Final Report

CSCI 113i – J

Crowdfunding Proposal
Feasibility Analysis

Business Goal

To optimize the screening process of crowdfunding proposals submitted to online crowdfunding platforms, as a great amount of manual effort is currently needed to screen them.

Given that crowdfunding projects in the Kickstarter platform only have approximately 30% success rate, there is a need to identify factors influencing a proposal's potential success so that it would have a higher chance of being approved. Through the simulations that would be executed using the model, we aim to achieve a potential success rate greater than 50%.

Data Mining Goal

To identify the most important and most impactful features that lead to a greater potential of success within each crowdfunding proposal.

Overview of the Dataset

Name

For this project, we combined two datasets.

- Kickstarter Projects
- 2016 Country Economic Profiles

Sources

Kickstarter Projects: It is an open-access dataset available via Kaggle, uploaded by Mickaël Mouillé.

2016 Country Economic Profiles: The organized dataset of basic economic indicators like population, poverty rate, human development index (HDI), life expectancy (in years), expected and mean schooling years, and the gross national income (GNI) were lifted from a separate Kaggle dataset linked with the analysis of crowdfunding successes for another platform, named Kiva. No information about the creator of this dataset is provided, aside from the fact that this was uploaded by a user named Beluga.

Both datasets can be used under the license CC BY-NC-SA 4.0.

Overview of the Dataset

Description

This dataset provides valuable insights for crowdfunding platform administrators, project creators, and potential backers by offering a data-driven approach to evaluate project viability and optimize the crowdfunding process.

Integrated with economic profiles of the countries where each project originated, we aim to see how these factors may affect a project's success or failure within the Kickstarter Platform.

Methodology

Tools Used

MS Excel was used to do some preliminary observations on the dataset, and to ensure that the dataset can be read as a CSV. Jupyter Notebook was used for the bulk of the exploration, data preprocessing, and other tasks done in this phase.

Data Size

The `kickstart_econ.csv` is 64.3 MB.
It contains 323 750 rows and 24 columns.

Data Quality Assessment

Data Accuracy

In the current context of our dataset which entails crowdfunding project information and country economic profiling, data accuracy must be ensured so that the assessment we will do in the latter parts of the project is reliable. Given that the sources for the project were presented, which are listed above in the *Collection Process*, it gives a minimum guarantee that the information embedded within the dataset is factual and realistic.

Data Completeness

Upon checking Kaggle's criteria on 'Completeness', we saw that both datasets (for Kickstarter Platform and 2016 Country Economic Profiles) only reported 33% ratings, garnering disapproval on the basis of Source/Provenance and Update Frequency.

Data Quality Assessment

Data Consistency

Data consistency is important to maintain so that, again, the reliability of our analysis after modeling would remain intact. Although no major inconsistencies were detected in the whole dataset, there are occasional exceptions wherein a row would have values that are not matched with the column it is referring to.

```
df.iloc[1563]
```

```
ID                                             1009317190
name                                       French Cuisine
category                           A Traditional Experience
main_category                                   Cookbooks
currency                                             Food
user_gender                                        female
deadline_date                                         USD
deadline_time                                         NaN
goal                                       9/8/2014 0:46
launched_date                                    8/3/1937
launched_time                                         NaN
pledged                                    8/9/2014 3:16
state                                                3984
backers                                            failed
country                                                CN
continent                                              AS
usd_pledged                                            US
population                                     1409517397
population_below_poverty_line                         3.3
hdi                                              0.737681
life_expectancy                                    75.963
expected_years_of_schooling                      13.53575
mean_years_of_schooling                           7.64184
gni                                           13345.47746
Name: 1563, dtype: object
```

Label and Label Description

Label	DataType	Description
ID	integer	Unique ID for each crowdfunding project.
name	string	Name of project.
category	string	Fine classification of the project nature.
main_category	string	General classification of the project nature.
currency	string	Currency of the crowdfunding project.
user_gender	string	Binary data. Either 'male' or 'female'.
deadline_date	date	Deadline date.
deadline_time	time	Deadline time at the specified date.
goal	integer	Target amount to be raised by the project.
launched_date	date	Launch date of project.
launched_time	time	Launch time at the specified date.
pledged	float	How much the project currently has raised.

state	string	Current state of project (can be success, failed, canceled, suspended, and live; but dropped the latter three to focus on success/fail rates).
backers	integer	Number of supporters.
country	string	Country of origin.
continent	string	Continent of country.
usd_pledged	float	Total amount pledged in USD.
population	integer	Population of the country.
population_below_poverty_line	float	Poverty incidence rate of the country.
hdi	float	HDI rating of the country.
life_expectancy	float	HDI metric.
expected_years_of_schooling	float	HDI metric.
mean_years_of_schooling	float	HDI metric.
gni	float	HDI metric.

Models Used

Random Forest Classifier

This is the primary supervised machine learning method used for this project. It is a commonly-used model because of its robustness, use for feature importance, and ensemble method.

XGBoost

One of the models touted for its high accuracy and usefulness for datasets with complex relationships among its features.

