Assignment 5 - Proposal

This document is a proposal for a Content Track project – an online course on cohort analysis for data analysts and data scientists.

Problem statement

Based on the literature and material available online, I have come to conclusion that there is a gap in the online courses that are focused on practical aspects of cohort analysis topic. The cohort analysis is an element of Google Analytics platform [14], and there are multiple online materials that cover this specific use case (see [15], [16] and [17] as references). After investigating blog posts and articles covering cohort analysis in my previous papers, I have come to a conclusion that there is no online course that covers cohort analysis which would give students the ability to 1) learn how to build the actual cohorts with Python, 2) use real-world datasets, and 3) learn how to extract actionable insights using cohort analysis.

I outline my solution to these three potential benefits of the course in the next section.

Course benefits

My proposal focuses on the three benefits outlined in the previous section. First, the course will be practical with hands-on programming exercises and reusable Python code snippets represented in a Jupyter Notebook format. The popularity of Python in the data science field has grown substantially, and a recent study done by Kdnuggets shows that it has taken over R programming language as the number one programming language for analytics, data science, and machine learning work [7]. Given that there are only several limited blog posts available on cohort analysis with Python, the choice of this programming language to teach this concept does look promising. While there are good resources on how to calculate cohorts with SQL like the one from Periscope data blog [8], Python is a more versatile language that enables the data scientist to use the insights and data from cohort analysis in a more general programming, and machine learning settings.

The second benefit is the use of a real-world dataset. With the rise of open data science competition platforms like Kaggle, real-world data has become more easily available. To run the cohort analysis we need to have a transactional dataset on a customer level. So an e-commerce or web traffic dataset would be the perfect option – there are multiple available datasets like “The Instacart Online Grocery Shopping Dataset 2017” [18], or YOOCHOOSE e-commerce dataset with clicks and purchases data published by YOOCHOOSE GmbH [19]. There are plenty others, but I plan to use the Online Retail Data Set from University of California, Irving Machine Learning Repository [2]. The reason being that it has been published for some time already, and there have been multiple papers, even dissertations written and analyses published based on this data, e.g. technical article on online retail industry data mining study [20]. I have also used this dataset for creating an online course on customer analytics with R that has been running for almost two years with statistics.com institute [1]. By preparing it, I have already developed a good knowledge and intuition of its structure and peculiarities which will help me bring the right level of depth to the future students of this new course.

Last benefit - extracting actionable insights - comes from my own experience working in analytics, data science, machine learning, and business intelligence fields for over a decade. I have used cohort analysis in multiple companies, and it has served and is still serving a purpose in my position as the head of data science in Amazon Devices – it continues to be an easy accessible tool to disseminate trends, understand them by assessing the changes in their behavior in detail. I can name dozens of cases where analyzing business trends with cohort analysis have helped identifying early concerning indications and manage them at the right time rather than missing them, and failing to do anything before it was too late. While the cohort analysis is just one tool in the data scientist’s toolbox, it’s a very important one. I have introduced it to multiple teams where I have worked in, and it has changed the way the those organization were looking at data, and making decisions. Just like I did by building a practical real-world customer analytics course with R on statistics.com platform [8], I plan to use the same approach with this course.

Description of content

This course will teach students how to practically implement cohort analysis on an online retail dataset from the University of California, Irvine Machine Learning Repository [2]. The course will taught in Python language which has recently surpassed R language as the most popular programming language for data science applications [7]. I plan to implement the different cohort categories described in the description developed by the Corporate Finance Institute – time based cohorts, segment based cohorts, and size based cohorts [3]. The project will first focus on exploring the dataset [2], understanding its structure by gaining insights into the distribution of variables, and doing some data preparation and cleaning. Afterwards, I will teach the basics of cohort analysis by building the most simple time based cohorts with Python and calculating one metric – either retention or average spend. Then I will increase the complexity of the course materials by introducing segment based cohorts and size based cohorts. Following that – after the different cohort categories have been introduced and developed – I will show how to implement different metrics and intersect them with the different cohorts. Finally, I will discuss the findings and insights that have been uncovered, and how they can be and are being used in the real-world. I will also present additional resources for study, give a list of potential more advanced use cases of cohort analysis, and other complementary analytical techniques.

Cohort analysis advantage

One of the main advantages of employing cohort analysis next to the other customer analytics metrics is that it allows to quickly understand the changes in the trend of some vanity metric [12], or spot a potential issue early on e.g. declining retention rate of old customers due to aggressive acquisition strategy and less investments to retain loyal customers [13].

