Anomaly Detection & Forecasting

Xiaoxin Xu

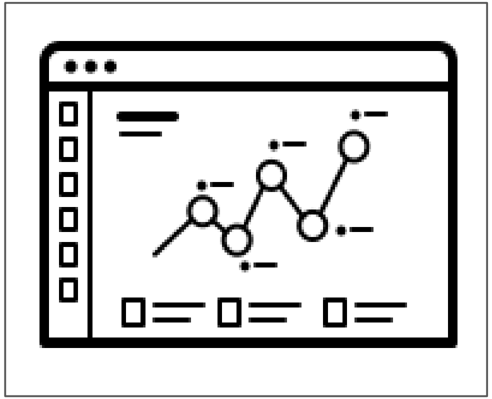
Executive Summary

Business Problem

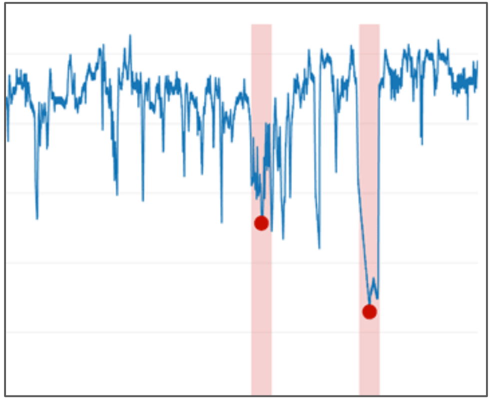
- A telecommunication major wanted to understand if there is a problem on its website
- Current process is manual and the expected state was to derive anomalous behavior across major KPIs in an automated manner

Solution Approach

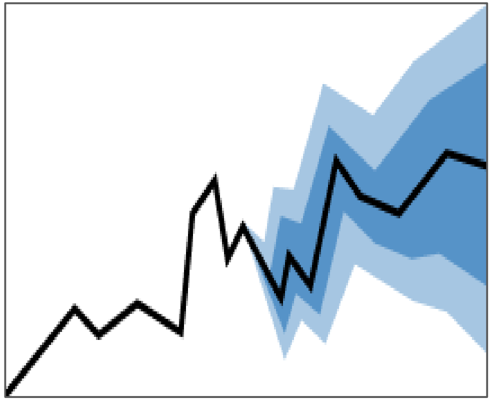
Identify KPIs → Anomaly Detection → Forecasting



Identify main KPIs: Visitors, Orders

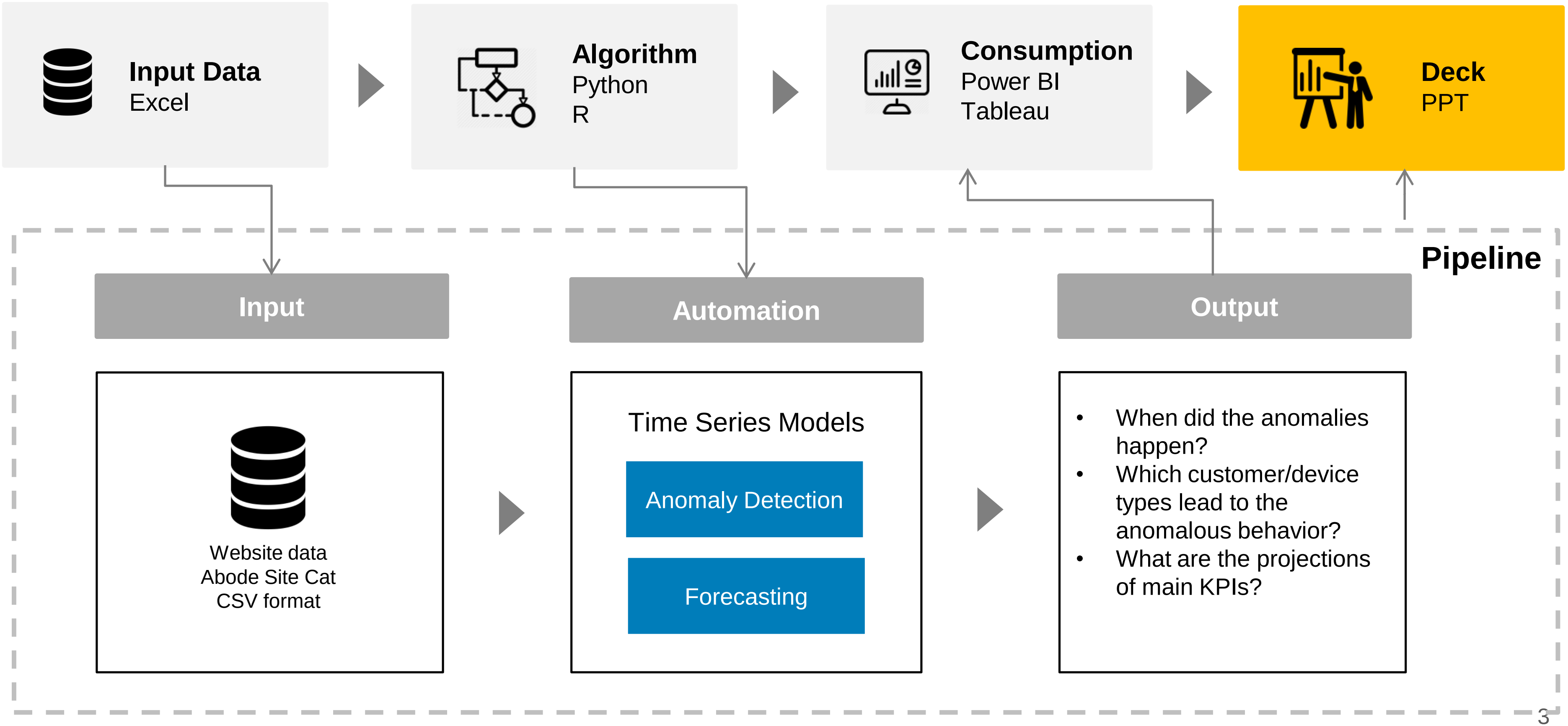


Find good and bad anomalies across key KPIs



Predict the following 7days and find outliers in the forecast

Architecture



Data & Statistics

59 Days Feb 1 – Mar 31	38M Total Visitors	369K Total Orders	0.96% Avg Conversion Rate Avg 3.33% in Telecom*

	Visitor			Order			Rate*		
	Desktop	Mobile	Tablet	Desktop	Mobile	Tablet	Desktop	Mobile	Tablet
Customer	7.6M	4.8M	0.78K	198K	49K	18K	2.59%	1.02%	2.39%
Prospect	1M	1M	0.18M	5K	5K	473	0.51%	0.45%	0.26%
Undetermined	9.3M	11.8M	1.3M	59K	27K	5K	0.63%	0.23%	0.39%

