CAPSTONE

Chloe Loh
GA DSI 38

CONTENTS

01

Context & Problem Statement

02

Data Collection & Cleaning

03

Exploratory Data Analysis (EDA)

04

Model & Evaluation

05

Conclusion & Recommendations

01

Context & Problem Statement

Context

Demand Forecasting

Critical for business success
Optimize operations + Maximize profits
Over-stocking => extra business cost
Under-stocking => loss of revenues

Importance

Operations Optimization

Resources allocation

Inventory Management

Storage & Logistics

Profitability

Cost efficiency & Cashflow

Customer Satisfaction

Growth & Sustainability

Challenges

Data

Availability & Quality

Market Uncertainty

Business dynamics & Economic conditions

Customer Behavior

Trends & Seasonality changes rapidly

Human Errors & Bias

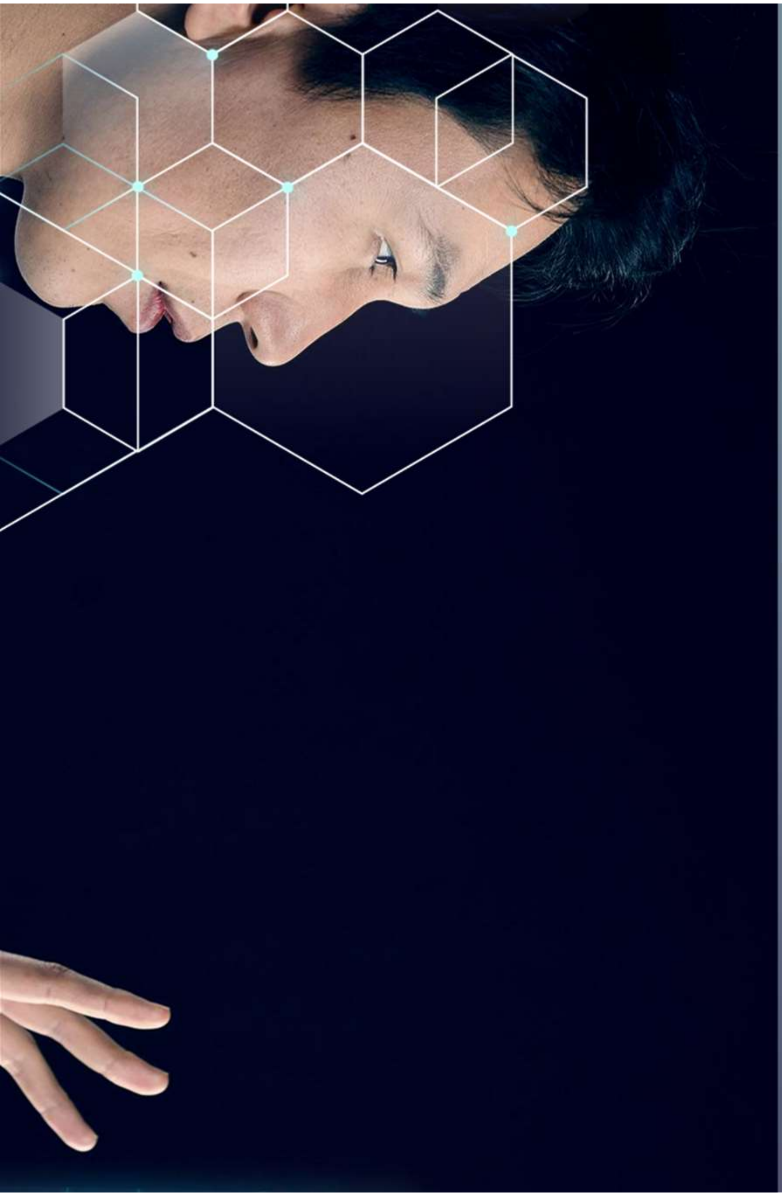
Personal judgement & stakeholders conflicting interests

Lack Expertise

Data Analysis, Statistics & Domain Knowledge

Complex Supply Chain

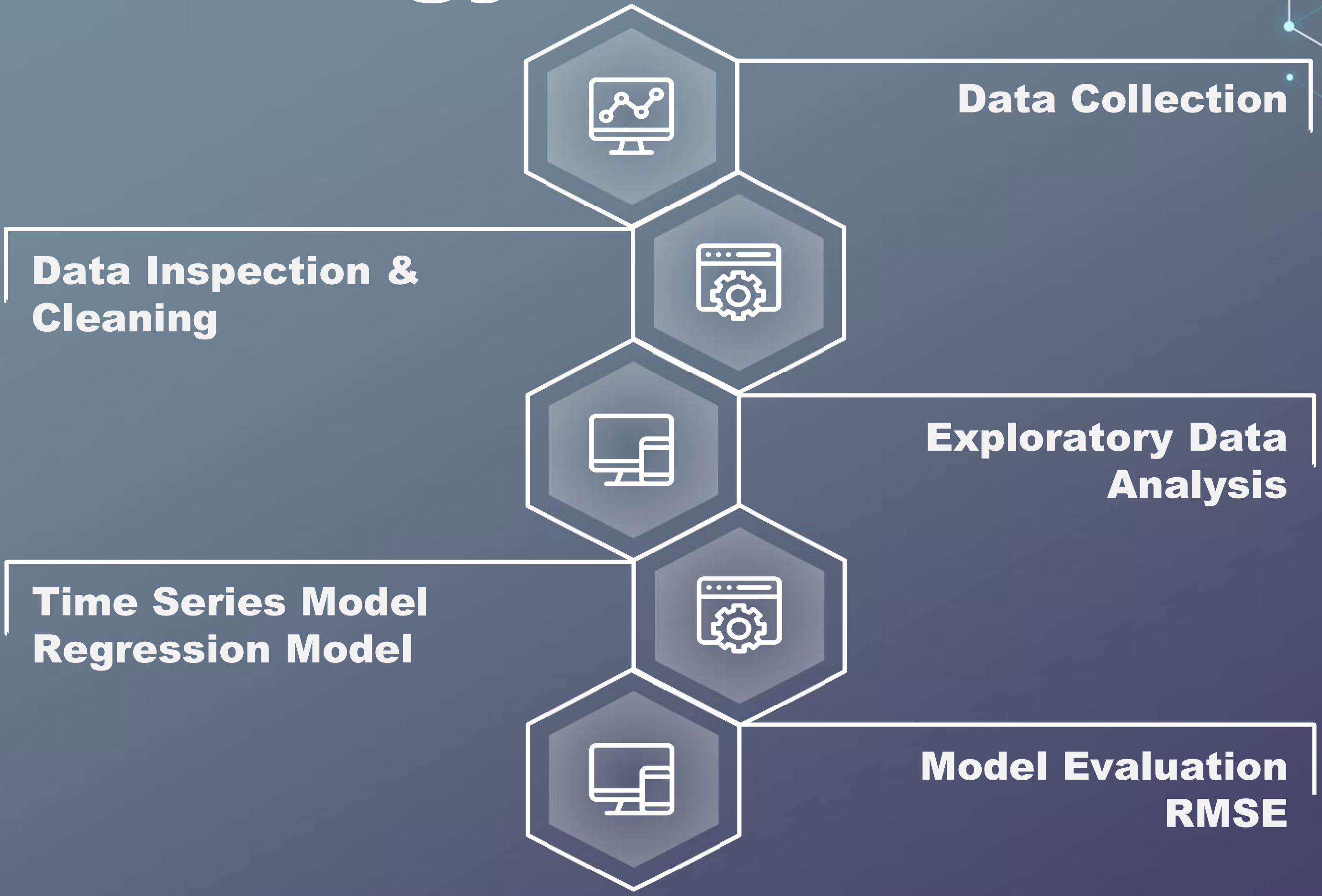
Global sourcing & interdependency



Problem Statement

Develop a short term forecasting model to help business effectively plan for demand surge in next 3 months in conjunction with festive season and special events, as current rule-based practice not able to predict seasonality

