ISyE 6740 – Summer 2025
Final Report

Team Member Names: Phat Dat Tieu (ptieu6@gatech.edu)

Project Title: What makes a successful Kickstarter campaign?

Introduction

Since its launch in 2009, Kickstarter has become one of the most well-known crowdfunding platforms, helping creators to release products across a wide range of categories, including art, design, technology, games, and more. With over hundreds of thousands of campaigns launched and billions of dollars raised to date, Kickstarter has established itself as a go-to site for funding creative ideas into realities. However, it's no guarantee that one can achieve their goal on Kickstarter. In fact, a significant proportion of campaigns fail to meet their funding target, resulting in zero funding due to the platform's all-or-nothing model.

Many creators have attempted to determine the "formula" for a successful campaign — from optimal funding goals, engaging videos to the careful launch timing. While anecdotal observations and informal strategy guides exist, few researches have systematically analyzed the comprehensive mix of campaign attributes that greatly affect the success using large-scale data.

This proposal suggests a data-driven study to investigate which features are most effectively associated with successful Kickstarter campaigns. The findings can guide creators and platform developers in prioritizing key factors during campaign planning and optimization.

Objective

The primary objective of this project is to identify and analyze key features linked with accomplished Kickstarter campaigns. Rather than predicting outcomes, the goal is to understand which factors matter most and why, through feature selection and interpretability-focused techniques. Some specific objectives include:

(1) Perform exploratory data analysis (EDA) to understand patterns of success across categories, time periods, and other campaign metadata.
(2) Visualize and interpret relationships between key features and campaign outcomes.
(3) Select and rank important features that influence campaign success, using both statistical tests and model-based importance analysis.
(4) Evaluate robustness of selected features.

Scope

We will analyze the data set of nearly 260,000 campaigns with 40 original variables, from 2009 to 2025 June, taken from latest dataset on Web Robots website. Disclaimer: not every

campagins data ever posted on Kickstarter will be collected. One can use more noble approach to gather data from well-established sources like Kaggle or scientific researches. Unfortunately, Kaggle most popular dataset last updated 7 years ago, and IEEE papers need authors' permission to work with.

Back Web Robots dataset, each campaign is classified as "failed", "successful" or "on going". This analysis will exclude "on going" products and geographic analysis, as America-based products hold nearly 63% of whole population, while more than half of listed countries only accounts for less than 1% each. Additionally, rare campaign categories representing fewer than 100 items will be excluded to reduce noise and ensure adequate sample sizes.

We will not discuss the following topics: external marketing efforts, social media coverage, effects of visual contents (photo, videos), full-text natural language processing of long descriptions, real-time user interaction such as backer behavior during live campaigns, number of questions asked, etc. Instead, the scope will focus on features available at or just before launch, to focus on what creators can influence during campaign preparation.

Methodology & Result

1. Data Preprocessing

1.1 Drop duplicates

Original data set is collected from Web Robot website, which collects data every month. Some projects having a long duration might be recorded more than once. We only keep the last record as it would be the latest update. All campaigns have their own id, so it's very easy to drop duplicates.

1.2 Time variables

1.2.1 Convert time data from unix to normal format

'created_at', 'launched_at', 'deadline' are 3 main time columns we are using, but they are in unix format, need to convert to normal datetime format for easier processing.

An interesting point is that there are a lot of dummy sample projects created from 1970s, which are irrelevant to our analysis, and those projects will be removed.

1.2.2 Adding more datetime variables

The main focus is 'launched_at', so we will extract as much information as possible. More variables extracted are:

- launched year
- launched month
- launched week
- launched weekday: Monday, Tuesday, Wednesday, ...
- is launched weekend: whether launched weekday is in weekend
- pre-launched time: time difference between `launched_at' and `created_at', in days
- duration: time difference between 'deadline' and 'launched_at', in days

1.3 State of a campaign

- Failed - projects that haven't reached the goal withing the deadline. It's interesting to note that the the time for completing a Kickstarter project is limited from 1 to 60 days.
- Live – active projects
- Successful – projects that have reached the amount of money they pledged
- Canceled – projects that were canceled by their creator. The reasons can be numeral, and the creator cannot relaunch the same project.
- Suspended – projects that have violated the terms of the platform. Main cases include providing false information, artificially increasing backers, the creator is presenting someone else's work as their own or is offering purchased items, claiming to have made them.

For simplicity, canceled and suspended projects are considered failed, and we only accounts for those whose results are already defined. Live and pre-launched campaigns are not in the scope of our work.

1.4 Category

More numerical variables are collected for main & sub categories, counts & average goal for each category

- main_category_count
- main_category_goal_mean
- sub_category_count
- sub_category_goal_mean

Onehot Encoding for all main categories. For example, those projects from Technology will have 'Technology' column value as 1, and those aren't from Technology will have 0.

1.5 Title, Description, Videos

The scope is not about NLP or LLM where the content of text is important, we just want to see how long / how many words they have.

- num_words_name: number of words in title
- num_chars_name: number of characters in title
- num_words_blurb: number of words in campaign description
- num_chars_blurb: number of characters in description
- has_video: does campaign have a video (1/0)

1.6 Goal amount

All goat amounts are converted to USD for consistency, and we also add log-transformed column to see how the distribution looks like

1.7 Columns to be used

Note that these columns are used for our classification model to work on feature extractions. Exploratory Data Analysis section would use all columns necessary

1.7.1 Categorical variables

'launched_month', 'launched_weekday', 'launched_is_weekend','quarter','main_category', 'sub_category','has_video'

1.7.2 Numerical variables

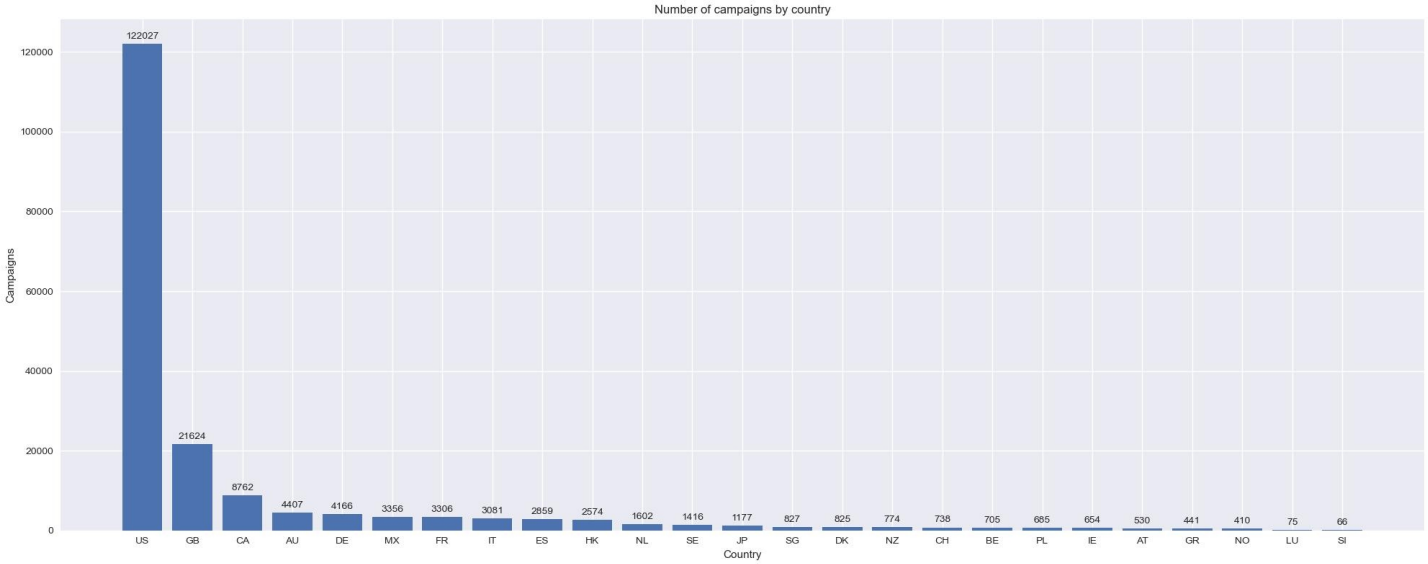
'pre_launch_time','duration','mean_main_category_goal', 'main_category_count','mean_sub_category_goal', 'sub_category_count','num_words_name', 'num_chars_name', 'num_words_blurb', 'num_chars_blurb', 'usd_goal', 'log_usd_goal'

2. Exploratory Data Analysis (EDA)

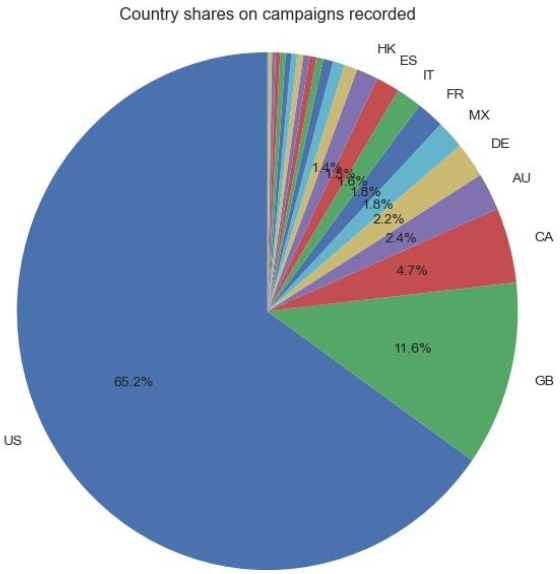
After data preprocessing, we have 187087 campaigns to analyze.