* Conversion rate = #Orders / #Visitors

* Visitors and orders from Gaming Console and E-Reader are low so they are excluded from analysis

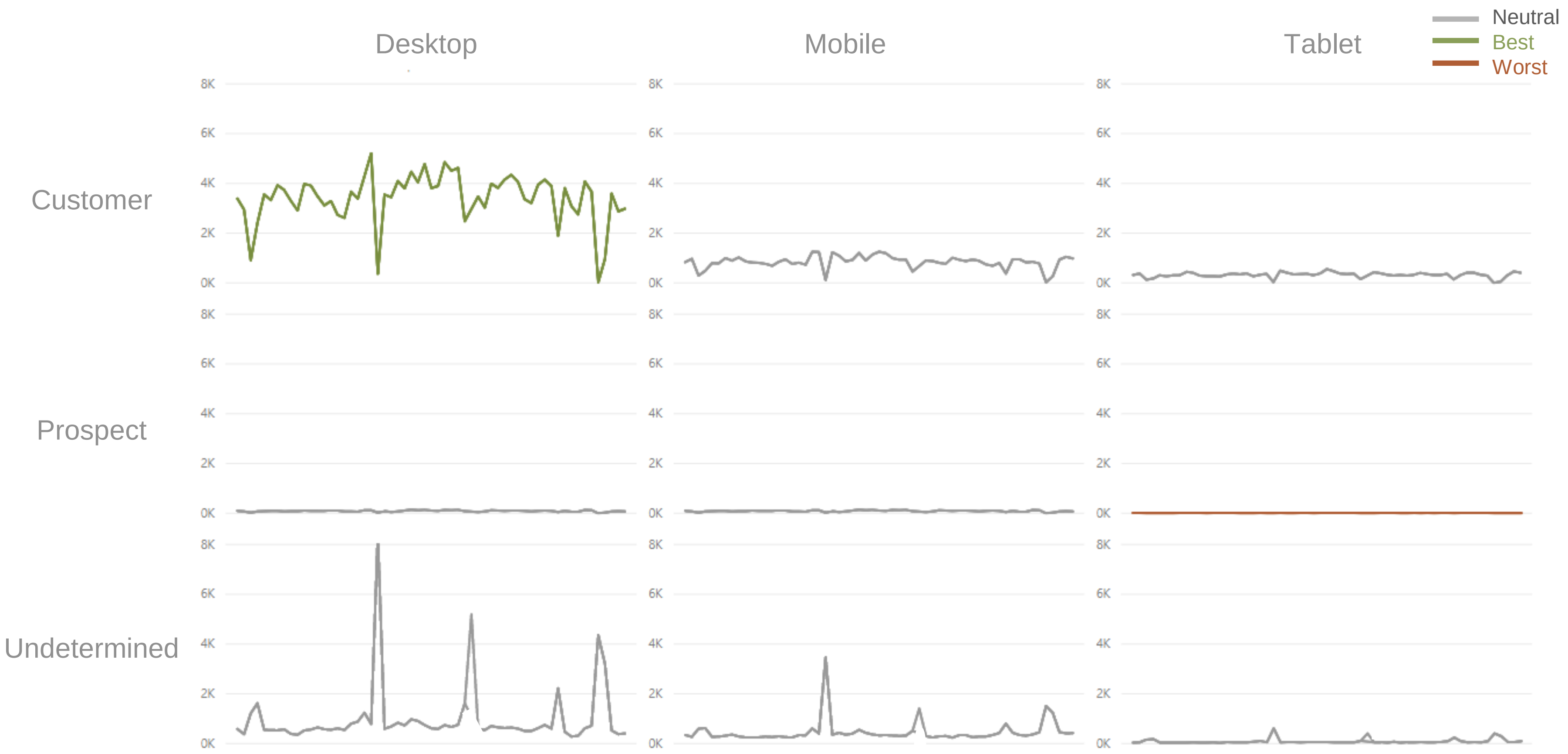
* Reference: https://www.wordstream.com/blog/ws/2018/08/13/google-ads-mobile-benchmarks

