Predicting the success of a crowdfunding campaign

Colby Shelton • Thinkful

Supervised Learning Capstone

Overview

- What is crowdfunding?
- Application of model
- Workflow

  - Exploratory Data Analysis
  - Feature Engineering
  - Model Selection
  - Evaluation

- Conclusion
- Next steps

What is crowdfunding?

Wikipedia states the answer well:

“

Crowdfunding is the practice of funding a project or
venture by raising small amounts of money from a large
number of people, typically via the Internet.
Online companies like Kickstarter and Indiegogo
as well as other non-profit platforms.

How much money has been raised you ask?

According to Fundly.com:

“

The crowdfunding industry has raised $34 billion
and it is projected to grow to over
$300 billion by 2025!

That’s a lot of money considering Fundly.com also stated that only 50% of crowdfunding campaigns get fully funded.

It’s obvious that there’s plenty of room to help improve that success rate.

How?

Using machine learning,
I believe it’s possible.

A model that can predict if
a campaign will succeed before
it’s launch would provide great value
to any crowdfunding product team.

How would this model help?

How would this model help?

- Time

How would this model help?

- Time
- Money

How would this model help?

- Time
- Money
- Frustration

Workflow

Exploratory Data Analysis

Campaign data sourced from Kickstarter.com and available at Kaggle.com.
Campaigns from 2May09 - 2Mar18

Included variables:

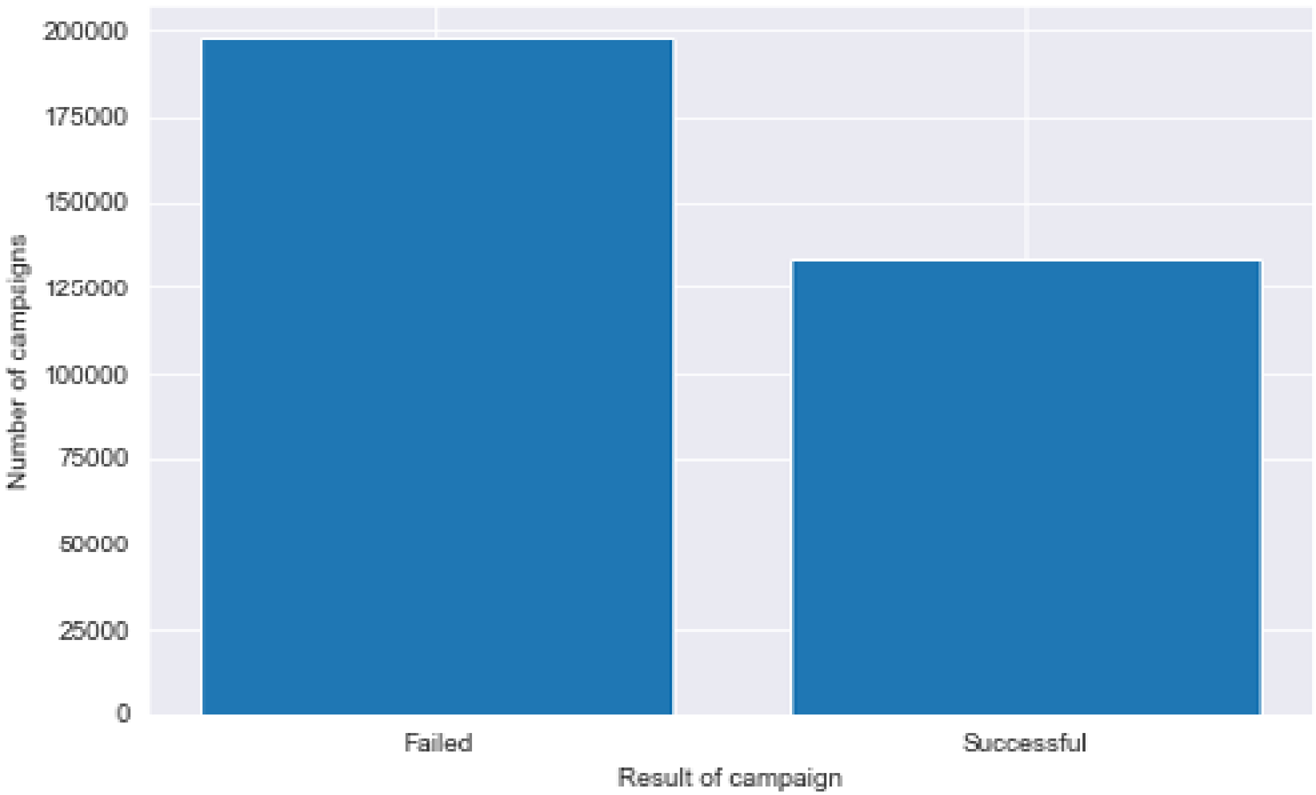
- ID
- Name
- Category (159)
- Main Category (15)
- Currency
- Deadline
- Goal
- Launched
- Pledged
- State
- Backers
- Country
- USD Pledged

Missing Data

- Dropped less than 5 data points on one necessary feature after dropping unnecessary variables

