10 Academy July 2020 Training - Weekly Challenge: Week 2

This case will be used for the week 2 of training for Batch 3.

User Analytics in the Telecommunication Industry - Overview

Situational Overview (Business Need)

You are working for a wealthy investor that specializes in purchasing assets that are undervalued. This investor's due diligence on all purchases includes a rich analysis of the data that underlies the business, to try to understand the fundamentals of the business and especially to identify opportunities to drive profitability by changing the focus of which products or services are being offered.

Your last role with this investor saw you do a rich analysis of a delivery company and you helped to identify that delivery to university students was the most profitable route to follow, and your analysis helped the investor purchase this delivery company and ramp up profits by 25% within 6 months through focussing on the most profitable aspect of the business. This was driven by university students always being hungry, awake at all hours, willing to purchase from a limited food menu and tending to live within a small geographical area.

The investor is interested in purchasing TellCo, an existing mobile service provider in the Republic of Pefkakia. TellCo's current owners have been willing to share their financial information but have never employed anyone to look at their data that is generated automatically by their systems.

Your employer wants you to provide a report to analyze opportunities for growth and make a recommendation on whether TellCo is worth buying or selling. You will do this by analyzing a telecommunication dataset that contains useful information about the customers & their activities on the network.

Data

- The data is here - extracted from a month aggregated data on xDR.

- The features described can be found here

Learning Outcomes

At the end of this challenge, Students will be able to:

- Technical Skills
    - Python
- Analysis
    - Perform various data wrangling techniques on the dataset
- Visualization
    - Plot self-explanatory visualizations that are rich in insights.
- Reporting back
    - Provide a comprehensive report on your analysis to management for decision making.

Team

Instructor: Jean-Henock

Main Tutor: Sebastian

Key Dates

- Discussion on the case - 1130 Rwanda time on Monday 27 July 2020. Use #all-week2 with #casebackgroundquestions to pre-ask questions.
- Interim Solution - 2000 Rwanda time on Tuesday 28 July 2020.
- Final Submission - 2000 Rwanda time on Saturday 1 August 2020

Grading for the week

There are 100 points available for the week.

20 points - community growth and peer support. This includes supporting other learners by answering questions (Slack), asking good questions (Slack), participating (not only attending) daily standups (GMeet) and sharing links and other learning resources with other learners.

25 points - presentation and reporting.

5 points - interim submission

5 - Requirements met, clear presentation

3 - Most requirements met, presentation acceptable

1 - Some effort made

20 points for the final submission. This is measured through:

- Clarity of graphs (5 points)
- Clarity of message (5 points)
- Professionalism/production value (free of spelling errors, use of same font, well produced) (5 points)

- Balance between being 'full of information' and 'easy to understand' (5 points)

55 points - data analysis and coding

10 points - interim submission

Validity of recommendations made (5 points)

Quality of code (including readability) (5 points)

45 points - final submission

Validity of recommendations made (25 points)

Quality of code (20 points)

Badges

Each week, one user will be awarded one of the badges below for the best performance in the category below.

In addition to being the badge holder for that badge, each badge winner will get +20 points to the overall score.

Visualization - quality of visualizations, understandability, skimmability, choice of visualization

Quality of code - reliability, maintainability, efficiency, commenting - in future this will be CICD

Innovative approach to analysis -using latest algorithms, adding in research paper content and other innovative approaches

Writing and presentation - clarity of written outputs, clarity of slides, overall production value

Most supportive in the community - helping others, adding links, tutoring those struggling

The goal of this approach is to support and reward expertise in different parts of the Data Scientist toolbox.

Late Policy

Our goal is to prepare successful learners for the work and submitting late, when given enough notice, shouldn't be necessary.

For interim submissions, those submitted 1-6 hours late will receive a maximum of 50% of the total possible grade. Those submitted >6 hours late may receive feedback, but will not receive a grade.

For final submissions, those submitted 1-24 hours late, will receive a maximum of 50% of the total possible grade. Those submitted >24 hours late may receive feedback, but will not receive a grade.

When calculating the leaderboard score:

- From week 4 onwards, your lowest week's score will not be considered.
- From week 8 onwards, your two lowest weeks' scores will not be considered.

Instructions

You're expected to go through each section and get answers to the questions.

The global objective is divided into 4 sub-objectives

- User Overview analysis
- User Engagement analysis
- User Experience analysis
- User Satisfaction analysis

Task 1 - User Overview analysis

The lifeblood of any business is its customers. Businesses are always finding ways to better understand their customers so that they can provide more efficient and tailored solutions to them. Exploratory Data Analysis is a fundamental step in the data science process. It involves all the processes used to familiarize oneself with the data and explore initial insights that will inform further steps in the data science process.

