Bibliometrics Analysis of Authors

Instructor: Christopher Asakiewicz
Authors: Jiahui Bi,
Hongyi Chen,
Shan Gao,
Yuyi Yan

Team Members

Yuyi Yan

Hongyi Chen

Jiahui Bi

Shan Gao

Introduction

- The overall US publishing industry is financially healthy, the average revenue of scientific journals has grown by nearly 60% from 2010 to 2017.
- We are looking for methods to identify future leading authors in a specific scientific field.

Project Goal

- Analyze the impact factor of the authors
- Analyze the Co-authorship
- Analyze funding, publication features and its correlation with impact factor
- Validate the top 5 authors in bio-material area in 2013-2015

Technology

- Python for cleansing the data collected from Web of Sciences.
- Python & Tableau for generating visualizations for EDA.
- Python for network analysis and modeling.
- VOSviewer for generating Co-authorship network visualization.

Data Source & Variables

- Web of Science core collection
- Topic: Biomaterials, from 1990 to 2019
- Total: 46409, 68 variables

AU	Authors
AF	Author Full Name
BA	Book Authors
BF	Book Authors Full Name
CA	Group Authors
GP	Book Group Authors
BE	Editors
TI	**Document Title**
SO	Publication Name
SE	Book Series Title
BS	Book Series Subtitle
LA	Language
DT	Document Type
CT	Conference Title
…	

- Conference Information
- Times Cited
- ISSN / ISBN
- Accession Number
- Author Identifiers

Publication Types

- B = Book
- J = Journal
- P = Patent
- S = Book in Series

Field Tags for ...

- Compounds
- Patents & INPI Records
- Reactions

Constructed the Author Dataset

AF
aAuthor_1;aAuthor_2;aAuthor_3;aAuthor4;…; aAuthor_n
bAuthor_1; bAuthor_2;…

First Author	Co-Author_1			Co-Author_n
aAuthor_1	aAuthor_2	…	…	aAuthor_n
bAuthor_1	bAuthor_2	…	…	…

First author

Author	Feature1	Feature2	…
aAuthor_1			
aAuthor_2			
…			
aAuthor_n			
bAuthor_1			

- Split the Author column.
- Use single author as the granularity of the author dataset.

Why Impact Factor?

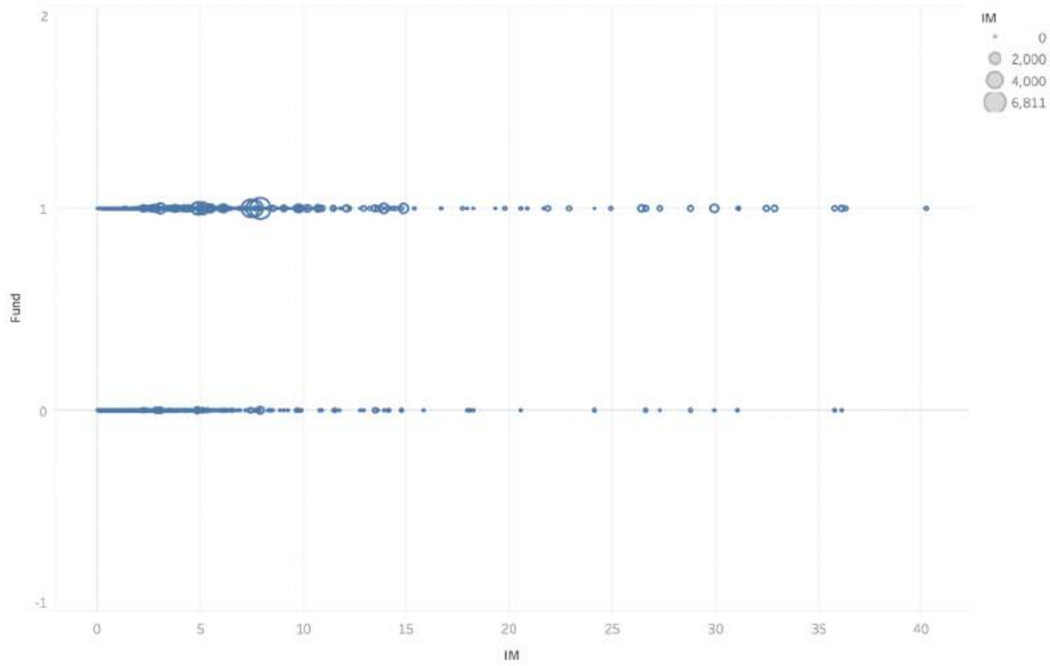
- Impact Factor is used to reflect the average number of citations divided by the total number of articles post on the journal recently. IF is frequently used as a proxy for relative importance of a journal. Higher IF indicates higher importance than lower ones:

$$\mathrm{IF}_y = \frac{\mathrm{Citations}_{y-1} + \mathrm{Citations}_{y-2}}{\mathrm{Publications}_{y-1} + \mathrm{Publications}_{y-2}}$$

  - We sum up the IF of the articles published by a certain author, considering the year of publication in the journal.
  - Total impact Factor of a certain author, during a time period, can reveal the quality of his published researches.

Why funding?