Target Variable

- Successful Campaigns
- Of 331,672 campaigns, 40% were successful

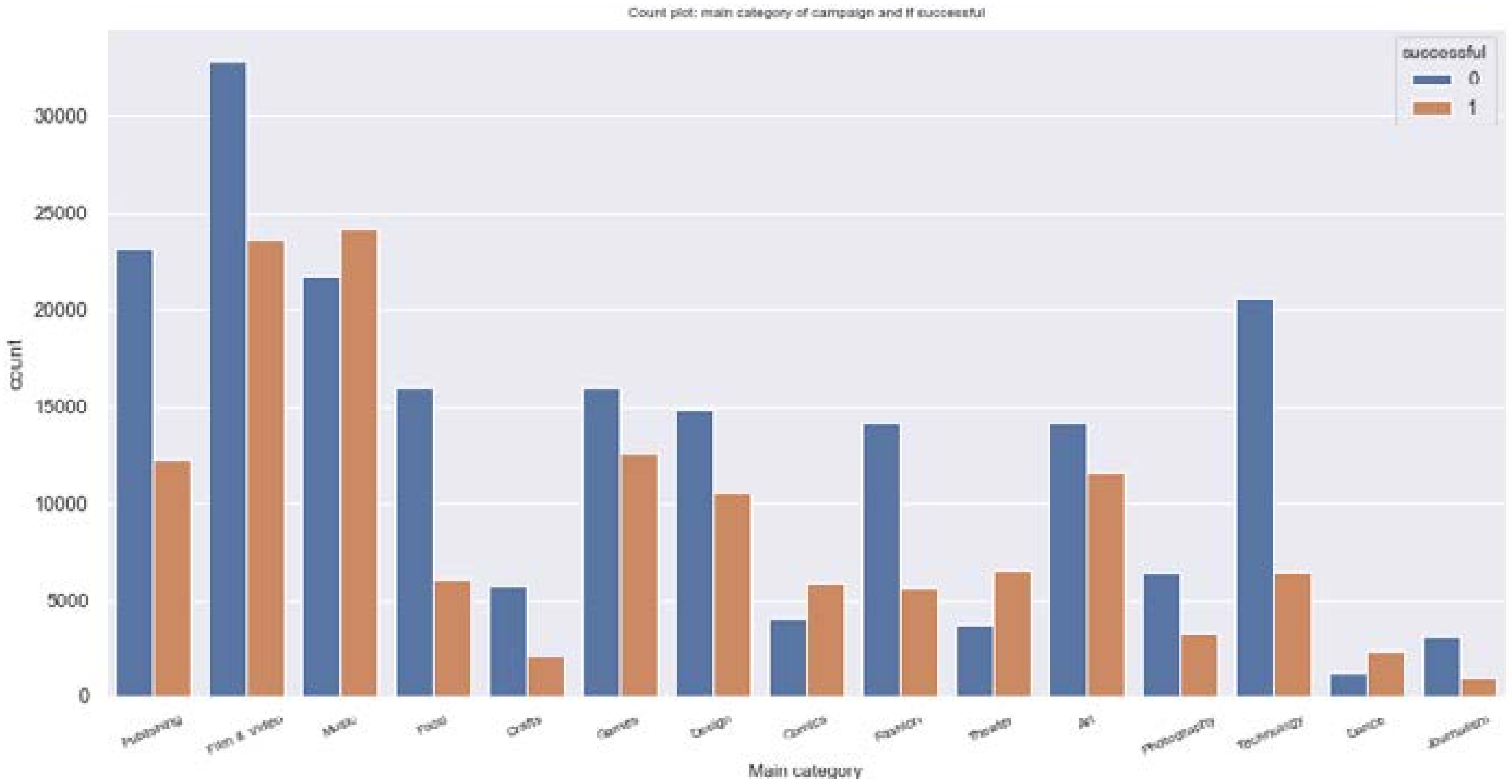


Categorical Features

(Required feature engineering)

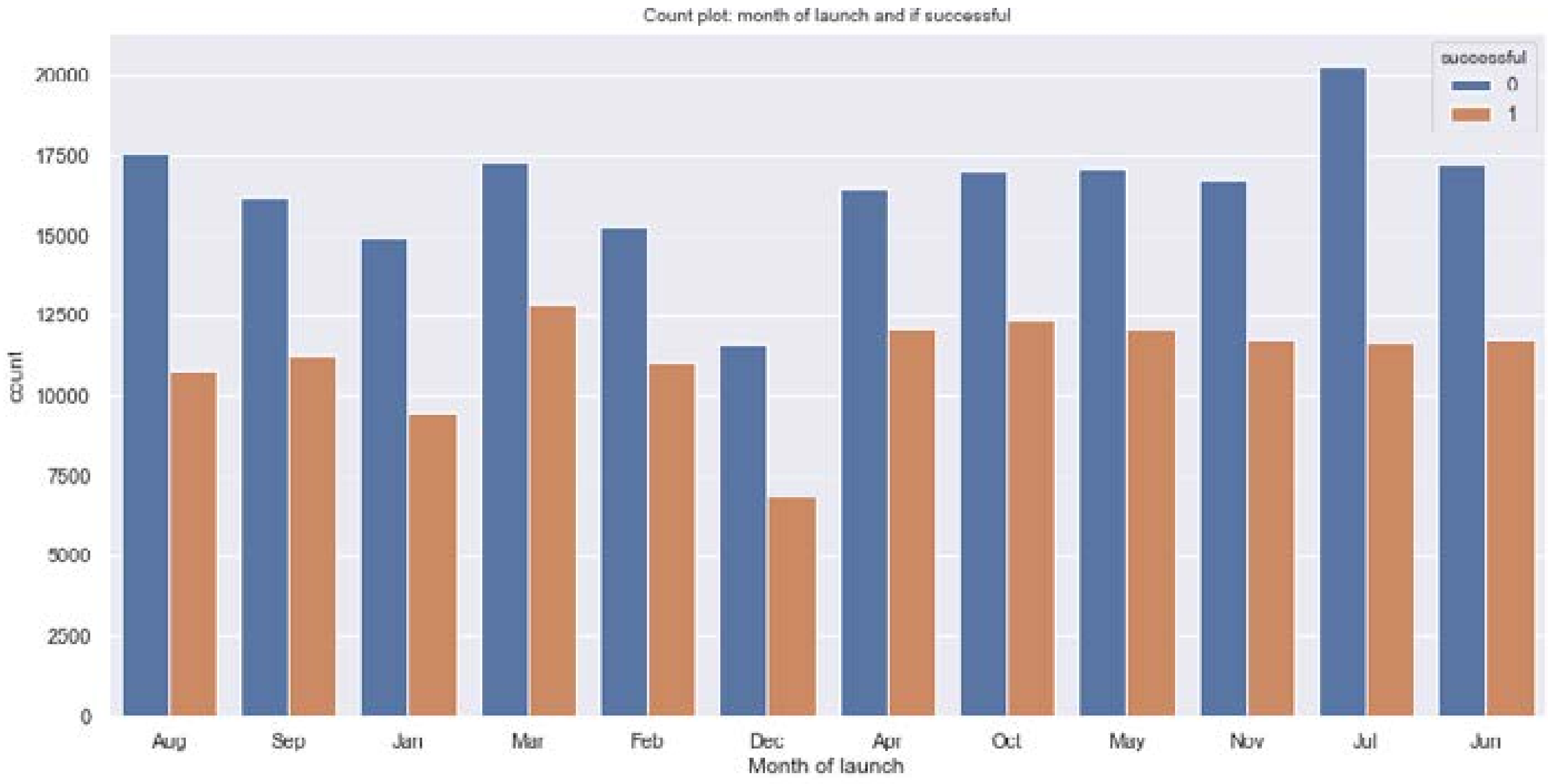
Main Category

- Music, Comics, Theater, & Dance are the only categories with more successful campaigns than failed.
- Technology raises the most money, but has the largest difference in successful vs. failed.