LightGBM, Gradient Boosting

Known for their efficiency and speed in prediction, making it suitable for large-scale datasets.

Results

Table 1. Evaluation metrics of different supervised machine learning models used in the project.

Model	Accuracy	Precision	Recall	F1-score
RandomForest	**0.7289**	0.6310	**0.5224**	**0.5716**
GradientBoost	0.6798	**0.6367**	0.4715	0.5418
LightGBM	0.6849	0.6377	0.4979	0.5593
XGBoost	0.6955	0.6005	0.4647	0.5386

Results

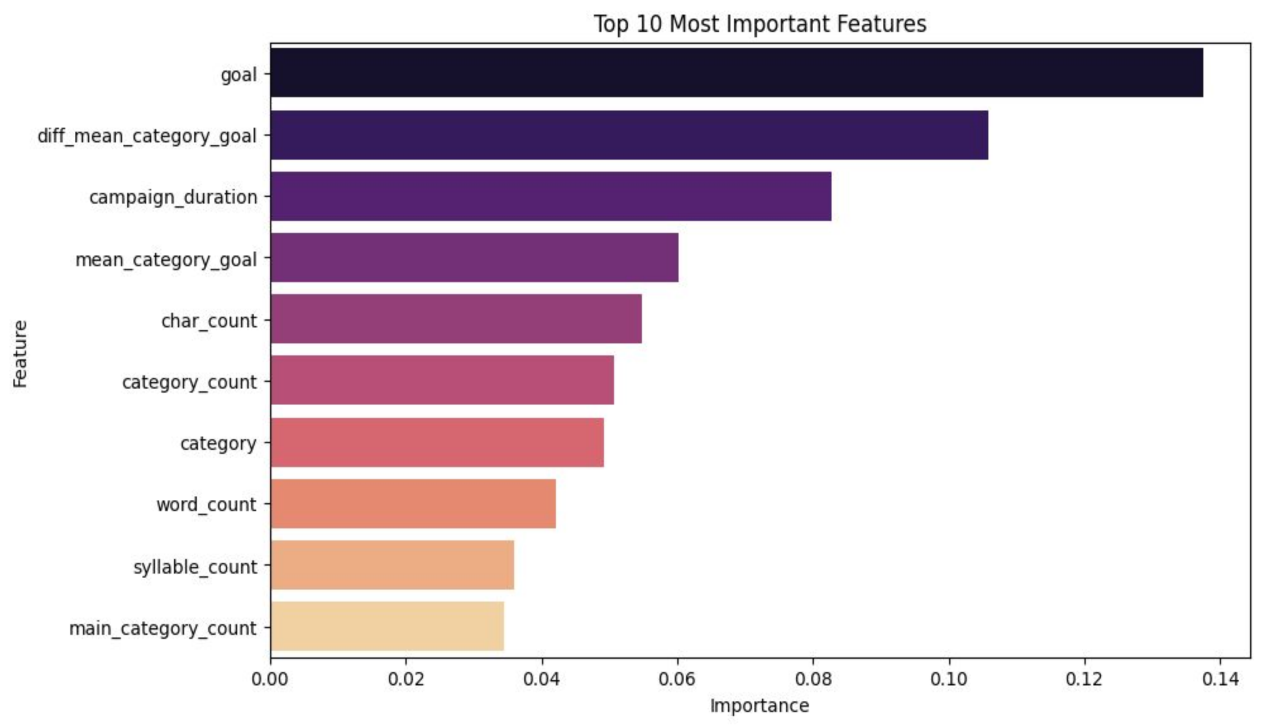

Figure 1. The ten most important features from all existing crowdfunding projects in the dataset.