It is always better to explore each data set using multiple exploratory techniques and compare the results. The goal of this step is to understand the dataset, identify the missing values & outliers if any using visual and quantitative methods to get a sense of the story it tells. It suggests the next logical steps, questions, or areas of research for your project.

For the actual telecom dataset, you're expected to conduct a full User Overview analysis & the following sub-tasks are your guidance:

- Start by identifying the top 10 handsets used by the customers.
- Then, identify the top 3 handset manufacturers
- Next, identify the top 5 handsets per handset manufacturer
- Make a short interpretation and recommendation to marketing teams

In telecommunication, CDR or Call Detail Record is the voice channel. XDR is the data channel. So here, consider xDR as data sessions Detail Record. In xDR, user behavior can be tracked through the following applications: Social Media, Google, Email, Youtube, Netflix, Gaming, Other .

Task 1.1 - Your employer wants to have an overview of the users' behavior on those applications.

- Aggregate per user the following information in the column - (Jupyter notebook):
    - number of xDR sessions
    - Session duration
    - the total download (DL) and upload (UL) data
    - the total data volume (in Bytes) during this session for each application

Task 1.2 - Conduct an exploratory data analysis on those data & communicate useful insights. Ensure that you identify and treat all missing values and outliers in the dataset by replacing by the mean of the corresponding column.

You're expected to report about the following :

- Describe all relevant variables and associated data types (slide).
- Analyze the basic metrics in your DataMart (explain) & their importance for the global objective. - (slide)
- Conduct a Non-Graphical Univariate Analysis by computing position & dispersion parameters for each quantitative variable and provide useful interpretation. - (jupyter notebook + slide)
- Conduct a Graphical Univariate Analysis by identifying the most suitable plotting options for each variable and interpret your findings. - (jupyter notebook + slide)
- Bivariate Analysis – explore the relationship between each application & the total DL+UL data using appropriate methods and interpret your findings. - (jupyter notebook + slide)
- Variable transformations – segment the users into top five decile classes based on the total duration for all sessions and compute the total data (DL+UL) per decile class. - (jupyter notebook + slide)
- Correlation Analysis – compute a correlation matrix for the following variables and interpret your findings: Social Media data, Google data, Email data, Youtube data, Netflix data, Gaming data, Other data - (jupyter notebook + slide)
- Dimensionality Reduction – perform a principal component analysis to reduce the dimensions of your data and provide a useful interpretation of the results (Provide your interpretation in four (4) bullet points-maximum). - (jupyter notebook + slide)

Task 2 - User Engagement analysis

As telecom brands are the data providers of all online activities, meeting user requirements, and creating an engaging user experience is a prerequisite for them. Building & improving the QoS (Quality of Service) to leverage the mobile platforms and to get more users for the business is good but the success of the business would be determined by the user engagement and activity of the customers on available apps.

In telecommunication, tracking the user activities on the database sessions is a good starting point to appreciate the user engagement for the overall applications and per application as well. If we can determine the level of engagement of a random user for any application, then it could help the technical teams of the business to know where to concentrate network resources for different clusters of customers based on the engagement scores.

In the current dataset you're expected to track the user's engagement using the following engagement metrics:

- sessions frequency
- the duration of the session

- the sessions total traffic (download and upload (bytes))

Task 2.1 - Based on the above:

- Aggregate the above metrics per customer id (MSISDN) and report the top 10 customers per engagement metric - (jupyter notebook + slide for top 10)
- Normalize each engagement metric and run a k-means (k=3) to classify customers in three groups of engagement. - (jupyter notebook)
- Compute the minimum, maximum, average & total non- normalized metrics for each cluster. Interpret your results visually with accompanying text. - (jupyter notebook + slide)
- Aggregate user total traffic per application and derive the top 10 most engaged users per application - (jupyter notebook + slide)
- Plot the top 3 most used applications. - (jupyter notebook + slide)
- Using *k*-means clustering algorithm, group users in *k* engagement clusters based on the engagement metrics:
    - What is the optimized value of *k*? - (slide)
    - Interpret your findings. - (slide)

Task 3 - Experience Analytics

The Telecommunication industry has experienced a great revolution since the last decade. Mobile devices have become the new fashion trend and play a vital role in everyone's life. The success of the mobile industry is by and large dependent on its consumers. Therefore, it is necessary for the vendors to focus on their target audience i.e. what are the needs and requirements of their consumers and how they feel and perceive their products. Tracking & evaluation of customers' experience can help the organizations to optimize their products and services so that it meets the evolving user expectations, needs, and acceptance.

In the telecommunication industry, the user experience is related, most of the time, to network parameter performances or the customers' device characteristics.

