Harnessing Helpfulness: Amazon Product Reviews to Elevate Marketplace Experience

DATA 270
Under guidance of Dr. Linsey Pang

Group 5

Presented by:

Eshita Gupta
Monica Lokare
Sneha Karri
Veena Ramesh Beknal

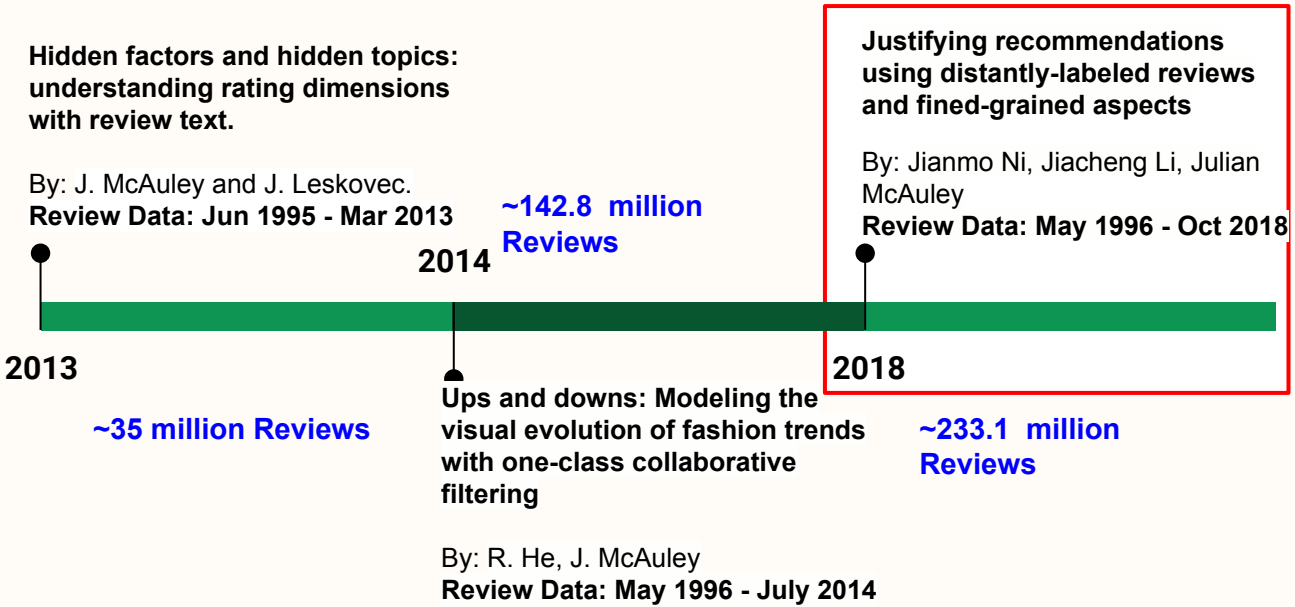
Chapter 2: Data & Project Management Plan

Data Management Plan

Data Collection Approaches

Amazon, is the primary destination for people looking to make online purchases across a wide spectrum of product categories making it the leading e-commerce platform in the United states. Amazon reviews play an integral part in the purchasing decisions made by customers as the listing for similar products can be quite overwhelming. We plan to use the **Amazon Review dataset published by Julian McAuley, UCSD (UC San Diego). We will use 2018 version of the dataset.**

Timeline



Dataset Variations

Raw: Raw data for 233.1 million reviews

5-Core: Selected data for 75.26 million reviews

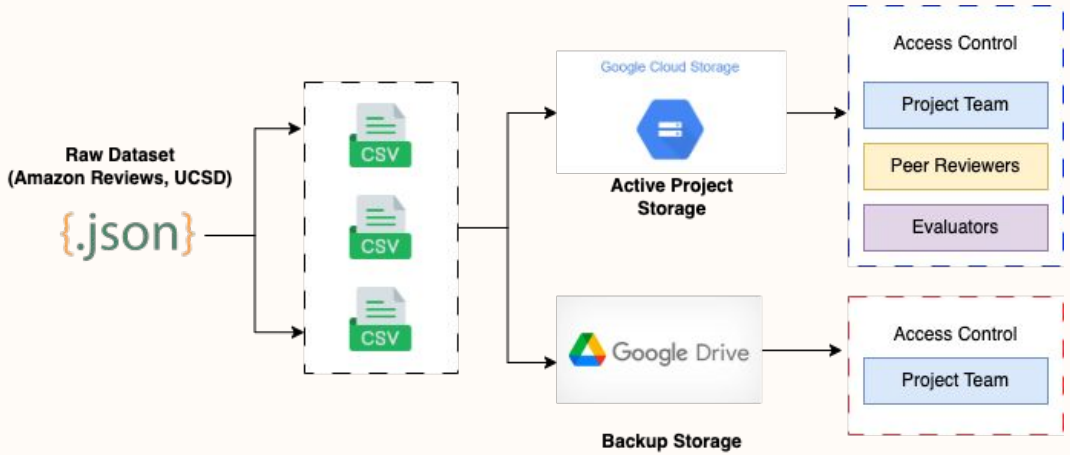
We will use a combination of the above two and filter the reviews from **2010-2018**. We will use the following categories:

