Take Home Challenge Presentation

University of California, Berkeley

Jinge Li

1. Dashboard

Build a dashboard via dash, which shows several items of time series topline table and visualizes time development plots of both original and log values for all variables. Please enter command line and execute ”python app.py” to open interactive webpage.

2. Anomaly Detection

(1). Data Cleaning Procedures

Remove redundant Date column;

Convert data type to datetime or numeric accordingly;

Deal with wrong split rows;

Replace all negative time spent (seconds) with 0.

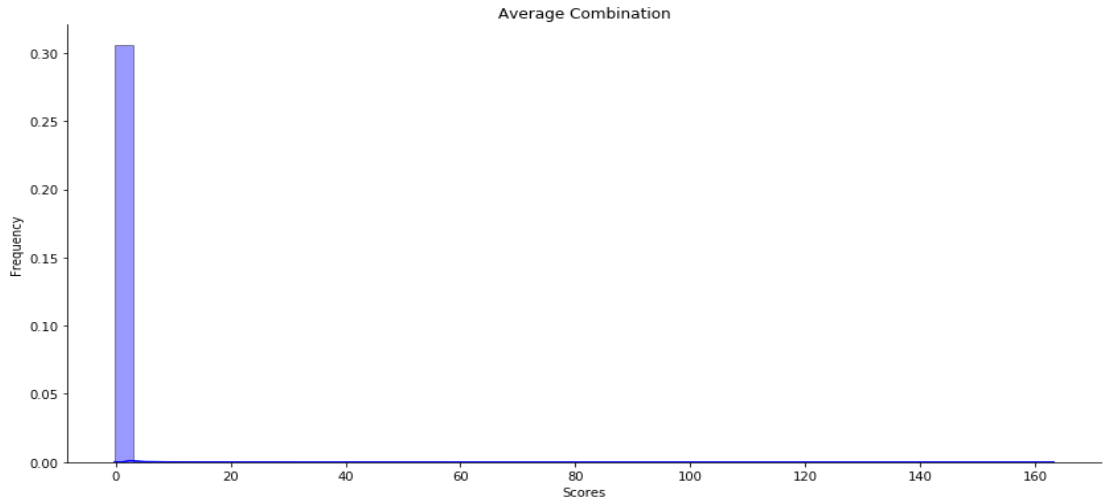
(2). Outlier Detection Via K Nearest Neighbors Method

Standardize all the features, compute standardized decision scores for every combination of K from 10 to 100 and observations. Detect anomaly observations under four different criterions.

Idea: treat observations with top 5% decision scores as anomaly points.

(a). Average of Average

Take average towards decision scores first over observations and then over K values.



