INFO 6210
Data Management and Database Design

Kickstarter Crowdfunding Recommendation Engine

Model Development

Submission Date: April 26, 2019

Members:

Name: Shaishav Shah
NU ID: 001447053

Name: Parth Rana
NU ID: 001417553

Name: Anshul Balamwar

NU ID: 001400625

Idea:

Kickstarter is one of the most popular crowdfunding platform on the internet. The aim of this project is to predict the success or failure of a Kickstarter campaign at launch time.

Abstract:

Crowdfunding is the practice of funding a project or venture by raising monetary contributions from many people. The majority of today's crowdfunding happens online through various websites and one of the most prominent is Kickstarter. The steps to start a Kickstarter project are; start a campaign, set the minimum funding goal, set reward levels, and choose a deadline. The most important aspect to know about launching a Kickstarter project is that if the project falls short of meeting its minimum funding goal, the project will not receive any fund. The projects analyzed in this project fall into one of 14 categories (Art, Comics, Dance, Design, Fashion, Film & Video, Film & amp; Video, Food, Games, Music, Photography, Publishing, Technology, Theater) and 51 subcategories. Only 55% of campaigns reach their funding goal thus it is extremely important for creators to know the factor(s) that might impact the outcome of their project before launch.

This project will take inputs from users using website and machine learning algorithms will provide various prediction / recommendations which are helpful to conduct the crowdfunding project.

Input from the Users on Website / Predictor for ML Algorithm

- Category and Subcategory of Project
- Location of the Project (City and State)
- Goal in Dollars
- Levels, Duration and No. of Update for the Project

Data Description:

We collected our data from Kaggle.com, the open online database. The data contain approximately 46000 mixed projects with limited information such as the goal, category, duration, number of comments, number of updates, levels, and duration.

Attributes:

The data comprises of dataset of 45, 957 observations and 17 columns.
Source: https://www.kaggle.com/parienza/kickstarter

1. **Project ID:** Unique id number assigned to each project in the data set.
2. **Project Name:** A list of the new project on Kickstarter.
3. **Url:** A list of individual sites containing all the information about the project.
4. **Category:** Each project is broadly classified into 14 categories like Art, Comics, Dance, Design, Fashion, Film & Video, Film & amp; Video, Food, Games, Music, Photography, Publishing, Technology, Theater.
5. **Location:** List of places where the projects are carried out.
6. **Status:** Each project had a status of either failed, successful, live, suspended or canceled. Since we are only interested in projects with status of either successful or failed, we decided to drop projects with status of live, suspended or canceled.
7. **Goal:** List of funding (in USD) each project requires according to their creators.
8. **Pledge:** contains list of funding (in USD) done for each project.
9. Funded Percentage: contains the data about funding for each project in percentage.
10. **Backers:** Total number of the people / organization who found the idea interesting and fund the project.
11. **Funded Date:** Day, date and time when a project's funding was supported.
12. **Levels:** Total number of segments in which a backer can fund the project.
13. **Reward Levels:** Amount which is assigned to each level.
14. **Updates:** Total number of updates provided for each project
15. **Comments:** Total number of comments given by the backers for the project.

Data Exploration and Preprocessing:

We collected our raw data set from Kaggle. It contained 45957 rows and 17 columns. The raw data set which we downloaded from Kaggle had many missing values. We

preprocessed our data set in three levels:

Level 1: The objective of your project was to predict whether a project is successful or a failed. We had different status in our data set like live, suspended and cancelled. We removed those rows from final data set.

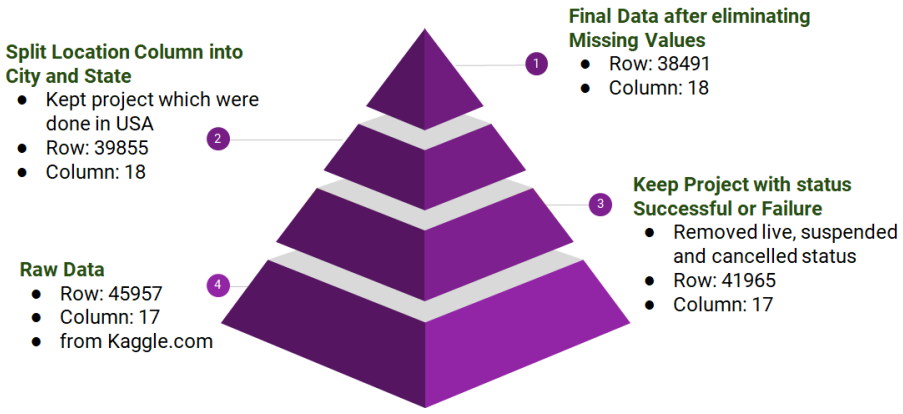
Rows: 45957 → 41965
Columns: 17 → 17

Level 2: We wanted to focus only on projects which were either successful or failure in United States of America. Keeping that in mind, we split or location column into City and State and kept only those states which belonged to USA.