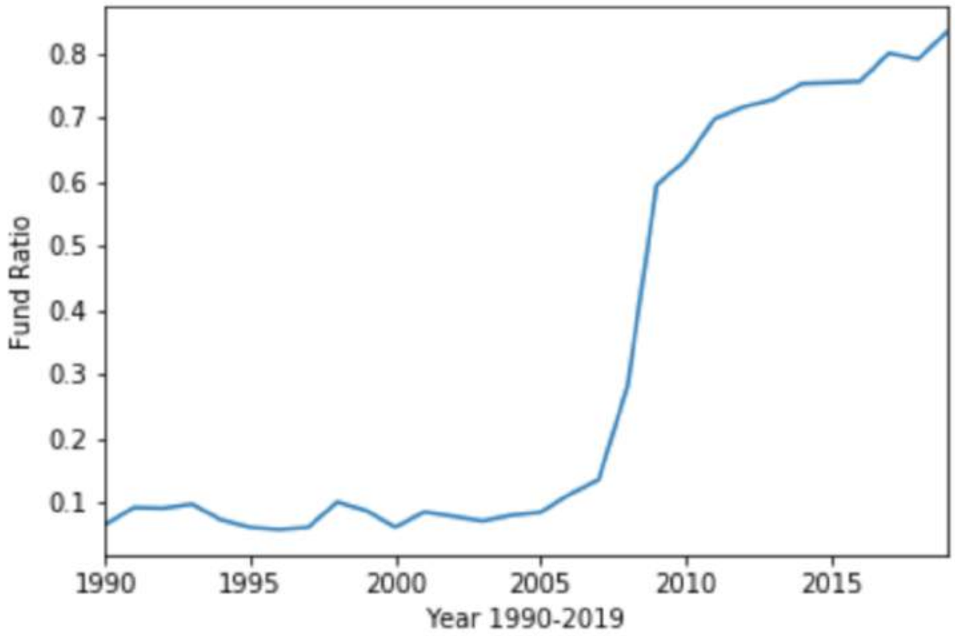
Funding & Impact Factor



- Funded articles are more likely published on a journal with high Impact factor.

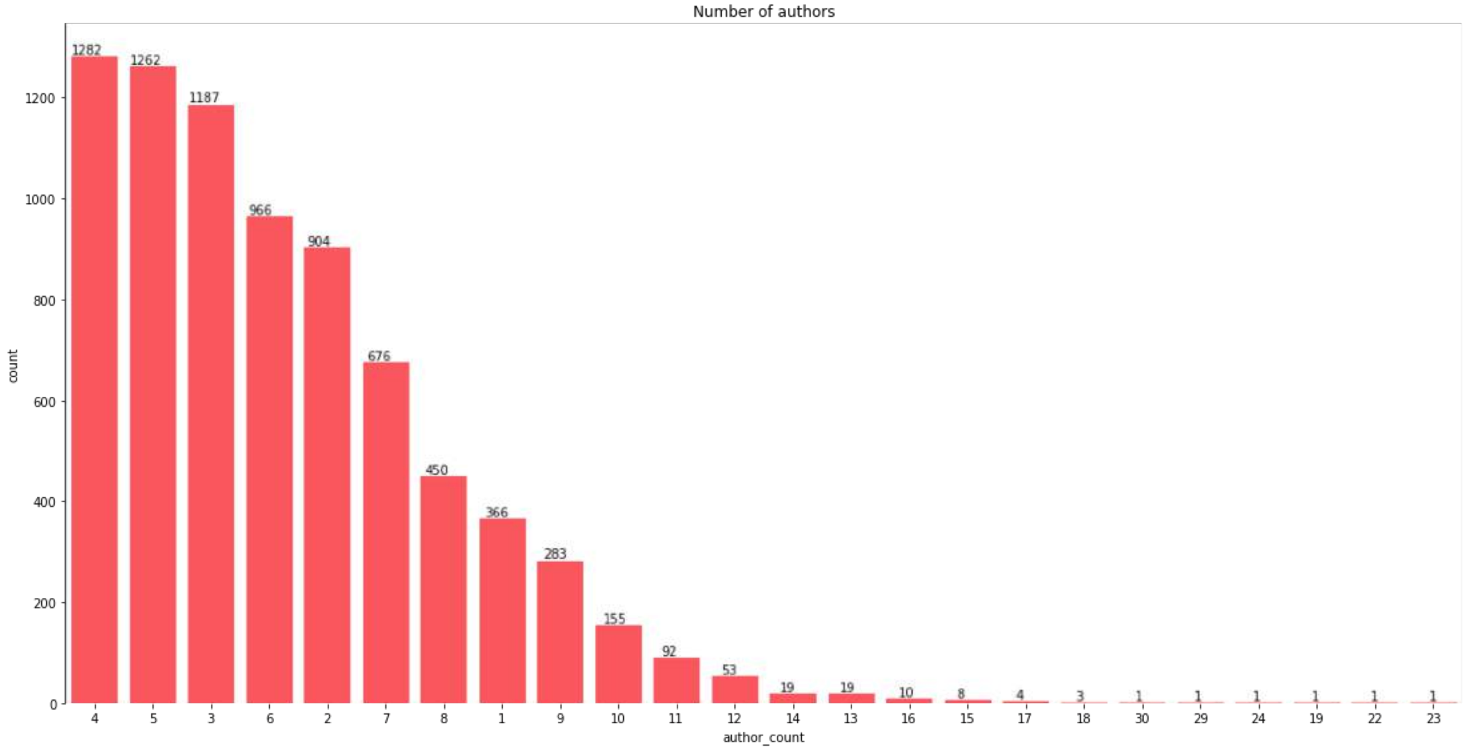
Why funding

Funding & Publication over year



- After 2010, more than 65% articles get funded.
- At 2009, funding & publication ratio has a cliff growth.

