Progress report on Business Intelligence Data Challenge

Done by Kirill Degtyarev

I have analyzed data related to e-commerce. Two CSV files represented data. The primary tools for discovering patterns and interesting insights was Python 3.6 Jupyter Notebooks, and for some visualization tasks, I used Tableau 2020.2.

Extracting and transforming data

Data was relatively clean, and no problems occur with extracting them to a Jupyter Notebook environment. The only column that contains NaN values was User_ID in table A. Since this column indicates string values of customers' ID, interpolation of missing data was impossible and was omitted during transformation. Other columns were clean, congaing appropriate to their description values. The only nuisance was that table A is presented in "wide" format, where each row corresponds to unique Conc_ID(ID of a transaction), while table B is in "long" format. For merging data, I wrote table_long_to_wide() function to transform table B into a wide format for convenient merging with table B with pd.merge() function. NaN values which appear to the IHC_channel was substituted by 0 since IHC_Conv value is in a range between 0 and 1 and summing up across all channels gives 1, then missing values could be assigned to 0.

KPIs

Before further analysis, I determine the main goals and KPIs that stakeholders might be interesting for stakeholders. In the circumstances of this challenge, it's not possible to ask those stakeholders directly. In the list of hints to this challenge, there are several KPIs to consider listed: **RevenueRevenue, number customers, a fraction of return customer**. I add to this list another KPIs that stakeholders might be interested in depending on their goals:

Market capture strategy:

- Number of customers
- Number of return customers
- Number of transactions

Increasing profitability/efficiency:

- Revenue
- Number of days between re-transactions(repeated)
- Effectivity of channels

Customer Retention Strategy:

- Increasing of loyalty
- A fraction of return customers

Descriptive analytics

My next part of the work is going through descriptive analytics. Since in this part, I have to consider different variables grouped by another variable (i.e. Conv_Date grouped by User_ID) I wrote a class DescriptiveStatistics that allows me to automate some work. A total number of

customers (unique User_ID values) equal to 79615, while customers bought something more than once are 8918 (fraction of return customers = 11,2%).

Using different aggregate functions for grouping Revenue by User_ID, I obtain several interesting insights:

- Histograms of all aggregated values have very long tails, meaning that there are abnormal values (it could be seen from the plot, as well as that value of 95% percentile is too small compared to a max value significantly positioned in the right tail.
- Empirical PDF skewed significantly to the right for each aggregated value (skewness >> 0), meaning that the majority of customers spends small sums of money (Mode < Mean). The probability mass distributed significantly on the right tail. Also, we observe that skewness of sum [RevenueRevenue] is much higher compare to other aggregated values. It means that the most noticeable contribution in the distribution in the right tail is made mostly by returned customers. So, from the stakeholders' perspective, it is better to try to have return customers (increase loyalty) than try to make the average sum of transaction bigger.
- Mean of the max '[Revenue]' (4596.476) is significantly lower than mean of the sum '[Revenue]' (29117.303), showing the difference between strategies aiming at increasing value of the transaction and increasing loyalty of clients (total of the RevenueRevenue from return customers).
- Kurtosis of the empirical PDF is much more significant than kurtosis of normal distribution (3), meaning that the peak of the ECDF is very sharp and therefore significant probability mass located in tails.

After I took a look at the Revenue of the return customers aggregated by the User_ID and highlighted:

- **Mean and max values of return customers of Revenue are higher than the mean and max values of the general sample**
- Kurtosis and skewness are lower for the return customers, meaning that the behaviour and predictions related to Revenue are more robust.
- **Comparison between the Revenue of the general sample and sample of return customers shows the importance to take into account the loyalty KPIs of customers rather than only KPIs related to absolute values.**

Diagnostic analytics and elements of predictive analytics

Performanceance and impact of channels over time

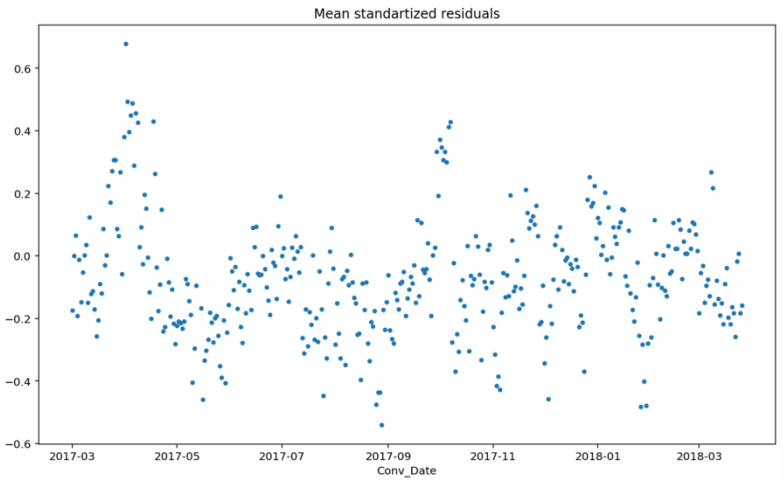
The impact of the channels could be understood as a measure of characteristics of channels(in our case IHC_Conv) on some KPI. For primary KPI, consider Revenue over time. Evaluation of such impact could be coefficients in linear regression, where independent variables are values IHC_Conv and the dependent variable is the Revenue.

Firstly, I considered histogram of the Revenue and their corresponding Q-Q plot and then applied log transformation to the depended variable (Revenue). After the log transformation, I got the Revenue distributed about normal.

Secondly, I evaluate an OLS model with IHC_Channel variables as independent variables and obtain such results:

- From the results of the Durbin-Watson test, we observe that a hypothesis of absence of autocorrelations between residuals does not reject (since Durbin-Watson values around 2).
- Unfortunately, both tests (Omnibus and Jarque-Bera) checking normality of residuals shows that residuals are non-normal distributed(since the p-value is about zero, meaning that we can reject null hypothesis without significant probability to get 1st oder failure(to accept the wrong hypothesis)). It means that exogeneous variables (in our case IHC_Conv of some channel) distributed non-normally.

Additionally, I did the Breusch-Pagan test to be sure in Homoscedasticity of residuals. Based on the p-values statistics, I can reject the null hypothesis on a 0.01 level of confidence, meaning that there is heteroscedasticity in residuals. To plot a representative graph (number of residuals - 79615, while time dates - only 396) I aggregated all residuals by Conv_Date and for each date calculate the mean value of standartized_residuals:



It could be explained quite intuitively, the more significant number of customers make transactions, the bigger would be differences between their sum of transactions.

To obtain trustful results, I applied robust evaluations of a covariance matrix and got quite good results:

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.991
Model:	OLS	Adj. R-squared (uncentered):	0.991
Method:	Least Squares	F-statistic:	3.817e+05
Date:	Tue, 23 Jun 2020	Prob (F-statistic):	0.00
Time:	16:43:42	Log-Likelihood:	-56782.
No. Observations:	79615	AIC:	1.136e+05
Df Residuals:	79593	BIC:	1.138e+05
Df Model:	22		
Covariance Type:	nonrobust		

As a measure of channel importance, I considered the values of a coefficient of the corresponding channels. The coefficients sorted in descending order **give a top list of most influential channels: T, A, C, E, B.**

In the last part, I did a customer segmentation. Customer segmentation is mostly essential for group separation, where objects(in our case customers) are similar to each other based on

some metrics. This group separation could be used further for building classification or predictive algorithms on each group separately. Therefore the quality of these algorithms should be higher than the quality of algorithm on a general sample.

From the previous steps of analysis and the list of KPIs, we conclude that the most important features for distinguishing customers between groups are:

- The mean revenue of transaction
- Number of transactions
- Mean value of IHC for each channel
- Number of days between the first transaction and the last transaction(for those who bought a product only once this value is 0)

I used K-means++ and DBSCAN algorithm. For the K-mean algorithm, I determined hyperparameter (k- number of customers) by plotting an Elbow graph and looking at the Silhouette graphs. Kmeans algorithm gives 23 clusters.

For the DBSCAN algorithm, I did a clustering analysis using Nearest Neighbour to determine eps parameter (radius of nest elements) for DBSCAN. DBSCAN shows 70 clusters and distinguishes anomaly values to dedicated class -1 (total number of them is 5914).

To the list of interesting insights, I would like to add results from providing cohort analysis in Tableau:

Cohort analyzis of retention

			Month to repeat transaction														
Year of Firs..	Quarter of ..	Month of Fi..	Null	0	1	2	3	4	5	6	7	8	9	10	11	12	Grand Total
2017	Q1	March	48,22%	8,27%	20,05%	6,99%	4,63%	2,88%	1,60%	1,80%	1,69%	1,64%	0,54%	0,49%	0,61%	0,58%	100,00%
	Q2	April	60,62%	5,84%	6,50%	5,73%	5,34%	2,48%	3,76%	4,09%	2,28%	0,87%	0,96%	0,72%	0,83%		100,00%
		May	74,33%	2,98%	4,89%	4,55%	2,76%	3,70%	2,29%	1,82%	1,00%	0,56%	0,72%	0,41%			100,00%
		June	84,66%	2,66%	3,90%	2,19%	2,21%	1,48%	1,19%	0,42%	0,53%	0,53%	0,24%				100,00%
	Q3	July	87,93%	2,02%	2,51%	2,34%	1,63%	1,58%	0,52%	0,59%	0,47%	0,42%					100,00%
		August	88,34%	2,01%	3,72%	2,51%	1,42%	0,53%	0,47%	0,56%	0,44%						100,00%
		September	91,04%	2,59%	2,15%	1,93%	0,55%	0,86%	0,55%	0,33%							100,00%
	Q4	October	93,91%	1,37%	2,56%	0,72%	0,70%	0,34%	0,40%								100,00%
		November	95,38%	1,49%	1,24%	0,75%	0,58%	0,56%									100,00%
		December	95,60%	1,07%	1,42%	1,16%	0,75%										100,00%
2018	Q1	January	96,16%	1,17%	1,75%	0,92%											100,00%
		February	97,20%	1,01%	1,80%												100,00%
		March	99,32%	0,68%													100,00%
Grand Total			84,10%	2,78%	4,26%	2,47%	1,79%	1,17%	0,98%	0,92%	0,59%	0,35%	0,23%	0,16%	0,16%	0,05%	100,00%

A number of return customers continuing to fall. It is related to each month. Thus, we can conclude that the company has problems with retention of their customers. And Journey statistics also showed as well as Cohort analysis of retention that different groups of return customers (with different level of retention) continuing to fall for the whole time.

The results presented in this reported are non-exhaustive, and others could be found in Jupyter Notebook and Tableau workbook.