Joe Hernandez

Data Mining – Final Project Report

16 JUNE 2018

IST 565

Data Mining Proposal

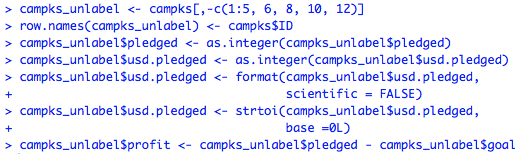
Kickstarter has been a revolutionary platform for innovative individual/group ideas using crowdfunding. Those who support the projects often receive an incentive for helping these entrepreneurs achieve their goals such as prototypes, merchandise, acknowledgements, and other means. Sometimes these projects receive sufficient funding and deliver. Other times they miss their goal and cannot deliver. The last outcome is they receive funding and do not deliver. Given such outcomes some categories of projects typically receive more funding than others. This begs the question if Kickstarter should continue attempting to support some of their unsuccessful categories. The proposal is to find the successful categories and find what criteria allowed for such campaigns to be successful.

Data

The data will consist of lists of Kickstarter campaigns in 2016. The records will be cleaned up to focus primarily on campaigns launched in 2016 with a deadline of the same year. This will consist of 50,000 rows of data. The variables this data will include: Category, Main Category, Goal, USD Pledged, State, and Backers.

Techniques

Through cluster analysis the large dataset of the 2016 kickstarter campaigns allow us to differentiate attribute types, understand data patterns, and deal with noise and outliers. In this analysis, kMeans was used to provide a top down approach to prevent overlapping. The following code provides the kMeans cluster.

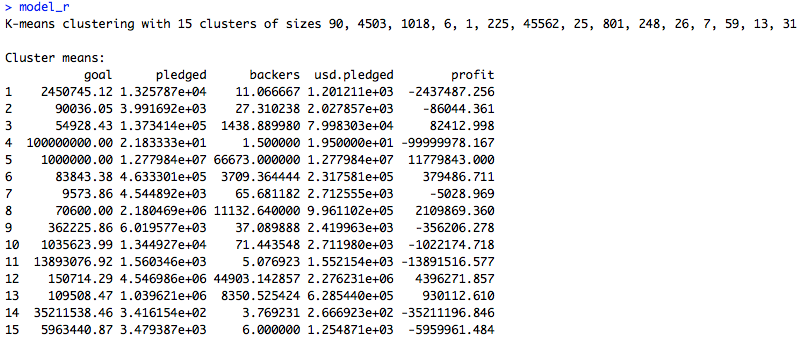


In this analysis 15 clusters were created to match the number of main categories in the campaign. The following code produced the clustering along with their assignments.

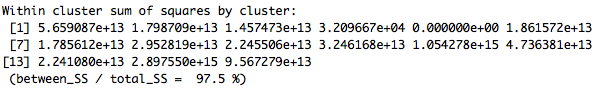




This clustering separated clusters based on the following cluster means for goal, pledged, backers, usd.pledged, and profit.



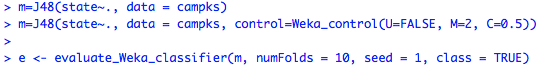
A good indicator was the high sum of square error:



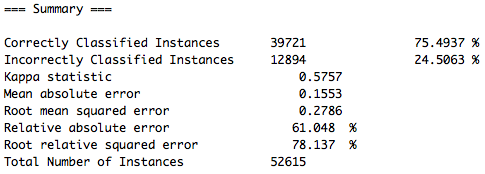
Modeling this produced the following linear plot:

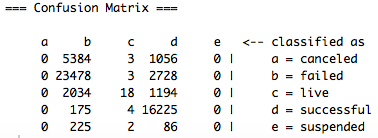


Another technique of analysis explored was the decision tree using J48.

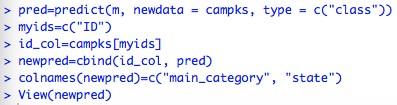


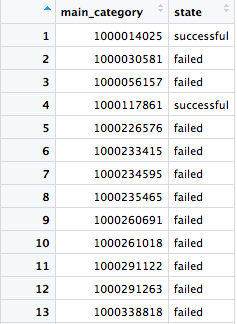
This tree diagram categorizes the data based on the success of a kickstarter program where the following provides the correctly classified instances and the following confusion matrix.





The following code creates a predictor for the classified state to the projects based on the ID code.





Summary

The overall observation of this analysis has shown an interesting perspective to analyze this data. Using kMeans allowed a way to look at data based on grouping and Decision Tree showed the appropriate campaign state classification. The prediction will be interesting to try with new input data for further observation.