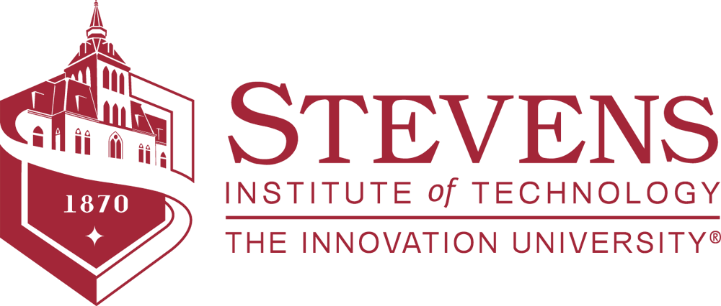
****

**Success Rate Prediction for Crowd Funding**

**BIA-678 Big Data Seminar**

**Instructor:**

**Dr. David Belanger**

By

Group 8:

Ritwin Kudumula

Suguna Bhargavi Bontha

Sucharitha Batchu

Haodong Zhao

Table of Contents

**Introduction3**

**About the Data5**

**Modelling and Results8**

**Efficiency and Run Time10**

**Run Time comparison using graphs11**

**Results13**

**Conclusion14**

**Future Work14**

**References15**

1. **INTRODUCTION**

**Background:**

Artists, musicians, filmmakers, designers, and other creators, everyone needs the right resources and support they need to make their ideas a reality. Till date, tens of thousands of creative projects — big and small — have not been able to come to life because of a lack of support. Luckily, the bank isn’t the only option these days. Crowdfunding sites abound on the internet which are very helpful. Many of our friends, and acquaintances may only be able to give small amounts, but if we can get a big enough crowd of them to donate, we can reach the funding goal. The problem is figuring out how to reach enough people without spending a lot of money on publicity and advertising. Crowdfunding websites solve this problem by offering a simple way to create attractive online fundraising pages that we can publicize through our email lists, Facebook pages, blogs, or any other social networks. We can also ask your friends to publicize our campaign through their networks. The more people we can ask, the more likely it is we will reach your goal. Crowdfunding can be used for either for-profit or non-profit projects. If the project has non-profit status, then one can offer his donors a charitable deduction on their income taxes for their gifts. If one does not have non-profit status, one cannot offer tax deductions, but he can offer other rewards or incentives.

Kickstarter is one such crowdfunding platform for all creative projects. No one will be charged for a pledge towards a project unless it reaches its funding goal. This way, creators always have the budget they scoped out before moving forward.  Doing A Crowdfunding Campaign involves Define the Project, Set the Funding Goal, create a Crowdfunding Web Page, Ask as Many People As Possible to Give Money and Spread the Word, Keep Asking Until the Goal is reached. If a project has met its goal, supporters’ credit cards are charged, and the project creators receive all the money raised (minus the credit card fees).  If a project has not met its funding goal, no one’s credit card is charged and the project receives no money. If the project is funded, the project’s creator is responsible for delivering the rewards pledged to supporters and finishing her project.

Kickstarter offers several ways for project creators to monitor their campaigns, including a real-time stream of comments, pledges, and other project activity; a list of all reward pledges; a message center to communicate with supporters; transaction history; and the ability to edit the project page and post updates for supporters as the campaign proceeds.  In addition, Kickstarter interfaces with Facebook, which can help simplify the campaign. Thus, they help bring creative projects to life.

1. **Objective:**

In this paper we took Kickstarter dataset from Kaggle and performed different analysis on the data as mentioned below as the analysis will be useful to both the users and people who are crowd sourcing. Users will be able modify their project based on factors influencing and funders will be able to make a decision on funding.

1. We use the method of EDA (Exploratory Data Analysis) to analyze the data we cleaned.
2. We step to the classification part; four classifiers are used to create data models. They are Logistic Regression, Decision Tree, Random Forest and Gradient Boost Tree.
3. The best fit model among the four classifiers is determined. The best model is used to predict if a project will be successful or not.
4. Analyzing this problem will be useful to both the users and people who are crowd sourcing. Users will be able modify their project based on factors influencing and funders will be able to decide on funding.
5. The four classifier models are run initially on local machine however since the data is very huge (about 0.2 million data rows) we later ran the models on cloud. Cloud is preferred because it overcomes the following short comes of running huge data on a local machine: Low hardware utilization, Lack of multi-tenancy support, No self-serve model, Slow onboarding new applications/users, Low bandwidth network, High OPEX, Lack of big data skills and expertise
6. Finally, we compared the results achieved on cloud to those achieved on local machine.

We achieved running the models on local machine and cloud using PySpark language.

1. **PySpark in brief:**

PySpark is the collaboration of Apache Spark and Python. It is a Python API for Spark that lets us harness the simplicity of Python and the power of Apache Spark in order to tame Big Data. Like spark, it is an open source data processing framework for performing Big data analytics on distributed computing cluster. The core data structure is RDD (Resilient Distributed Dataset), this enables it to run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk. PySpark is a great language for performing exploratory data analysis at scale, building machine learning pipelines, and creating ETLs for a data platform. There are several different options for getting up and running with Spark as follows:

* Self-Hosted: We can set up a cluster yourself using bare metal machines or virtual machines Example, Apache Ambari
* Cloud Providers: Most cloud providers offer Spark clusters: AWS has EMR and GCP has DataProc.
* Vendor Solutions: Companies including Databricks and Cloudera provide Spark solutions, making it easy to get up and running with Spark.

In this project we ran classifications in PySpark on cloud using databricks.

1. **ABOUT THE DATA**

**A. Data Source and Columns:**

We obtained the dataset from Kaggle. It is a csv file and contains around 200000 rows and 20 columns.

Following is the brief explanation of the attribute in the Dataset:

ID: internal Kickstarter ID

name: name of project

currency: currency used to support

main category:  albums, books, or films etc.

subcategory: comic, thriller

launched at: the date the project was launched

duration: Duration required to raise funds for the project

goal\_usd: The funding goal is the amount of money that a creator needs to complete their project.

City: city pledged from

state: state pledged from

country: country pledged from

blurb\_length: length of the description of the project

name\_lenth: number of words in the name of the project

status: success or fail

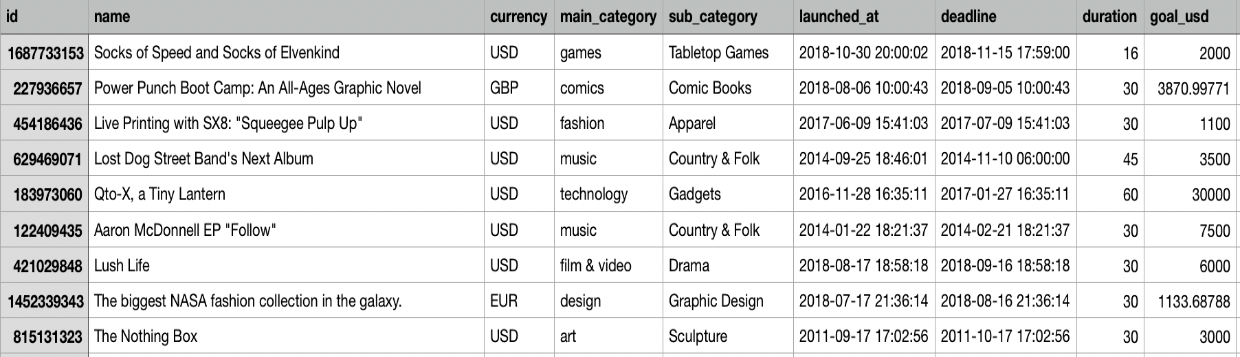
start\_month: the month the project was launched

end\_month: the month the project was completed

start\_Q: the quarter the project was launched

end\_Q: the quarter the project was launched

usd pledged: amount of money pledged

****

****

Figure 1: Data File Snippet

**B. Data Understanding**

Exploratory data analysis was performed on the data by finding the correlation between each factor. The factors with less correlation with the independent variables and high correlation with the dependent variable “Status” were remained for further analysis. The rest of the variables that did not match this criterion were removed from our research study. Variables that satisfied the criteria and that are thus used for the classification are listed as follows: Duration, goal\_usd, blurb\_length, name\_length, start\_month, end\_month, Category, Country, StartQ, EndQ and Status .Status is the dependent variable and the rest are the independent variables for the classification and regression models developed later.

**C. 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 countries and 145 states. We categorize them into big categories i.e., continents 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.

**D. Modelling**

We successfully implemented machine learning algorithms of Logistic Regression, Decision tree, Random Forest and Gradient Boost Tree. 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 thus, the most accurate model was described as the best fit model.

**3. MODELLING AND RESULTS**

As discussed previously we used machine learning algorithms to predict if the project is a success or not. They are as listed below:

**A. Algorithms Used:**

* Logistic Regression
* Decision tree
* Random Forest
* Gradient Boost Tree

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.

**B. ROC Curve:**

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.

True positive rate is also known the sensitivity or recall which is the probability of detection in machine learning. The false positive rate is known as the fall-out which is the probability of false alarm (False alarm is when we predict the project to be success but the project actually is a fail).

**A close up of a map

Description automatically generated**

Figure 2 : ROC Curve

The ROC curve obtained is as shown above. According to the ROC curve, it is clear that the Random Forest classifier got the highest accuracy with the value of 74.81%. Gradient Boost also has nearly the same accuracy and thus even this model is equally efficient in prediction

**C. Cloud service used for the classifications**

**Databricks**

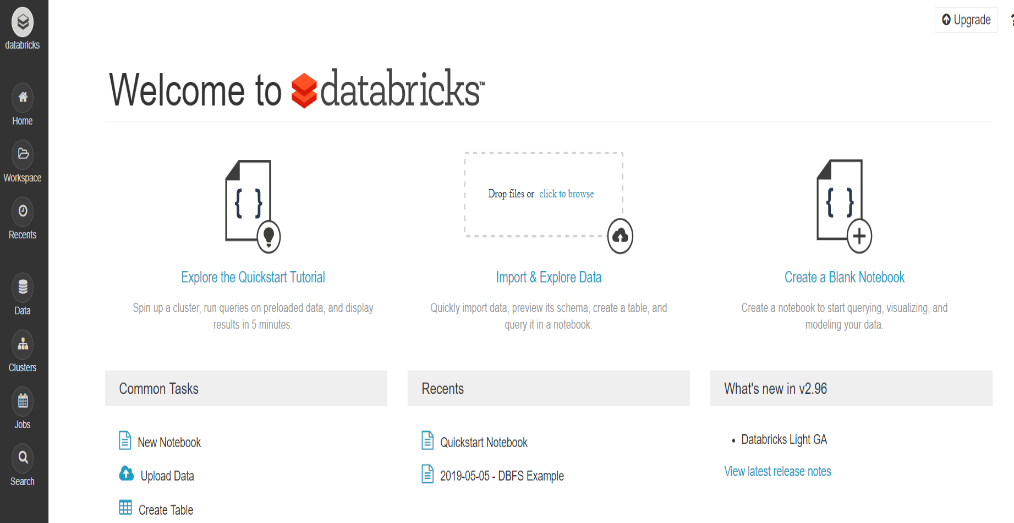
There are different options for running spark on cloud- Self Hosted, Cloud Providers, Vendor Solutions. Databricks is a Vendor Solutions providing Company. It is very user friendly and has inbuilt spark on it.

Figure 3 : Databricks Home page snippet

Above is the snippet of the home page is shown below. It involves three simple steps: Creating a table for the input data file, creating a cluster and a notebook to run the program.

To read data on databricks, one should upload the data file to Data-Bricks File System (DBFS) and create the data table with UI to access the data. Below is the snippet of the sample data in Data-bricks.

A close up of text on a white background

Description automatically generated

Figure 4 : Databricks Data Table

**4. EFFICIENCY AND RUN TIME**

**On a Local Machine :**

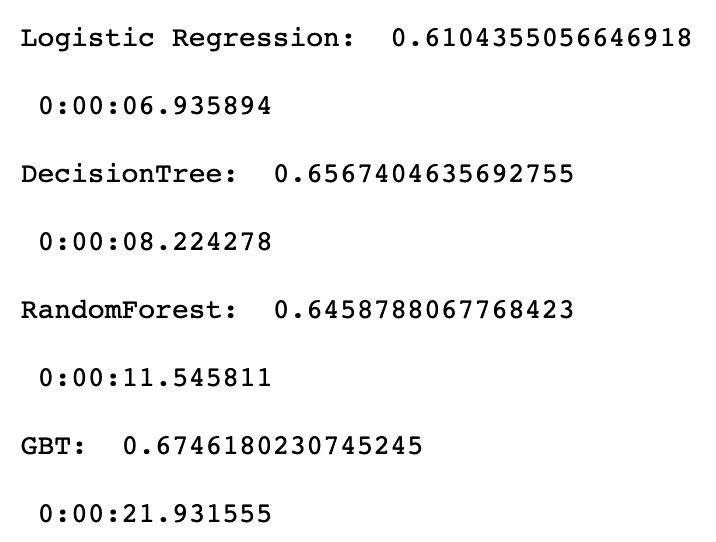
The local machine used to run our PySpark program is a MacBook Pro with an Intel i7 processor. Below snippet shows the accuracy and runtime of different algorithms used.

Figure 5 : Local Machine Accuracy & Runtime

**On the Cloud:**

The cloud used in the project is a cluster provided by Databricks and the local machine used to access the cloud is and HP spectre 360 with Intel i7 processor. Below snippet shows the accuracy and runtime of different algorithms used.

**A screenshot of a cell phone

Description automatically generated**

Figure 6 : Databricks Accuracy & Runtime

**5. RUN TIME COMPARISION USING GRAPHS**

**A. Logistic Regression**

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

Figure 7: LR Local Machine

Figure 8: LR Coud

From above we can see for Logistic regression, the run time on local graph is going down after a certain point. The reason for this can be local machine’s memory and the overall tendency of run time on the cloud is going up. Which means, more the data, more the run time. The run time here is in 10\* seconds. i.e where it reads 0.9, it means 9 seconds.

**B. Decision Tree**

**A close up of a map

Description automatically generatedA close up of a map

Description automatically generated**

Figure 9: DT Local Machine

**Cluster:**

Figure 10: DT on Cloud

From above we can see for Decision tree, the run time on local as well as cloud, the overall tendency of is going up. Which means, more the data, more the run time. It’s taking more time on cloud because the local machine used for Pyspark is a high end machine with a good processor. The run time here is in 10\* seconds. i.e where it reads 0.9, it means 9 seconds.

**C. Random Forest**

**A close up of a map

Description automatically generatedA close up of a map

Description automatically generated**

Figure 12: RF Cloud

Figure 11: RF Local Machine

**Cluster:**

From above we can see for Random Forest, the run time on local graph is going down after a certain point. The reason for this can be local machine’s memory and the overall tendency of run time on the cloud is going up. Which means, more the data, more the run time. The run time here is in 10\* seconds. i.e where it reads 0.9, it means 9 seconds.

**D. Gradient-Boost-Tree**

**A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated**

Figure 14: GBT Cloud

Figure 13: GBT LOcal Machine

**Cluster:**

In Our project we see that the run time of GBT is the highest on Local Machine as well as cloud. The Runtime here is in seconds.

**6. RESULTS**

* Models which have higher AUC take high computational time on both local and cluster
* Gradient Boosting outperforms other models in AUC but takes maximum time as well
* Random Forest has the right balance of computational time and AUC – It is robust across local machine and cluster
* We observed more time taken on cloud. However we are assuming that Local Internet speed/Bandwidth can be a reason and the performance and processor of the local machine is very good as it is an high end device.

**7. CONCLUSION**

Our prediction model has accuracy about 70 which is very good in machine learning. Hence, this model can be used in predicting the success of real time projects in kickstarter. It will be helpful for the project developer to know the success of project before hand and he can improvise in the project to get a funder. Prediction can help the funder in deciding whether fund in the project or not.

In this project Random Forest is the recommended model based on above analysis for success rate prediction for crowd funding. As scaling increases, the performance on cluster improves. Thus, we concluded that the models perform better on cluster vs the local if the data is big.

**8. FUTURE WORK**

In the rum time graphs of Logistic regression and Random Forest the graph goes up upto a certain extent and then goes down after that extent. Our future work is to find out the exact reason behind the runtime going down.

**9. REFERENCES**

1. <https://fundsforwriters.com/grants/>

2. <https://www.kaggle.com/datasets>

3. <https://en.wikipedia.org/wiki/Random_forest>

4. <https://en.wikipedia.org/wiki/Decision_tree>

5. <https://docs.databricks.com/user-guide/tables.html#create-table-ui>

6 <https://towardsdatascience.com/a-brief-introduction-to-pyspark-ff4284701873>

7.<https://learn.launchboom.com/5-reasons-why-crowdfunding-is-the-best-way-to-launch-your-new-product-f080f4f6e176>

8. <https://www.womenarts.org/skills/crowdfunding/>

9. <https://medium.com/datadriveninvestor/big-data-in-the-cloud-why-and-how-ac19bb025bc1>