The University of Manchester

ABC Grocery Case

Group Coursework

BMAN73701 Programming in Python for Business Analytics

Student ID: 11351804

Agenda

- Introduction
- Methodology
- Exploratory Data Analysis
- Data Preprocessing
- Feature Engineering & Selection
- Modelling & Evaluation
- Prediction & Evaluation
- Results & Analysis
- Conclusion & Contingency

1

Introduction

Introduction

Background

ABC
South America

54 Stores

33 Products

With a consistent positive sales growth

Objective

Predict sales for each product type in each store between 31 July 2017 and 15 August 2017.

Accurate Projection | **Customer Satisfaction** | **Increase Revenue**

Scope & Limitation

- No new marketing and operational initiatives
- All stores work with the same characteristic and strategies
- External factors, such as. Force majeure, government intervention, does not apply

2

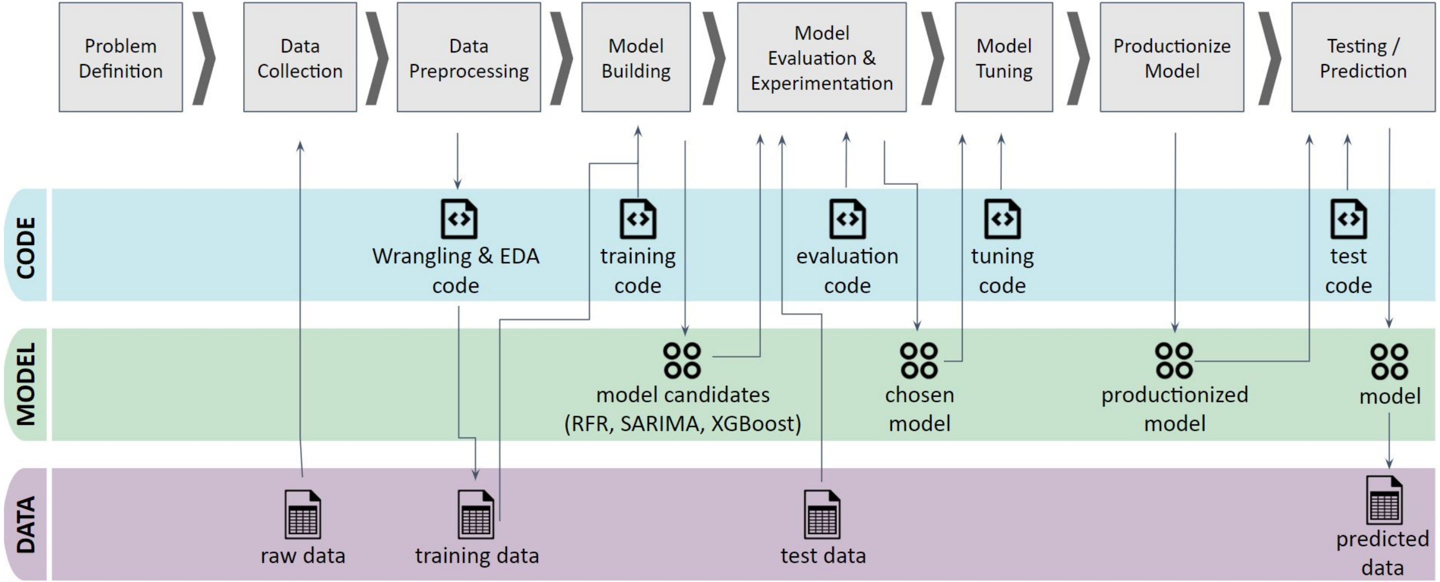
Methodology

Generally, the research consists of 8 stages from defining the problem until predicting the sales. RFR, SARIMA, and XGBoost were chosen as model candidates

Model Candidates

- Random Forest Regressor
- SARIMA
- XGBoost

Machine Learning Pipeline



3

Exploratory Data Analysis

Exploratory Data Analysis

Check NULL

```
date            0
store_nbr       0
product_type    0
sales           0
special_offer   0
dtype: int64
```