Methodology



Data Collection

Data Inspection &
Cleaning

Exploratory Data
Analysis

Time Series Model
Regression Model

Model Evaluation
RMSE

02

Data Collection & Cleaning

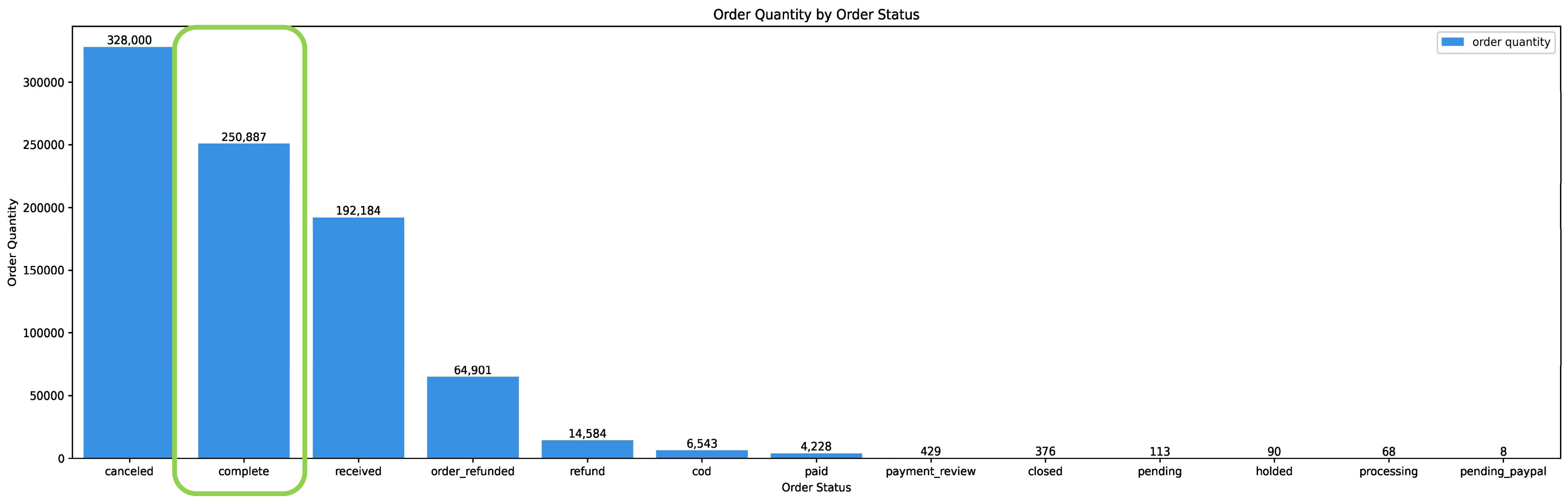
Data Collection & Cleaning

1. Data source: online ecommerce sales datasets 286,000 rows, 35 features
2. Period: October 2020 to September 2021 (12 months of daily sales)
3. Inspection for nulls, duplication, data types
4. Standardization of columns
5. Convert date columns to datetime index
6. Further inspection on mathematical accuracy, nulls, duplication after cleaning

03

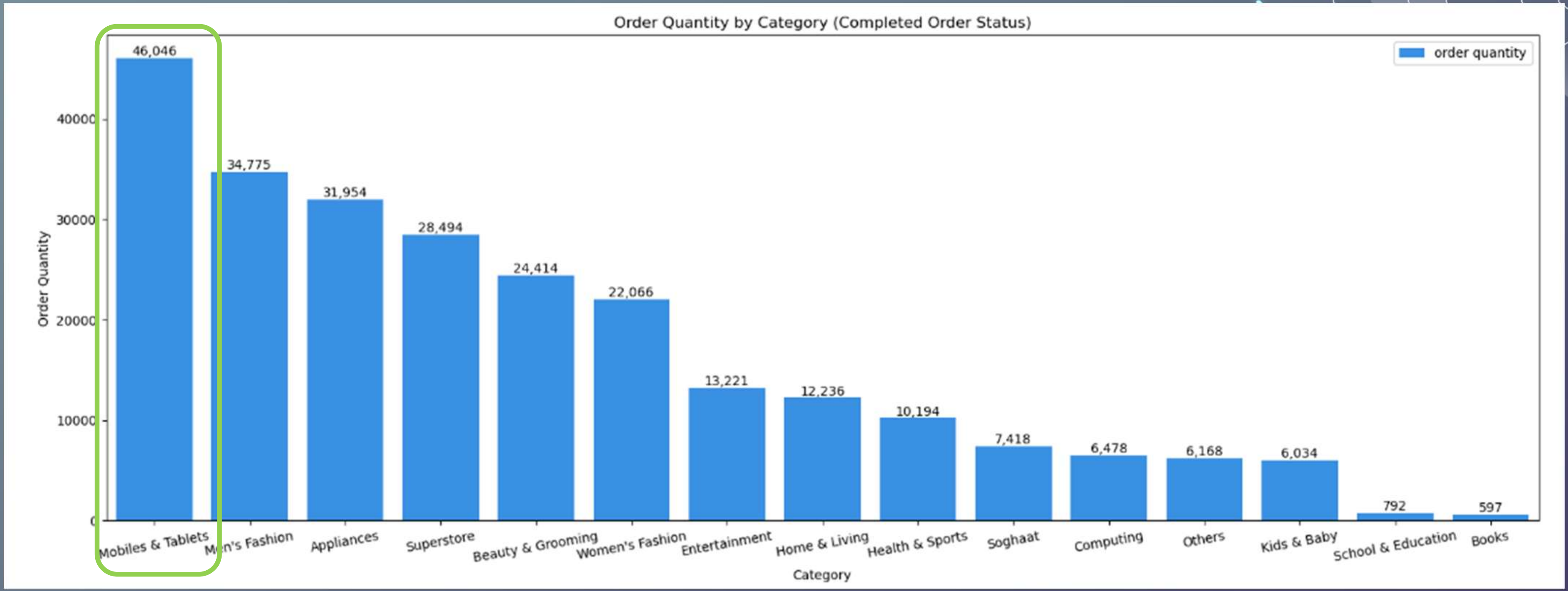
Exploratory Data Analysis (EDA)

Overall Order Status



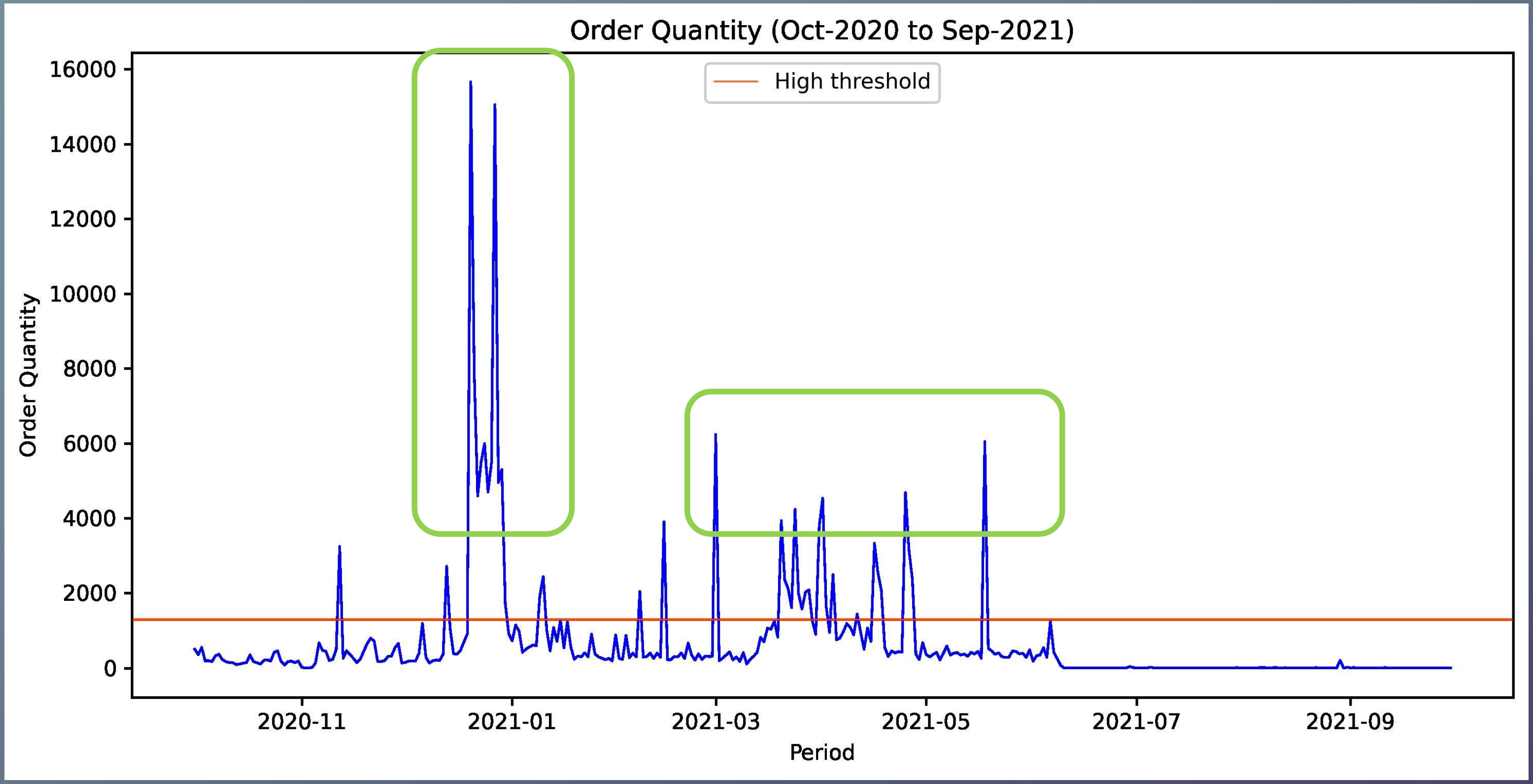
- Order status indicated as ‘completed’ means product is delivered and payment is received
- Cancelled order status is significant hence business should gather data on reasons of cancellation