Co-authorship

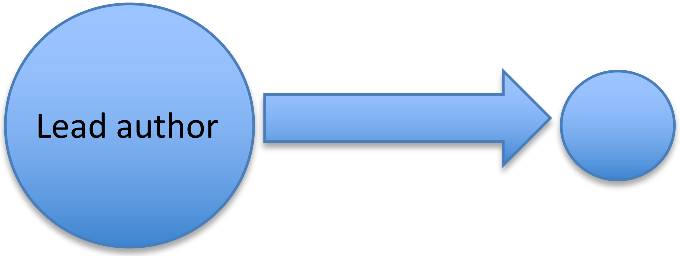


- The article with 4 or 5 authors is most common.
- Total 366 articles have only one author, considering 7745 articles during 2010-2012.

Co-author Network

Co-author Directed Network

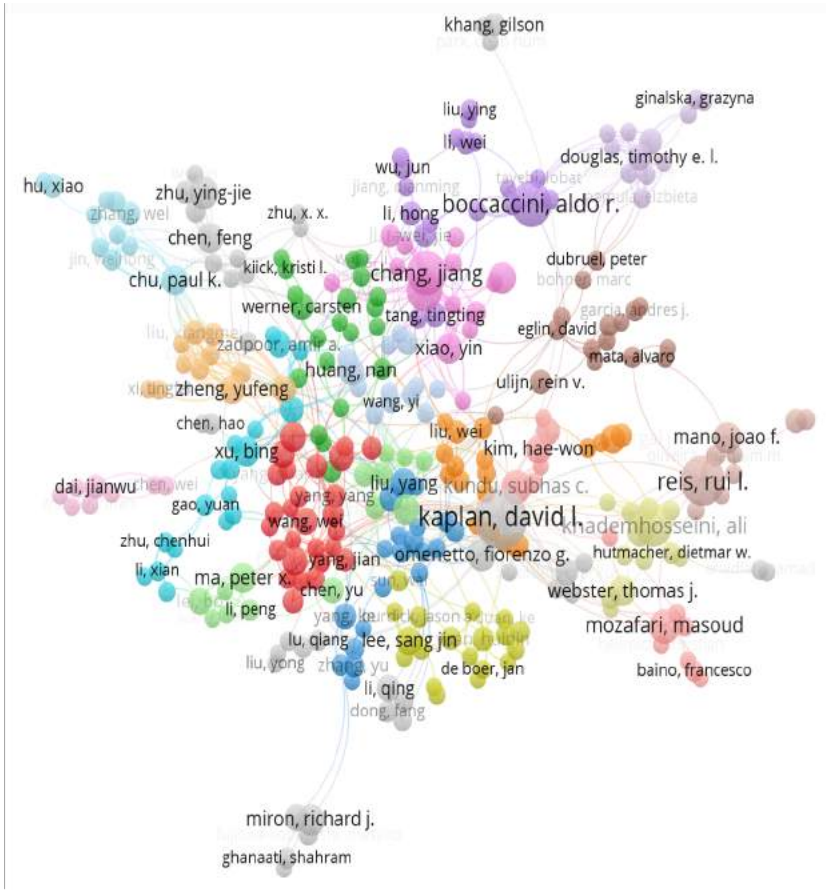
- The Co-authorship network is formed if two authors(node) co-authoring an article together(edge). The edges are directed from the lead author to the other authors. The larger the node is, the more paper the author published.



(Ying Ding, Scientific collaboration and endorsement: Network analysis of co-authorship and citation networks, *J Informetr.* 2011 January 1; 5(1): 187–203.)

Co-authorship Analysis

- The authors with high centrality are always active corresponding authors. It’s a good indicator to consider their co-authorships to detect potential authors.
- As it is a directed network, the outdegree and indegree of a certain node, reflect the information about the times the author act as the first author and the number of the research participated in.



