Imperial College London

Imperial College London

Department of Life Sciences, Faculty of Natural Sciences

Predicting Academic Funding Success Using Machine Learning Approach

Author:

Jingkai SUN

CID:

01991822

Supervisors:

Dr. David Orme

Dr. James Rosindell

Dr. Samraat Pawar

Email:

ks3020@ic.ac.uk

April 7, 2021

Keywords

Machine Learning (ML), Text Mining, Biology, Natural Language Processing (NLP), Scientific Funding Success, Quantitative Methods, Statistical Analysis

1 Introduction and Project Ideas

Academic funding provides essential support for scientists. With an increase in the number of researchers and scientists all over the world, the competition of the academic job market is increasingly fierce, so that the probability of funding success is decreasing. Hence, it is important to understand what determines funding success and how it will be biased due to different reasons.

So far, numerous quantitative measures in the performance of researchers had been proposed [4], such as h-index [2] and impact factor [5]. Also, due to the development of big data and machine learning, many researchers started to study determinants of scientific success using statistical and machine learning methods[1][4][5]. However, no research has been done for studying the potential key factors and prediction on funding success using such quantitative methods. Therefore, the object of this project is to study the determinants of funding success. The relevant questions are:

- What are the most influential factors to determine the probability of funding success in biological science?
- How do these factors affect the funding success?
- Are funding success predictable using the Machine Learning approach?
- Are predicted outcomes robust in other subject areas?
- Do phrases in the paper improve the accuracy of the predicted outcomes? (to be discussed if time is available)

2 Proposed Methods

All text data, split into training and test datasets, will be obtained from the database of UKRI, NSF, NIH and ERC. I plan to extract features from text data and build a statistical model in training sets to find the key factors of which weights of the model are statistically significant. i.e.

I will analyse all funded papers and find the common factors among them, then using a statistical model to see whether they are statistically siginificant. Then these features will be used to fit different machine learning models. Afterwards, I will use the best-fitted model to predict testing sets. Also, if time is available, I plan to use the NLP method to analyse the phrases in the funded paper and to see if the phrases have strong predictive power.

3 Anticipated Outcomes

In this project, I assume there are lots of factors determining the funding success. Thus, after analysis, many factors that have a siginificant effect to funding success should be found, so that I can use them to make prediction. Also, phrases has been proved to have stongly predictive power on Kickstarter, which is a website of which entrepreneurs look for funding [3]. Hence, I assume that phrases will also boost predicted accuracy on my research purpose, which is another outcome I expect to obtain.

4 Timeline and Feasibility

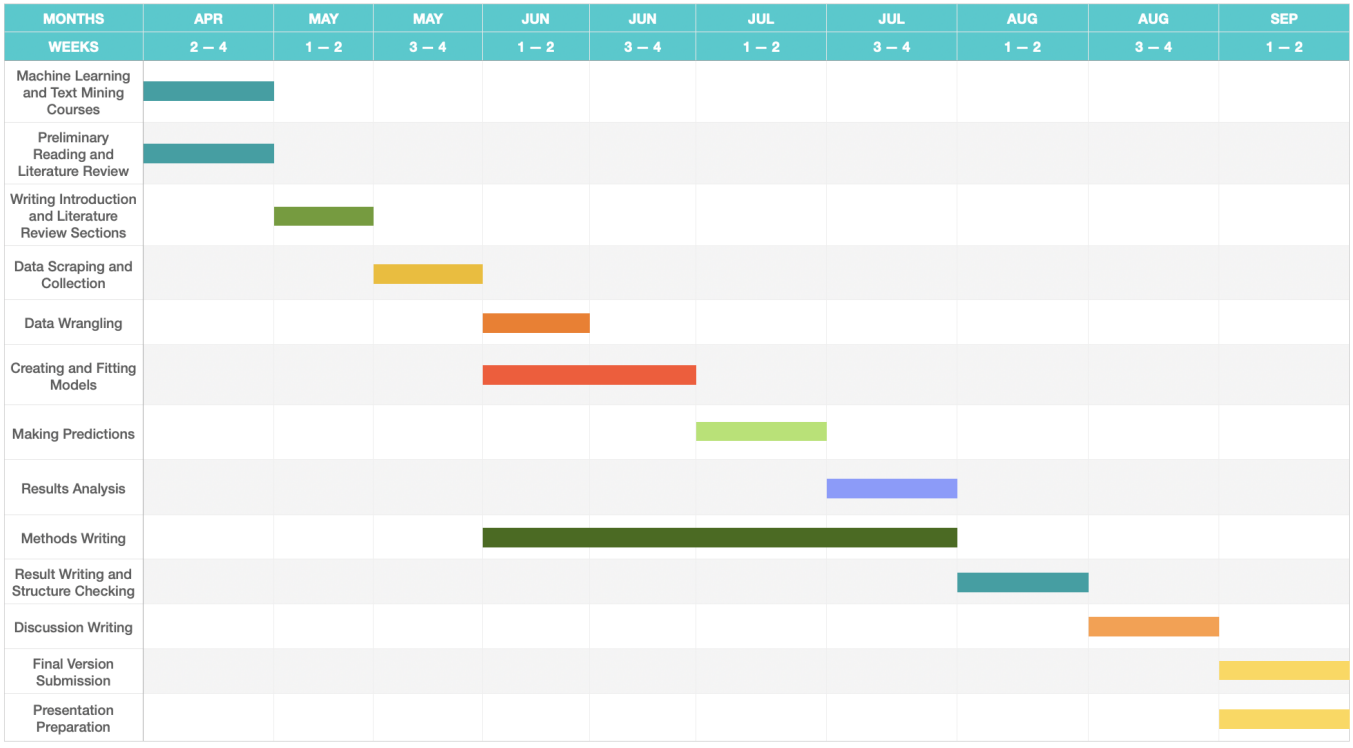

Figure 1: Project Timeline

5 Budget

None

References

[1] Daniel Acuna, Stefano Allesina, and Konrad Kording. Future impact: Predicting scientific success. *Nature*, 489:201–2, 09 2012.

[2] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[3] Tanushree Mitra and Eric Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. pages 49–61, 02 2014.

[4] Emre Sarigöl, Rene Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 02 2014.

[5] David van Dijk, Ohad Manor, and Lucas B. Carey. Publication metrics and success on the academic job market. *Current Biology*, 24(11):R516–R517, 2014.