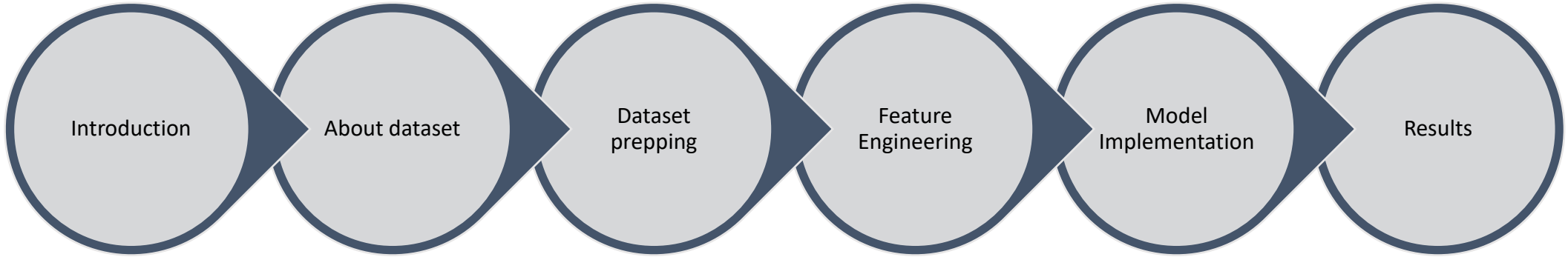
Capstone Project: ML-Regression

Rossmann Sales Prediction

Presented by,
Vivek CP.
Data Science Trainee, at Alma Better

Flow Of The Presentation



Introduction

Rossmann

- Drug Store Chain.

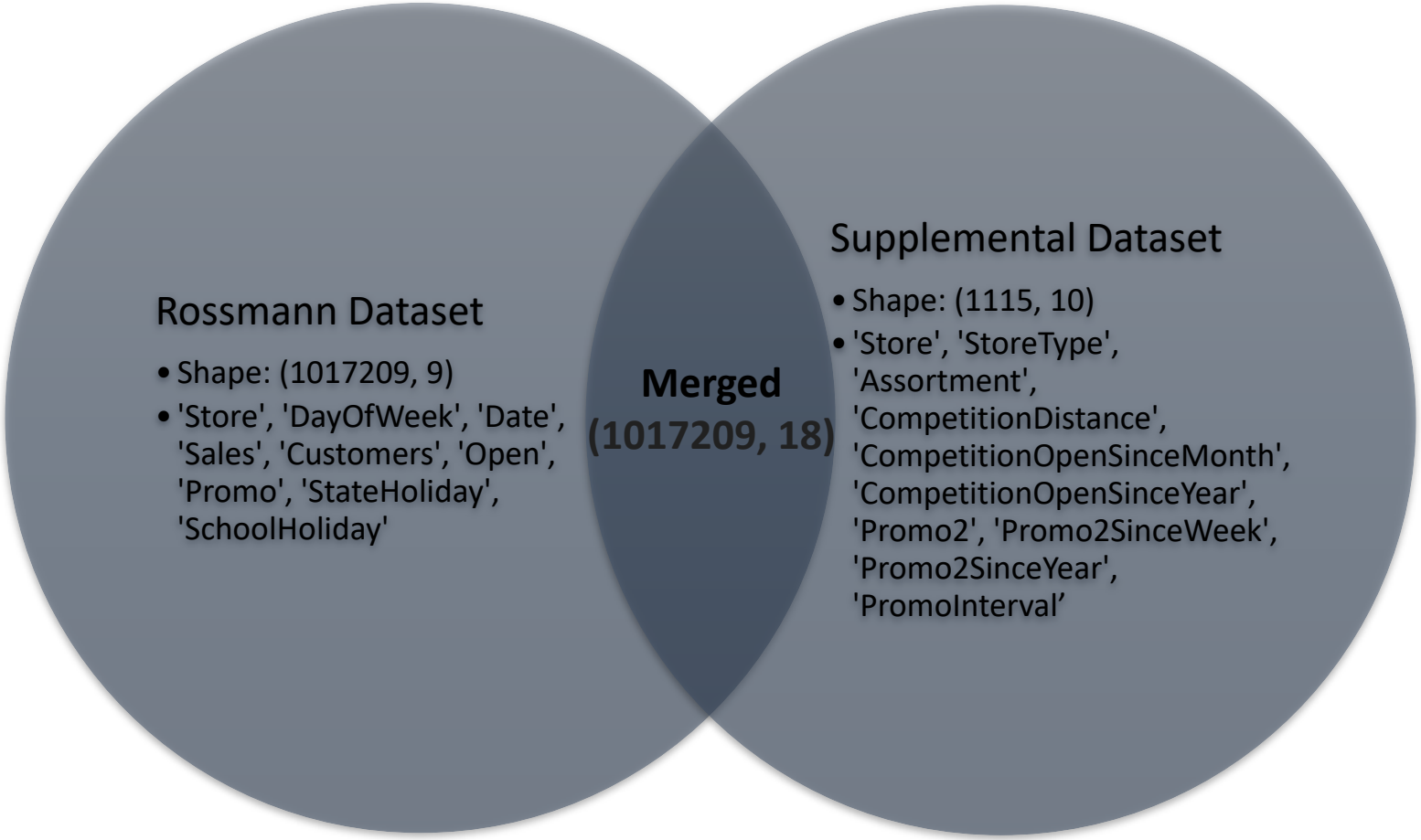
Objective

- Predict sales for next 6 weeks.

