Data Science Project

10/17/2016

Goal of the Project:

- Company XYZ is a food delivery company. They have been relying significantly on online ads.

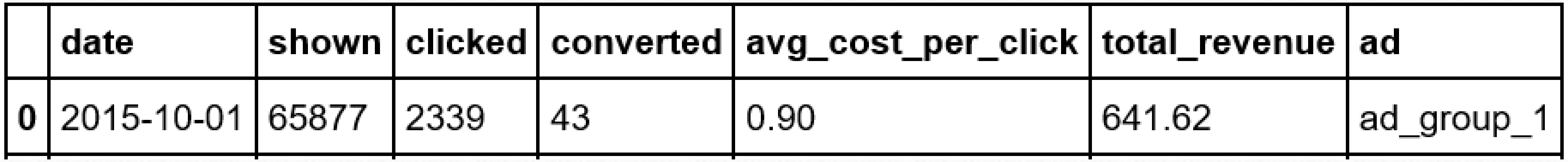
- At the moment, they are running 40 different ad campaigns and would like to understand their performance.

Specific Tasks:

- Identify the 5 best ad groups
- Predict how many ads will be shown on Dec, 15 for each campaign
- Cluster ads into 3 groups based on the trend of avg_cost_per_click

Data Exploration:

- Data Shape: (2115,7)



	date	shown	clicked	converted	avg_cost_per_click	total_revenue	ad
0	2015-10-01	65877	2339	43	0.90	641.62	ad_group_1