Regular sample points: 711552

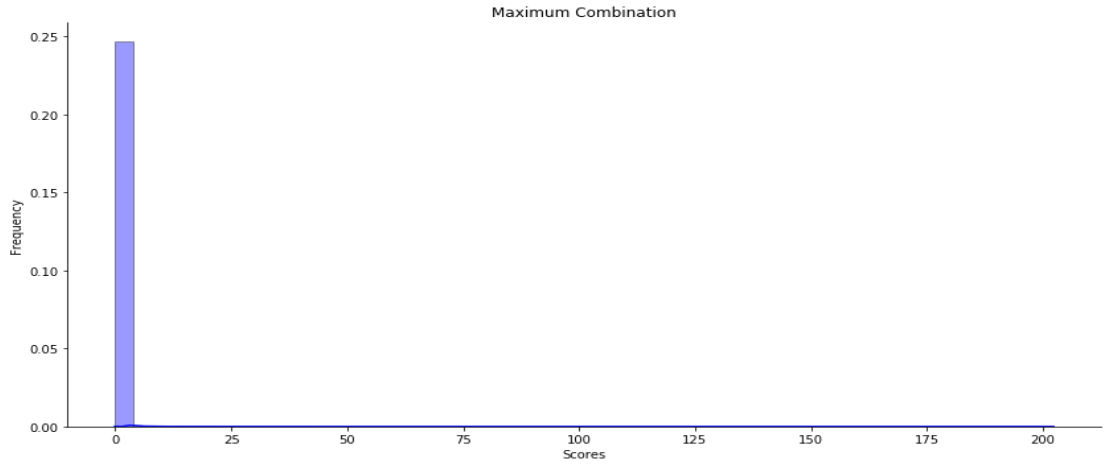
Anomaly sample points: 36712

Mean of Two Classes

class	TRU	DAU	Items	Trans	Talk	CF	RC	TS	Score
0	665872	16405	45555	3567	2.78	4690	40	122	-0.06
1	2175874	52378	154312	11555	5.24	12906	42	132	1.26

(b). Maximum of Maximum

Take maximum value of decision scores first over observations and then over K values.



Regular sample points: 703794

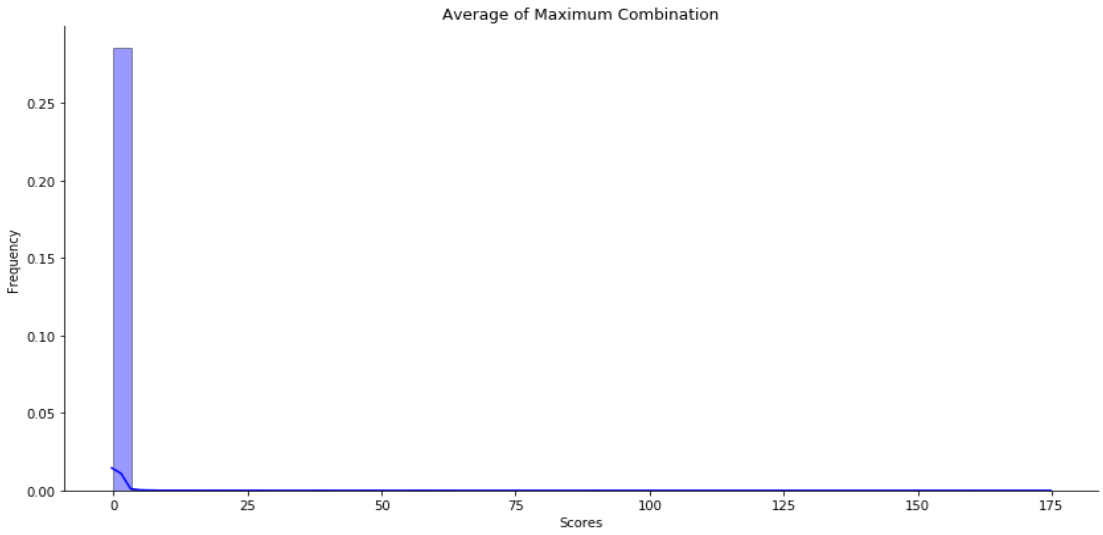
Anomaly sample points: 44470

Mean of Two Classes

class	TRU	DAU	Items	Trans	Talk	CF	RC	TS	Score
0	662350	16300	45276	3550	2.77	4670	40	122	-0.06
1	1968182	47750	139760	10420	4.95	11792	42	133	1.33

(c). Average of Maximum

First take average over observations and then take maximum of K values.