2.1 Checking country campaigns

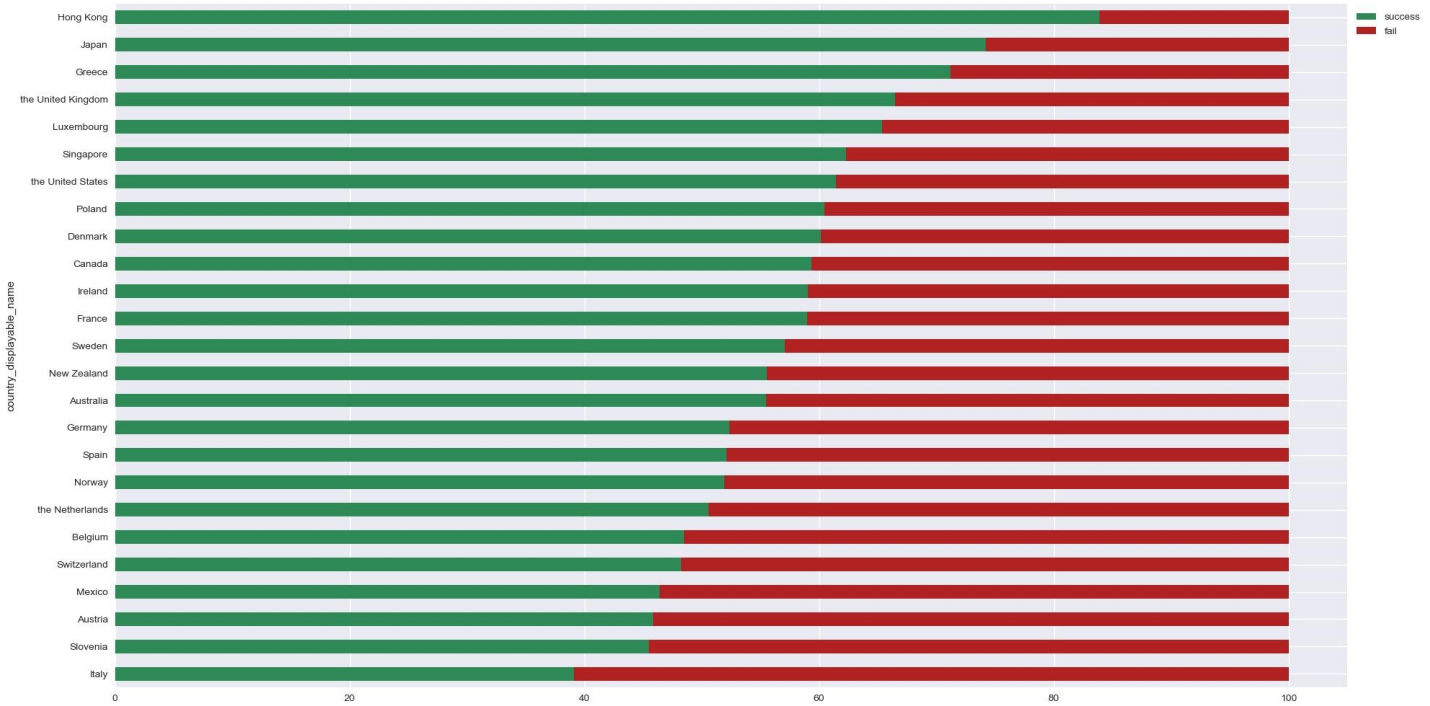
2.1.2 Number of campaigns by country



United States has the most projects, with 123,027 (65%) campaigns, almost 6 times more than the second most country / region, Great Britain. Nearly half of countries listed have less than 1000 projects. Due to this imbalance, country is not considered a major factor in our feature extraction model.



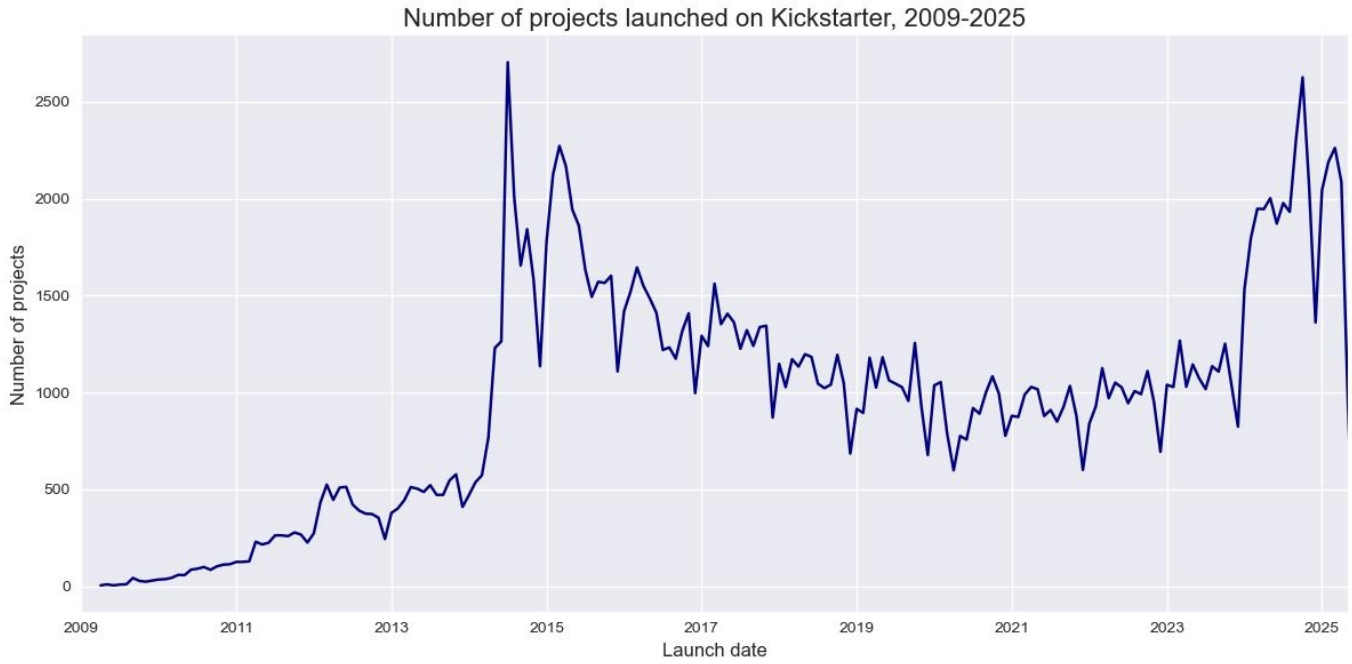
2.1.2 Country success rate



More campaigns doesn't directly mean that country would have better successful rate. US-based campaigns have around 60% successful rate, while the highest rate countries are HongKong and Japan, sitting at 83% and 74%. Kickstarter seems to be not very popular in some European countries, where 4 out of 5 lowest rate is from EU, with less than 50% successful rate.

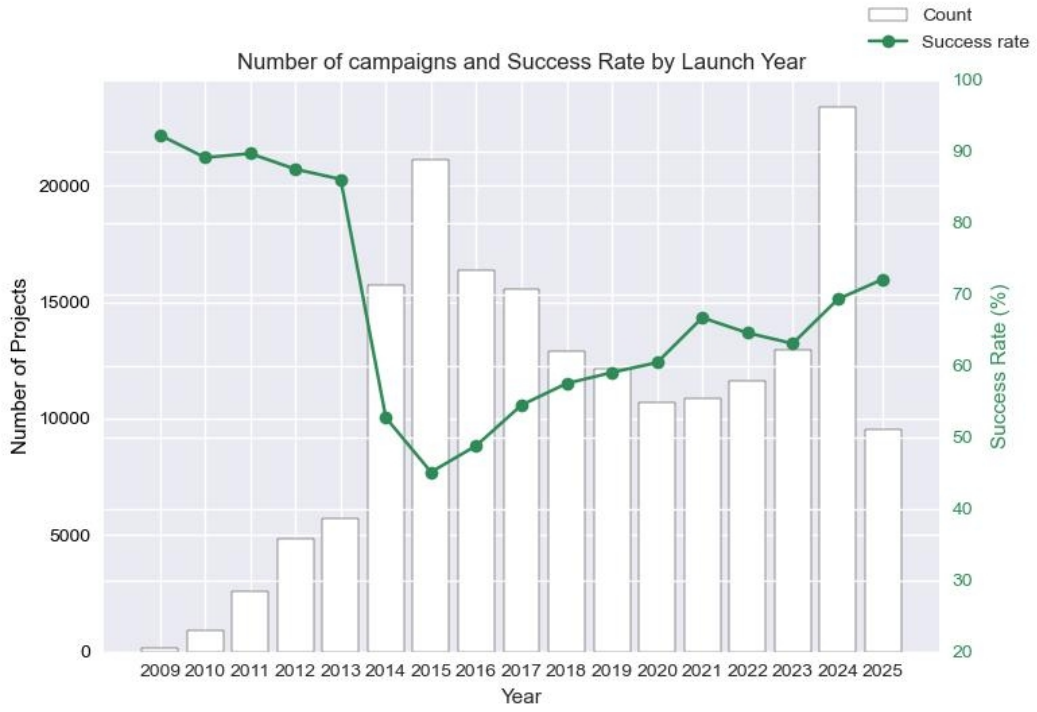
2.2 Checking campaigns by time

2.2.1 Number of campaigns by year



2014-2015 have a lot of campaigns recorded. Not sure if it's just because website manage to collect unusualy high amount of campaigns or it just happens. During COVID time 2020-2021, productions drops a little bit. After that Kickstarter starts to gain back the numbers, and spikes in 2024. So far in the first quarter of 2025, the project amount shows promising signs to overcome 2024. Last end of the graph doesn't show the correct number, as all ongoing projects are removed.