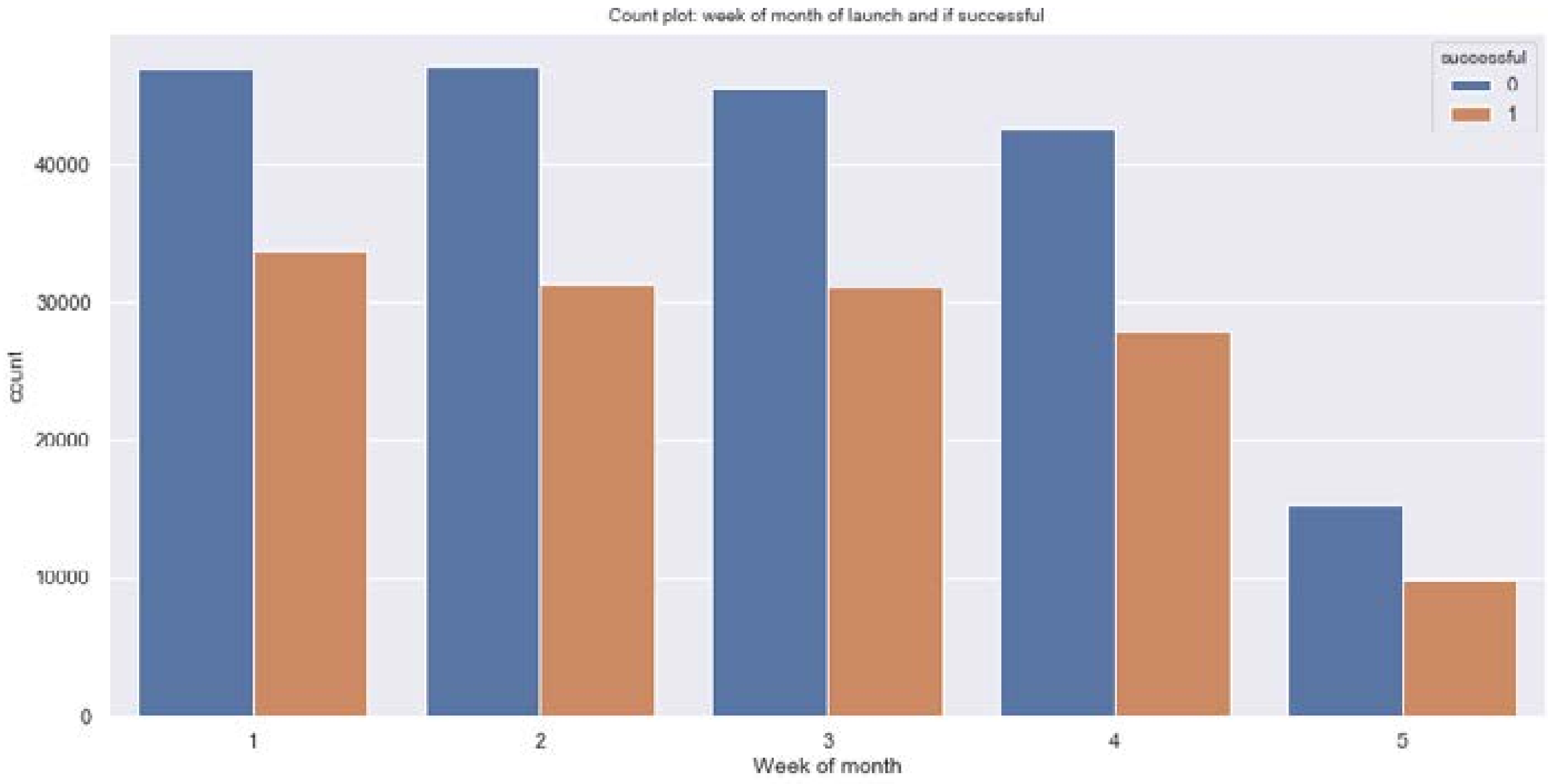
Month of Launch

- The summer months show the highest rate of launches.
- Dec sees the least amount of launches.



