Improving Criteo's Retention by Churn Analysis and Customer Clustering

MBD - Research Methods in Business Analytics II

Summary

- **Main Objectives:**
    - Assessing Criteo's client churn rate based on two measurements of the concept: i) quarterly, via cohort analysis; ii) across the time frame of the dataset (i.e. from Q1 2017 to Q1 2020).
    - Predicting churn date of any given customer in the dataset via a churn analysis model.
    - Identifying customer retention strategies by analyzing the top-performing clients by tenure (duration with Criteo) in the US, France, and Japan. The methodology and main insights found can be extended to other geographies and industries.
    - The ultimate goal is to develop a tailored retention strategy for each of Criteo's clients.

1) Problem and Motivation

Churn analysis is relevant for Criteo for two main reasons. First, customer retention ensures recurring revenues, therefore provides stability to the business. Second, customer retention offers significant opportunities to upsell and cross-sell Criteo's product portfolio.

Focussing on upselling opportunities first. Loyal customers tend to spend more, therefore delivering more ads. This translates to the increase in data collection and, in turn, improved customer experience as Criteo keeps developing and enhancing its solutions. This has, of course, potential value for cross-selling, too. Indeed, satisfied, loyal customers might adopt other Criteo's products (e.g., awareness). In addition, new customers (e.g., new verticals) might be willing to use Criteo's solutions.

Thus, we believe that reducing attrition will positively impact two corporate priorities: "Strengthen the core" and "Expand product portfolio". Customer loyalty might also constitute a major source of competitive advantage for Criteo vis-à-vis its direct competitors.

2) Approach

The rationale behind our ideation approach was to identify a relevant business issue for Criteo and develop a tailored strategy/solution with the highest business value possible. We started working on three distinct sources (i.e., Criteo's datasets, corporate presentation, and annual report) at the same time to establish our focus, relying heavily on brainstorming and fact-based communication to generate and approve ideas.

Mining the data we realized there was an opportunity to calculate customer attrition and develop more in-depth analyses on the topic such as the prediction of churn date and the classification of clusters of clients based on variables related to industry and country. During the ideation phase we also verified the business relevance of churn according to Criteo's corporate priorities and annual report, which refers to churn as "a useful indicator of the stability of our revenue base and the long-term value of our client relationships" (Criteo, 2019, p. 104) as well as a major growth driver (p. 11). We here found the link with Criteo's corporate priorities to "Strengthen the core" and "Expand product portfolio". In other words, we decided to work on churn because we found evidence of value for the company and an excellent opportunity to work on the analytical skills and tools developed during the year.

The main steps we followed were:

1. Identification of the business problem (described above)
2. Data exploration to visualize the root problem regarding churn
    a. Total % of churn by total clients, country and industry
    b. Retention cohort
    c. Opportunity loss (in days) by country and industry
3. Data preparation
4. Modelling
5. Further analysis of the output from the models to create business strategies

3) Datasets

Three datasets were used to perform the required analysis. The data was shared by criteo as:

1. **criteo_data**: This dataset gives information about criteo performance from 03/2017-01/2020. Some calculated fields created by us are as follows:
    a. Unique Client: 'Account_id' and 'Client country' combined gives the unique client name.
    b. CPO (Cost per order): How much a client spends to get on sale driven by criteo. It is calculated as: spend_criteo_euro/sales_criteo
    c. COS (cost of sales): How much a client spends with criteo to get one euro back in revenue. It is a percentage and is calculated as: spend_criteo_euro /revenue_criteo_euro
    d. AOV: The average price of a sale driven by Criteo. It is calculated as: revenue_criteo_euro/sales_criteo
2. **onsite_data**: This dataset shows what happens on a monthly basis on the client's website/app. This is not Criteo related but contains information related to client sales, revenue, country etc.

3. **catalog_data**: This dataset shows how many products a client has in its catalog and can be sold. The field 'Products' gives the number of products available in the catalog, which can be sold by the client.

4) Tools & Analytics

Data exploration and visualization (Tableau and R)

1. Tableau to analyse retention cohort and number of churn by country, industry, and client type. In order to do so, we created several new calculated fields such as 'new/existing customer', 'start_date', 'churn_date', and 'tenure', using Level of Detail (LOD) function in Tableau.
2. R to visualize the churn rate, using 'survival', 'survminer', 'ranger',' ggplot2', 'randomForestSRC', and 'ggRandomForests' libraries. The model we used is called Kaplan-Meier, one of the non-parametric survival models. Taking 'churn' as censor and 'churn_date' as interval, we plotted the survival curve by country, industry, and cluster.