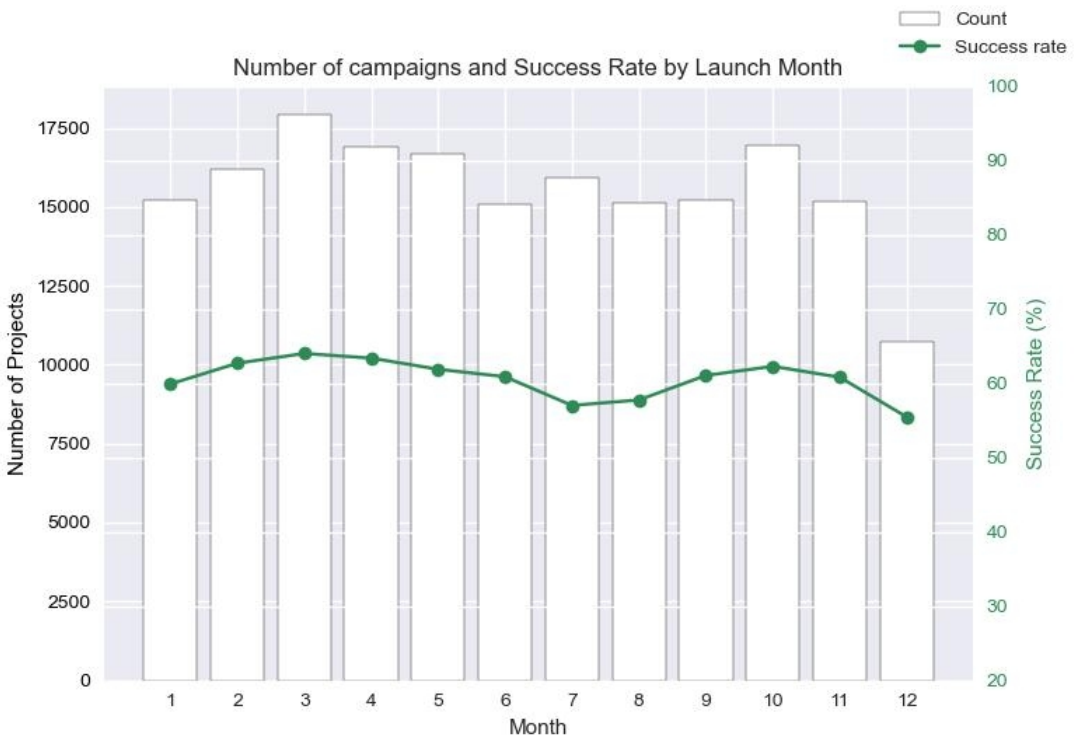
2.2.2 Success rate over the year



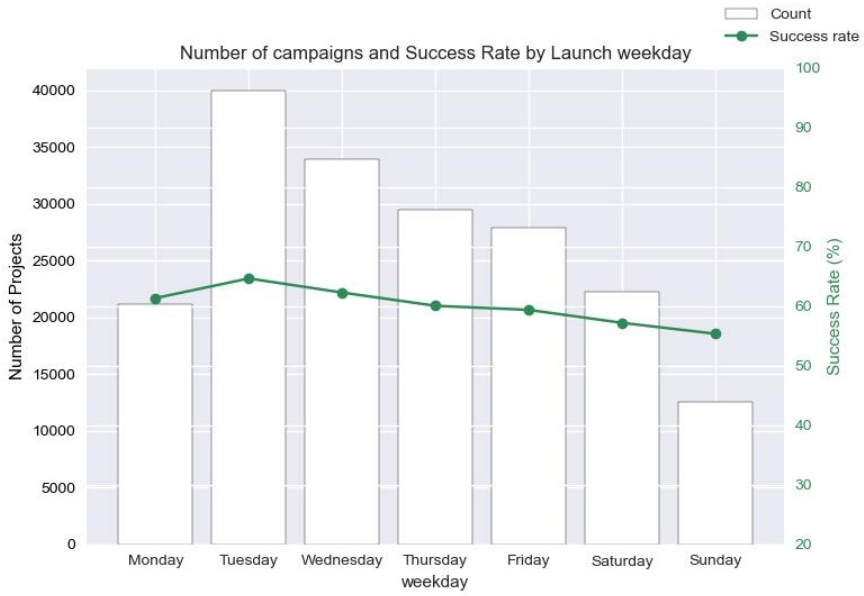
At the start of Kickstarter, the number of projects was low, but the success rate was really high, more than 85%, might due to several reasons, the quality of product was better, the price was lower compared to market, online shopping was not as competitive as in recent time, etc. 2014-2015 marked the significant progress on sheer quantity of campaigns, after enterperneurs realized this platform is hidden gem. Quadruple projects on Kickstarter in 2015, compared to 2013, which caused the success rate plummeted to 43%. Much more projects, but not much more buyers to back it up. In the next 8 years, number of campaigns reduced gradually, especially in COVID period. Last year 2024, Kickstarter returned to its prime where more than 22000 projects were launched, and success rate was reasonably high with 70%.

2.2.3 Number of campaigns and success rate by months

Most of the months have similar number of campaigns, with some small peaks in March and October. The holiday winter season December surprisingly has the lowest number by a far margin, 30% lower than the second last. Success rate also doesn't change much across all months, just a little bit lower in december. Seems like people enjoy other platforms where sales are more common during this time.



2.2.4 Number of campaigns by weekday



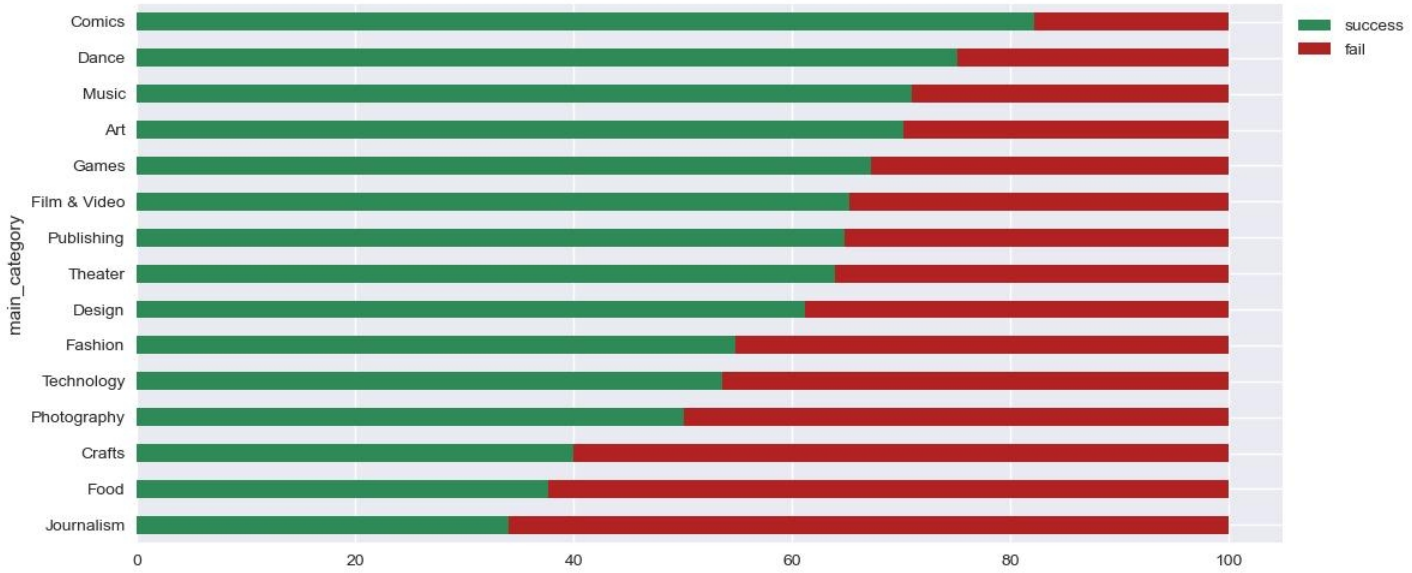
Another surprising charts where the days off have the inferior number, while mid week have much superior amounts of campaigns launched. Not sure if it's creator intention to buffer a few days before weekend start, so that their project would be more likely to pop-up on the front page of its category, but it's very interesting to see such low number on weekends. Success rate seems to be similar through the week.

2.2.5 Pledged and Goal amount by year

2009-2013 is a very healthy start for Kickstarter, the average goal was always increasing, and average pledge was also more than the goal. Things began to get much tougher from 2014 until 2024. Latest years are not collected fully, and we don't include ongoing projects, which might make the pledge amount not as higher than goal as it is in the graph.

2.3 Checking campaigns by category

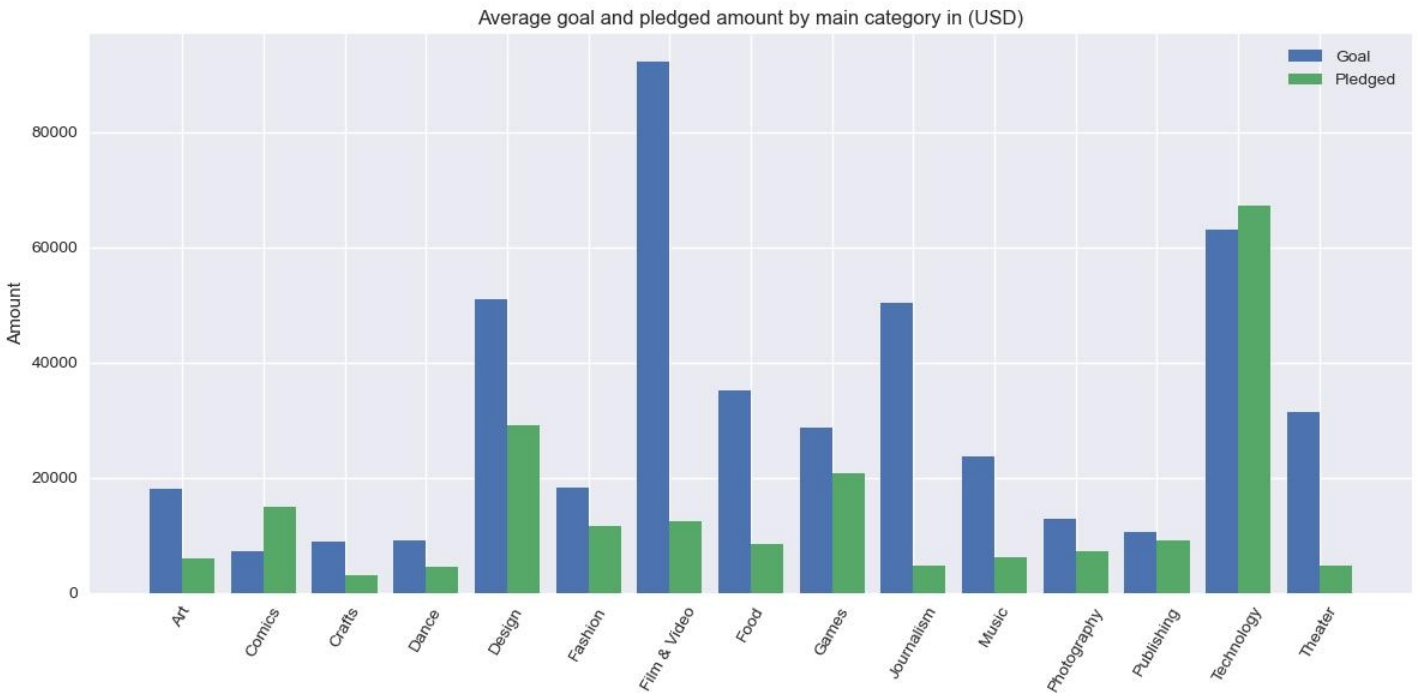
2.3.1 Success rate by category



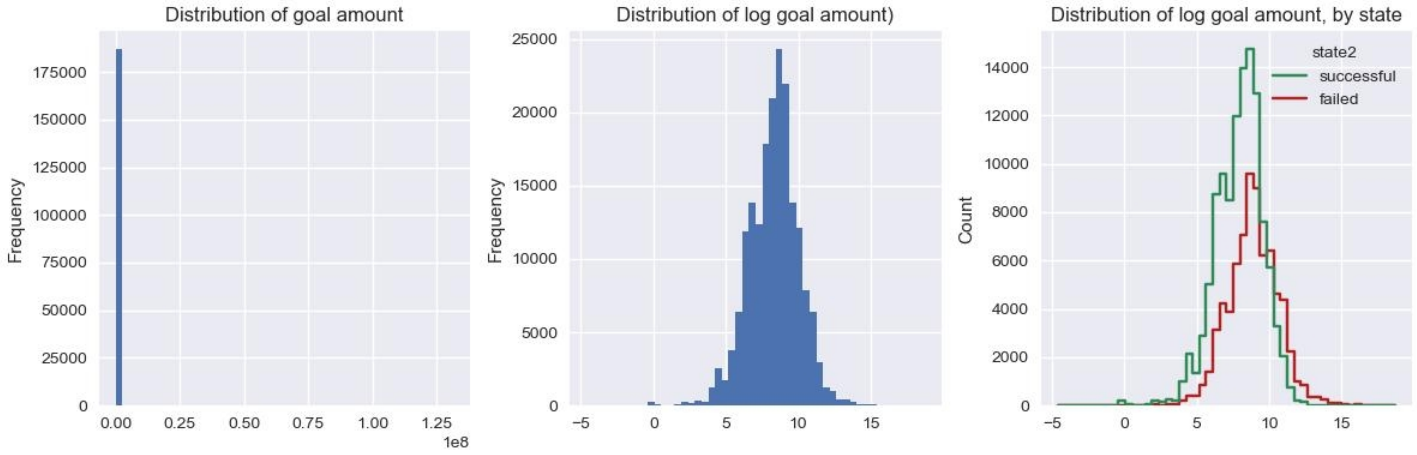
Hobby-related categories have the highest success rate, while long reading books, news, documents are not the crowd's favourite. Food-related products also not very successful, as people are more likely to purchase them in physical stores or trustworthy, proven sources.