Regular sample points: 709542

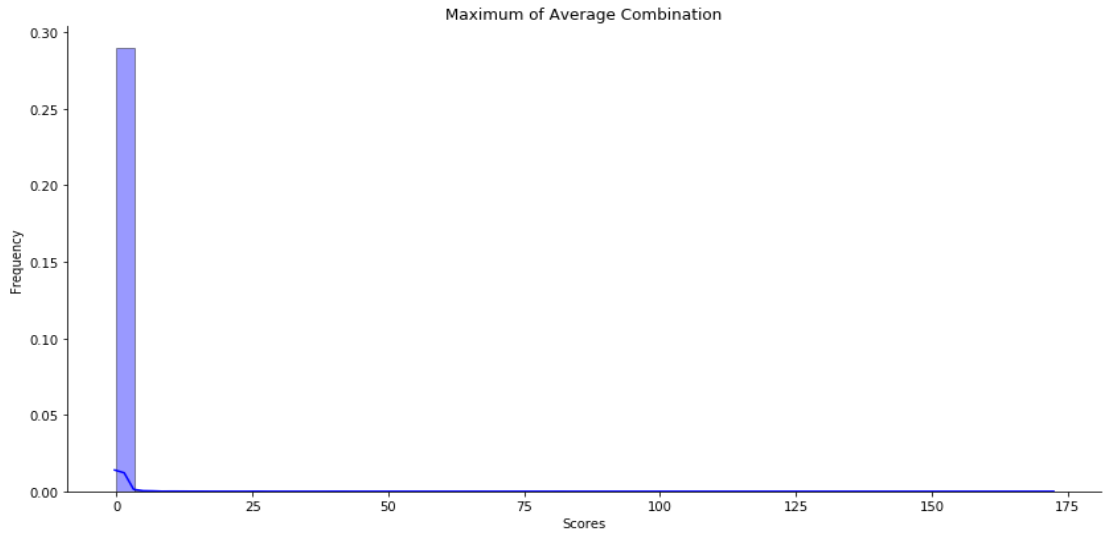
Anomaly sample points: 38722

Mean of Two Classes

class	TRU	DAU	Items	Trans	Talk	CF	RC	TS	Score
0	664867	16372	45483	3562	2.78	4685	40	122	-0.06
1	2115911	51110	149999	11219	5.17	125773	42	133	1.29

(d). Maximum of Average

First find maximum over observations and then take average of K values.



Regular sample points: 708109

Anomaly sample points: 40155

Mean of Two Classes

class	TRU	DAU	Items	Trans	Talk	CF	RC	TS	Score
0	664131	16353	45420	3559	2.78	4681	40	122	-0.06
1	2077100	50208	147367	11012	5.12	12370	42	133	1.30

Generally, taking maximum value is more likely to treat sample points as anomaly, since we are choosing sample points with higher decision scores.

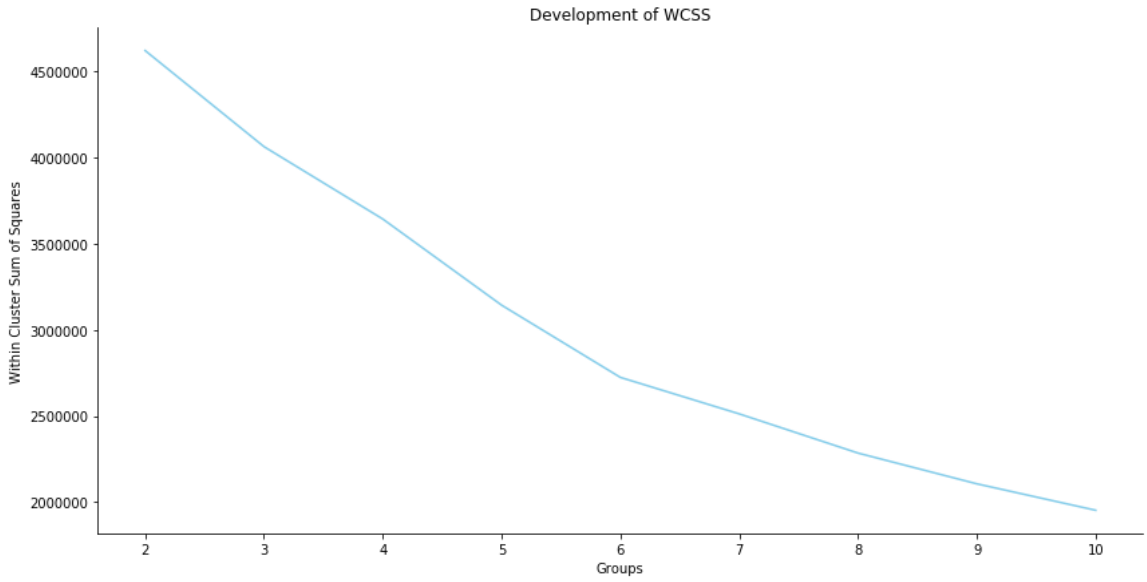
3. Clustering and Prediction

(1). K-means Clustering

Since Items per Trans and Items per DAU are not consistent with Items, Trans and DAU, we do not take them into consideration when clustering.