Data Preparation (Python)

1. Change the rows to the columns and aggregate by spend to feed the model with spend by devices, OS, environment, and criteo product. Removed outliers which were bigger than 3 z-score

Modelling (Python)

1. **Churn Date Prediction**:
   Used 'XGBRegressor' from 'xgboost' package and found the best parameter combination, such as number of estimators, maximum depth of trees, and minimum child weight, with Grid Search Cross Validation to minimize the RMSE. We rescaled the churn month to year so that the regression model can predict correctly and then scaled back to normal month form after the prediction. We got normalized RMSE with 0.389.

2. **Clustering by Country by Industry**:
   Used FAMD to reduce the dimensionality of our mixed data to 2, then ran K-means clustering. To find the optimal number of clusters, we created a function and maximized the silhouette score automatically. However we chose the number of clusters by looking at the scatter plot to make the distribution of each cluster better. Silhouette score is around 98-99%.
   Then, created a function to automatically identify the best and average client, the one closest to the cluster centroid, in terms of tenure within each cluster. Tenure is calculated by predicting the churn date with our previous model and subtracting the starting date. Finally, created a function to automatically display the difference between the best and average client.

5) Results

The data provided by Criteo revealed that customer churn is one of the biggest issues faced by the company, with over 40% churning in the three year period. To address this problem we constructed a model to predict churn dates. This model was created in Python by implementing a regression tree using the XG Boost methodology and resulted in a Mean Squared Error of 0.38.

The final output of this model is a tool in which the client can access any of its Account IDs and get the predicted churn date in return.

Due to the limitations of the data provided, it is difficult from a business point of view to understand the reason behind the churn, and therefore create strategies to increase retention. This problem was addressed by clustering the data in order to prioritize clients and create personalized strategies that may help improve retention.

After addressing all the previous quantitative issues, we concluded that Criteo should address four (4) key points from the business standpoint in order to reduce churn and increase retention:

1. Understand the reasons behind the churn: It is crucial for Criteo to understand the rationale behind why clients stop working with them. Possible reasons could be unachieved client expectations, speed to respond to client's requests, limited personalized offer of products, amongst others.

2. Identify clients at risk: By analyzing the patterns through machine learning models, the churn prediction model presented in this project can easily detect clients at risk, their current profitability for Criteo and their estimated churn date. This data will facilitate the prioritization of clients towards offering them personalization or increased customer service.

   Some of the important metrics to measure customer value and identify priority accounts are:

   - Cost to Serve (CTS): An analysis of all costs associated with delivering the service, compared with the profitability of the contract.
   - Lifetime Value (LTV): Prediction of how much money a customer will bring in over the course of its contract with Criteo.
   - Share of Wallet (SOW): calculates the percentage of a customer's spending that is allocated to a service. Indicates how much more can the account potentially spend with Criteo.

   This analysis will allow Criteo to recognize which clients are valuable and vulnerable, thus focusing retention techniques on them.

3. Group customers by needs groups:
   Criteo's data sets and customer base are so broad that it is not possible to have a one-size-fits-all approach to customer retention. Strategies to reduce churn should be tailored specifically to what Criteo knows about the individual account - hence the importance of qualitative feedback regarding the reasons for low retention. The latter is one of our major recommendations for the improvement of the models presented above.

   Once the churn reasons have been identified and transformed into a structured format, together with the data provided for the development of this project, the clustering model presented in this project could be improved. The final result is gathering customers in cluster groups that will facilitate the development of retention strategies in a semi-personalized way.

4. Develop customized digital strategies for each cluster:
   As customer offers and personalization is a tough task to accomplish client by client, clustering becomes a vital part of generating retention strategies that can be applied to

the overall client base. Two bibliographical sources were consulted for broadening our understanding of this topic:

According to a January 2018 survey of 200 US senior decision marketers conducted by Verndale, personalization is most important for increasing sales and improving customer satisfaction and retention. Yet, 84% of survey respondents agreed that the potential of personalization has not been fully realized, especially using data-driven technologies.