Methodology

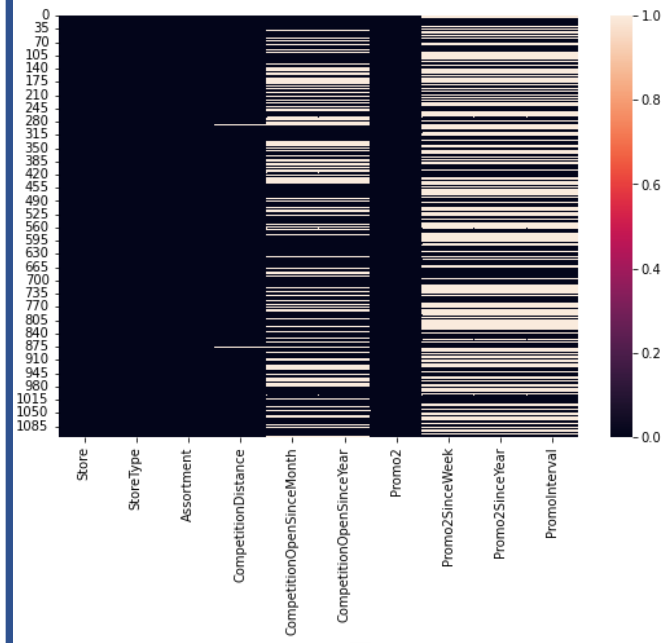
- Linear Regression
- Random Forest Regression
- Time Series Analysis.

Dataset


Rossmann Dataset
• Shape: (1017209, 9)
• 'Store', 'DayOfWeek', 'Date', 'Sales', 'Customers', 'Open', 'Promo', 'StateHoliday', 'SchoolHoliday'
Merged
(1017209, 18)
Supplemental Dataset
• Shape: (1115, 10)
• 'Store', 'StoreType', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2', 'Promo2SinceWeek', 'Promo2SinceYear', 'PromoInterval'

Dataset Clean-up

NaN Value & Outlier Clean-up



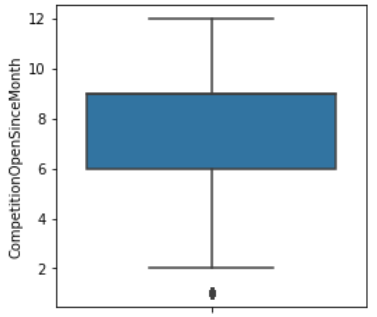
NaN value:

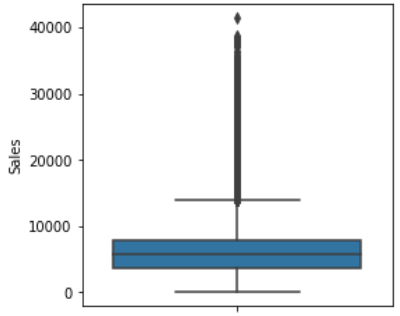
- CompetitionDistance- Median
- CompetitionOpenSinceMonth- Mode
- CompetitionOpenSinceYear- Mode
- PromoInterval- 0/'None'
- Promo2SinceYear- 0/'None'
- Promo2SinceWeek- 0/'None'