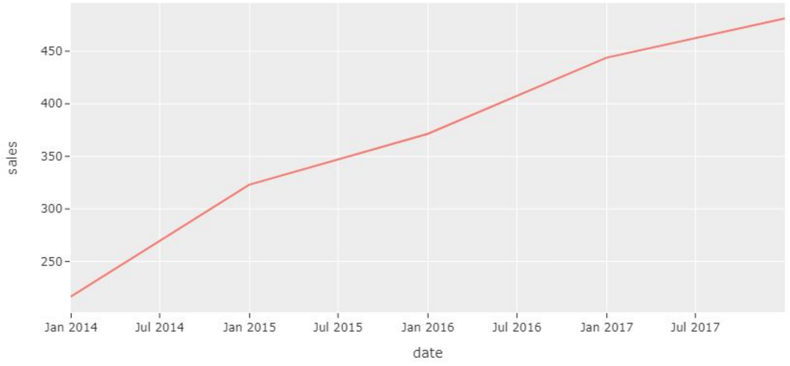
Details of the dataset

Category	Value	Remark
Count of Unique Stores	54	
Count of Unique Product Types	33	
Number of Days or Daily Records	1,684	For each product in each store
Total Number of Records	3,000,888	

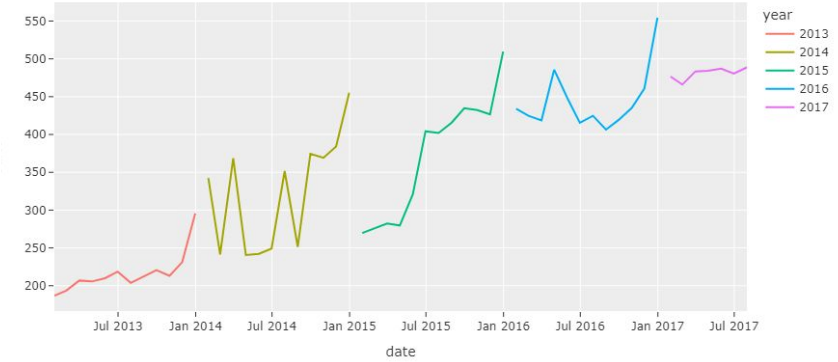
Exploratory Data Analysis

A. Sales Trends

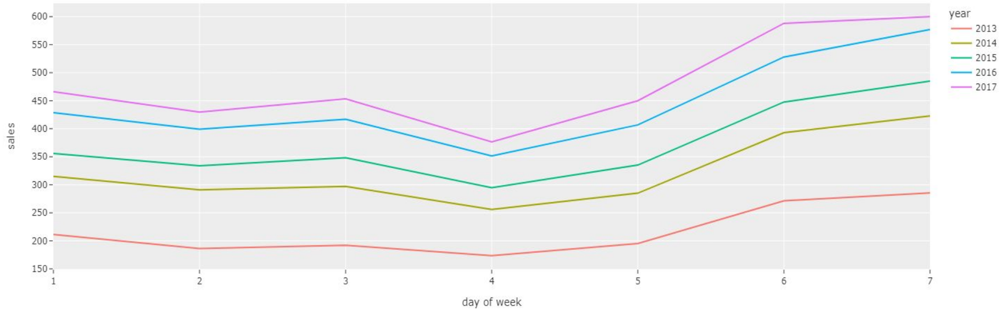
- Annually



- Monthly



- Day of Week

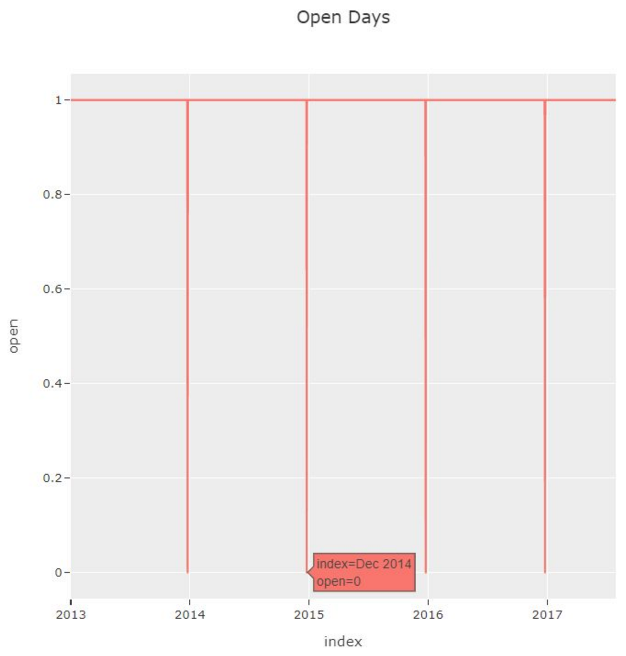


MANCHESTER 1824
The University of Manchester

Exploratory Data Analysis

B. Holidays

Christmas



New Year

Number of Zero Sales on New Year's Day

	m-d	store_nbr	sales
0	01-01	1	0.0
1	01-01	2	0.0
2	01-01	3	0.0
:	:	:	:
22	01-01	23	0.0
23	01-01	24	0.0
25	01-01	26	0.0
:	:	:	:
33	01-01	34	0.0
34	01-01	35	0.0
36	01-01	37	0.0
:	:	:	:
49	01-01	50	0.0
50	01-01	51	0.0
51	01-01	53	0.0
52	01-01	54	0.0

Only store 25 and 36 operated on New Year's Day

	store_nbr	date	sales
0	25	2013-01-01	2511.618999
1	25	2014-01-01	4992.534400
2	25	2015-01-01	12773.616980
3	25	2016-01-01	16433.394000
4	25	2017-01-01	12082.500997
5	36	2013-01-01	0.000000
6	36	2014-01-01	3609.531004
7	36	2015-01-01	0.000000
8	36	2016-01-01	0.000000
9	36	2017-01-01	0.000000