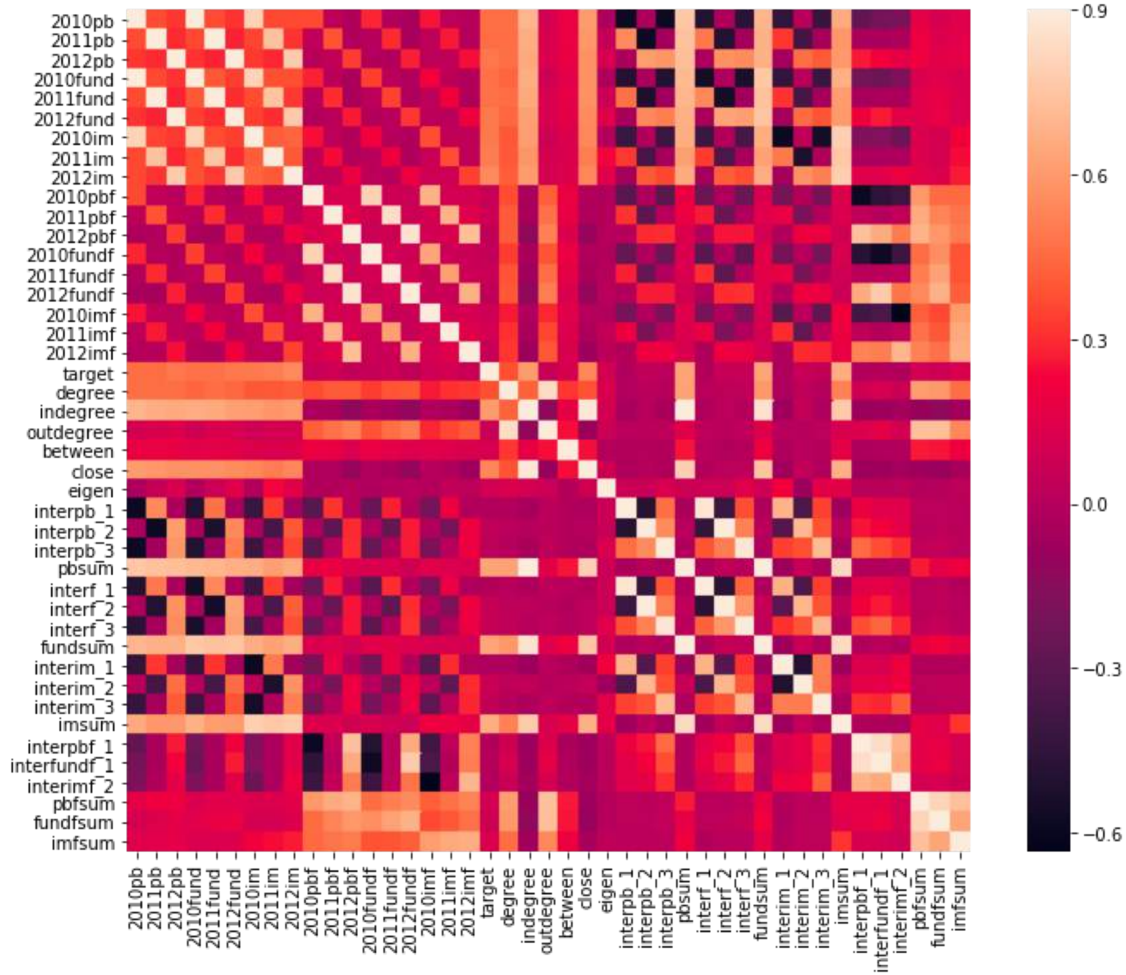
Time Gap

- **Feature**: Constructed from author dataset from **2010 to 2012,** by two groups, all authors and first author.
- **Res Variable**: The sum of impact factor from **2013 to 2015**.
- Predicted future top authors by using current features.

- Used articles published after 2010 to construct author dataset.
- Considering Funding & Publication ratio after 2010, which is remarkably different.
- After 2010, the new published articles have exceeded more than 2000 every year.