Exploratory Analysis - Visitor

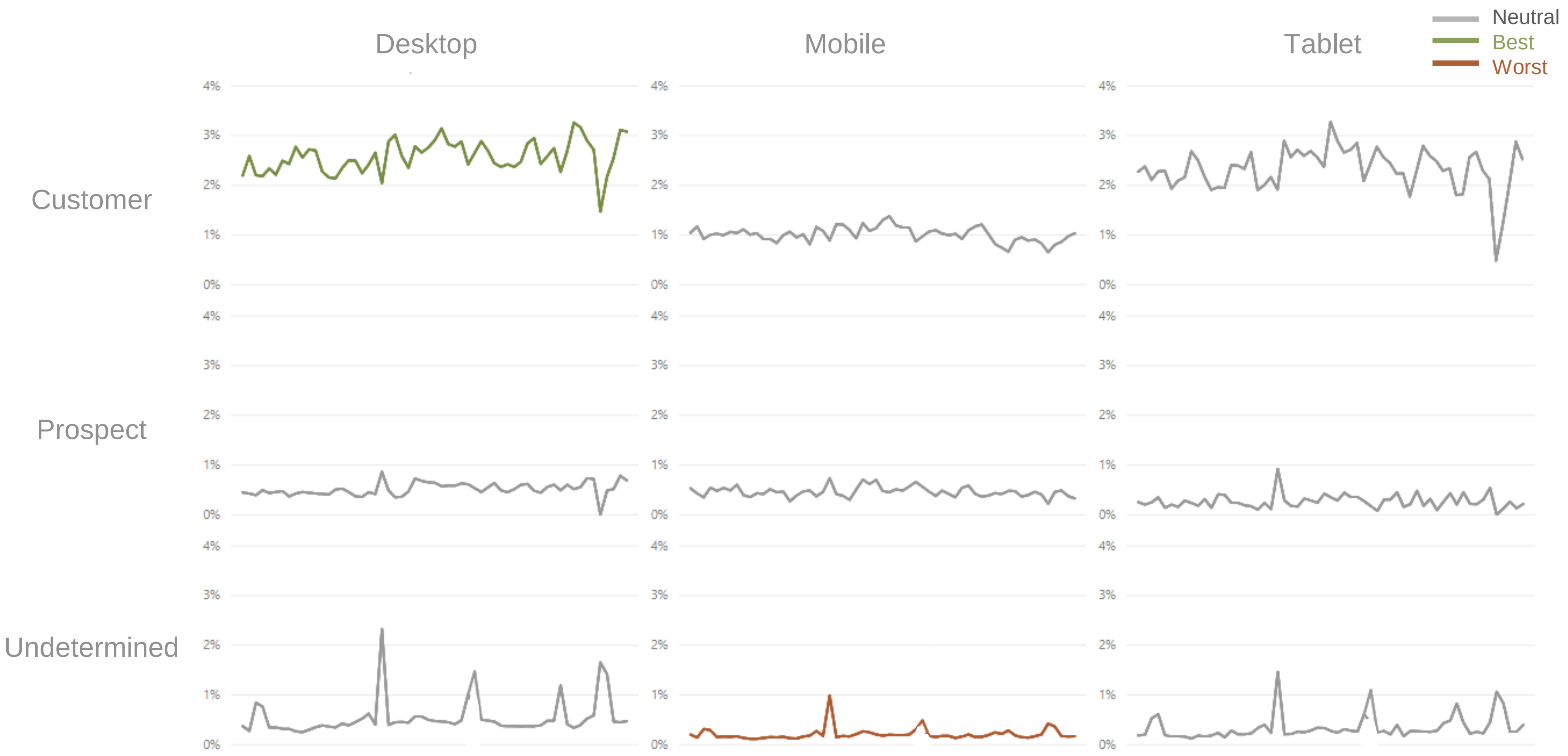
Customer tagging problem

Exploratory Analysis - Visitor

Exploratory Analysis - Order

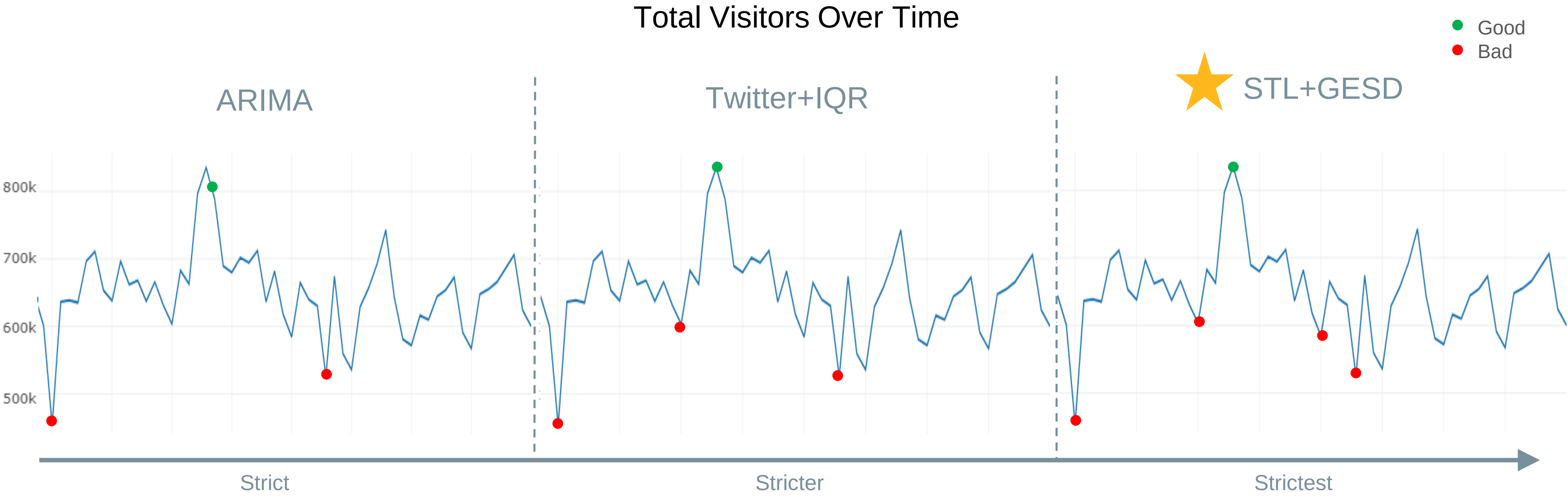


Exploratory Analysis - Rate



Anomaly Detection

Model Selection



	ARIMA	Twitter+IQR	STL+GESD
Pros	• Well-studied • Parameters(p, d, q) -> more accurate	• Catch seasonal component • No loops – easily scaled	• More sensitive to capture anomalies • Less resistant to outliers
Cons	• Stationarity assumption • Parameter tuning	• No long term trend • IQR is easily skewed by outliers	• Iterative – more expensive

