**Crowdfunding Success Forecast**

Instructor: Amirhossein Gandomi

Authors: Chenkai Hang, Haihan Hu, Haodong Zhao

BIA 652 Multivariate Analysis

**Abstract**

By the rapid increase in global finance, there is an exponential growth in a large number of people that want to start a business. Crowdfunding is the most important part when starting a business, and we want to understand the crowdfunding process by applying the knowledge that we learned from this class. To be specific, the real-world application of the classification technique to predict the success of web funding project and the correlation between project features and project success rate.

Our team, composed of three BI&A master students from Stevens Institute of Technology, are interested in researching the area of Multivariate Analysis and Machine Learning Methods. The self-scraped data and the python coding package are used to create a dedicated result for this study. The goal of our research has achieved; however, the ambition of our team will go beyond the class content and reach the real finance world.

***Keywords***

Crowdfunding, Predicting, Classification, Correlation



**Introduction**

Since we are researching the funding projects on Kickstarter, we have two ways to get the data. Use the data from Kaggle that created by someone else or gather the data by ourselves. Although the data produced by others may be better organized, we decide to collect data directly from Kickstarter by ourselves.

We gathered the data by using the method that we learned from other class, using the web mining method to scrape the funding projects’ data. The data could help us understand the projects’ current progress, project created location, and many other features that relate to our research. How to process and analyze these data are the most critical steps that we need to consider during our study carefully.

Data is cleaned and reorganized before doing the analyze. We use the method of EDA (Exploratory Data Analysis) to analyze the data we cleaned, which includes visualization of the correlation between each feature and the dependent binary variable. While we step to the classification part, four classifiers are used to create data models. Linear Discriminant Analysis, Logistic Regression, K-Nearest-Neighbor, and Naive Bayes are the four classifiers. The Ensemble method is also used to create a classification model in our project. Comparison of these models is the final step in our data analyzing. The best fit model will be the model that we want to study further.

**Problem Description**

The funding projects on the Kickstarter is roughly defined in three groups; Successfully funded project, in funding progress projects, fail to be funded projects. We want to group these projects into only two groups, success or failure.

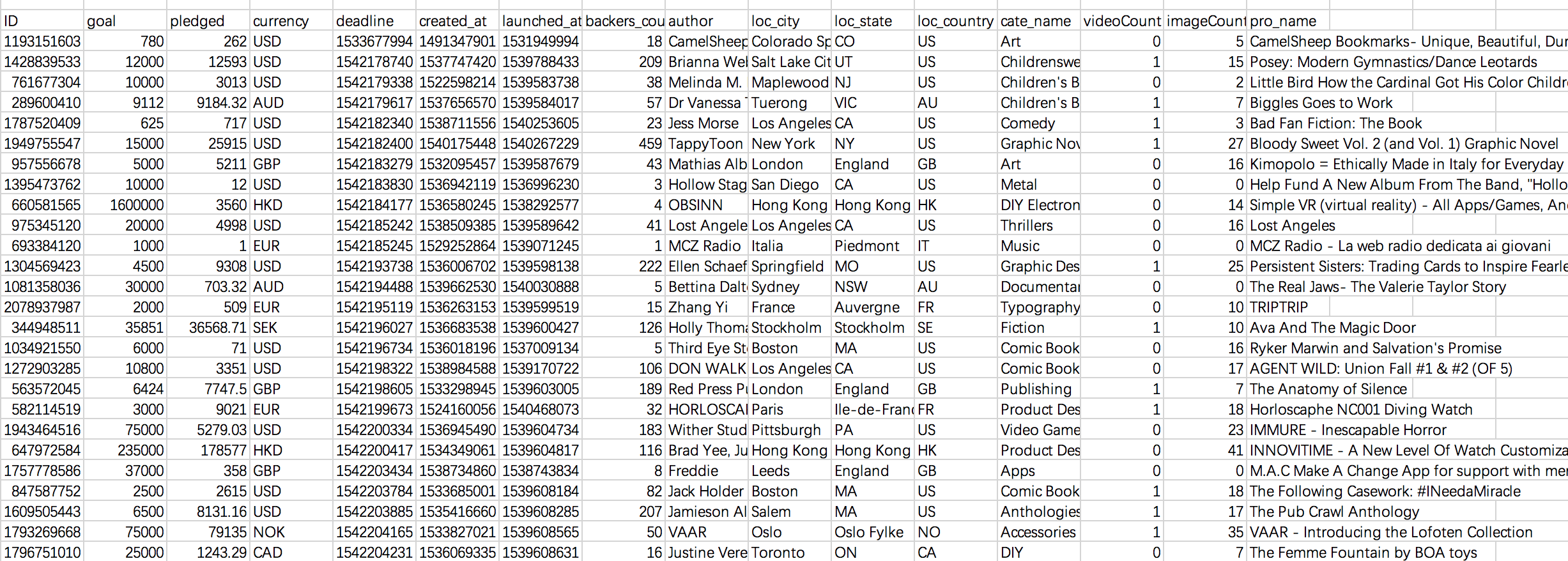
Therefore we need to determine how to define a project that is still in progress to be successful or not. Our definition of failure for those in-progress-projects is, the percent of time used for the project is greater than the current money it already gathered. Therefore, the in-progress-projects will be defined as a success if the situation is the other way around.

Another problem is which features are more critical to a funding project compare to the others. Analyzing this problem could help projects creator to decide where to put their focus on and which feature they want to show to the public the most to make their project to be a success.