Feature Selection

Feature correlations

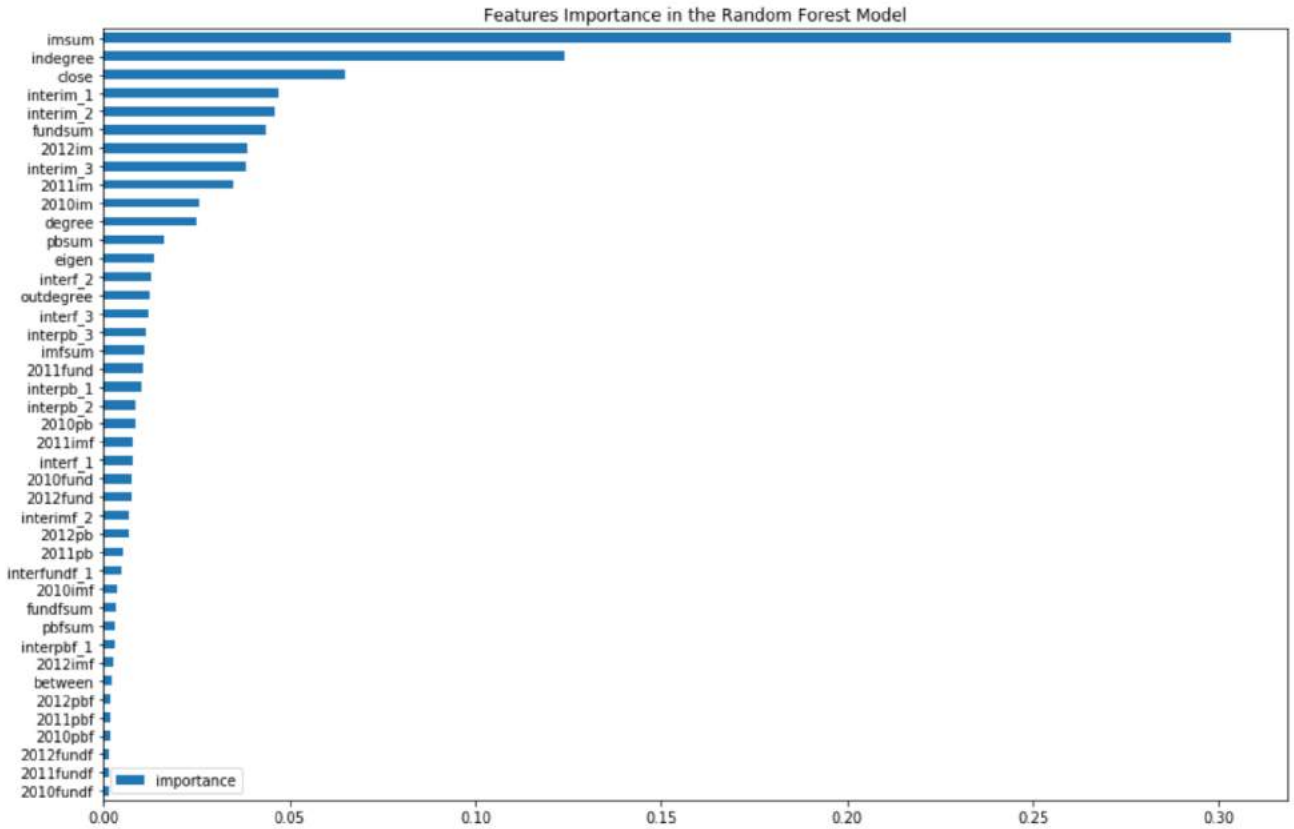


Modeling

- Linear Regression with Ridge, Lasso, and ElasticNet, comparing with Random Forest.