Two Callouts

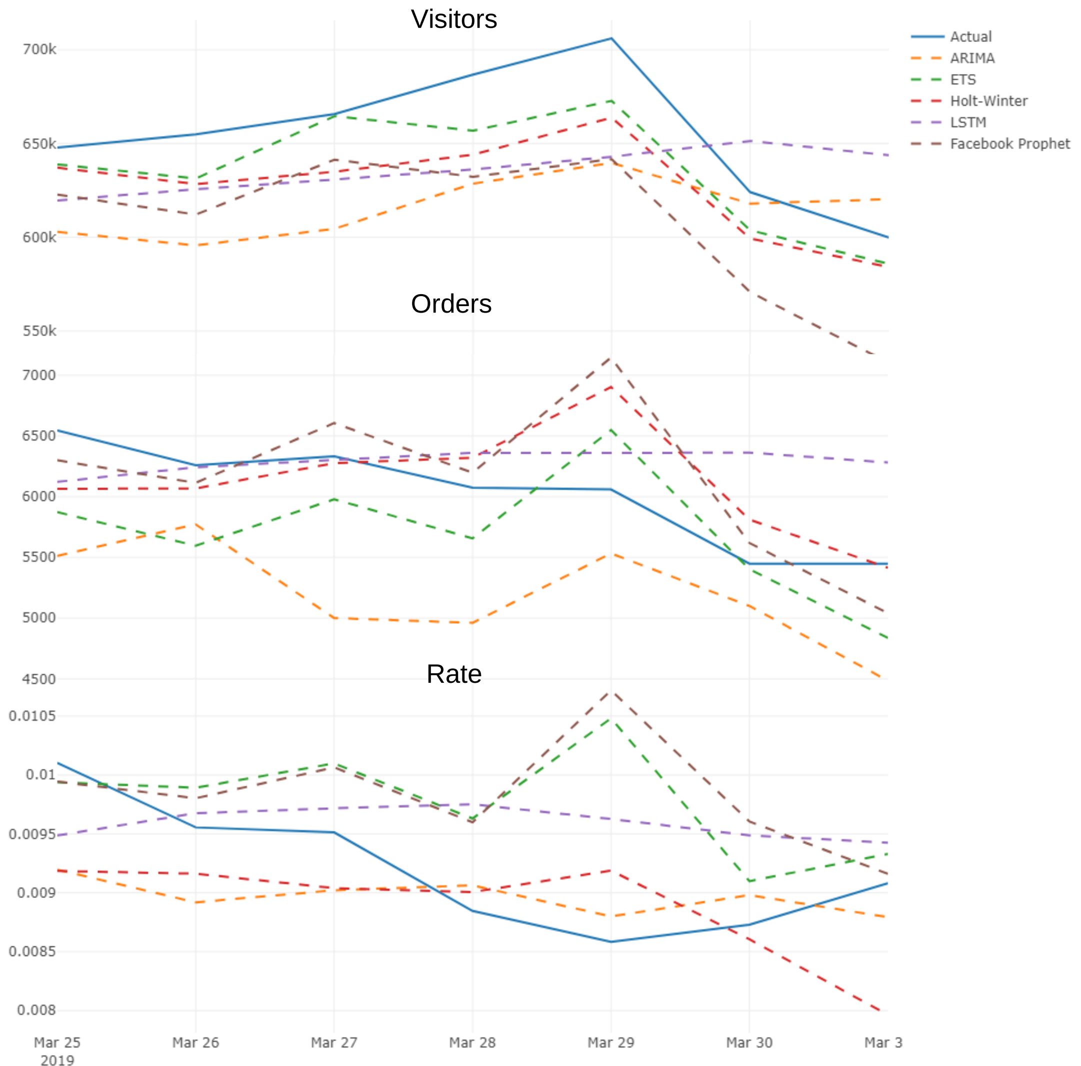
Feb 3rd
All metrics show bad anomalies

Feb 22nd
All metrics show good anomalies

Forecasting

Model Selection - Forecasting

Pros	Cons
ARIMA	
• Parameters(p, d, q) -> more accurate	• Model assumption – stationary • Parameter tuning
ETS	
• No requirement for stationarity	• Parameter selection
Holt-Winter	
• Triple exponential smoothing	• Model complexity • Overfitting
LSTM	
• Time dependency • Perform well with large data	• Hyperparameter Tuning • Require large data
Facebook Prophet	
• Seasonal effects • Robust to missing data and outliers	• Overfitting • Require large data



Metric Evaluation

	Visitors			Orders			Rate		
	RMSE	MAPE	MASE	RMSE	MAPE	MASE	RMSE	MAPE	MASE
ARIMA	132365.1	7.3%	0.5	2376.4	16.3%	1.6	1.5e3	5.1%	0.5
ETS	66286.1	2.9%	0.2	1343.5	8.2%	0.9	2.2e3	6.3%	0.7
HW	78978.5	4.4%	0.3	1093.8	5.0%	0.3	1.7e3	6.2%	0.7
LSTM	109966.7	6.2%	0.5	1373.2	6.4%	0.8	1.7e3	5.9%	0.6
FB	132491.2	7.9%	0.5	1242.6	5.5%	0.6	2.5e3	7.7%	0.8

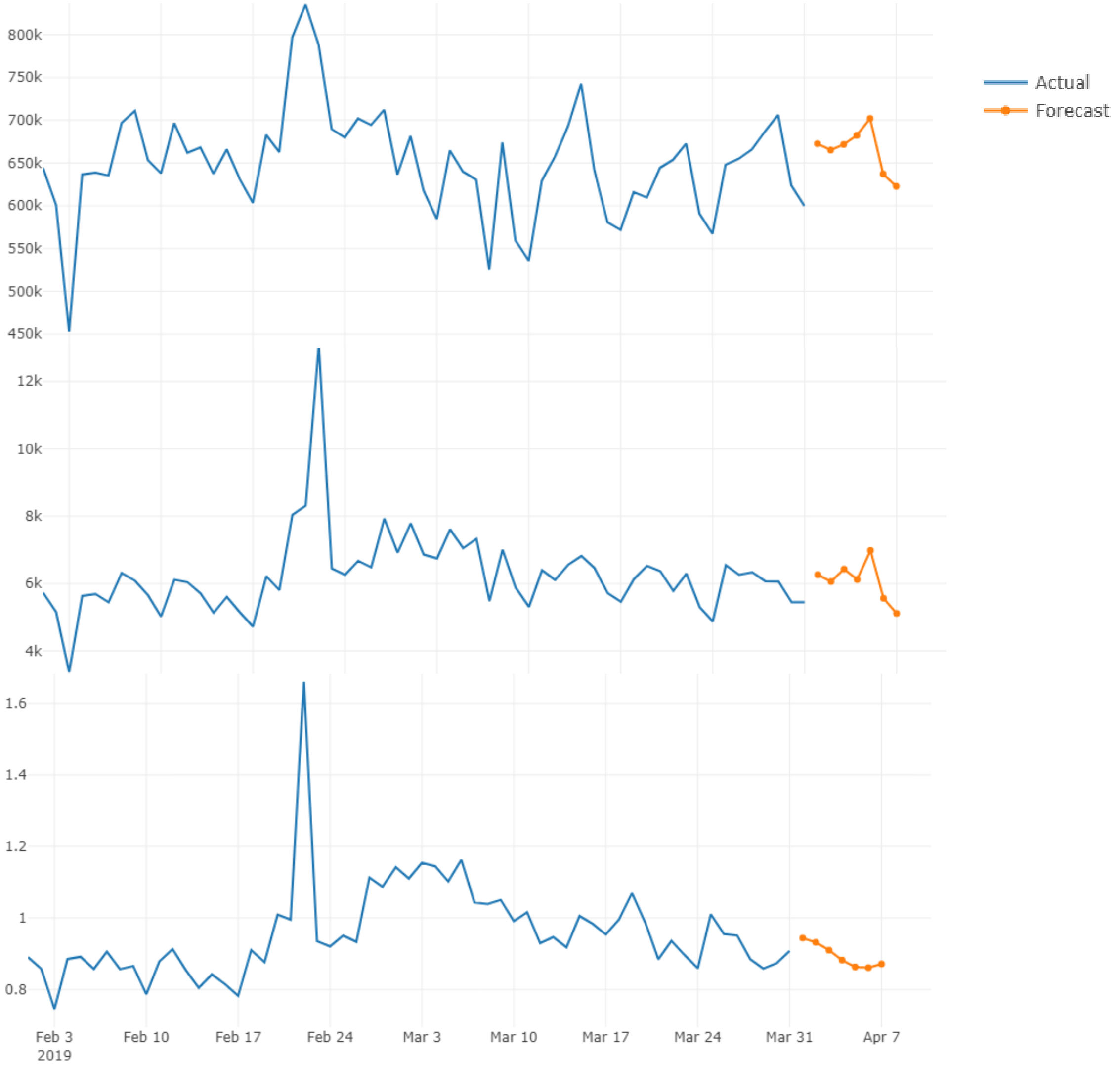
* See metrics across customer and device type in appendix