**Evaluation of Database**

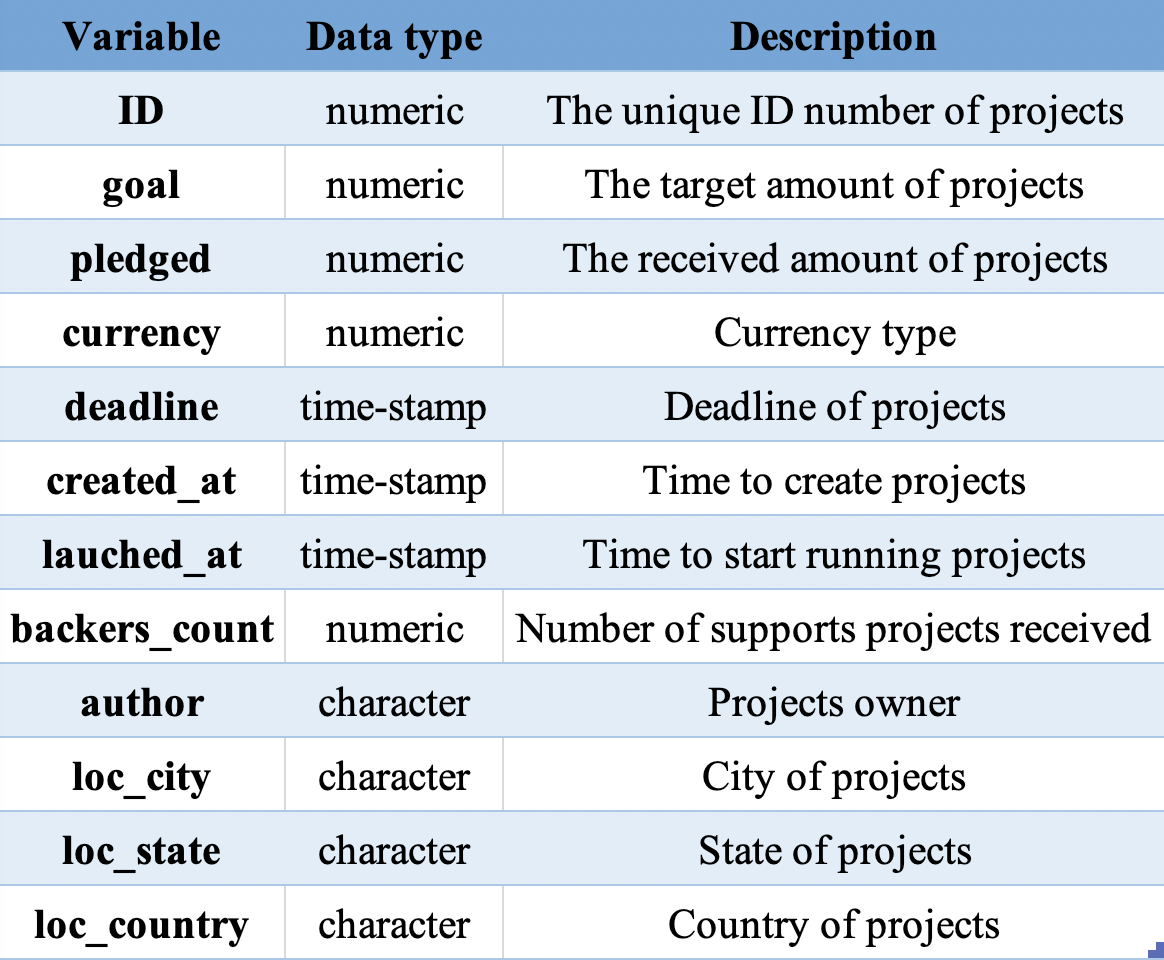
The data we gathered contains around 2400 rows and 16 columns, which means we have nearly 2400 different funding projects and 16 independent variables that could be used to analyzing and modeling.

The structure of the raw dataset is listed in ‘Figure 1’. Each project contains the following features, id, project name, goal, pledged, currency, deadline, created time, launched time, number of backers, author, city, state, category name, number of videos, number of images. A capture of the raw data is shown in the picture, ‘Figure 1’, below and a detailed explanation of each project would be given in the table, ‘Table 1’.



The Capture of Raw Data Set from Kickstarter (**Figure 1**)

Detailed Explanation of Each Variable (**Table 1**)

****

The Table shows the description of each variable and the data type of each variable.

**Data Processing and Preparation**

**Step 1:** Data Cleaning

* Check dataset and remove N/A value

**Step 2:** Data normalization

Currency

* We coded a function to convert all currencies to USD.

Time

* The original time format in our data is in timestamp format. We change the time format to YYYY-MM-DD for easy calculating.

Re-category

* In the original dataset, there are over 50 ‘loc\_country’ and 145 ‘cate\_name’. We categorize them into big categories based on scientific judgement.

Dummy value

* Since the Country and category data is not numeric, we set them to dummy values.

Data standardization

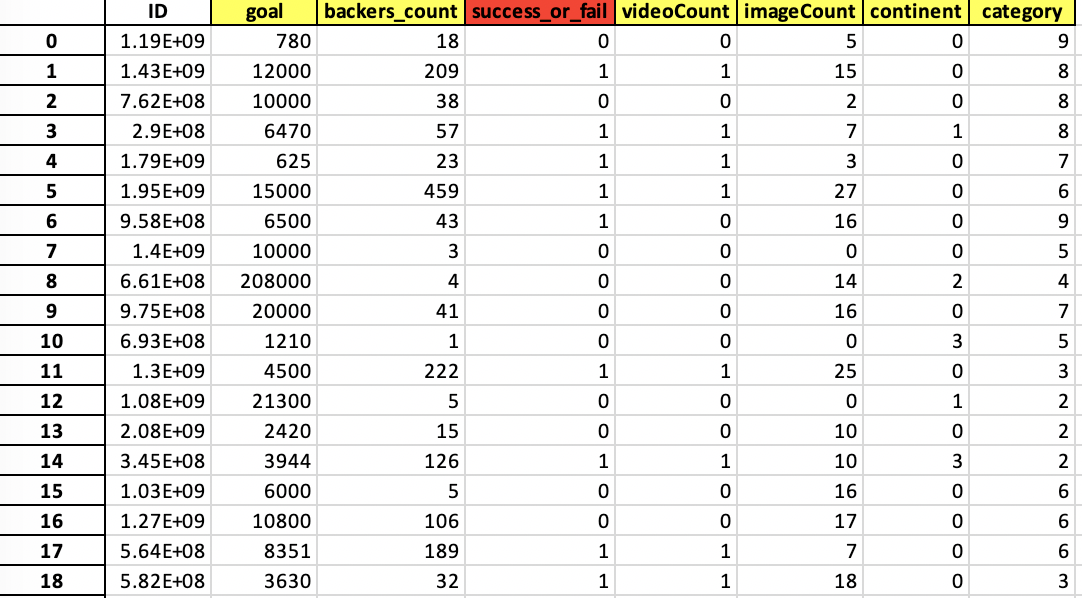
* There is a visibly big mathematical difference between ‘goal’ and other variables. We standardize the data to limit this gap.

**Step 3:** Definition of Dependent Value

* Success is defined as percentage of ‘pledged’ over ‘goal’ greater than percent of time used.

**Step 4:** Variable Selection (Detailed Show in Figure 2)

* Yellow columns are independent variables (“goal”, “backers\_count”, “continent”, “category”, “videoCount”, “imageCount”).
* Red Column is dependent variable (“success\_or\_fail ”).



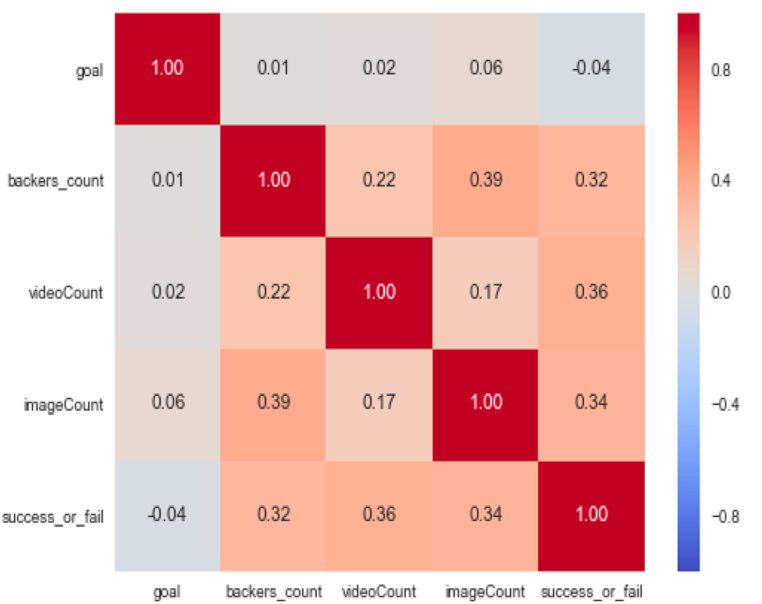
The Capture of Processed Data Set from Kickstarter (**Figure 2**)

**Step 5:** Training and Testing Dataset Split

* The data is separated into training and testing data sets.
* Training data: 75% of the data is used to develop the model.
* Testing data: 25% of the data is used to evaluate the model.