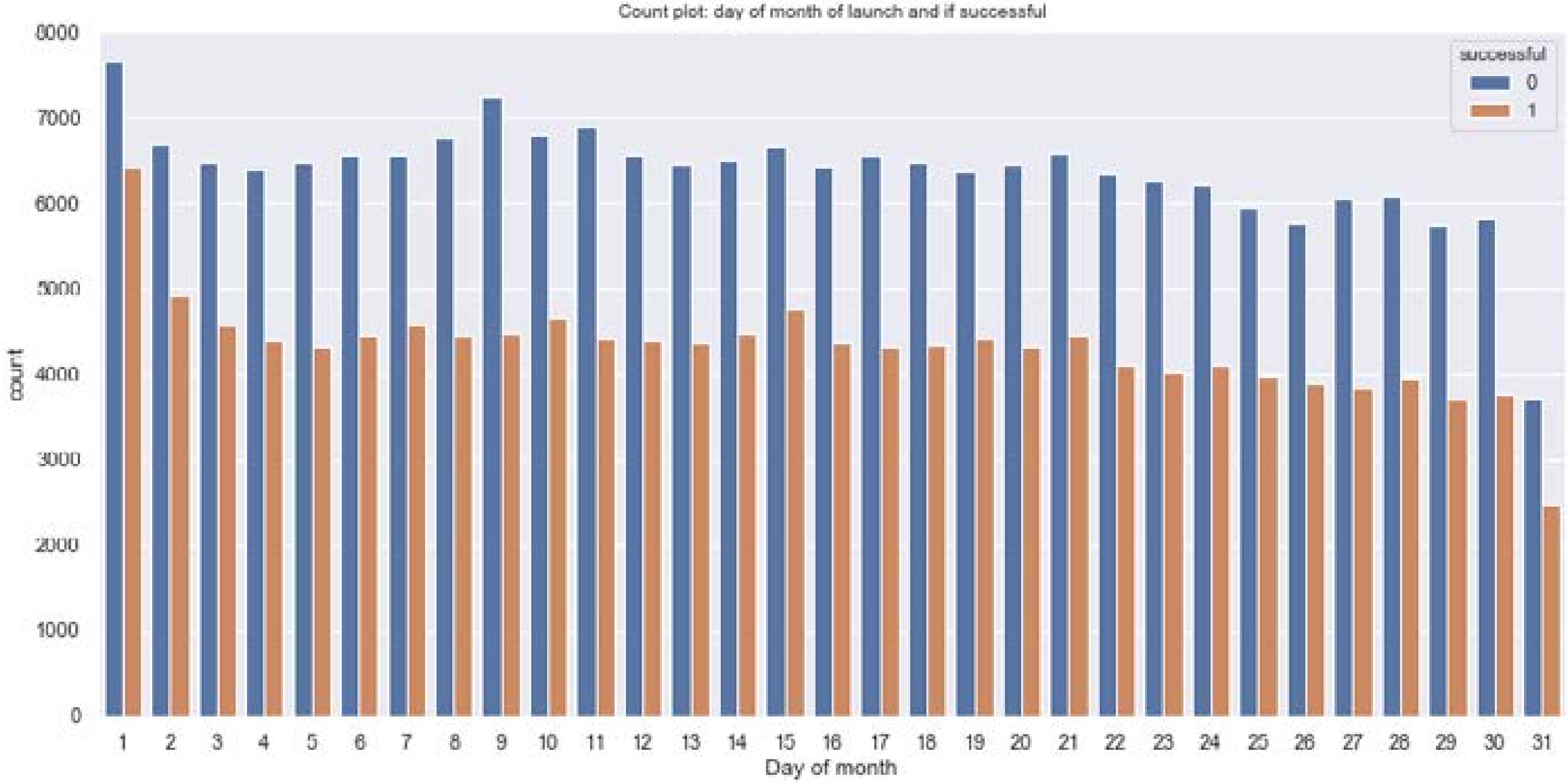
Week of Month of Launch

- There’s a pretty good balance except in week 5
- There’s only 4-5 fifth weeks per year



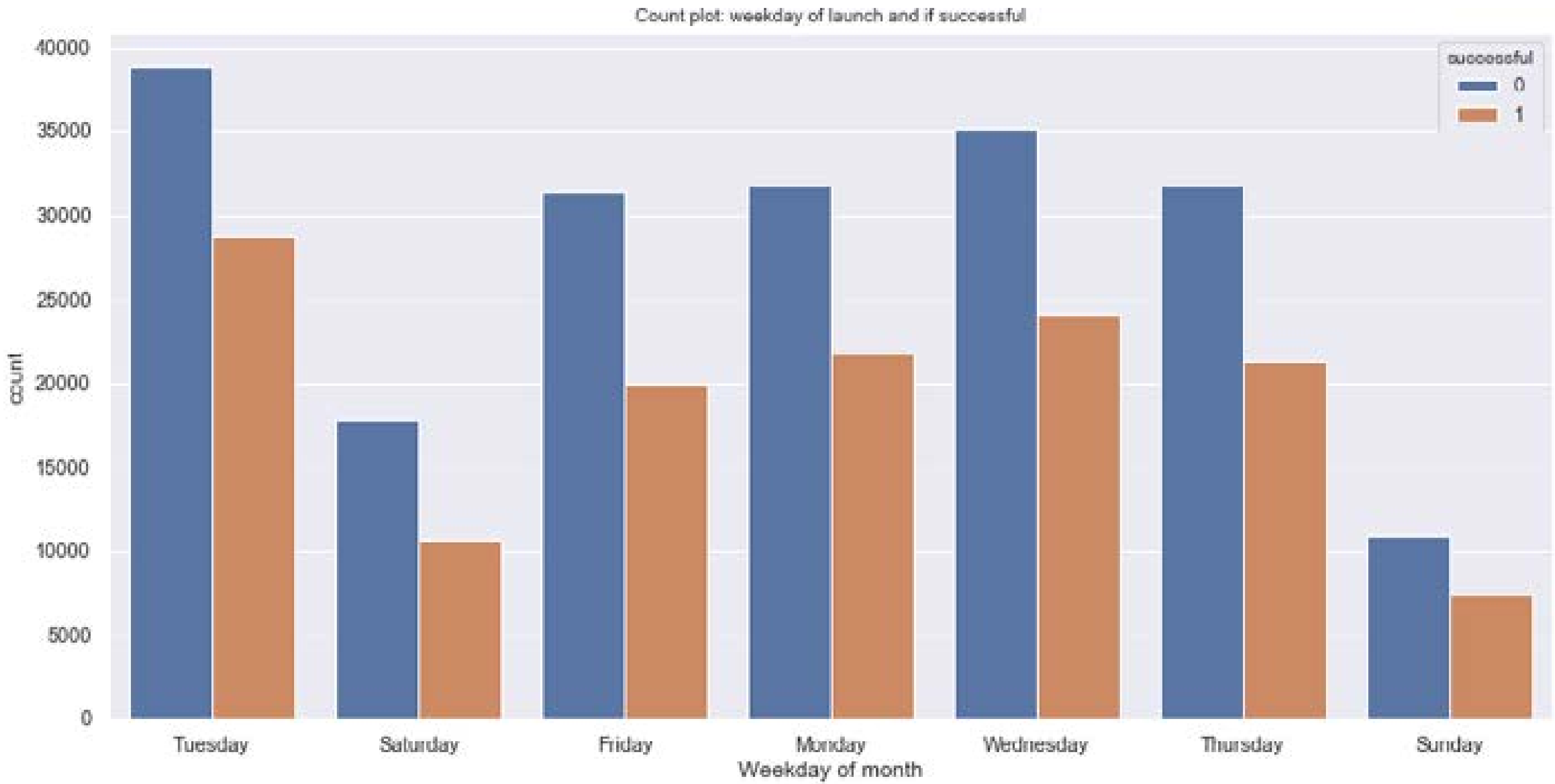
Day of Month of Launch

- More launches on first payday of each month. (1st)
- Only six 31st's a year so you see a significant dip