Forecast

Visitors

Orders

Rate

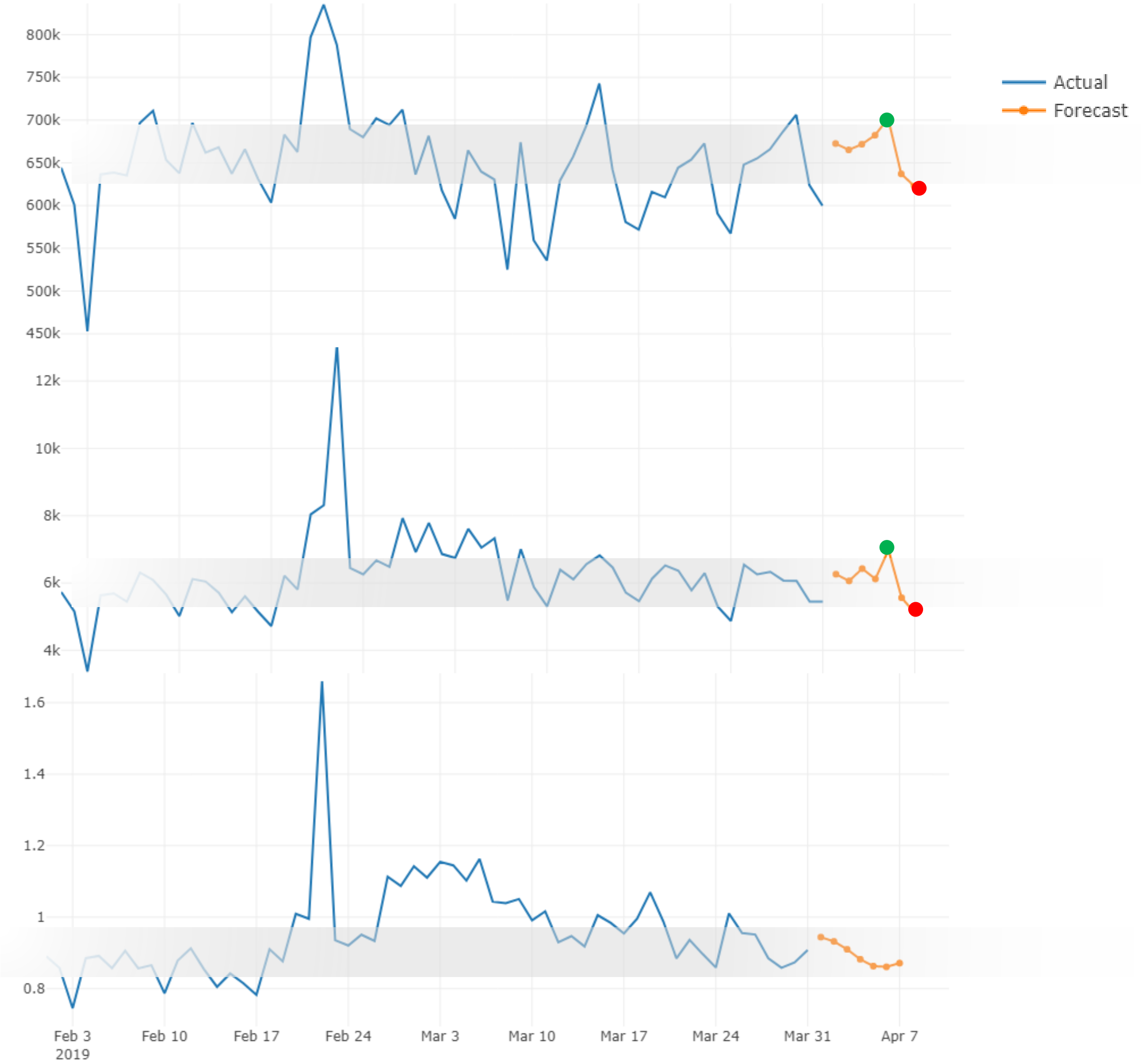


Outliers in the forecast?