Rows: 41965 → 39855
Columns: 17 → 18

Level 3: We had many missing values in our data set. Since our data set is comparatively big in size, we decided to remove those records from our final data set.

Rows: 39855 → 38491
Columns: 18 → 18



Exploratory Data Analysis:

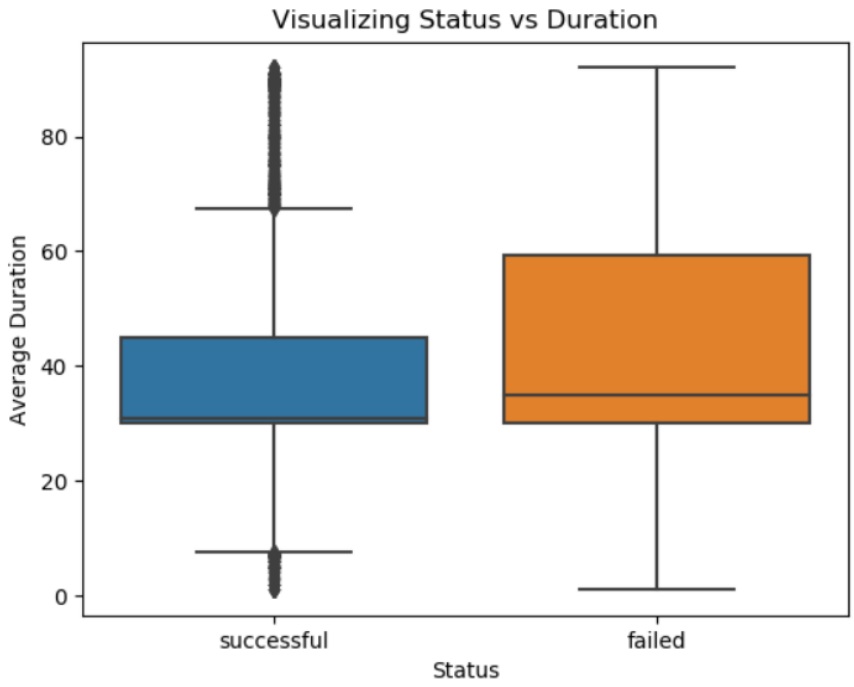
After preprocessing your data set, it is very to explore the data set for better and accurate analysis. This helps to understand questions like what features are important and what will be their contribution towards final prediction.

Exploratory Data Analysis (EDA)			
	Successful	Failed	Total
Total Projects	22869	18996	41865
Project Proportion (%)	54.73	45.27	100
Project Total Goal ($)	125.95M	310.59M	436.55M
Project Goal Average	5483.82M	16350.59M	16350.59M
Amount Pledge ($)	197.96	17.05	215.01
Average Pledge Amount ($)	8656.35	897.34	9553.69
Goal vs Pledge (%)	157.16	5.48	162.64
Average Number of Updates	6.687	1.496	8.183
Average Duration of Project (Days)	37.96	42.97	80.93
Average Number of Backers	119.37	12.56	131.93
Average Number of Reward Levels	8.49	7.29	15.78

Data Visualization:

Statistical analysis can be carried out to understand correlation of success with the funding goal, duration and engagement on the page. Combination of goal, success/failure and project category can be visualised by plotting scatter plot. Performance of crowdfunding can be related with geographic location of each project using geographic map. These basic statistical approaches and visualization methods can give basic understanding of data.