Order Quantity by Category



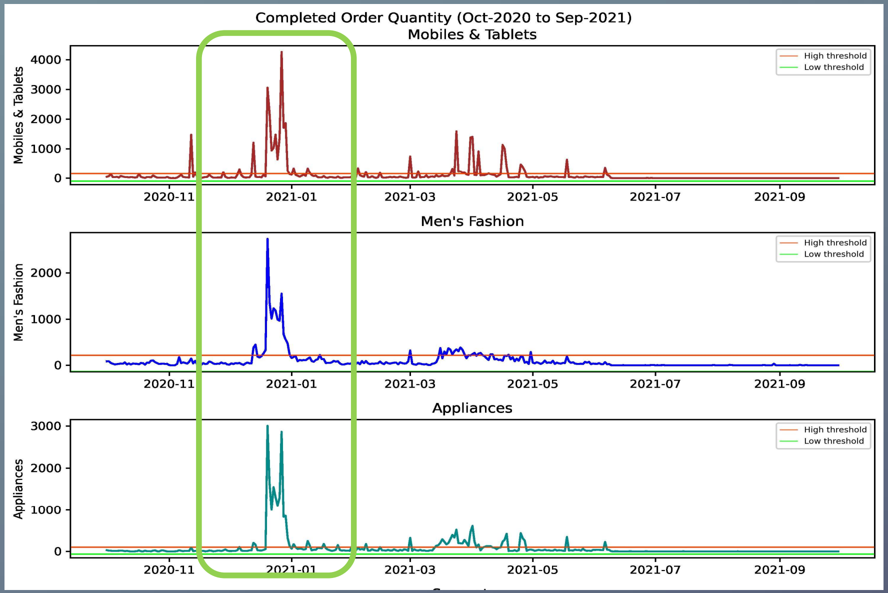
- Mobiles & Tablets is top sellers and has more daily sales, hence it is selected for modelling
- Top 3 categories made up 45% of total completed orders

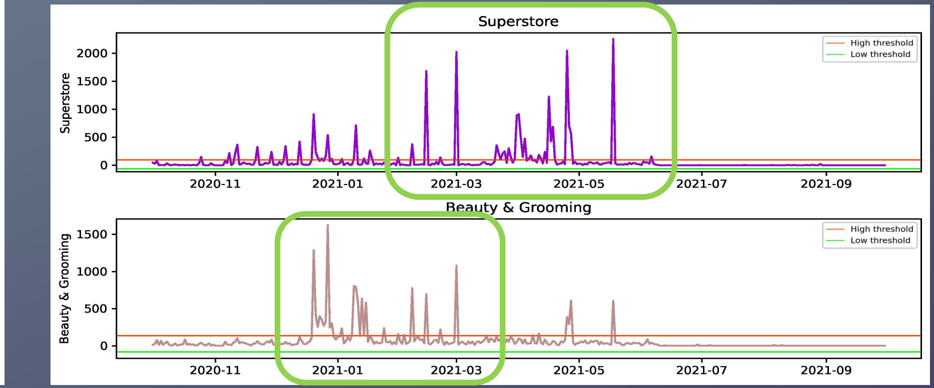
Trends of Order Quantity



- Trend for all category
- Outliers = peaks
- Dec-2020 = Christmas
- Threshold = 1299 units
- Low demand
- Cyclical business

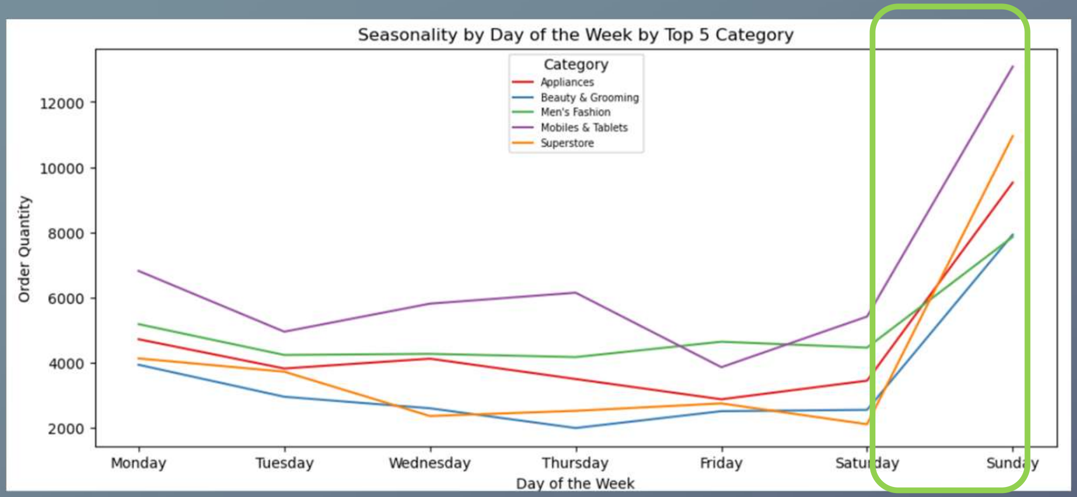
Trends by Top 5 Category



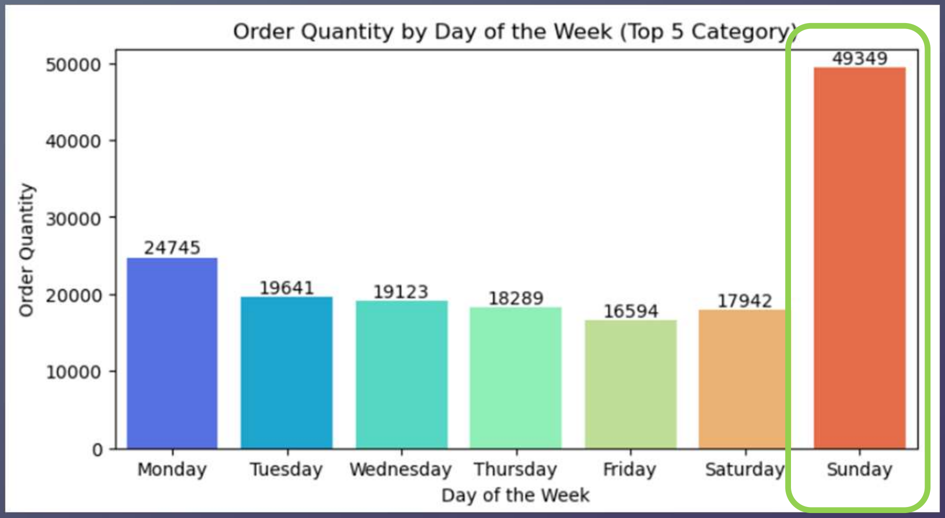


- Trend for top 3 category are similar and follows closely with all category
- Top 3 = Mobiles & Tablets, Men’s Fashion and Appliances
- Next 2 category follows a different trend, hence one product category is chosen for modelling.
- Flat demand from Jun to Sep 2021