In a study made in April 2018 by Data & Marketing Association (DMA) and Winterberry Group, a majority (53.2%) of marketers in North America believed enhancing customer experience would lead to loyalty and retention, second only to increasing engagement.

Based on the reasons exposed above, as one major conclusion for this project we believe that creating retention techniques focused on customer experience is crucial for Criteo's client retention. However before doing so, it is important to understand the reasons behind the churn, which unfortunately wasn't information given for this particular project and hence couldn't be analyzed in its scope.

6) Contributions and Uniqueness

As for modelling, there are mainly two unique values. Firstly, the models can break down the root cause of the problem by country, industry, and cluster. Secondly, they can further provide the actionable data by automatically identifying the best performing client and the average client (a client closest to the cluster centroid) within the cluster in terms of tenure, and calculating all the differences between the two, such as spend by environment, device, criteo product and etc.
This idea is based on a hypothesis that the more similar the clients are, the more related their strategies can be. Since the result of clustering is 98% silhouette score, the score which represents the similarity within each cluster, we can conclude that the clients can benchmark the best performing client in terms of tenure and can follow the same strategy. We believe that it is a valuable solution for Criteo, because it can radically reduce the time of analyzing and identifying the problem by the automation, instead of using visualization tools or excel, and quickly adopt the strategies at granular level which are based on decent logic supported by data.

In terms of our analysis, since countries and industries behave completely differently, we decided to first group all the existing clients per country, and then per industry per country. We decided to select 3 different countries, to do this we first calculated which countries had the highest average churn rate, and which countries generated the highest amount of revenue, all three of US, Japan, and France were part of the top 4 countries for both criteria. Following this, we also discovered that the apparel industry was the industry with the highest churn rate as well as the highest amount of revenue generated for Criteo. The result was 4 different clusters for France and Japan and 5 different clusters for the US.

Our next step was to select one cluster for each country and compare the highest performing client in terms of tenure compared to the average of the cluster, in order to compare their differences and see what can be done to retain other similar clients as long as possible with Criteo.

After performing the aforementioned analysis, we generated four sets of rules. The more the client is paying criteo, the longer they stay. The lower the Cost per Order, the longer the client stays. The higher the revenues generated by Criteo, the longer the client stays. And finally, the higher the ratio between the revenue generated by Criteo and the client's overall revenue, the longer the

client stays. When looking at the lowest performers in terms of duration for these clusters, the results also followed the same pattern. The main issue for these three countries appears to be that certain clients are spending too much for too little revenue generated by Criteo.

A recommendation for these customers with a very low tenure and a close by predicted churn date would be to focus on the following products: awareness and consideration. Indeed these products are very underused, are less costly to the clients and would be the first stepping stone in becoming a well known online apparel brand. We suggest that Criteo places more emphasis on these other two products to the customers with the lowest tenure before getting them to overspend on the conversion products, which will in all likelihood be unsuccessful and lead to their churn based on the data we have analysed.

The main issue with the analysis we came up with, is we compared the same industry in three developed countries, meaning the top and average performers in between these selected countries are very much the same and the strategy employed should consequently be similar. However, we believe our same methodology can be used for different industries and for different countries in order to acquire more insights based on those selected criteria.

7) Appendix: Visualization / analytics summary

7.1 Data exploration and visualization

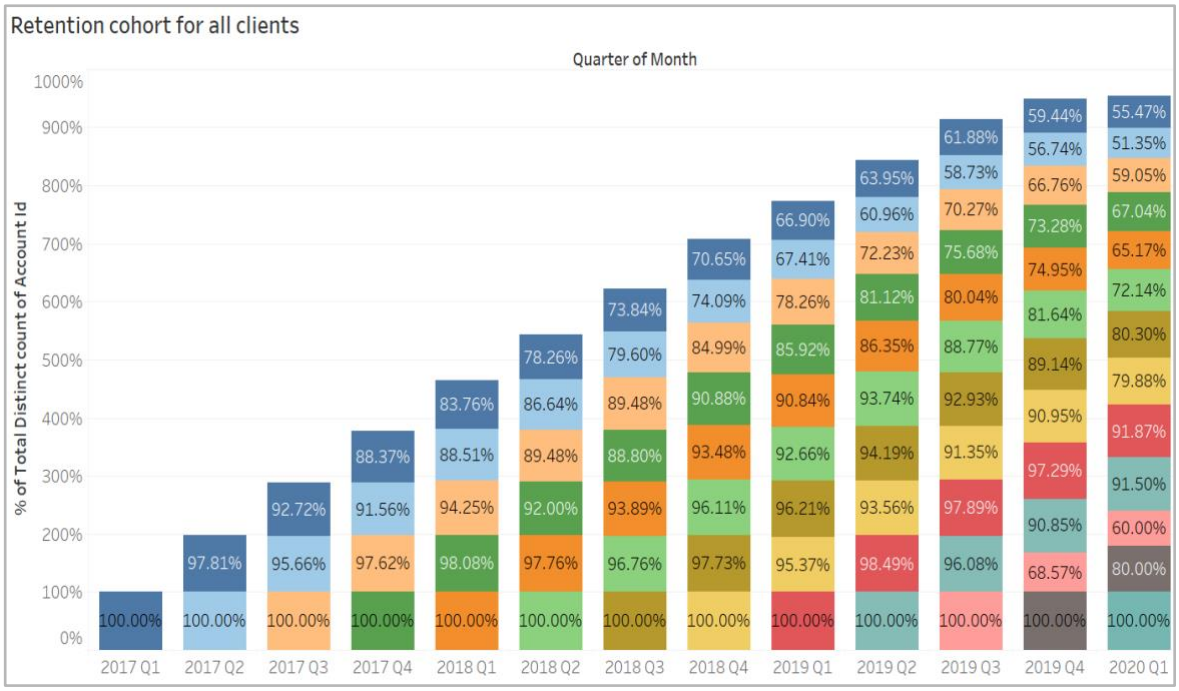

Fig.1. Retention cohort for all clients

**It can be filtered by country, industry, and product on Tableau*

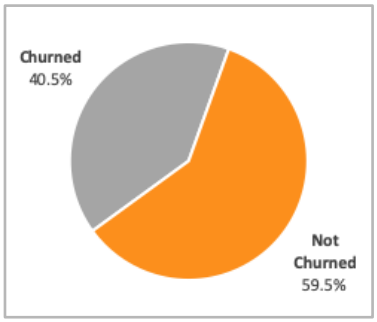

Fig. 2. Overall client churn rate (%)

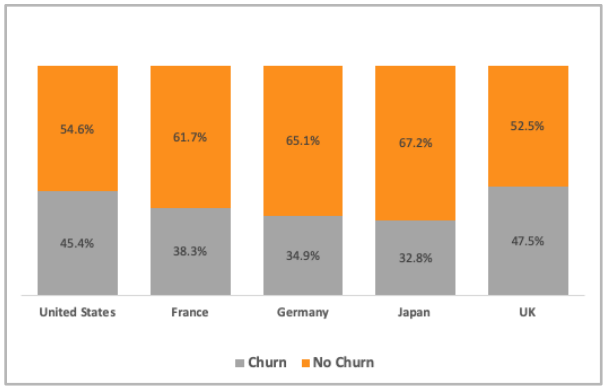

Fig. 3. Clients churned by country (%)

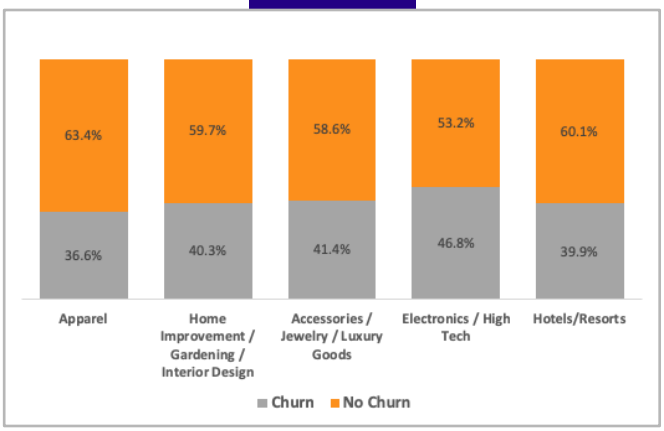

Fig. 4. Clients churned by industry (%)

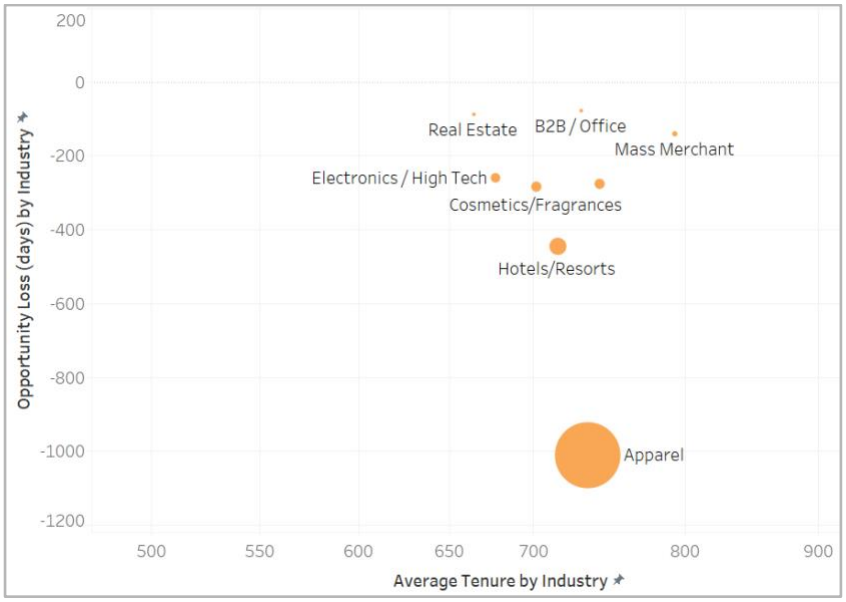

Fig. 5. Opportunity loss (days) vs Average tenure (days) by Industry

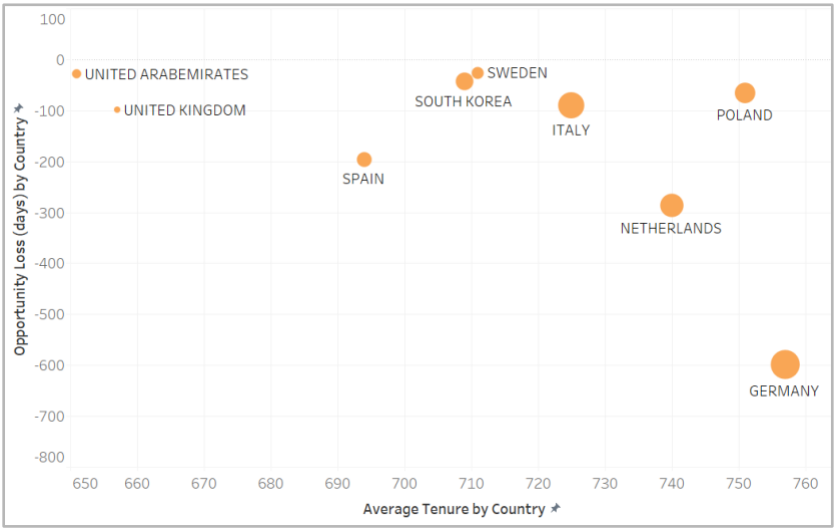

Fig. 6. Opportunity loss (days) vs Average tenure (days) by Country

