10 Academy Batch 4 - Weekly Challenge: Week 1

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 detailed 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. You will deliver insights you managed to extract to your employer through an easy to use web based dashboard and a written report.

Data

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

- The features described can be found here

Competency Mapping

The tasks you will carry out in this week's challenge will contribute differently to the 17 competencies 10 Academy identified as essential for job preparedness in the field of data science, and Machine Learning engineering. The mapping below shows the change (lift) one can obtain through delivering the highest performance in these tasks.

MC0: Marginal contribution - causes no significant change
MC1: Minor contribution - recognised for routine performance gain
MC2: Measurable contribution - will contribute a value towards portfolio and job readiness metric
MC3: Major contribution - best performance of these types of tasks at least three times within our training leads one to attain job ready level along that competency dimension.

Competency	Contribution	Potential contributions from this week
Business Understanding	MC2	Understanding and reasoning the business context. Thinking about suitable analysis that matches the business need. Thinking about clients and their interests.
Data Engineering	MC1	Thinking about how to store data for easy analysis, and what format to use to build responsive dashboards.
Data Understanding	MC3	This is the main focus of the project - to understand the data provided and extract insight. You will have to explore different techniques, algorithms, statistical distributions, sampling, and visualisation techniques to gain insight.
Dashboard & Visualization	MC3	Building a dashboard to explore data as well as to communicate insight. Advanced use of modules such as plotly, seaborn, matplotlib etc. to build descriptive visualizations. Reading through the modules documentation to expand your skill set.
Mathematics and Statistics	MC3	Thinking about statistical distributions, sampling, bias, overfitting, correlations.

MLOps & Continuous Delivery	MC1	Using Github for code development, thinking about feature store, planning analysis pipeline, making docker containers to deploy dashboard.
Modelling and evaluation	MC2	Feature importance calculation requires using different ML models. Choosing evaluation metrics.
Python programming	MC3	A lot of modular and object oriented python code writing. Python package building.
SQL programming	MC1	Building feature stores using MySQL or NOSQL databases.
Fluency in the Scientific Method	MC2	Thinking about evidence. Generating hypothesis, testing hypothesis. Thinking about different types of errors.
Ethics	MC0	No significant input to data privacy or usage of data.
Statistical & Critical Thinking	MC2	Thinking about the difference between causal vs chance correlation. Giving reasonable recommendations. Thinking about uncertainties.
Software Engineering & Dev Environment	MC1	Reading articles on software project planning. Unit testing.
Impact & Lifelong learning	MC2	Learning new concepts, ideas, and skills fast, and applying them to the problem at hand.
Professional Culture & Communication	MC2	Writing a well formatted presentation with no mistakes, formatted nicely.
Social Intelligence & Mentorship	MC2	Asking for help early, providing help for those who need it, avoiding being stuck.
Career Thinking	MC1	

Team

Instructors: Yabebal, Abubakar, Mahlet, Kevin,Usman, Paulcy

Key Dates

- Discussion on the case - 09:30 UTC on Monday 12 July 2021. Use #all-week1 to pre-ask questions.
- Interim Solution - 2000 UTC on Wednesday 14 July 2021.
- Final Submission - 2000 UTC on Saturday 17 July 2021

Leaderboard 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 (RC), asking good questions (RC), participating (not only attending) daily standups (GMeet) and sharing links and other learning resources with other learners.

20 points - presentation and reporting.

5 points - interim submission

5 - Requirements met, clear presentation

3 - Most requirements met, presentation acceptable

1 - Some effort made

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

- Clarity of graph interpretation (3 points)
- Clarity of message (4 points)
- Professionalism/production value (free of spelling errors, use of same font, well produced, well formatted graphs) (5 points)
- Balance between being ‘full of information’ and ‘easy to understand’ (3 points)

20 points - dashboard code, screenshot, and cloud deployment

10 points - screenshot & dashboard code submission

10 points - Dockerfile to build the dashboard as docker container or code deployed

40 points - data analysis and coding

10 points - interim submission

Preprocessing & EDA (4 points)

Generating insightful and quality plots (2 points)

Frequent github commits, multiple branching, and pull request (2 points)

Modularity and quality of code (including readability) (2 points)

30 points - final submission

Preprocessing & EDA (15 points)

Generating insightful, novel, and quality plots (10 points)

Advanced github use, modularity, and quality of code (5 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 for the leaderboard 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 Machine Learning Engineer toolbox.