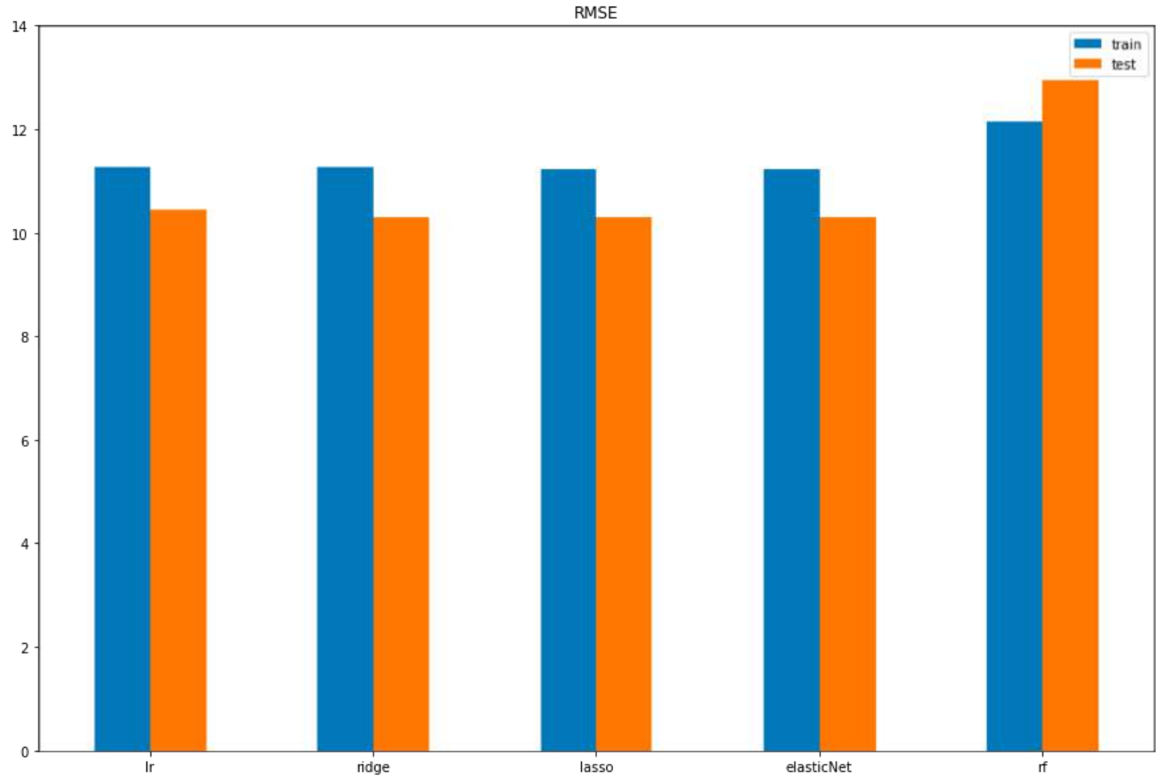
Feature Importance

Random Forest



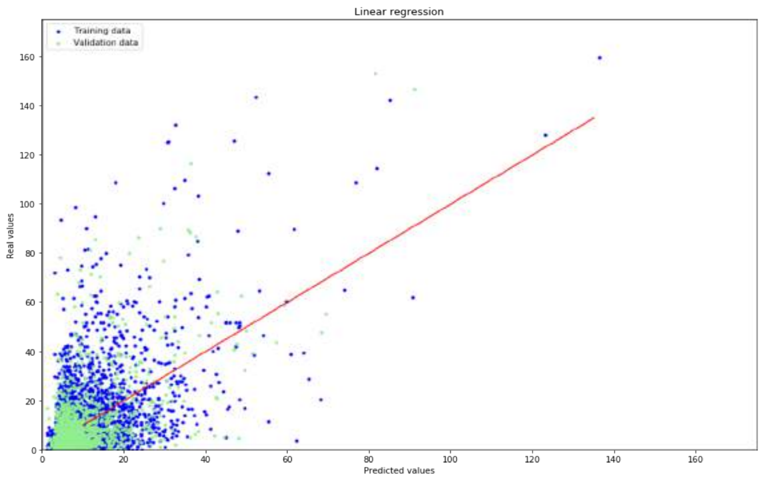
- Impact factor related features
- Co – author network features

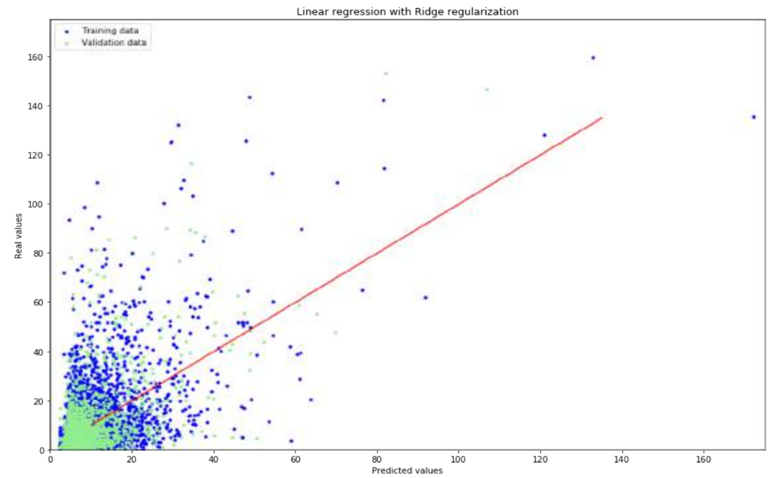
Modeling

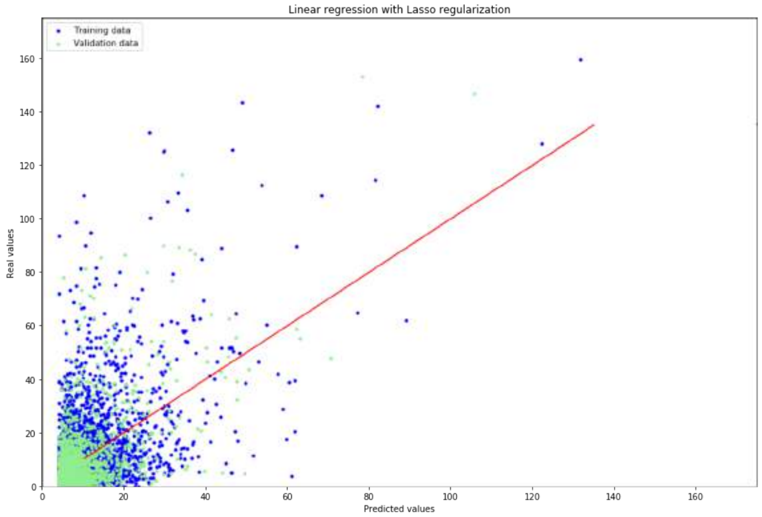


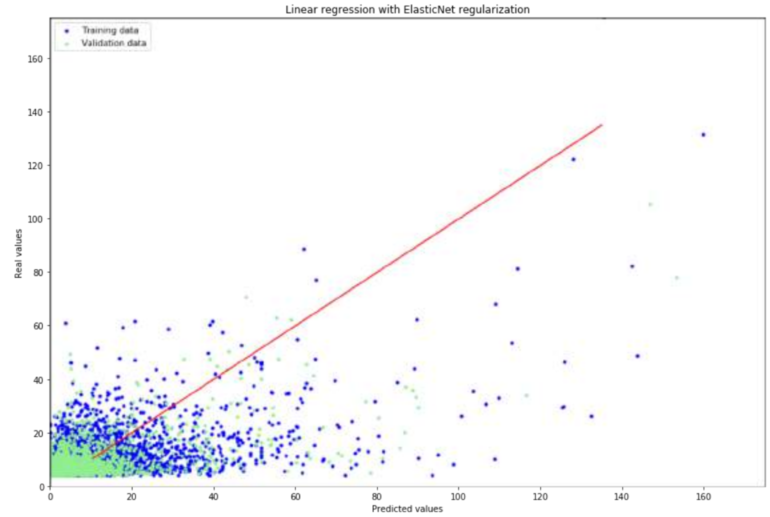
	train	test
lr	11.2634	10.4566
ridge	11.2684	10.3060
lasso	11.2331	10.2956
elasticNet	11.2292	10.2937
rf	12.1611	12.9575