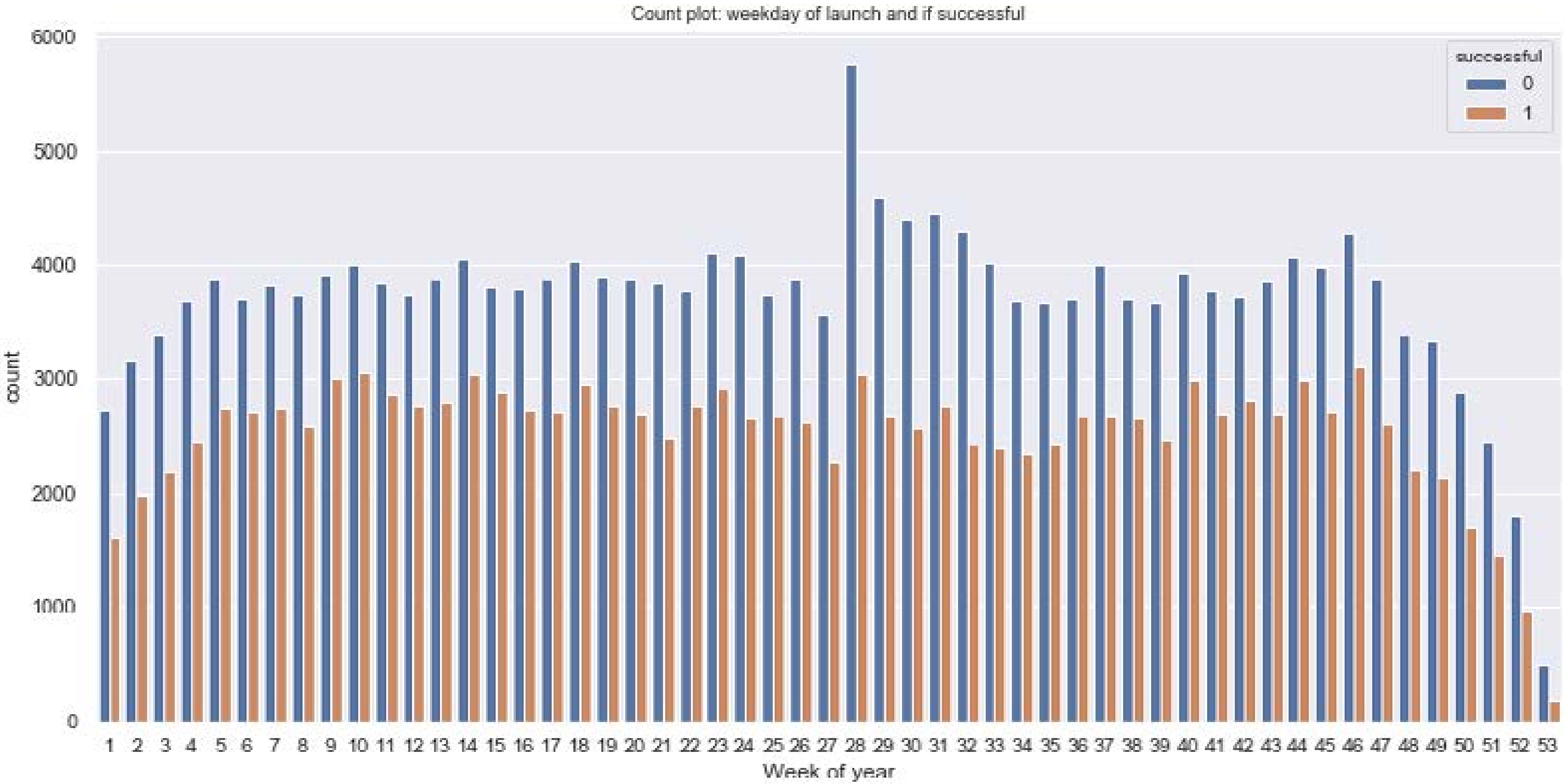
Weekday of Launch

- A majority of launches happen on Tuesday’s
- We see that weekends see less launches



Week of Year of Launch

- Peak in the 28th week for campaign failures
- Downward trend starting after Thanksgiving/Black Friday into New Years where it begins to stabilize around Feb

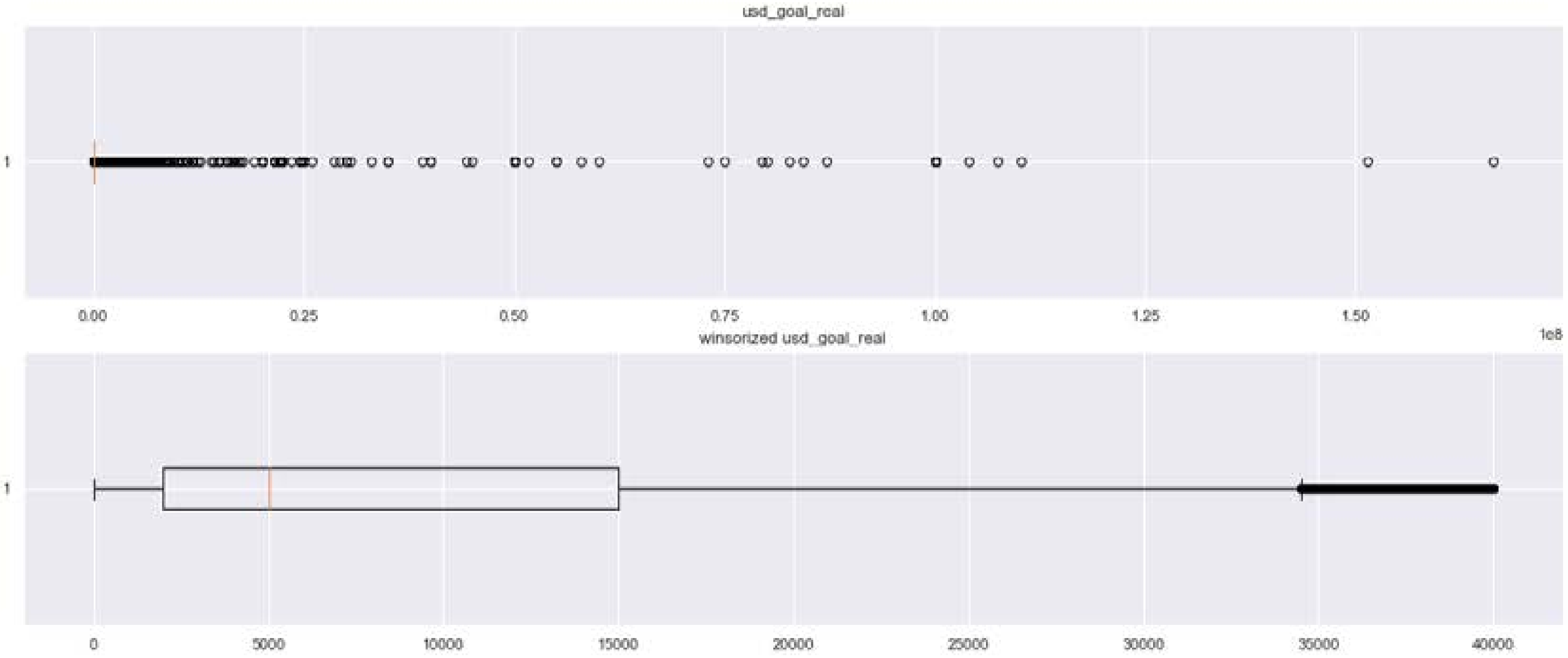


Numerical Features

(Required feature engineering)