Results

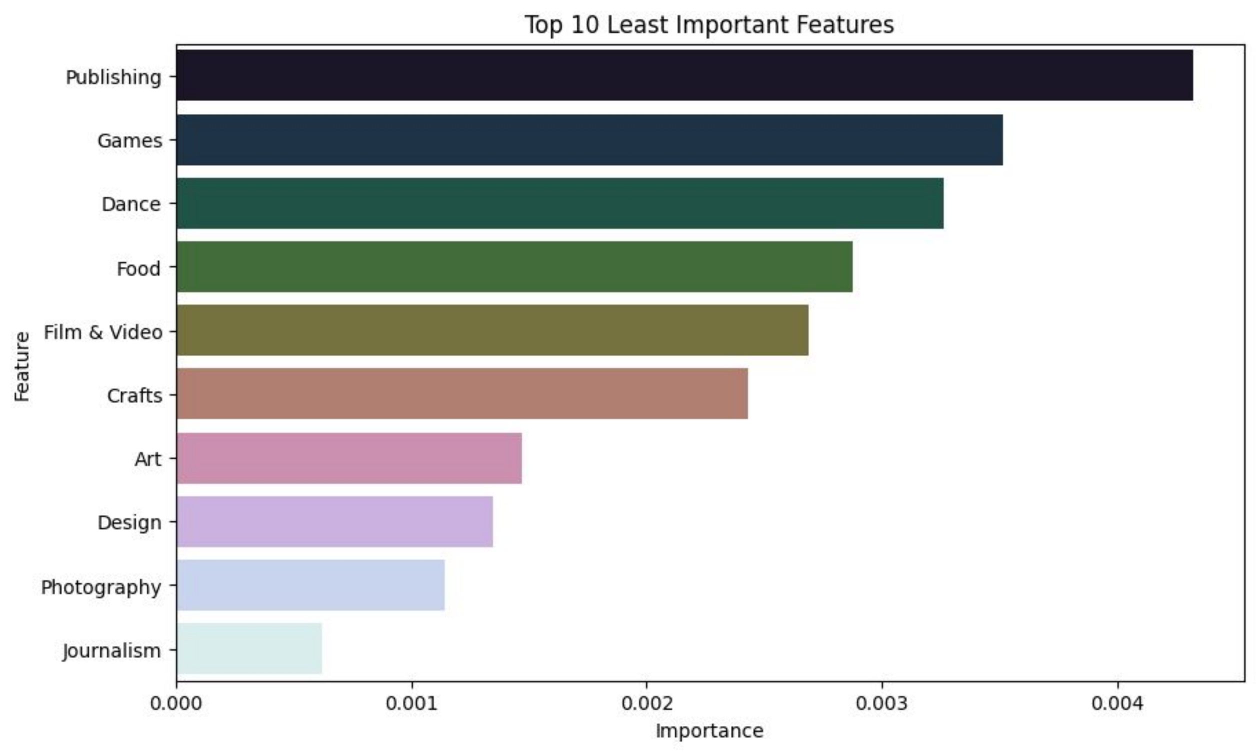

Figure 2. The ten least important features from all existing crowdfunding projects in the dataset.

Results

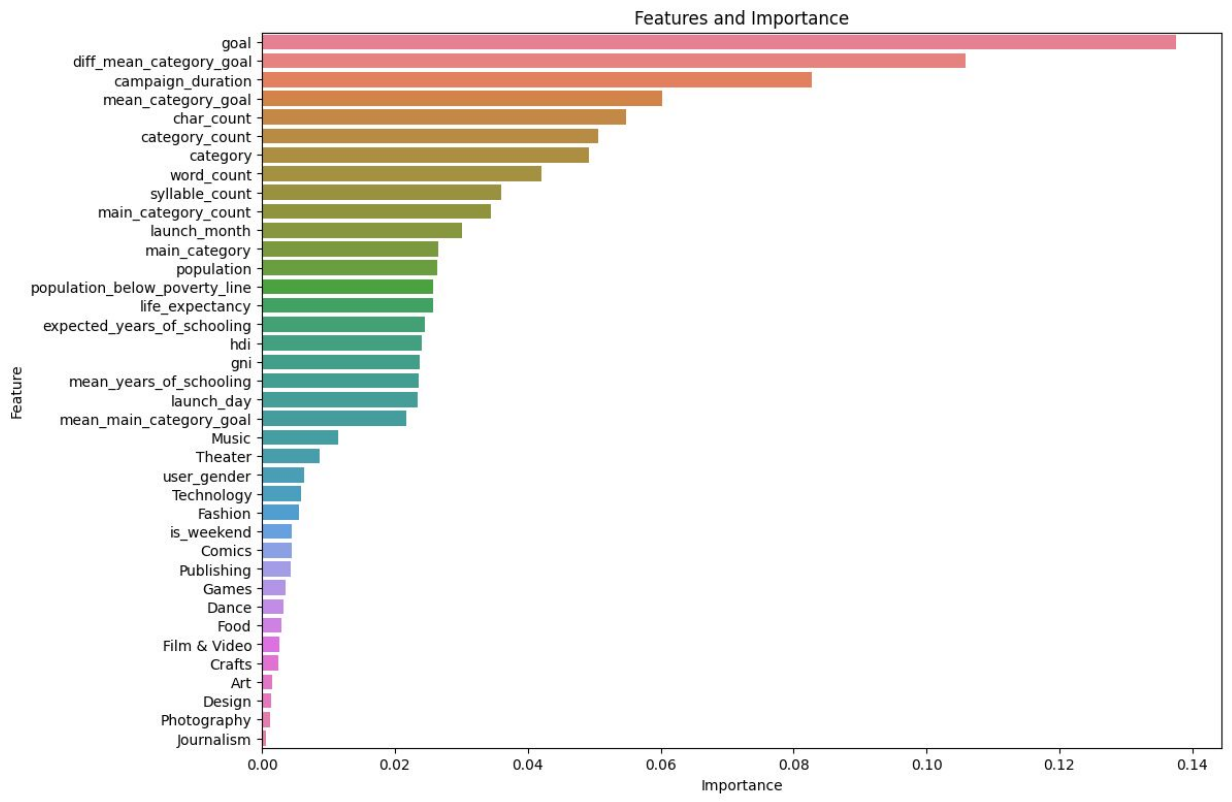

Figure 3. A bar graph reflecting all features and their importance values. Note that variables related to a project category, as well as economic factors, were deemed to be (at the maximum) moderately important.