2.3.2 Pledged and goal amount by category

Most categories have average goal higher than average pledged amount, the worst among them is 'Film & Video'. Only 'Comics' and 'Technology' have average pledged amount higher than average goal, where Technology also has the second highest goal as well.

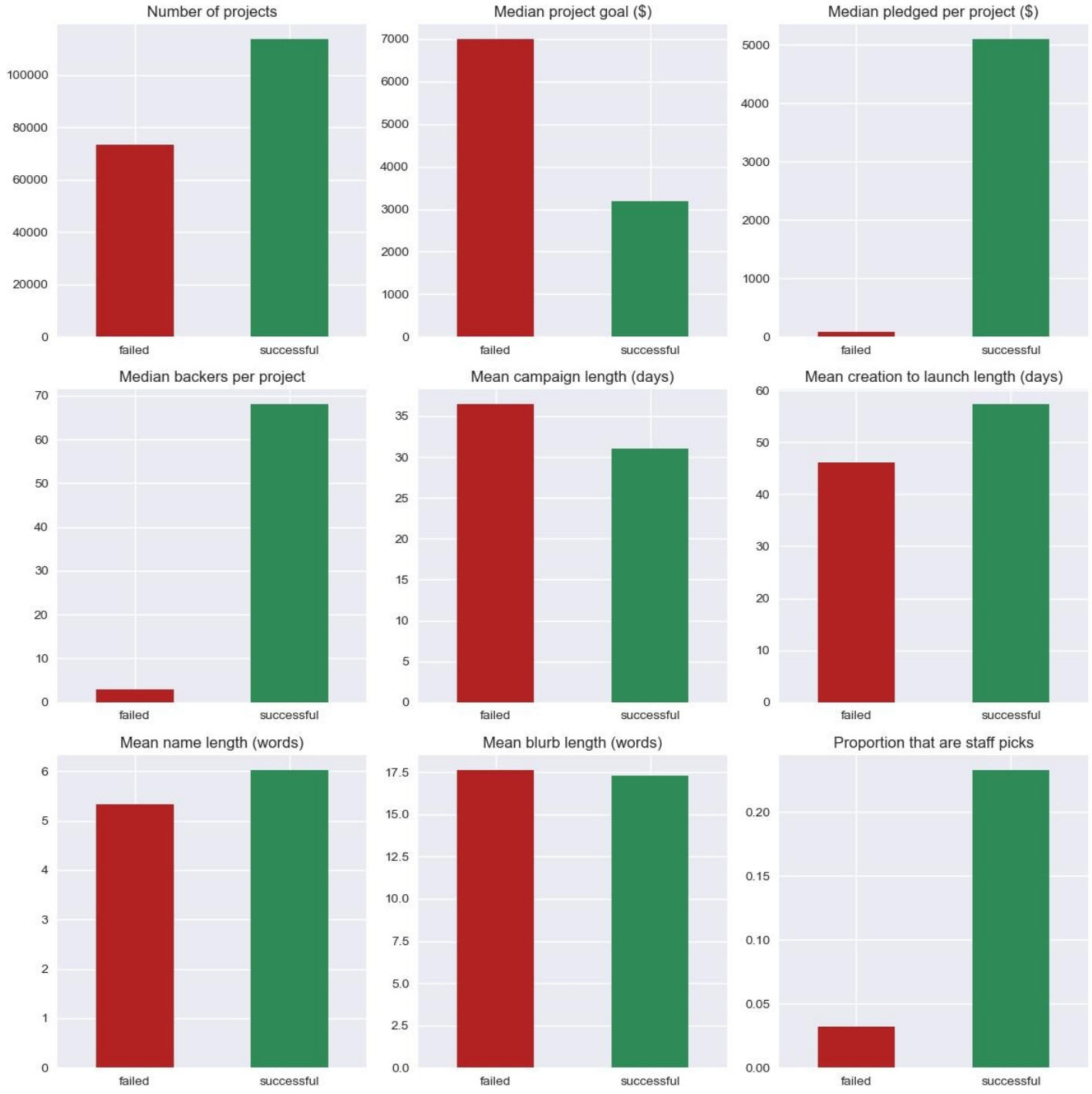


2.4 Goal distributions

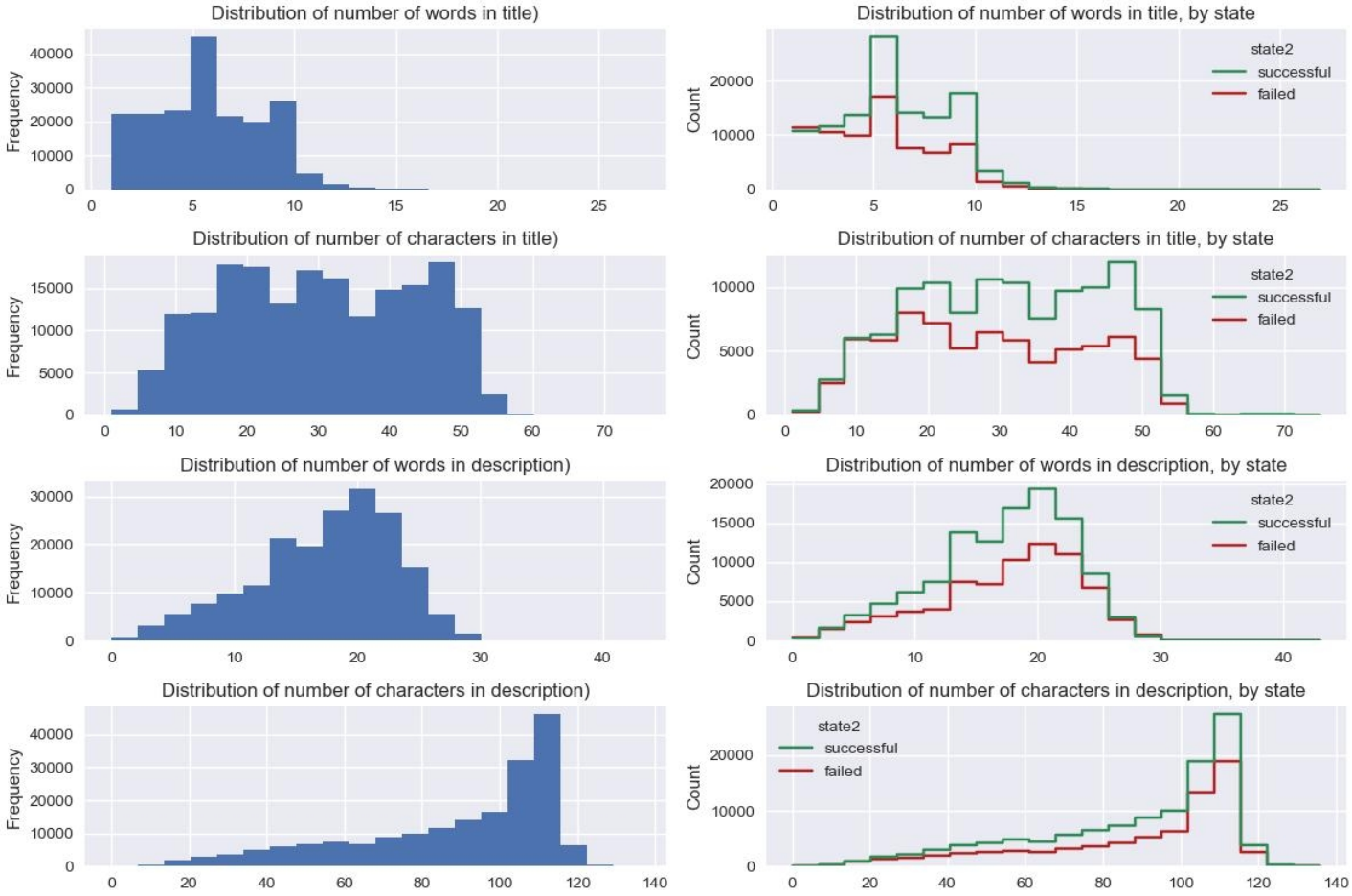


Raw goal distribution is heavily skewed by big projects. When taking log-transform on goal, it becomes normal distribution, look pretty good for analysis. Successful projects also have less goal than failed ones, which makes sense, considering Kickstarter is all-or-nothing funding, only when creators receive enough pledge then they can only earn the money.

2.5 Succesful rate among other metrics



2.6 Basic Text data analysis



The image above shows distributions of words and length of title / description for all / successful / failed recorded projects. A common trend here is that more words, higher chance of success, just don't make it too long.

3. Feature Selection and Ranking

3.1 Chi-squared test for categorical data

Categorical variables are all transformed by one-hot encoding to feed into scikit-learn chi2, especially all main categories have their own columns. For example, those projects from Comics will have 'Comics' column value as 1, and those aren't from Comics will have 0. Top 10 categorical factors are:

1. 'main_category'
2. 'sub_category'
3. 'Art'
4. 'Comics'

5. 'Crafts'
6. 'Food'
7. 'Journalism'
8. 'Music',
9. ‘has_video'
10. ‘launched_weekday'