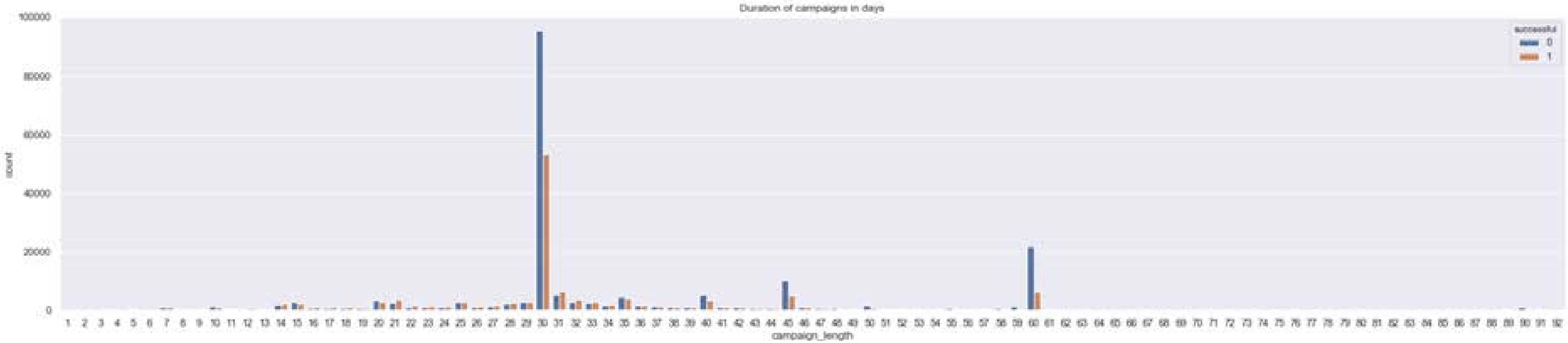
Monetary Goal of Campaign

- Due to outliers, I decided to use winzorization and kept all data points within the 85th percentile.
- The average goal was $5000.



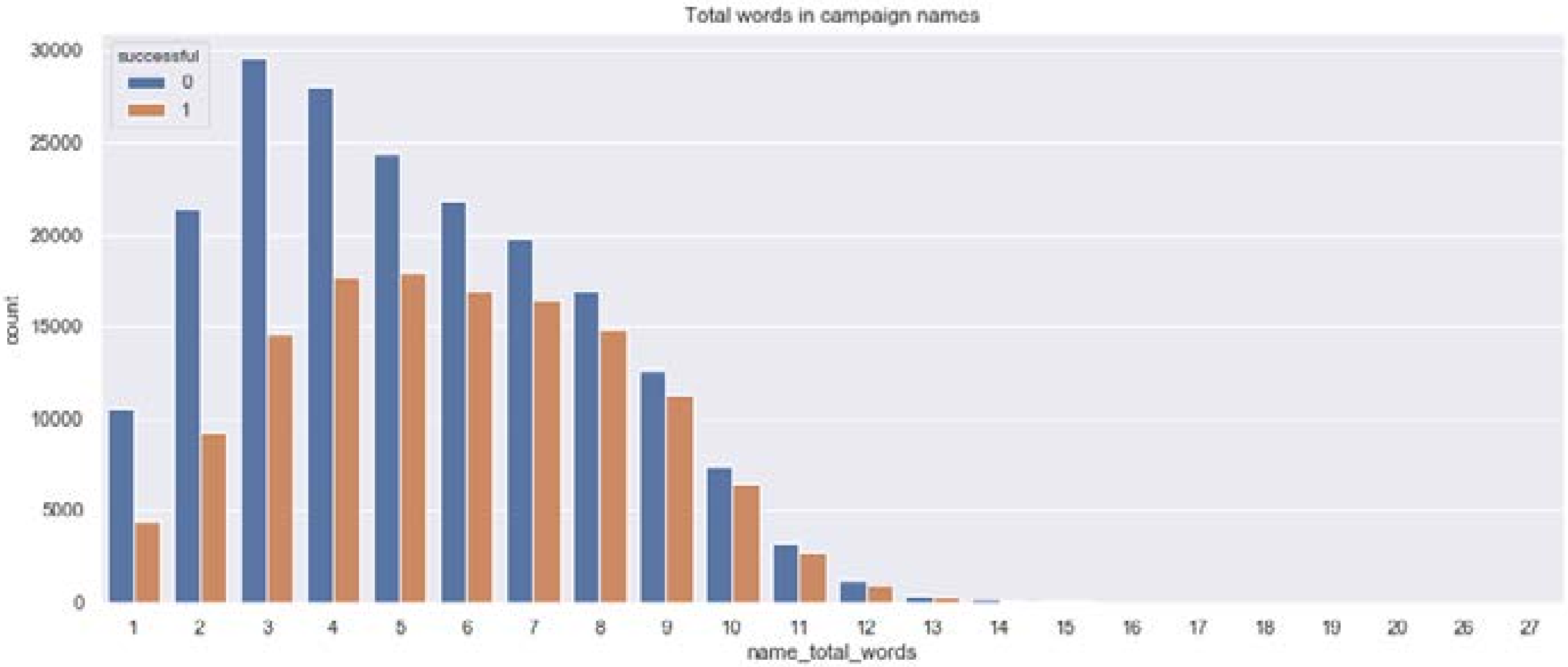
Campaign Duration

- A vast majority of campaigns run for 30 days.
- Other popular durations are 45 and 60 days.



Number of words in campaign name

- Names with 5 words have the most success
- Names with 3 words have the most failures



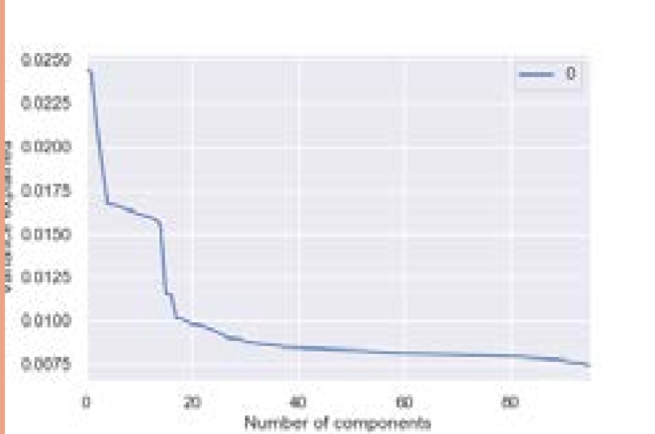
Feature Engineering

Feature Engineering

- Removed outliers: usd_goal_real using winsorization.
- Created name_word_length feature using name of campaign after cleaning name column
- Create launch date variables using launch date & deadline features.
- Used get_dummies(true/false) on categorical features: main_category & all campaign launch date related features.
- Dropped features causing data leakage (ie. backers, usd_pledged)

