10 Academy Cohort B - Weekly Challenge: Week 1

User Analytics in the Telecommunication Industry - Overview

# Situational Overview (Business Need)

You are working for a wealthy investor that specialises 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 analyse opportunities for growth and make a recommendation on whether TellCo is worth buying or selling. You will do this by analysing 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

* We have sourced the data from a month's aggregation of xDR records.
* The description for attributes can be found [here](https://docs.google.com/spreadsheets/d/1TxwrOabpuFZVVWYe2bfIr2Uu44WM-oC7H0vG8qeO8vA/edit?usp=sharing)
* This data should be extracted from a PostgreSQL database, and we've included both the database schema and the corresponding SQL file [here](https://drive.google.com/file/d/1BjIbrz5nWl-8ZIVErIMEmxDOqy6qNVOg/view).

## Learning Outcomes

* Understanding and reasoning the business context. Thinking about suitable analysis that matches the business need. Thinking about clients and their interests.
* Understanding the data provided and extract insight. You will have to explore different techniques, algorithms, statistical distributions, sampling, and visualisation techniques to gain insight.
* Understand the data structure and algorithms used in EDA and machine learning pipelines
* 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 visualisations. Reading through the modules documentation to expand your skill set.
* Thinking about statistical distributions, sampling, bias, overfitting, correlations.
* Modular and object oriented python code writing. Python package building.

# Competency Mapping

The tasks you will carry out in this week’s challenge will contribute differently to the 11 competencies 10 Academy identified as essential for job preparedness in the field of Data Engineering, and Machine Learning engineering. The mapping below shows the change (lift) one can obtain through delivering the highest performance in these tasks.

| **Competency** | **Potential contributions from this week** |
| --- | --- |
| Professionalism for a global-level job | Articulating business values |
| Collaboration and Communicating | Reporting to stakeholders |
| Software Development Frameworks | Using Github for CI/CD, writing modular codes, and packaging |
| Python programming | Advanced use of python modules such as Pandas, Matplotlib, Numpy, Scikit-learn, Prophet and other relevant python packages |
| SQL programming | Extensive use of SQL on PostgreSQL |
| Data & Analytics Engineering | data filtering, data transformation, and data warehouse management |
| MLOps & AutoML | Pipeline design, data and model versioning, |
| Deep Learning and Machine Learning | NLP, topic modelling, sentiment analysis |
| Web & Mobile app programming | HTML, CSS ,Flask, Streamlit |

## Team

Tutors:

* Yabebal
* Emtinan
* Rehmet
* Natnael

## 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.

**Visualisation** - quality of visualisations, understandability, skimmability, choice of visualisation

**Quality of code** - reliability, maintainability, efficiency, commenting - in future this will be [CICD](https://en.wikipedia.org/wiki/CI/CD)

**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.
* Estimated code complexity (running time and memory size order of magnitude estimate) for the key part of your code. To do so you would need to explicitly identify and profile the data structures and algorithms in your code.
* Code connected using scikit pipeline or other form of chaining multiple EDA steps
* Beautiful and insightful Streamlit dashboard that shows your findings.
* SQL database as feature store which can be used to store selected features for dashboard visualisation and model training.
* Your project folder should mirror as close as possible the example [here](https://github.com/streamlit/streamlit/tree/develop). In particular
  + Code is installable via pip
  + Has unit tests with good test coverage
  + Has CI/CD setup - using Github Actions
  + Has Dockerfile to build it as docker image
  + Python codes are written in the style and structure of [Streamlit python source code](https://github.com/streamlit/streamlit/tree/develop/lib) - reading and understanding source codes of well-known open source packages will help you learn advanced python programming.

The global objective is divided into 4 sub-objectives

* User Overview analysis
* User Engagement analysis
* User Experience analysis
* User Satisfaction analysis

## Task 1 - Report on Streamlit understanding

1. **Streamlit Source Code Overview**: Provide a one-page comprehensive overview of [Streamlit source code](https://github.com/streamlit/streamlit/tree/develop/lib) structure. Your report should emphasise on
   * How the python source code is structured
   * The different advanced python programming techniques used in the code base
   * How unit and integration tests are written
   * What extra packages are used - check the requirements.txt file
   * How doc strings, logging, and error handling are written
   * How @property and other builtin and custom decorators are used in the source code
   * Any other observations that interested you.
2. **Identifying the data structure and algorithm** used in the key component of the data loading part of the streamlit package.

## Task 2 - 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 familiarise 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.

Start by applying your streamlite understanding to prepare your code with modular design, best practices in coding.