- Appliances (Raw)
- Automotive (K-Core)
- Cell phones and accessories (K-Core)
- Tools & Home improvement (K-Core)

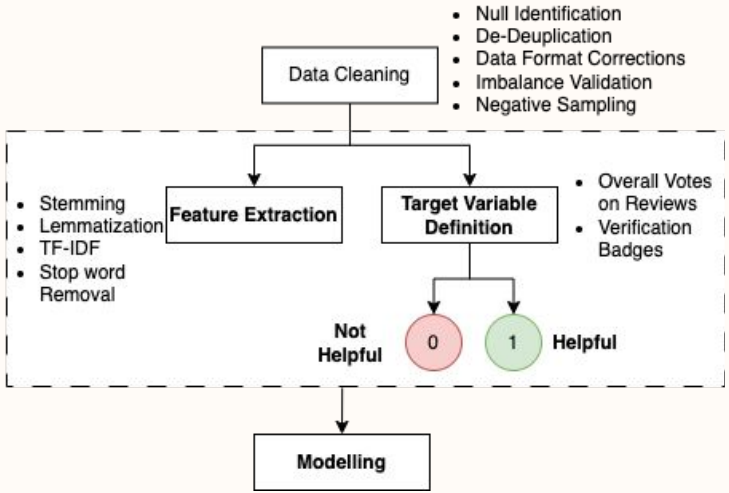
Data Management & Storage Methods

UCSD dataset is gigabytes in size due to the sheer volume of the reviews. The raw data is available in JSON format and using the data in its original form is not desirable or efficient. We will therefore use GCS (Google Cloud Storage) to efficiently handle the large volumes of data. Raw JSONs will be converted into CSV files and stored on GCS and Google Drive for backup. Access will be controlled using Google authentication and security policies. Data will be cleaned, sanitized, pre-processed, reviewed and then used for modelling.

Storage



Processing



Data Usage Mechanisms

The Amazon dataset is a publicly available dataset and data usage is governed by the rules laid out by the original publishers and usage of this data requires explicit citations.

- The data is originally available in JSON format and we will convert them into CSV format for facilitating convenient parsing and pre-processing.
- The converted files will be placed on GCS (Google Cloud Storage), however, as the data is publicly available on the original publisher's webpage, we will not be opening access to this data on GCS.
- We will store the data only until the project completion on GCS and post completion the data will be purged, data will be persistent on our personal Google Drives for future reference.
- Any access privileges extended for peer reviews will be revoked from GCS.

Amazon Review Dataset (2018)
Publisher: Jianmo Ni (UCSD)
URL: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/
Citations:
Justifying recommendations using distantly-labeled reviews and fined-grained aspects
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019

Project Development Methodology

Business Understanding

- During this phase, we will focus on the project's **objective** and its **deliverables**.
- **Our goal is to improve the helpfulness of reviews on e-commerce sites such as Amazon.com** where users post product reviews
- **Our project aims to enhance the utility of product reviews on e-commerce platforms like Amazon, which is a leader in the US market with a 38% share and $574.8 billion in net sales in 2023**
- By focusing on areas where reviews can be objective, **excluding subjective categories like literature and fashion**
- **77% of consumers seek out websites with ratings and reviews**, and **98% consider them essential in their decision-making process**
- Enhancing review helpfulness **benefits both buyers**, by aiding in more informed decisions, **and sellers**, by providing valuable feedback and improving customer satisfaction
- Encouraging the inclusion of beneficial attributes in reviews and product listings can make the marketplace more vibrant and improve the overall experience for all parties involved

Data Understanding

- In this phase, **we focus on data exploration and understanding by selecting objective review categories from the Amazon Reviews dataset**
- Categories covered: Appliances, Tools and Home Improvements, Automotive, and Cellphone and Accessories.
- **The raw JSON data is converted to CSV** and **deduplicated** to ensure uniqueness
- The analysis **includes reviews from 2010 to 2018 to reflect the period after the mass adoption of the internet and e-commerce**
- We **examine the data types, distributions, patterns, and check for any inconsistencies** like missing information or duplicates
- Additionally, we **analyze the distribution of ratings across categories and review volumes over time** to detect trends such as seasonality and identify peak buying periods
- This comprehensive data understanding and cleaning process sets the foundation for effective model building in later stages

Data Preparation

- In the data preparation phase, we **aim to define a binary dependent variable by combining helpful votes and verification status** from the raw review data
- **Reviews from verified buyers with a significant number of helpful votes are labeled as helpful (1)**, while **others are considered not helpful (0)**, framing our problem as binary classification
- We utilize 3 categories of features for prediction: **raw features** (e.g., ratings), **metadata features** (e.g., review text length, word counts), and **tokenized review features** (created through text processing techniques like TF-IDF, stemming, and lemmatization)
- **Feature engineering is employed to refine these features** by eliminating redundancy, reducing sparsity, and maintaining review integrity with minimal imputation
- **High dimensionality, especially from TF-IDF matrices, necessitates dimensionality reduction** (via techniques like PCA or t-SNE), **balancing between feature volume and explainability**
- This meticulous preparation leads to a dataset ready for the modeling phase