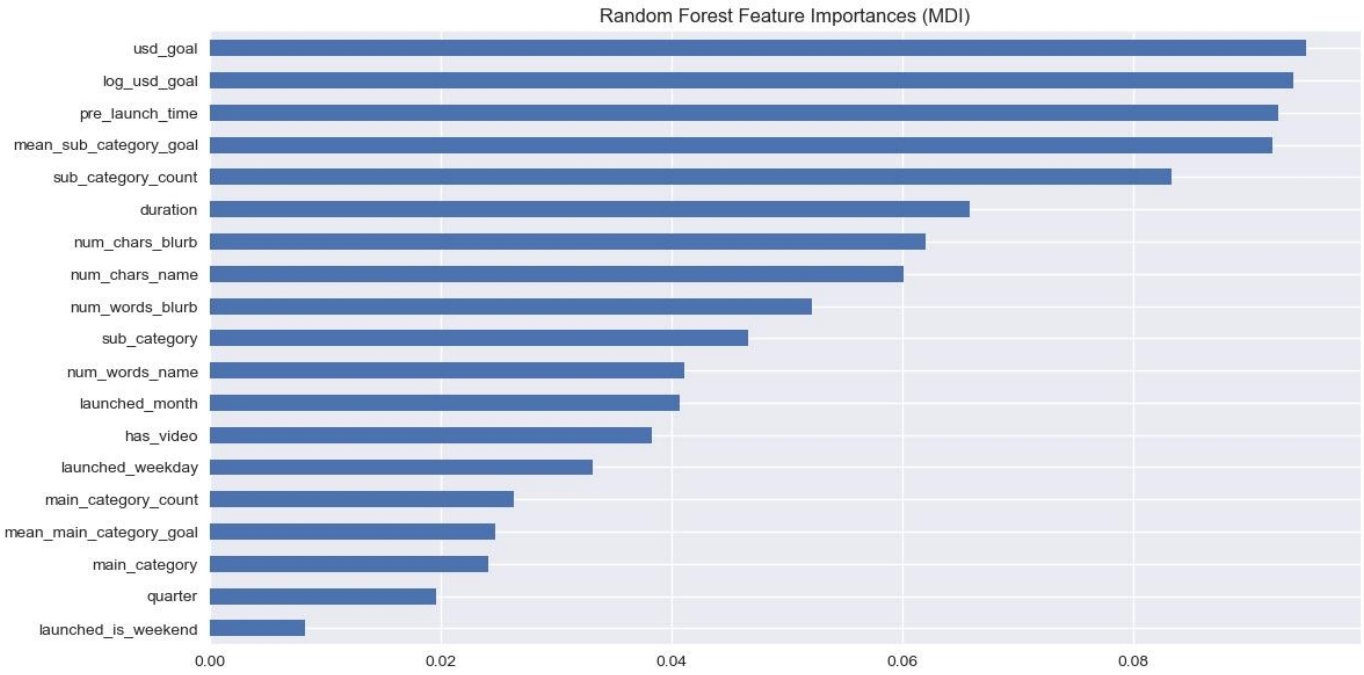
3.2 Tree’s Feature Importance from Mean Decrease in Impurity (MDI)

Data is devided into training (85%) and test set (15%), and feed into Random Forest (RF)

3.2.1 Basic tree model, default scikit-learn setting

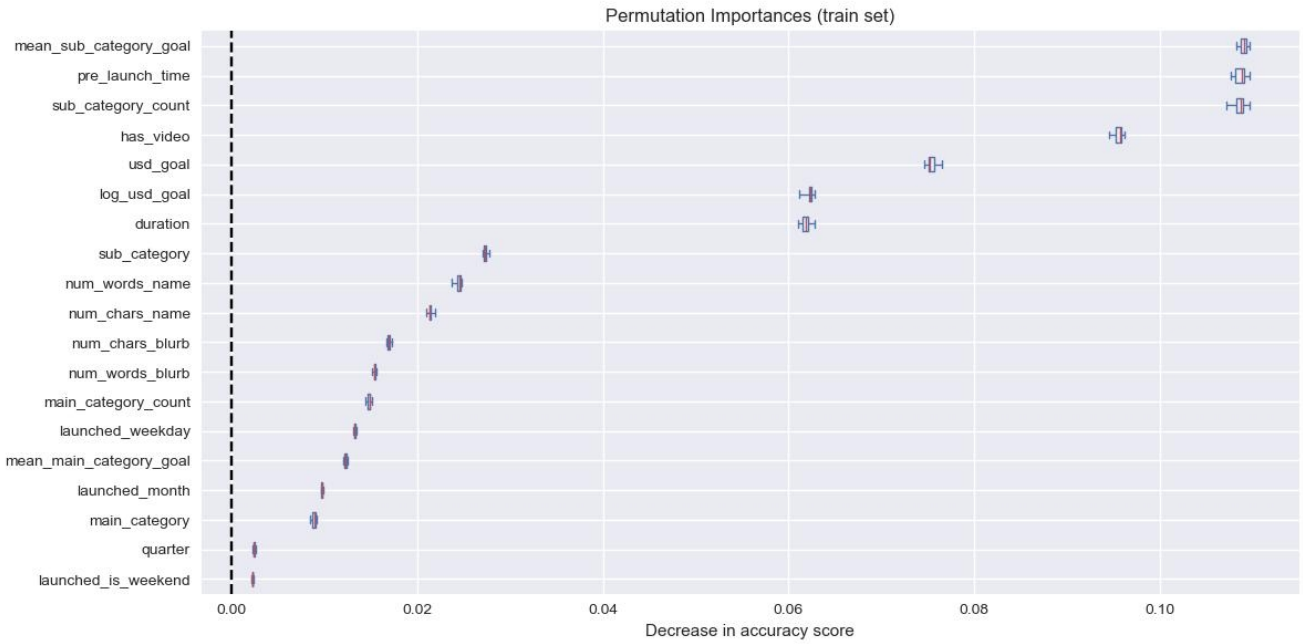
Training accuracy: 1.0

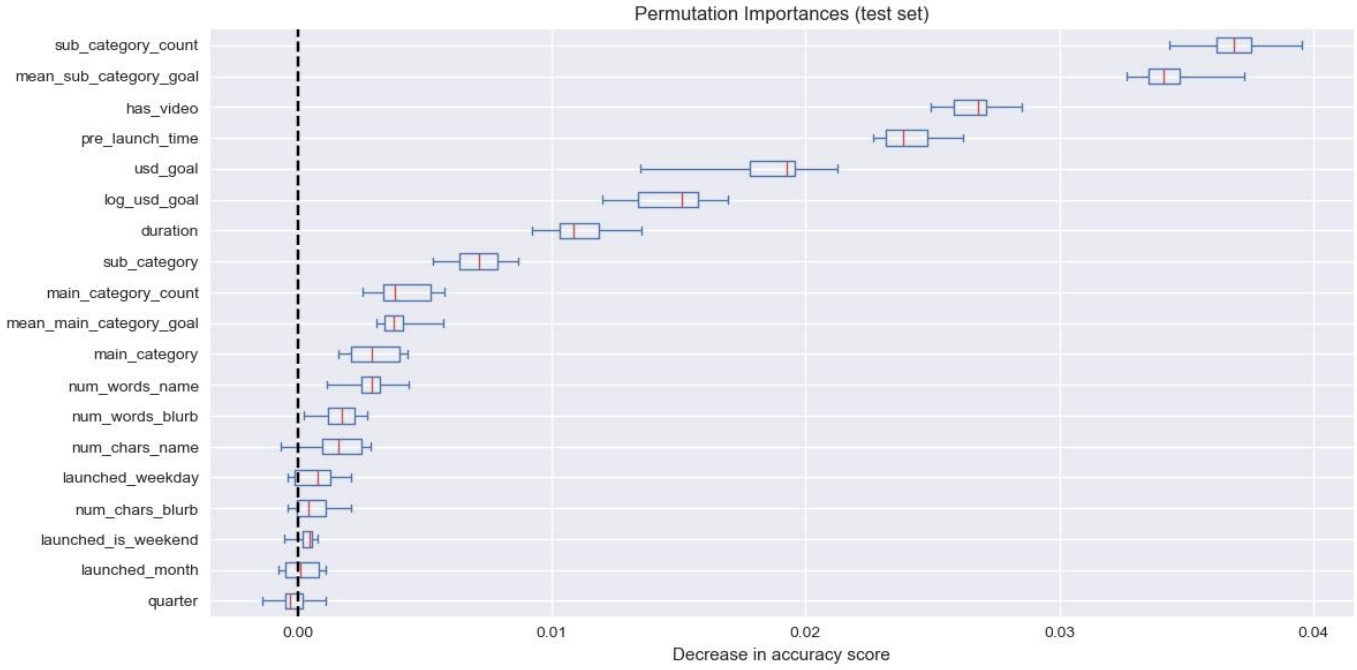
Test accuracy: 0.79



This bar chart presents feature importances based on Mean Decrease in Impurity (MDI) from a basic Random Forest model. It shows how much each feature contributes to reducing impurity (e.g., Gini index) across the trees. Features like usd_goal, log_usd_goal, and pre_launch_time are deemed most important by the model's internal splitting criteria. Surprisingly, whether the project is launched in weekend or not doesn’t affect the outcome too much.

Another interesting point is that this MDI analysis heavily favor continuous or high cardinality features. Top 10 factors from the graph are all numerical values This effect can be explained as these features offer a much larger number of potentiasl split points compared to low cardinality features, increase chance of ‘lucky split’, which significantly reduce impurity in training data.





Top Features in Train and Test:

mean_sub_category_goal, pre_launch_time, sub_category_count, has_video are consistently important in both the training and test sets.

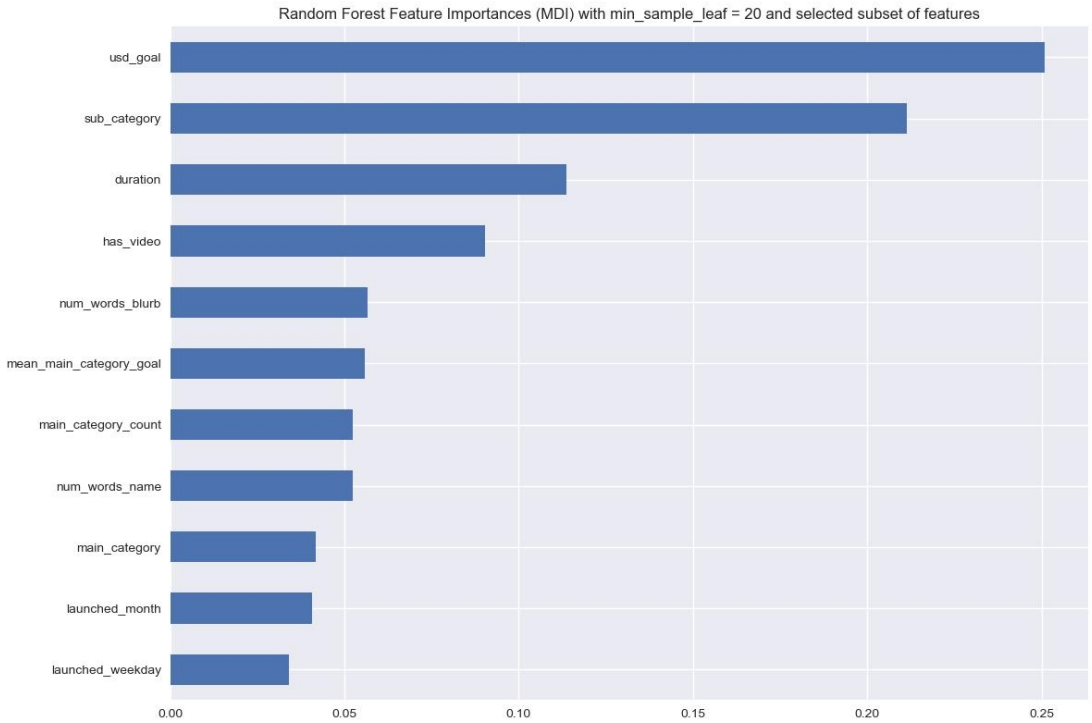
Overfitting Indicators:

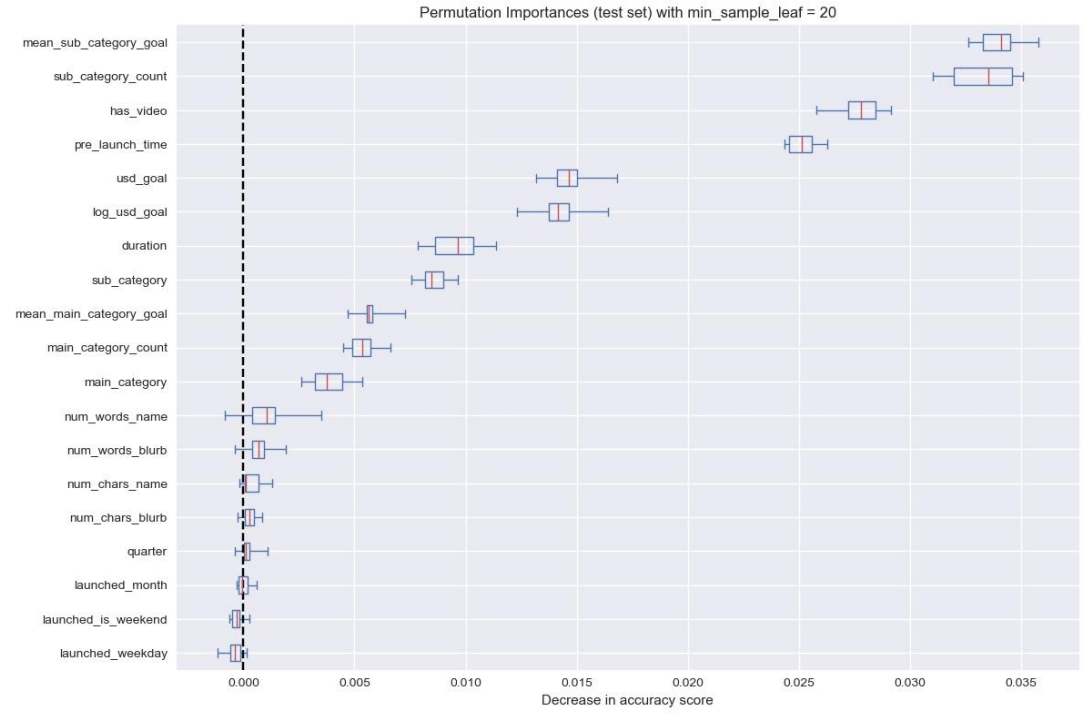
Features like usd_goal and log_usd_goal have much higher importance in the training set but lower importance in the test set. This suggests overfitting (also shown by train & test accuracy) — the model may rely too heavily on these variables during training, but they don't generalize as well Time factors (which quarter, is_weekend, or launched_month) seem do not change the result of a campaign at all, indicates that they are arguably redundant or irrelevant.

3.2.2 Set min_samples_leaf = 20

Training accuracy: 0.83

Test accuracy: 0.79

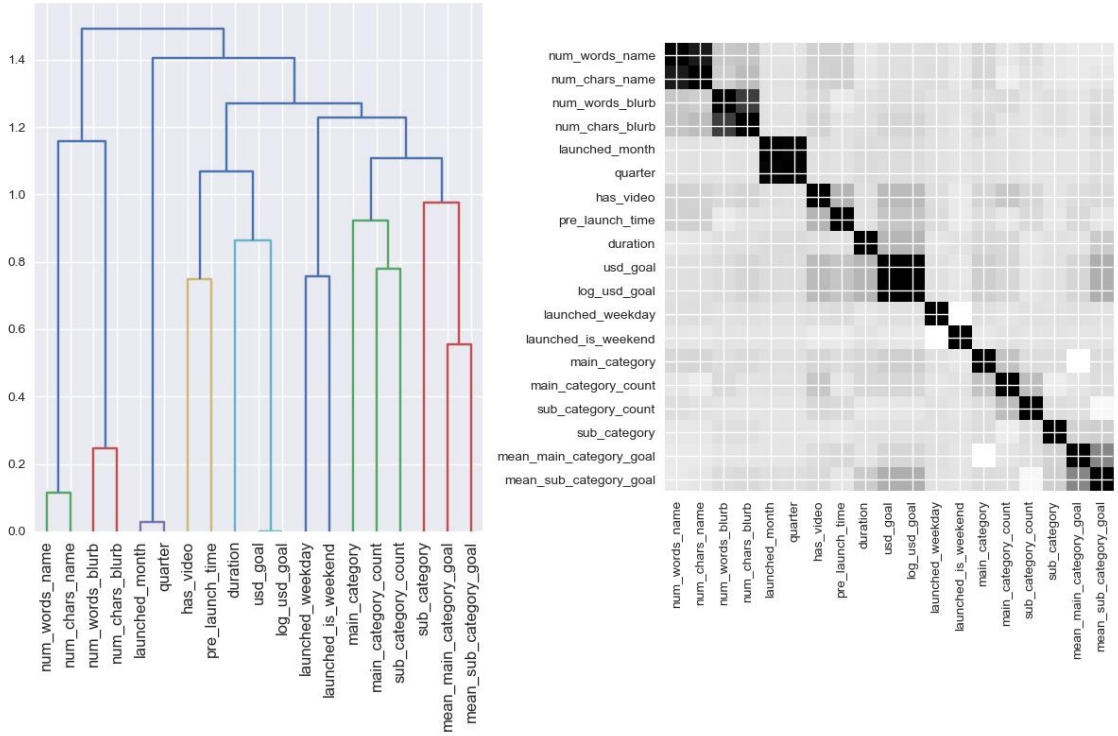




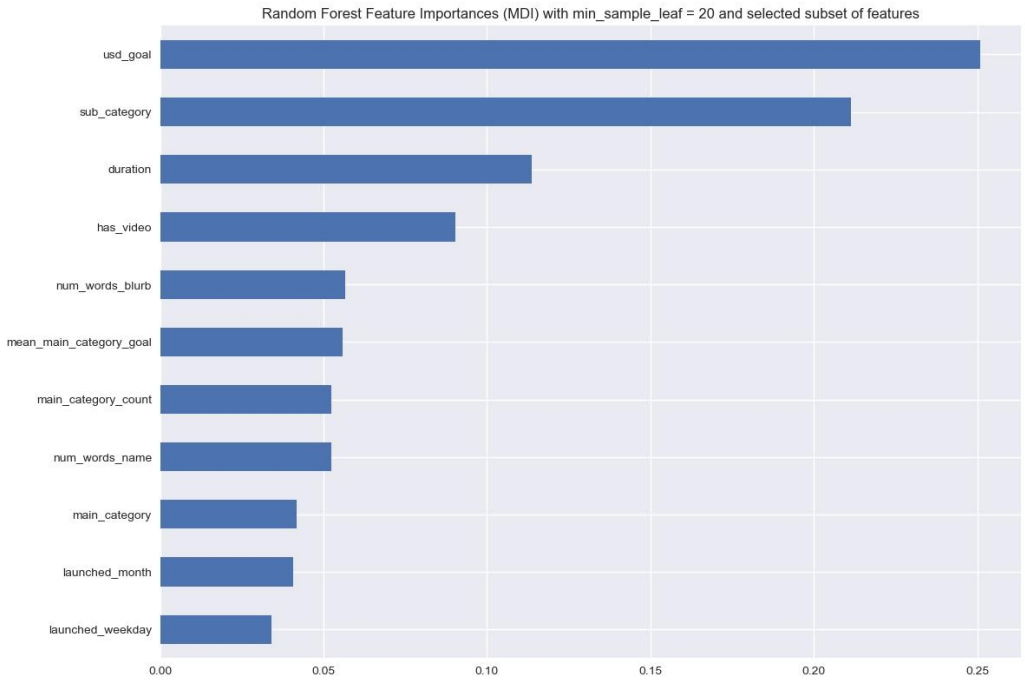
By setting the min_sample_leaf = 20, we remove some splits having very little data to calculate, which also affects some features only appear in those nodes. This version of RF reduces a big number of features in MDI, only 11 factors are included, compared to 19 factors in the base model. Less factors also increase their effects on MDI, but the order of top ranking factors is still similar.

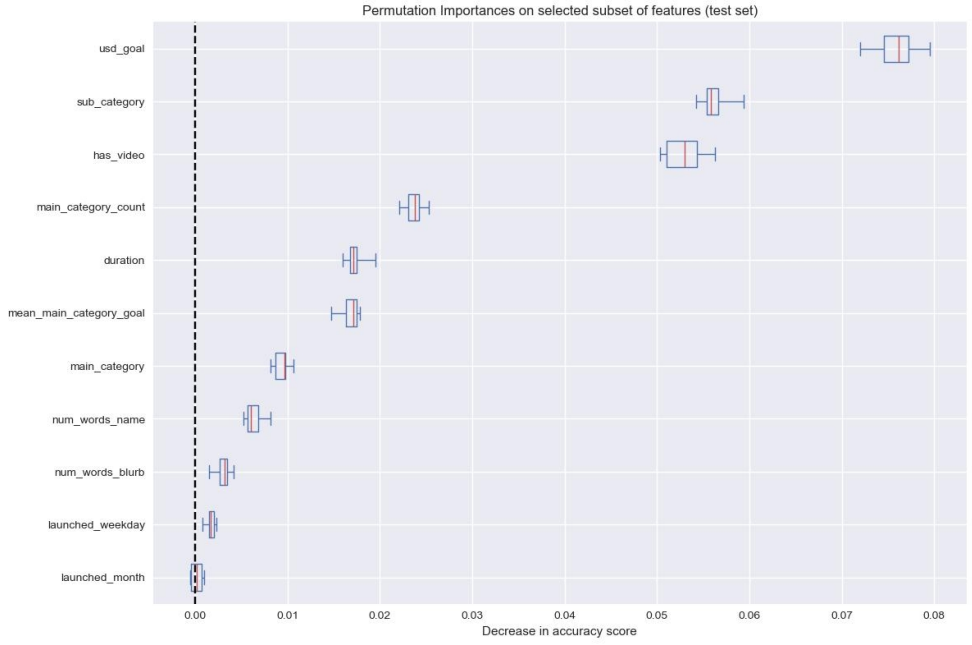
With permutation importance, the decrease in accuracy scores when each feature is randomly permuted are almost the same between these 2 models. mean_sub_category_goal, sub_category_count, has_video, pre_launch_time are top 4 most impactful features.

3.3 Handling Multicollinear Features



One way to handle multicollinear features is by performing hierarchical clustering on the Spearman rank-order correlations, picking a threshold, and keeping a single feature from each cluster. We choose threshold = 0.8 for this analysis, mostly because it's the highest decimal that seperate duration and usd_goal. There are 11 features left after running this hierarchical clustering, and their are exactly the same as 11 features in second version of RF with min_samples_leaf = 20.