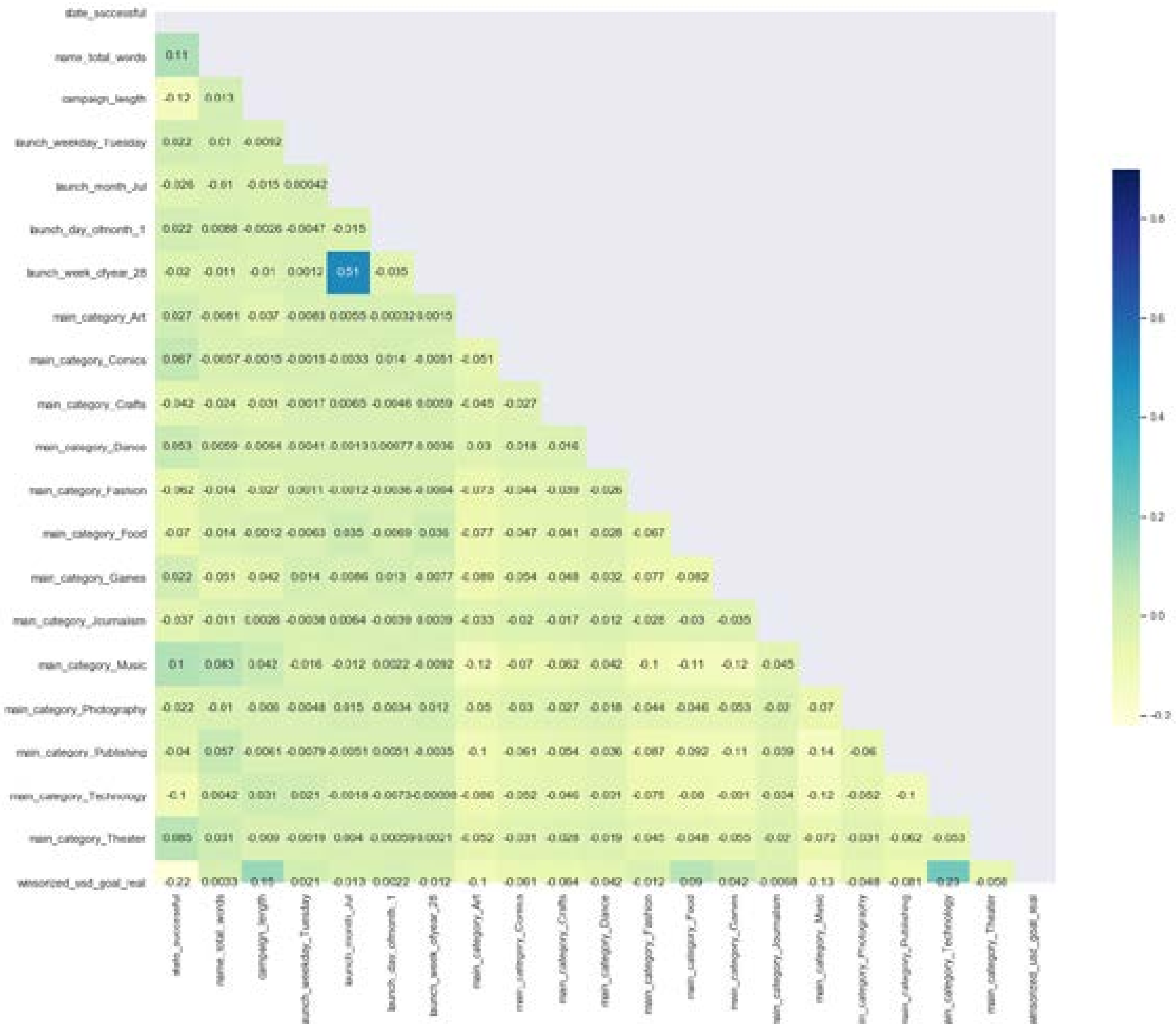
Feature Selection

- Principal Component Analysis (PCA): Needed to keep 96 of 126 features to explain 95% of the variance.



- So I went with **SelectKBest** that selected 20 features with the highest accuracy scores.

Correlation Heatmap: 20 K Best features



Model Selection

Classification

Successful/Failed

- **K-Nearest Neighbors**

  - Classified from votes based on proximity to k-nearest neighbors.

- **Logistic Regression**

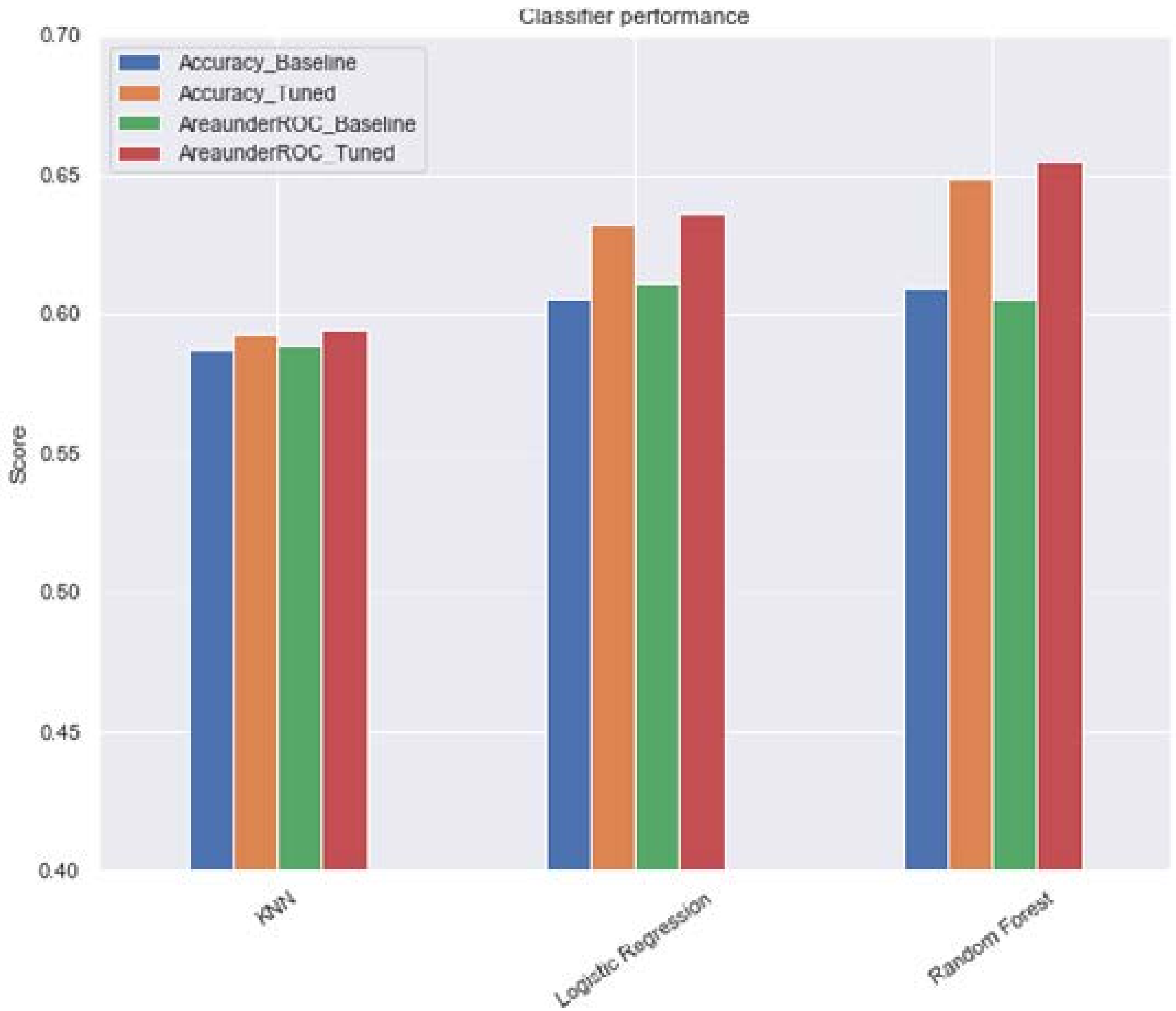
  - Find a relationship between features and probability of particular outcome.