Store 36 only opened on New Year's Day 2014

NB: store 52 only operated on Apr 20th 2017

Exploratory Data Analysis

C. Special Offer

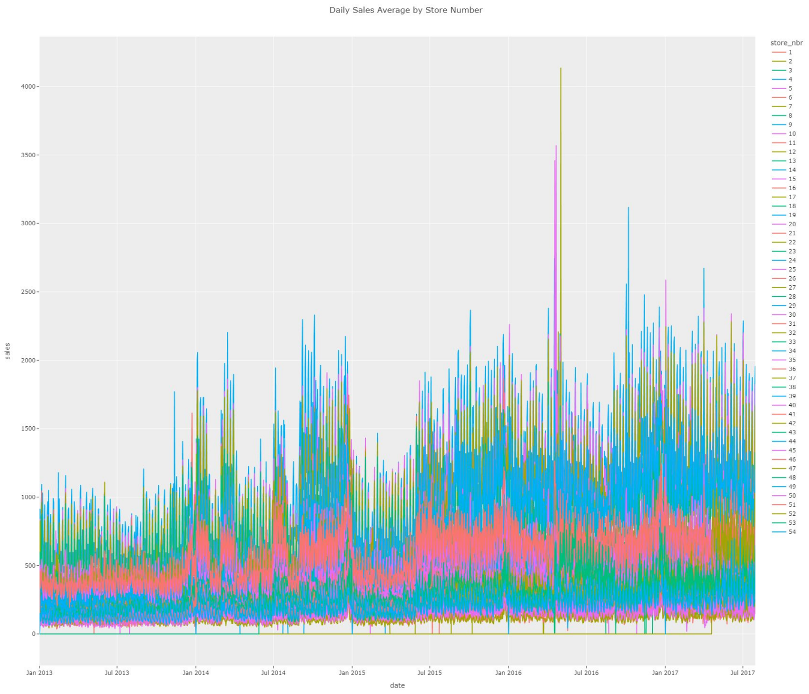
Trend

Correlation to Sales

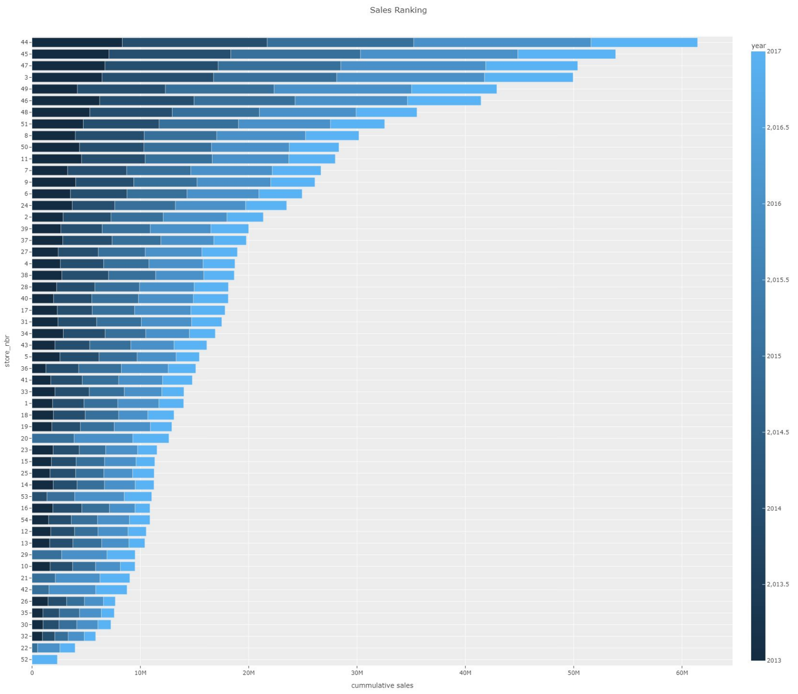
Exploratory Data Analysis

D. Store Performance

Daily Sales Average



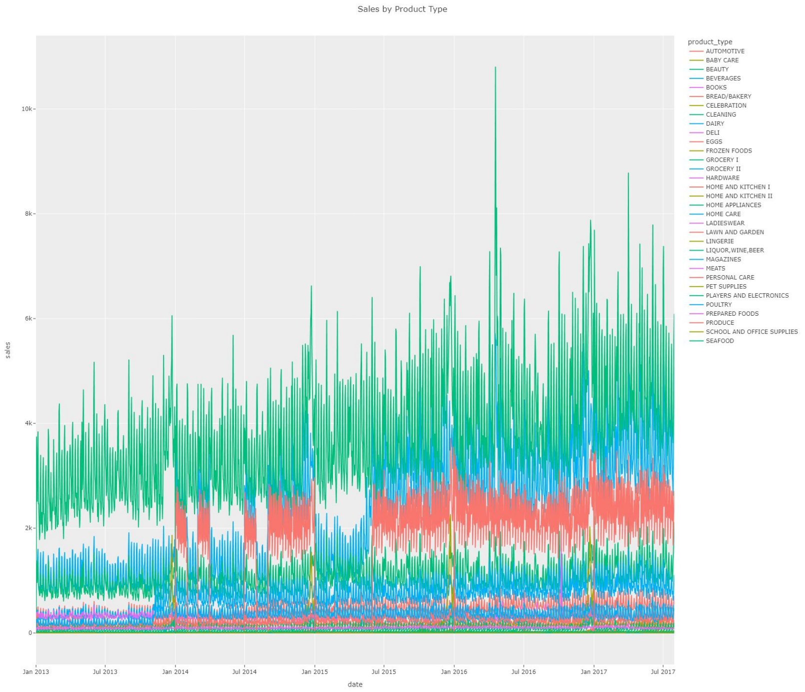
Sales Ranking



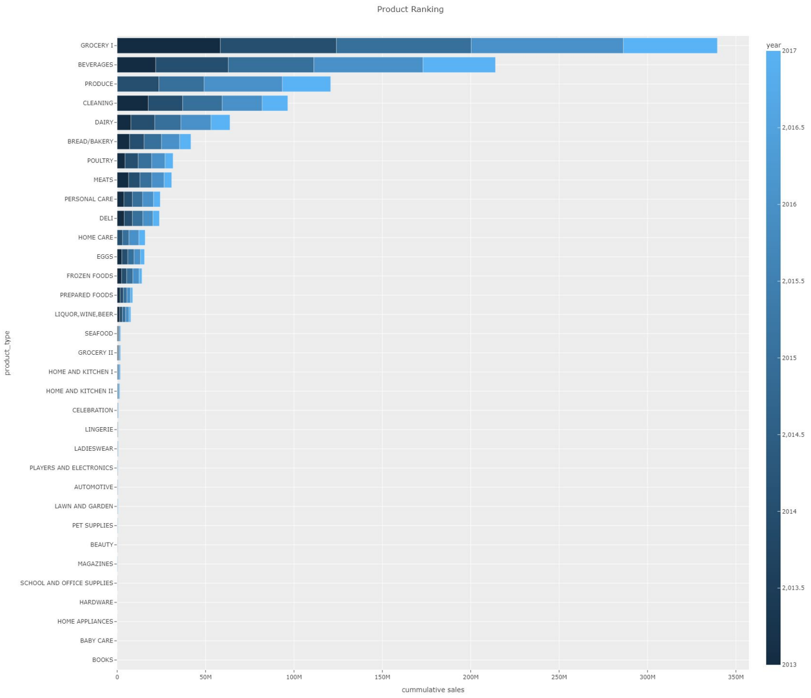
Exploratory Data Analysis

E. Product Type Preference

Daily Sales Average



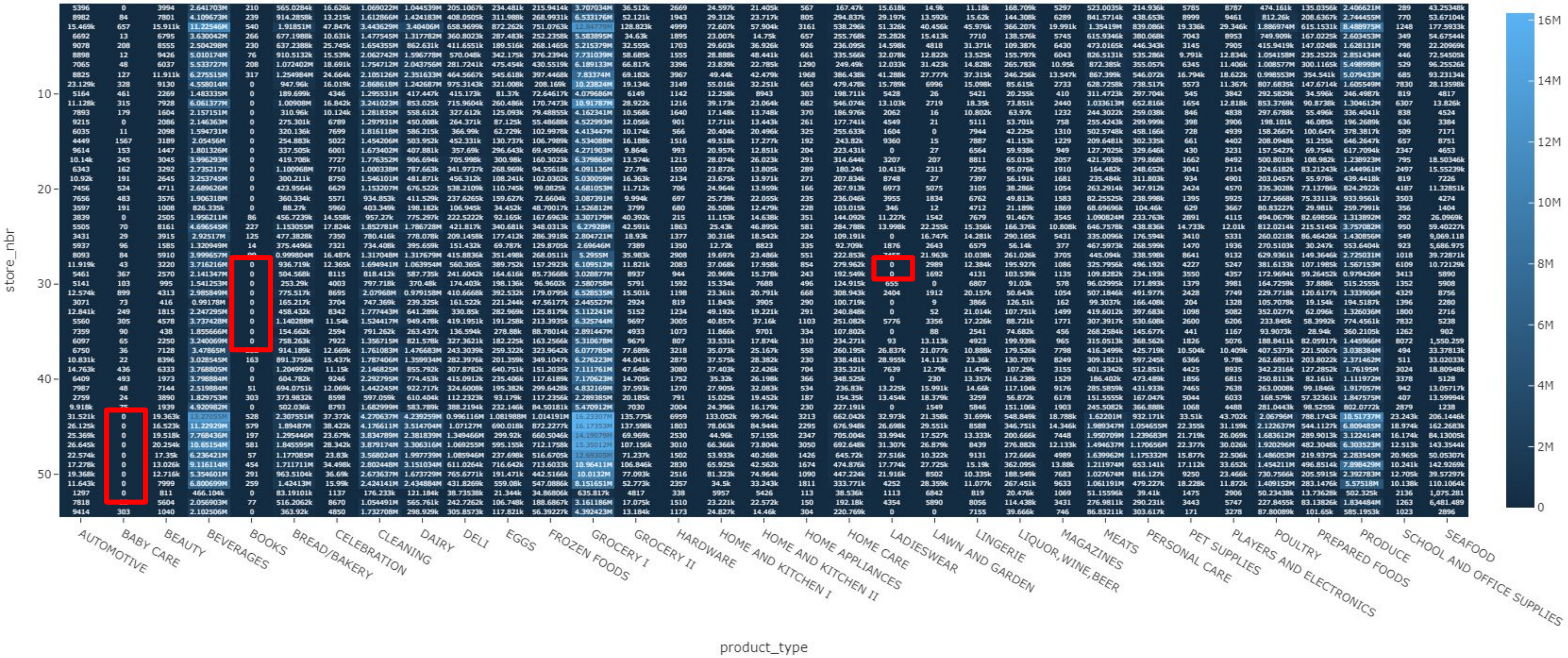
Sales Ranking



MANCHESTER 1824
The University of Manchester

Exploratory Data Analysis

F. Product Sold in Stores



4

Data Preprocessing

Data Preprocessing

Delete Unnecessary Records

- Delete products not sold in some stores