Visitors

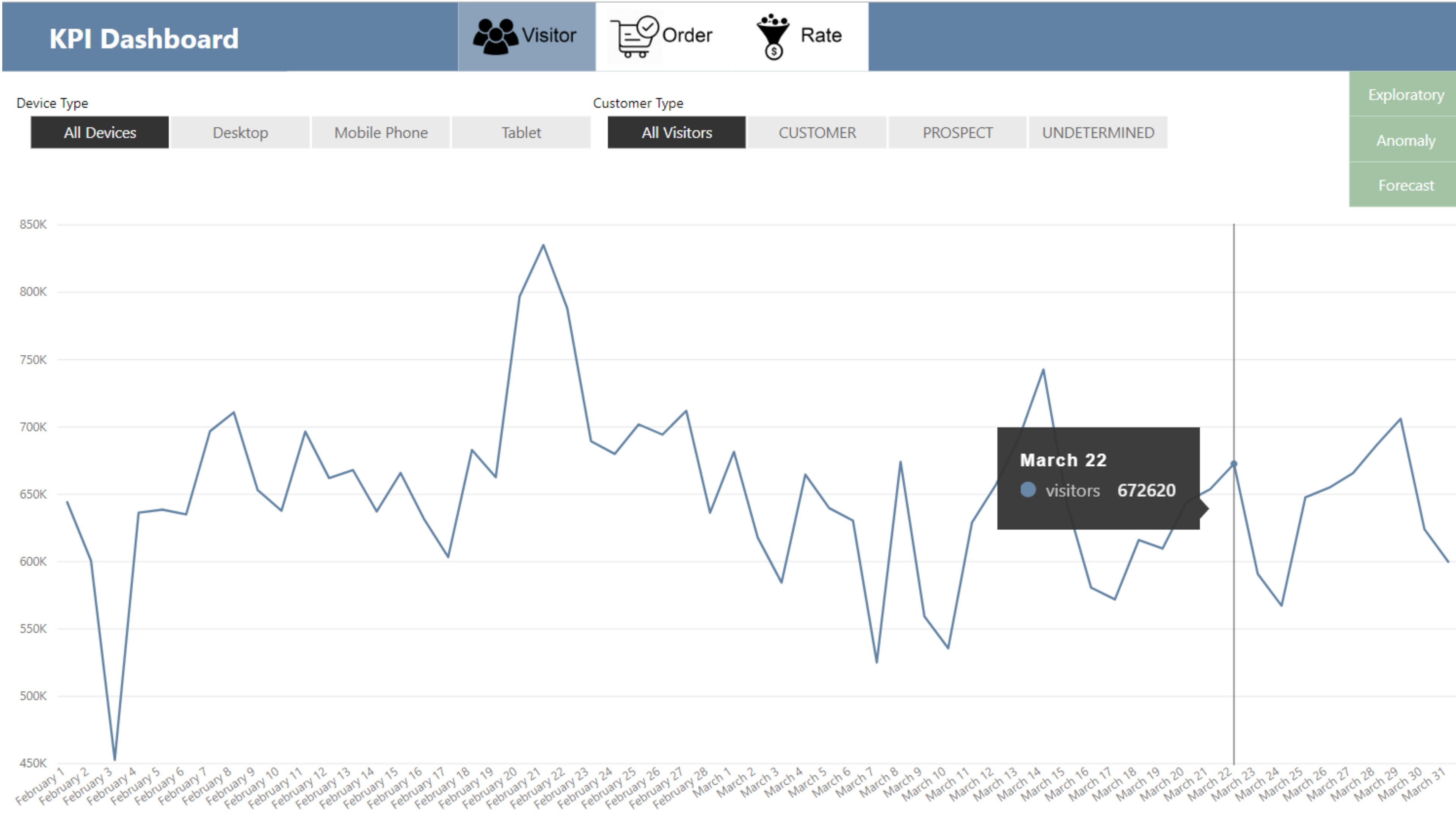
Orders

Rate



* Outliers: points lie out of ±1.5 standard deviation

KPI Dashboard Demo


KPI Dashboard
Visitor
Order
Rate
Device Type
All Devices
Desktop
Mobile Phone
Tablet
Customer Type
All Visitors
CUSTOMER
PROSPECT
UNDETERMINED
Exploratory
Anomaly
Forecast
850K
800K
750K
700K
650K
600K
550K
500K
450K
March 22
visitors 672620
February 1
February 2
February 3
February 4
February 5
February 6
February 7
February 8
February 9
February 10
February 11
February 12
February 13
February 14
February 15
February 16
February 17
February 18
February 19
February 20
February 21
February 22
February 23
February 24
February 25
February 26
February 27
February 28
March 1
March 2
March 3
March 4
March 5
March 6
March 7
March 8
March 9
March 10
March 11
March 12
March 13
March 14
March 15
March 16
March 17
March 18
March 19
March 20
March 21
March 22
March 23
March 24
March 25
March 26
March 27
March 28
March 29
March 30
March 31

Appendix

Metric Evaluation – using selected model

	Visitor ETS			Order Holt-Winter			Rate ARIMA		
	RMSE	MAE	MASE	RMSE	MAE	MASE	RMSE	MAE	MASE
All	66286.08	20686.66	0.235732	1093.834	310.1625	0.269539	0.001493	0.000511	0.832543
Customer x Desktop	73263.96	23571.93	0.899388	865.6737	253.26	0.922383	0.00494	0.001532	1.647831
Customer x Mobile	86510.63	30973.98	0.727294	3745.477	1068.289	0.942647	0.014045	0.004057	0.680417
Customer x Tablet	9435.26	3168.783	0.713414	353.9444	89.99719	0.548286	0.022752	0.006194	0.77245
Prospect x Desktop	80504.49	24909.79	0.497514	1274.202	336.9946	0.894224	0.002757	0.000761	0.783244
Prospect x Mobile	130128.7	39952.52	0.617765	3954.163	1001.093	0.876723	0.014254	0.003275	0.817562
Prospect x Tablet	19274.86	6028.2	1.240991	370.9294	95.74592	1.012419	0.008957	0.002382	0.954046
Undetermined x Desktop	13608.78	4135.544	0.465872	112.8573	32.89702	0.903055	0.002962	0.00083	1.17678
Undetermined x Mobile	14486	5197.83	1.120878	115.6508	36.62473	0.989858	0.006676	0.001979	1.013532
Undetermined x Tablet	3193.753	927.2324	0.726834	13.16669	4.430497	0.939802	0.004344	0.00128	1.045439