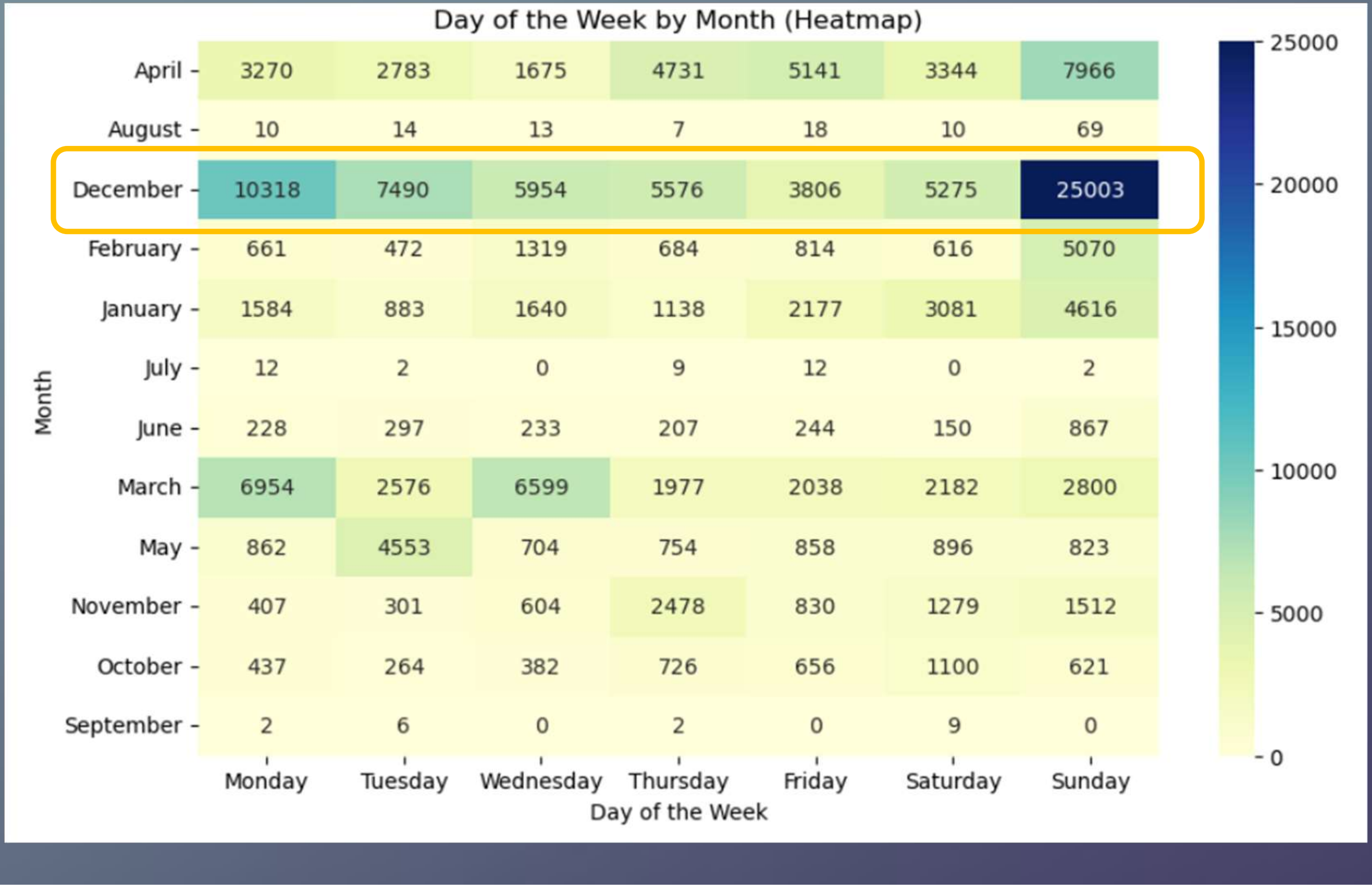
Day of Week Analysis



- Top 5 category shows similar trend for weekly trend
- Sundays are the popular day to do online shopping
- Fridays are the lowest due to socialising



Day of Week Analysis

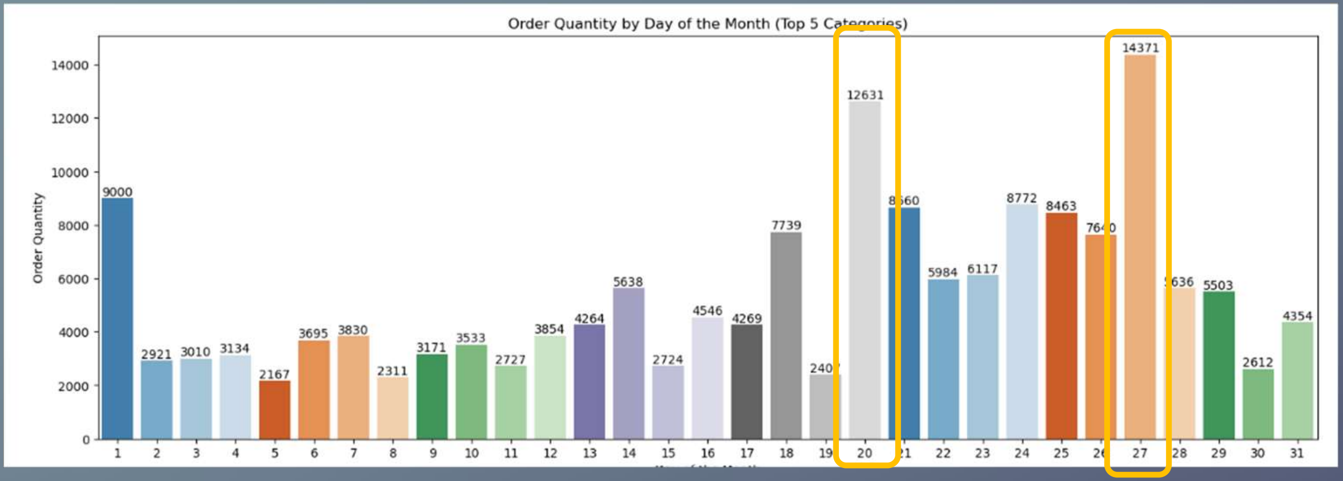


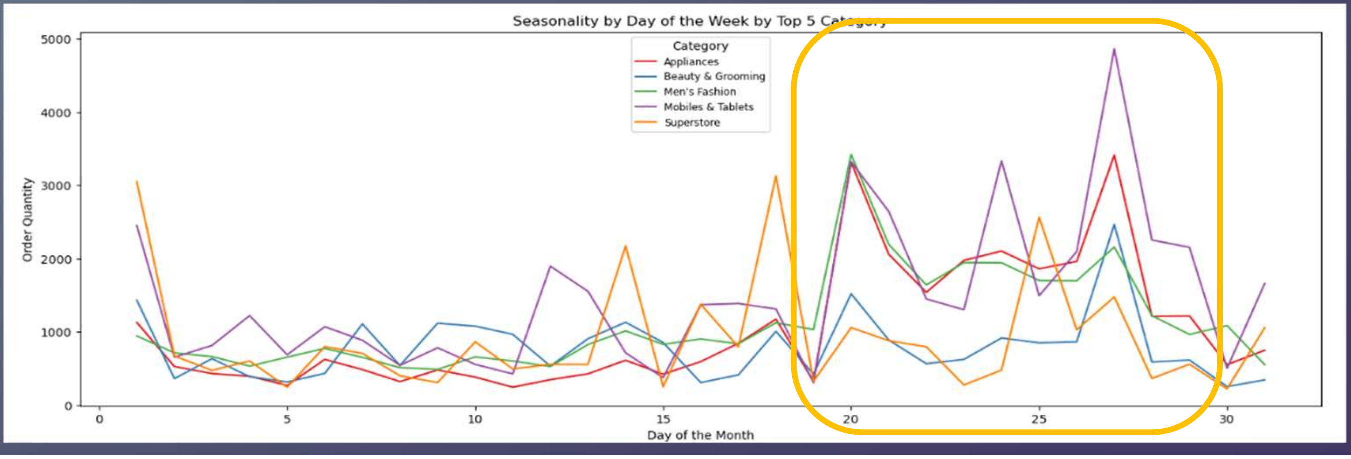
Day of the Week by Month (Heatmap)

Month	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
April	3270	2783	1675	4731	5141	3344	7966
August	10	14	13	7	18	10	69
December	10318	7490	5954	5576	3806	5275	25003
February	661	472	1319	684	814	616	5070
January	1584	883	1640	1138	2177	3081	4616
July	12	2	0	9	12	0	2
June	228	297	233	207	244	150	867
March	6954	2576	6599	1977	2038	2182	2800
May	862	4553	704	754	858	896	823
November	407	301	604	2478	830	1279	1512
October	437	264	382	726	656	1100	621
September	2	6	0	2	0	9	0

- Correlation between day of week and month
- Top 5 categories
- Sundays in December exhibit strongest correlation with orders