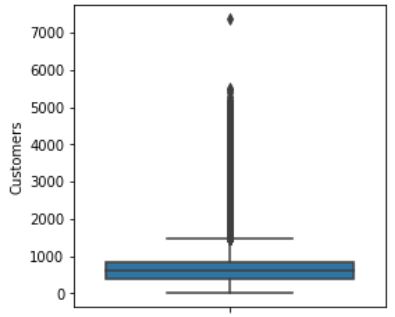
IQR method

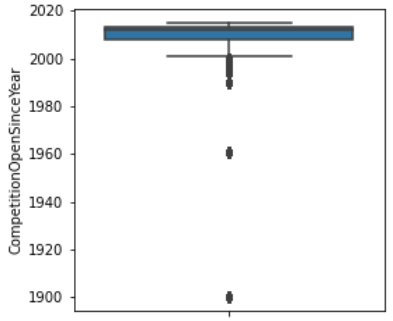
Upper Limit: Q3+1.5IQR

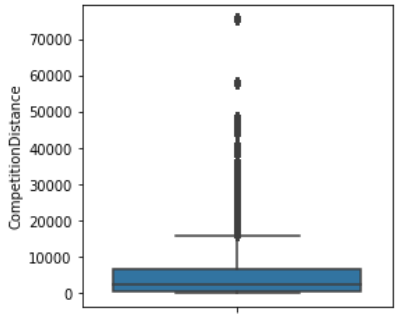
Lower Limit: Q1-1.5IQR







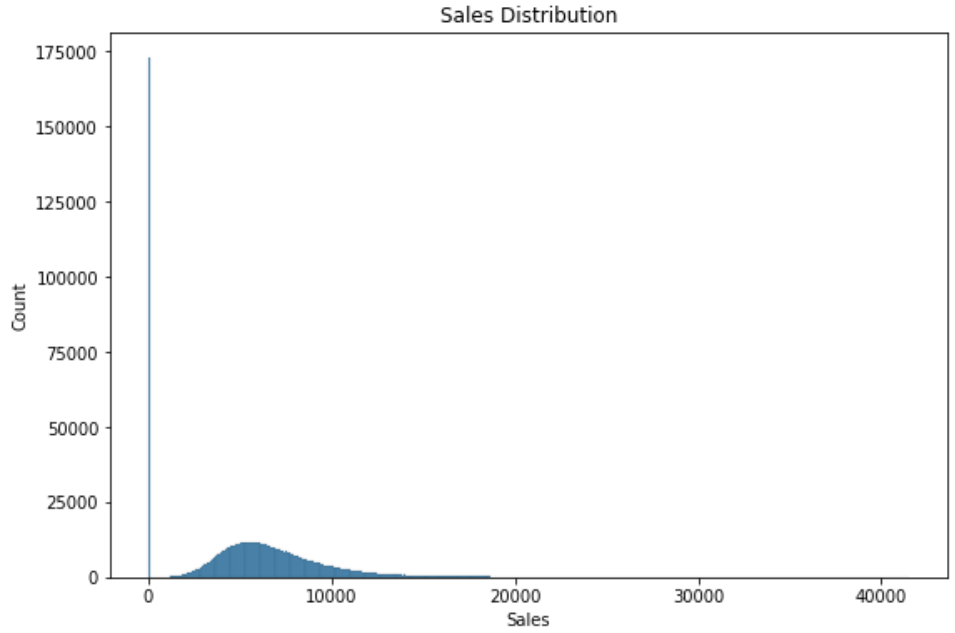


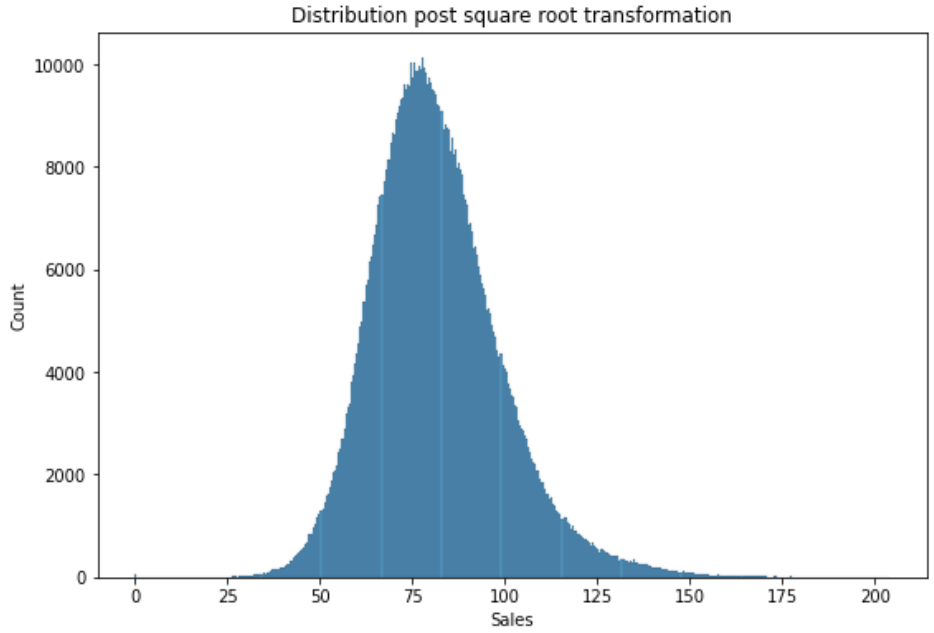


Feature Engineering

Target Feature Conditioning

Normalising Target Feature





Categorical Feature Encoding

One Hot Encoding

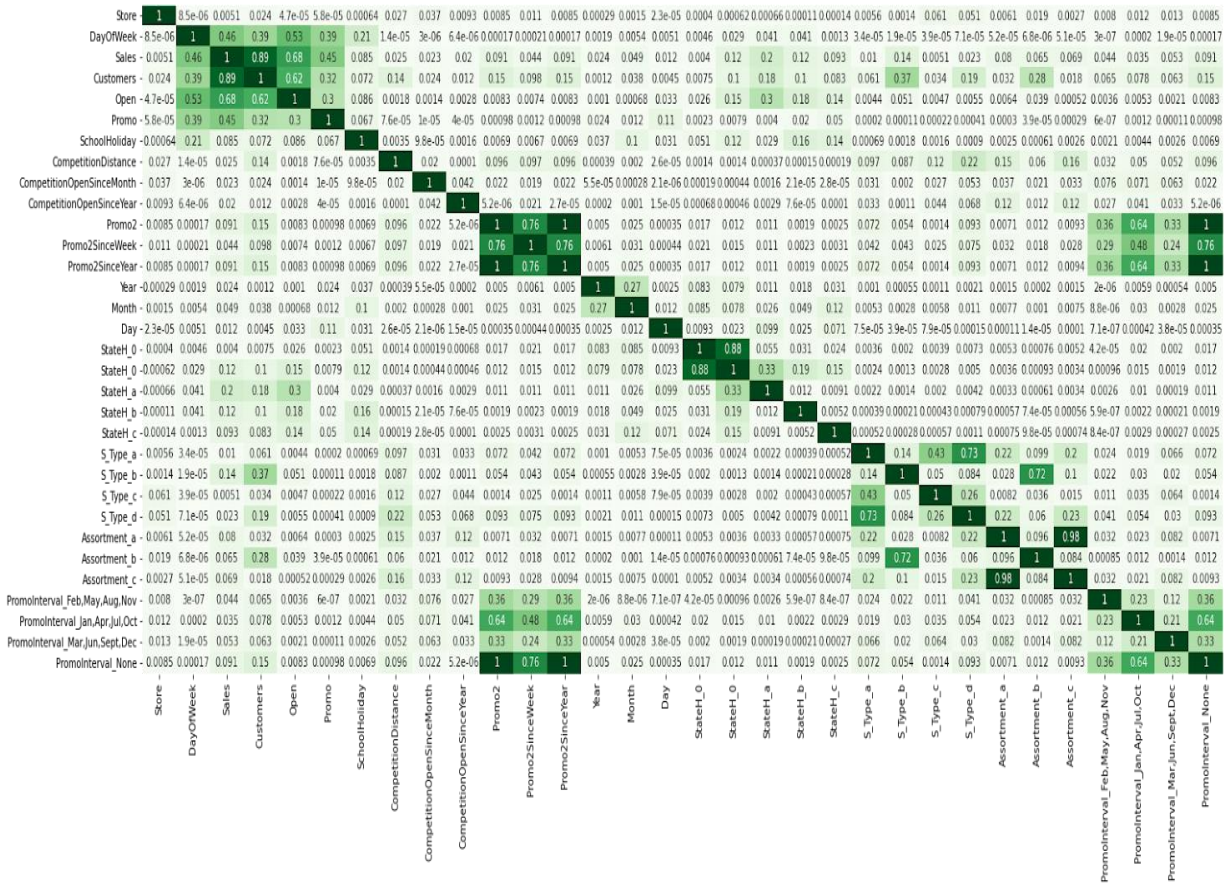
'StateHoliday','StoreType','Assortment','PromoInterval'

StateH_0	StateH_0	StateH_a	StateH_b	StateH_c	S_Type_a	S_Type_b	S_Type_c	S_Type_d	Assortment_a	Assortment_b	Assortment_c	PromoInterval_Feb,May,Aug,Nov	I
0	1	0	0	0	0	0	1	0	1	0	0	0	
0	1	0	0	0	0	0	1	0	1	0	0	0	
0	1	0	0	0	0	0	1	0	1	0	0	0	
0	1	0	0	0	0	0	1	0	1	0	0	0	
0	1	0	0	0	0	0	1	0	1	0	0	0	

Feature Selection

Multicollinearity

'Store','Assortment_b','S_Type_b','S_type_c','StateH_0','S_Type_c','Promo2','CompetitionDistance','CompetitionOpenSinceMonth','CompetitionOpenSinceYear','Promo2SinceWeek','Day','S_Type_d','PromoInterval_Jan,Apr,Jul,Oct','Sales','Promo2SinceYear','Promointerval_None','PromoInterval_Feb,May,Aug,Nov','PromoInterval_Jan,Apr,Jul,Oct','PromoInterval_Mar,Jun,Sept,Dec','S_Type_a','Date', ‘assortment c’.