- Delete holidays

Add Time Series Features

Added Features:
- sales lag 21 days
- sales lag 365 days
- rolling means 7 days (shifted 21 days)

NB : 21 days is chosen since we should predict 16 days ahead and match the data with the same day of week

Encode

Apply label encoder to product_type

Splitting Dataset

	Training Set	Test Set
Time Period	1 January 2017 - 30 July 2017	31 July - 15 August 2017
Number of Unique Records	363,122	28,512

Data Preprocessing

Training set example

	store_nbr	product_type	date	sales	special_offer	day of week	sales_lag21	sales_lag365	rolling_means7	encoded_product_type
2505641	25	AUTOMOTIVE	2017-01-01	5.000	0	7	5.000	4.000	2.571429	0
2505642	25	BABY CARE	2017-01-01	2.000	0	7	0.000	0.000	0.000000	1
2505643	25	BEAUTY	2017-01-01	3.000	0	7	2.000	13.000	2.285714	2
2505644	25	BEVERAGES	2017-01-01	4008.000	38	7	2448.000	5104.000	1516.142857	3
2505645	25	BOOKS	2017-01-01	0.000	0	7	0.000	0.000	0.142857	4
2505646	25	BREAD/BAKERY	2017-01-01	490.573	3	7	304.747	680.952	292.831143	5

Test set example

Sales from previous 21 days

Sales from previous 365 days

7-day average sales from previous 21 days

	store_nbr	product_type	date	sales	special_offer	day of week	sales_lag21	sales_lag365	rolling_means7	encoded_product_type
2896417	54	MAGAZINES	2017-08-15	2.000	0	2	1.000	4.000000	1.285714	23
2896418	54	MEATS	2017-08-15	57.842	0	2	61.225	62.073997	55.087143	24
2896419	54	PERSONAL CARE	2017-08-15	169.000	5	2	125.000	151.000000	162.714286	25
2896420	54	PET SUPPLIES	2017-08-15	0.000	0	2	0.000	0.000000	0.142857	26
2896421	54	PLAYERS AND ELECTRONICS	2017-08-15	2.000	0	2	3.000	6.000000	2.285714	27
2896422	54	POULTRY	2017-08-15	59.619	0	2	50.686	124.472000	65.829858	28
2896423	54	PREPARED FOODS	2017-08-15	94.000	0	2	65.000	60.000000	88.285714	29
2896424	54	PRODUCE	2017-08-15	915.371	76	2	914.959	578.231000	672.206000	30
2896425	54	SCHOOL AND OFFICE SUPPLIES	2017-08-15	0.000	0	2	0.000	0.000000	0.000000	31
2896426	54	SEAFOOD	2017-08-15	3.000	0	2	7.000	4.000000	2.714286	32

5

Modelling & Evaluation

Modelling & Evaluation

(1) Deciding the models

In examining the case of ABC Grocery, our objective is to predict future sales using past sales data for each store and product type. This task is typically referred to as **"Time Series Forecasting"**.