Linear Regression Result

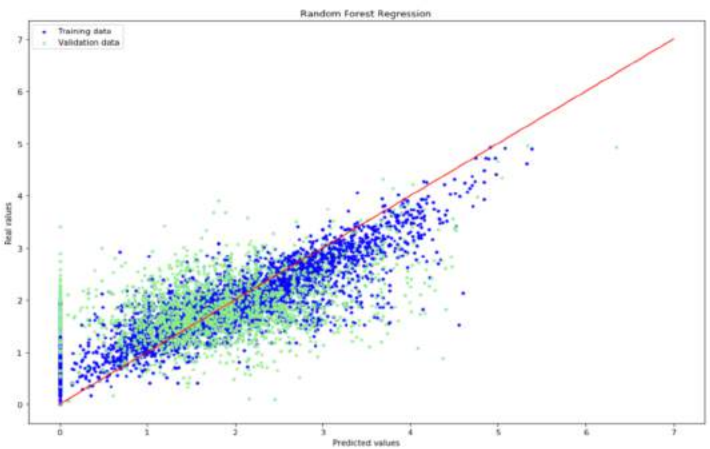


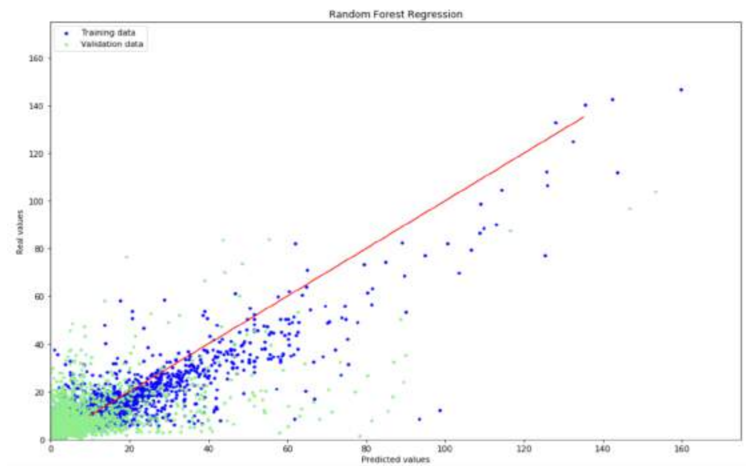


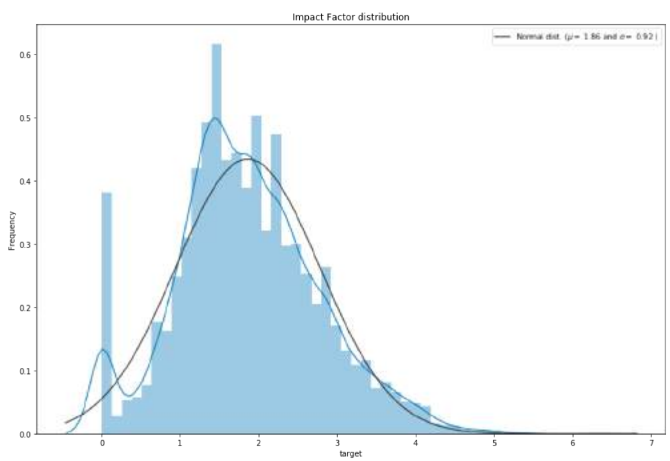




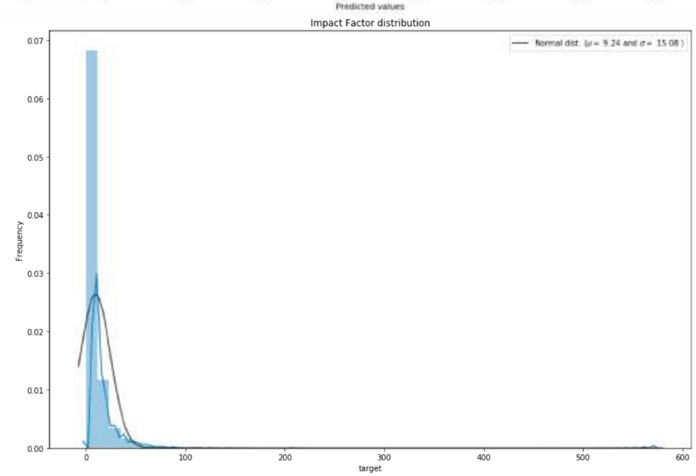
Random Forest Result





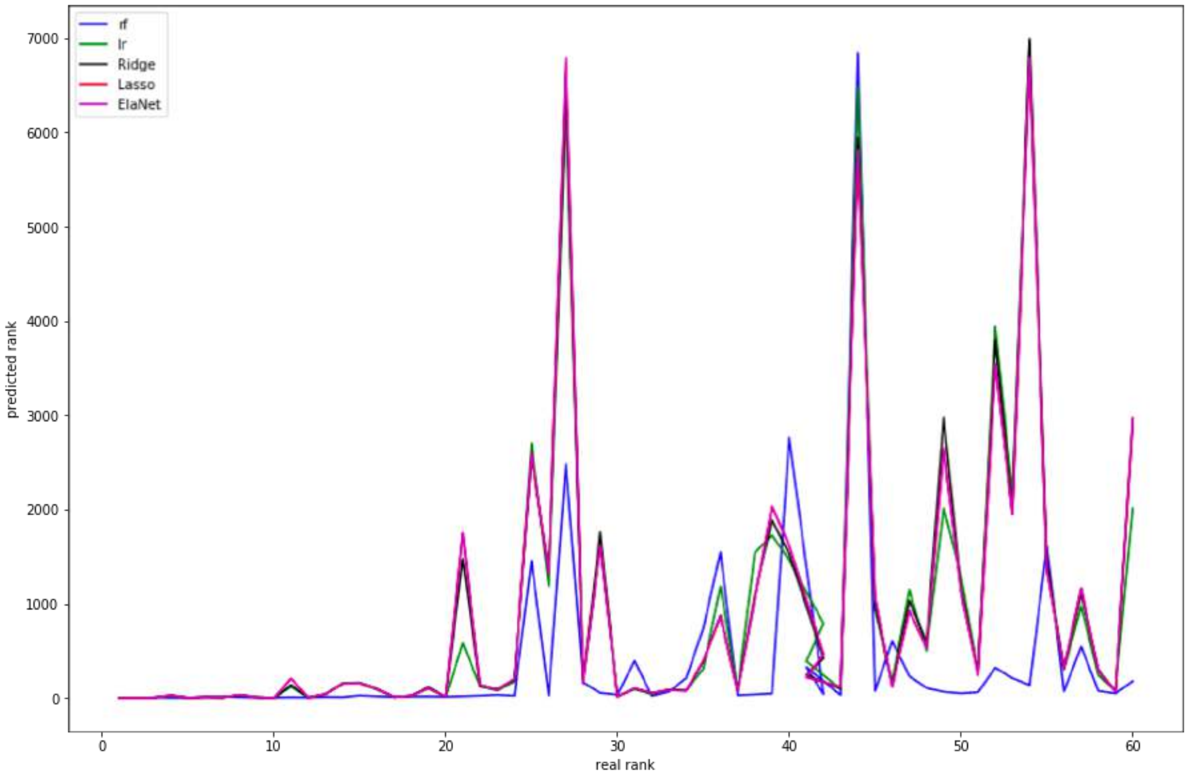


Log transform on Res variable



Original distribution of Res variable

Considering Ranking



- Theses model can make great prediction on top 25 authors.
- Random Forest seems more stable and has a better performance than linear regression.

Considering Ranking

	Top 10	Top 20	Top 50	Top 100
rf	7	19	32	57
lr	7	11	17	36
ridge	7	11	16	33
lasso	7	11	16	35
elasticNet	7	11	16	35

- The table shows the number of seats models predicted on each top N list, considering real ranking list.
- As N increasing, the ratio predicted is decreased. The ratio decreasing in Linear regression model is more significant.

Top 20 Ranking

author	rank_2013	rank_2015	rank_rf	rank_lr	rank_ridge	rank_lasso	rank_elasticNet
Kaplan, David L.	1	1	2	1	1	1	1
Mano, Joao F.	4	2	1	3	3	3	3
Reis, Rui L.	6	3	8	5	5	5	5
Mooney, David J.	41	4	4	25	31	29	29
Khademhosseini, Ali	5	5	3	4	4	4	4
Chu, Paul K.	14	6	14	11	9	11	11
Lendlein, Andreas	15	7	16	7	7	7	7
Burdick, Jason A.	42	8	11	30	36	33	33
Omenetto, Fiorenzo G.	7	9	5	9	11	9	9