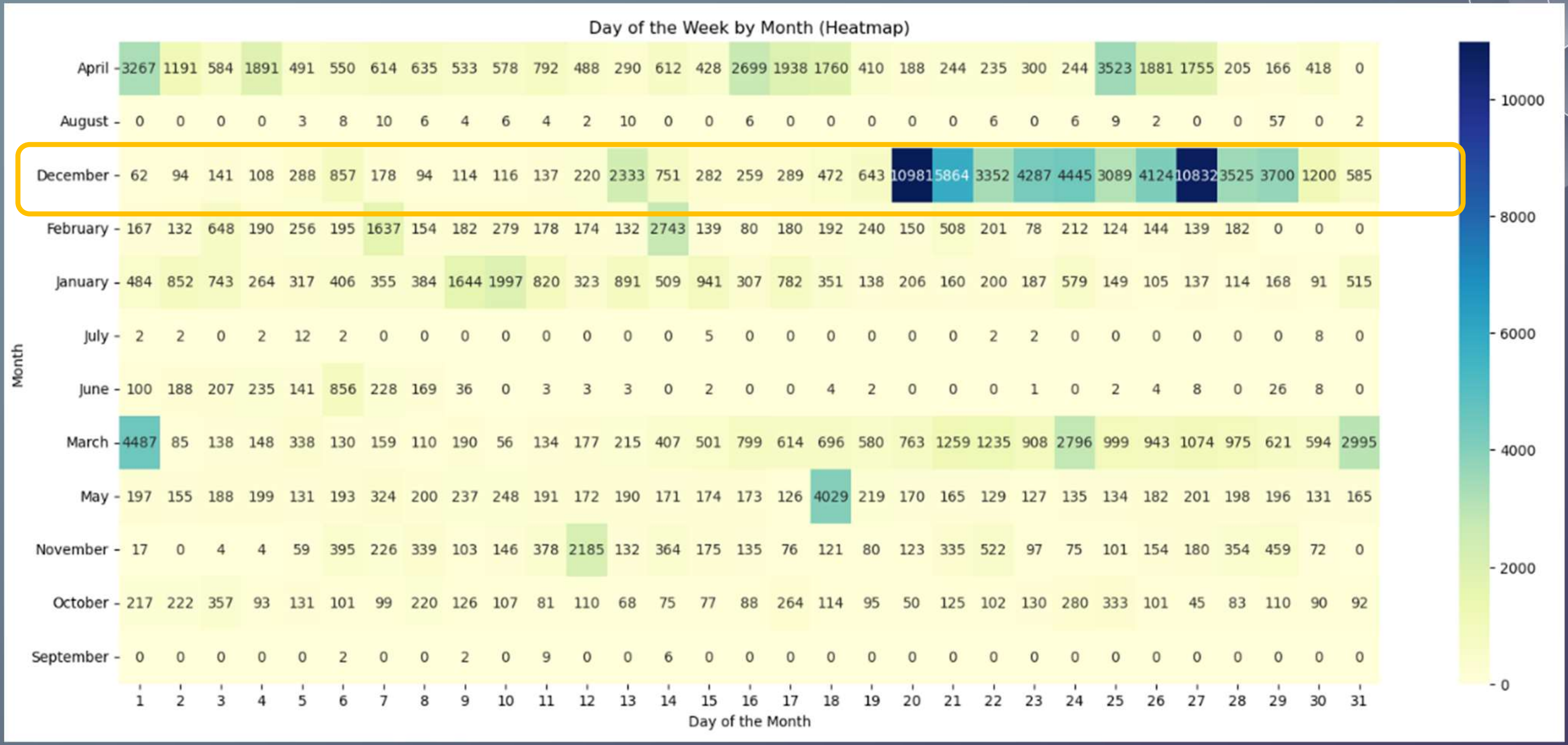
Day of Month Analysis





- 27th and 20th = Christmas, weekends
- Top 5 category shows similar uptrend towards end of month

Day of Month Analysis



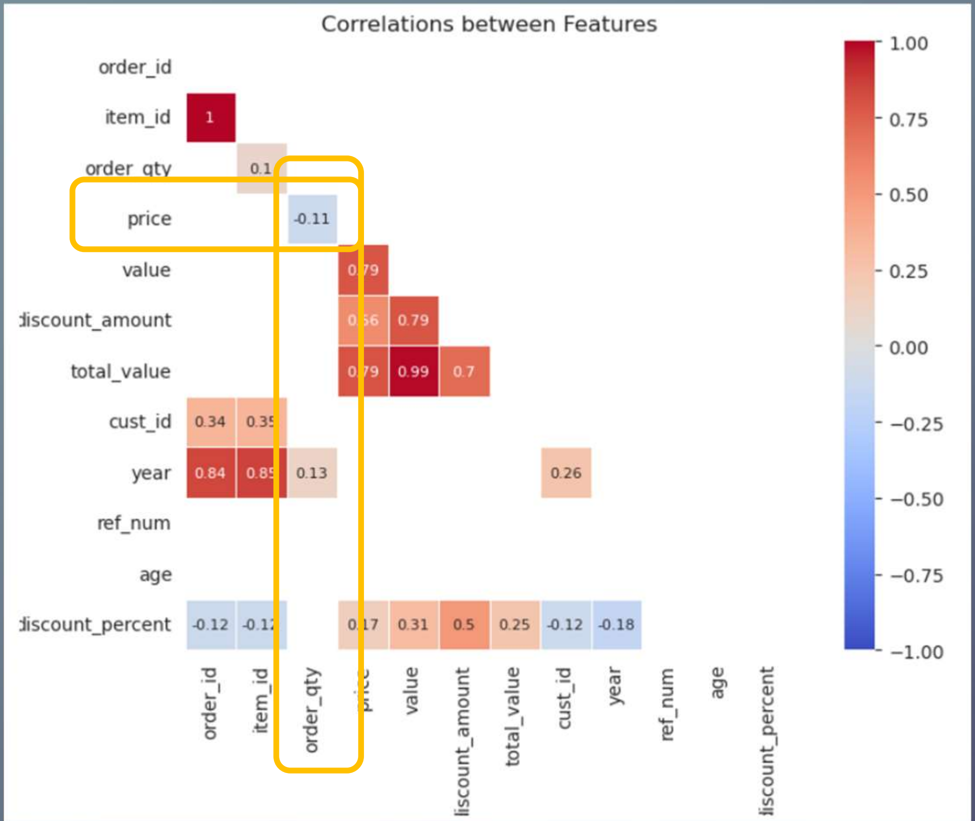
Day of the Week by Month (Heatmap)

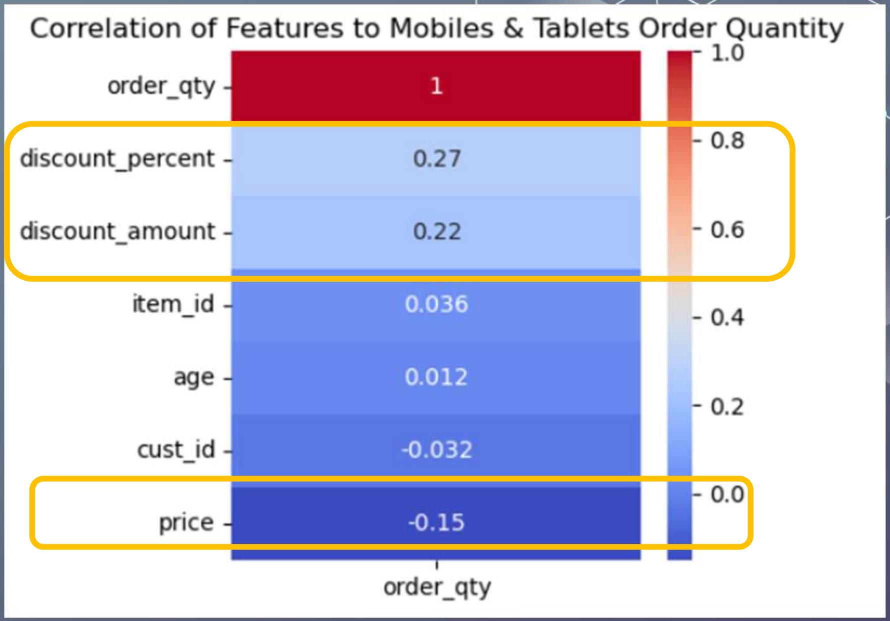
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
April	3267	1191	584	1891	491	550	614	635	533	578	792	488	290	612	428	2699	1938	1760	410	188	244	235	300	244	3523	1881	1755	205	166	418	0
August	0	0	0	0	3	8	10	6	4	6	4	2	10	0	0	6	0	0	0	0	0	6	0	6	9	2	0	0	57	0	2
December	62	94	141	108	288	857	178	94	114	116	137	220	2333	751	282	259	289	472	643	10981	5864	3352	4287	4445	3089	4124	10832	3525	3700	1200	585
February	167	132	648	190	256	195	1637	154	182	279	178	174	132	2743	139	80	180	192	240	150	508	201	78	212	124	144	139	182	0	0	0
January	484	852	743	264	317	406	355	384	1644	1997	820	323	891	509	941	307	782	351	138	206	160	200	187	579	149	105	137	114	168	91	515
July	2	2	0	2	12	2	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	2	2	0	0	0	0	0	0	8	0
June	100	188	207	235	141	856	228	169	36	0	3	3	3	0	2	0	0	4	2	0	0	0	1	0	2	4	8	0	26	8	0
March	4487	85	138	148	338	130	159	110	190	56	134	177	215	407	501	799	614	696	580	763	1259	1235	908	2796	999	943	1074	975	621	594	2995
May	197	155	188	199	131	193	324	200	237	248	191	172	190	171	174	173	126	4029	219	170	165	129	127	135	134	182	201	198	196	131	165
November	17	0	4	4	59	395	226	339	103	146	378	2185	132	364	175	135	76	121	80	123	335	522	97	75	101	154	180	354	459	72	0
October	217	222	357	93	131	101	99	220	126	107	81	110	68	75	77	88	264	114	95	50	125	102	130	280	333	101	45	83	110	90	92
September	0	0	0	0	0	2	0	0	2	0	9	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Correlation between day of month and month

