Sales Forecast for Super Cue Boba Shop

Tom Cheng

Datas Science Career Track Capstone, May 20^{th} 2020 Cohort

Why Forecast?

Sales forecast helps control **inventory**, efficient **staffing**, and **predict profits**.

Most of boba shops are small shops, and thus everything counts. Especially staffing, **an extra staff per day can take away up to 15% of daily revenue**!

What Boba Shops?

And many more…

Super Cue BOBA + BITES

Plentea

PurpleKow A place for comfort dessert, coffee and tea

happy lemon

Quickly

green bubble

Sharetea Bubblicious

Tpumps Good tea to share

BOBA GUYS EST. 2011

TEASPOON

TapiocaExpress

The Original Hot Honey Boba and Iced Milk Drinks

TEA FOR U

GONG CHA tea.coffee.juice

What's needed to do sales forecast?

- Use **common sense and logics** to find **factors** that may influence sales, in most cafes/restaurants: **TIME**
- **Data**: past sales and factors

Where my data came from?

- I was able to get 2 years (2016&2017) of hourly sales data from Super Cue Café.
- The files were in csv format; two files, with 12 sheets (one month per sheet).

	A	B	C	D	E	F	G	H	I	J	K
1		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	AVG		
2	Time	1/4/2016	1/5/2016	1/6/2016	1/7/2016	1/8/2016	1/9/2016	1/10/2016			
3	11:00	16.39	22.45	33.59	8.40	27.25	15.70	52.70	25.21		
4	12:00	36.27	27.75	41.48	23.30	86.00	88.04	40.33	49.02		
5	13:00	78.68	7.25	111.56	54.49	48.34	79.59	109.70	69.94		
6	14:00	51.44	30.64	92.00	42.28	65.21	142.73	100.45	74.96		
7	15:00	57.70	100.67	94.83	116.13	186.20	161.16	140.85	122.51		
8	16:00	148.93	149.72	94.27	101.65	158.67	75.23	88.12	116.66		
9	**17:00**	**74.55**	**43.14**	**45.26**	**52.04**	**93.76**	**110.65**	**136.35**	79.39		
10	**18:00**	**50.34**	**68.53**	**70.35**	**47.96**	**117.30**	**66.70**	**133.73**	79.27		
11	**19:00**	**56.02**	**93.65**	**57.22**	**128.00**	**143.23**	**74.14**	**109.64**	94.56		
12	**20:00**	**79.35**	**75.90**	**52.53**	**77.01**	**105.70**	**89.94**	**72.14**	78.94		
13	**21:00**	**58.34**	**27.45**	**46.87**	**91.02**	**182.96**	**78.05**	**64.06**	78.39		
14	**22:00**	**31.68**	**31.70**	**49.12**	**75.42**	**89.71**	**89.77**	**35.55**	57.56		
15	**23:00**					**64.12**	**27.70**		**45.91**		
16	**Total**	**739.69**	**678.85**	**789.08**	**817.70**	**1368.45**	**1099.40**	**1083.62**	**939.54**		
17	AM	389.41	338.48	467.73	346.25	571.67	562.45	532.15	458.31		
18	PM	350.28	340.37	321.35	471.45	796.78	536.95	551.47	481.24		
19											
20		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	AVG		
21	Time	1/11/2016	1/12/2016	1/13/2016	1/14/2016	1/15/2016	1/16/2016	1/17/2016			
22	11:00	11.35	10.85	16.95	11.90	24.20	43.51	35.50	22.04		
23	12:00	38.55	65.56	44.65	28.83	49.40	34.35	73.23	47.80		
24	13:00	61.45	70.70	84.48	33.79	83.09	115.03	60.04	72.65		
25	14:00	91.62	72.27	99.70	41.60	32.10	108.92	101.64	78.26		
26	15:00	138.37	87.73	97.60	83.27	156.48	117.72	93.58	110.68		
27	16:00	88.73	115.79	149.71	106.12	141.42	169.85	105.19	125.26		

Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec

What should time series data look like?