Group Work Policy

This submission is to be done individually. Collaborative learning is encouraged, but each person must have his or her own submissions.

Late Policy

Our goal is to prepare successful trainees 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

At the end of this week you are expected to to have a complete project that has

- Reusable code for data preparation and cleaning.
- Dashboard that shows your findings.
- Reusable feature store which can be used to store selected features for later usage on similar problems
- Your project folder should mirror as close as possible the example here. In particular
    - Code is installable via pip
    - Has unit tests with good test coverage
    - Has CI/CD setup - using Travis or Github Actions
    - Has Dockerfile to build it as docker image

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 top 3 handset manufacturer
- Make a short interpretation and recommendation to marketing teams

In telecommunication, CDR or Call Detail Record is the voice channel and XDR is the data channel equivalent. 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
    - 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 using python script and slide :

- Describe all relevant variables and associated data types (slide).
- Analyze the basic metrics (mean, median, etc) in the Dataset (explain) & their importance for the global objective.
- Conduct a Non-Graphical Univariate Analysis by computing dispersion parameters for each quantitative variable and provide useful interpretation.
- Conduct a Graphical Univariate Analysis by identifying the most suitable plotting options for each variable and interpret your findings.
- Bivariate Analysis – explore the relationship between each application & the total DL+UL data using appropriate methods and interpret your findings.
- 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.
- 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
- 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).

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 submit python script and slide :

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

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 largely 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. For this task you need a python script that includes all solutions to tasks.

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

- Average TCP retransmission
- Average RTT
- Handset type

- Average throughput

Task 3.2 - Compute & list 10 of the top, bottom and most frequent:

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

Task 3.3 - Compute & report:

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. (The description must define each group based on your understanding of the data)

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) (Euclidean Distance)
b. experience score to each user. Consider the experience score as the Euclidean distance between the user data point & the worst experience's cluster.

Task 4.2 - Consider the average of both engagement & experience scores as the satisfaction score & report the top 10 satisfied customer

Task 4.3 - Build a regression model of your choice to predict the satisfaction score of a customer.

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

Task 4.5 - Aggregate the average satisfaction & experience score per cluster.

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 query output on the exported table.

Task 4.7 Model deployment tracking- deploy the model and monitor your model. Here you can use Docker or other MlOps tools which can help you to track your model's change. Your model tracking report includes code version, start and end time, source, parameters, metrics (loss convergence) and artifacts or any output file regarding each specific run. (CSV file, screenshot)

Deliverables

Interim Submission **(Due Wednesday 14.07.2021 2000 UTC)**

1. 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 repository

Feedback

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

Final Submission **(Due Sat 17.07.2021 2000UTC)**

2. 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.
3. A Github link to your dashboard code and a screenshot of your dashboard. To build your dashboard you can use Streamlit or Flask or any other web-based Framework that you are familiar with. The important element is that your plots and insights should be easily navigable in a remote browser.
4. A Github link to your Data analysis code.

Feedback

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

References

- 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
- https://www.statology.org/deciles-in-python/
- Use cases and challenges in telecom big data analytics paper (PDF) Use cases and challenges in telecom big data analytics
- https://github.com/10-Academy-Self-Learning-Resources/Data-Understanding
- https://github.com/10-Academy-Self-Learning-Resources/DataVisualization
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html
- https://strategyanalytics.medium.com/pandas-read-excel-removed-support-for-xlsx-files-426e4acfde89
- https://www.mlflow.org/docs/latest/tutorials-and-examples/tutorial.html