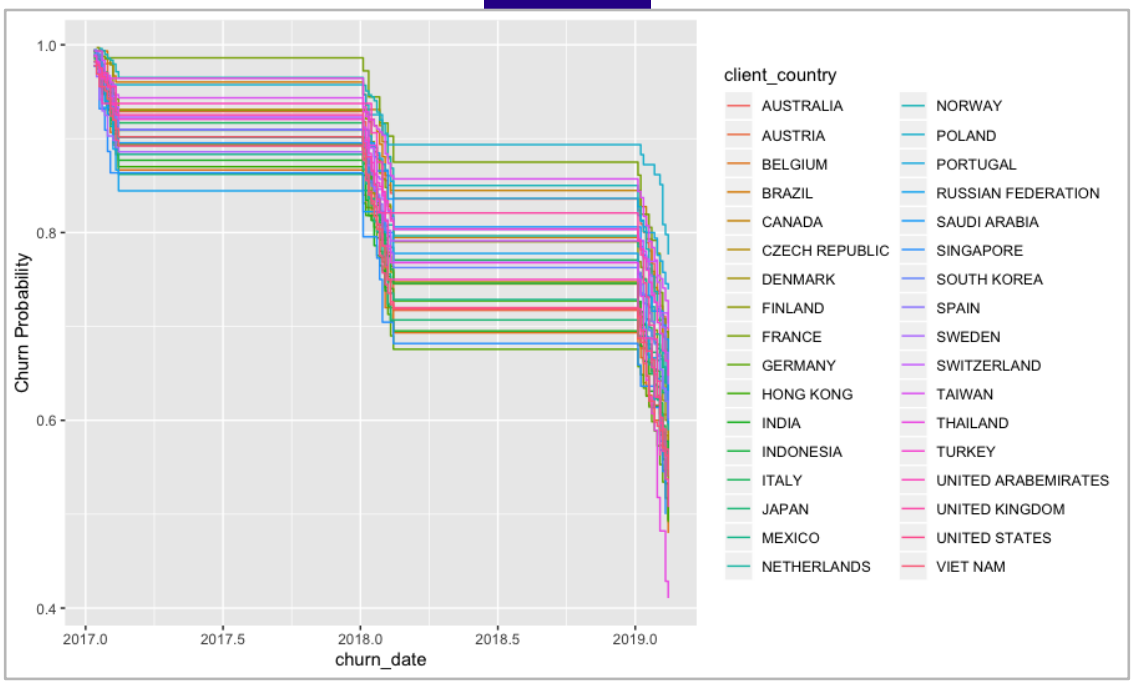

Fig. 7. Churn rate by country using Kaplan-Meier survival model

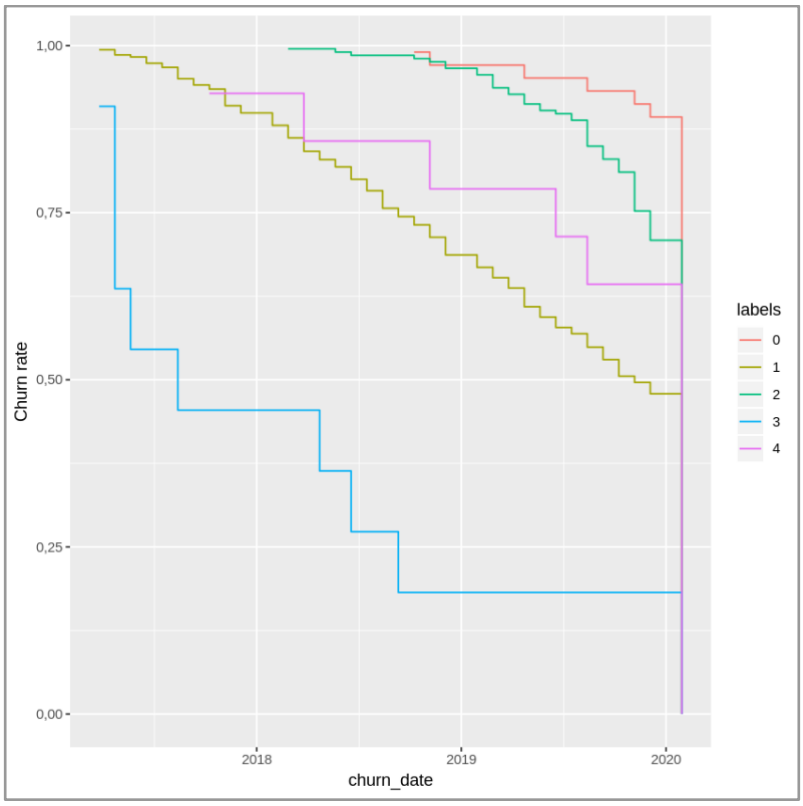

Fig. 8. Churn rate by cluster using Kaplan-Meier survival model (US)