(2) Splitting Train/Test Datasets

Dataframe

80% 20%

Train Test

(3) Model Fitting

Random Forest Regressor

XGBoost

SARIMA

Lengthy Computation times

(4) Hyperparameter Tuning

8 combinations of:
- N_estimators
- Max_features
- Max_depth

Search type: Random

(5) Cross Validation

TSCV Cross Validation

Modelling & Evaluation

Random Forest Regressor Performance Metric Evaluation

MAE	Training set : 30.03 Test set : 84.07
RMSE	Training set : 113.41 Test set : 301.35
Adjusted R2	Training set : 0.99 Test set : 0.94

XG Boost Performance Metric Evaluation

MAE	Training set : 50.05 Test set : 87.91
RMSE	Training set : 143.88 Test set : 321.40
Adjusted R2	Training set : 0.98 Test set : 0.93

Modelling & Evaluation

Feature Importance

Importance level	Feature (Random Forest)	Importance
1	Rolling_means7	0.290
2	Sales_lag21	0.288
3	Sales_lag365	0.211
4	Special_offer	0.115
5	encoded_product_type	0.037
6	store_nbr	0.034
7	day of week	0.023

Feature Selection

‘Store_nbr’ and ‘day_of_week’ are dropped since they show the least significance

Importance level	Feature (Random Forest)
1	Sales_lag21
2	Rolling_means7
3	Sales_lag365
4	Special_offer
5	encoded_product_type

Modelling & Evaluation

Random Forest Regressor Hyperparameter Tuning

Before Tuning		After Tuning
Training set : 30.03 Test set : 84.07	**MAE**	Training set : 61.10 Test set : 83.77
Training set : 113.41 Test set : 301.35	**RMSE**	Training set : 194.02 Test set : 296.49
Training set : 0.99 Test set : 0.94	**Adjusted R2**	Training set : 0.98 Test set : 0.95

6

Prediction & Evaluation

Prediction and Evaluation

(1) Model Decided for Test dataset

Random Forest Regressor

(2) Apply the model to Predict the Test Dataset (31 July - 15 August 2023)

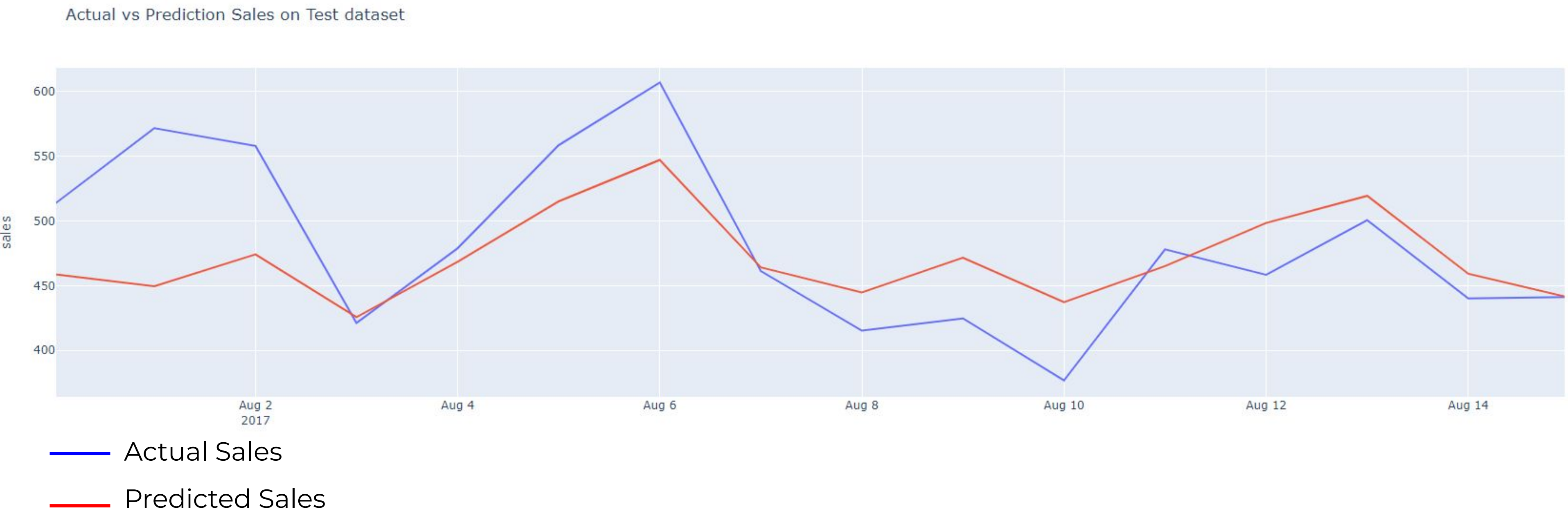
RFR Model	.predict	(df_test[[*selected features*]])