- **Random Forest**

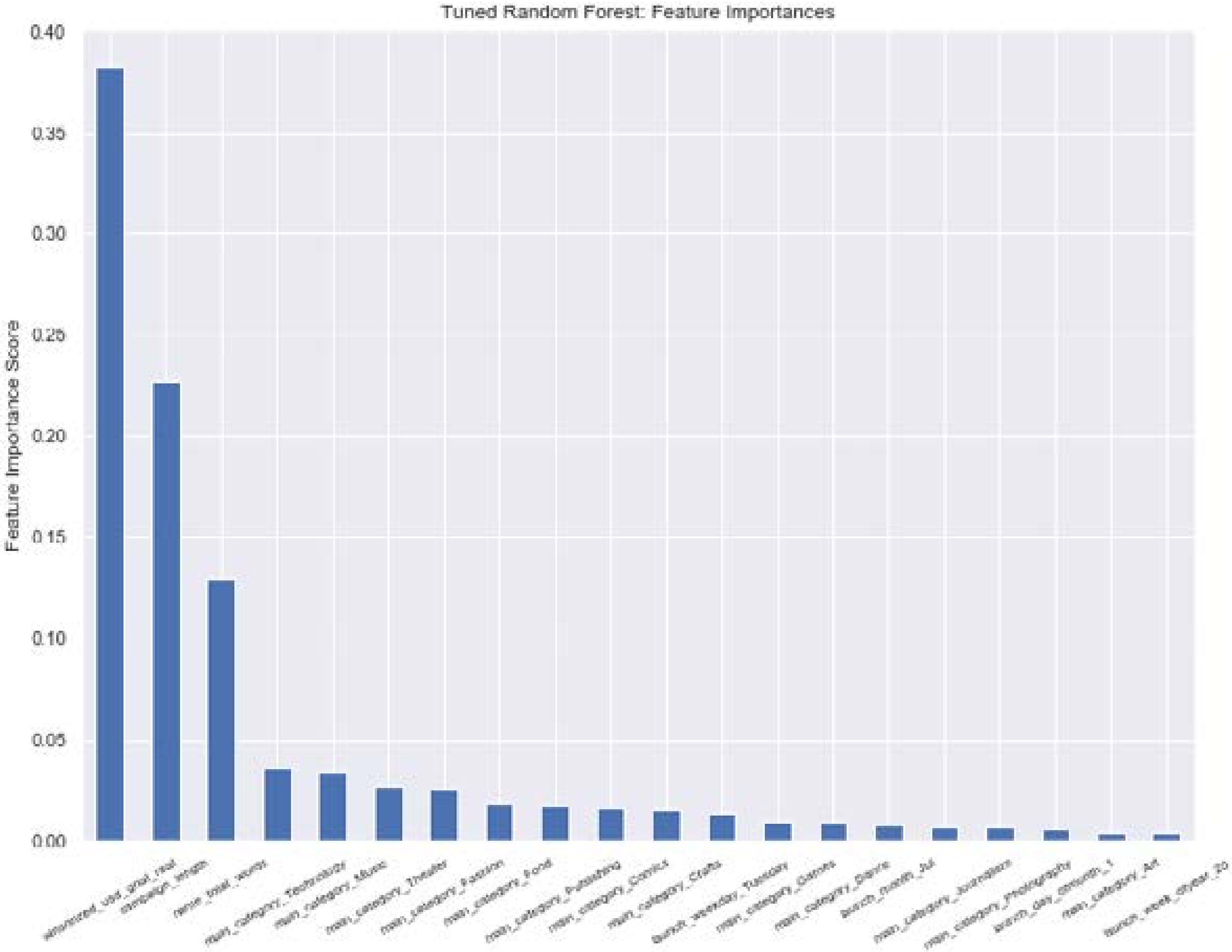
  - Multiple decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Evaluation

Evaluation of performance before/after tuning



Feature Importance: Tuned Random Forest



Conclusion

Results

- Logistic Regression and Random Forest had similar baseline accuracy.
  - (%60.6 vs %60.9)

- Random Forest had the greatest accuracy imporvement after tuning hyperparameters.
  - (%64.9)

- Random Forest took much longer in terms of run time compared to Logistic Regression.
  - (RF = 35 minutes vs. LR = 12 seconds)

- Random Forests are great predictors; more difficult to explain than Logostic Regression.

Results

- With campaign success odds currently sitting at 50/50 (Fundly.com), I believe this model is already showing promising results.

- With accuracy currently at %64.9, there's definitely room to improve. But, I believe I'm onto something in terms of providing value to a crowdfunding product team.

- With further feature engineering, like that of using campaign name word count, there's more creative insights to be discovered that will help improve model accuracy and give campaign managers insight on how to market their products (i.e. what to name the product knowing that the name word length and success rate are fairly correlated.)

Next Steps

Next Steps

Useful model but needs more data!

- Was pre-launch marketing performed? PPC?
- How many emails were gathered?
- How many updates were made to backers before/launch?
- Average age of backer? (Know who to market to)
- How many times was the campaign shared? (fewer than 2 shares == 97% chance of failure)
- Who raised more than their goal?
- Product description
- Social following stats
- Focused model on campaigns that raised 2x-20x their goal. Why/How?
- Further experiments of campaign names and success (word difficulty, syllable count, does name explain product)
- Perform ensemble methods on each model

Questions?

Thank you!