In this section, you're expected to focus on network parameters like TCP retransmission, Round Trip Time (RTT), Throughput, and the customers' device characteristics like the handset type to conduct a deep user experience analysis. The network parameters are all columns in the dataset. The following questions are your guidance to complete the task.

Task 3. 1 - Aggregate, per customer, the following information (treat missing & outliers by replacing by the mean or the mode of the corresponding variable) -(jupyter notebook):

- Average TCP retransmission
- Average RTT
- Handset type
- Average throughput

Task 3.2 - Compute & list 10 of the top, bottom and most frequent - (jupyter notebook + slide):

a. TCP values in the dataset.
b. RTT values in the dataset.
c. Throughput values in the dataset.

Task 3.3 - Compute & report - (jupyter notebook + slide):

d. The distribution of the average throughput per handset type and provide interpretation for your findings.
e. The average TCP retransmission view per handset type and provide interpretation for your findings.

Task 3.4 - Using the experience metrics above, perform a *k*-means clustering (where *k* = 3) to segment users into groups of experiences and provide a brief description of each cluster. - (jupyter notebook + slide)

Task 4 - Satisfaction Analysis

Assuming that the satisfaction of a user is dependent on user engagement and experience, you're expected in this section to analyze customer satisfaction in depth. The following tasks will guide you:

Based on the engagement analysis + the experience analysis you conducted above,

Task 4. 1 - Write a python program to assign:

a. engagement score to each user. Consider the engagement score as the Euclidean distance between the user data point & the less engaged cluster (use the first clustering for this) - (jupyter notebook)
b. experience score to each user. Consider the experience score as the Euclidean distance between the user data point & the worst experience's cluster. - (jupyter notebook)

Task 4.2 - Consider the average of both satisfaction & experience scores as the satisfaction score & report the top 10 satisfied customer - (jupyter notebook + slide)

Task 4.3 - Run a regression model of your choice to predict the satisfaction score of a customer. - (jupyter notebook)

Task 4.4 - Run a k-means (k=2) on the engagement & the experience score - (jupyter notebook).

Task 4.5 - Aggregate the average satisfaction & experience score per cluster. - (jupyter notebook + slide)

Task 4.6 - Export your final table containing all user id + engagement, experience & satisfaction scores in your local MySQL database. Report a screenshot of a select on the exported table. (jupyter notebook + slide)

Interim Submission **(Due Tuesday 28.07 20hr Rwanda time)**

- Your employer wants a quick meeting after you've done a first quick pass of the data and wants to know whether further investigation is useful. To achieve this, summarize your findings from Task 1 in seven slides - no need for a title slide - this is just an interim submission. The variables we would like to analyze in the task 1 are:
    - Number of xDR sessions, Session duration, the total download (DL) and upload (UL) data , the total data volume (in Bytes) during this session for each application (Social Media, Google, Email, YouTube, Netflix, Gaming).
        - Slides 1-3: Non graphical Univariate analysis - For each of the above variables describing the customers, report in a table the minimum value, the maximum value, the average, the 1st, 2nd & 3rd quartile and provide useful interpretations.
        - Slides 4-6: : Graphical Univariate Analysis - For each of the above variables, report plots which show the distribution of the corresponding variable in the whole dataset and provide a one sentence comment per plot.
        - Slides 7 : For each of the data consumption applications (Social Media, Google, Email, YouTube, Netflix, Gaming), report a bivariate plot where the application is represented on x axis & the total data (UL+DL) is represented on y axis- comments your results.
    - Link to your GitHub code

Feedback

You may not receive detailed comments on your interim submission, but will receive a grade.

Final Submission **(Due Sat 01.08 20hr Rwanda time)**

- Summarize your findings from all of the 4 Tasks (Customers Overview, User Engagement, Experience and Satisfaction Analysis). Your employer demands no more than 20 slides, including a title page and references.
    - Ensure that you make a recommendation to your employer on the growth potential of the company (positive or negative) based on the data.
    - Ensure that you share the data and slides with justifying your recommendation with data and graphs
    - Ensure that you outline the limitations of your analysis.
    - Ensure that you make a recommendation on whether your employer should purchase this company.
- Link to your Github code that includes your Jupyter notebook..

Feedback

You will receive comments/feedback in addition to a grade.

References

We should have resources for learners to self-teach - to point them to the right places for key pieces of information. I think this would include the following:

- Exploratory Data Analysis In Python
- Non Graphical Univariate Analysis 1
- Non Graphical Univariate Analysis 2
- Univariate and Bivariate Analysis
- How to define an outlier
- How to Correlation Analysis
- How to do PCA (Video)
- Define telecoms QoS
- An Oracle Data Science Case Study in Telecom
- Use cases and challenges in telecom big data analytics paper (PDF) Use cases and challenges in telecom big data analytics