Modeling

- During the modeling phase, **we will build supervised classification machine learning models** using **Python libraries such as numpy, pandas, NLTK, sklearn, textblob, wordcloud, and xgboost**
- After we label the data, we will **randomly split the data into training and test sets**, with the **training set being 70% while test set is 30%**
- Once the data is split, we can apply various models for classification, such as **Logistic Regression, Support Vector Machine (SVM)**, as well as tree-based ensemble models like **Random Forest** and **XGBoost** on the training dataset
- We will **build a baseline model with default parameters for the chosen categories**
- Next, we will **find optimal hyperparameters using the methods like grid search, random search and cross validation for tuned model**
- We will **compare the results of the baseline and tuned model and analyze model performance improvements**
- We **anticipate the XGBoost model will perform better than all the other models** since it is an ensemble modeling technique that uses boosting to improve performance while reducing the possibility of overfitting
- **Best model will be chosen for each product category and analyzed**, and the results will be compared with all the other models for each category

Evaluation

- **Evaluation metrics** used in our project will be the **Receiver Operating Characteristics curve (ROC), Area under the Curve (AUC), Accuracy, Precision, Recall, and confusion matrix**
- For a binary classifier, ROC is a graph that acts as a performance indicator
- The area under the curve captures the area under the ROC
- Confusion matrix is a visualization in the form of the actual labels versus the model's predictions
- Accuracy is a measure of the overall correctness of the model
- Recall is a true positive rate. Precision is a positive predicted value
- F1 score is the harmonic mean of recall and precision which gives us a general view of performance
- **Feature importance can be computed only for Logistic Regression** (using coefficient), **Random Forest** and **XGBoost** (using Gini importance)
    - In Logistic Regression, the dependent variable for the model would be the log-odds of helpfulness, and hence the coefficients of the features can be interpreted as their importance

Deployment

- In the deployment phase, we will take critical steps to implement our project. We will begin by uploading all the code, designed to evaluate the helpfulness of Amazon reviews, into a private GitHub repository.

- This will enable us to document our work in a centralized manner and have further enhancements and collaboration by team members.

- Following testing, we will document the performance of the codebase and conduct a demonstration of the final models.

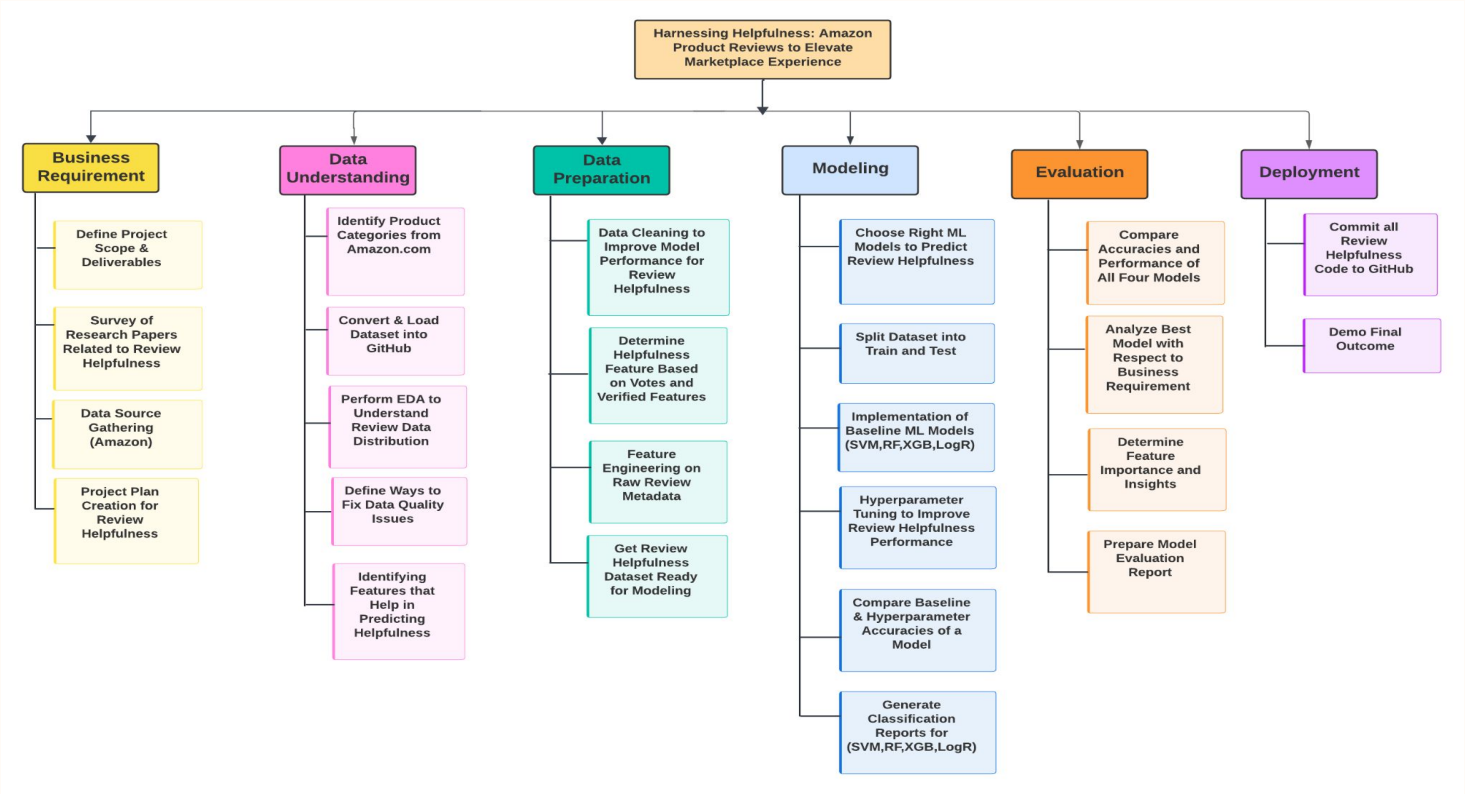
- We will monitor the deployment to quickly address any issues that arise and to fine-tune the model based on user feedback.

