1. Dataset Pre-processing

Invalid variables were removed. No duplicated or inconsistent inputs were detected. No N/A records were detected except for in column ‘Category’.

2. Manual Feature selection

In the dataset provided, many discrete variables should not be seen simply as numerical variable, such as weekday and month. Also, categorical variables have more than 10 categories. Even if no data exploration technique was taught in this course, I decided that it is necessary to analyze some variables with graphs in order to generate valuable predictors.

Categorical variables were analyzed using boxplots. Sample shown as Appendix 1 (100 outliers were removed for better illustration). A large number of ‘Musical’ and ‘Plays’ projects have relatively low ‘usd_pledged’ and high failing rate. Therefore, new variable ‘bad_category’ was created to substitute the original variable ‘category’. Three variables weekday-related were analyzed with the same process.

Discrete variables were analyzed using scatterplots. For example, projects that have deadline in different months have similar success number and distribution in ‘usd_pledged’. Therefore, variable ‘deadline_month’ is discarded. Note that records with value ‘True’ were transformed to 0, which is opposite to normal rules.

3. Linear Regression

First of all, sklearn model was generated for a base MSE (=1.36e+18). Then, Ridge Regression was performed with different alpha (best MSE is 5.58e+17, alpha =5,000). LASSO Regression was also performed with the same method (best MSE =2.22e+10, alpha =5,000). The increase of alpha from 1000 to 5000, however, doesn’t decrease MSE largely compared to the base MSE scale. Therefore, model needs to be further tested with ‘Kickstarter-Grading-Sample’ dataset to avoid overfitting.

4. Classification Regression

Firstly, sklearn model was generated (base accuracy score =0.65). Random Forest was used for feature selection. It selected all the numerical variables (importance >0.05), but zero categorical variables. Another sklearn model was generated using the features selected (accuracy =0.65), the selection is ineffective. RFE was then performed (accuracy =0.66), and sklearn model was generated using the features selected (accuracy =0.65). So, feature selection was discarded for classification. Other classification algorithms were also tested. Among them, CART (accuracy =0.69, ‘max_depth’ =4) and GBT (accuracy =0.70, ‘min_samples_split’ =7) showed relatively better performance.

5. Overfitting Test

Variance of optimal model alternatives were calculated in order to choose the final optimal model. For regression model, LASSO models with different alpha shows similar MSE of around 3.4e+10. Considering the performance on the original dataset, variables’ number and coefficient scales, a final alpha of 1500 was chosen. The model has 7 predictors, where the increase of ‘name_len_clean’ contributes the most to ‘usd_pledged’, and being ‘bad_category’ decreases ‘usd_pledged’ the most. For classification model, CART, GBT, and random forest methods were tested with the optimal hyperparameters gained above. GBT was chosen for it has the best performance (accuracy=0.69).

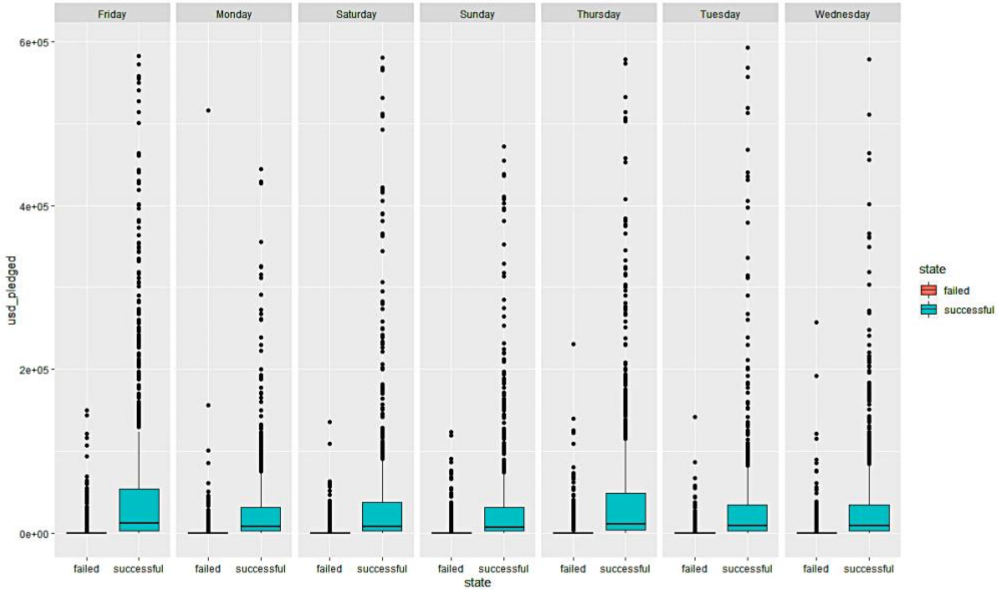
These predictive models can inform project creators on market preferences and success factors. They can also assess if Kickstarter is a good funding option for their project.

6. Clustering

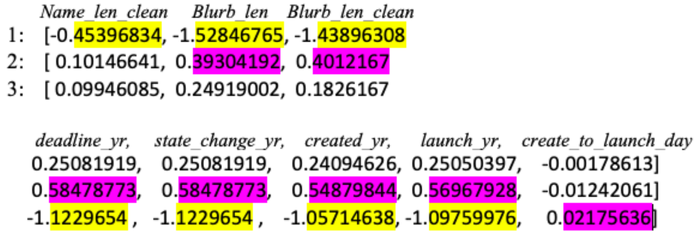
Here I analyzed feature groups for failed project. Binary variables were not considered, as Euclidean distances for (0,1)s aren't very meaningful. To start with, all numerical variables were included. Then each variable was discarded to see its effect on silhouette score. Through trial-and-error, an ideal variable set was gained. Then I chose cluster number of 3 for it presents the most reasonable separation and insights, as shown in Appendix 2 (silhouette score =0.27).

Low values are marked by yellow and high ones are pink. Each cluster shows distinct features. Cluster-1 projects have short names and blurbs. Cluster-2 projects has long blurbs, but they started funding in later years. Cluster-3 projects started funding relatively early, but they took a long time from creation to launch.

Appendix 1: Boxplot Sample: 'deadline_weekday' - 'usd_pledged', 'state'



Appendix 2: Clustering Center


```
    Name_len_clean    Blurb_len    Blurb_len_clean
1:  [-0.45396834, -1.52846765, -1.43896308
2:  [ 0.10146641,  0.39304192,  0.4012167
3:  [ 0.09946085,  0.24919002,  0.1826167

  deadline_yr,   state_change_yr,   created_yr,    launch_yr,   create_to_launch_day
  0.25081919,    0.25081919,    0.24094626,  0.25050397,    -0.00178613]
  0.58478773,    0.58478773,    0.54879844,  0.56967928,    -0.01242061]
 -1.1229654 ,   -1.1229654 ,   -1.05714638, -1.09759976,     0.02175636]
```