Visitors

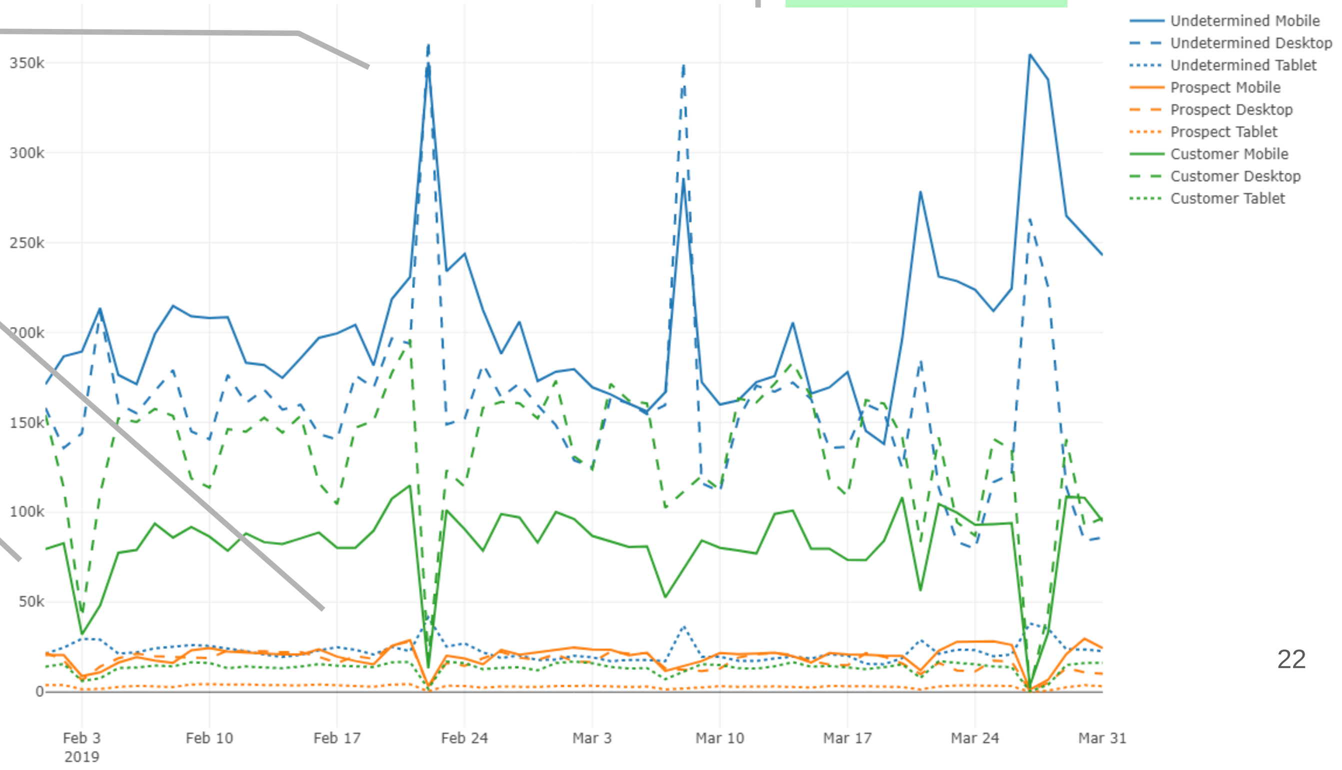
customer tagging issue?