- This is a phased implementation for ongoing checks and enhancements which makes sure our project performs smoothly.

- By planning carefully and executing adaptively, we aim to deliver an enhanced shopping experience for customers and sellers that values and highlights the most helpful product reviews.

Project Organization Plan

Work Breakdown Structure



We will be using CRISP-DM (Cross-Industry Standard Process for Data Mining) for our project, which is highly beneficial framework as it focuses on improving Amazon's review experience.

- In the first phase, business requirements, we will define the project scope, deliverable, conduct a literature survey and gather all required data sources and identify tools, technologies, and methodologies required to carry out our project. We create a project plan to execute our entire project, which establishes the project's purpose, deliverables, and key metrics that we can use for evaluating the project.

- The second phase is data understanding, which mainly focuses on exploring data and understanding the dataset thoroughly. We convert JSON to CSV for the product categories that as subjective and perform data exploration to understand the feature distribution, the types of variables, their data types, and dependencies. Here we also identify data quality issues and define ways to address these issues in order to prepare data for the data preparation phase.

- The third phase is the crucial phase, where we prepare our data for predictive modeling. Tasks involved are removing duplicate records, standardizing, imputing missing data, and creating a dependent variable based on votes and verified features. Feature engineering is performed on raw data which includes review metadata, extracting useful features from text, and plotting graphs for data understanding. The TF-IDF vectorizer is created to get our dataset ready for modeling.

- In the Modeling phase, once we finalize the machine learning models to predict review helpfulness we split the dataset into train and test, and implement the models. We calculate accuracy and other evaluation metrics and also display classification report. We find optimal hyperparameters and perform hyperparameter tuning to improve the performance of our models.

- In the Evaluation phase we analyze each member's machine learning model and compare the accuracies of all four classification models. We identify the highest accuracy model and provide a proper justification for selecting the model with best accuracy.

- In the final Deployment phase, we push all the codes to a version control system like GitHub and demo our final results, findings, and project methodology.

Project Resource Requirements and Plan

Tools and Cost Estimation

- For our project we will utilize the Python Jupyter Notebook (6.5.4 version) in combination with visualization libraries such as Matplotlib and Seaborn
- Windows machine with 16 GB RAM and a 64-bit processor for this project
- The project files will be stored in GitHub, while Google Cloud storage can be used to store the dataset and other project files
- For model training, testing, and deploying machine learning models, we will be using Vertex AI service
- Machine learning frameworks like sklearn, wordcloud, nltk, and xgboost libraries are used in the our project
- Additional software such as MS Office Professional Plus, Draw.io, Github, ClickUp, and Lucid software will be utilized
- We have estimated a total cost of $1000 for our project

Hardware Requirements

Hardware	Configuration	Purpose
Cloud Storage	2 GB	Store the CSV files
Google Drive	2 GB	Store the CSV files, Raw JSON Files
Vertex AI Training	e2-standard-4(VCPUs:4, RAM:16)	Model training
Vertex AI Prediction	e2-standard-4(VCPUs:4, RAM:16)	Model deployment
Local Windows Machine	16 GB RAM, 64-bit processor	Preprocessing, Model Development and Testing

Software Requirements

Resource	Version	Purpose
Python Jupyter Notebook	6.5.4	Project Development
NumPy	1.24.3	Preprocessing, cleaning
Pandas	2.0.3	Preprocessing, cleaning
scikit-learn or sklearn	1.3.0	Model Development
nltk	3.8.1	Preprocessing, cleaning
wordcloud	1.9.3	Preprocessing
xgboost	2.0.2	Model Development
GitHub	--	Project code management
ClickUp	--	Project Management Tool
MSOffice Professional Plus	2021	Creating and editing reports
Draw.io	--	Creating Flow diagrams
Lucid Chart	--	Creating work breakdown structure

Resources and Cost Estimation

Utility	Resource Type	Tool/Application	Duration	Cost Estimation
Data Storage	Hardware	Cloud Storage	2 months	Free (Student Credits)
Data Preprocessing	Software	Python Jupyter notebook	2 months	Free
Machine Learning Frameworks	Software	sklearn, wordcloud, nltk, xgboost	2 months	Free
GitHub	Software	Data Files and Project Files	2 months	Free
Local Machine	Hardware	64-bit Version	2 months	$1000
Machine Learning Modeling	Software	Vertex AI	2 months	Free (Student Credits)
Visualization Tool	Software	Looker	2 months	Free (Student Credits)

Project Schedule

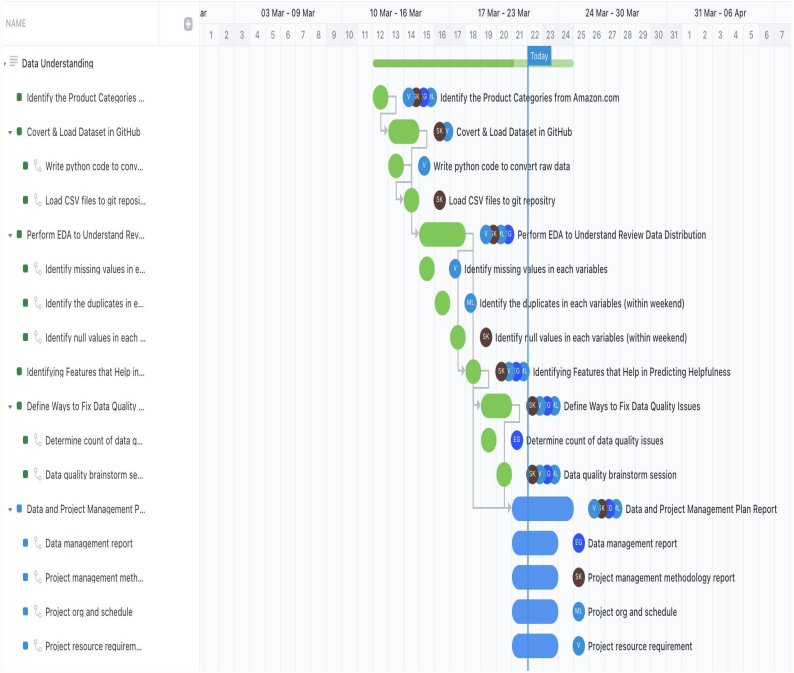
Gantt Chart

- Gantt chart is used to track tasks and the status of the project, which is displayed in the form of a bar chart. We wanted the flexibility of moving back and forth between phases, due to which we've structured our Gantt chart to match the CRISP-DM framework, along with an agile methodology.
- Using this approach, we can understand the task dependencies, duration, and individual responsibility. The x-axis represents the project's timeline, and the y axis represents the tasks and subtasks. We read the graph from left to right to keep track of the task’s status.
- We have broken down each phase into multiple sprints, which last 2 weeks, and will be delivered at the end of each sprint. Each sprint has tasks and sub tasks allocated to group/individual team members.
- There are seven sprints in total. A subtask duration is estimated by the difficulty level and amount of work involved. At any point, we can go back to the previous phase /tasks/ subtasks to change or modify the existing information.
- After each sprint, we have a retrospective meeting to understand the challenges faced and areas of improvement for upcoming sprints.

Gantt Chart and it’s Dependencies of Business Understanding

Task Number	Task Name	Depends On	Start Date	End Date	Duration
TS1	Define Project Scope & Deliverables	–	Jan 29	Feb 9	12
TS2	Prepare Project Abstract Report	TS1	Feb 6	Feb 9	4
TS3	Survey of Research Papers Related to Review Helpfulness	TS2	Feb 10	Feb 24	15
TS4	Data Source Gathering(Amazon)	TS1	Feb 6	Feb 9	4
TS5	Project Plan Creation for Review Helpfulness	TS4	Feb 10	Feb 14	5
TS6	Create Project Introduction Report	TS1, TS2,TS3	Feb 25	Mar 9	14