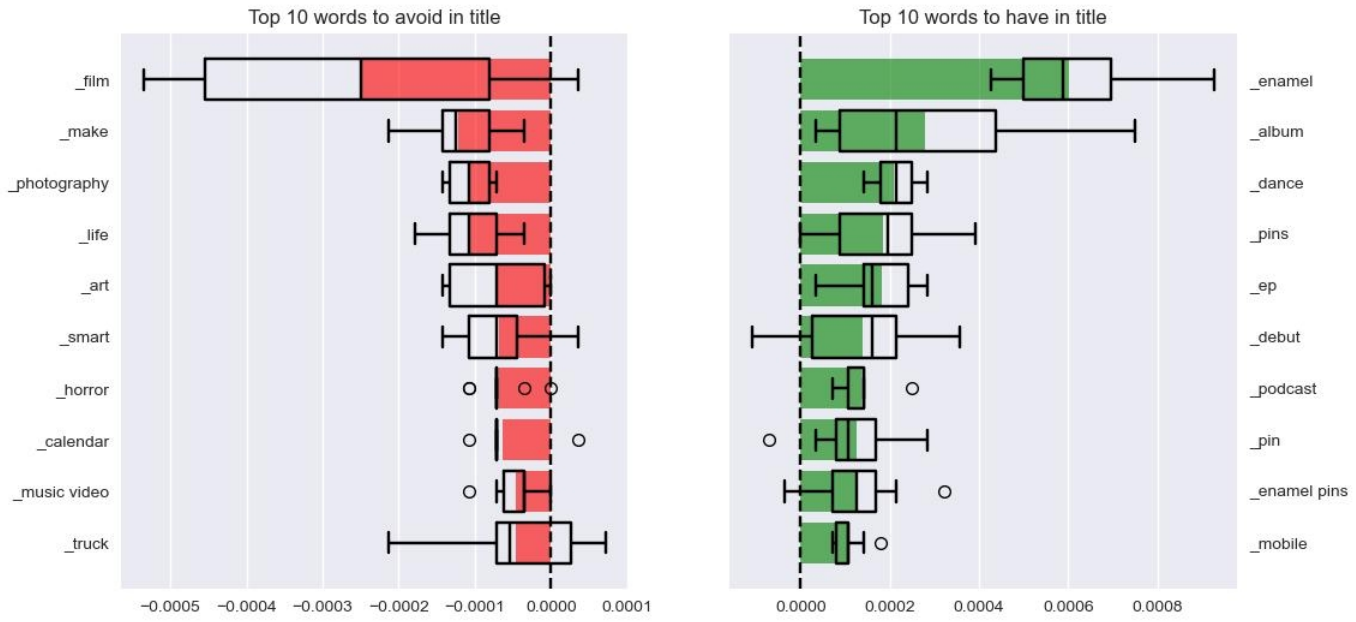


With the new set of features, we easily see that usd_goal and sub_category are the top 2 features. Although has_video is low on MDI, it shows the importances on permutation, as it's very close to the second factor. Same results can be interpreted from the bottom features, launched_weekday and launched_month don't change the outcome very much.

3.4 Presence of which keywords makes the biggest impact in the predictions

Words are vectorized using TfidfVectorizer from scikit-learn, ignoring lowercase / uppercase, and remove english stopwords. There are stopwords from other languages as well, but the numbers are low, so we keep it in our 'vocabulary'. Each word will have its own feature, added to train data as categorical variables, feed into same model as section 3.2.2 (min_sample_leaf=20)

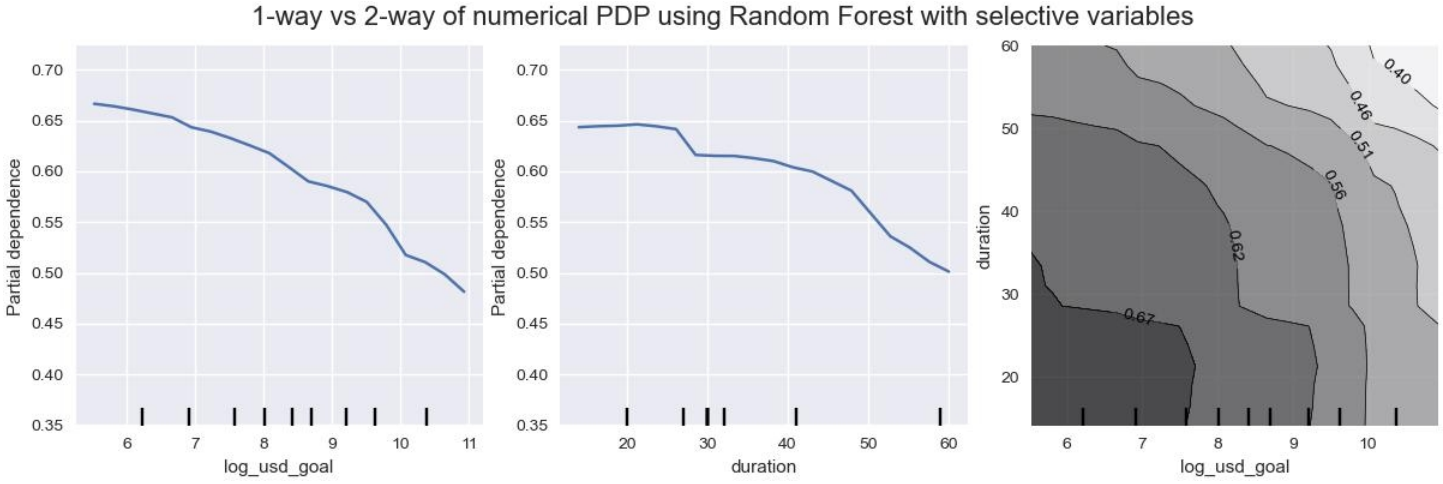
Having these words in the title can increase or decrease your chance of successful campaigns.



3.5 How does changes in features lead to changes in model outcome? (Partial Dependencies)

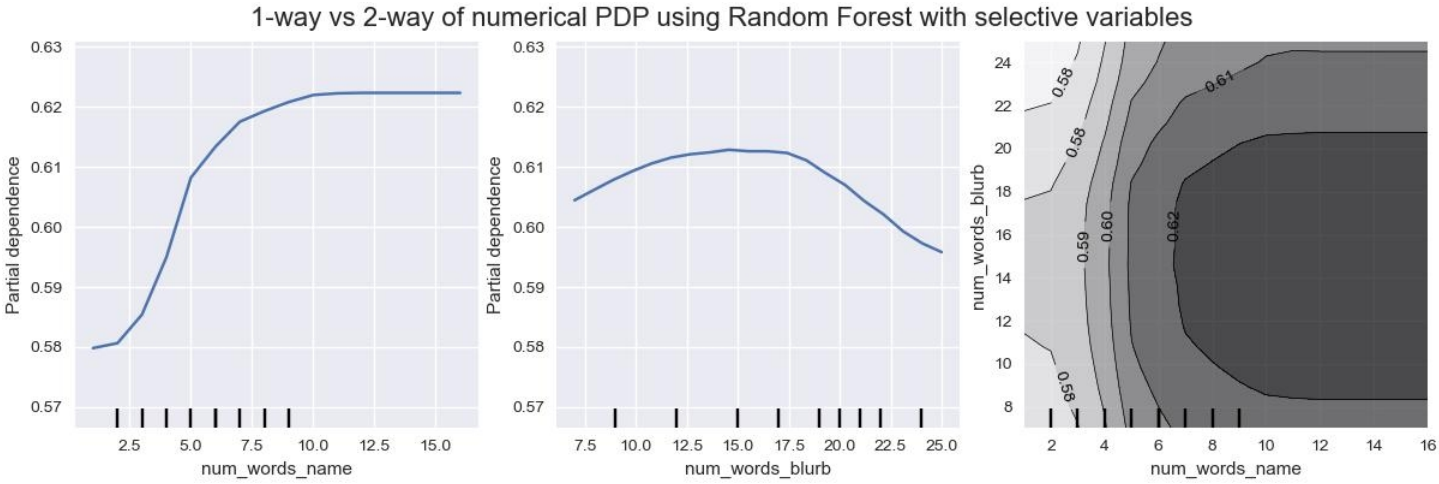
(3rd graph, the darker, the higher success. Partial Dependency Plot = PDP)

log_usd_goal vs duration:



- As log_usd_goal / duration increases, the predicted success probability decreases.
- Projects with lower funding goals are predicted to succeed more often.
- The vertical bars show data distribution (tick marks indicating where training examples lie).
- Projects with both low log_usd_goal and short duration have the highest predicted success (bottom-left corner).
- As either goal or duration increases, the probability drops (top-right is the lowest).
- There is a negative interaction between the two variables: longer campaigns with higher goals are penalized more heavily by the model.

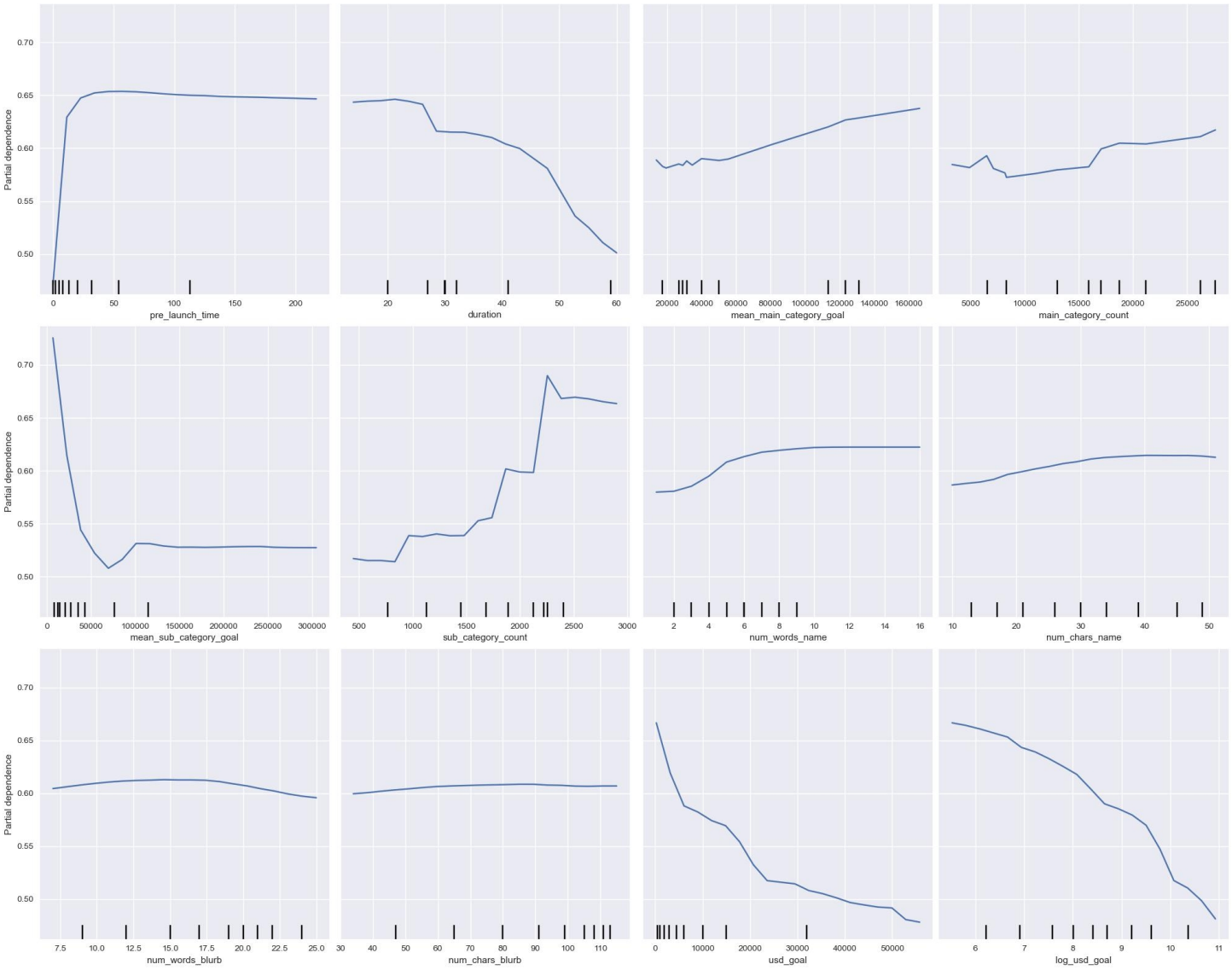
Number of words in title (num_words_name) vs in description (num_words_blurb)