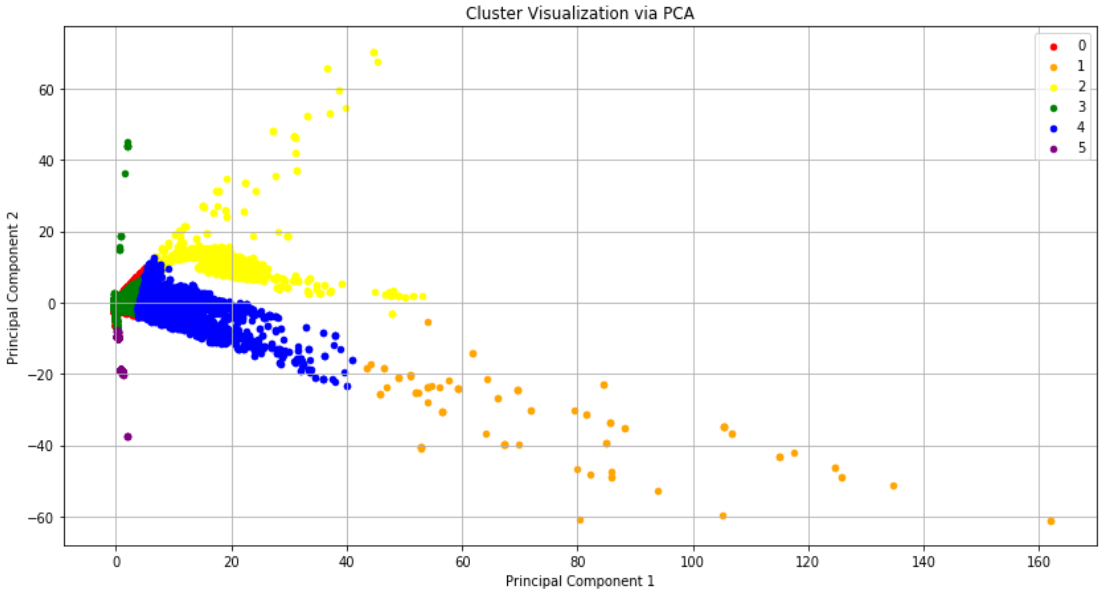
Try clusters from 2 to 10 and select the number of groups based on Within Cluster Sum of Squares. (magnitude: 10^6)

Numbers	2	3	4	5	6	7	8	9	10
WCSS	4.6	4.1	3.6	3.1	2.7	2.5	2.3	2.1	2.0



Assign 6 groups according to turning point in plot above, separate sample points to 6 clusters based on K-means algorithm.

For visualization, use Principal Component Analysis to extract first and second principal components. Together these two principal components roughly contain 53.35% information of whole dataset.



(2). Model Prediction

Our sample points do not have original labels, so we record cluster information in previous question as labels and use Support Vector Machine to classify.

Split training data and validation data as roughly 6:1, train SVM in training dataset and compute accuracy score in both training and validation dataset.

The accuracy rate in training dataset is 0.99935.

The accuracy rate in validation dataset is 0.99906.

SVM works fairly well, accuracy score of training dataset is almost 1, which means training data is linearly separable.