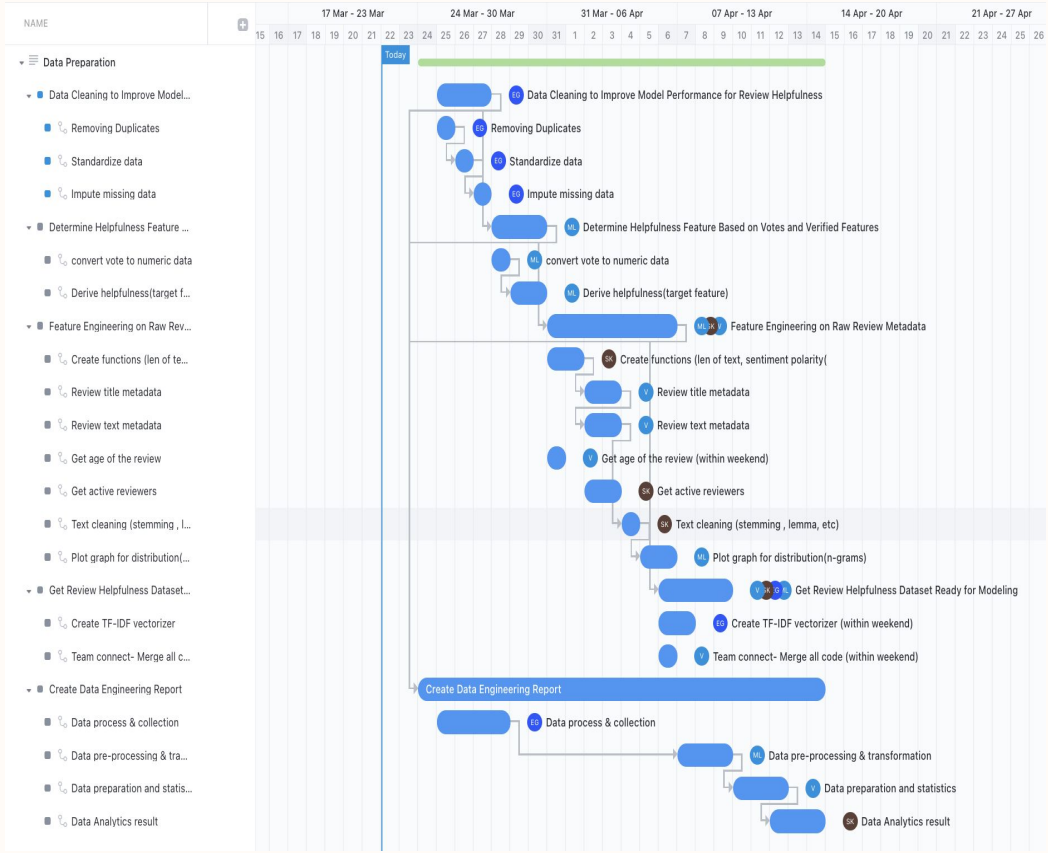
Gantt Chart and it's Dependencies of Data Understanding



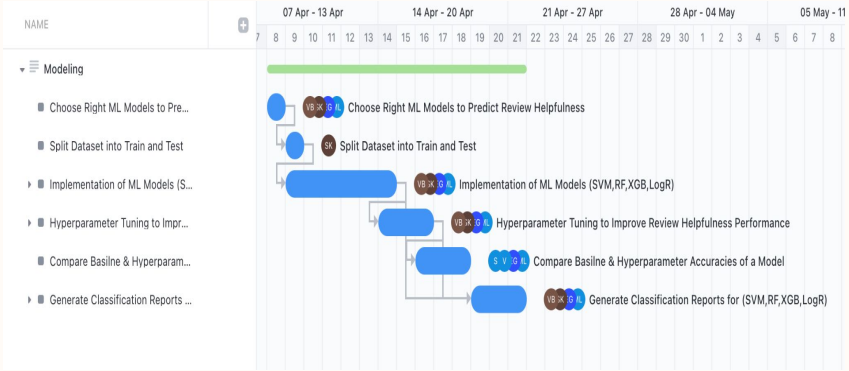
Task Number	Task Name	Depends On	Start Date	End Date	Duration
TS1	Identify the Product Categories from Amazon	--	March 12	March 12	1
TS2	Convert & Load Dataset in GitHub	TS1	March 13	March 14	12
TS3	Perform EDA to Understand Review Data Distribution	TS2	March 15	March 17	3
TS4	Identifying Features that Help in Predicting Helpfulness	TS3	March 18	March 18	1
TS5	Define Ways to Fix Data Quality Issues	TS4	March 19	March 20	2
TS6	Data and Project Management Plan Report	TS3, TS5	March 21	March 24	4

Task Number	Task Name	Depends on	Start Date	End Date	Duration
TS1	Data Cleaning to Improve Model Performance for Review Helpfulness	--	March 25	March 27	3
TS2	Determine Helpfulness Feature Based on Votes and Verified Features	TS1	March 28	March 30	3
TS3	Feature Engineering on Raw Review Metadata	TS2	March 31	April 6	7
TS4	Get Review Helpfulness Dataset Ready for Modeling	TS3	April 6	April 7	2
TS5	Create Data Engineering Report	TS1,TS2,TS3,TS4	March 24	April 14	22