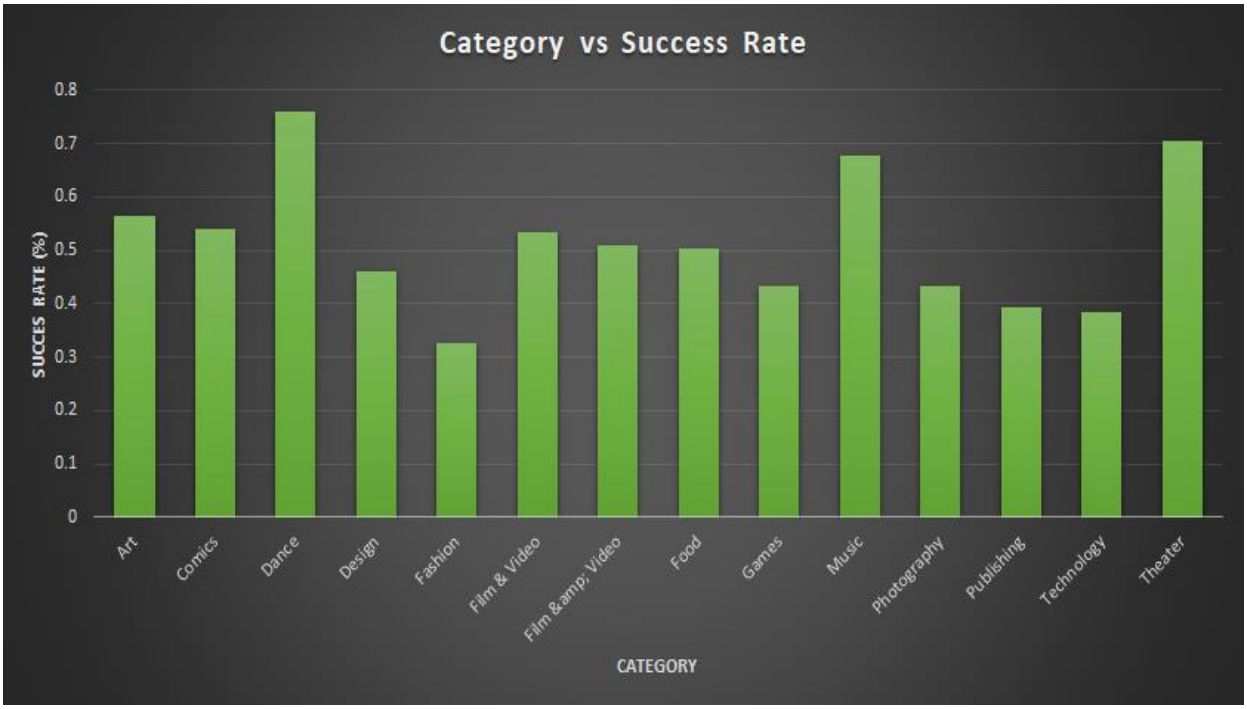
Creating box plot to find the variance in duration of the project:



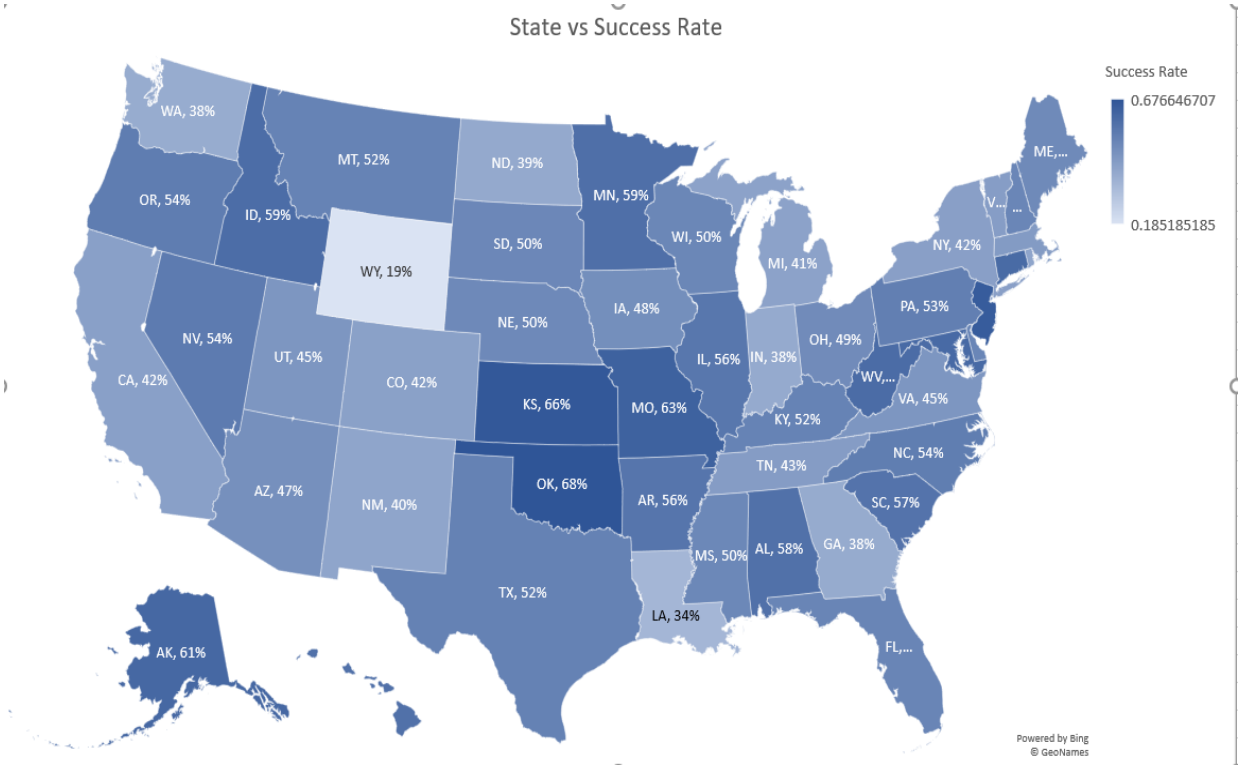
- Average duration for a successful project: 38 Days
- Average duration for a failed project: 43 Days