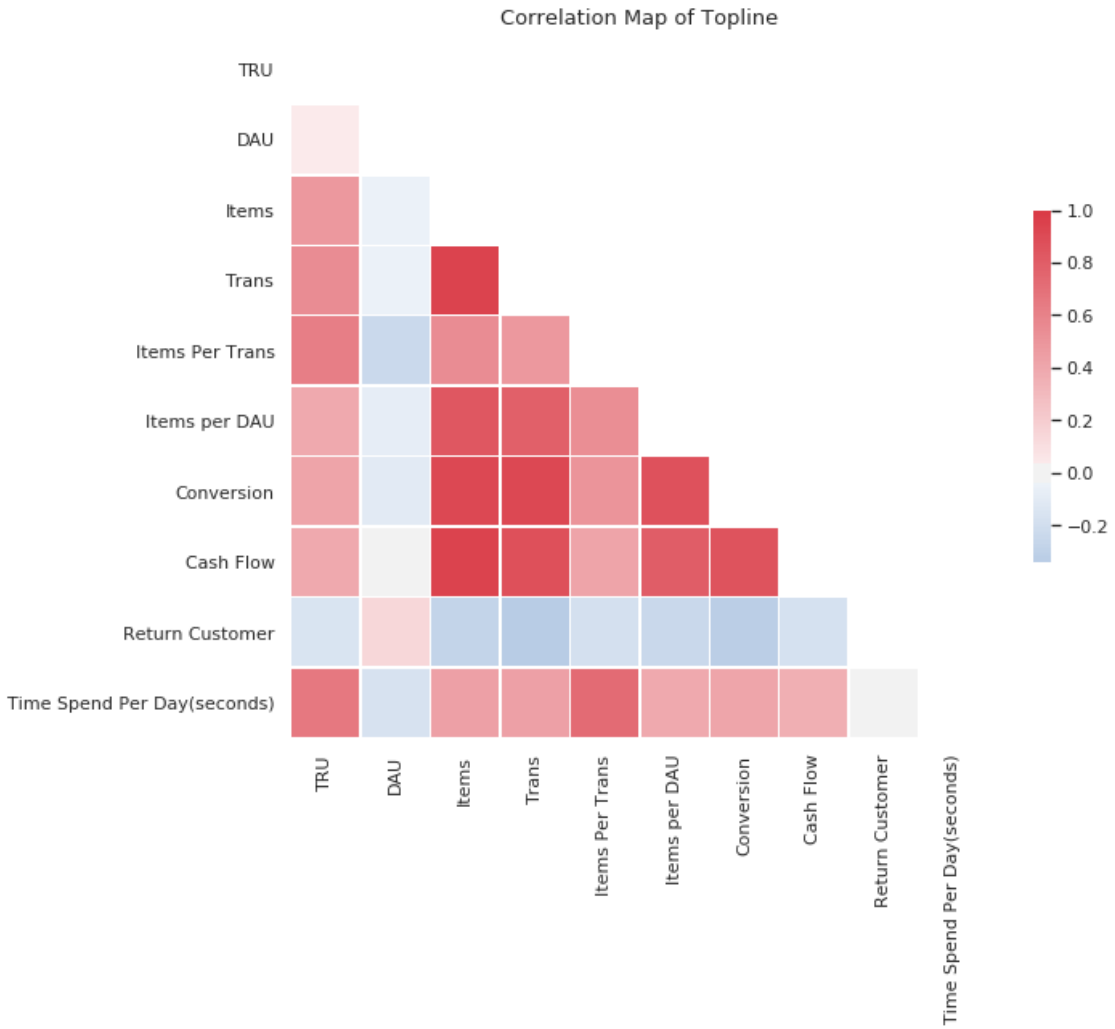
Whenever a new sample point appears, plug in SVM decision function to predict which class does this sample point belong to.

4. Time Series Causation Graph Analysis

(1). Correlations

Take average over all regions in a given date to construct time series data.