Feature Selection

Variance Inflation Factor

Variable	VIF
DayOfWeek	8.553595
Customers	4.969602
Open	13.777009
Promo	1.995819
SchoolHoliday	1.346284
Year	32.935062
Month	4.212767
StateH_a	1.225258
StateH_b	1.118830
StateH_c	1.075896
Assortment_a	2.123727
PromoInterval_None	2.072111

→

Variables	VIF
DayOfWeek	3.210494
Customers	4.969416
Open	7.482442
Promo	1.902145
SchoolHoliday	1.336638
Month	3.818019
StateH_a	1.070407
StateH_b	1.061704
StateH_c	1.060964
Assortment_a	2.046793
PromoInterval_None	2.020194

→

Final Features: 'DayOfWeek','Customers','Open','Promo','SchoolHoliday','Month','StateH_a','StateH_b','StateH_c','Assortment_a','PromoInterval_None'

→

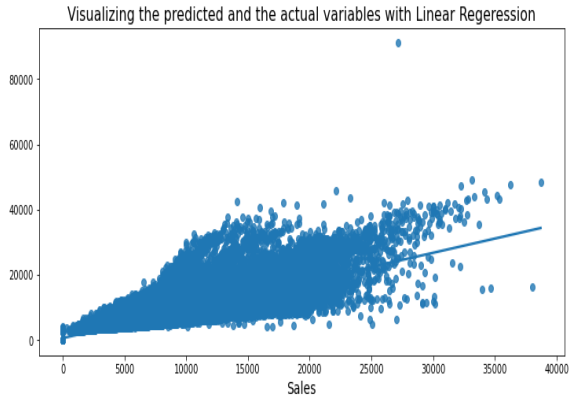
Standard Scaler

Test size for train-test split- 30%

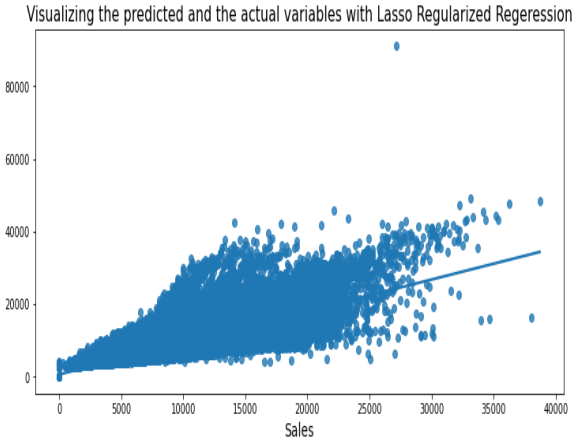
Model Implementation

Model Implementation

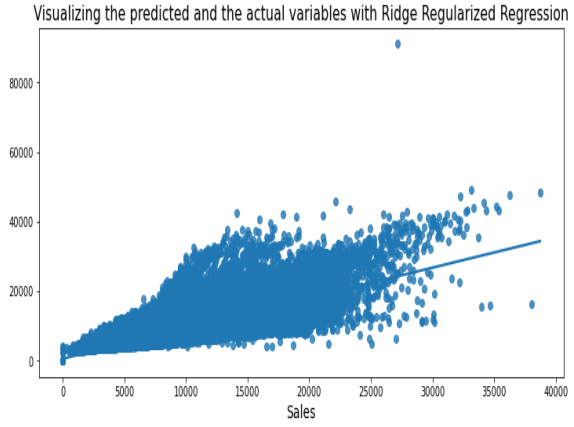
Linear Regression:
Regression Score: 93.86%
MSE : 2650046.85
RMSE : 1627.89
R2 : 0.8207
Adjusted R2 : 0.8207



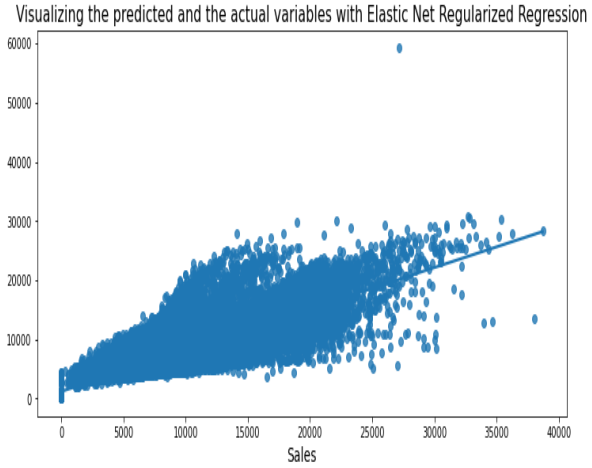
Lasso Regularized Regression:
Regression Score: 93.86%
MSE : 2650046.44
RMSE : 1627.89
R2 : 0.8207
Adjusted R2 : 0.8207



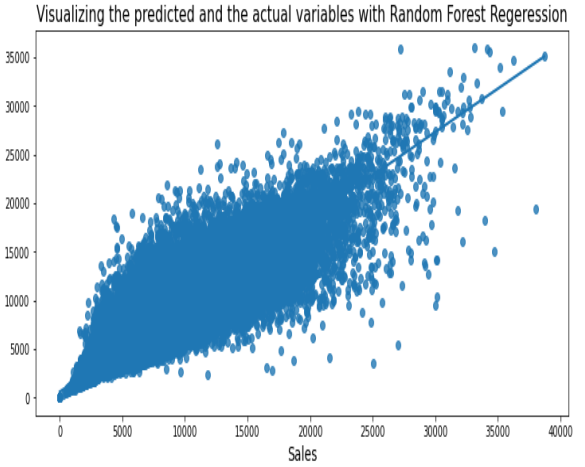
Ridge Regularized Regression:
Regression Score: 93.86%
MSE : 2650037.11
RMSE : 1627.89
R2 : 0.8207
Adjusted R2 : 0.8207