Langer, Robert	2	10	6	2	2	2	2
Higuchi, Akon	93	11	9	123	140	208	209
Anderson, Daniel G.	3	12	7	6	6	6	6
Chang, Jiang	85	13	12	47	40	42	42
Ling, Qing-Dong	94	14	10	153	160	153	153
Hubbell, Jeffrey A.	140	15	28	158	164	156	157
Chang, Yung	83	16	19	91	102	99	99
Stupp, Samuel I.	8	17	13	10	10	10	10
Kundu, Subhas C.	58	18	17	27	26	26	26
Umezawa, Akihiro	84	19	18	102	121	108	108
Boccaccini, Aldo R.	24	20	15	12	13	14	14

- Random Forest has a greater performance on picking up potential top authors.

Conclusion & Future Work

- **Conclusion：**
  - We used two kinds of regression models. Linear regression with 4 different regularizations is more conservative than Random Forest, which explained a lower RMSE.
  - Random Forest with log transform is more stable than using original target distribution. However, when considering pick up top authors, log transform did not helped.
  - Random Forest model are more likely pick up potential top authors.
  - These models have advantages in picking top authors. (under top 50)
- **Future Work**
  - Considering more features, citation network, academic age etc.
  - We can expand the time span by adding features from other aspects, thus more author can be considered.

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu