmas.
- Store 14 Department 92 recorded the highest total sales, accumulating a revenue of $26,101,497.
- Stores 30, 33, 38, and 44 had the lowest average weekly sales.

Recommendations:

- Given the pronounced seasonal and holiday effects, strategic resource allocation and promotions during peak sales windows(Thanksgiving and Christmas) can maximize revenue.
- Leveraging the insights from the different models for individual store/dept can aid in more accurate sales forecasting, allowing for better inventory management and resource optimization.
- SHAP analysis helps understand global importance of different features as well as the reliance of weekly_sales on different features.

Future Work

- Clustering analysis before doing individual time series modeling
  - reduce complexity of the data
  - reveal hidden structure or similar sales pattern
  - save resources for decision making and strategy development as well as prediction accuracy
- Alternative package for autoML such as AutoTS
- Establish web-based model system with integrated workflow could facilitate model update with refreshed dataset, therefore improving model accuracy