Degree Centrality

Definition 1. Degree centrality is the total number of connections linked to a vertex. It can be thought of as a kind of popularity measure, but a crude one that does not recognize a difference between quantity and quality.

This kind of visualization is useful for identifying key factors and understanding the complex interplay between different variables in a multifaceted dataset.

Results

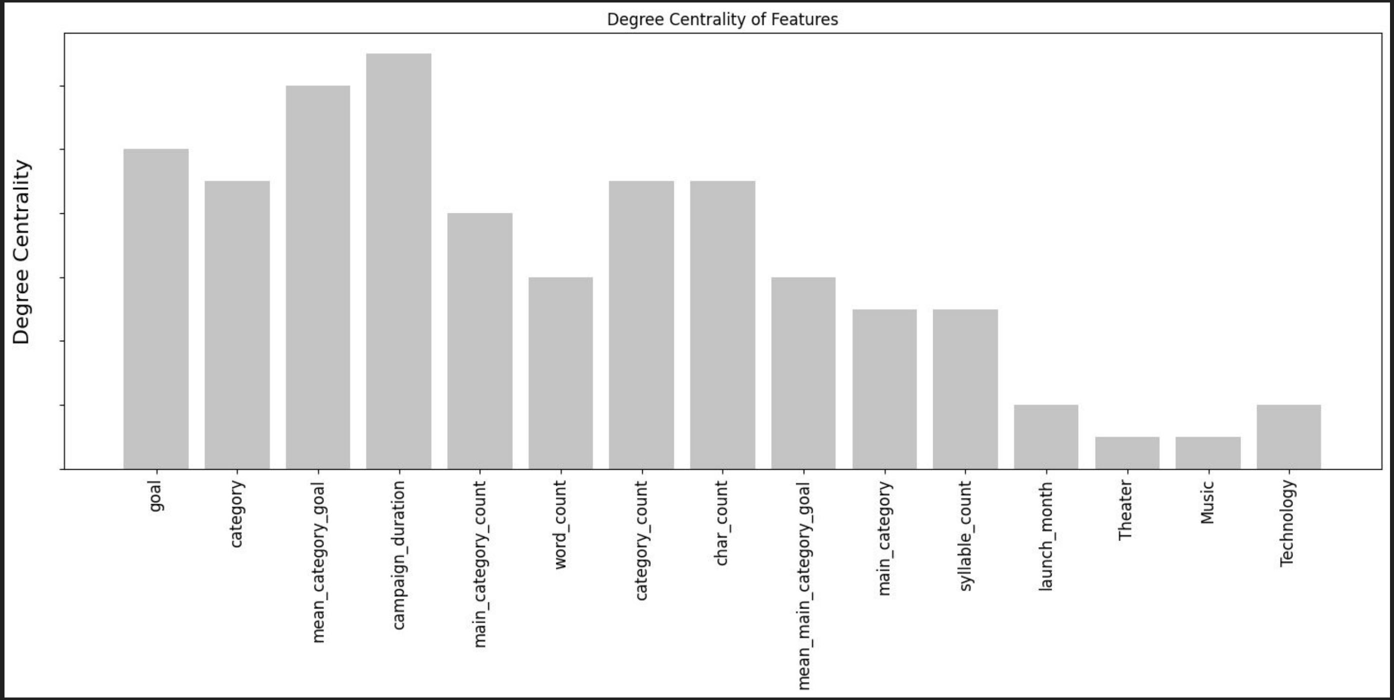

Figure 4. The degree centrality graph of selected features.