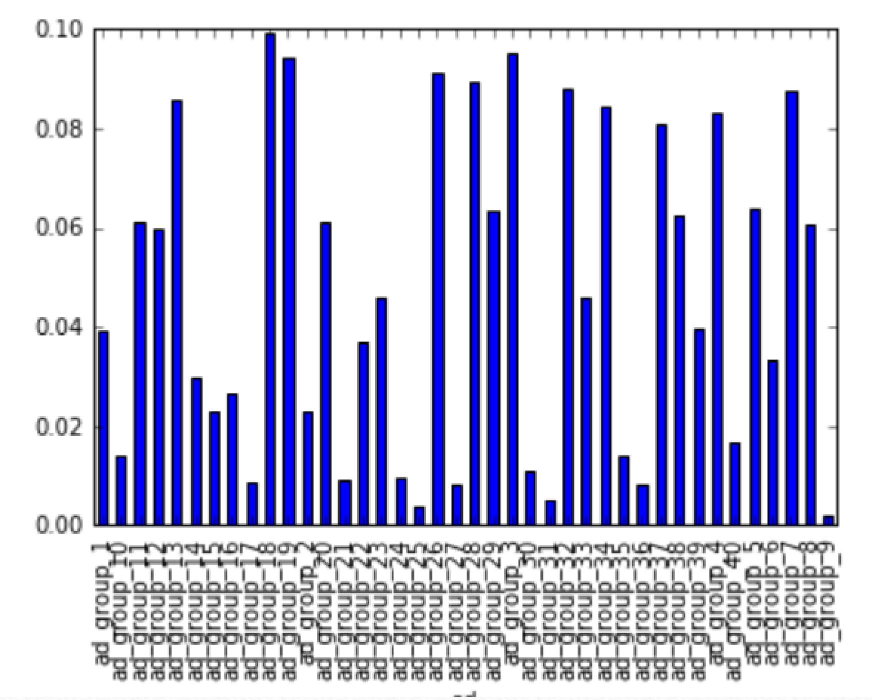
- Missing Values
- Outlier

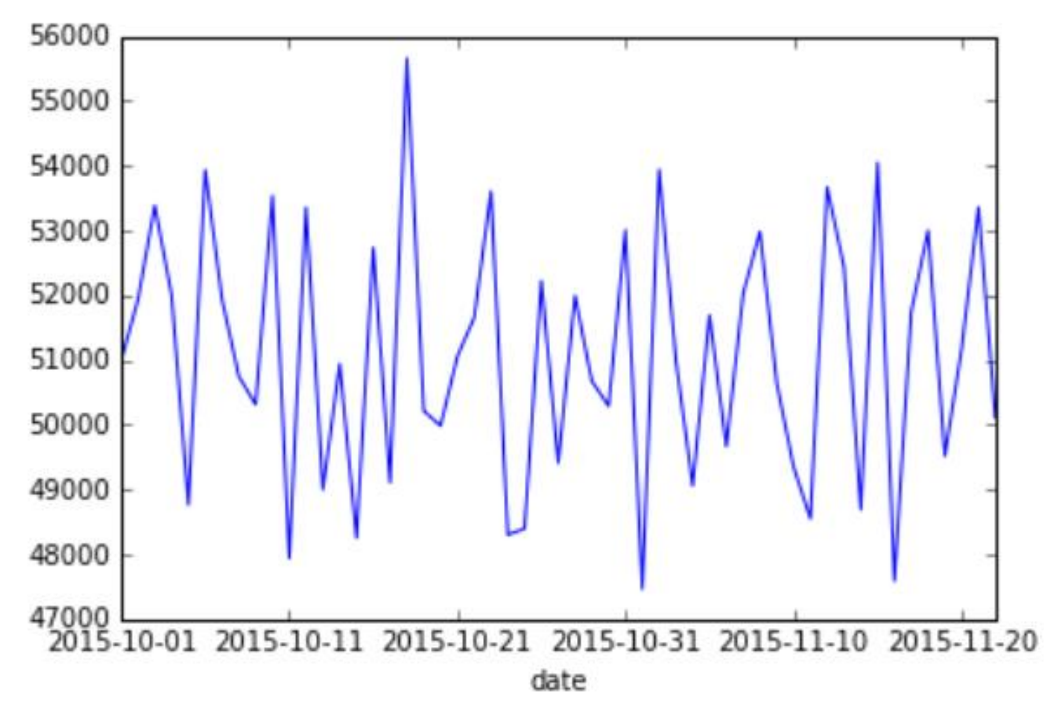
Task 1: Identify the 5 Best Ad Groups

- Metric:
  - CRT = $\frac{Clicks}{Impressoin} * 100\%$

- Measures how many people see your ads and whether they engage with the ads

Pros & Cons:

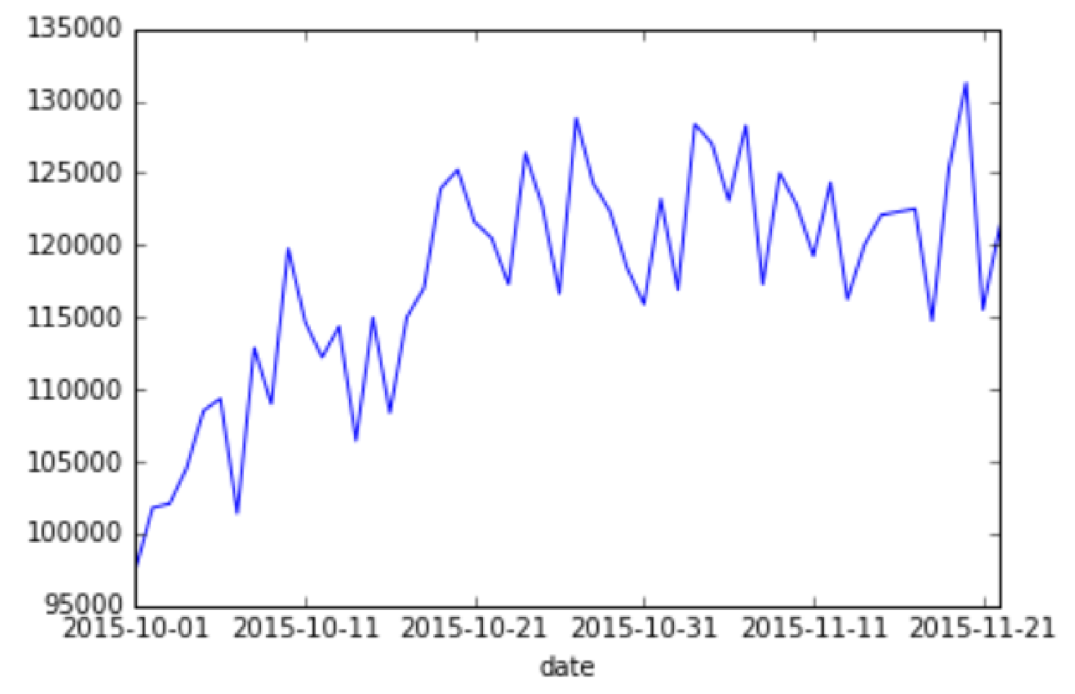

0.10
0.08
0.06
0.04
0.02
0.00
ad_group_1
ad_group_10
ad_group_11
ad_group_12
ad_group_13
ad_group_14
ad_group_15
ad_group_16
ad_group_17
ad_group_18
ad_group_19
ad_group_2
ad_group_20
ad_group_21
ad_group_22
ad_group_23
ad_group_24
ad_group_25
ad_group_26
ad_group_27
ad_group_28
ad_group_29
ad_group_3
ad_group_30
ad_group_31
ad_group_32
ad_group_33
ad_group_34
ad_group_35
ad_group_36
ad_group_37
ad_group_38
ad_group_39
ad_group_4
ad_group_40
ad_group_5
ad_group_6
ad_group_7
ad_group_8
ad_group_9


56000
55000
54000
53000
52000
51000
50000
49000
48000
47000
2015-10-01
2015-10-11
2015-10-21
2015-10-31
2015-11-10
2015-11-20
date

Task 2:Predict #of ads to be shown on Dec, 15

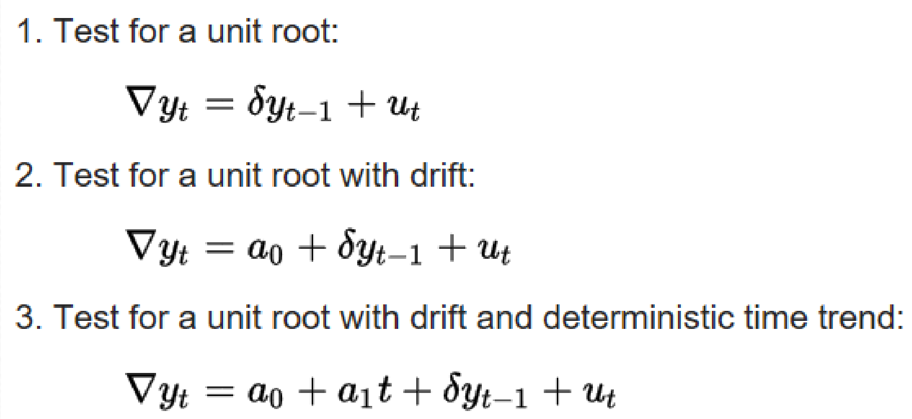
- Data type
  - Cross-sectional
  - Time series
  - Panel data

Plot Series Data



Statistical Tests Performed:

- Augmented Dickey–Fuller test
- adf=sm.tsa.adfuller(ads_s, maxlag=None, regression='nc', autolag='AIC', store=False, regresults=False)



1. Test for a unit root:

$$\nabla y_t = \delta y_{t-1} + u_t$$

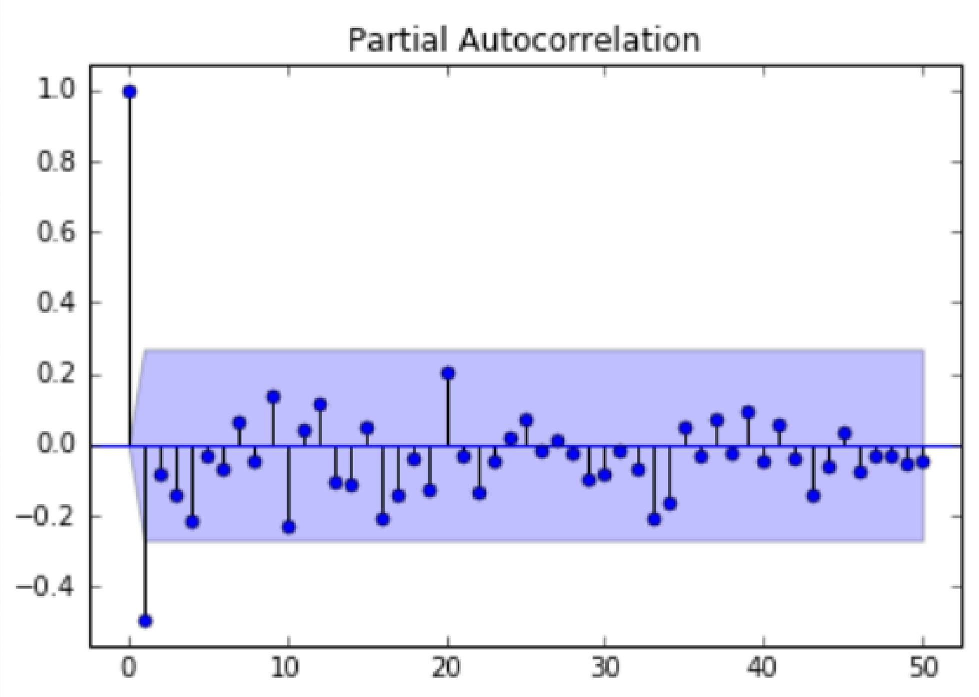
2. Test for a unit root with drift:

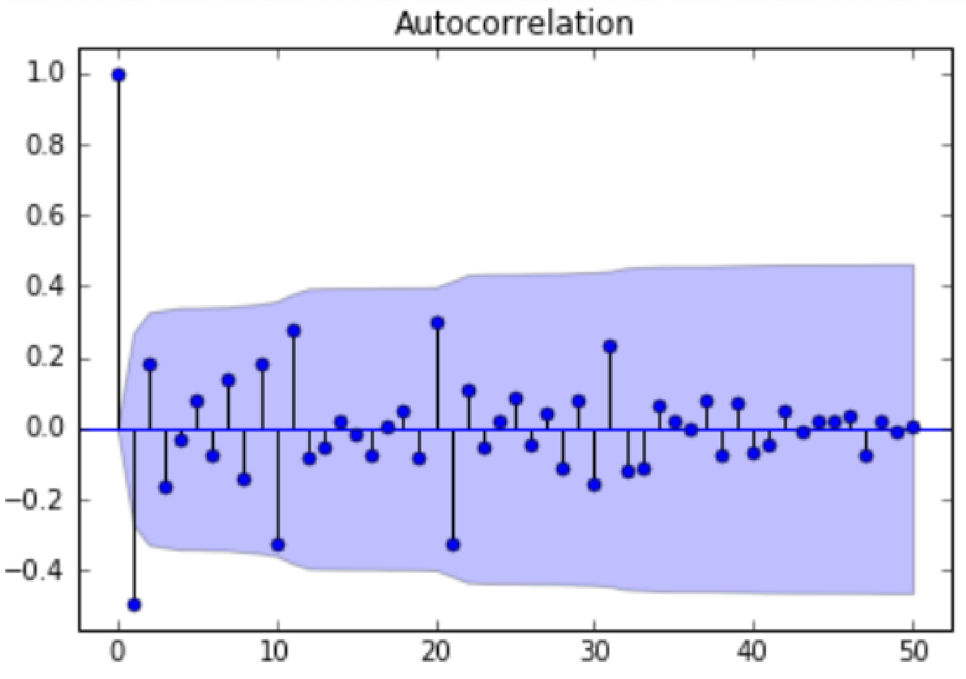
$$\nabla y_t = a_0 + \delta y_{t-1} + u_t$$

3. Test for a unit root with drift and deterministic time trend:

$$\nabla y_t = a_0 + a_1 t + \delta y_{t-1} + u_t$$

Check for Autocorrelations:





Model Selection:

- Criteria:

$$AIC = -2\log(L) + 2(p+q+k+1)$$

$$AICc = AIC + (2(p+q+k+1)(p+q+k+2))/(T-p-q-k-2)$$

$$BIC = AIC + (\log(T) - 2)(p+q+k+1)$$

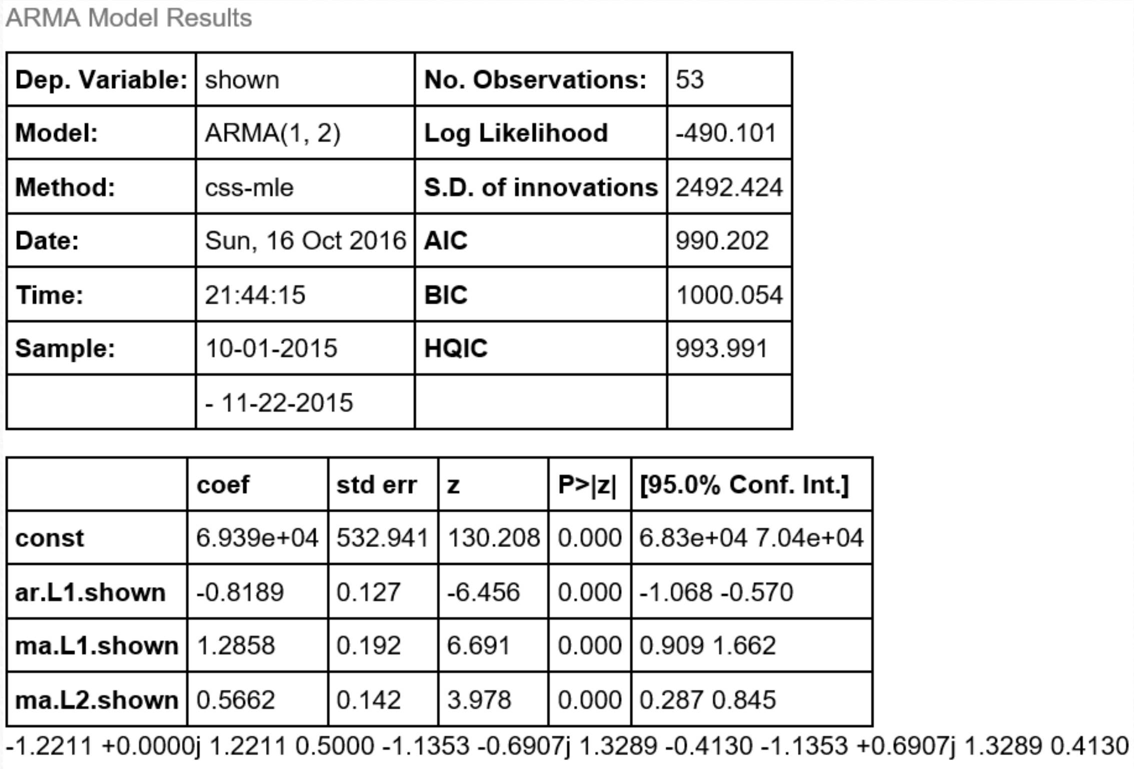
Model Selection:

- ARIMA(p,d,q): x13_arima_select_order
  - Perform automatic seasonal ARIMA order identification using x12/x13 ARIMA

- AMRM(p,q): arma_order_select_ic

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Model Output:



ARMA Model Results

Dep. Variable:	shown	**No. Observations:**	53
Model:	ARMA(1, 2)	**Log Likelihood**	-490.101
Method:	css-mle	**S.D. of innovations**	2492.424
Date:	Sun, 16 Oct 2016	**AIC**	990.202
Time:	21:44:15	**BIC**	1000.054
Sample:	10-01-2015	**HQIC**	993.991
	- 11-22-2015		

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
const	6.939e+04	532.941	130.208	0.000	6.83e+04 7.04e+04
ar.L1.shown	-0.8189	0.127	-6.456	0.000	-1.068 -0.570
ma.L1.shown	1.2858	0.192	6.691	0.000	0.909 1.662
ma.L2.shown	0.5662	0.142	3.978	0.000	0.287 0.845

-1.2211 +0.0000j 1.2211 0.5000 -1.1353 -0.6907j 1.3289 -0.4130 -1.1353 +0.6907j 1.3289 0.4130

Model Performance:

- In sample fit:
  - P-value
  - $\varepsilon_t \sim N(\mu, \sigma^2)$, iid
  - Mean Absolute Error
- Out of Time Prediction

Potential Improvement:

- Data transformation
- ARCH/GARCH
- VAR

Task 3: Cluster ads into 3 groups

- Pearson Correlation Coefficient
- $\rho = \frac{Cov(\text{avg\_cost\_per\_click, days\_past})}{\sigma_{\text{avg\_cost\_per\_click}}\sigma_{\text{days\_past}}}$
- Scipy
  - stats.pearsonr (avg_cost_per_click, days_past)

Q & A:

Thank you!