7.2.1 Modelling Results - Churn Date Prediction

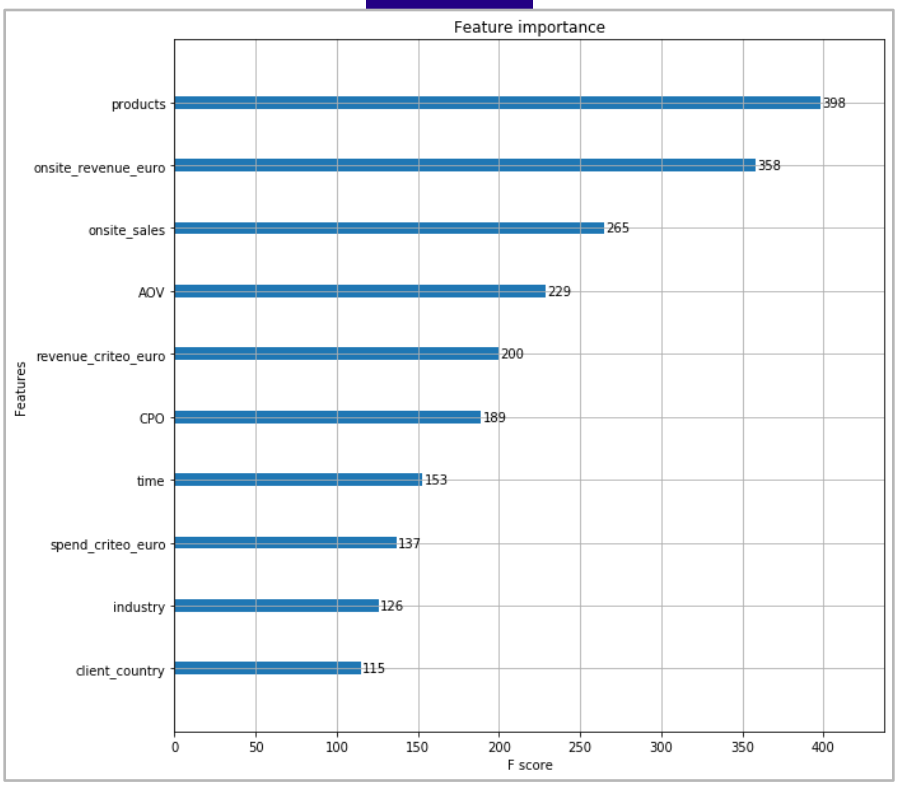

Fig. 9. Important features for prediction of customer churn date

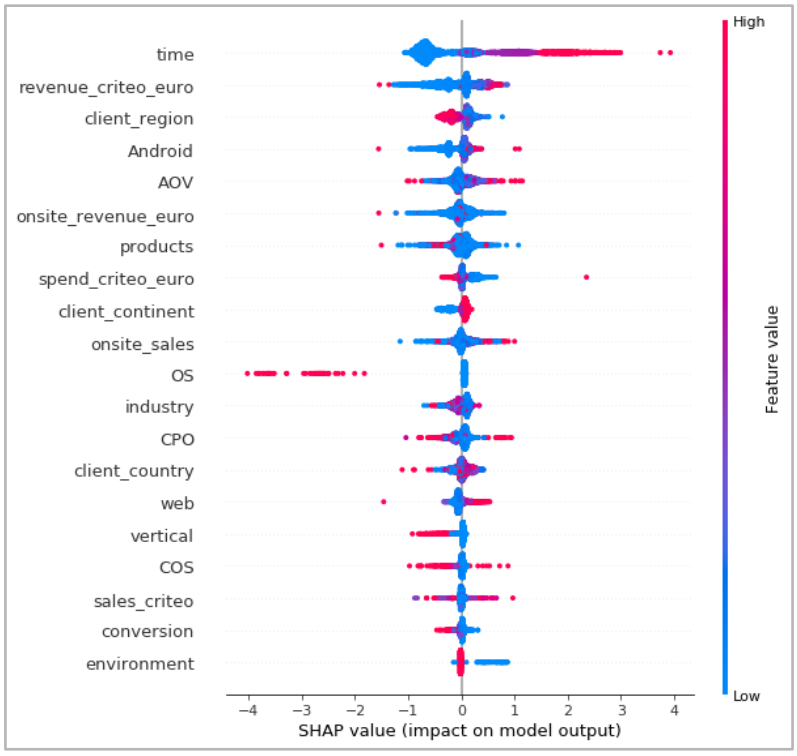

Fig. 10. SHAP value for prediction of customer churn date

7.2.2 Modelling Results - Clustering

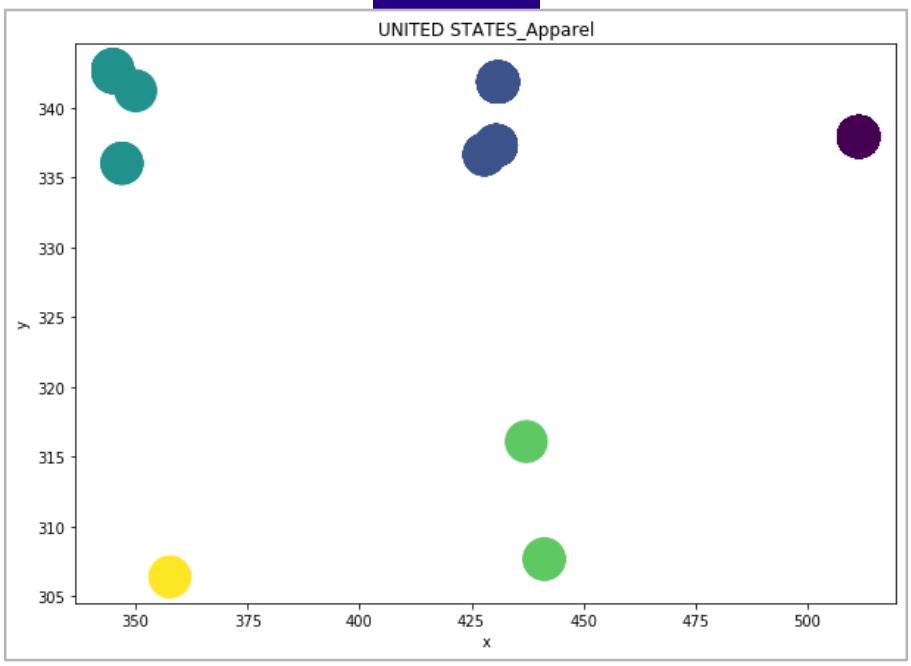

Fig. 11. Clustering scatter plot of US Apparel industry (Silhouette score 98-99%)

```
change value spend_criteo_euro to  107523
change value sales_criteo to  4485
change value CPO to  24.0
change value revenue_criteo_euro to  420841
change value AOV to  94.0
change value Android to  107523.0
change value Mobile to  107523.0
change value web to  107523.0
change value conversion to  107523.0
change value products to  3228.0
change value onsite sales to  10019.0
```