20th-29th Dec
1st Mar
18th May

Correlation





- All category: order quantity vs price = weak negative correlation

- Mobiles & Tablets: order quantity vs discount – weak positive correlation
- Mobiles & Tablets: order quantity vs price – weak negative correlation

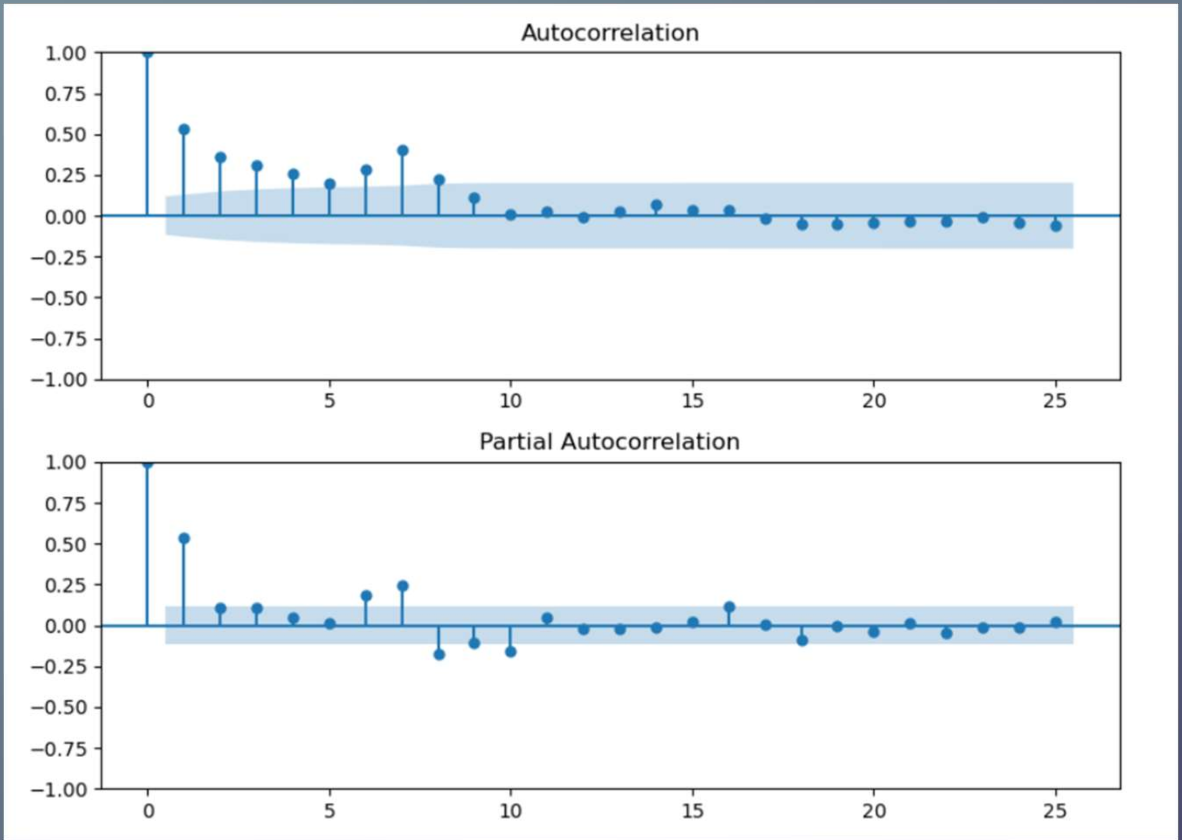
04

Model & Evaluation

Time Series Model

ARIMA
ARIMAX
SARIMA
SARIMAX

Stationarity, AR and MA

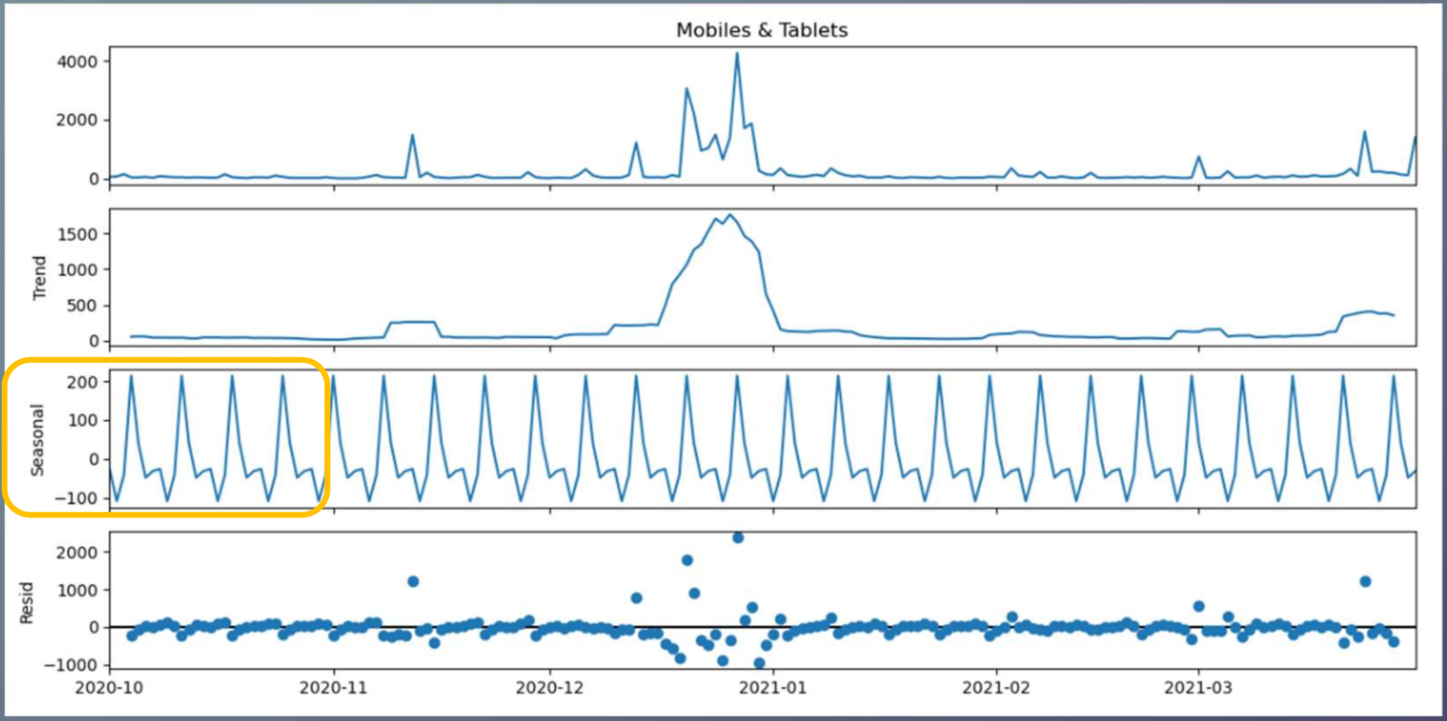


- Augmented Dickey–Fuller test (ADF Test)
- Kwiatkowski-Phillips-Schmidt-Shin test (KPSS Test)

- p-value for ADF test < 0.05 = reject null => stationary
- p-value for KPSS test > 0.05 = => stationary

