**Columbia University QMSS Practicum Project**

**Project Description**

Lovelytics is a data consulting firm focusing on creating end-to-end analytics solutions using Databricks, Tableau Software, and other platforms. Epsilon is a marketing and advertising firm focused on outcome-based marketing, determined to drive growth with some of the world’s top brands, agencies, and publishers. For this project, students will work together with Lovelytics to develop an analytical read out centered on customer segmentation for marketing purposes.

**Project Goal**

The goal of the project is to work with Lovelytics team members to develop a comprehensive customer segmentation solution that can be utilized for Epsilon’s marketing purposes. During the engagement, Lovelytics and the students will focus on developing the Databricks pipelines and algorithms with data visualizations in Tableau to showcase the results in a clean and client-ready read out.

**Project Roadmap**

* Set up Databricks Community Edition and download Tableau Desktop (via student free license)
* Ingest Epsilon data files provided by Lovelytics. Create bronze tables within Databricks. Create silver tables from the bronze tables by cleaning and feature engineering (Reference Databricks Medallion Architecture)
* Run a statistical analysis on top of the silver tables. Explore relationships between the variables. Should the data be standardized, why or why not? Throughout the process of feature engineering, feel free to create new variables and/or drop variables as found fit. Develop visualizations within Databricks and summary metrics to support your findings.
* Develop clustering algorithms to explore possible household profiles.
  + Two clustering (segmentation) algorithms
    - One partitional algorithm (i.e k-means)
    - One hierarchical algorithm (i.e. agglomerative or divisive)
  + There are several parameters you will have to tune:
    - The final number of clusters, which measure of dissimilarity to use
    - The variables to use
    - Variable transformation
* Contrast benefits and setbacks of both algorithms. Choose one of the algorithms to create the final clusters. Make sure to provide a brief justification for each step of your thinking process.
* Using the final clusters, draw conclusions on possible segments of households. Create a dashboard or dashboards in Tableau. Select what you consider are the most important variables to create visualizations on Tableau that provide insights into the customers found within the data
* Presentation to Epsilon and Lovelytics Management based on work and findings.

**Project Deliverables**

|  |  |  |
| --- | --- | --- |
| Focus Area | Tasks | Success |
| Getting familiar with Databricks and Data Preparation. | Load datasets. Create bronze table \*combine several data sets. Create silver table \*Data cleaning, filling missing values, enforcing correct schema, etc. | To become familiar with Databricks UI. Databricks File Storage (DBFS). Cluster creation and notebook development. Ingest the data (Bronze Tables) and clean the data (Silver Tables) |
| Data Exploration | Run a statistical analysis on top of the silver tables. Explore relationships between the variables. Should the data be standardized, why or why not? Throughout the process of feature engineering, feel free to create new variables and/or drop variables as found fit. Develop visualizations within Databricks and summary metrics to support your findings. | Data preprocessing, statistical analysis of processed data, and visualize summary metrics in Databricks |
| Develop clustering algorithm | Develop two clustering (segmentation) algorithms to explore possible household profiles. One partitional and one hierarchical algorithm | Development of two clustering algorithms |
| Algorithm Evaluation | Contrast benefits and setbacks of both algorithms. Choose one of the algorithms to create the final clusters. | Final clusters are selected for profiling with documented evaluation and reasoning |
| Tableau dashboard | Using the final clusters, draw conclusions on possible segments of households. Create a set of dashboards in Tableau. Select what you consider are the most important variables to create visualizations on Tableau. | Tableau Dashboard is developed, and customer insights or other findings are presented in the visualizations |
| Client Read Out | Presentation to Lovelytics and Epsilon based on work and findings | Presentation and Client Read Out |

**Datasets**

|  |  |  |
| --- | --- | --- |
| **Dataset Name** | **Relatable Columns** | **Description** |
| Epsilon\_1\_ (1-5) | Personal Sequence number (psn) | Demographic and Psychographic data (Household level) |
| Epsilon\_2\_ (1-32) | Personal Sequence number (psn) | Epsilon’s credit card based transactional database (Household level) |

Note: Lovelytics will provide a data dictionary for the datasets provided. This data is specific for Maryland, Virginia, and Washington D.C.

**References**

**Databricks Medallion Architecture**  
[https://databricks.com/notebooks/delta-lake-cdf.html](about:blank)

**Segmentation with Databricks**  
[https://databricks.com/blog/2021/03/16/segmentation-in-the-age-of-personalization.html](about:blank)

**Project Mentors and Communication**

**Mentor 1**

**Name: Nate Buesgens**  
**Title:** VP, Databricks Services  
**Department/Division:** Databricks Services  
**Email Address:** [Nate.Buesgens@Lovelytics.com](about:blank)  
**Phone Number:** (410) 929-6449  
**Availability:** 1 hour per day

**Mentor 2**

**Name:** Christian Bartholomew  
**Title:** Lead Consultant  
**Department/Division:** Databricks Services  
**Email Address:** [christian.bartholomew@lovelytics.com](about:blank)  
**Phone Number:** [(732) 609-5796](about:blank)  
**Availability:** 1 hour per day