Elastic Net Regularized Regression:
Regression Score: 87.68%
MSE : 3168146.48
RMSE : 1779.92
R2 : 0.7856
Adjusted R2 : 0.7856



Random Forest Regression:
Regression Score: 98.31%
MSE : 1997817.68
RMSE : 1413.44
R2 : 0.8648
Adjusted R2 : 0.8648



Model Implementation

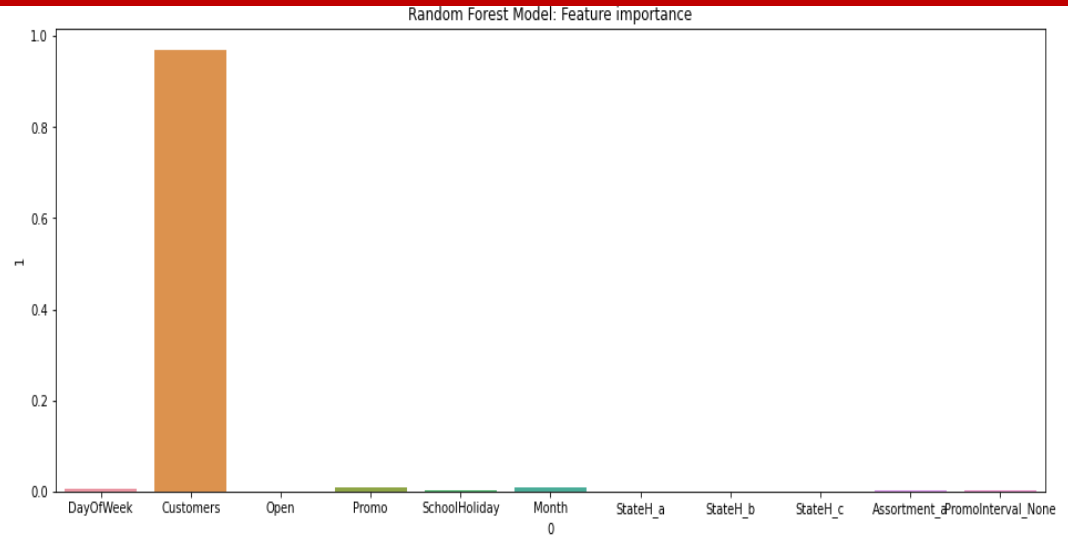
Regression: Remarks

Major flaw with the model

- Dependency on footfall

Model metrics after removing Customer feature:

- MSE : 6038777.217198744
- RMSE : 2457.392361264018
- R2 : 0.5915207858485739
- Adjusted R2 : 0.591507399758554 7



Linear Regression Coefficients

Feature	Coefficient
'DayOfWeek'	-0.99871
'Customers'	19.857750
'Open'	13.966787
'Promo'	4.188622
'SchoolHoliday'	0.130462
'Month'	0.502287
'StateH_a'	-0.706267
'StateH_b'	-0.535271
'StateH_c'	-0.184255
'Assortment_a'	-1.249139
'PromoInterval_None'	-0.271236