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 behaviour can be tracked through the following applications: Social Media, Google, Email, Youtube, Netflix, Gaming, Other.

**Task 2.1** - Your employer wants to have an overview of the users’ behaviour 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 2.2** - 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 3 - 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 3.1** - Following the above, perform the following tasks:

* 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 4 - 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](https://www.saminiir.com/lets-code-tcp-ip-stack-5-tcp-retransmission/#tcp-retransmission), [Round Trip Time (RTT)](http://www.on-time.com/rtos-32-docs/rtip-32/programming-manual/tcp-ip-networking/tcp/round-trip-time-rtt.htm), [Throughput](https://blog.thousandeyes.com/a-very-simple-model-for-tcp-throughput/#:~:text=A%20Simple%20Model%20for%20TCP%20Throughput&text=Where%20WindowSize%20is%20the%20amount,%2Dto%2Dend%20network%20path.), 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 4. 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 4.2** - Compute & list 10 of the top, bottom and most frequent:

* 1. TCP values in the dataset.
  2. RTT values in the dataset.
  3. Throughput values in the dataset.

**Task 4.3** - Compute & report:

* 1. The distribution of the average throughput per handset type and provide interpretation for your findings.
  2. The average TCP retransmission view per handset type and provide interpretation for your findings.

**Task 4.4** - 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 5 - 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 5. 1** - Write a python program to assign:

* 1. 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](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.euclidean_distances.html))
  2. experience score to each user. Consider the experience score as the Euclidean distance between the user data point & the worst experience’s cluster.

**Task 5.2** - Consider the average of both engagement & experience scores as the satisfaction score & report the top 10 satisfied customer

**Task 5.3** - Build a regression model of your choice to predict the satisfaction score of a customer.

**Task 5.4** - Run a k-means (k=2) on the engagement & the experience score .

**Task 5.5** - Aggregate the average satisfaction & experience score per cluster.

**Task 5.6** - 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 5.7** 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 artefacts or any output file regarding each specific run. (CSV file, screenshot)

# 

# Tutorials Schedule

## Overview

* Monday: Understanding week 1 challenge and Data Pre-processing
* Tuesday: Data Exploration and Insights communication
* Wednesday: Data modelling
* Thursday: Dashboard Development

In the following, the colour **purple** indicates morning sessions, and **blue** indicates afternoon sessions.

## Monday: Data Pre-processing

Here, students will understand this week's challenge and learn how to prepare data for modelling.

* Introduction to week1 challenge(YF)
* Going through the Streamlit source code (Rehmet)

Key Performance Indicators:

* Understanding advanced code base
* Efficient and modular coding technique
* Understanding data in DB – connecting with DB
* Ability to help others

## Tuesday: EDA

Here, students will learn how to understand data and communicate insights from the data.

* Data Extraction, Cleaning, Transforming and formatting using modular python (Emtinan)
* Working with PostgreSQL DB using SQL, Pandas, and SQLAlchemy (Emtinan And Rehmet)

Key Performance Indicators:

* Proactivity to self-learn - sharing references
* Intermediate to advanced SQL techniques
* Data Understanding and Exploration

## Wednesday: Data Modelling

Here, students will learn how to model the data.

* Modelling (Emtinan)

Key Performance Indicators:

* Modelling
* MLOps
* Proactivity to self-learn - sharing references

## Thursday: Dashboard Development

Here, students will learn how to visually communicate data and insights using dashboards.

* Dashboard development (Rehmet)

Key Performance Indicators:

* Design Thinking
* Dashboard Design
* Proactivity to self-learn - sharing references

# 

# Deliverables

**MAKE SURE YOUR SUBMISSION IS IN A PDF FORMAT WHEN ASKED TO UPLOAD A WORD/PPT FILE.**

## Interim Submission **(Due Wednesday 24.04.2024 2000 UTC)**

1. A PDF containing concise and comprehensive descriptions of the key data structure and algorithm you used in your code to address **Task 2 and Task 3** in 1 or 2 pages. If you use pandas, scikit-learn etc. in your code, report on the data structure and algorithm used in the part of those packages you used. Comment on how long (time) and how much memory (RAM size) will it take your code if the data size grows by 10x, 1000x, 100000x, 1000000x. Comment on the suitability of the data structure and algorithm you used for large data sets.
2. 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 27.04.2024 2000 UTC)**

1. Report for Task 1 - Streamlit source code review (1-3 pages).
2. Summarise your findings from Task 2 to 5 (Customers Overview, User Engagement, Experience and Satisfaction Analysis). Your employer demands no more than 30 slides - 1hr presentation.