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Data%20Wrangling.ipynb

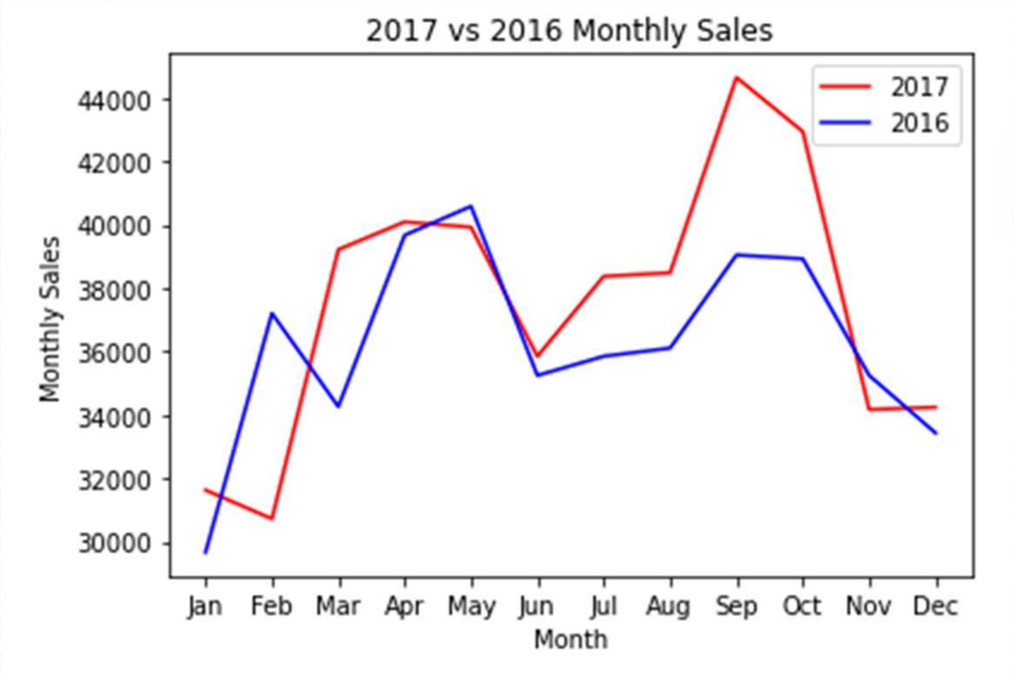
- Ultimately, we will want a time series data with the following format:
  - Row index in time series format
  - Columns with factors that influence sales (if any), and most importantly: Sales

	A	B
1	Date	sales
2	1/4/2016 11:00	16.39
3	1/4/2016 12:00	36.27
4	1/4/2016 13:00	78.68
5	1/4/2016 14:00	51.44
6	1/4/2016 15:00	57.7
7	1/4/2016 16:00	148.93
8	1/4/2016 17:00	74.55
9	1/4/2016 18:00	50.34

Exploratory Data Analysis (EDA)

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20EDA.ipynb

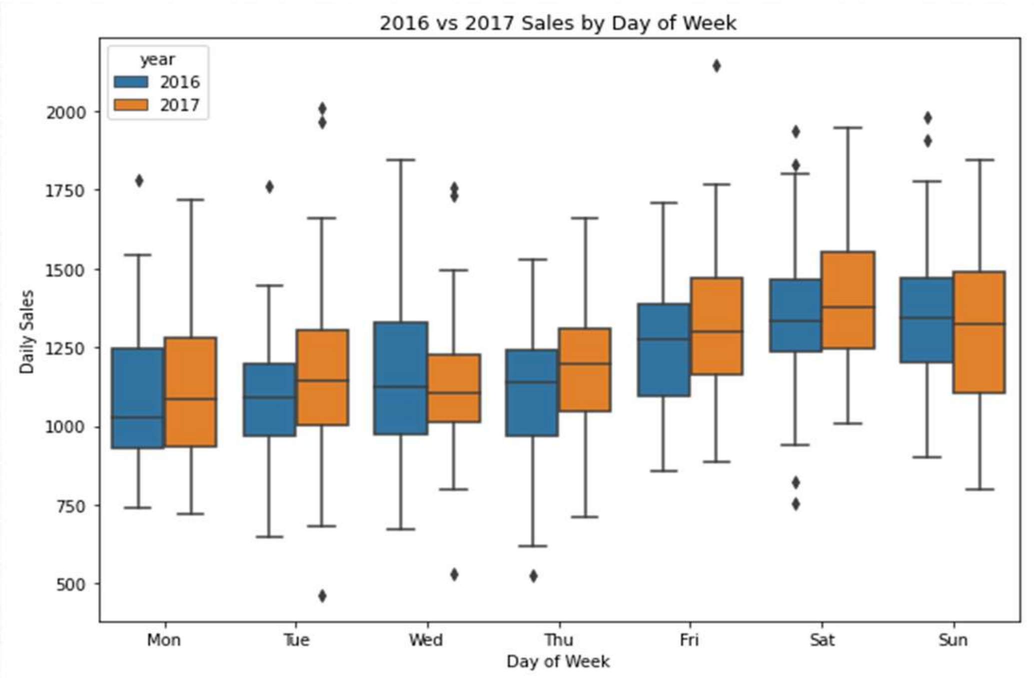
- Comparison between 2016&2017
  - By **Month** (12)
  - By **Week** (52x7)
  - **Day of Week** (7)
  - **Hour of Day** (11AM~Close)



EDA cont.

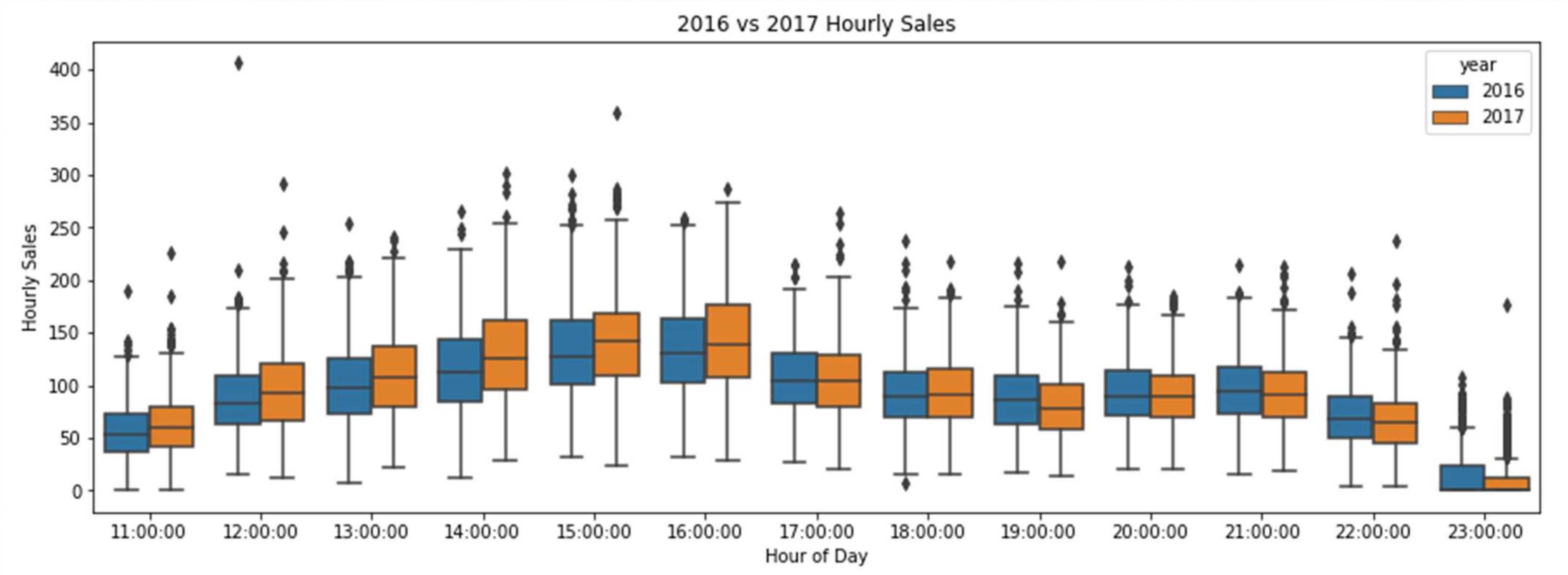
https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20EDA.ipynb

- 2016 vs 2017 average daily sales boxplot
- Why boxplot?
  - Median
  - IQR
  - Shows Outliers



EDA cont.

