User Analytics for Strategic Business Acquisition

Overview

Situational Overview (Business Need)

You are working for a wealthy investor who specializes in purchasing undervalued assets. This investor's due diligence on all acquisitions includes a detailed analysis of the underlying business data to understand the fundamentals and identify opportunities to drive profitability by focusing on the most lucrative products or services.

In your previous role with this investor, you conducted a comprehensive analysis of a delivery company and identified that deliveries to university students were the most profitable segment. Your analysis led to the acquisition of the delivery company and increased profits by 25% within six months by concentrating on this high-margin segment. This success was driven by university students' consistent demand, late-night activity, preference for a limited product menu, and residence within a small geographical area.

The investor is now interested in purchasing **TellCo**, an existing service provider in the Republic of Pefkakia. TellCo's current owners have shared their financial information but have never employed anyone to analyze their system-generated data.

Your employer has tasked you with providing a report to analyze growth opportunities and recommend whether TellCo is worth buying or selling. You will achieve this by analyzing a **telecommunication dataset** containing valuable information about customers and their network activities. Your findings will be delivered through an easy-to-use web-based dashboard and a comprehensive written report.

Data

- **Dataset:** Available here – extracted from a month of aggregated data on xDR (Data Sessions Detail Record).
- **Features Description:** Link to Features

Instructions

By the end of this project period, you are expected to have a complete project that includes:

- **Reusable Code for Data Preparation and Cleaning**
- **A Dashboard Displaying Your Findings**

- **A Reusable Feature Store** for storing selected features for future similar problems

Global Objectives

Divided into 4 sub-objectives:

1. **User Overview Analysis**
2. **User Engagement Analysis**
3. **User Experience Analysis**
4. **User Satisfaction Analysis**

Detailed Tasks

Task 1 - User Overview Analysis

Understanding customers is crucial for any business. Exploratory Data Analysis (EDA) is essential to familiarize yourself with the data and uncover initial insights that guide further analysis.

Sub-Tasks:

1. **Identify the Top 10 Handsets Used by Customers.**
2. **Identify the Top 3 Handset Manufacturers.**
3. **Identify the Top 5 Handsets per Top 3 Handset Manufacturer.**
4. **Provide a Short Interpretation and Recommendation to Marketing Teams.**

Note: In telecommunication, CDR (Call Detail Record) is equivalent to voice data, while XDR (Data Sessions Detail Record) pertains to data sessions. User behavior can be tracked through applications like Social Media, Google, Email, YouTube, Netflix, Gaming, and Others.

Task 1.1 - User Behavior Overview

Aggregate per user the following information:

- Number of xDR sessions
- Session duration
- Total download (DL) and upload (UL) data
- Total data volume (in Bytes) during each session for each application

Task 1.2 - Exploratory Data Analysis

Conduct EDA on the aggregated data and communicate useful insights. Ensure to identify and treat all missing values and outliers by replacing them with the mean of the corresponding column.

Deliverables:

- **Python Script and Slides Including:**
    - Description of all relevant variables and their data types.

- Analysis of basic metrics (mean, median, etc.) and their importance.
- Non-Graphical Univariate Analysis with dispersion parameters and interpretations.
- Graphical Univariate Analysis with suitable plots and one-sentence comments.
- Bivariate Analysis exploring the relationship between each application and total DL+UL data.
- Variable Transformations: Segment users into top five decile classes based on total session duration and compute total DL+UL data per decile.
- Correlation Analysis: Compute and interpret a correlation matrix for Social Media, Google, Email, YouTube, Netflix, Gaming, and Other data.
- Dimensionality Reduction: Perform PCA to reduce data dimensions and provide interpretations in four bullet points.

Task 2 - User Engagement Analysis

User engagement and activity are critical determinants of a company's success. By tracking user activities through database sessions, you can assess overall engagement and identify where to focus resources.

Engagement Metrics:

- Sessions frequency
- Duration of the session
- Session total traffic (download and upload in bytes)

Task 2.1 - Engagement Metrics Analysis

Deliverables:

- **Python Script and Slides Including:**
  - Aggregate metrics per customer ID (MSISDN) and report the top 10 customers per engagement metric.
  - Normalize each engagement metric and run k-means clustering (k=3) to classify customers into three engagement groups.
  - Compute minimum, maximum, average, and total non-normalized metrics for each cluster.
  - Visual interpretation with accompanying explanatory text.
  - Aggregate user total traffic per application and identify the top 10 most engaged users per application.
  - Plot the top 3 most used applications with appropriate charts.
  - Determine the optimized value of k using the elbow method and interpret findings.

