

**Team Project Brief: Understanding Customer Reviews**

**MMA 865 2023S   
Moez Ali  
Last Updated: December 7, 2022**

# The Case

Product reviews play a crucial role in the purchasing decision of a customer. Amazon.com, Inc. is one of the largest retailers in the history of the planet with revenues of USD 386 billion in 2020. Each year, customers leave millions of product reviews on the Amazon.com website. Understanding these reviews is crucial for Amazon and for the customer experience. In this project, you will explore and model these customer reviews to find out which reviews are helpful.

# The Data

We have collected millions (Big Volume) of Amazon reviews (Big Variety and Big Velocity) from the past decade. An example review is:

|  |
| --- |
| {  "reviewID": "15632",  "overall": 5.0,  "verified": true,  "reviewTime": "09 13, 2009"  "reviewerID": "A2SUAM1J3GNN3B",  "asin": "0000013714",  "reviewerName": "J. McDonald",  "reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!",  "summary": "Heavenly Highway Hymns",   "unixReviewTime": 1252800000,  “label”: 1 🡸 This is the target column } |

# Your Mission

Your mission consists of two parts.

In Part 1, you will describe trends/patterns/insights in the data and answer specific questions below.

In Part 2, you will use ML to build a predictive model and submit your model’s predictions to the course’s private Kaggle Competition.

## Part 1: Trend Exploration

You are to use Azure Databricks (Apache Spark) to explore and analyze the data to identify patterns and trends. You will focus on the following key question: **What makes a review helpful?**

## Part 2: The Kaggle Competition

You will compete against fellow teams in an attempt to build the best prediction model.

Using the labeled training data (found in the Databricks table named ***default.reviews\_train****,* you are to build a model to predict whether a review is helpful (i.e. *label=1*) given the other features (including any features you engineer) of the review. You may use any preprocessing techniques you’d like, and you may use any classification technique you deem appropriate. Once built, you will use your model to predict the helpfulness vote of a held-out test set with unknown (to you) labels. This data is in the table named ***default.reviews\_test*** in the Databricks workspace. You will submit your predictions to the course project’s ongoing Kaggle Competition by the end date to see how your model ranks against your peers’.

The Kaggle Competition is private, and must be accessed through the private link, which I will put on the portal. Please register your team on competition here: [LINK TO BE PROVIDED LATER]. Your marks for Part II will be calculated based on the highest submission made until **Feb 16, 2023 (11:59 PM EST)**

# Deliverables

You will be responsible for the following deliverables:

* At least one submission to the Kaggle competition by the competition end date **(Feb 16, 2023)**
  + Please name your Kaggle team the same as your assigned team name for the program (e.g., “Team Paris” or “Team Richardson”).
  + You will earn points based on your overall model performance as follows:
    - Level 1: AUC of 0.82 or higher *(Example notebook provided on Databricks achieve this score)*
    - Level 2: AUC of 0.84 or higher
    - Level 3: AUC of 0.86 or higher
    - Level 4: AUC of 0.88 or higher
  + You may earn bonus points as follows:
    - Highest Kaggle score in section: 2 bonus points
    - Top-3 Kaggle score in section: 1 bonus point
* A 15-minute presentation, to be delivered live in class, that includes:
  + The insights/results for Part 1
  + A description of your modeling journey for Part 2.
  + An estimate for how much it cost to develop your insights and models, in terms of both cloud computing costs, and developer/analyst time.
  + A PDF of your presentation slides, named: 00\_TeamName\_Presentation.pdf.
* You are not to submit source code used during the project. However, you should be prepared to supply your source code, should I request it. Your code should be fully reproducible, well-documented, and ready for inspection.

Only the PDF file shall be submitted to the course portal.

# Databricks

Teams can use Microsoft Azure Databricks (which uses Apache Spark) for creating their analysis and modeling pipelines. This cloud-based tool is the state-of-the-art method for working with Big Data. Databricks provides a Notebook-like interface for writing SQL, Python, R, and/or Scala code to work with datasets.

To access Databricks, head to <https://canadacentral.azuredatabricks.net/login.html> and use your Queen’s username and password to log in. A Spark cluster has been created for each team. Clusters are made of Standard\_DS12\_v2 (28 GB, 4 CPU) nodes and can scale out to 4 worker nodes. Teams can start and stop (called “terminate”) their clusters on their own (i.e., IT is not needed). Clusters will auto-terminate after 30 minutes of inactivity.

The typical workflow will be to sign into Databricks, start your cluster (which takes 3-5 minutes), open a Notebook (either new or saved), and write/run code. When you are done for the day, you close your Notebook and terminate the cluster.

The MMA program will be covering the costs of using the tool, but please be mindful of how much you are using/costing. For example, please don’t leave a Notebook stuck in a loop for 4 days straight, incurring costs the entire time. See more pricing details at the [Azure Databricks Pricing](https://azure.microsoft.com/en-us/pricing/details/databricks/) page. The school uses the Standard Tier, West US Region, pay-as-you-go model.