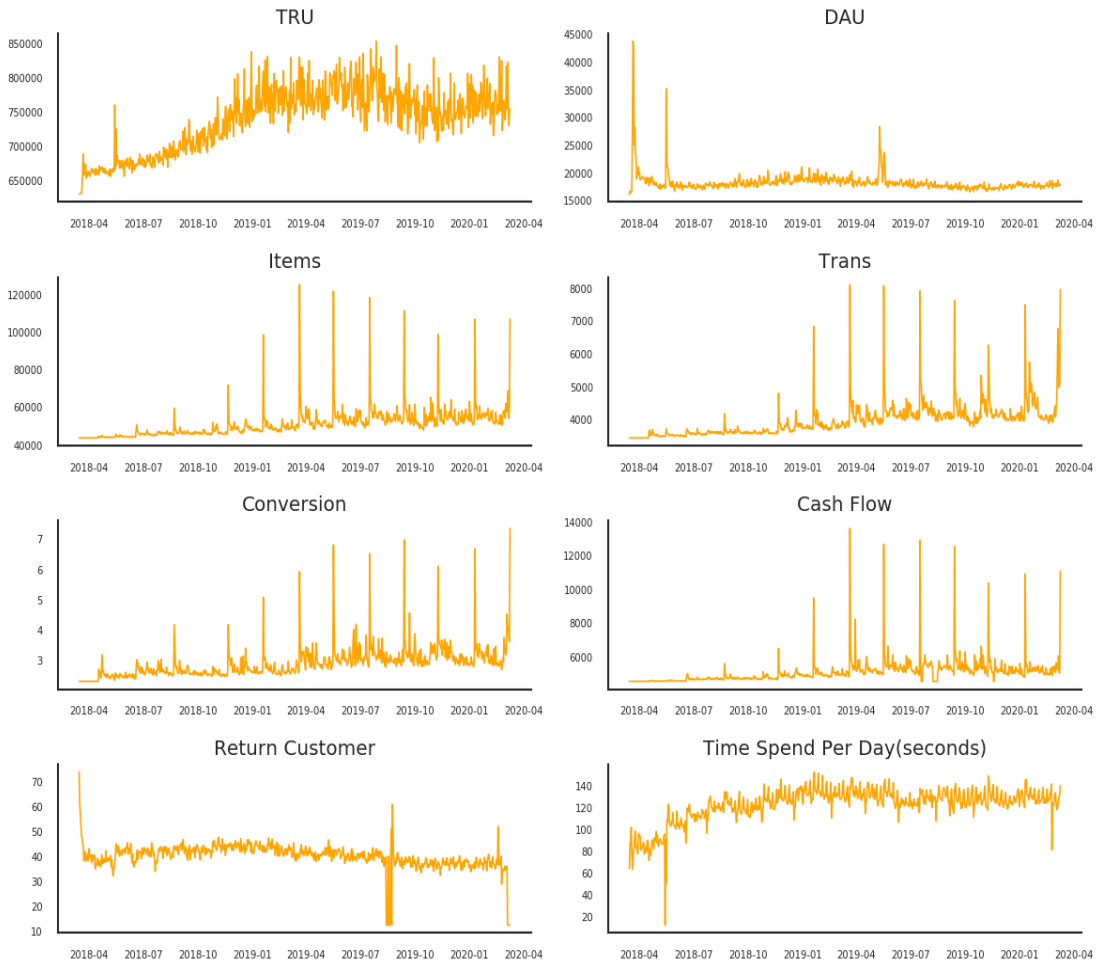
Draw correlation map among all variables.



Correlation map may give us intuitions about causations.

For example: Items, Trans, Conversion could exert positive influence on Cash Flow because they have significant positive correlations.

(2). Time Series



From the time series' patterns we can generally draw a conclusion:

Items, Trans, Conversion, Cash Flow could interchangeably cause each other.