* 1. Ensure that you include a title slide in the beginning and a reference slide at the end.
  2. Ensure that you make a recommendation to your employer on the growth potential of the company (positive or negative) based on the data.
  3. Ensure that you share the data and slides with justifying your recommendation with data and graphs
  4. Ensure that you outline the limitations of your analysis.
  5. Ensure that you make a recommendation on whether your employer should purchase this company.

1. A Github link to all your codes and a screenshot of your dashboard. Make sure to include as many screen shots as possible so that we are able to see all the functionality of your dashboard - your evaluation for the dashboard wireframe and functionality will be entirely based on the screenshots (or deployed version) you submitted. Make an effort to deploy your dashboard live in [Streamlit Cloud](https://docs.streamlit.io/streamlit-cloud/get-started/deploy-an-app) Heroku, [Netlify](https://netlify.app/) or any one of the free web app hosting services you can find [here](https://bugfender.com/blog/10-best-app-deployment-platforms/).

### Feedback

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

# 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 (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

2 - Evidence of clear understanding of the business context and data

2 - Evidence of clear plan to complete the project and what is done

till date

1 - Style of report

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

4 - Style and quality of report (e.g. error free, font and format consistency)

4 - Creative articulation, clarity of content, and objective communication[[1]](#footnote-0)

10 - Clear sections on

1 - objectives of the project and the intended business value

1 - data size, type, format and other details (e.g. missing values)

1 - methods and algorithms used

1 - details on pipeline, automation, (code, data, model) versioning

2 - well produced supporting figures and graphics

2 - result and discussion

2 - summary of what has been achieved, its implication, and weather

objectives of the project are met or not and why

2 - Balance between being ‘full of information’ and ‘easy to understand’

20 points - dashboard code, screenshot, and cloud deployment

10 points - screenshot & dashboard code submission

2 - functionality of dashboard to address business need

2 - multiple screenshot or live access

2 - proof of wireframe design

2 - CI/CD deployment

10 points - dashboard code and deployment

2 - Dockerfile to build the dashboard

2 - Github actions and unit test

2 - Quality and structure of code

2 - incorporating javascript/css to enhance dashboard

2 - sql schema of feature store database

40 points - data analysis and coding

15 points - interim submission

4 - Preprocessing & EDA

2 - Generating novel plots - insightful and quality plots

2 - Frequent github commits, multiple branching, and pull request

3 - Modularity and quality of code (including readability)

4 - use of scikit pipeline or other pipeline approach

30 points - final submission

10 - Depth of preprocessing & EDA

5 - Pipeline driven analysis

5 - plots and notebook - insightful plots and well commented sections

6 - Advanced github use, modularity, and quality of code

4 - Github actions and unit test

# 

# References

General references

* [Exploratory Data Analysis In Python](https://medium.com/@ramdittakavi_61179/exploratory-data-analysis-a-structured-approach-da4d74646445)
* [Non Graphical Univariate Analysis 1](https://online.stat.psu.edu/stat504/node/11/)
* [Non Graphical Univariate Analysis 2](https://www.ocsc.go.th/sites/default/files/attachment/page/3502_exploratory-data-analysis.pdf)
* [Univariate and Bivariate Analysis](https://towardsdatascience.com/how-to-use-python-seaborn-for-exploratory-data-analysis-1a4850f48f14)
* [How to define an outlier](https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm#:~:text=Definition%20of%20outliers,what%20will%20be%20considered%20abnormal.)
* [How to Correlation Analysis](https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html)
* [How to do PCA](https://sebastianraschka.com/Articles/2014_pca_step_by_step.html) ([Video](https://www.youtube.com/watch?v=FgakZw6K1QQ))
* [Define telecoms QoS](https://www.eolss.net/Sample-Chapters/C05/E6-108-14-00.pdf)
* [An Oracle Data Science Case Study in Telecom](https://blogs.oracle.com/datascience/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://www.researchgate.net/publication/311628734_Use_cases_and_challenges_in_telecom_big_data_analytics/link/587dcdf508ae9275d4e898fc/download)
* <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>

Design

* [App Layout & Style Tips | Designing Apps for User (Part II) (streamlit.io)](https://blog.streamlit.io/designing-streamlit-apps-for-the-user-part-ii/)
* [Marvel | Create your free account (marvelapp.com)](https://marvelapp.com/signup)

1. [Kinds and Objectives of Report writing (theintactone.com)](https://theintactone.com/2019/09/01/bc-u3-topic-5-kinds-and-objectives-of-report-writing/) [↑](#footnote-ref-0)