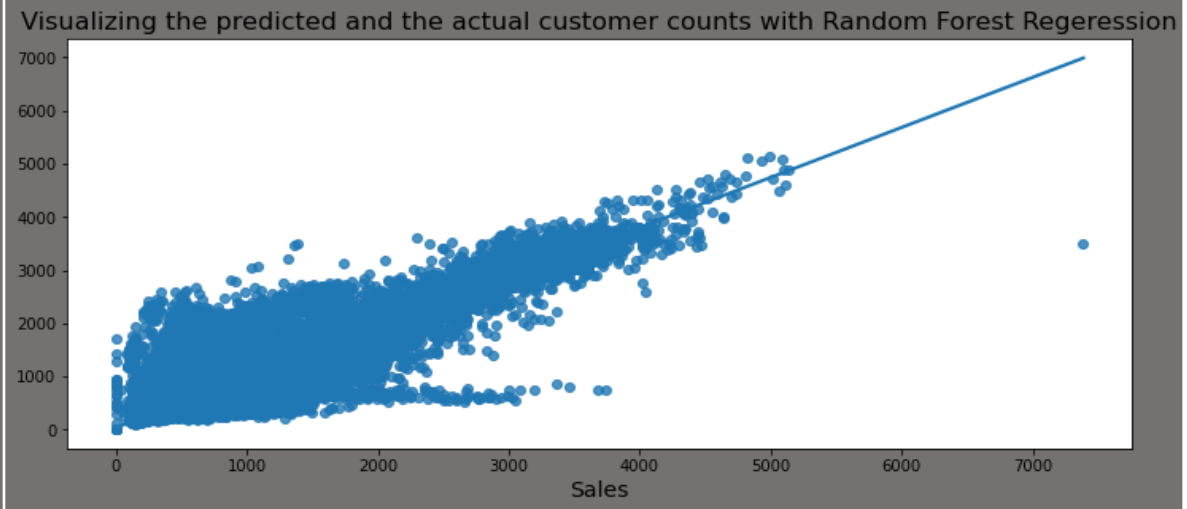
Model Implementation

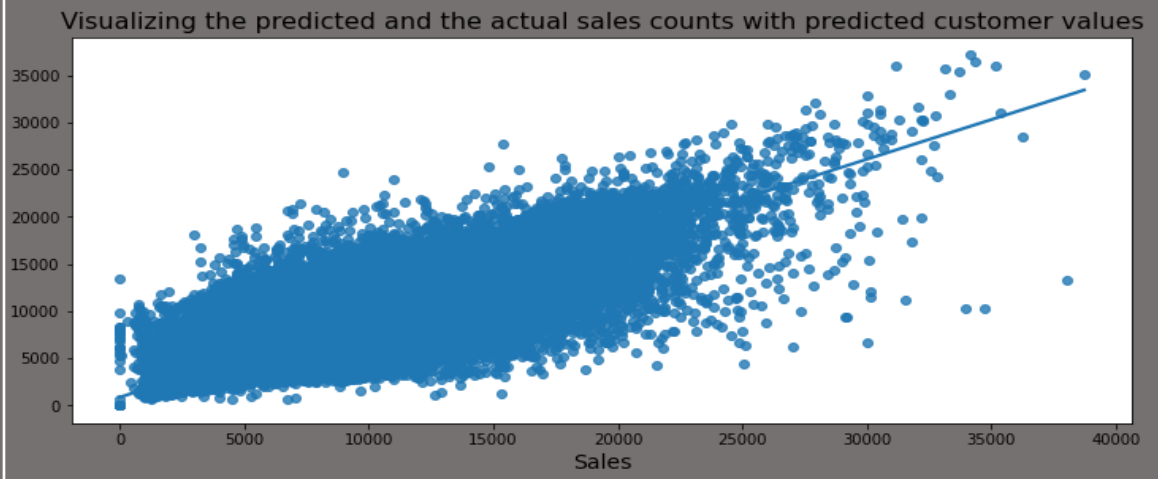
Solution 1: Predict Customers

Customer prediction using Random Forest:
Regression Score: 99.78%
MSE : 23081.65
RMSE : 151.92
R2 : 0.8931
Adjusted R2 : 0.8931

Features Used:
'Store','DayOfWeek','Open','Promo','SchoolHoliday','CompetitionDistance','CompetitionOpenSinceMonth','CompetitionOpenSinceYear','Promo2','Promo2SinceWeek','Promo2SinceYear','Month','Day','StateH_0','StateH_0','StateH_a','StateH_b','StateH_c','S_Type_a','S_Type_b','S_Type_c','S_Type_d','Assortment_a','Assortment_b','Assortment_c','PromoInterval_Feb,May,Aug,Nov','PromoInterval_Jan,Apr,Jul,Oct','PromoInterval_Mar,Jun,Sept,Dec','PromoInterval_None'

Sales prediction using Random Forest with predicted customer values:
Regression Score: 99.02%
MSE : 2868683.40
RMSE : 1693.71
R2 : 0.8059
Adjusted R2 : 0.8059