**Feature Analysis**

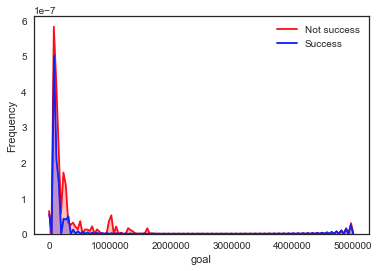
First of all, we conducted feature analysis, making the correlation matrix between numerical values (‘goal’, ‘backers\_count’, ‘video\_count’, and ‘image\_count’) and dependent variable. The correlation matrix is shown in Figure 3.



The Correlation relation of four features (**Figure 3**)

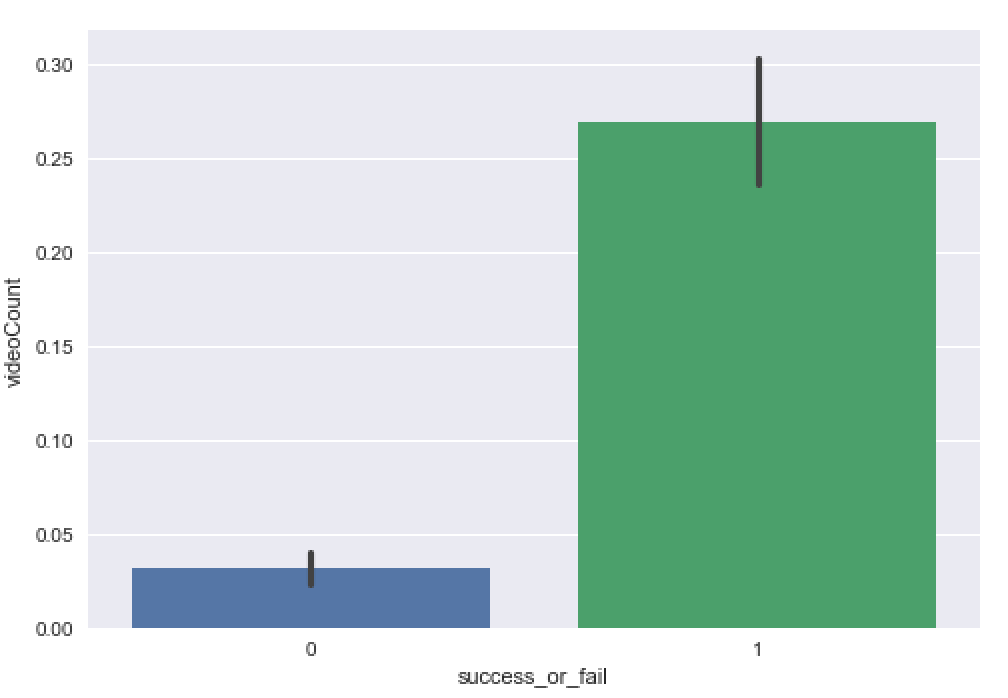
* ‘video\_count’ is the most influential feature with the highest value of 0.36
* ‘goal’ is the least important feature with the value of -0.04
* ‘Backers\_count’, ‘videocount’, and ‘imageCount’ has small difference in value. Meaning the other features is also critical to a successful project

We are going to explore the other features in the next stage of our research.

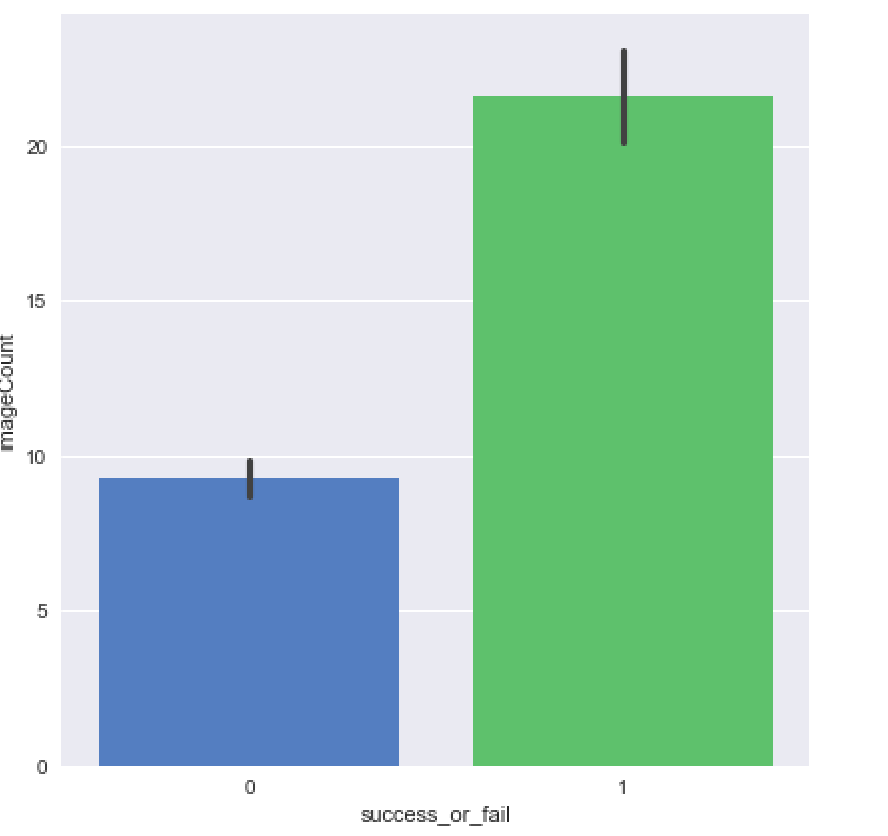


The relationship between goal and success (**Figure 4**)

Obviously, most projects have the goal under 1M dollars. The line of success and the line of not success almostly coincide, so the goal is not highly related to the results of the projects.