What is more, the cohort analysis is not just a one dimensional way to triangulate the data – companies like one of the most successful Israel's start-up Optimove have developed methodologies for advanced cohort analysis [13] where the cohorts are created on multiple data points such as acquisition month (time-based cohort), recent purchase patterns (behavioral cohort), discounting (price cohort), amount spent with the company compared to other customers (value cohort). In this way the business managers can understand performance of different cohorts over time and discover patterns that are unaccessible by using vanity metrics or other approaches.

Description of tools

The development and deployment on this course will done on a Teachable platform [4]. Teachable provides an easy and intuitive learning management system (LMS) which has multiple features such as effortless setup, exceptional learning experience, simple yet powerful website customization, fully optimized for web and mobile, user-friendly website builder, and advanced developer customization [5]. Teachable has a tiered pricing structure, with the cheapest plan priced at 39 U.S. Dollars per month. While the Teachable platform does offer a free plan, my plan is to opt in for the paid plan and use my current blog cyborgus.com [6], transfer my blog posts and launch an analytics learning website which first course will be the one on cohort analysis.

Alternative sources

Based on the research of the available online courses and materials I can conclude that there are no online courses on cohort analysis. The only available materials are online resources in a tutorial format such as a blog post from Periscope on how to build cohort analysis with Structured Query Language (SQL) [8], overview of cohort analysis principles and best practices from Custora [9], or even an extensive list from Project BI startup of online resources on cohort analysis, mostly consisting of blogs and Quora questions [10].

Content outline and structure

In this part I will describe in detail the outline and the structure of the content. It will cover end of week deliverables for weeks 6 through 11. While this week's assignment requires outlining milestones for the start of week 5 through end of week 11, but since week 5 is where this proposal is submitted, I have excluded it from the content structure. The structure of this course was influenced by the study by Lynna J. Ausburn called "Course design elements most valued by adult learners in blended online education environments: an American perspective" [11]. Following the findings from the study, I will focus on making the course parts independent, although sequentially designed so the students can pick up the course chapters according to their needs, while also follow along the whole course as well. On top of that, it will be self paced and not require to submit homework at a certain time, although some recommendations around timelines will be provided.

Week 6

Status Check 1

In this first status check I will do an introduction to cohort analysis, provide useful links, and explore the dataset to understand it better, make data preparation and cleaning. I have used the Online Retail dataset [2] in one of my online published courses with Statistics.com called "Customer Analytics with R" [1], and have explored it in depth – it has a lot of systemic outliers which have to be removed, and the dataset itself cleaned before it can be used for analysis.

Week 7
Status Check 2

The second status check will provide an in-depth example of the cohort analysis on the cleaned Online Retail dataset [2]. In this part I will focus on outlining an in-depth notebook and materials with documented Python code on how to build time-based cohorts, and calculate retention rate for them. Additionally, I will show different ways of presenting the data – in the table format, plotting heatmap, or using line charts, all of which serve different purposes.

Intermediate Milestone 1

For this milestone I will aim to have my Teachable account set up [4], and linked with my current domain Cyborgus.com where my blog is hosted [6]. In it, I plan to have already the course materials posted from first and second status checks. On top of that, I will record one or several instructional videos with presentations, and create appealing informational charts describing the lesson plans in depth, while also serving as marketing material for promotion or maybe even crowd-sourcing future courses in the future.

Week 8
Status Check 3

In this status check, I will continue developing the course materials and code. This week I will focus on building the different cohort categories other that time based cohorts – segment based cohorts, and size based cohorts. I will introduce the implementation in Python for these use cases, and provide background and in-depth use cases for the real world how such cohorts are used.

Week 9
Status Check 4

The fourth status check will utilize the different cohorts developed in the previous week, and show how to use them to calculate other metrics than retention – average monthly spend, average number of products purchased, average number of days of purchase per month, and others. In this week's status check I will focus on the practical aspects on the cohort analysis and how to use it to generate insights. Also, I will showcase multiple practical capabilities of combining the different cohort analysis categories and metrics to get to actionable recommendations. I will also present several examples from the real world to give students a good overview of the applicability of the material.

Intermediate Milestone 2

For the second milestone I will introduce the course content, with some of the informational graphs developed, some of the videos recorded, and most of the code written and documented. As a stretch, I would expect all videos recorded and informational charts developed and leave some time for the next week to focus on the trailers and presentations to attract students.

Week 10
Status Check 5

The focus in this week will be on finalizing any material left from the previous week – developing informational charts, and recording videos. While this could take a day or two, the bulk of the work will be building trailers and presentations, and reviewing the materials in full, to make sure the course flow is easy and intuitive. Also, I will build quizzes, and finalize the content, so there are no significant gaps or jumps within or in between the course parts.

Week 11
Final Project

For the final project I will upload all of the course material either in a Jupyter Notebook format, or provide a link to a fully operationalized free online course on my Teachable online data science school. This depends if the setup and deployment will be successful.