Gantt Chart and it's Dependencies of Data Preparation

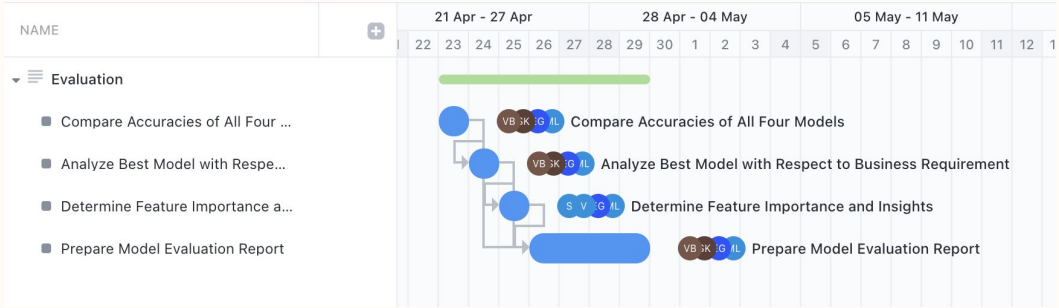


Gantt Chart and it’s Dependencies of Modeling



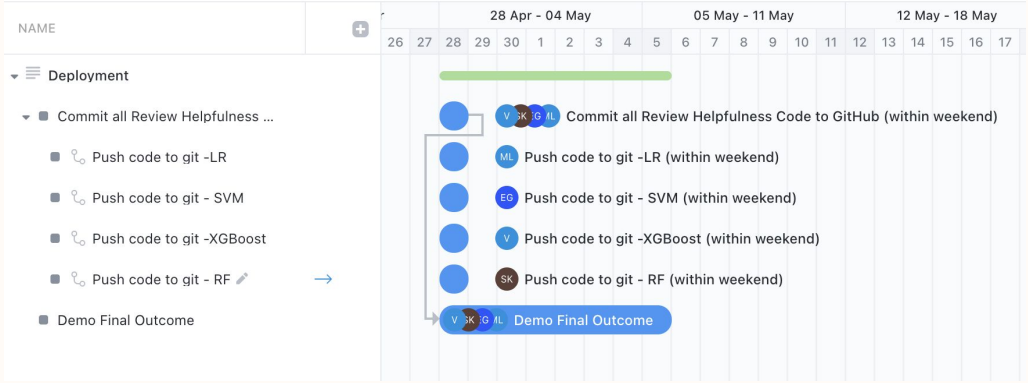
Task Number	Task Name	Depends On	Start Date	End Date	Duration
TS1	Choose Right ML Models to Predict Review Helpfulness	--	April 8	April 8	1
TS2	Split Dataset into Train and Test	TS1	April 9	April 9	1
TS3	Implementation of ML Models (SVM,RF,XGB,LogR)	TS2	April 9	April 14	6
TS4	Hyperparameter Tuning to Improve Review Helpfulness Performance	TS2, TS3	April 14	April 16	3
TS5	Compare Baseline & Hyperparameter Accuracies of a Model	TS4	April 16	April 18	3
TS6	Generate Classification Reports for (SVM,RF,XGB,LogR)	TS4,TS3	April 16	April 21	6

Gantt Chart and it’s Dependencies of Evaluation



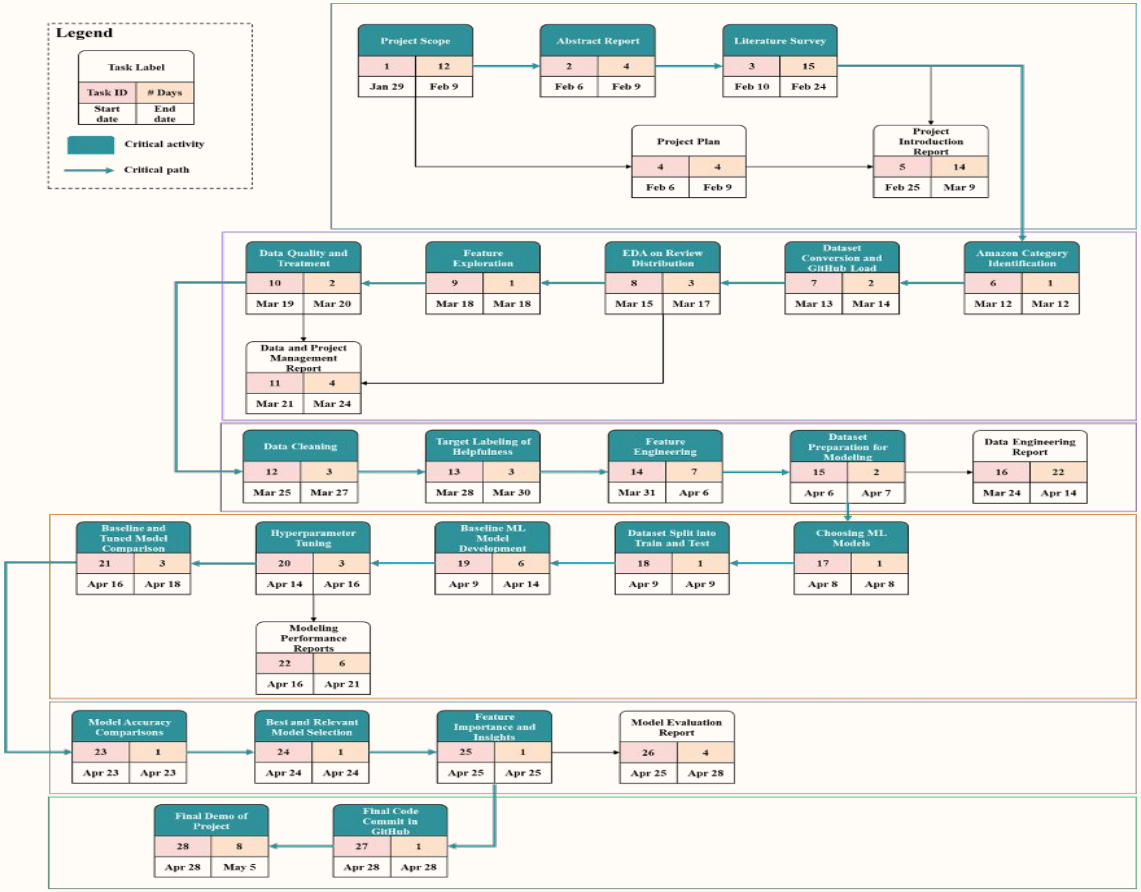
Task Number	Task Name	Depends On	Start Date	End Date	Duration
TS1	Compare Accuracies of All Four Models	--	April 23	April 23	1
TS2	Analyze Best Model with Respect to Business Requirement	TS1	April 24	April 24	1
TS3	Determine Feature Importance and Insights	TS2	April 25	April 25	1
TS4	Prepare Model Evaluation Report	TS1, TS2, TS3	April 25	April 28	4