- Autocorrelation (ACF) – lags not in negative zone – no differencing

Seasonality

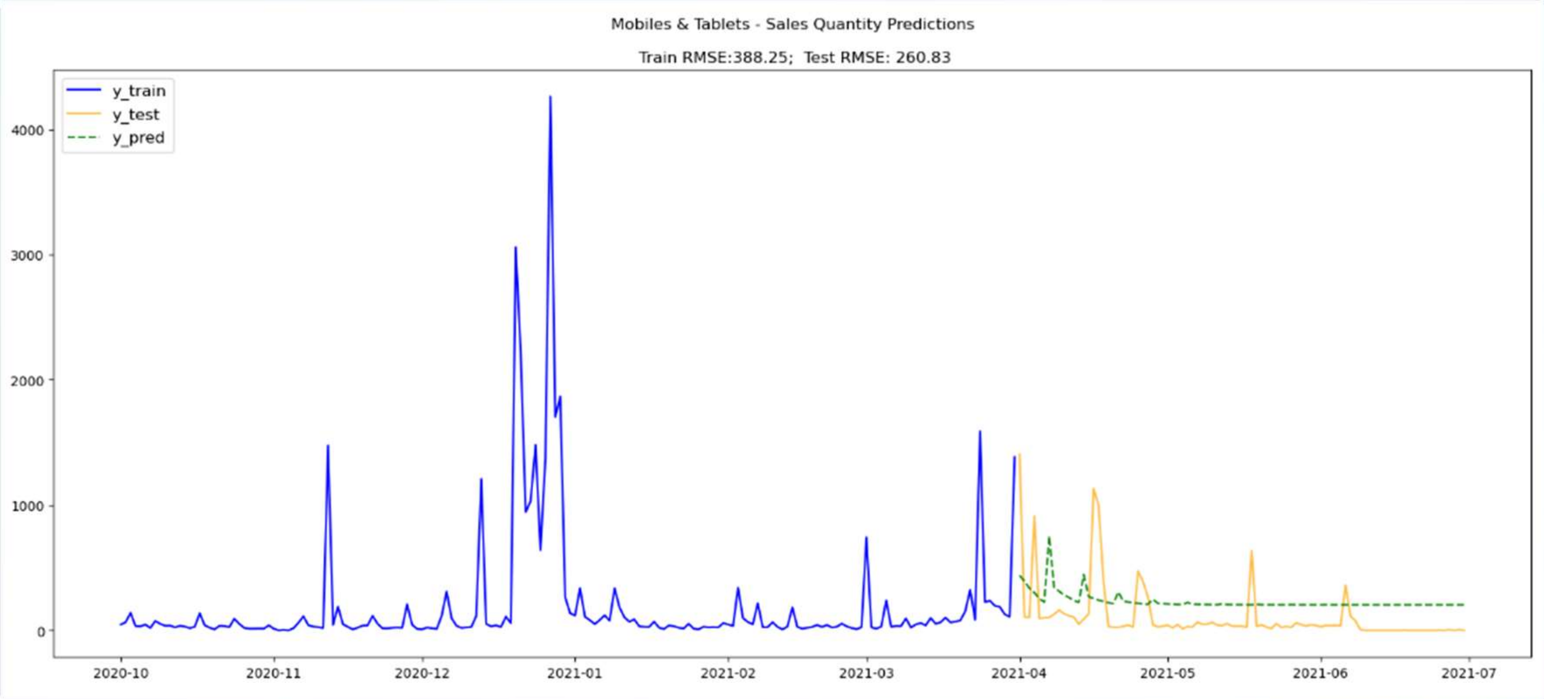


- Trend is not linear

- Seasonality within a month, each cycle lasts a week

- Residuals are high in Dec-2020

SARIMA

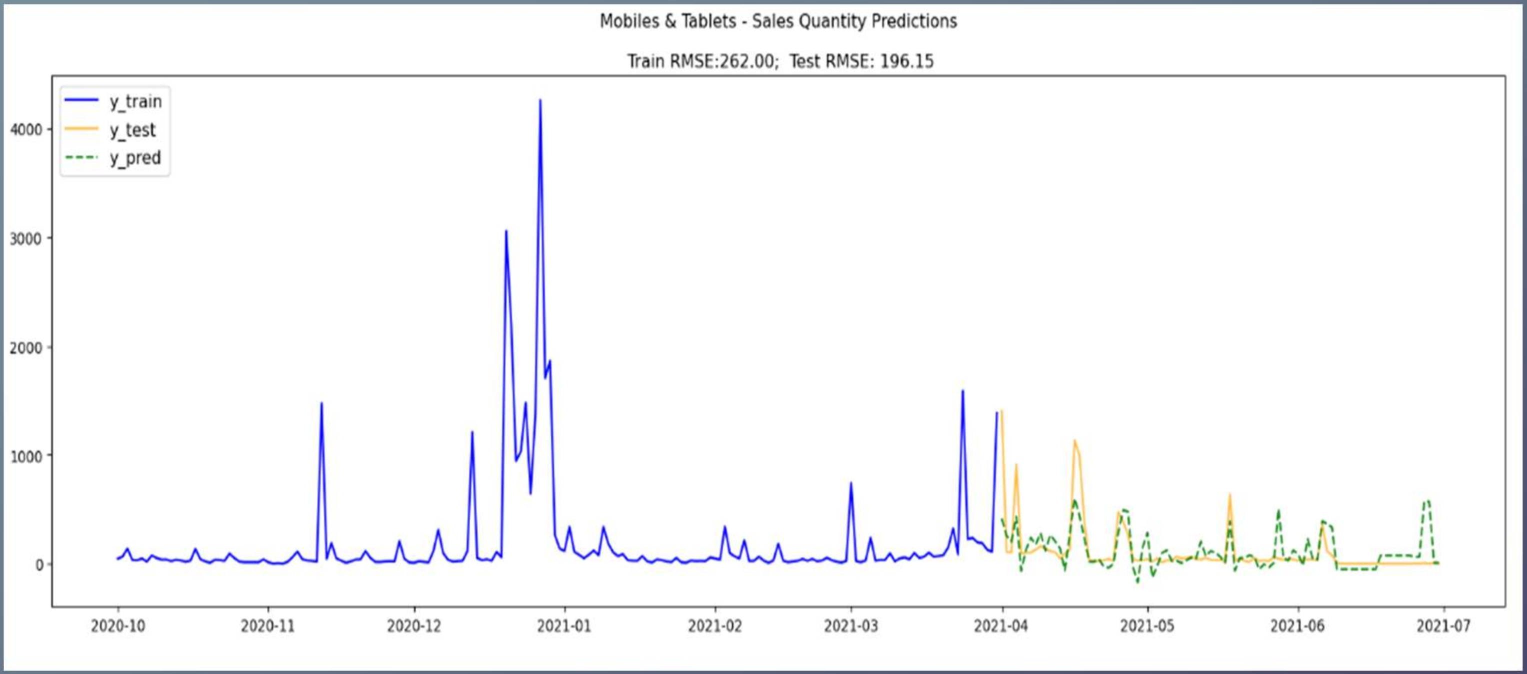


- SARIMA overfitting

- Train RMSE: 388
- Test RMSE: 261

- SARIMA not managed to predict well

SARIMAX



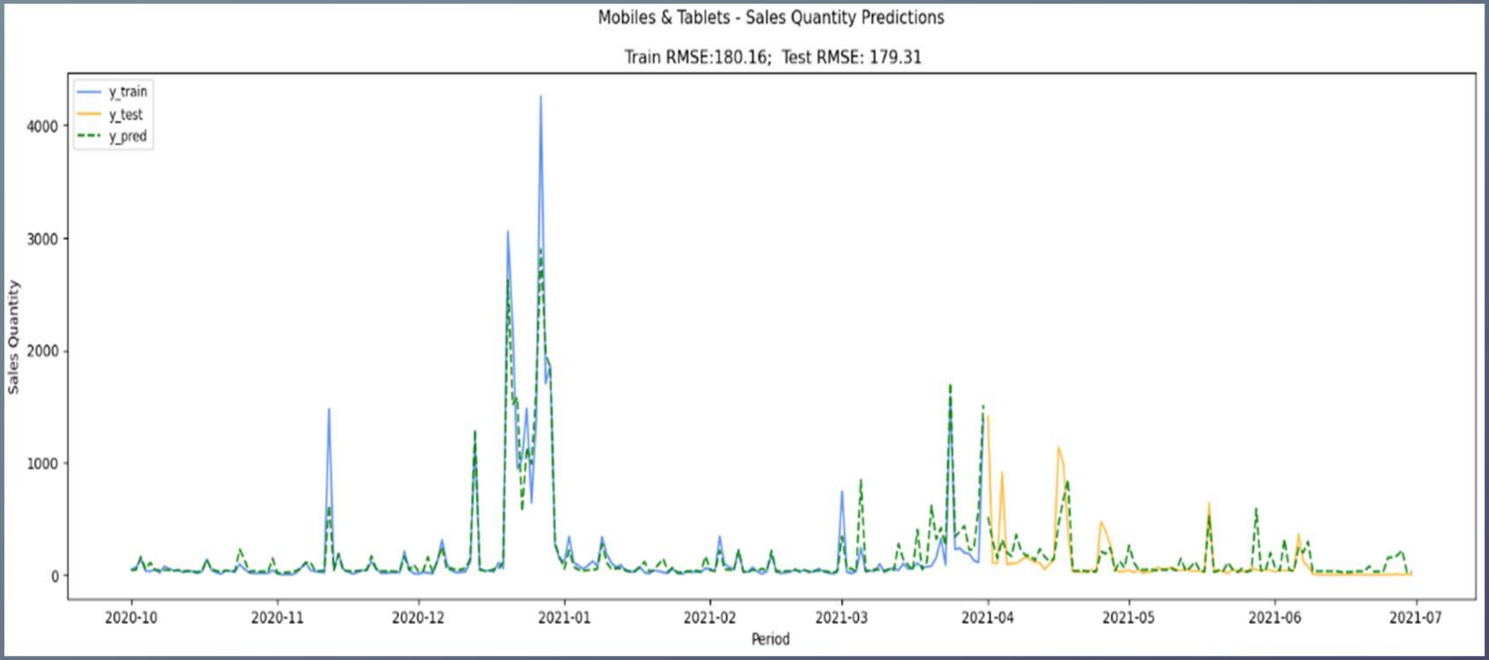
- Less overfitting with exogenous variables
- Train RMSE: 262
- Test RMSE: 196
- SARIMAX able to predict better than SARIMA