Final Paper

For the final paper my aim is to summarize my learnings from the development of the course and research into a well structured 6-page document in SIGCHI Proceedings format. There are several goals for this paper. First, is to outline the benefits of the content, and learnings from developing an online course on it. Second, I will outline the alternative complementary courses that could be developed as a follow-up to this initiative which can be partially inferred from the answers to the Qualifier Question. Lastly, it will have a business plan like format, where I will outline the next steps for deployment, as well as investment required, and potential financing sources, e.g. Kickstarter, venture capital, and alternative sources.

Final Presentation

For the final presentation I plan to record a video presenting the course materials, and have a well written promotional course material, with informational graphics, benefits and other marketing elements of the course. Preferably, the full presentation material will be deployed to my Teachable online data science school, but if that turns out not to be easily achievable, I will present a separate video where I promote the course, and give an in-depth walkthrough of it.

References

1. Statistics.com Institute for Statistics Education. Customer Analytics in R. Retrieved 18:40. June 2, 2018, from https://www.statistics.com/customer-analytics-in-r/
2. UCI Machine Learning Repository. (2015, November 6). Online Retail Data Set. Retrieved 13:14, June 1, 2018, from https://archive.ics.uci.edu/ml/datasets/Online+Retail
3. Corporate Finance Institute. What is Cohort Analysis?. Retrieved 21:31, May 25, 2018, from https://corporatefinanceinstitute.com/resources/knowledge/other/cohort-analysis/
4. Teachable. Teachable platform features. Retrieved 19:45, June 15, 2018, from https://teachable.com/features
5. Mirasee Courses and Education. Teachable Review: A Simple, Easy-to-Use Learning Management System. Retrieved 20:15, June 16, 2018, from https://mirasee.com/blog/teachable-review/
6. Cyborgus data science blog. Retrieved 16:30, June 16, 2018, from https://cyborgus.com/
7. Kdnuggets News. (2017, August). Python overtakes R, becomes the leader in Data Science, Machine Learning platforms. Retrieved 12:20, June 2, 2018, from https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html
8. Periscope Data blog. (2014, June 4). How To Calculate Cohort Retention in SQL. Retrieved 21:19, May 25, 2018, from https://www.periscopedata.com/blog/how-to-calculate-cohort-retention-in-sql
9. Custora University. Cohort Analysis. Retrieved 20:33, May 25, 2018, from https://university.custora.com/for-data-scientists/cohort-analysis/basic
10. ProjectBI. The Complete List of Resources on Cohort Analysis. Retrieved 11:21, June 14, 2018, from http://www.projectbi.net/cohort-analysis-complete-list-resources/
11. Lynna J. Ausburn (2004) Course design elements most valued by adult learners in blended online education environments: an American perspective, Educational Media International, 41:4, 327-337, DOI: 10.1080/0952398042000314820
12. Kissmetrics Academy. Metrics, Metrics On The Wall, Who’s The Vainest Of Them All? Retrieved 12:29, June 9, 2018, from https://blog.kissmetrics.com/vainest-metrics/
13. Optimove Learning Center. Cohort Analysis. Retrieved 13:30, June 9, 2018, from https://www.optimove.com/learning-center/cohort-analysis
14. Google Analytics Help. The Cohort Analysis report. Retrieved 15:16, June 2, 2018, from https://support.google.com/analytics/answer/6074676?hl=en
15. Yoast tech blog. (2017, September 8). Google Analytics – What is Cohort Analysis? Retrieved 19:11, June 2, 2018, from https://yoast.com/cohort-analysis-google-analytics/
16. EduPristine blog. (2015, April 7). Cohort Analysis in Google Analytics Tutorial. Retrieved 15:22, June 2, 2018, from https://www.edupristine.com/blog/cohort-analysis-in-google-analytics
17. SEMrush blog. (2015, June 19). How to Use Cohort Analysis in Google Analytics to Drive More Business? Retrieved 13:07, June 2, 2018, from https://www.semrush.com/blog/cohort-analysis-google-analytics-drive-more-business/
18. The Instacart Online Grocery Shopping Dataset 2017. Retrieved 19:39, June 1, 2018, from https://www.instacart.com/datasets/grocery-shopping-2017
19. YOOCHOOSE e-commerce data set by YOOCHOOSE GmbH. Retrieved 14:54, June 1, 2018, from http://recsys.yoochoose.net/challenge.html

20. Chen, Daqing & Laing Sain, Sai & Guo, Kun. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management. Retrieved 19:43, June 1, 2018, from https://www.researchgate.net/publication/263329040_Data_mining_for_the_online_retail_industry_A_case_study_of_RFM_model-based_customer_segmentation_using_data_mining