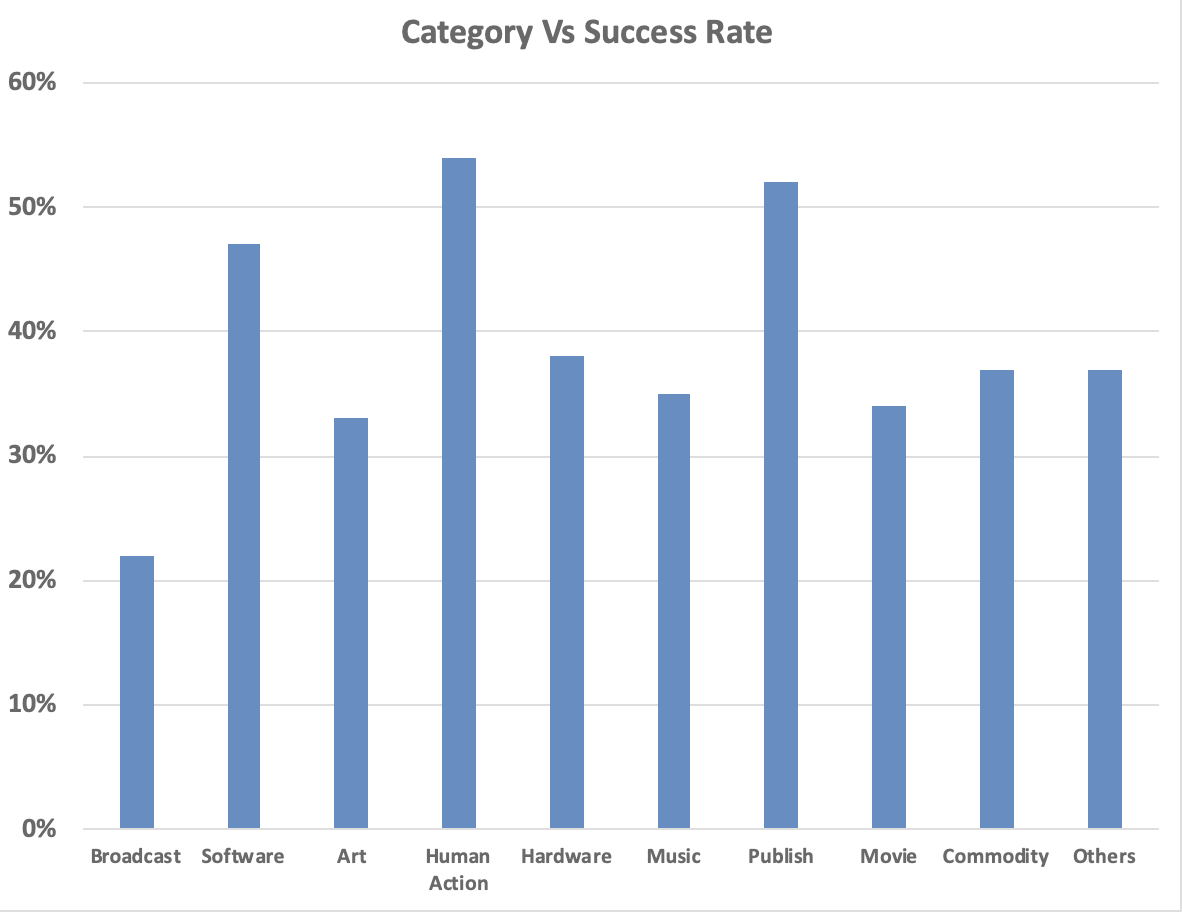


The relationship between VideoCount and Success (**Figure 5**)

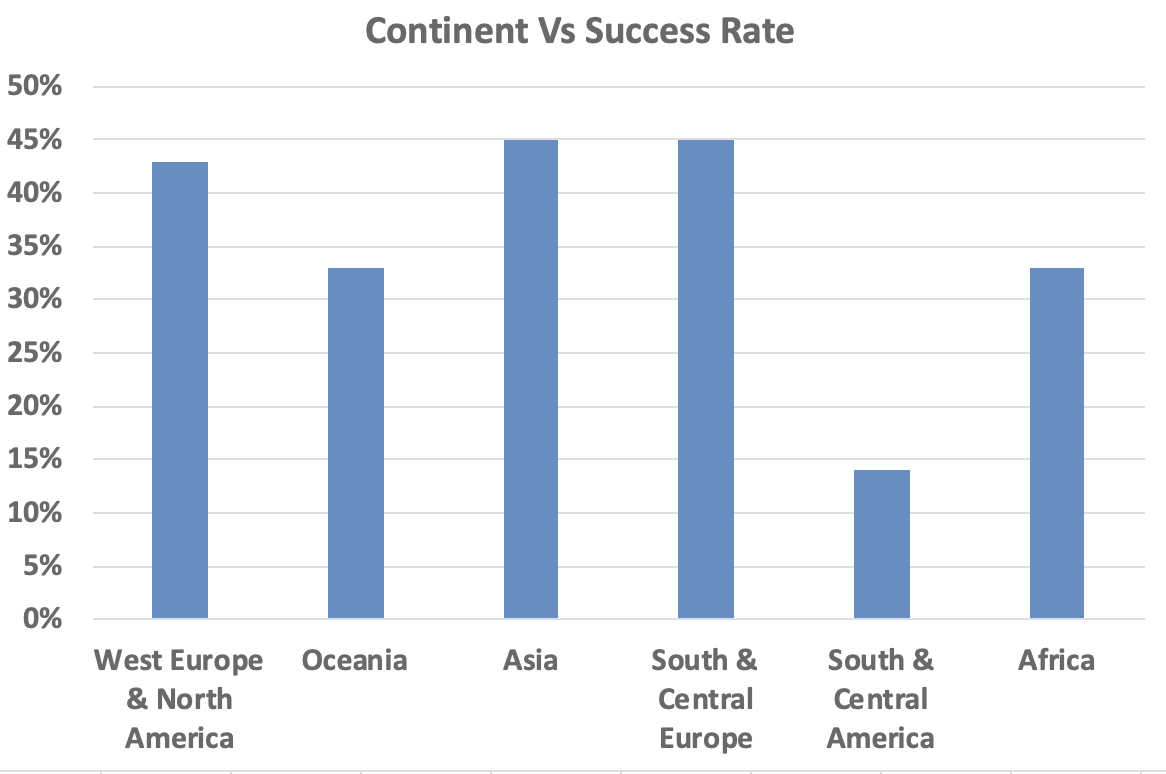


The relationship Between ImageCounta and Success (**Figure 6**)

We particularly want to see the correlation between the image, video counts and the success rate of the funding projects. Since the ‘ImageCount’ and ‘VideoCount’ have the highest and second highest correlation scores.



