Executive Summary

Overview

The goal of this project is to be able to predict revenue per customer by analysing the data generated by the GStore platform, Google's branded products shop that operates solely online through an e-commerce.

The 80/20 rule is a principle that is fairly common in business as it is in computing, economics or sports: 80% of the OPEX come from 20% of the product lines, 80% of the customers come from 20% of the customer segments or, in this case, 80% of the revenues come from a small percentage of customers. It is critical for the decision maker to be aware of what areas of their business are the main revenue drivers and data science can become a powerful tool to identify these.

Results

The analysis identified a high peak in GStorere's revenues around mid December, which could be explained by the Christmas season, widely followed by North America's population.However, that could potentially introduce a bias toward the region of North America.

Also, North America not only has a higher number of counts but also has the highest number of non-zero revenue counts. Comparatively, Asia and Europe have lower number of non-zero revenue counts. By characterizing customers based on the traffic source, we would that Youtube has high counts but very low in the non-zero revenue counts.

Conclusion

As we are treating an ecommerce platform that operates 100% online, collecting information about the user (demographics like age, gender, location, etc) about the transaction (items bought, time of the purchase, price) and the customer journey (time spent on the webpage, number of items seen, number of clicks, etc) will be straightforward. Specific efforts should be done to get a well organized data set where variables are well labeled and the numbers collected aren't flawed/wrong. We do not foresee any special challenge in the data collection as the sector is luckily a "data/metrics centric" space.

Recommendations

Our recommendations are that all the high selling/ high willingness-to-buy demographics and the high-revenues dates and product categories are subject to several A/B testings with some of the features highlighted by the prediction model. Examples of this A/B testing could be to prioritize the first 5-6 products seen by the customer depending on the amount of clicks they did on specific categories and similar products. This A/B testing could be deployed on a few thousand customers and could be useful to validate the model's predictions with real transactions before implementing it at a large scale.

The GStore could use those learnings to fine-tune their monetization strategy: they would be able to be more efficient in their advertising, positioning and web-design strategies. At the same time, if the learnings received from the analysis can be extrapolated to other e-commerce platform, this intelligence could be sold under the form of consulting projects and external advisory.

Project Description

Project Overview

The objective of this project is to understand how a thorough analysis of the Google Shop's can give insight on how the customer behaves when using the platform, what products are they most interested in and which are the main drivers of their purchasing decision.

Model development and operational details

We used a Light GBM model to go through the data gathered in our final case. As Light GBM model grows vertically, the best ways to avoid overfitting it is to decrease the maximum depth of the tree. Due to its sensitivity to overfitting, we can use the fact that the number of leaves should be less than equal to 2^(maximum depth). For faster speed parameters, bagging fraction is used along with feature fraction to allow only selected iterated parameters in the next round. The model allows for faster training considering the data each for test and training is 1 GB of size. There is lower memory usage and parallel learning is supported.

There are many ways to allow for better prediction. Firstly, attaining more data. This will allow for more clarity in what we are predicting. Secondly, cleaning and resampling data will allow us to work with smaller but more precise information for better modelling. Another technique that can be implemented is transforming the data like the way it has been done here. Instead of taking the data the way it comes, we can use a function like the log or exponential to work with simpler numbers.

Although we haven't implemented much feature selection or engineering in this problem, there is immense scope in doing so to improve our performance. Lastly, revising the scope of the problem helps look at data from a different perspective which is always needed in this field. If all else fails, there are algorithmic techniques like baseline performance which we have tried implementing here by looking a random weak algorithm on the data and setting that as the calibrated zero for any model to perform better than that. Parameters have been tuned using bagging fraction which is very specific to the model. We also plan to look at ensemble models to make more complex decisions about the data.

A better prediction can always lead to overfitting and the above methods mention in some way how to avoid it. The metric used is RMSE and as the LGBM model is sensitive to overfitting there is always a possibility of that happening. To keep it as fair as possible, we witness the trainman and test set and make a few analysis to see how distinct they are. This will allow us to test the model on a test set not seen before allowing for a fair evaluation.

Key Results

When a company is able to identify which are the relevant factors affecting their business problems it can harness the power of data to improve the management's decision-making process.

Some businesses (and particularly their managers) do not completely trust a data-based approach, since they doubt the quality and the sources of the data are optimal.

From the transaction revenue analysis, the 80/20 rule identified can be used to help decision makers from the marketing teams to make investments in promotional strategies. In our case, we predicted the natural log of the sum of all customer transactions. In our data analysis, we evaluated the ratio of revenue generating customers to customers with no revenue and the ration of unique visitors and common visitors.

To make better data-driven decisions, we characterized our data sets by different categories to better predict a more accurate revenue generation, such as the device information (i.e. device browser distribution, device operating system). We also did data exploration by using the training data from August 1st, 2016 to July 31st, 2017 to see the non-zero revenue count over the year. We also characterized data by the geographic information and network domain to better do the revenue prediction.

Future Recommendations

Our main recommendation is that Google Shop tries to translate this analysis learnings into a real-life A/B testing phase: testing if changes in the product categories, if giving the highest selling categories more relevance on the web or if targeted advertising campaigns right before special dates (Thanksgiving, Christmas, Black Friday, etc) could help in boosting the shops revenues.
As we have mentioned several times in class, data is king, and further data collection could allow us to achieve deeper levels of granularity, which would in turn become highly detailed recommendations and impact assessments on the potential actions our client, GShop could act upon.

An idea that was recurrent after going through our analysis is that it would be very interesting if we could cross the e-commerce data with additional browsing history gathered through cookies. If the websites visited, the amount of time spent in them and the number of clicks/path to these websites could be monitores, this might provide an extremely valuable insight to the e-commerce revenue optimization.

Learning Component

The major learning after going through this final case was starting to use our recently-acquired data-science knowledge when executing or auditing an analytical project that is closely related to a financial challenge.

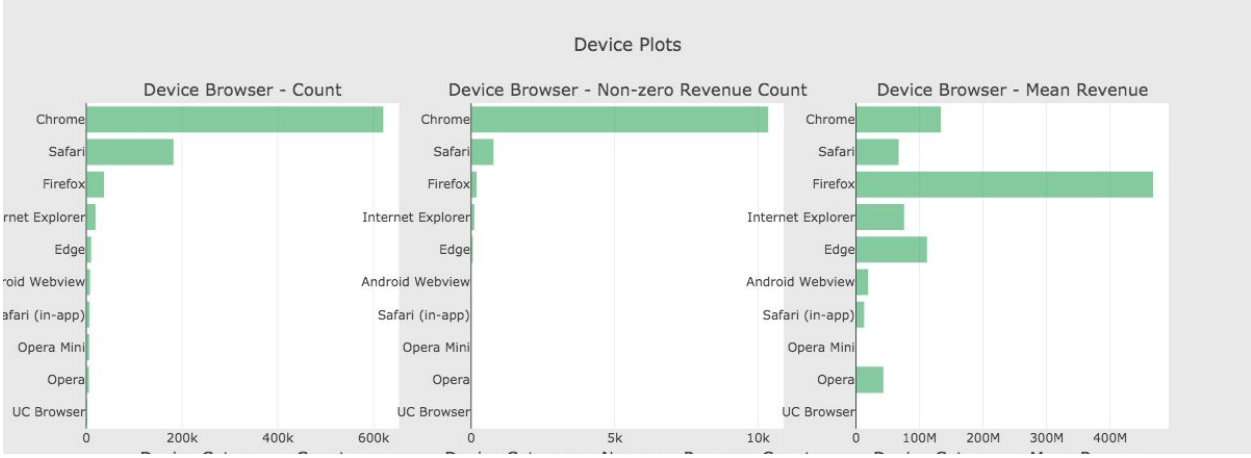
Working in a group of engineers and business people was an extremely enriching experience, as both profiles happened to be particularly compatible when approaching this final case. While the technical side was mastered by Prithiv and Crystal, engineers with an deep expertise in data analysis and python programming, Oriol brought the business and financial acumen to understand how the problem needed to be presented and how the solutions had to be shown.