Model Implementation

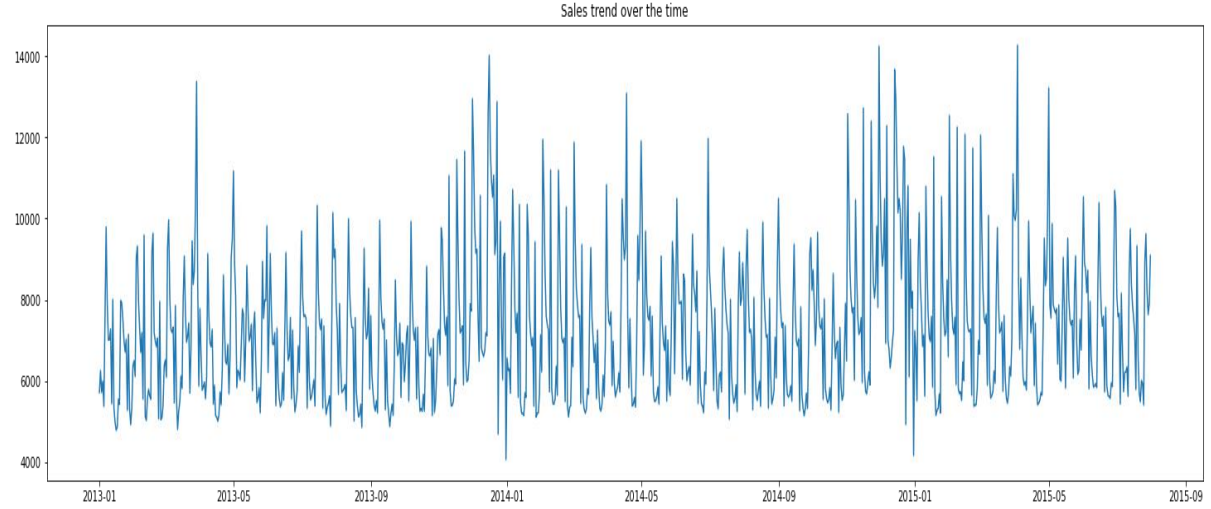
Solution 2: Time Series Analysis

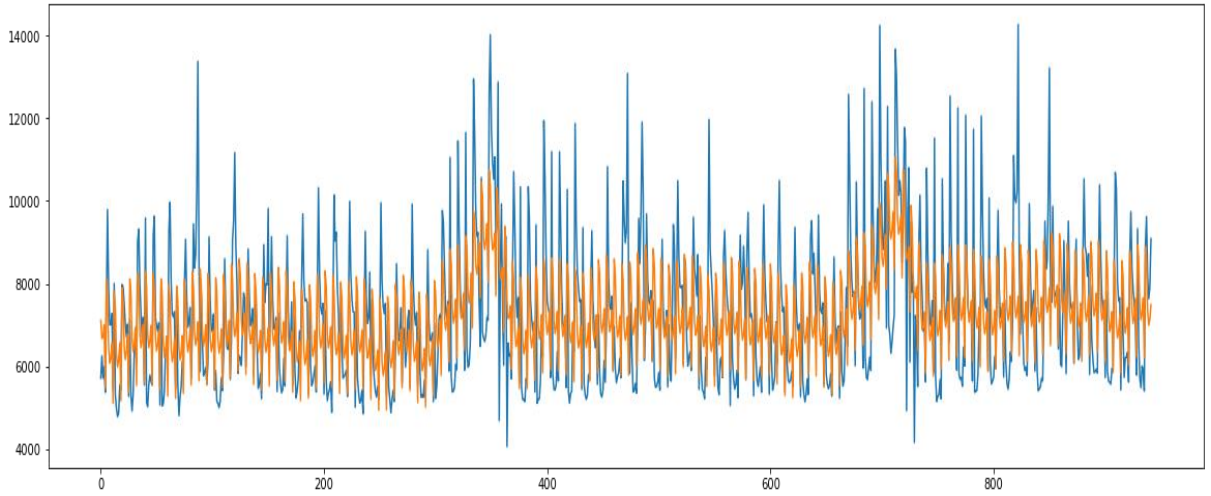
Facebook Prophet

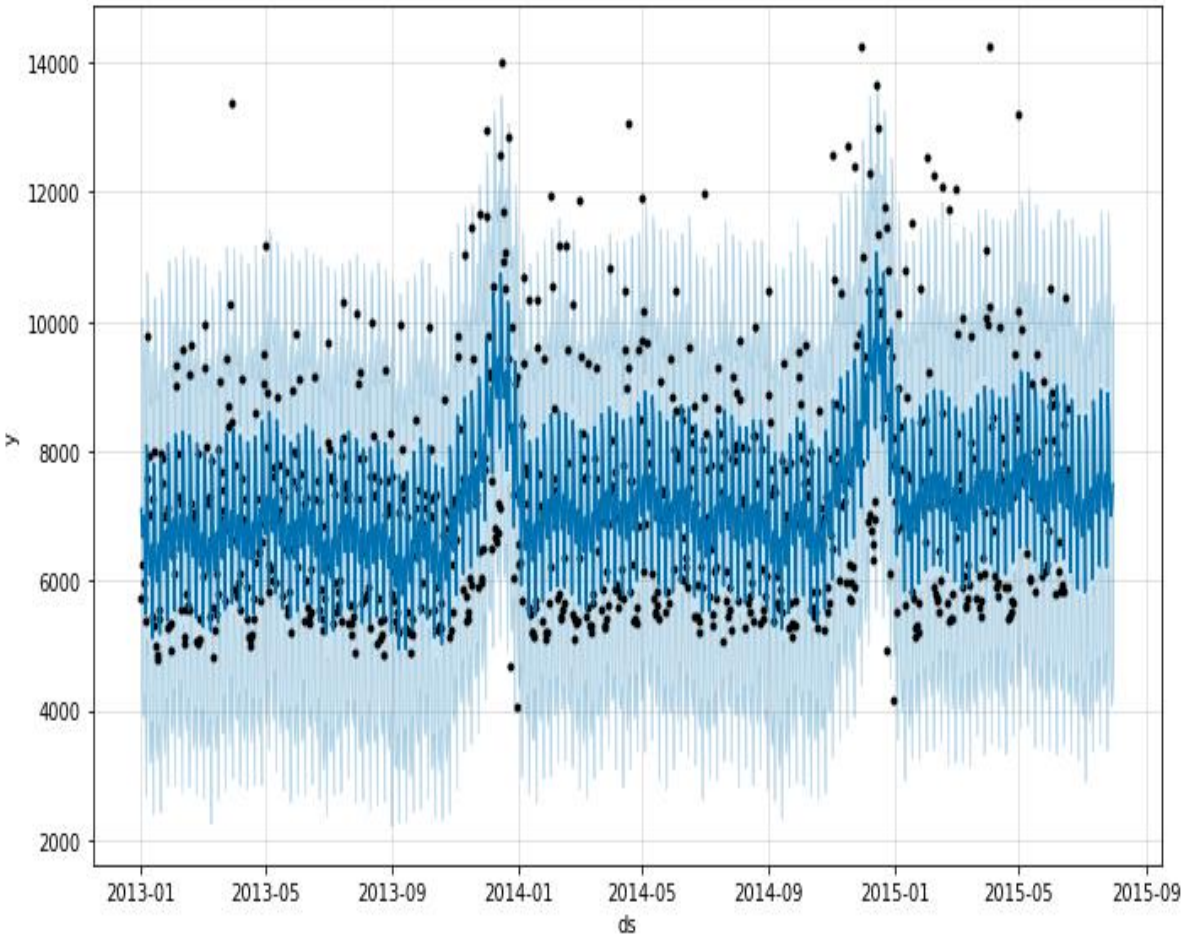
- Trend analysis and forecast tool by Facebook.
- Uses only two features at its most basic level: ‘ds’ and ‘y’

