Milestone 4 - Initial Results

Team ID: 21

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

- Feature Engineering
    - Launched month
        - Extracted the month in which the project is launched for funding
    - Launched year
        - Extracted the year in which the project is launched for funding
    - Launched quarter
        - Extracted the quarter in which the project is launched for funding
    - Project name length
        - Extracted the length of the name of the project submitted
    - Project name first character alphabet
        - Group projects by alphabet on the first character.
    - Keyword extraction
        - Buzzword count: the count for 25 most common startup keywords in each project description
        - Keywords selected: *['app', 'platform', 'technology', 'service', 'solution', 'data', 'manage', 'market', 'help', 'mobile', 'users', 'system', 'software', 'customer', 'application', 'online', 'web', 'create', 'health', 'provider', 'network', 'cloud', 'social', 'device', 'access']*
- Data Visualization
    - Library used: seaborn

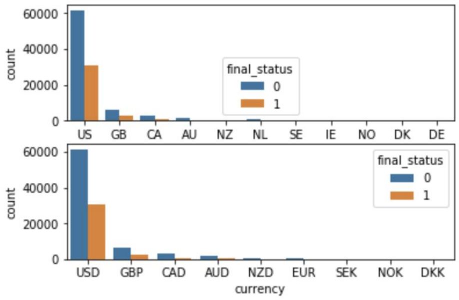

Image 1. Country and currency data with respect to final status (1 - funded, 0 - not funded)

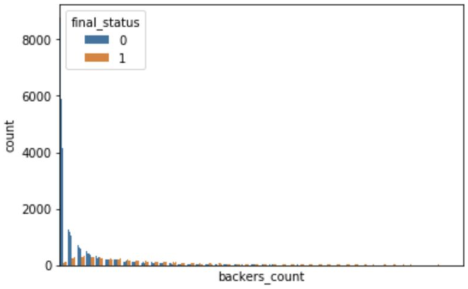

Image 2. Backer count with respect to final status (1 - funded, 0 - not funded)

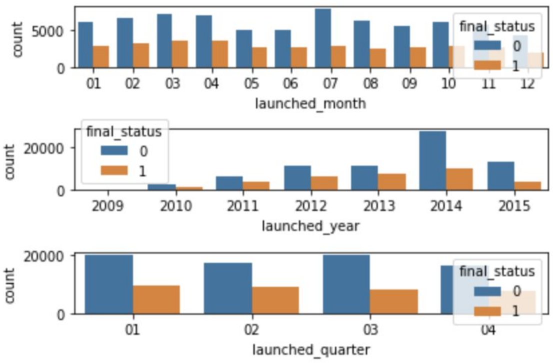

Image 3. Launched time with respect to final status (1 - funded, 0 - not funded)

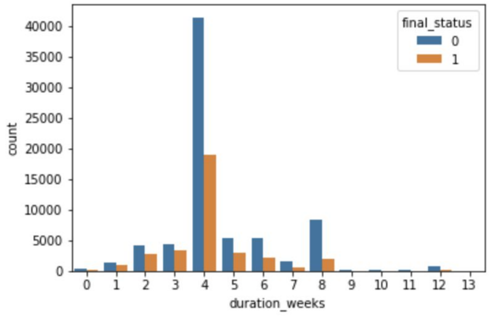

Image 4. Funding period duration in weeks with respect to final status (1 - funded, 0 - not funded)

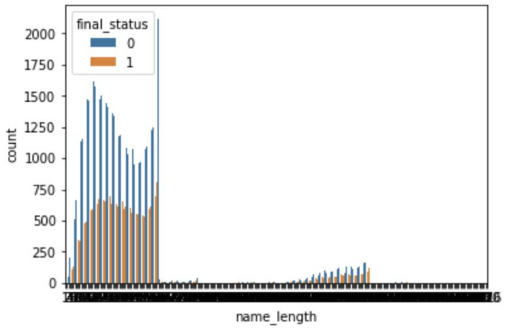

Image 5. Length of project name with respect to final status (1 - funded, 0 - not funded)

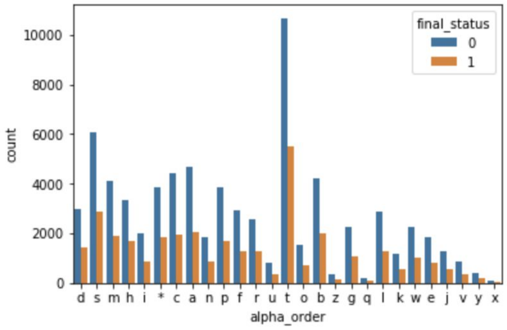

Image 6. The projects grouped by alphabetical order on the first character with respect to final status (1 - funded, 0 - not funded)

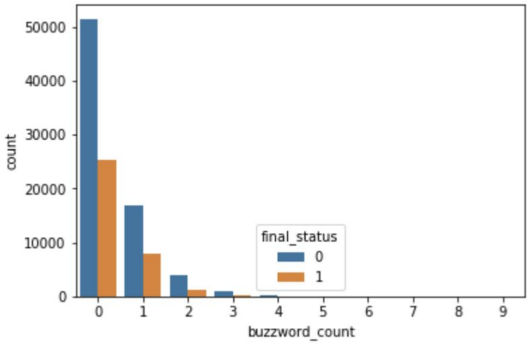

Image 7. The count of buzzwords in project description with respect to final status (1 - funded, 0 - not funded)

    - Based on the results of these analysis, we have used combination of these features to perform logistic regression
    - Some have been found to have stronger correlation with the final status than others and more features will be extracted to select the best set of features that may be used across all three algorithms
- Data Split
    - Train-Test: 80-20
- Logistic Regression
    - Library used: sklearn
    - L1 Penalty
        - Solver: saga

Trial	Features	Acc
1	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘duration_weeks’]	0.677
2	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks’]	0.678
3	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘name_length’, ‘alpha_order’]	0.781
4	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘name_length’, ‘alpha_order’, ‘buzzword_count’]	0.781
5	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘buzzword_count’]	0.678

    - L2 Penalty
        - Solver: lbfgs

Trial	Features	Acc
1	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘duration_weeks’]	0.797
2	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks’]	0.800
3	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘name_length’, ‘alpha_order’]	0.801
4	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘name_length’, ‘alpha_order’, ‘buzzword_count’]	0.800
5	[‘log goal’, ‘country’, ‘currency’, ‘backers_count’, ‘launched_year’, ‘launched_month’, ‘duration_weeks, ‘buzzword_count’]	0.801

- Next Steps
    - Features
        - Description
            - Augmented a sequence RNN (LSTM) to analyze the description and its relationship to the final status of the project
            - Sentiment analysis of description
        - Holiday season
            - Extract common holidays in the countries presented to look for connection of funding with respect to holiday season
            - Will take the week before and after the holiday as measuring metrics
        - Currency appreciation / depreciation
            - Look for connection between currency appreciation / depreciation and the funding outcome
            - Will need additional data input for currency
        - Sports season
            - Look for connection between popular sports season and the funding outcome

    - Algorithms
        - Neural Network
        - SVM (benchmark)