It was a great way to test our understanding of feature selection, overfitting and prediction model implementation with a problem that might be particularly complex for our levels of expertise.

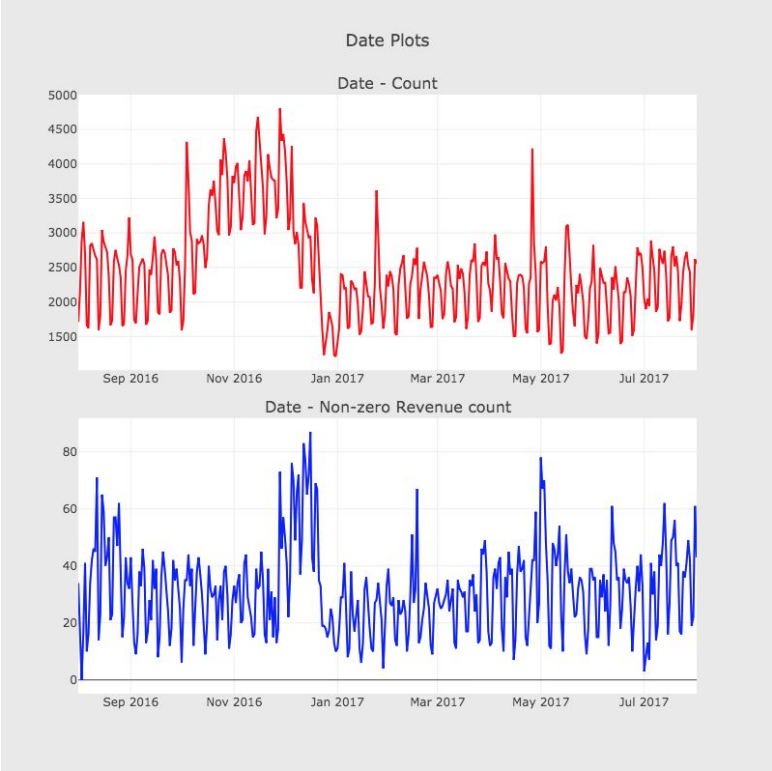
EWMBA & MBA 296 LEC 5B – Data Science Applications in Finance and Accounting
Prof. Donatella Taurasi
Team Epsilon

Exhibits

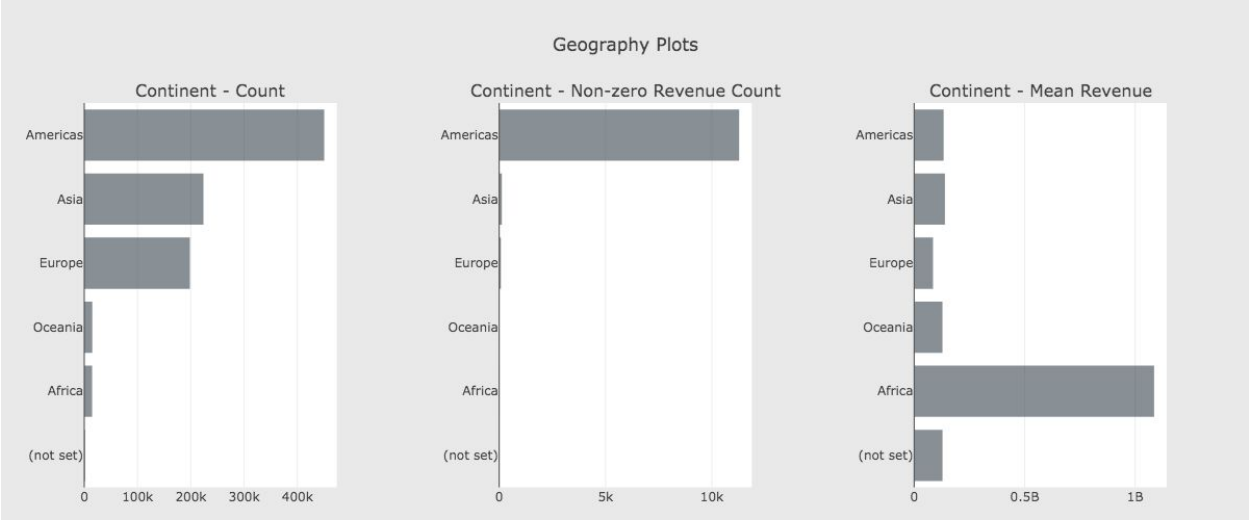
Device plots based on count, non-zero revenue and mean revenue:



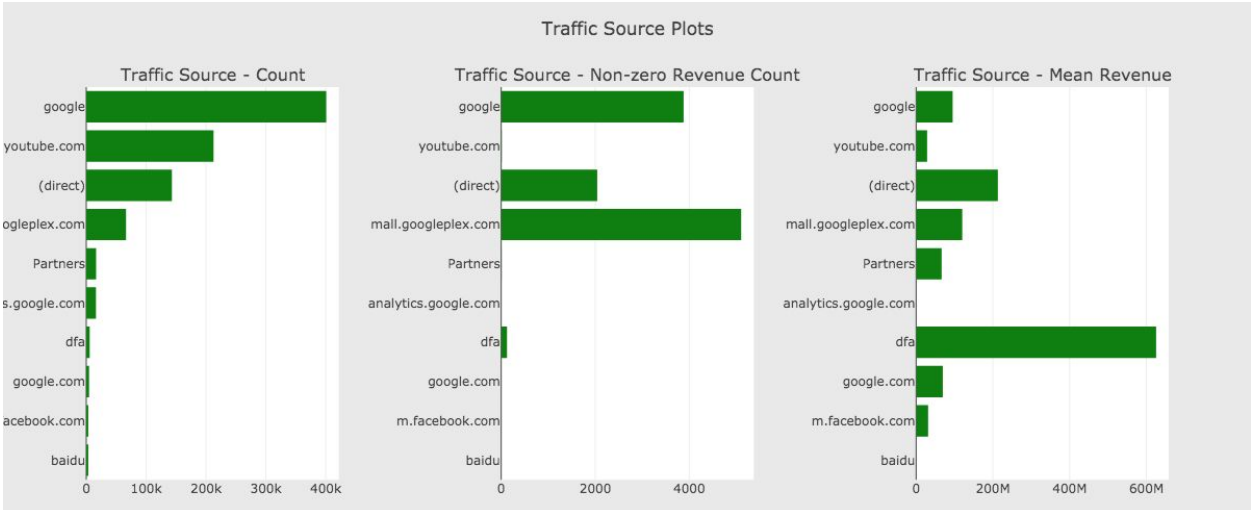
Date plots based on on count and non-zero revenue count:



Geography plot based on count, non-zero revenue and mean revenue:

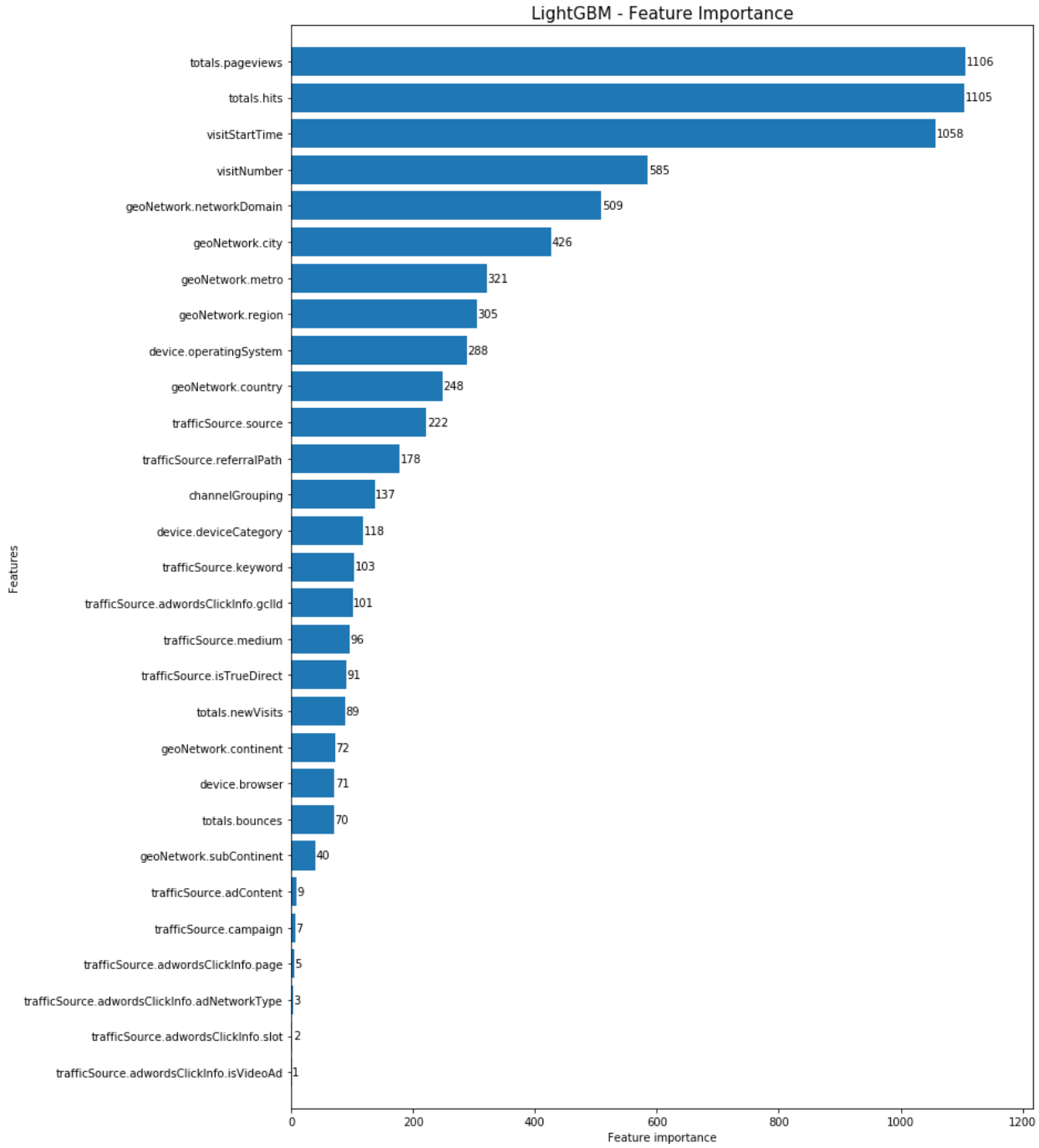


Traffic source plot based on count, non-revenue count and mean revenue:



EWMBA & MBA 296 LEC 5B – Data Science Applications in Finance and Accounting
Prof. Donatella Taurasi
Team Epsilon

Feature importance after running LightGBM model (+15 features):



Explanation of the main features used in this model:

- *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.
- *geoNetwork* - This section contains information about the geography of the user.
- *socialEngagementType* - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
- *totals* - This section contains aggregate values across the session.
- *trafficSource* - This section contains information about the Traffic Source from which the session originated.
- *visitId* - An identifier for this session. This is part of the value usually stored as the _utmb cookie. 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 (expressed as POSIX time).
- *hits* - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
- *customDimensions* - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
- *totals* - This set of columns mostly includes high-level aggregate data.