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20EDA.ipynb



Machine Learning Modeling

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

- Type: Time Series Forecast
- Tools: pandas, numpy, matplotlib, seaborn, statsmodels.api (for ACF), sklearn.metrics, datetime
- Methods/Models:
  - Autoregressive Integrated Moving Average (**ARIMA**)
  - Seasonal Autoregressive Integrated Moving-Average (**SARIMA**)
  - Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (**SARIMAX**)
  - Holt Winter's Exponential Smoothing (**HWES**)
  - **TBATS** (Trigonometric Exponential Smoothing State Space model with Box-Cox transformation, ARMA errors, Trend and Seasonal Components)

Modeling Steps

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

- Data Preprocessing: time series format, no missing values, add exogenous if needed (SARIMAX).
- Split to train/test, 50/50 for hourly dataset, 80/20 for daily.
- Use grid search to find best hyperparameters for each model
- Use MAE and fitting time to evaluate the model's performance

Modeling (Day)

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

Top models by ForeMAE

	model	MSE	MAE	AIC	ForeMSE	ForeMAE	p	d	q	P	D	Q	fit_time(s)	fit_time_per_row(s)
451	AA_DM3FT	67322.726891	192.798886	NaN	21133.293238	112.767890	NaN	NaN	NaN	NaN	NaN	NaN	32.80	0.0541
452	AA_DM4FT	69428.451047	194.963525	NaN	21098.806273	112.940502	NaN	NaN	NaN	NaN	NaN	NaN	19.43	0.0321
450	AA_DM2FT	66073.422118	190.475741	NaN	21330.794984	113.681134	NaN	NaN	NaN	NaN	NaN	NaN	25.02	0.0413
84	ARIMAX	65873.074606	193.495806	8029.195843	22576.327364	116.106628	5.0	0.0	3.0	NaN	NaN	NaN	NaN	NaN
393	SARIMAX	65873.074606	193.495806	8029.195843	22576.327364	116.106628	5.0	0.0	3.0	0.0	0.0	0.0	NaN	NaN

TBATS, ARIMAX, SARIMAX, auto_arima were the only models that were able to capture yearly seasonality (visually). And this table concludes the top model metric scores based on 2018 MAE (ForeMAE).
Winner: AA_DMFTs

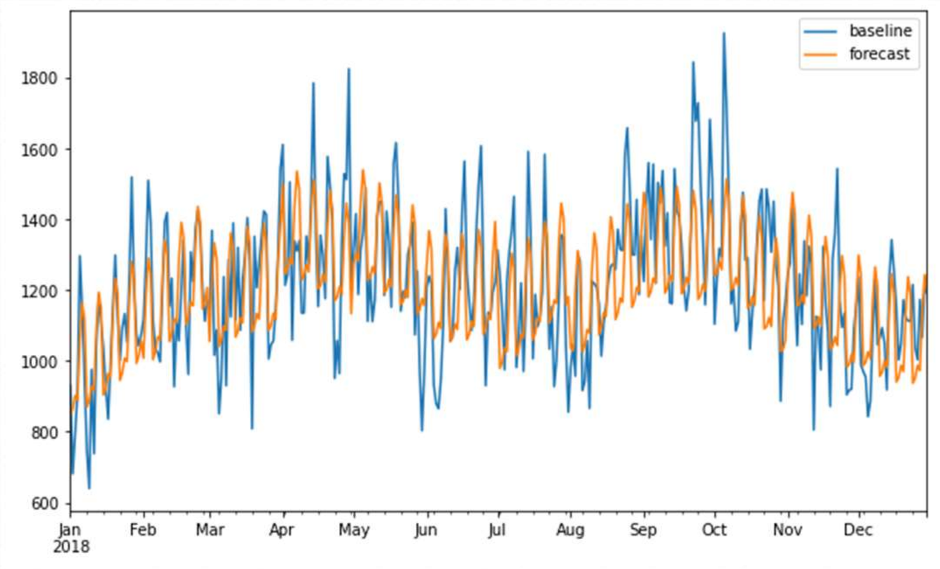
Modeling (Day) cont.