Results

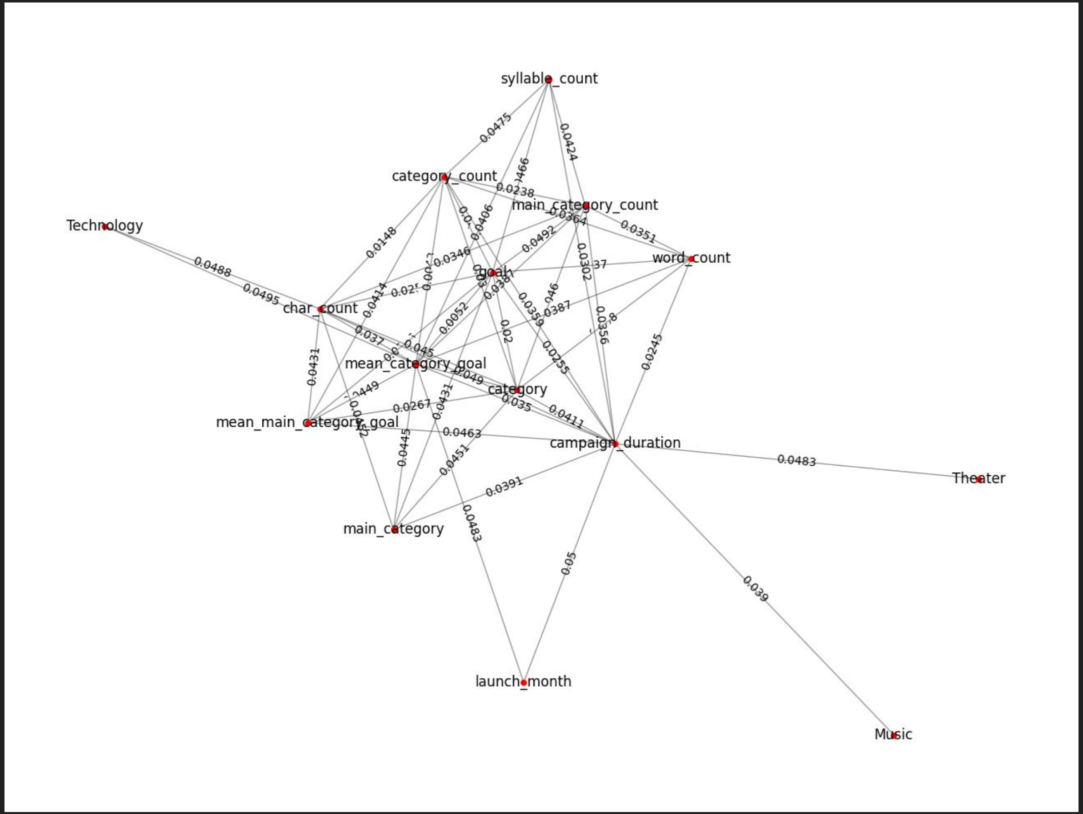

Figure 5. Network graph representing the relationships and correlations between various features in the dataset.

Permutation Importance

Definition 2. Permutation feature importance measures the contribution of each feature to a fitted model's statistical performance on a given tabular dataset.

This technique is particularly useful for non-linear or opaque estimators, and involves randomly shuffling the values of a single feature and observing the resulting degradation of the model's score. By breaking the relationship between the feature and the target, we determine how much the model relies on such particular feature.

Results

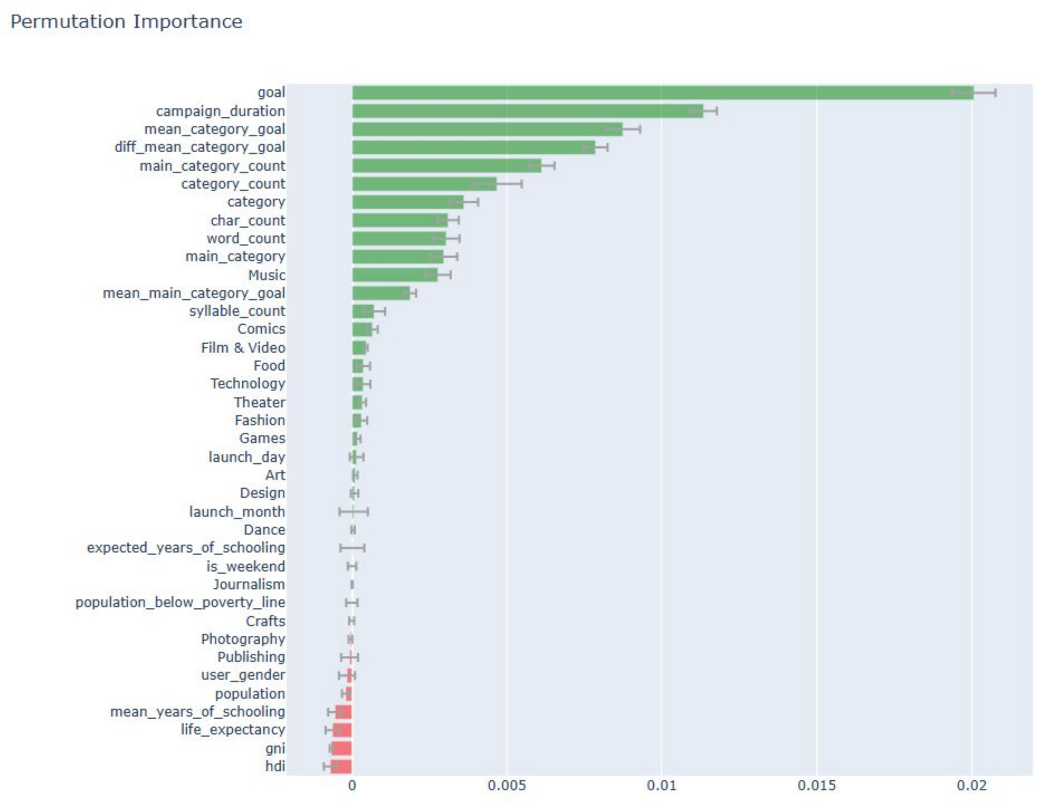

Figure 6. Permutation importances of the extracted features from crowdfunding projects.