Model Implementation

Actual VS Forecasted Values



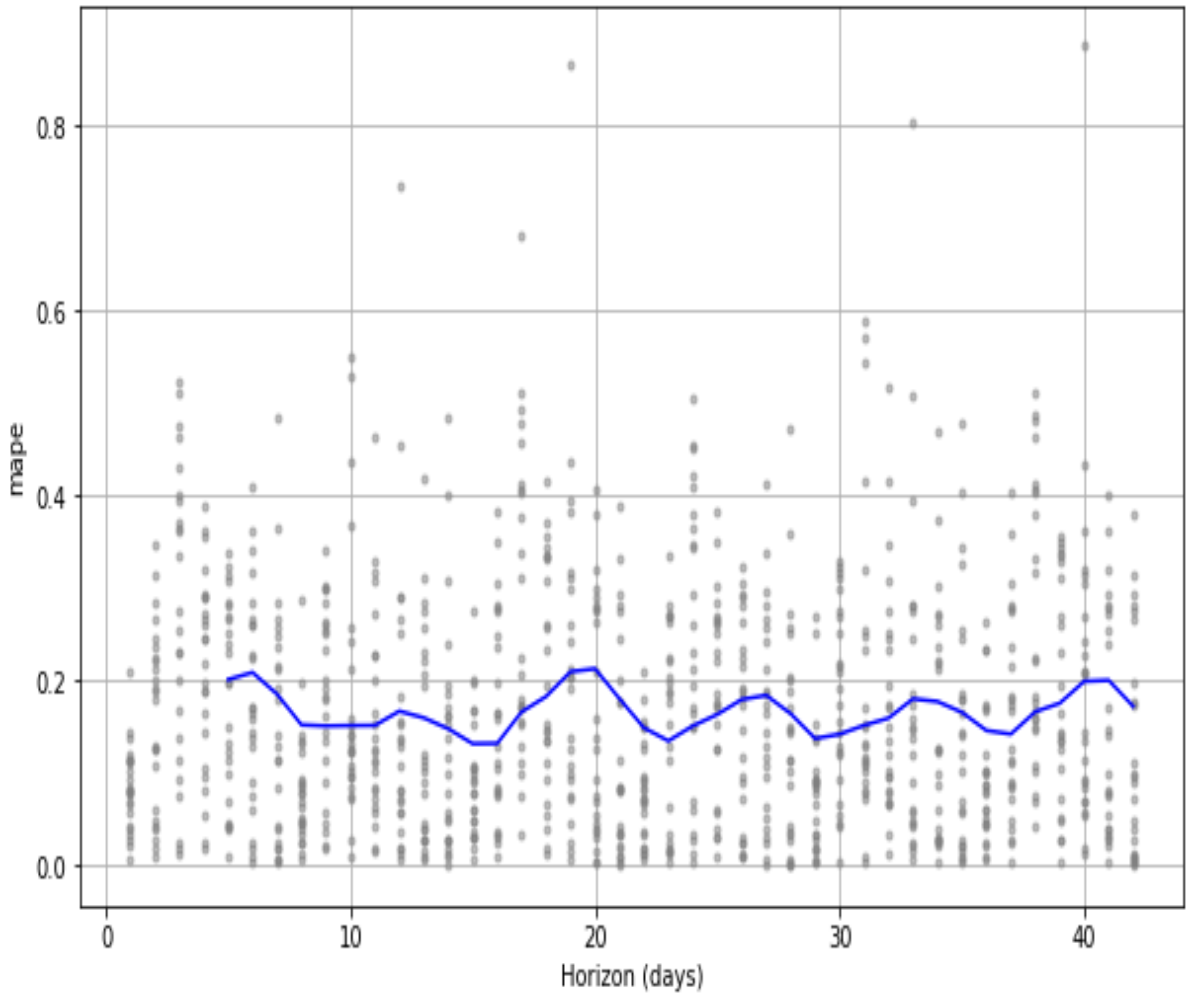




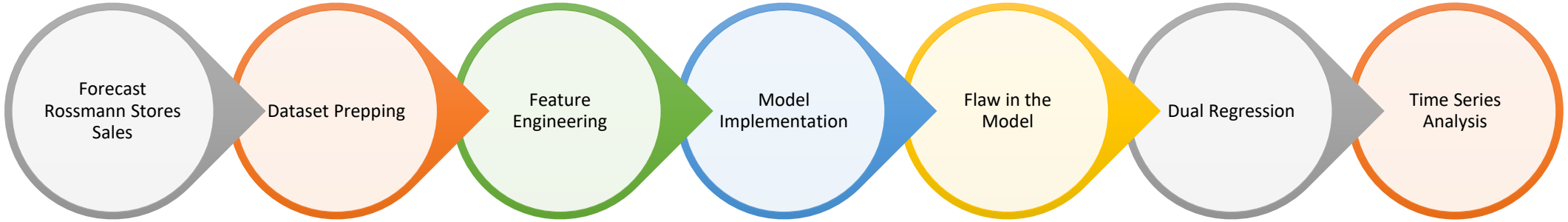
Model Implementation

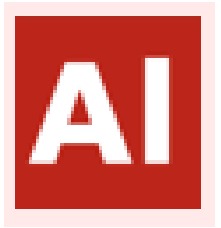
Model Diagnosis

horizon	mse	rmse	mae	mape	mdape	coverage
5 days	3.305448e+06	1818.089167	1525.276989	0.200589	0.198907	0.880000
6 days	3.146887e+06	1773.946716	1496.153488	0.208230	0.207968	0.901667
7 days	2.763578e+06	1662.401263	1316.886535	0.184775	0.191111	0.941667
8 days	2.162627e+06	1470.587238	1070.321673	0.151516	0.131131	0.958333
9 days	2.704998e+06	1644.687760	1179.803732	0.150309	0.133051	0.918333
10 days	2.700173e+06	1643.220412	1210.181368	0.150747	0.129637	0.916667
11 days	2.401443e+06	1549.659119	1199.711464	0.150925	0.124833	0.908333
12 days	2.594802e+06	1610.838914	1292.370776	0.166338	0.134940	0.890000
13 days	2.143142e+06	1463.947278	1174.306656	0.159153	0.123500	0.921667
14 days	2.363538e+06	1537.380194	1116.534996	0.147279	0.113773	0.926667
15 days	2.000387e+06	1414.350358	969.471411	0.131164	0.098860	0.946667
16 days	2.400054e+06	1549.210612	1073.900712	0.131444	0.103231	0.916667
17 days	3.012283e+06	1735.592896	1294.884622	0.165568	0.133098	0.888333
18 days	2.662488e+06	1631.713279	1314.875175	0.182242	0.152867	0.906667
19 days	2.902022e+06	1703.532096	1433.578369	0.209129	0.170921	0.890000
20 days	2.527123e+06	1589.692623	1345.455356	0.212302	0.170921	0.921667
21 days	1.787871e+06	1337.113121	1090.853916	0.179858	0.144512	0.953333
22 days	1.292215e+06	1136.756242	877.370232	0.148295	0.096869	0.960000
23 days	1.650145e+06	1284.579517	943.753460	0.134167	0.092726	0.926667
24 days	2.330513e+06	1526.601787	1142.308310	0.150366	0.126251	0.918333
25 days	2.546547e+06	1595.790299	1250.678272	0.162860	0.153718	0.918333
26 days	2.860183e+06	1691.207512	1382.923933	0.178998	0.179391	0.900000
27 days	2.709656e+06	1646.103315	1347.135247	0.183471	0.179391	0.921667
28 days	2.431545e+06	1559.341071	1193.278050	0.163604	0.171565	0.926667



Conclusion


Forecast Rossmann Stores Sales
Dataset Prepping
Feature Engineering
Model Implementation
Flaw in the Model
Dual Regression
Time Series Analysis



Thank you

Queries/Feedback:
cpvivek1@gmail.com
+91 9446472273 (Whatsapp/call)