Gantt Chart and it’s Dependencies of Deployment



Task Number	Task Name	Depends On	Start Date	End Date	Duration
TS1	Commit all Review Helpfulness Code to GitHub	--	April 28	April 28	1
TS2	Demo Final Outcome	TS1	April 28	May 5	8

PERT Chart


Legend
Task Label
Task ID
Days
Start date
End date
Critical activity
Critical path
Project Scope
1
12
Jan 29
Feb 9
Abstract Report
2
4
Feb 6
Feb 9
Literature Survey
3
15
Feb 10
Feb 24
Project Plan
4
4
Feb 6
Feb 9
Project Introduction Report
5
14
Feb 25
Mar 9
Data Quality and Treatment
10
2
Mar 19
Mar 20
Feature Exploration
9
1
Mar 18
Mar 18
EDA on Review Distribution
8
3
Mar 15
Mar 17
Dataset Conversion and GitHub Load
7
2
Mar 13
Mar 14
Amazon Category Identification
6
1
Mar 12
Mar 12
Data and Project Management Report
11
4
Mar 21
Mar 24
Data Cleaning
12
3
Mar 25
Mar 27
Target Labeling of Helpfulness
13
3
Mar 28
Mar 30
Feature Engineering
14
7
Mar 31
Apr 6
Dataset Preparation for Modeling
15
2
Apr 6
Apr 7
Data Engineering Report
16
22
Mar 24
Apr 14
Baseline and Tuned Model Comparison
21
3
Apr 16
Apr 18
Hyperparameter Tuning
20
3
Apr 14
Apr 16
Baseline ML Model Development
19
6
Apr 9
Apr 14
Dataset Split into Train and Test
18
1
Apr 9
Apr 9
Choosing ML Models
17
1
Apr 8
Apr 8
Modeling Performance Reports
22
6
Apr 16
Apr 21
Model Accuracy Comparisons
23
1
Apr 23
Apr 23
Best and Relevant Model Selection
24
1
Apr 24
Apr 24
Feature Importance and Insights
25
1
Apr 25
Apr 25
Model Evaluation Report
26
4
Apr 25
Apr 28
Final Demo of Project
28
8
Apr 28
May 5
Final Code Commit in GitHub
27
1
Apr 28
Apr 28

PERT Chart

- PERT chart is used widely in organizations to analyze the project tasks and time required to complete the project. Here we calculate the minimum amount of time needed to finish the project. It gives a visual representation of events in the project timeline.

- Each component has a task label, task id, estimated days, start date and end date. The task ids are sequential. The critical task labels have been highlighted while non critical tasks don't have a highlight in the labels. Most of the report generation tasks have more estimated number of days than the other tasks.

- In the first phase more time was spent on defining project scope followed by literature and technology survey which is 12 and 15 days respectively. This phase start date is January 29th and end date was March 9th which is two sprints a total of 4 weeks.

- In the data understanding phase the tasks estimated days were between 1 to 4 depending on the complexity. It took one sprint to complete this phase with a start date of March 12th to March 24th.

- Since we started generating the data engineering report on day one of sprint 5 for the data preparation phase, it has the highest estimated days, which is March 24 to April 14.

- In the modeling phase from April 8th to April 21st, Baseline model takes 6 days followed by other tasks estimated at 3 days and 6 days. It takes a week for evaluation and another week for deployment during the last sprint. Tasks are given 1 day each except for the report and final demo.

PERT Chart (Contd..)

- The first step in the project is to define the scope of the project based on which we draft the project abstract followed by literature and technology survey. Amazon category identification in the data understanding phase can be started as soon as the research survey is done. It doesn't really depend on the project plan and project introduction reports which is why these are marked as non-critical tasks.

- On the review dataset, EDA is performed and the significance of each feature is analyzed thoroughly, which aids in identifying the target feature. Also, we identify data quality issues and define methods for resolving them. All these steps mentioned are critical in order to move to the next phase which is data preparation.

- Data is cleaned by identifying the missing, null and duplicate values. Target variable is created based on the vote and verified feature of the dataset. Feature engineering is performed on the review metadata. Finally the dataset is made ready for the modeling phase by creating a TF-IDF vectorizer.

- Moving to the modeling, we identify the right machine learning model, split the cleaned dataset into train and test, baseline model for hyperparameter tuning and compare the baseline model with tuned model for accuracy comparison in the evaluation phase. Model performance report is the non critical task hence upon model comparison we choose the best model and determine the insights for evaluation.

- Model evaluation report is a non critical task here due to which we push the code to GitHub as soon as insights are drawn from the best model. Lastly we submit our final report and presentation as a completion of the project.

Thank You