Results

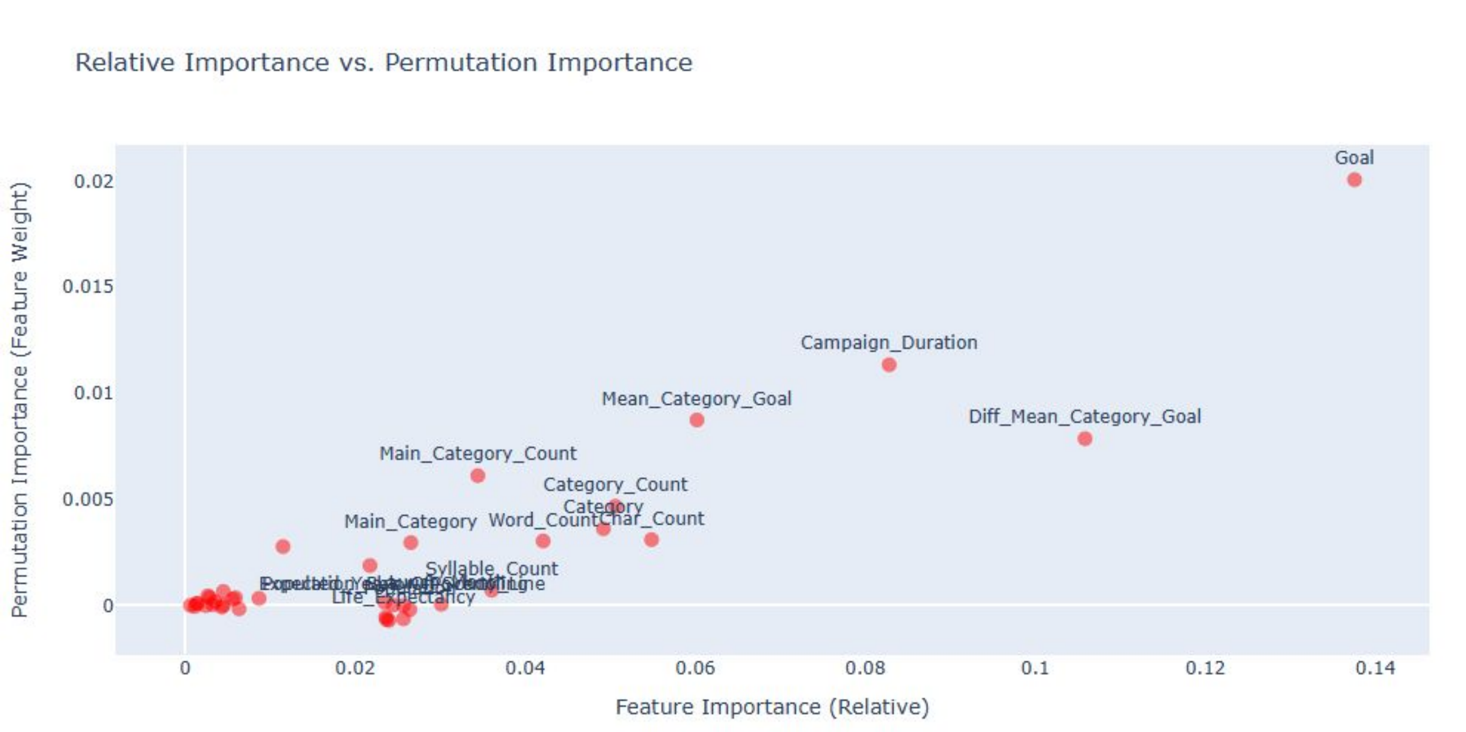

Figure 7. Relative importance-vs-Permutation importance graph.