The Reorganized of ‘Cate\_name’ (**Figure 7**)



The Reorganized Countries (**Figure 8**)

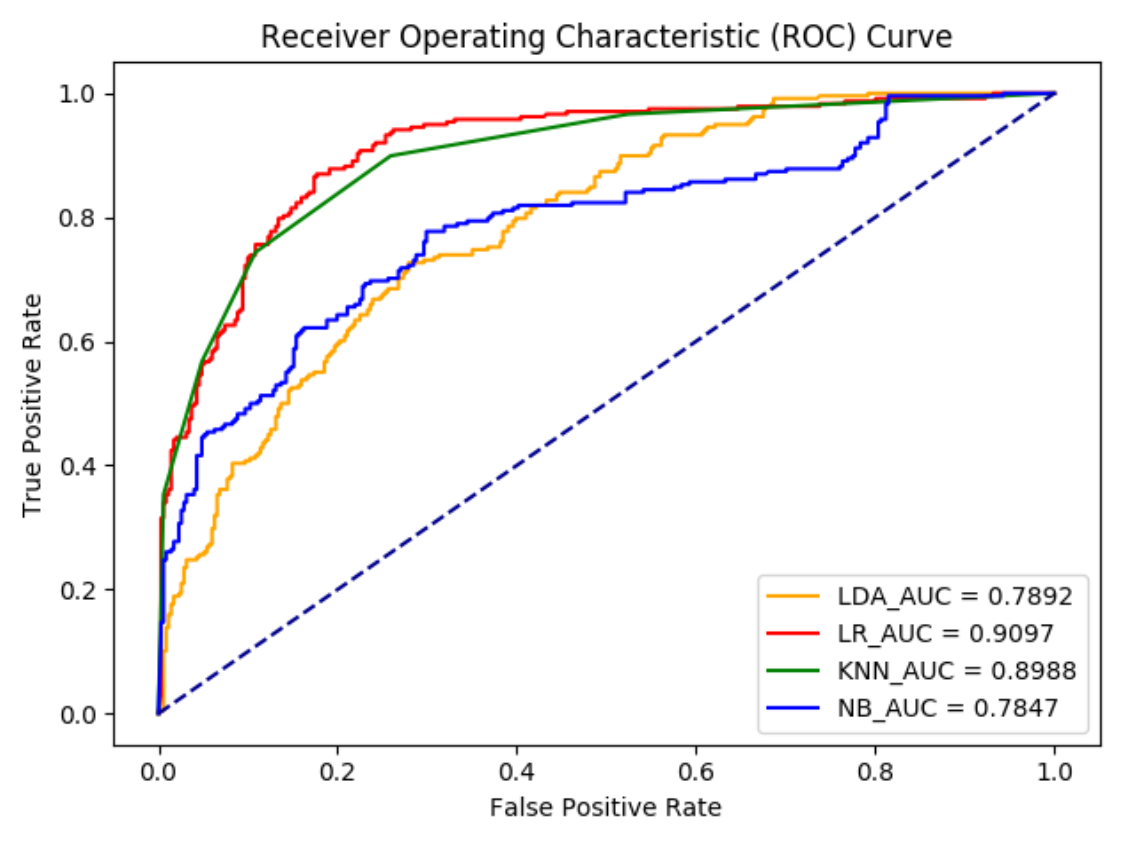
The Figure 7 shows all 10 features that we used to categorize the ‘cate\_game’. The ‘other’ features includes the ‘cate\_name’ like fantasy, and family that hard to categorize into other features. The original ‘loc\_country’ is categorized into continents in Figure 8. The separation and Combination of Europe and North America is based on the current political and economy situation.

**Modelling & Results**

After data clean up and feature selection, we selected six independent variables: ‘Goal’, ‘backers\_count’, ‘videoCount’, ‘imageCount’, ‘continent’ and ‘category’, the dependent variable is ‘success\_or\_fail’.

We successfully implemented machine learning algorithms of Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbors, and Naïve Bayes Classification.

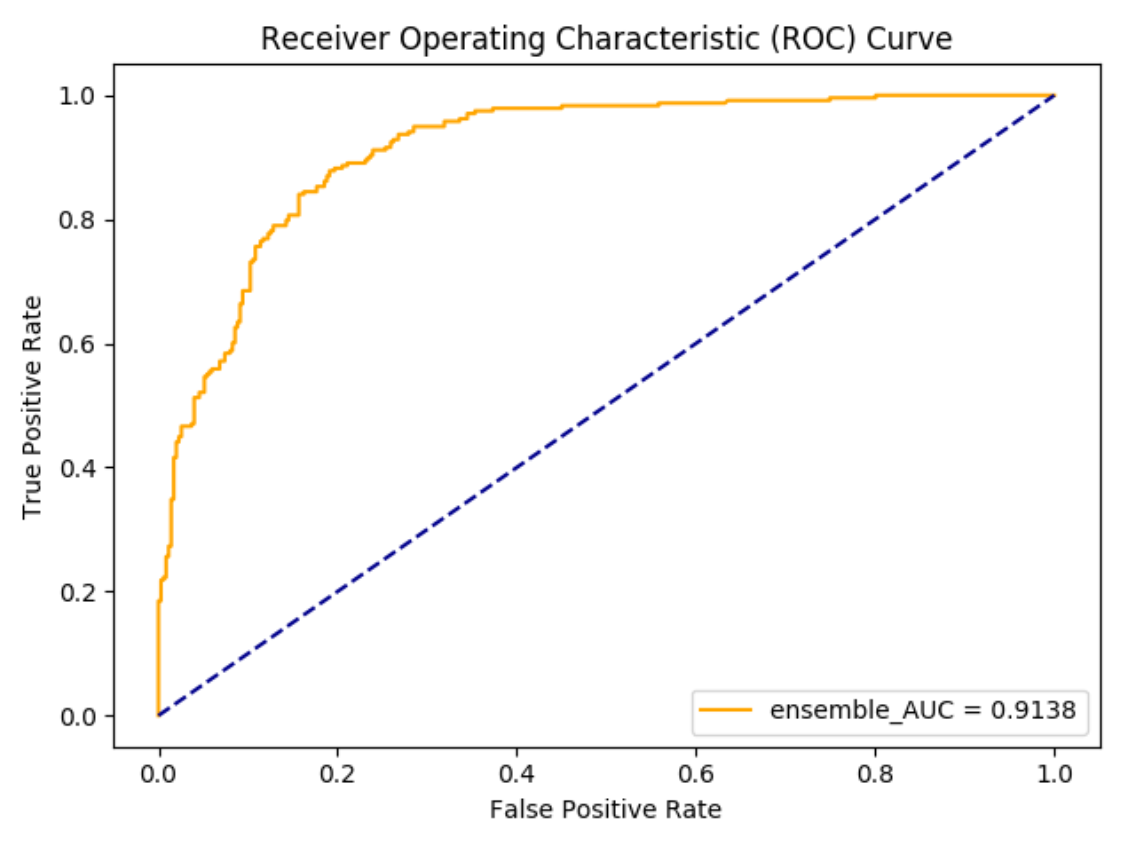
Accuracy can be computed by comparing actual test set values and predicted values. We could see how accurately the classifier or model can predict the success of projects. And we also implemented voting ensemble method to find the ensemble performance of the previous four classifiers.



The ROC curve of four classifiers without PCA (**Figure 9**)

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The larger the area under the curve, the higher the accuracy rate the model has.

