CSE 519 -- Data Science (Fall 2018)
Prof. Steven Skiena
Homework 3: Data Integration and Modeling
Due: Tuesday, October 23, 2018

This homework will investigate data integration and model building in IPython. It is based on the Google Analytics Customer Revenue Prediction Kaggle challenge, revolving around predicting how much the GStore customer will spend. You are charged with predicting the natural log of the sum of all transactions per user. For every user in the test set, the target will be:

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

The dataset has the following fields:

fullVisitorId- A unique identifier for each user of the Google Merchandise Store.
channelGrouping - The channel via which the user came to the Store.
date - The date on which the user visited the Store.
device - The specifications for the device used to access the Store. This field is a JSON with additional information such as browser, operatingSystem, deviceCategory etc.
geoNetwork - This section contains information about the geography of the user. This JSON field contains subcolumns such as continent, country, region etc.
sessionId - A unique identifier for this visit to the store.
totals - This section contains aggregate values across the session. This is also a JSON field with a column transactionRevenue whose value is to be predicted.
trafficSource - This section contains information about the Traffic Source from which the session originated.
visitId - An identifier for this session. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
visitNumber - The session number for this user. If this is the first session, then this is set to 1.
visitStartTime - The timestamp when the session started

Many of the tasks mirror those of the previous assignment, as practice makes perfect. As in the previous assignment, you will need to submit all your results in a single google form and your code files in three different format (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you needed to produce the resulting tables and figures.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Tasks (100 pts)

1. Take a look at the training data. There may be anomalies in the data that you may need to factor in before you start on the other tasks. Make a note of the anomalies that you notice. Clean the data first to handle these issues. Explain what you did to clean the data (in bulleted form). (10 points)
2. Generate a heatmap and two other plots (with a subset of variables) visualizing interesting positive and negative correlations. Explain the reason for your choice for these variables and any interesting results associated with them. (15 points)
3. Cluster the data based on geographic information available with a subset of variables that you find relevant. Include a visualization plot. Describe your inferences from the clustering and discuss their significance. (15 points)
4. Define a buying score or probability function for each user, which predicts the likelihood of a user buying a product from the GStore. Rank the ten most likely users as who will buy a product from the store. Does it seem that you that it produces good results? Report why or why not. (15 points)
5. Identify at least one external data set which you can integrate into your transaction prediction analysis to make it better. Discuss/analyze the extent to which this data helps with the prediction task. (10 points).
6. Finally, build the best prediction model you can to solve the Kaggle task. Use any data, ideas, and approach that you like. Submit the results of your best models on Kaggle. Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation. (20 points)
7. Do a permutation test to determine whether your model really benefits from each input variable you use. In particular, one at a time, for each relevant input variable, permute the value of this variable and see how they impact the accuracy of the results. Run enough permutations per variable to establish a *p*-value of how good your predictions of log of sum of transactions per user are. You can use whatever metric you wish to score your model (like mean absolute error). (15 points)