(3) Aggregate and Visualize the prediction on general overview, each store, and each product

(4) Evaluate the Predicted Sales vs the Actual Sales

Prediction - Actual vs Prediction Sales (Overview)

Actual vs Prediction Sales on Test dataset



Prediction - Actual vs Prediction Sales (Stores)

Example of stores with high accuracy

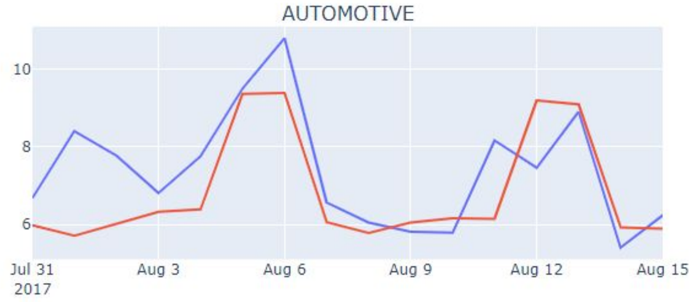
Example of stores with moderate accuracy

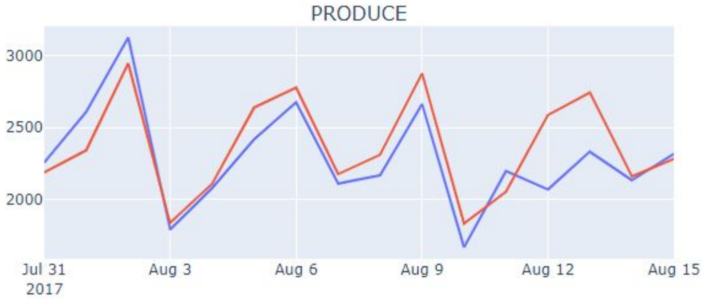
Actual Sale

Predicted Sales

Prediction - Actual vs Prediction Sales (Products)

Example of products with high accuracy

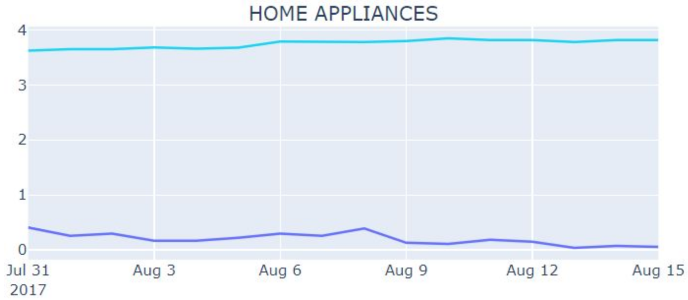




Actual Sale

Predicted Sales

Example of product with low accuracy



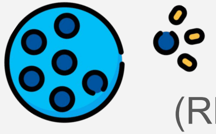
7

Results & Analysis

Result & Analysis

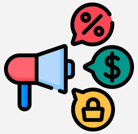
Key Points

Moderately good projection (MAE ~84) ≈ 16% deviation



high outliers among products (RMSE score)

The higher of the sales, **The better** accuracy



Special offer has a **positive impact** on boosting sales

Differentiate the inventory management based on the projection accuracy rate

8

Conclusion & Contingency

Conclusion & Contingency

Conclusion

- As positive correlation exists, in business, the company can give additional special offers in stores to boost sales next time around.
- Demand uncertainty in particular items are identified based on the error values of the model.

Contingency

- Distinct models for products and store due to high variations
- Further study to see the best time lagging of projection
- Store-specific modeling
- Comprehensive analysis for product categories

Thank you

MANCHESTER
1824
The University of Manchester