Task 3 - Experience Analytics

User experience is influenced by various performance parameters and user device characteristics. Analyzing these factors helps optimize products and services to meet user expectations.

Focus Areas:

- **Network Parameters:** TCP retransmission, Round Trip Time (RTT), Throughput
- **Device Characteristics:** Handset type

Task 3.1 - Aggregate Experience Metrics

Deliverables:

- **Python Script Including:**
    - Aggregate per customer:
        - Average TCP retransmission
        - Average RTT
        - Handset type
        - Average throughput
    - Treat missing values and outliers by replacing them with the mean or mode of the corresponding variable.

Task 3.2 - Top and Bottom Metrics

Deliverables:

- **Python Script Including:**
    - Compute and list the top 10, bottom 10, and most frequent values for:
        - TCP values
        - RTT values
        - Throughput values

Task 3.3 - Distribution Analysis

Deliverables:

- **Python Script and Slides Including:**
    - Distribution of average throughput per handset type with interpretations.
    - Average TCP retransmission per handset type with interpretations.

Task 3.4 - Clustering User Experiences

Deliverables:

- **Python Script and Slides Including:**
    - Perform k-means clustering (k=3) based on experience metrics.

- Segment users into experience groups and provide a brief description of each cluster based on data insights.

Task 4 - Satisfaction Analysis

Customer satisfaction is influenced by both user engagement and experience. This section involves analyzing satisfaction in depth based on previous analyses.

Task 4.1 - Assign Scores

Deliverables:

- **Python Program Including:**
  - Assign an **engagement score** to each user as the Euclidean distance between the user data point and the least engaged cluster (from Task 2).
  - Assign an **experience score** to each user as the Euclidean distance between the user data point and the worst experience cluster (from Task 3).

Task 4.2 - Satisfaction Score

Deliverables:

- **Python Script and Slides Including:**
  - Calculate the average of engagement and experience scores as the **satisfaction score**.
  - Report the top 10 satisfied customers.

Task 4.3 - Predicting Satisfaction

Deliverables:

- **Python Script Including:**
  - Build a regression model of your choice to predict the satisfaction score of a customer.

Task 4.4 - Clustering Satisfaction and Experience

Deliverables:

- **Python Script and Slides Including:**
  - Run k-means clustering (k=2) on engagement and experience scores.

Task 4.5 - Aggregate Scores per Cluster

Deliverables:

- **Python Script and Slides Including:**
  - Aggregate the average satisfaction and experience scores per cluster.

Task 4.6 - Export to MySQL

Deliverables:

- **Python Script Including:**
    - Export the final table containing all user IDs, engagement scores, experience scores, and satisfaction scores to a local MySQL database.
    - Provide a screenshot of a select query output from the exported table.

Task 4.7 - Model Deployment Tracking

Deliverables:

- **Model Tracking Report Including:**
    - Code version
    - Start and end time of each run
    - Source and parameters used
    - Metrics (e.g., loss convergence)
    - Artifacts or output files related to each specific run (e.g., CSV files, screenshots)
    - **Deployment Tools:** Use Docker or other MLOps tools for tracking model changes.

Final Submission - January 28, 2025

Provide a comprehensive summary of your findings from all four tasks: **User Overview**, **User Engagement**, **User Experience**, and **User Satisfaction Analysis**.

Final Presentation Requirements

Slides:

- **Maximum:** 15 slides, including a title page and references.
- **Content Should Include:**
    - Summary of findings from each task.
    - Recommendations to the employer on the growth potential of **TellCo** (positive or negative) based on data insights.
    - Data and graphs justifying your recommendation.
    - Outline of the limitations of your analysis.
    - Final recommendation on whether the employer should purchase **TellCo**.

Dashboard:

- **GitHub Link:** Provide a link to your dashboard code.
- **Screenshot:** Include a screenshot of your deployed dashboard.
- **Deployment:** Ensure the dashboard is built using a web-based framework (e.g., Streamlit, Flask) and is easily navigable via a remote browser.

Data Analysis Code:

- **GitHub Link:** Provide a link to your complete data analysis code repository.

Best of luck with your project! If you have any questions or need further assistance, feel free to reach out through the chat or during the project consultation sessions.