2/22
Undetermined ↑
Customer ↓
Prospect ↓

2/17 All ↓
2/28 All ↓
3/7 All ↓

2/3
Customer ↓
Prospect ↓

3/8 & 3/27
Undetermined ↑
Customer ↓
Prospect ↓

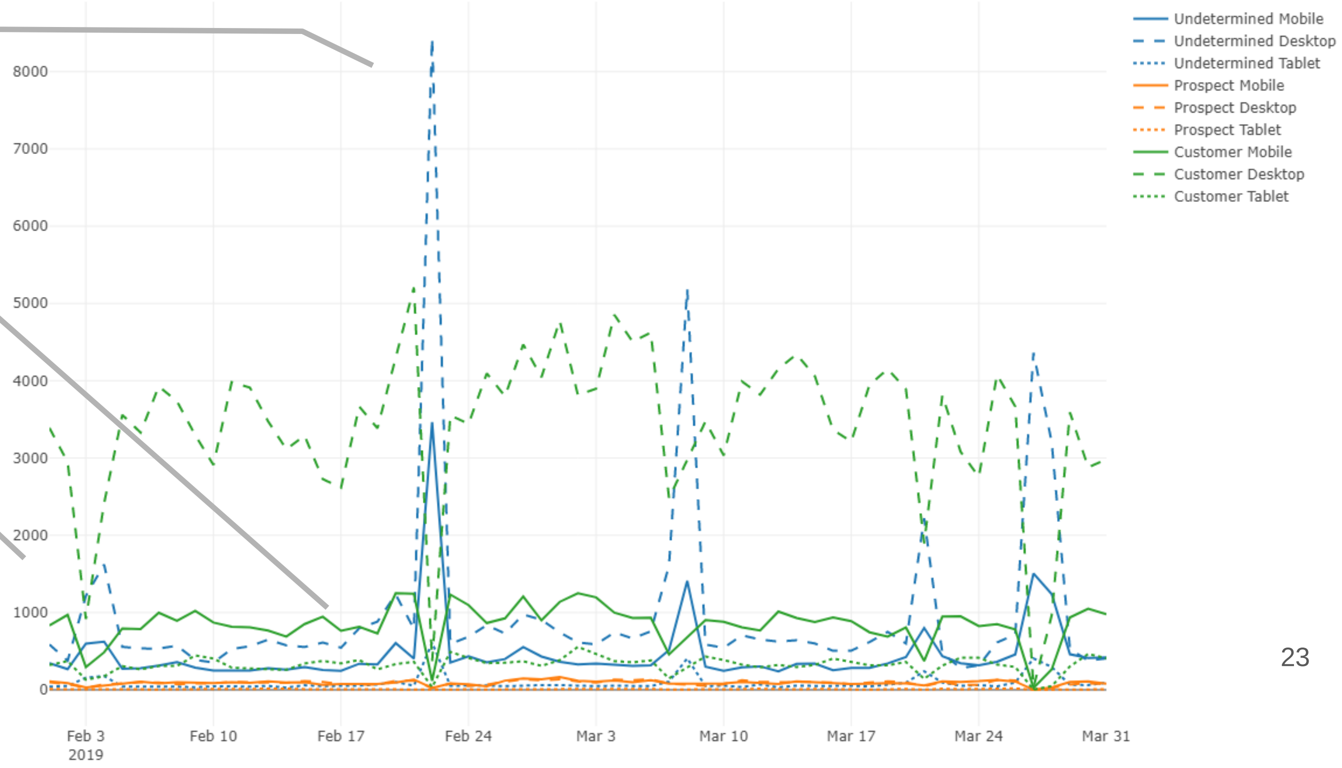


Orders

2/22
Undetermined ↑
Customer ↓
Prospect ↓

2/17 All ↓
2/28 All ↓
3/7 All ↓

2/3
Customer ↓
Prospect ↓

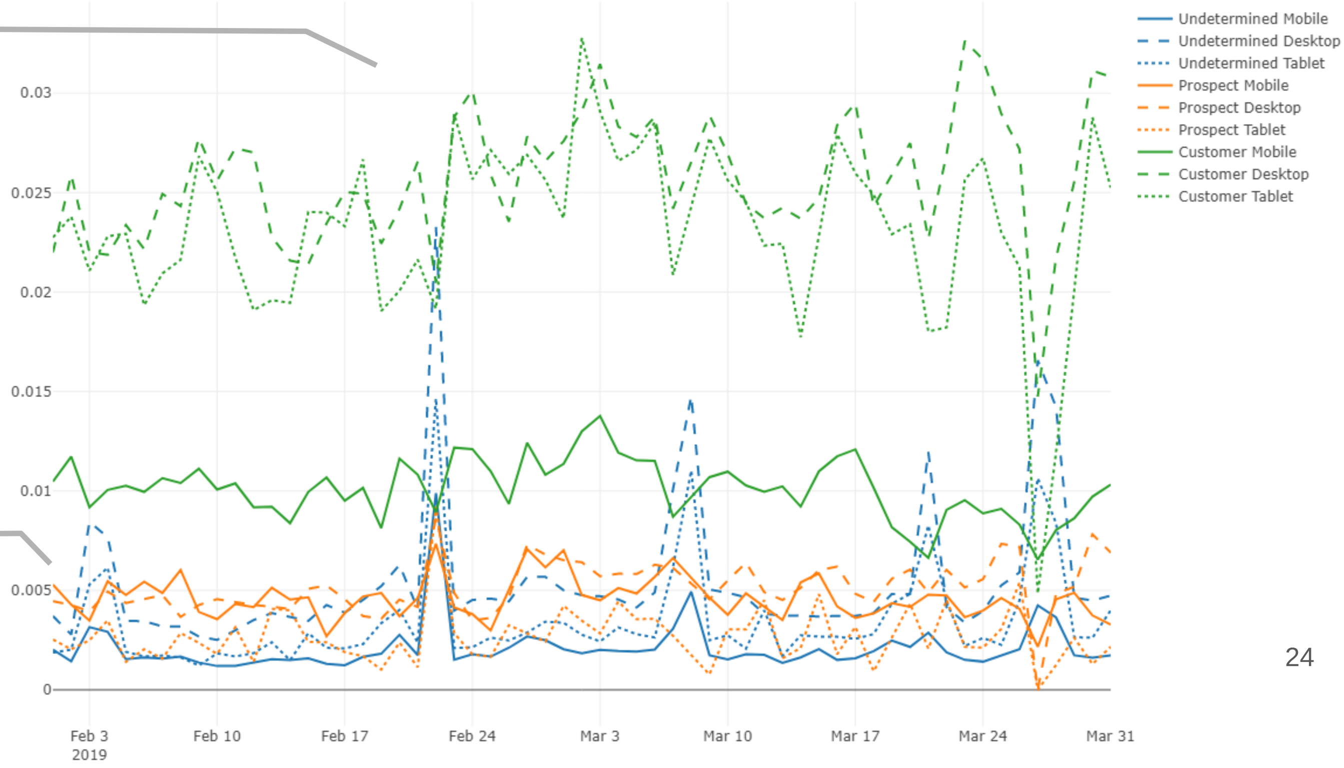


Conversion Rate

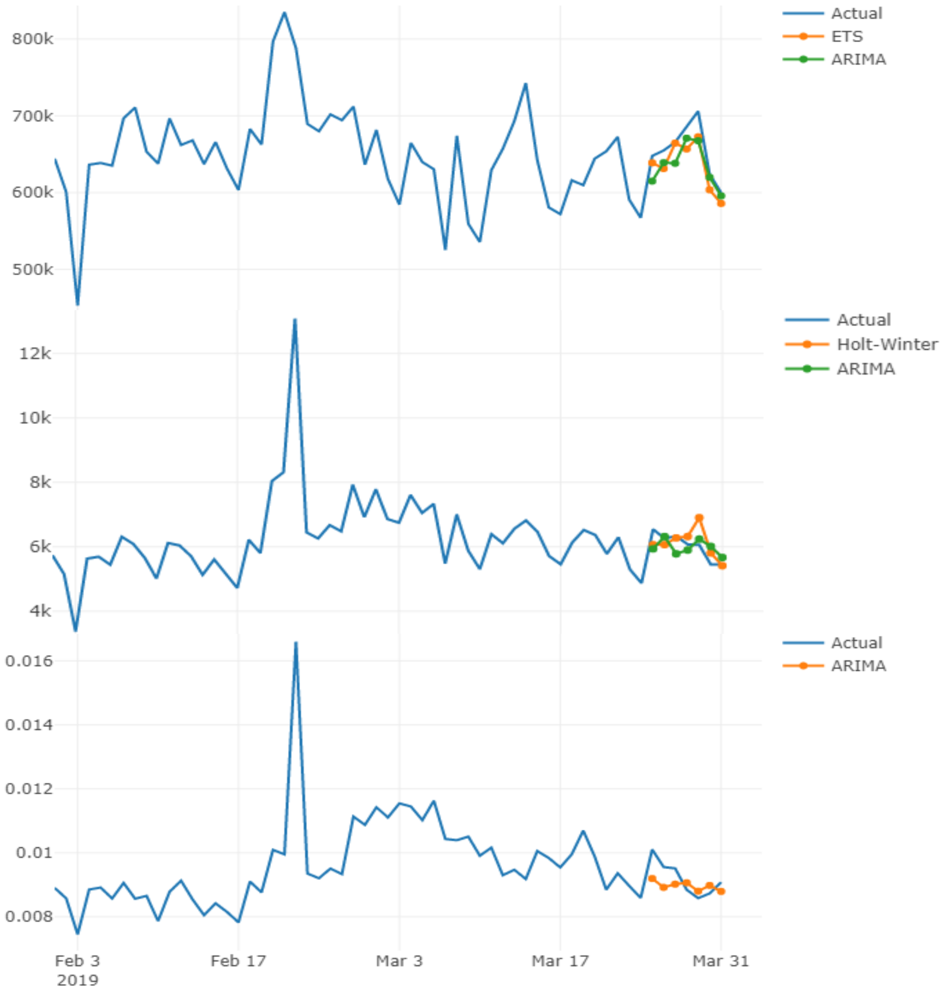
3/27
Undetermined ↑
Customer ↓
Prospect ↓

2/22
Undetermined ↑
Customer ↓
Prospect ↑

2/3
Undetermined ↑
Customer ↓
Prospect ↓



ARIMA outperforms with proper tuning



Anomaly Detection on Predictions