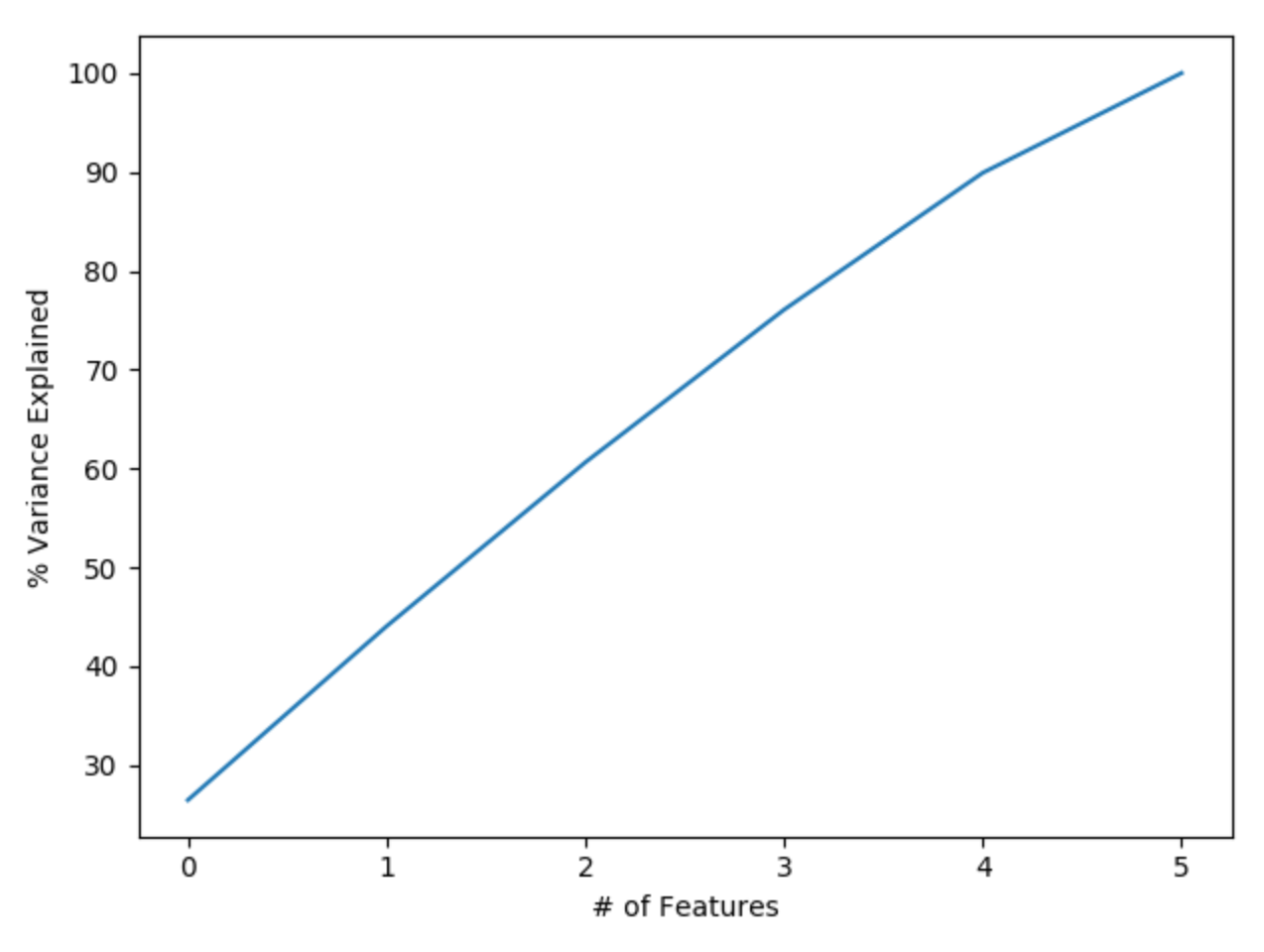
According to the ROC curve, it is clear that the LR classifier got the highest accuracy with the value of 90.37%.



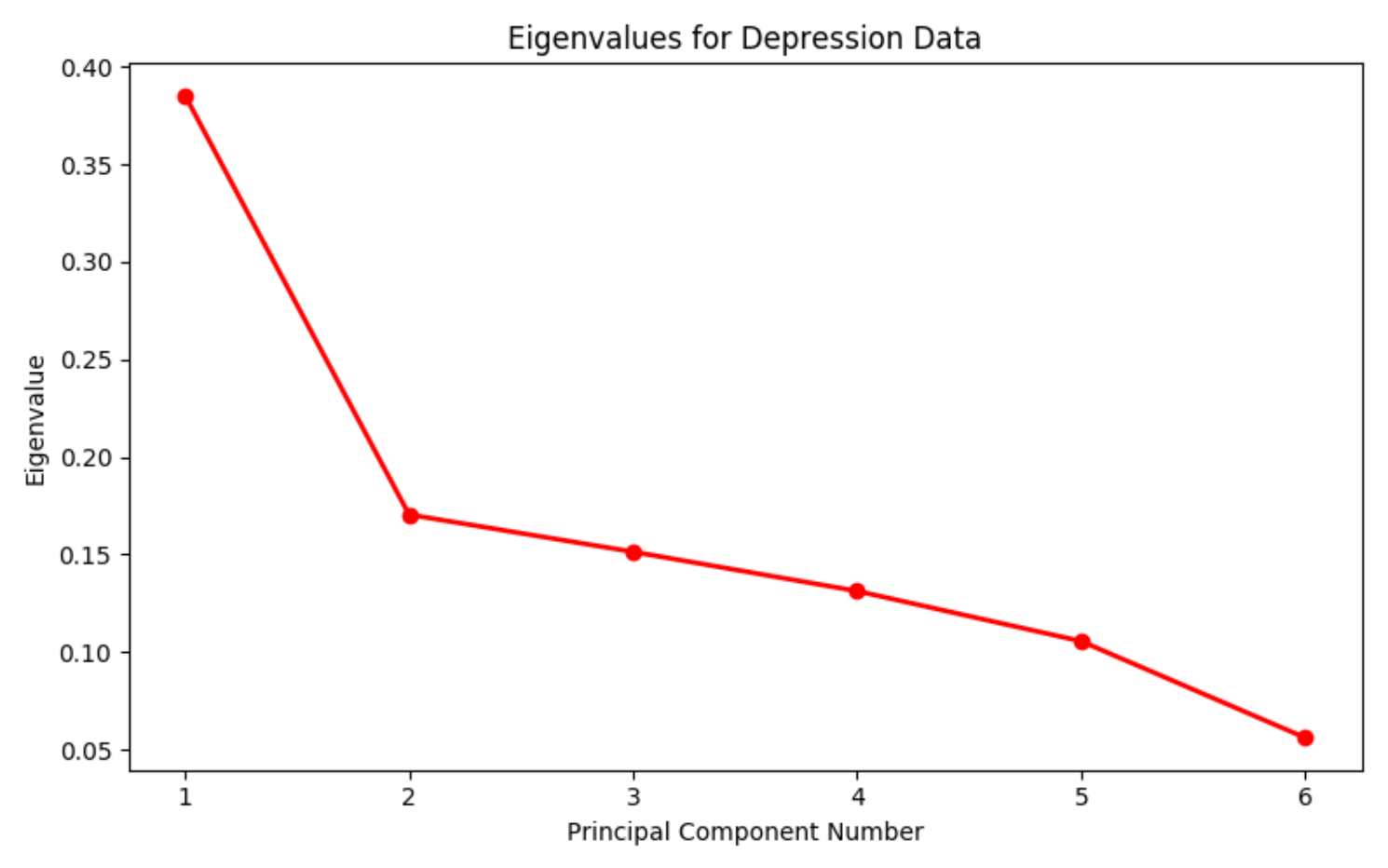
The ROC curve of Ensemble without PCA (**Figure 10**)

To make more accurate predictions, we used the voting ensemble modeling, which is a voting classifier to combine the predictions from multiple models to optimize the results. By using cross-validation, we got a classification rate of 91.38%, considered as good accuracy.