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

What is AA_DM2FT

- AA for auto_arima, which is a form of SARIMAX that automatically finds the optimal hyperparameter for SARIMAX
- DM for the exogenous variable: Day of Week, and Month.
- 2FT for 2 Fourier-term

AA_DF2FT 2018 baseline vs pred



Modeling (hour)

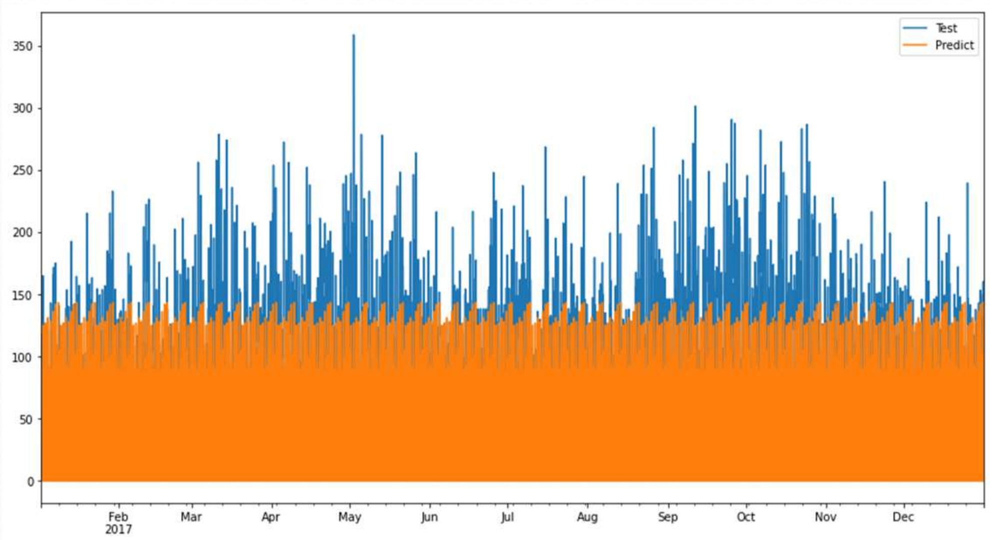
https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

Top model scores by model type

model	2017RMSE	2017MAE	fit_time_per_row	total_fit_time(min)
TBATS-hourly2017	28.538750	15.545161	0.0849	12.36144
AA3X-hourly	28.208743	15.430962	1.2739	185.47984
AA3N-hourly	28.721891	15.685231	0.5904	85.96224
AA3R-hourly	28.459946	15.537691	0.5217	75.95952

Pay attention to total_fit_time in minutes, there is an obvious winner. TBATS, but one problem: it doesn't capture yearly seasonality (it gets hour of day and day of week).

TBATS 2017 hourly sales test vs pred



Modeling (hour)

https://github.com/tc18fwd/SpringBoard/blob/master/Capstone%20Two/Capstone%202%20Finalized%20Codes.ipynb

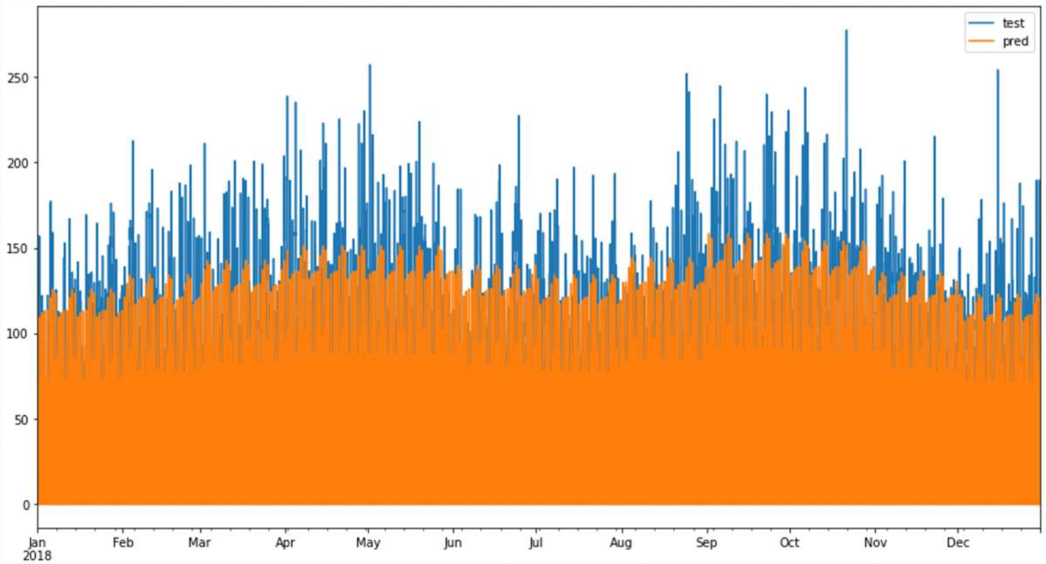
TBATS with yearly seasonality

model	2017RMSE	2017MAE	fit_time_per_row
AA_TBATS2017	28.118901	15.366706	NaN

model	2017RMSE	2017MAE	fit_time_per_row
TBATSxR_2017	28.113221	15.361412	NaN

Yearly seasonality added by multiplying ratios derived from monthly sales prediction by auto_arima (AA) and by manual calculation (R).

