# Title: Amazon Product recommendation system using customer review analysis

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Simran Anil | Mhaske | smhaske@hawk.iit.edu |
| Yashraj Sanjay | Diwate | ydiwate@hawk.iit.edu |
| Pratik | Patil | ppatil24@hawk.iit.edu |
| Ishan | Prabhune | iprabhune@hawk.iit.edu |
| Harshal | Sawant | hsawant1@hawk.iit.edu |
| Hrishikesh | Pawar | hpawar@hawk.iit.edu |

Important Notes:

* Each group must submit ONLY one copy by a single team member!
* Do not worry! Since you will list all team members in the table above.

1. **Introduction**

Introduce the background of your application and give me the motivations why you want to do that.  
  
Our project focuses on analyzing Amazon customer reviews using a recommendation system. The motivation behind this project stems from the increasing reliance on online reviews for purchasing decisions. With the vast amount of reviews available on Amazon, it becomes challenging for users to sift through them to find relevant and trustworthy information. By implementing a recommendation system, we aim to provide users with personalized recommendations based on their preferences and interests, thus enhancing their shopping experience on the platform.

1. **Data Sets**

Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, what is your target variable, and so forth

Tell where the data is, such as giving Kaggle URL

Tell the size of the data

Tell the variables

Tell the target variables, if you are going to perform predictive tasks  
  
We have obtained our dataset from amazon reviews github, specifically the Amazon Customer Reviews dataset. The dataset contains customer reviews from various product categories on Amazon, collected over a certain period. The size of the dataset is approximately 130 million records, with multiple features such as review rating, title, text, images, asin, parent\_asin, user\_id, timestamp and helpful\_vote. Our target variable for this project is to predict the likelihood of a user finding a review helpful based on the review content and associated metadata.

- Data Source: https://snap.stanford.edu/data/web-Amazon.html

- Size: Approximately 130 million records

- Variables: rating, title, text, images, asin, parent\_asin, user\_id, timestamp and helpful\_vote

- Target Variable: rating

1. **Research Problems**

List your research problems, that is, what kinds of the problems you want to solve.  
You cannot simply say I want to explore the data and find the patterns  
If you decide to work on a classification task, you must identify labels.  
If your project is involved with multiple data mining tasks, you should clearly mention each problem and why you want to do that.  
You should provide finer-grained research problems that can be solved by data analysis/mining techniques. If it is an implementation project, you should introduce the challenges in implementing or development and how you will evaluate them

We aim to predict the helpfulness of Amazon reviews based on various factors such as review text sentiment, product category, and reviewer characteristics. This problem is crucial for guiding users towards informative reviews and improving overall user experience for better purchasing. This can be done by using techniques such as collaborative filtering or content based filtering.

Various data mining tasks will be performed on the dataset such as - finding null values, understanding the data needs and datatypes, discretization, normalisation , etc. All of these is required to perform to get the recommendation system working.

1. **Potential Solutions**

For each problem you list above, figure out feasible solutions, and introduce your plan to perform experiments

For recommendation purpose, we plan to leverage machine learning techniques such as sentiment analysis to analyze review text sentiment, topic modeling to extract key themes from reviews, and feature engineering to incorporate metadata like product category and reviewer information. We will experiment with classification algorithms such as logistic regression, random forest, and gradient boosting to build predictive models.

1. **Evaluations**

There could be multiple solutions for a same problem, You must figure out how to evaluate them and the details about your evaluations, for example, hold-out or N-folds evaluation?, which metrics you will use for evaluations.  
  
 - For recommendation system, we will evaluate the performance of our models using metrics such as accuracy, precision, recall, MAE, RMSE and F1-score. We will employ techniques like cross-validation or hold-out validation to ensure robustness and generalization of the models.

- For evaluation we plan to use hold-out evaluation along with few hyperparameters as the data size is huge and to compare the results with different settings.

1. **Expected Outcomes**

Introduce your expected outcomes for your project

For this project, our goal is to leverage customer-provided reviews and sentiment analysis to recommend similar products to users. By analyzing the sentiment expressed in the reviews, we can understand customers' feelings towards certain products. In the provided review data, both the review title and text convey positive sentiments, indicating satisfaction with the product. This positive sentiment aligns with the high rating of 5.0 given by the reviewer, suggesting a positive evaluation overall. By applying sentiment analysis techniques, we can identify patterns in customer sentiment and use this information to recommend products with similar positive attributes to users. In essence, sentiment analysis helps us extract valuable insights from customer feedback, enabling us to make meaningful product recommendations tailored to user preferences.