Fig. 12. Indicate difference between best and average client within a cluster

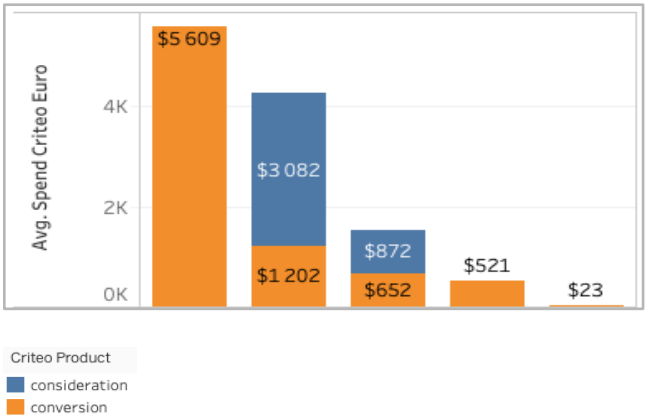

Fig. 13. Client clusters by spend criteo product

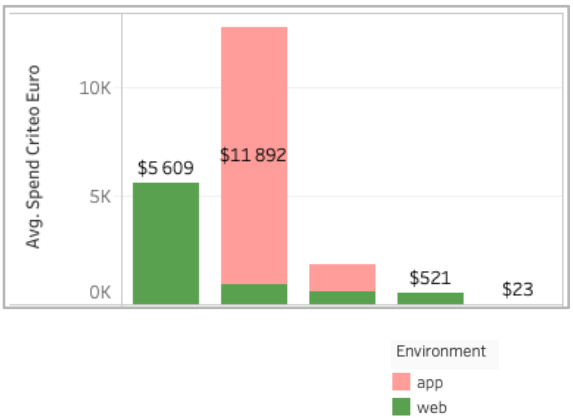

Fig. 14. Client clusters by spend environment

8) References

- *Bennes, Ross (2018). Why Marketers struggle with Data-Driven Personalization. Available in eMarketer.*
- *Criteo. (2019). 2019 Annual report on Form 10K. Retrieved from https://criteo.investorroom.com/annual-reports.*
- *eMarketer (2018). Which customers are your most valuable? Available in eMarketer.*
- *Garcia, Krista (2020). Challenges and Solutions for Customer Retention. Available in eMarketer.*
- *Ryan, Jillian (2018). Customer Growth Marketing: How B2B Deepen Relationships Through Retention, Loyalty and Advocacy Strategies. Available in eMarketer.*