Muvaffak Onus

How to do hypothesis testing on sales data?

This post is the step 2 in the project mentioned in this post. We will do a hypothesis testing to see if promotions affect sales of the given day.

For simplicity we take a sample with n =50,000 because data has over 1 million rows.

Null hypothesis: Promo feature doesn't affect sales.
Alternative: Sales are dependent on promo feature.

Following is an example table from the data:

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	539	6	2015-07-25	13022	1182	1	0	0	0
1	150	1	2014-12-29	9990	1060	1	0	0	1
2	163	2	2014-01-07	8055	952	1	1	0	0
3	995	2	2014-03-04	10792	864	1	1	0	0
4	109	3	2015-02-11	5910	705	1	0	0	0
5	121	3	2013-10-23	4354	494	1	1	0	1
6	569	3	2015-04-29	4888	706	1	1	0	0
7	750	7	2014-04-20	0	0	0	0	0	0
8	1022	7	2013-11-03	0	0	0	0	0	0
9	236	2	2015-06-02	9521	1065	1	1	0	0
10	177	6	2013-09-07	1215	191	1	0	0	0
11	949	4	2015-06-04	6363	521	1	1	0	0
12	1053	1	2013-05-13	12037	1300	1	1	0	0
13	290	3	2014-08-27	5626	613	1	0	0	0
14	148	6	2015-06-20	8572	695	1	0	0	0
15	513	3	2013-10-30	17036	2038	1	0	0	0
16	1001	6	2013-07-20	2846	370	1	0	0	0
17	134	2	2015-07-14	5649	578	1	1	0	0
18	77	7	2013-02-10	0	0	0	0	0	0
19	554	3	2013-10-16	3627	490	1	0	0	0
20	291	4	2015-04-23	6148	686	1	0	0	0

We will apply an OLS Linear regression to see if Promo has effect on Sales. The following is the code piece to run the model:

```
import pandas as pd
import statsmodels.api as sm
df = pd.read_csv("dataset/train-sample.csv", sep=',', parse_dates=[2])

# LR input
inputData = pd.DataFrame(df, columns=["Promo"])
inputData = sm.add_constant(inputData, has_constant='add')
targetData = pd.DataFrame(df, columns=["Sales"])

# Run LR
model = sm.OLS(targetData, inputData).fit()
model.summary()
```

The result is as follows:

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.203
Method:	Least Squares	**F-statistic:**	1.271e+04
Date:	Thu, 29 Mar 2018	**Prob (F-statistic):**	0.00
Time:	19:17:03	**Log-Likelihood:**	-4.7775e+05
No. Observations:	50000	**AIC:**	9.555e+05
Df Residuals:	49998	**BIC:**	9.555e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4392.2704	19.442	225.921	0.000	4354.165	4430.376
Promo	3542.7601	31.427	112.730	0.000	3481.163	3604.357

Omnibus:	7528.180	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20096.463
Skew:	0.831	**Prob(JB):**	0.00
Kurtosis:	5.624	**Cond. No.**	2.43

As seen on the summary table, P-value for Promo variable is ~0. It means that the probability of null hypothesis happening is 0 with confidence level of 95%. So, we reject the null hypothesis and say

Promo affects sales with 95% of confidence.

This is pretty much how you'd conduct an hypothesis testing. Finding such features and enhancing the model with them is the next step to get closer to predict how Sales is affected and how we can predict the future!


monus / March 29, 2018

Leave a Reply

Your email address will not be published. Required fields are marked *

COMMENT

NAME *

EMAIL *

WEBSITE

POST COMMENT

PREVIOUS

Forecasting Sales: Daily Demand Prediction of Products in Rossmann Stores

Search …

RECENT POSTS

- How to do hypothesis testing on sales data?
- Forecasting Sales: Daily Demand Prediction of Products in Rossmann Stores
- How to detect emotion of the player in Unity?

RECENT COMMENTS

ARCHIVES

- March 2018
- February 2018
- January 2017

CATEGORIES

- Uncategorized

META

- Log in
- Entries RSS
- Comments RSS
- WordPress.org