**Dimension Reduction Method: PCA**

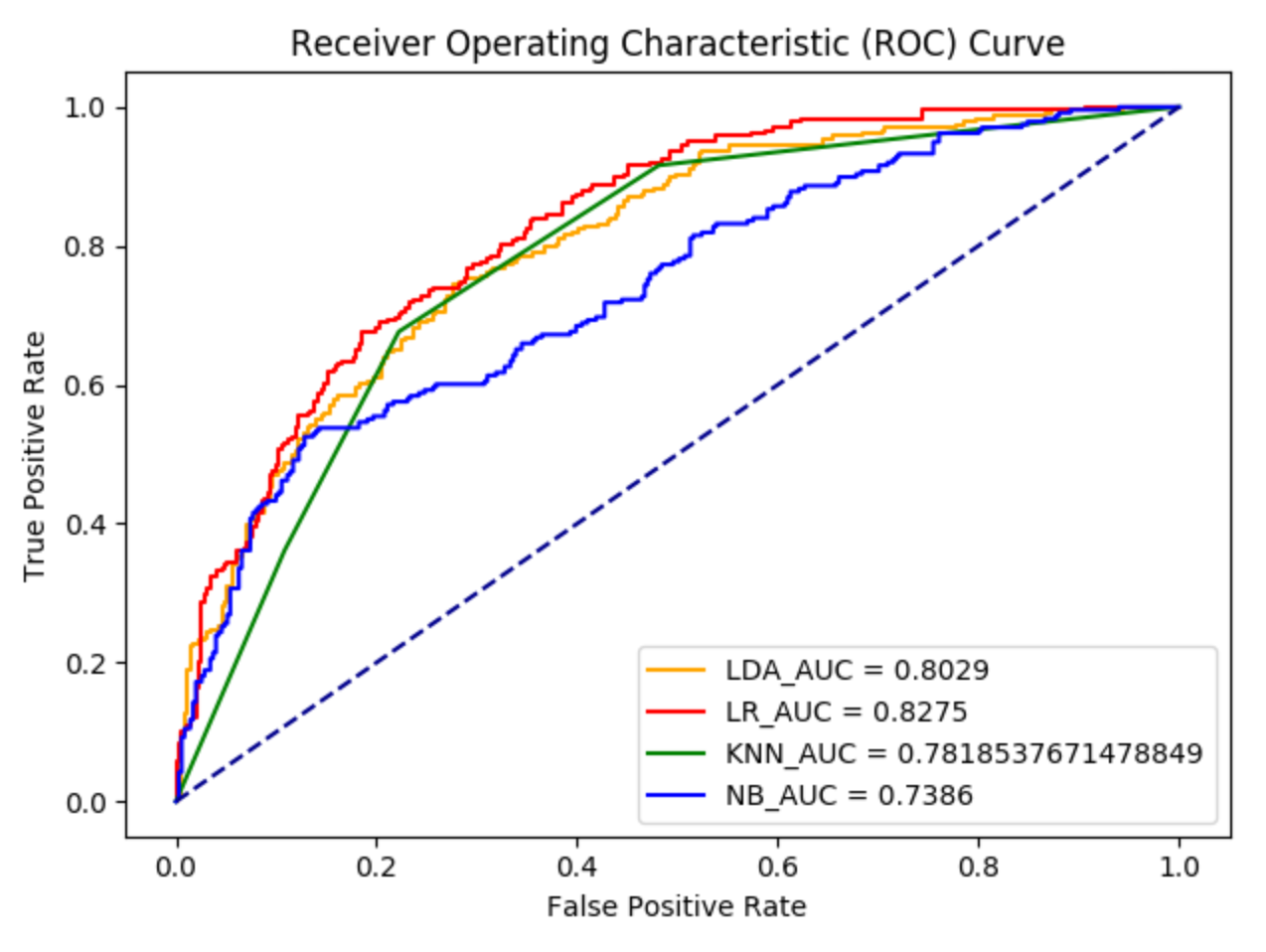


Cumulative Percentages of Total Variance (**Figure 11**)