Creating a histogram to find to which category had the maximum success rate:

- Categories with maximum success rate: Dance, Theatre and Music.
- Categories with Minimum success rate: Fashion, Publishing and Technology.

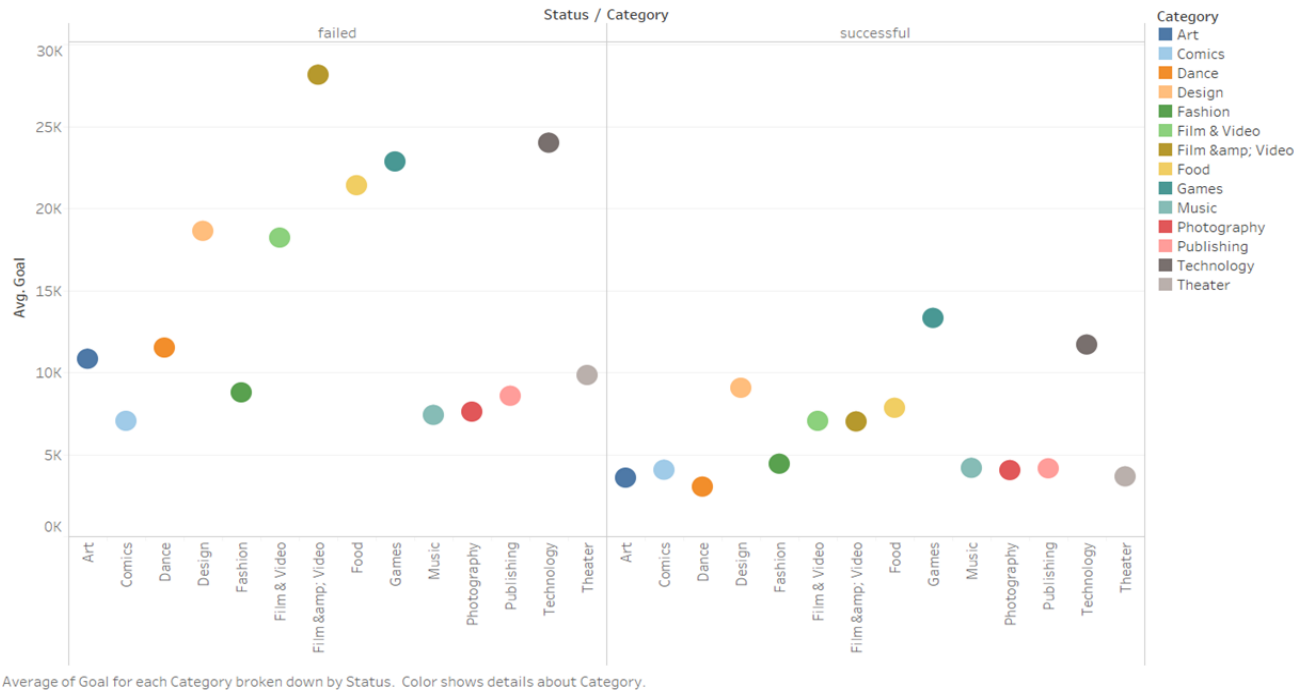


Creating a histogram to find to which state had the maximum success rate:



- USA States with maximum success rate: OK - 68%, KS - 66%, NJ - 65%
- USA States with minimum success rate: DC - 40%, LA - 34%, WY - 18.5%

Creating a Scatter plot to find to average goals for a successful project and failed project in different categories:



- Maximum successful projects in a category have an average goal between $5000 to $7500.
- Category with maximum average goal for a successful project is Games.
- Failed projects in a category have more average goals than successful projects.
- Category with maximum average goal for a failed project is Film & amp and technology.

Model Assumptions:

1. All the data present in the data set is collected from a reliable source and is an unbiased data set.
2. We have converted categorical data into numerical data set for classification.
3. We have considered that the person who will post their projects will mention number of updates he will provide during his project to backers.

Model Development:

1. Divided the data set in 75% to 25% training and testing data set.
2. Implemented Random Forest, Logistic Regression and K- Nearest Neighbour Classification models.
3. Evaluated the results obtained by creating confusion matrix, finding their accuracy, error, F1 score, their classification report and ROC curve.
4. Carried out feature selection to improve the accuracy of the models.

Model 1: Random Forest:

1. We implemented random forest model without feature selection, considering all the necessary features for the models.
2. We used Gini Index to measure the impurity in a node.
3. We used 500 Decision trees to train our data set.
4. Accuracy obtained was 80.2% with an error of 19.8 %.
5. Some of the metrics obtained are listed below:
    a. F1 Score: 0.834
    b. Precision: 0.82
    c. Sensitivity: 0.85
6. Confusion Matrix obtained is given below:

    3330 1026

    764 4503

Model 2: Logistic Regression:

1. We implemented logistic Regression model without feature selection, considering all the necessary features for the models.
2. We used Liblinear Solver to find the weights and L2 penalty for regularization.
3. We also set the inverse regularization parameter to 1.
4. Accuracy obtained was 78% with an error of 21.9 %.
5. Some of the metrics obtained are listed below:
    a. F1 Score: 0.803
    b. Precision: 0.82
    c. Sensitivity: 0.79
6. Confusion Matrix obtained is given below:

    3407 936
    1052 4228

Model 3: K Nearest Neighbour:

1. We implemented K Nearest Neighbour model without feature selection, considering all the necessary features for the models.
2. We used 7 number of neighbours and uniform weights for them.
3. We used Euclidean Distance Metric to calculate the distance.
4. Accuracy obtained was 63.5% with an error of 36.5 %.
5. Some of the metrics obtained are listed below:
    a. F1 Score: 0.683
    b. Precision: 0.65
    c. Sensitivity: 0.71
6. Confusion Matrix obtained is given below:

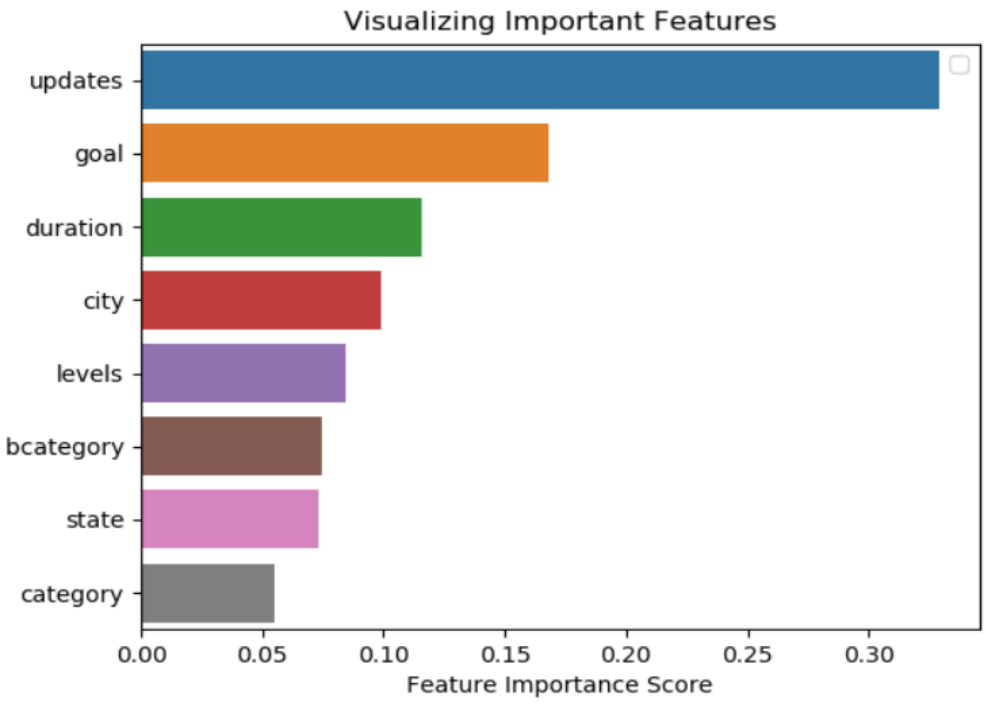
    2355 1994
    1515 3759

Feature Selection:

Feature Selection is a technique which is used to find the important variables or features which contribute towards or accurate prediction of the model.

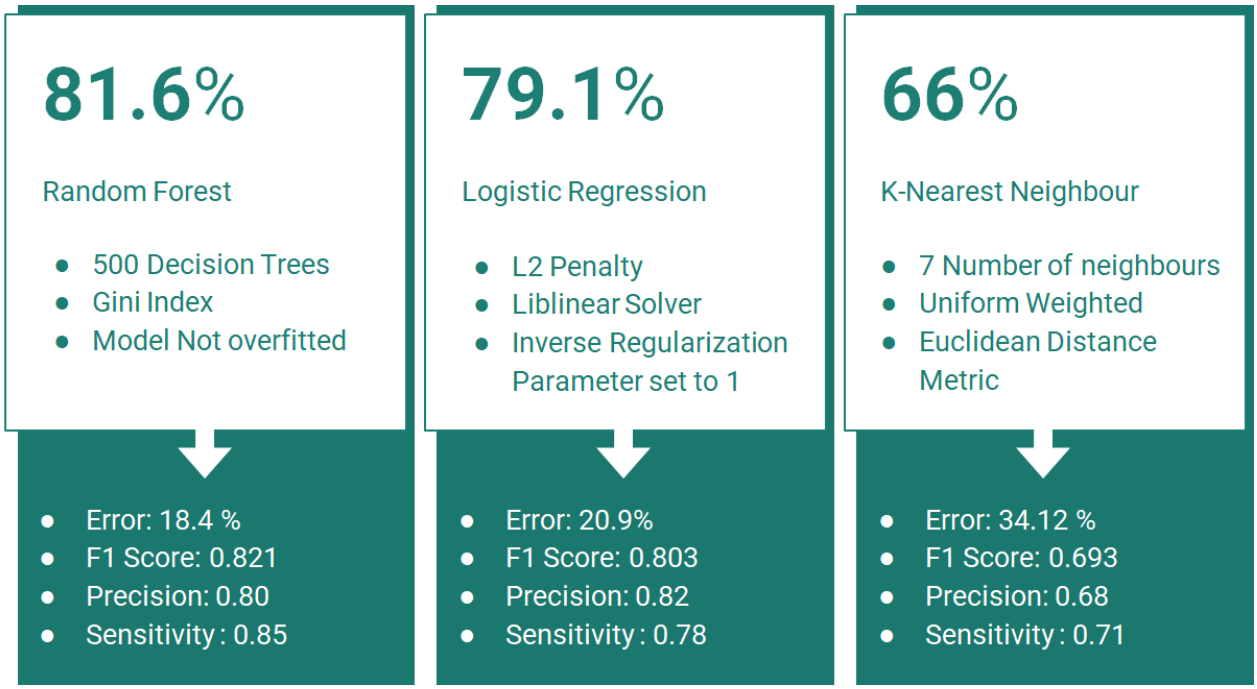
After performing feature selection and plotting a bar plot, we can conclude that following attributes were important for classification prediction in descending order:

1. Update (Maximum)
2. Goal
3. Duration
4. City
5. Levels
6. Subcategory
7. State
8. Category (Minimum)



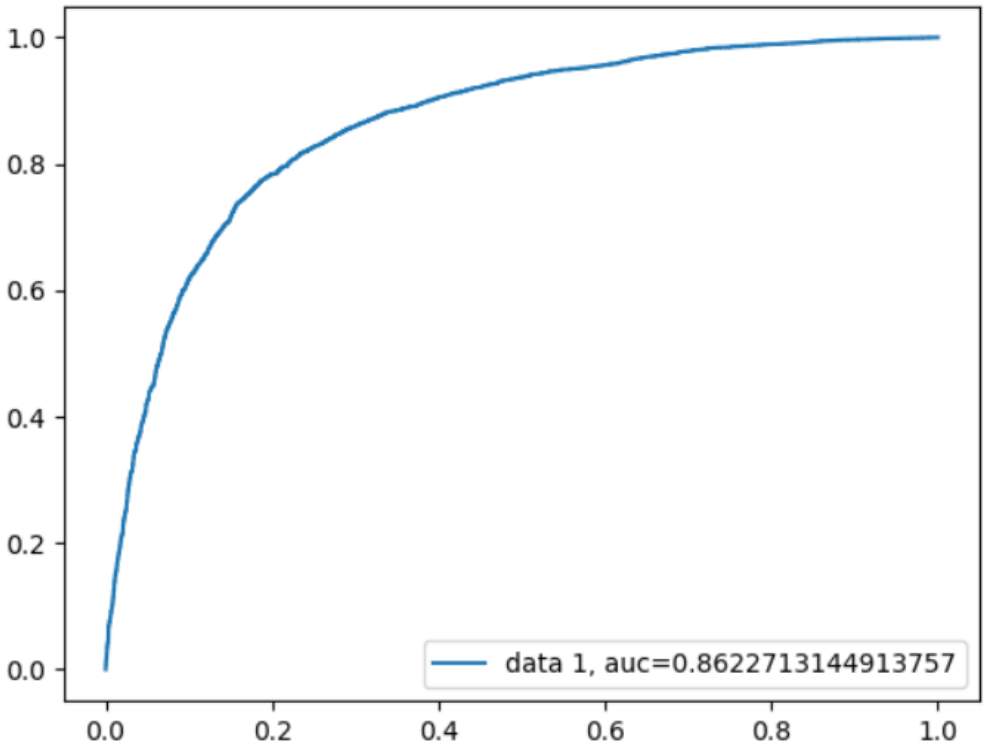
Model Development after feature Selection:

After performing feature selection, following results were obtained keeping the basic parameters same for each of the three models:



ROC Curve for Logistic Regression:

Area under the curve generally provides the accuracy of the model to separate the two classes (in our model: Successful and Failed Projects). We plot the ROC curve for logistic Regression and found that the model separates the two classes with 86.22 % accuracy.



Conclusion:

- Out of the three models, Random Forest Model proved to be the best classifier followed by logistic Regression.
- Update, goal, duration and city were the most important features for the classification models.
- Oklahoma had the best success rate while Oregon, New Hampshire and Connecticut had maximum projects.
- Maximum Successful projects were carried out in category dance, theatre and music.
- Projects having a average goal between $5000 to $10000 have high chances of becoming successful.
- Projects providing regular updates have better chances of becoming successful.
- Successful projects are generally completed in around 38 days.

Impact of the Project Outcomes:

This analysis can give prediction regarding success or failure of a particular crowdfunding of a project. By changing parameters of the crowdfunding, the possibility of success can be increased. If crowdfunding is successful in initial stages of project then I can give a great push in the project growth. So these predictions are very helpful to the people who want to convert their imagination into reality with the use of crowdfunding.

Future Prospects:

- Developing interactive website which will help users, and will increase chances of success of their crowdfunding campaign.
- Recommendations for the fail predicted projects can be given by analysing similar successful projects. These recommendations can increase the chances of success of the crowdfunding.
- Recommendations for the hashtag for social media campaign can be given by analysing description and using social media API like Tweepy, Instaloader, Google Trend, etc.

References:

- **Kickstarter** (https://www.kickstarter.com/)
- **Predicting the success of Kickstarter campaigns (**https://towardsdatascience.com/predicting-the-success-of-kickstarter-campaigns-3f4a976419b9**)**
- **Kaggle (**https://www.kaggle.com/parienza/kickstarter**)**