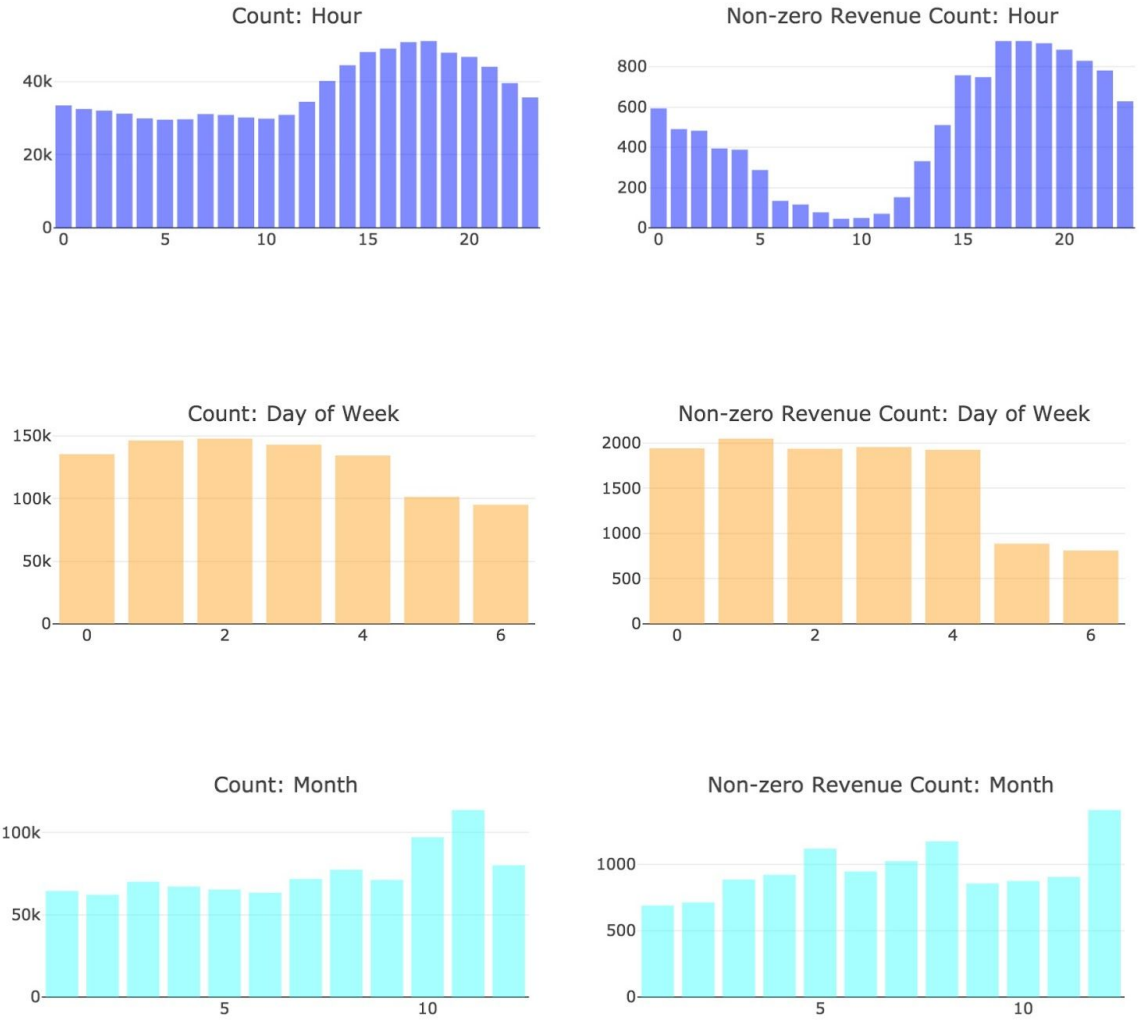
Task 1:

1) Explain what you did to clean the data. List each step/method as a separate item.

- There are 4 columns in the Train and Test Data which are in JSON format. Normalized these columns and added the different extracted columns as features.
- The column Campaign Code that is present in the Train Data but absent in the Test Data. Hence, dropped this column from the Train Data as well.
- The column Session ID uniquely identifies all the tuples and hence does not add any new information to each entry. Hence, dropped this column from both Train and Test Data.
- There are 19 columns in the data which have the same value for the entire dataset. Dropped all these columns from both Train and Test Data.
- There are many Categorical Columns in the data. Converted these columns to Integer Values.
- Converted Transaction Revenue to Float.
- Converted other numerical values in Object form to Float.
- Replaced the NAN values in Transaction Revenue column with 0.
- Extracted the Date column to obtain ‘Hour’, ‘Day of Week’ and ‘Month’. Added these columns to the feature set.

Task 2:

2.1.1) Upload your heatmap (or other exciting plot).

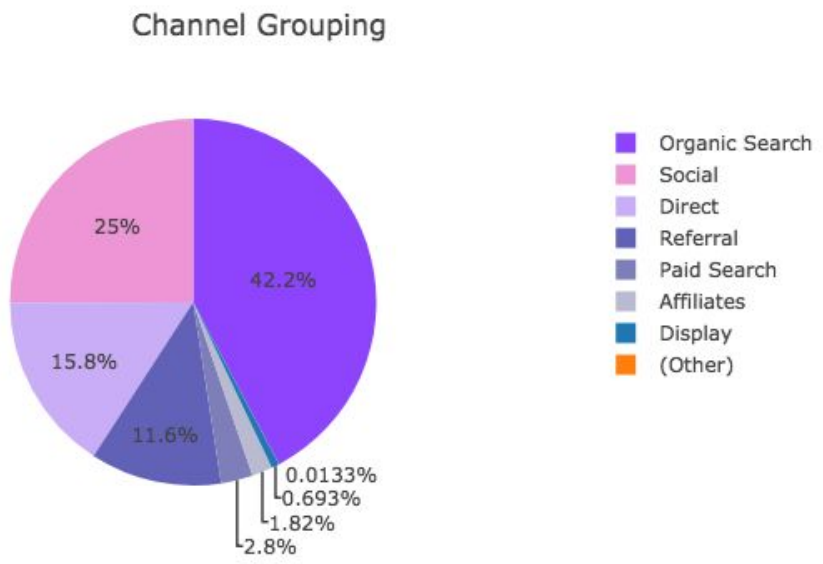


2.1.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.1.1)

- The above plot describes the Count (no.of times the user visited) and Non-Zero Revenue Count (no.of times a user actually made a transaction) Vs Device Category and Device Operating System.
- Device Category
    - An interesting observation here is that users visit from all the 3 categories Desktop, Mobile and Tablet but transactions are usually made from Desktop.
    - This aligns with the usual scenario where people generally scroll through both desktop and mobile but prefer to make purchases via desktop rather than Mobile/Tablet due to possible reasons like Comfort, Security reasons etc.
- Device Operating System

- A very interesting observation here is that the No.of Views is highest in Windows rather than Max and also higher in Android rather than IOS. But, the no.of transactions of Mac users is very high compared to Windows users. Similarly, no.of transactions of IOS users is higher than Android.
- This may generally shows that Mac, Iphone owners may be a little more able financially as compared to those of Windows and Android owners. Being more able financially the no.of times their visits actually turn to transactions is higher whereas many users owning Windows and Android make a lot of visits but not much of them convert to actually transactions.

2.2.1) Upload a second plot of your choice.



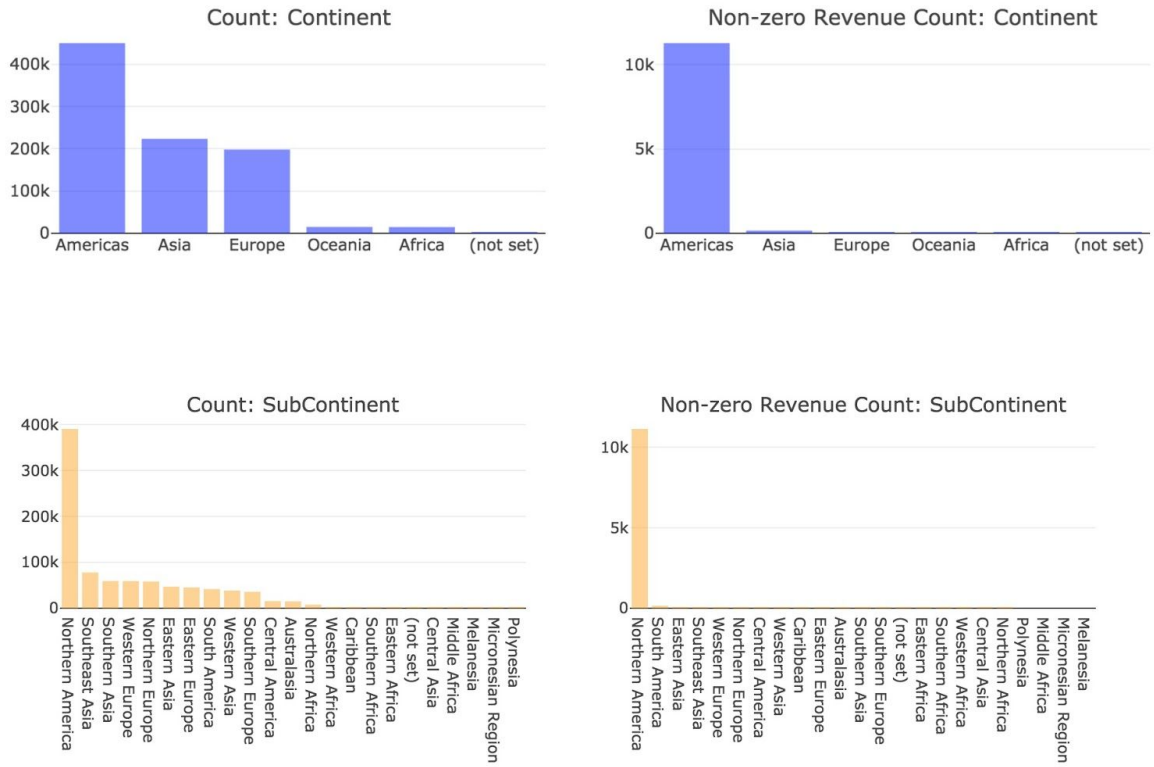
2.2.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.2.1)

- The above plot describes the Count (no.of times the user visited) and Non-Zero Revenue Count (no.of times a user actually made a transaction) Vs Hour of Day, Day of Week and Month.
- Hour of the Day
  - We see that there are more no.of visits and more no.of transactions in the evening. This seems like a very plausible pattern which could be due to the fact that most people are free during the later part of the day (considering work during the day)
  - Also, we can see that there are more no.of visits as compared to the no.of transactions during the afternoon part of the day. This could be that many people

view and choose their options during their small leisure time during the afternoon but don't actually make the transaction until they are back home.

- Day of the week
    - There is quite a similar pattern here between the no.of visits vs the no.of transactions, but interesting thing is that during the weekend the no.of visits is higher than the no.of transactions.
    - The decrease in no.of transactions during the weekend could possibly be because people tend to spend time away from home during the weekends (say on a vacation or so).
- Month
    - We can clearly see that the no.of transactions is very high during December, which could be due to Christmas and New year etc.
    - Also, we can see that during November the no.of visits is very high, which is during Thanksgiving/Black Friday and the no.of transactions is very high taking into account November end and December start which overlaps with Thanksgiving/Black Friday.
    - We also see that during May and August the no.of transactions is higher than the no.of visits which could be due to the start of summer holidays during May and start of schools/college during August.

2.3.1) Upload a third plot of your choice

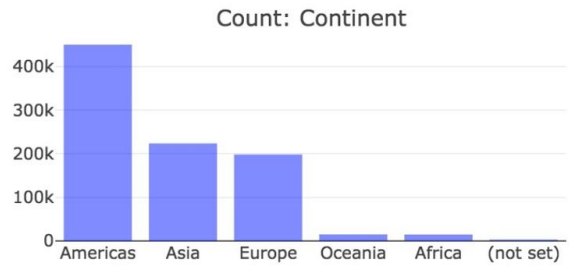


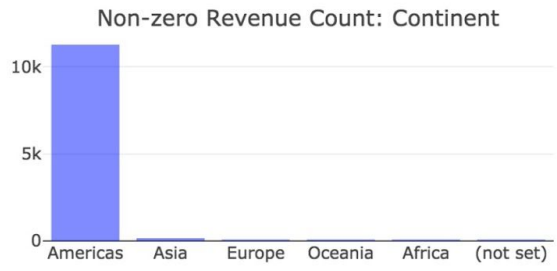
2.3.2) Explain the reason for your choice for these variables and any interesting results associated with them. (based on 2.3.1)

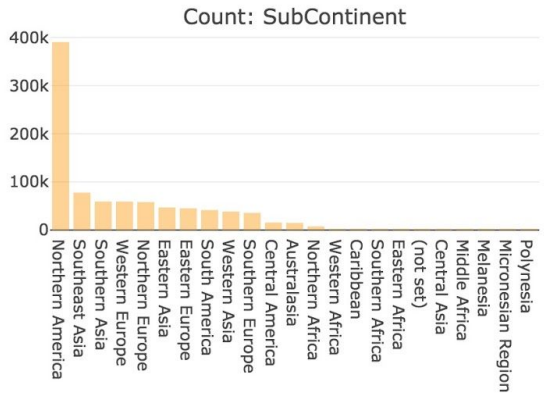
- The above plot described shows the different Channeling Groups through which people visit the store.
- We see that most no.of people (42.2%) visit the store using 'Organic Search' which is through unpaid search results. And second highest (25%) is via Social like Facebook, Twitter, Youtube etc.
- The other two higher methods are through Direct (15.8%) where users navigated via URL and also Referral (11.6%) which is users who landed on Google Store by clicking on a link in some other site other than the main ones.
- All the above are true in a day to day scenario where most traffic to any website is through the above 4 methods, proving this observation to be quite interesting.

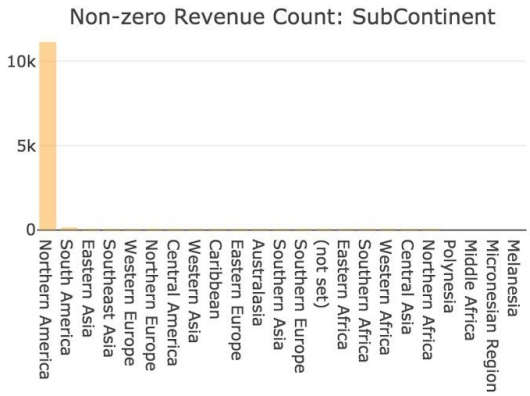
Task 3:

3.1) Upload a visualization of the clustering.









3.2) Describe your inferences from the clustering and discuss their significance. (based on 3.1)

- The above plot describes the Count (no.of times the user visited) and Non-Zero Revenue Count (no.of times a user actually made a transaction) Vs Continent, SubContinent.
- Continent
    - From the above plot we can observe that America has both higher no.of visits as well as no.of non-zero transactions.
    - Asia and Europe also have high no.of visits but the no.of non-zero transactions is very less.
    - This inference is plausible owing to the fact that America has higher GDP (meaning there are high no.of people who are able financially) as compared to that of Asia and Europe.
- SubContinent
    - Here as well we see a similar pattern as above where North America has both high no.of visits and high no.of transactions.
    - A similar inference as above can be done here as well.

Task 4:

4.1) Rank the ten most likely users who will buy a product from the store. List them here.

fullVisitorId	count	count of non-zero revenue	mean
0608915197735218105	17	13	2.335615e+08
4984366501121503466	24	16	5.946188e+08
3857043812510146001	9	6	8.878500e+07
2411322974724385937	11	7	2.970000e+07
0777922178356486144	10	6	5.311667e+07
8657427332734176422	10	6	1.259100e+08
7463172420271311409	16	9	8.027889e+08
119870259714905967	9	5	7.978420e+08
2446685875964479851	11	6	6.934600e+08
6147396474895233852	10	5	6.530560e+08

fullVisitorId	probability
0608915197735218105	0.764706
4984366501121503466	0.666667

3857043812510146001	0.666667
2411322974724385937	0.636364
0777922178356486144	0.600000
8657427332734176422	0.600000
7463172420271311409	0.562500
119870259714905967	0.555556
2446685875964479851	0.545455
6147396474895233852	0.500000

4.2) Does your model produce good results? Why or why not? Explain the rationale behind your buying probability function.

- Grouped the users based on Full Visitor ID and considered the probability of them making a transaction as
    - Total no.of Non-zero revenue transactions / Total no.of Visits.
- The rationale behind using the above probability function is that a user who makes a purchase most of the times he visits is most likely to buy a product from the store rather than someone who visits many times but makes a transaction very few times.
- One issue with the above function is that a user who visits the store 1 time and makes the transaction during his visit, he gets a probability of 1 making him seem like a user who would buy a product from the store.
- But intuitively, we do not have much data about this user to say that he will likely be among the to 10 ranked users. Hence, to overcome the above considered users who made transaction atleast a certain no.of times (say 8, reported using this above). Then upon these users whose the probabilities are sorted and the top 10 probabilities among them are considered as the top 10 users.

Task 5

5.1) Identify at least one external data set which you can integrate into your transaction prediction analysis to make it better.

- No.of Internet Users per Country
  https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users
    - Our current problem to predict a Google Store customer's revenue deals with Internet as all the transactions happen over the internet.

    - Hence, trying to integrate the no.of internet users per country could be a significant insight into the no.of transactions being performed by that country in turn helping us produce our predictions better.
    - In short, a country with less no.of internet users could mean that it produces less no.of transactions on the whole.
- Population per Country

  https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)
    - Dealing with people, it could be possible that a higher population in a country could lead to a higher no.of transactions.
    - Hence, trying to integrate this into our feature set might provide some interesting results.

Obtained both the above datasets from Wikipedia

5.2) Discuss/analyze the extent to which this data helps with the prediction task.

- Integrated both the above features into my model but observed that they did not improve the results substantially.
- Reason for the above could be the our available data is very skewed with respect to the country with nearly 80% of the entries being from USA meaning that most of the data samples get the same value for the No.of Internet Users and Population columns.
- Thus, they may not significantly add anything to the existing feature set.

6.1) Report your best achieved rank

6.2) Report the score you received for your best rank

6.3) Report the total number of entries you made during the course of this challenge

6.4) Include a snapshot of your best score on the leaderboard as confirmation

Task 7:

7.1) For multiple (not necessarily all) relevant input variables, permute the value of each variable and see how it impact the accuracy of the results. Report the results below.

pval for totals_hits

[0.0]

pval for visitId, totals_newVisits, geoNetwork_continent, trafficSource_adwordsClickInfo.isVideoAd
[0.0, 0.76, 0.48, 0.85]
pval for totals_pageviews and trafficSource_campaign
[0.0, 0.79]

7.2) Explain your process during the permutation test as well as your findings.

- Considered top few and bottom few features based on their significance. (The significance of the features is obtained after running LGBM and obtaining the base RMSE.)
- For each of these columns, permuted the values of the column (with the rest of the data remaining the same) and ran LGBM on this permuted new data.
- Ran 100 permutations on each column considered above and obtained their RMSE values.
- Now, compared these RMSE values with that of the base RMSE (which was obtained with the original correct data).
- Counted the no.of times the newly obtained RMSE is lesser that the base RMSE and obtained the p-value as
    - (No.of times the RMSE of the permutations is <= Base RMSE) / Total No.of Permutations
- Ideally, we would want that none of the permuted RMSE's are lesser than the base RMSE as that would mean that they are not truly contributing to the prediction.
    - P-value <= 0.05 implies that the feature did not occur by chance, else it occured by chance.
- Observed that the p-values for highly significant features like are 0 (saying that none of their permutations produce better RMSE meaning they are highly significant), whereas least significant features produce really high p-value meaning that many of their permutations generate better RMSE meaning that it occured by chance.