Results

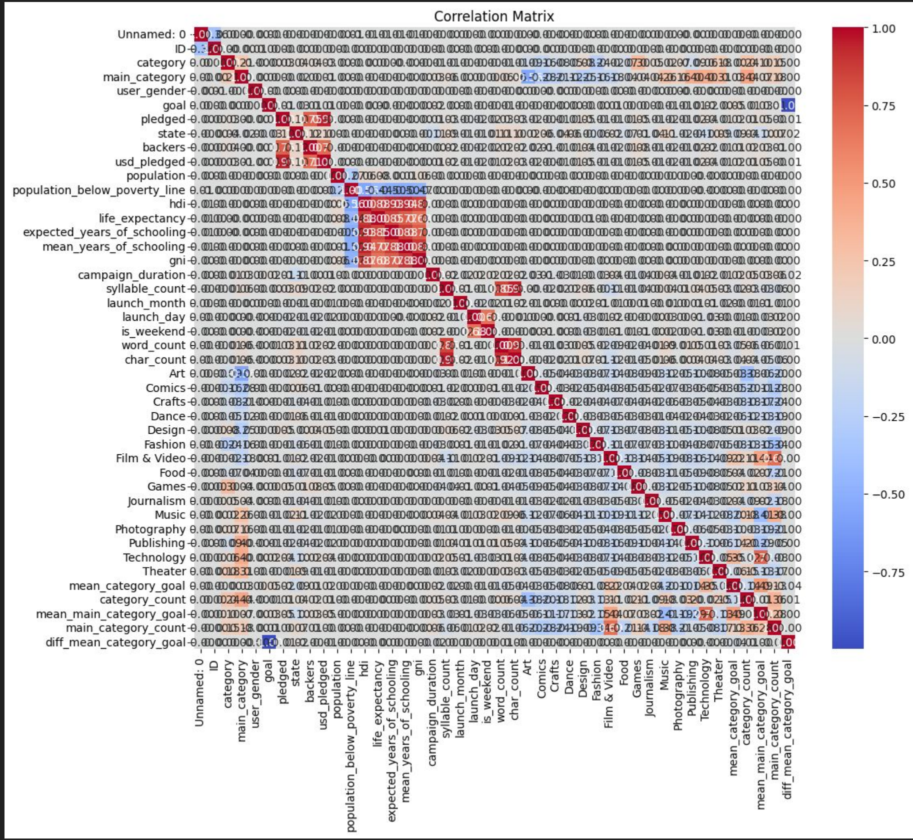

Figure 8. Correlation matrix of all variables in the dataset. Most are really uncorrelated, with few exceptions.

Partial Dependency

Definition 3. Partial dependence plots (PDP) show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the 'complement' features).

Intuitively, we can interpret the partial dependence as the expected target response as a function of the input features of interest.

Results

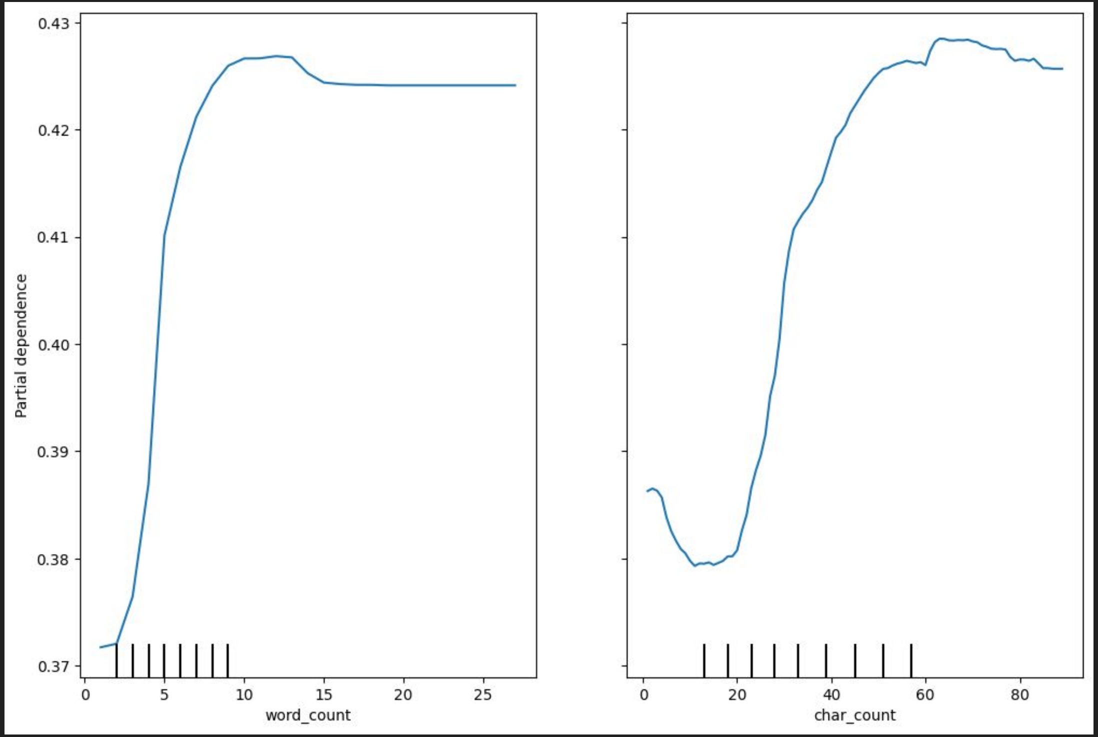

Figure 9. Partial dependency graph for `word_count` and `char_count`.

Results

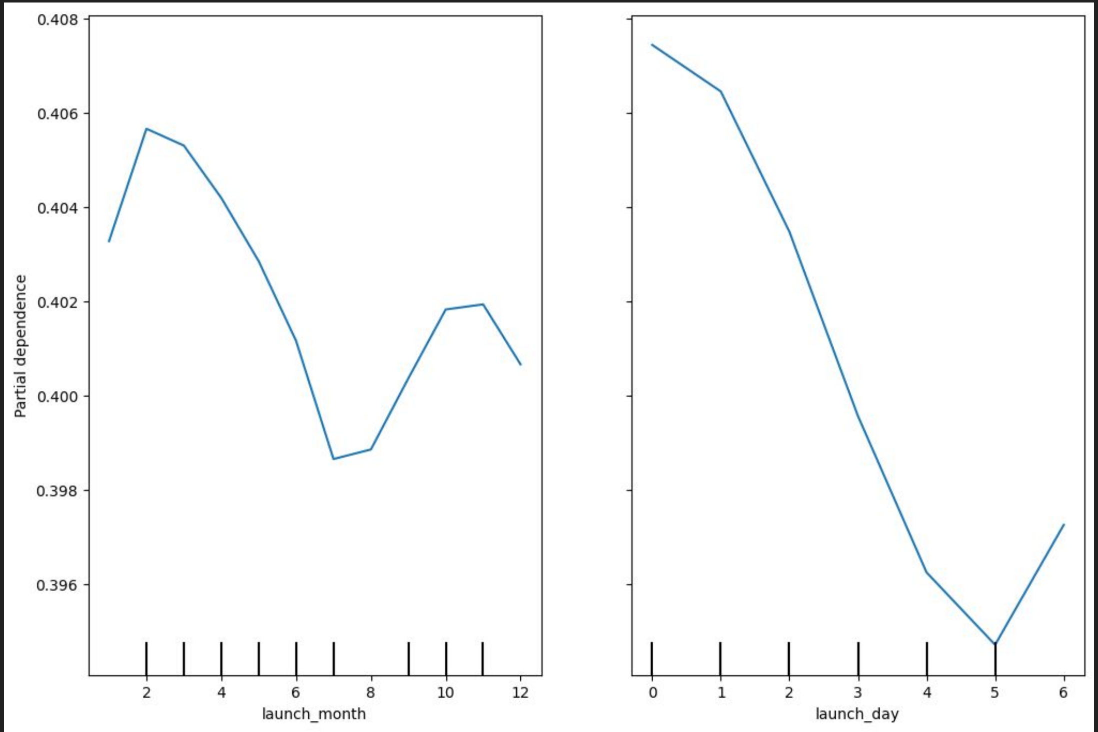

Figure 10. Partial dependency graph for `launch_month` and `launch_day`.

Results

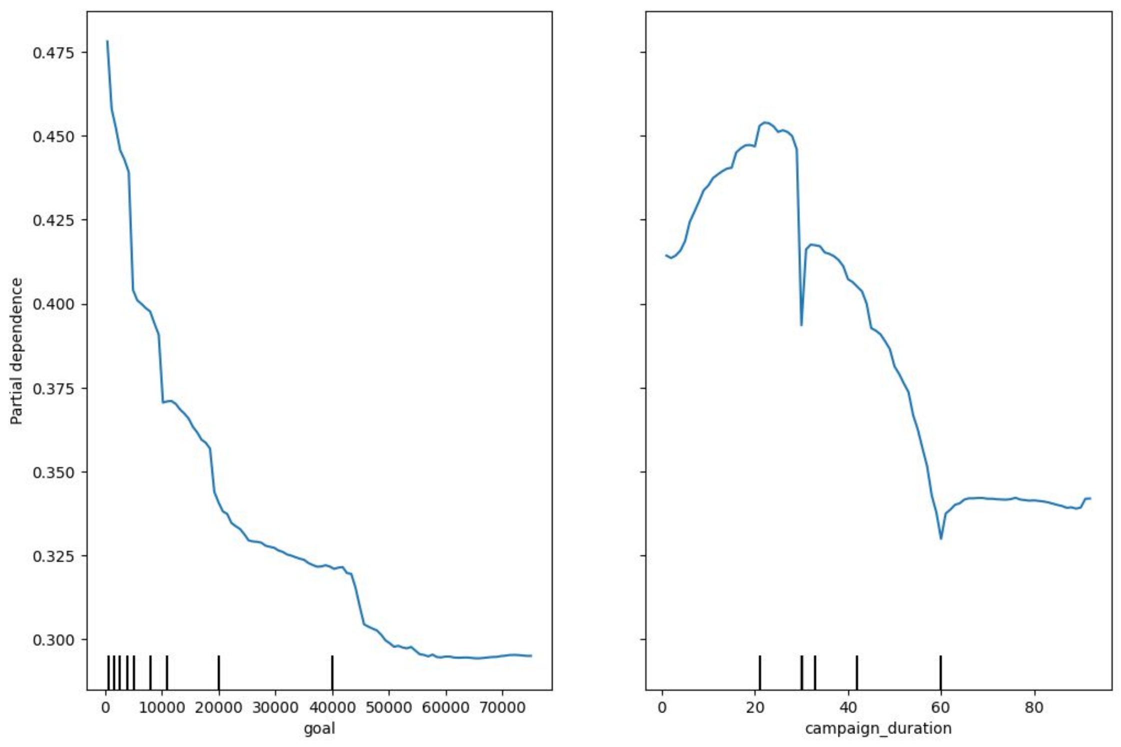

Figure 10. Partial dependency graph for `goal` and `campaign_duration`.

Evaluation

- Using the test data, we were able to create predictions regarding the success rate of the crowdfunding projects within this partition.
    - Out of 69 216 entries, only 20 163 were deemed **successful** by the random forest classifier.
    - This translates to roughly **29.13% success rate.**
    - Given that real-world data is benchmarked at around 30%, the model reflected the general trend.
- With its accuracy score seeking to be improved, we recommend devising other methods to potentially improve viability of crowdfunding success.
- Despite this, one of the goals of this exploration was also realized: the determination of the most important features within a certain crowdfunding project.

Thank you.