Milestone 1 - Dataset and Goals

Team ID: 21

Team Members: Phuong Pham, Jing Lin, Shang-Yun (Maggie) Wu

- Data to Use: Kickstarter Project Funding Prediction (https://www.kaggle.com/dilipajm/kickstarter-project-funding-prediction/data)
    - Public data repository to predict whether or not a kickstarter project will be funded (funding is successful only when exact amount is raised)
    - The data will be used for a binary classification problem (funded vs. not funded)
    - Possible features include, but are not limited to:

Features	**Description**
Project ID	Unique ID of the project (string + int)
Name	Name of the project (string)
Description	Description of project (string)
Goal	Goal (amount) required for the project (int)
Keyword	Keywords which describe the project (string)
Disabled Communication	Communication option disabled for donors (boolean)
Country	Country of the project author (string)
Currency	Currency in which goal is required (string)
Deadline	Deadline to achieve the goal by (unix time format)
State Change At	Date of change in current status (unix time format)
Created At	Date when project is published (unix time format)
Launched At	Date when project donation goes live (unix time format)
Backers Count	Number of donors (int)
Final Status	Whether or not project is funded (0/1)

Table 1. Features and description of the dataset

- Possible Questions to Ask
    - Whether or not a project is funded based on features
    - What is the best time to launch a project
    - What kind of description leads to a successfully funded project
    - How likely is a project to be funded based on duration of donation
    - Relationship between goal amount and final status
- Pre-processing Data
    - Only training data is provided but a subset of it will be used as test set
    - Cross validation will be used to compare the performance of the model
    - Donation period is extracted from date-related features
    - Fill in missing data / eliminate data with many missing features
- Plan of Action
    - Split dataset into training and test set
    - Prepare and clean the data
    - Analyze, identify patterns and explore the data
        - Finding correlation between features and prediction
        - Finding correlation among numerical and categorical features
        - Obtain more features from given feature (eg. funding duration = deadline - launch time)
        - Parse project description and analyze project description sentiment
    - Model, predict and solve the problem
        - Explore various models (eg. Naive Bayes, Random Forest, Decision Tree, SVM, etc)
        - Building neural network and tunings on hyperparameters
        - Cross-validate to compare the performance of different models
    - Evaluate on test set
    - Visualize, report and present the problem solving steps and final solution
        - Backward engineering to reason about the significance of certain features
        - Reason about errors in different models