From 2009, these four variables will dramatically increase and then back to normal level every three months

Tru may have causation with these four variables, but the relationships are not obvious

(3). Granger's Causality Test

For every combination of two variables, run Granger's Causality Test and record p values in table below.

y\x	TRU	DAU	Items	Trans	Talk	CF	RC	TS
TRU	1	0.0002	0.0003	0.0037	0	0.0084	0.0024	0
DAU	0	1	0.1155	0.0649	0.0285	0.2884	0	0
Items	0	0.0292	1	0	0.0137	0	0	0
Trans	0	0.0224	0	1	0.0971	0	0	0.0002
Talk	0	0.0113	0.0782	0	1	0	0	0.0015
CF	0	0.1260	0	0	0	1	0	0
RC	0.007	0.0966	0.0187	0	0.0002	0.4996	1	0.0071
TS	0	0	0	0	0	0	0.0005	1

The null hypothesis in (i, j) element is that lagged j th variable does not explain the variation in i th variable.

Take p-value 0.05 as critical value, roughly all variables could interchangeably cause each other except DAU and Return Customer.

DAU and Return Customer are more independent.

5. Finding and Story

(1). Summary Statistics

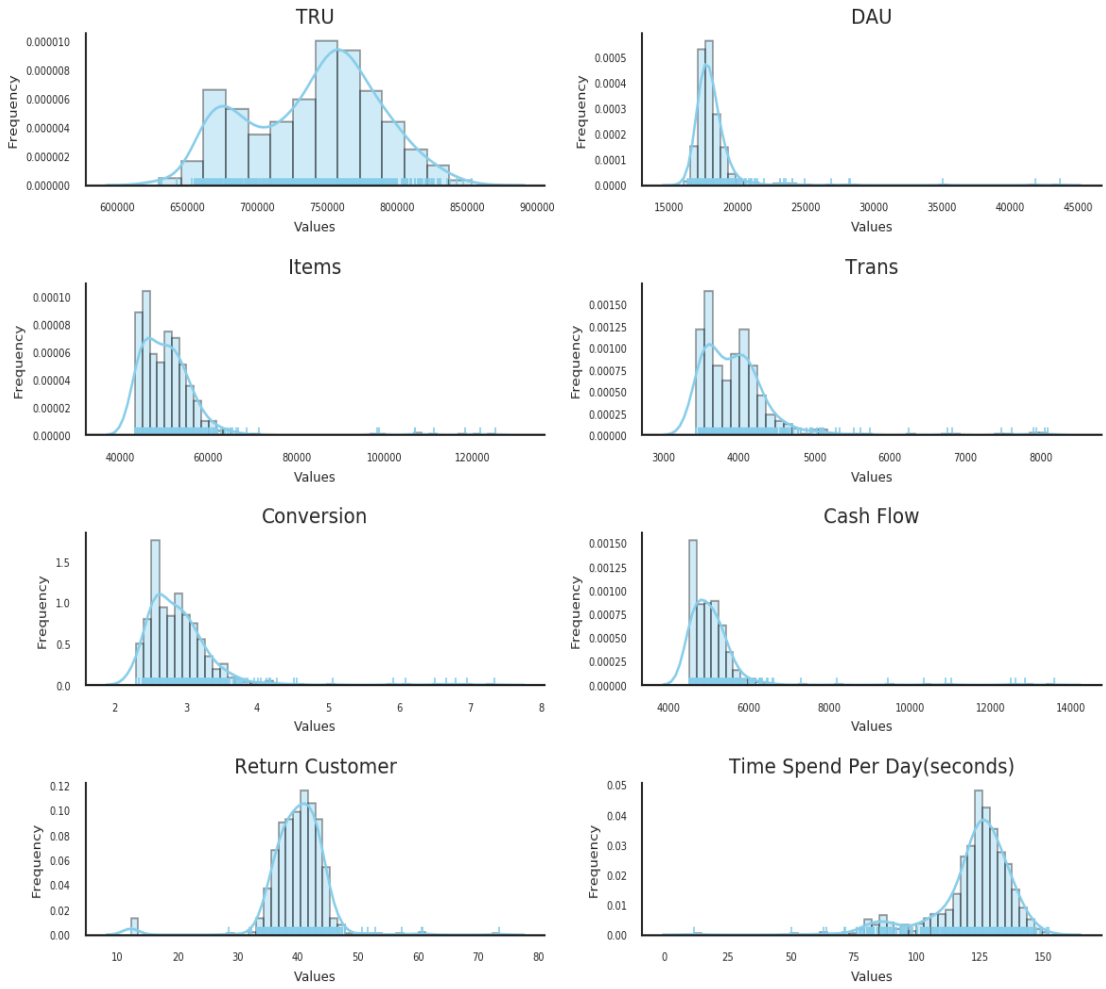
	mean	std	min	25%	50%	75%	max
TRU	739957	677660	628648	628810	631445	657473	26776754
DAU	18169	18108	15292	15297	15374	16177	2141935
Items	50891	46102	43429	43429	43521	44660	4387412
Trans	3958	3656	3428	3428	3433	3523	434847
I/Trans	31.57	23.63	17.58	17.58	28.09	36.08	1744.87
I/DAU	1.41	1.21	1.30	1.30	1.32	1.39	444.24
Talk	2.90	2.59	2.30	2.30	2.56	3.00	232.63
CF	5093	4713	4527	4527	4533	4611	819825
RC	39.80	24.80	12.21	28.66	39.43	48.92	3912.09
TS	122.61	43.31	0.00	98.95	118.18	139.37	1390.06

Time spend per day has mean 122.61 and standard deviation 43.31, which is comparably stable.

Other variables have similar mean and standard deviation, indicating high volatility.

A little question about inconsistency in Items, Trans and Items Per Trans and DAU, Items Per DAU.

(2). Distributions



TRU and Time Spend Per Day(seconds) follow bimodal distribution, indicating there might exist two classes.

Return Customer follows normal distribution with mean 40 and variance 615.

Items, Trans, Conversion, Cash Flow could follow gamma distribution with similar parameters but different magnitude.

DAU may follow gamma distribution with different parameters.

We could use Items, Trans, Conversion to predict Cash Flow.

(3). Regression

Run OLS regression with full model.

OLS Regression Results

Dep. Variable:	Cash Flow	**R-squared:**	0.912
Model:	OLS	**Adj. R-squared:**	0.911
Method:	Least Squares	**F-statistic:**	1056.
Date:	Sun, 15 Mar 2020	**Prob (F-statistic):**	0.00
Time:	23:31:43	**Log-Likelihood:**	-5004.8
No. Observations:	724	**AIC:**	1.003e+04
Df Residuals:	716	**BIC:**	1.006e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	276.8374	191.304	1.447	0.148	-98.747	652.422
TRU	-0.0011	0.000	-3.738	0.000	-0.002	-0.001
DAU	0.0095	0.005	1.817	0.070	-0.001	0.020
Items	0.1058	0.004	28.046	0.000	0.098	0.113
Trans	-0.1178	0.062	-1.905	0.057	-0.239	0.004
Conversion	78.1021	48.145	1.622	0.105	-16.419	172.624
Return Customer	14.0061	1.991	7.035	0.000	10.098	17.915
Time Spend Per Day(seconds)	-1.7952	0.842	-2.131	0.033	-3.449	-0.141

Omnibus:	433.186	**Durbin-Watson:**	1.243
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	47025.983
Skew:	1.735	**Prob(JB):**	0.00
Kurtosis:	42.330	**Cond. No.**	1.56e+07

Adjusted R square reaches 0.91 which means model fits pretty well.

F test is rejected, indicating overall variables are statistically significant.

TRU, Items, Return Customer and Time Spent Per Day reject T test under significance level 0.05.

They exert statistically significant influence on Cash Flow.

(4). Strategy

To increase Cash Flow:

Launch features which could increase Items, Return Customer and decrease TRU, Time Spent Per Day.

Make use of three month's seasonality trend in Cash Flow, considering causations among Items, Trans and Conversation.

Implement different strategies according to different clusters

Thank you for your time!