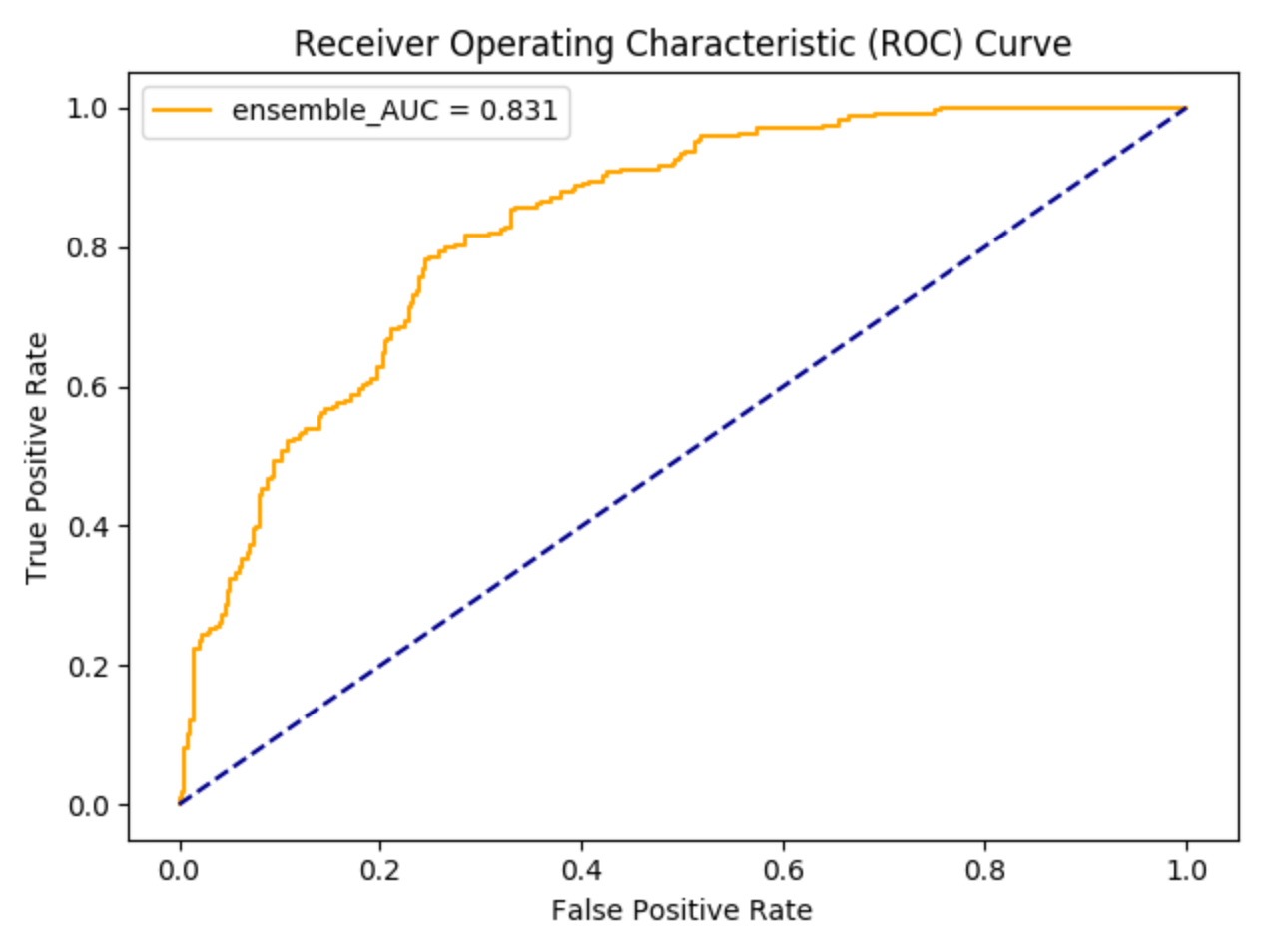


Scree plot of PCA (**Figure 12**)

PCA shows all the six components are essential. With five features we can preserve something around 90% of the total variance of the data. In the scree plot, after arranging the eigenvalues in the descending order, we can use five principal components which eigenvalues are over 0.1. In the following modeling part, we compared the results of models with six variables and that of 5 principal ones.



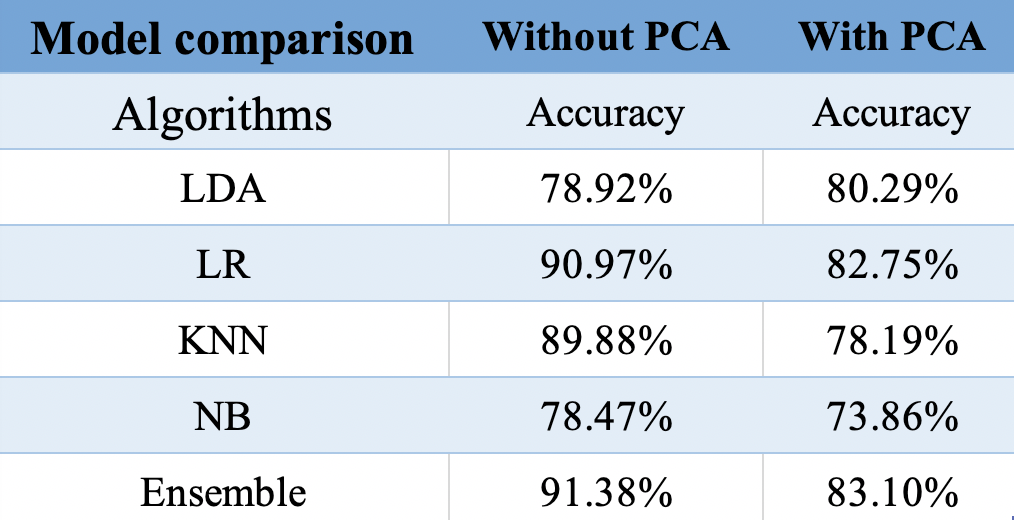
The ROC curve of four classifiers with PCA (**Figure 13**)



The ROC curve of Ensemble with PCA (**Figure 14**)

We got a classification rate of 83.1%, which is actually lower than the accuracy in our previous model with the 6 variables.

Model Performance Comparison (**Table 2**)



It is obvious that Logistic Regression algorithms have the closest accuracy as that of voting, and Logistic Regression shows the highest accuracy while Naive Bayes shows the lowest accuracy.

It seemed surprising that there was not too much improvement of the prediction accuracy of the success of the projects after we applied the ensemble method.

With PCA, the accuracy of LDA improved and that of other models decreased. It suggested that goal, backers\_count, imageCount, videoCount, category and continent are all very important features that are closely correlated to the success of the projects.

**Conclusion**

The result of our project shows, among the four classifiers, the Logistic Regression has the best AUC score (Area Under Curve). And we have an even better rating if we do the Emsebling for these four classifiers.

The reason why Logistic Regression gives the highest AUC score probably is that the nature of Logistic Regression is using one or several nominal, ordinal, and interval independent variables to predict the binary dependent variables, which is precisely what we are doing in our project.

The nature of the ensemble method is constructing a set of Classifiers and then classify new data points by taking a vote of their prediction. So the bias and the variance are decreased, and the prediction score is increased in the new model. Thus the ensemble will return us the highest AUC score.

**Future Research**

Both the dependent variables and the independent variables are defined by ourselves. For example, the success or fail is simply defined by the percent of time greater than the percent of money gathered. Another example is we categorize the “loc\_countries” and the “cate\_name” by the subject judgment of ourselves. We compare the similar data set on Kaggle by others, though the categorized features are mostly the same, the data set still has small differences. We think we could think about this more carefully and scientifically in the future.

Another thing is we only using six features to do the prediction model, and there are more features on or not directly on the Kickstarter websites that are worth researching. One important feature is the social media effects of these projects. The feature means the exposure rate of these projects to others on the website other than Kickstarter. This feature is like a web advertisement and will have a critical role in the success rate of this funding project. Due to the difficulty of doing this research in our project, we are not going to discuss it further. However, this could be a next stage study for the prediction of funding projects on the Kickstarter and the funding creator.

**References**

Ensemble Methods in Machine Learning, *Thomas G Dietterich 2013 Oregon State University,* [*http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf*](http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf)

Logistic Regression, *Wikipedia, the definition of Logistic Regression Model,* [*https://en.wikipedia.org/wiki/Logistic\_regression*](https://en.wikipedia.org/wiki/Logistic_regression)

Kickstarter Data Sets, *Kaggle by Mickael Mouille 2018,* [*https://www.kaggle.com/kemical/kickstarter-projects*](https://www.kaggle.com/kemical/kickstarter-projects)

LDA package Coding Reference, *Linear Discriminant Analysis Scikit-learn,* <https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html>

KNN package Coding Reference, *K-Nearest-Neighbor Scikit-learn,* <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

Naive Bayes package Coding Reference, *Naive Bayes Scikit-learn,* <https://scikit-learn.org/stable/modules/naive_bayes.html>