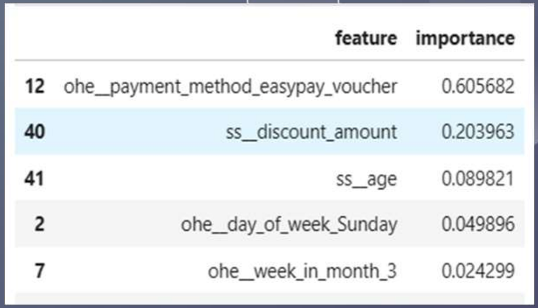
Regression Model

Random Forest

xTreme Gradient Boosting
(XGBoost)

Random Forest Model





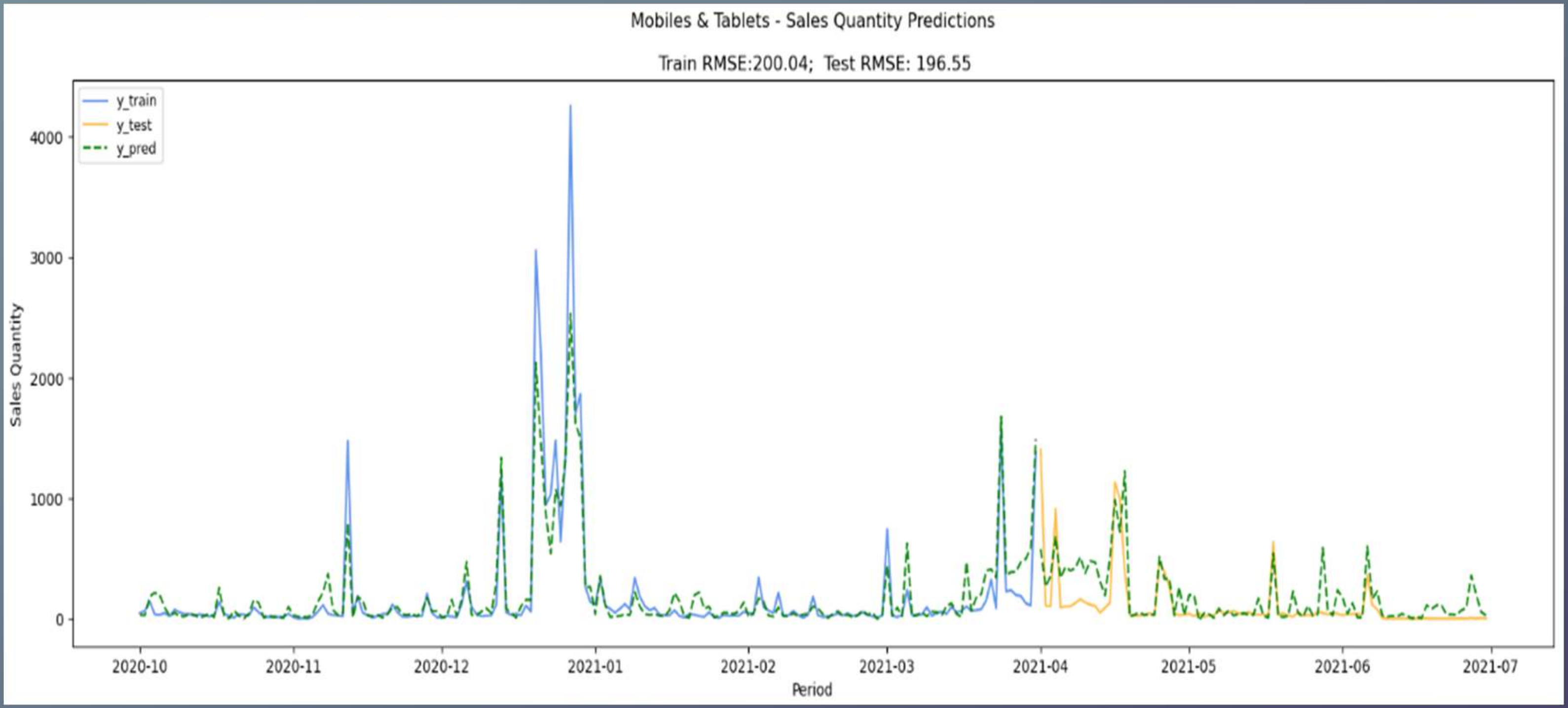
	feature	importance
12	ohe__payment_method_easypay_voucher	0.605682
40	ss__discount_amount	0.203963
41	ss__age	0.089821
2	ohe__day_of_week_Sunday	0.049896
7	ohe__week_in_month_3	0.024299

- Good fit
- Predicted trends very well
- Lowest RMSE

- Train RMSE: 180
- Test RMSE: 179

- Feature importance: payment method, discount, age, Sundays, week 3 of the month

XGBoost Model



- Good fit

- Train RMSE: 200
- Test RMSE: 197

- Regression models predicted trends better than time series models

Model Comparison

1

SARIMAX

Train RMSE: 262
Test RMSE: 196

2

Random Forest

Train RMSE: 180
Test RMSE: 179

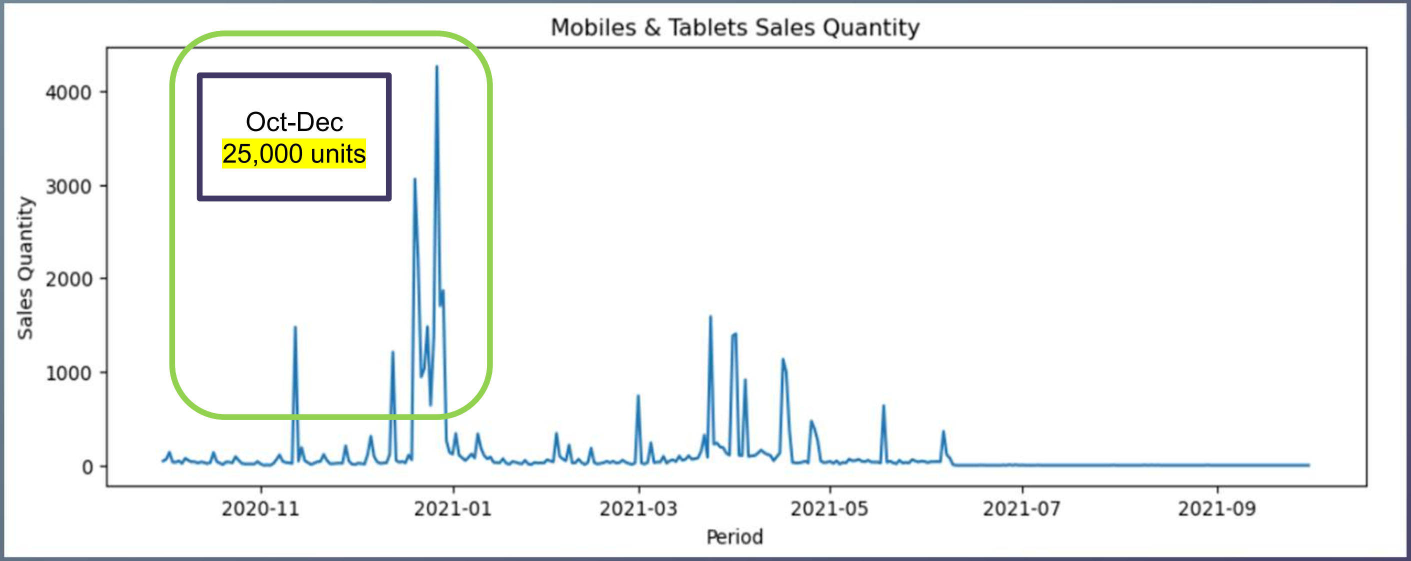
3

XGBoost

Train RMSE: 200
Test RMSE: 197

Cost-Benefit Analysis

- Mobiles & Tablets average price = $712
- Mobiles & Tablets margin = 43%
- Mobiles & Tablets average cost = $306 [$712 * 43%]
- Current state historical mean = 169 units
- Oct-Dec forecasted qty = 169 * 90 days = 15,210 units



Oct-Dec
25,000 units

- Loss of business = 10,139 units
- Product margin loss = $3,136,587 !!!
- Future business loss

04

Conclusion & Recommendation

Conclusion

Best Model

Random Forest

Good fit
Hyperparameter tuning
Feature importance

Importance of Forecasting

Loss of current orders

Loss of FUTURE business

Data

Quality & Quantity
Prediction Accuracy
Features Selection

Recommendations

Data Collection

More Historical Data
Product Attributes

Feature Engineering

Interactive terms
Market Basket Analysis
Holidays
Special Events

Model Selection

Different Models
Seasonality
Trends

The End

Thank you!