TBATSxTBATS(month) 2018 baseline vs prediction



Use the model to find optimal store hours

Time where pred < 40

	pred
2018-01-01 11:00:00	39.429802
2018-01-05 23:00:00	8.360880
2018-01-06 23:00:00	10.373761
2018-01-08 11:00:00	39.431212
2018-01-12 23:00:00	8.360880
...	...
2018-12-21 23:00:00	8.172995
2018-12-22 23:00:00	10.140643
2018-12-24 11:00:00	38.545117
2018-12-28 23:00:00	8.172995
2018-12-29 23:00:00	10.140643

113 rows × 1 columns

11:00 where pred < 40

	pred
2018-01-01 11:00:00	39.429802
2018-01-08 11:00:00	39.431212
2018-01-15 11:00:00	39.431212
2018-01-22 11:00:00	39.431212
2018-01-29 11:00:00	39.431212
2018-12-03 11:00:00	38.545117
2018-12-10 11:00:00	38.545117
2018-12-17 11:00:00	38.545117
2018-12-24 11:00:00	38.545117

11pm and pred < 40

	pred
count	104.000000
mean	10.514056
std	1.389341
min	8.172995
25%	9.441219
50%	10.239730
75%	11.597632
max	13.054621

Recommend to open from 11AM to 11PM everyday. This may have saved about 104*27=2808 dollars per year, and have happier staffs.

Use the model to find peak hours

Hours with sales above 149

	pred
2018-04-07 15:00:00	151.185653
2018-04-07 16:00:00	149.381177
2018-04-14 15:00:00	151.185653
2018-04-14 16:00:00	149.381177
2018-04-21 15:00:00	151.185653
2018-04-21 16:00:00	149.381177
2018-04-28 15:00:00	151.185653
2018-04-28 16:00:00	149.381177
2018-05-05 15:00:00	151.185653
2018-05-05 16:00:00	149.381177
2018-05-12 15:00:00	151.185653
2018-05-12 16:00:00	149.381177

No. of time sales was over 150 at 3pm

	counts
4	4
5	4
9	14
10	8

April, May, Sep, and October.

2 hr before/after peak hour

```
                     pred
2018-10-06 13:00:00  118.485713
2018-10-13 13:00:00  118.485713
2018-10-20 13:00:00  118.485713
2018-10-27 13:00:00  118.485713
                     pred
2018-10-06 14:00:00  136.491407
2018-10-13 14:00:00  136.491407
2018-10-20 14:00:00  136.491407
2018-10-27 14:00:00  136.491407
                     pred
2018-10-06 17:00:00  118.188856
2018-10-13 17:00:00  118.188856
2018-10-20 17:00:00  118.188856
2018-10-27 17:00:00  118.188856
                     pred
2018-10-06 18:00:00  104.183166
2018-10-13 18:00:00  104.183166
2018-10-20 18:00:00  104.183166
2018-10-27 18:00:00  104.183166
```

Recommendation for Super Cue

- It's recommended for Super Cue to have 3 staffs during these hours
- April and May's Saturday: 1PM to 5PM
- September's weekends (Fri to Sun): 2PM to 6PM
- October's Saturday and Sunday: 1PM to 5PM
- And have store open from 11AM to 11PM everyday

Limitations & Ideas to Improve Prediction

- All the models did not account holidays, events, temperature, and weather. Which are all factors to forecast sales, especially for drinks. So please do expect differences if the above mentioned factors were to be considered, e.g. you should expect more sales on a hot day, and less when it's raining.
- If the factors mentioned above could be implemented in the model, the prediction will definitely be better. However, models with multivariable factors will be very time consuming and data demanding.

THANK YOU!

Super BOBA + BITES Cue
Super
Super BOBA + BITES Cue
Super BOBA + BITES Cue