Interpretation:

- Projects with long descriptions are less likely to succeed.
- Longer title seems not having much effects on the outcome of campaigns after a certain amount of words.
- The vertical bars show data distribution (tick marks indicating where training examples lie).
- The best combination of length is 7-10 words in title, and 13-17 words in the description.

1-way of PDP for all features:



3.6 Understanding the decisions made by the Model (using SHAP)

Here we use 1 sample of test data to predict where it will be successful or not.

launched_month = 8 (August)

launched_weekday = Tuesday

launched_is_weekend = False

quarter = 3

main_category = Technology

sub_category = Software

has_video = True

pre_launch_time = 1 days

duration = 54 days

mean_main_category_goal = 113533.376247

main_category_count = 18748

mean_sub_category_goal = 128134.196898

sub_category_count = 1615

num_words_name = 8

num_chars_name = 40

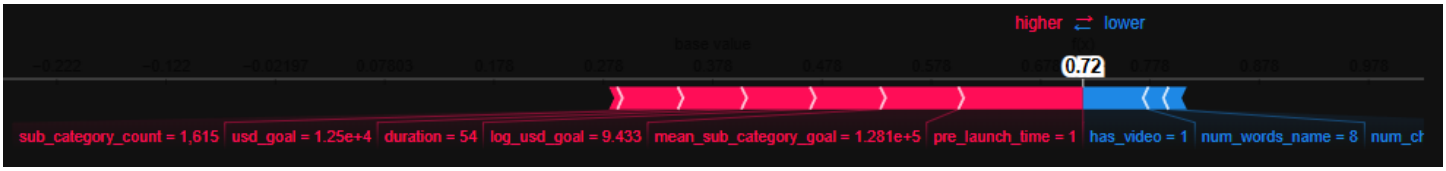
num_words_blurb = 18

num_chars_blurb = 93

usd_goal = 12500

log_usd_goal = 9.433484

The image below shows the chance of this project being 'failed', 72%, meaning our model predict the probability of this project being successful is 28%. Each segment is a Shapley value from SHAP python library, as it measures how much each factor contribute to the overall result.

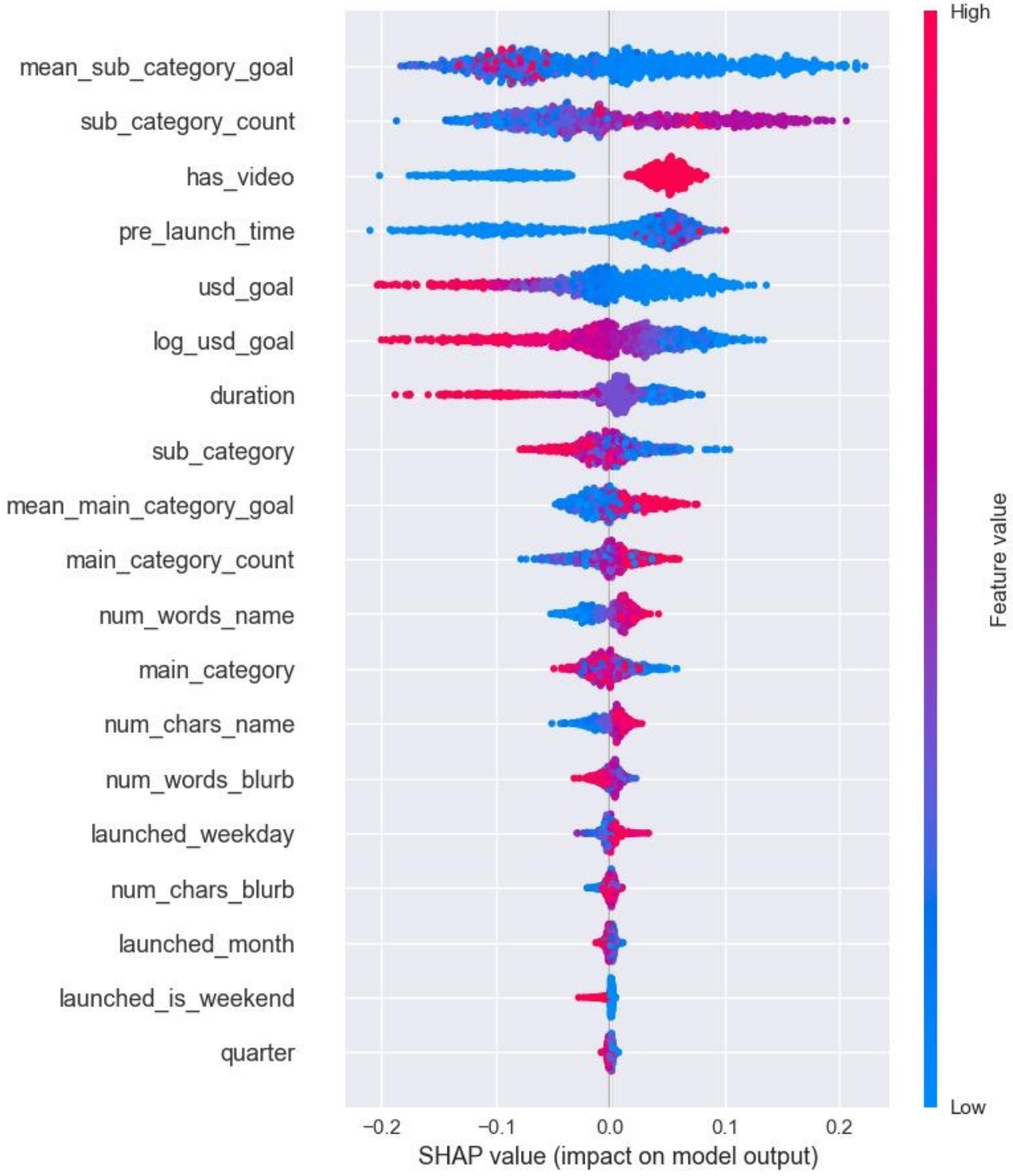


The main reason for the model to judge 'failed' is little pre_launched time, very high average main category goal, log_usd_goal and long duration. Some arguments for a successful campaign is that it has video and good title length, but it's not enough to outweigh the negative effects.

Note that features have more impact within 1 campaign (both positive and negative) are closer to the middle ground (0). The same idea applies for the summary charts below, there 1000 test samples are evaluated for how much impact is from their features.

The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature

from low to high. Overlapping points are jittered in the y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance. In the summary plot, we see first indications of the relationship between the value of a feature and the impact on the prediction. A video contributes greatlly to the success. The summary plot also shows that mean_sub_category_goal has the widest range of effects .



Potential Impact

This research offers significant implications for both Kickstarter developers, creators and the crowdfunding ecosystem as a whole:

- For creators: The ability to prioritize top campaign factors, improve planning, and lower the risk of failure.
- For Kickstarter (Platform-Level): Gaining insights into feature significance can help direct UI/UX improvements, offer targeted suggestions during campaign setup, and improve creator support tools.
- For researchers and analysts: The project advances data science methods for feature selection and interpretability within real-world business platforms.
- For educators and entrepreneurs: Offers a evidence-based reference for teaching best practices in crowdfunding, digital entrepreneurship, and product marketing.

What could have done more

Use different models than Random Forest: boosting techniques like XGBoost, AdaBoost, or LASSO L1 Regularization, recursive feature elimination (RFE)
Checking other dataset if possible
Mean category goal, mean sub category goal can use only 1 year window frame, meaning instead of calculating average of all campaigns in the history, only campaigns launched within 1 year before launched date are included for the calculation. This will help the model grasp a more accurate goal & pledge amount, prevent inflation effects where projects in early years are much cheaper.
To further improve the model's performance and interpretability, consider experimenting with alternative algorithms beyond Random Forest, such as XGBoost, AdaBoost, or LASSO (L1 regularization) might offer better predictive accuracy and insights, especially with high-dimensional data. You may also implement Recursive Feature Elimination (RFE) to systematically identify the most influential features, just for Exploratory Data Analysis.
Additional suggestions include:

- Explore other datasets, if available, to enhance model generalizability and robustness.
- When calculating mean_category_goal and mean_sub_category_goal, use a rolling 1-year window instead of aggregating across all historical data. This approach ensures the averages reflect recent trends, reducing the risk of inflation bias from outdated or irrelevant campaigns. Limiting to a 1-year window helps the model better capture contemporary funding expectations and goal-setting behavior, aligning predictions with current market conditions.

Conclusion

In summary, this study provides useful practical insights for different stakeholders involved in crowdfunding. By combining careful analysis with practical applicability, it sets the stage for more informed, user-focused, and data-driven decisions across crowdfunding platforms.

The feature analysis revealed that the most impactful features for campaign success are: **goal**, **mean_sub_category_goal**, **pre_launch_time, sub_category_count**, **has_video** and **duration**. In particular, the funding goal and sub-category goal averages provide insight into financial expectations and benchmarks. Engagement and competition levels are reflected by the campaign duration and the number of campaigns within a sub-category. Additionally, the presence of a video enhances user engagement, while pre-launch preparation time indicates the level of planning and readiness. These findings can guide future campaign strategies and feature selection for predictive and prescriptive modeling.

Reference

Github
Kaggle 1, 2, 3, 4
Scikit learn documentation: 1, 2, 3, 4, 5
SHAP documentation: 1