

Predicting Paper Retractions

William Davis and Jack Kendrick

						LETTER	Vid Add of Add
	The Part Avenue (2000 ML Non-one TECHNICAL, ADVANCE	Articles		Rady report		INC Evolutionary Biology	essential for hypoxia-induce
persion of postnatal human manuae	An enhanced transient expression system in plant: suppression of gene silencing by the p19 protein of bushy stunt virus			Real-lymphold-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children A returned by Matti, A Antony, Linest De Casso, Mittel, Mitsecett, A FONIS, Mit Tonas, Panes, Panes, Panes	I LE JACQ"	tot graphical analysis environment	Ania Mentari, Nala Danhalari, Orinta ya
MANUE UNITABLE	War Values? I want was the set of the set of the program products of the set of the	Cardiac stem cells in patients with ischaemic 🛛 🖓 🦒 cardiomyopathy (SCIPIO): initial results of a randomised	Note has been networked. N Engl J Med 2012(178(21))2441.2	Example Endpaned VV instructor at measure and a second of the second data with the second da	Original Paper	Gangolf Jobb+1, Amdi von Hausen	b mail many "plates all()
Contracting the second se	And and a second and a second			Note: 10 online (new price price price) (and the second se	Predictive Validity of a Medic: Adherence Measure in an Outr	Albert Operand (Specific and Specific and Specific and American Specific and American Specific and Specific and Specific and American Specific and America	Tenderical Social Socia
And a statement with and added restatements And a tradementation being	way way way way way way way way	8 dogs block that Statisfaction Equation Free Annual Summary Independent Association, Basegonappin confer, etcas, only (SSG), Improve peek for Conference on the CON- manual Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Independent Association Statistical Statistica	In EDICINE	Insuburnation and Darpa sampling, respective memory insurgar (Monto), exceptionality (2010, est toleration of the sampling of	Donald E. Marieky, ScD, MSPH ¹ Alfonso Ang, PhD, ² Marie Kros Horry J. Ward, MD ³	Companying Kannon Limit and Kannon Lim	W. In
Providence of a data in the second se	In the description of the plant have been beind and a sublight suprementation of the plant and	Increment candidate/opting (SC.1710) of animologies CSCs for the treatment of heart for treatment or the first of the treatment of heart for treatment or the first of the treatment of the treat	resultary Prevention of Cardia	material. The second se		The web and a final sector of the CANNEL TO A TANK of the sector of the	
In a start of the starting of	Annual state of the state of th	before commay after papas guiling were concentrively model in the Lot al and contractory. In singe 3, and an another partners were madeway assigned in the treatment or cannot group in a 1 mile by use of a comparisonment. Mede and eminimize schemes 1 million matslageous GGCs were administend by intractory in the lot of a TreVision 10 (1990) and the treatment of the lot of a TreVision 10 (1990) and the treatment of the lot of a TreVision 10 (1990) and the treatment of the lot of a TreVision 10 (1990) and the treatment of the lot of a TreVision 10 (1990) and the	In Construction, PLD, Dalama Canda, D.D., And Salar Salaria and Annual and	inflation is all banks and another thanks in the second se	This shady cannot no to psychowskic properties and nort due suncavero and production addremote manare to patient with foportunions. The authors also assauds norms produces of theorements of addremments of addremments of addremments of addremments produces.	Reading TREEPCER is an every low one of a sign factor of the and requirements over a warr a world'y	
Mill Charles and Collection and Coll 2018	The provide of the state of the	13. dags (15.4) after sagery, contrain were set piers aus transmosting beingd. If Tardy was upon Bald, the obsciedingshipsking being strain and strain and the piers and transmosting of strain and the strain and strain and the strain and strain and the strain	O A TA See Meeting And Darks, Add, Janki, K.D., Allow, A.D., Allow,	that deciding is any details and used at 35 bits and the specific of the proton order number of the specific order of the proton of the specific order of	an hosoledge, notal support, athlustion softe care, and complexity of the medical regimers. A third of 1267 patients participated in the study mean age near 52.5 years, 44.5% server study. ACVS server black, then then isless	the indicated full-block are used related to approximate the second seco	
And Y. Markani, S. Matt. Statistical and an approximate of the statistical and appr	and a section of the	proug, no GC-related above effects our region of GC- of patients who were analysed, DTF increased in hyperson effects our region of GC- of patients who were analysed. DTF increased in hyperson effects our region of GC- of patients who were analysed. DTF increased in hyperson effects our region of the same analysed of the s	AASTRACT	And a set function of speeds in a grant of program of the speeds of the	50.8% graduated from high school, 20% were new real, and 54.75 and scores 455,000. The 8 down real-control address and the score 455,000 and the score address registrantly associated with Mood processor control and processor with score, correspond with	delarititi peri tare increases (U(3)) Stationana de Cali a sono entitado oni on the Vanciona (delocati entito) entitado oni on the Vanciona (delocati entito)	F Carry parties 1 41 42 101
Control of the second sec	And Market And Andreas			Memoralise Event Phone Theorem The Descent Phone	(Pull)). Using a compart of a d, for sen hard, the resource to identify patient such poor or control and exclosing do by 15°, and for a characteristic mar 35%. The weak atom does not sent to be rescaled at the sent sent mar atom of the sent sent sent sent to be rescaled at the sent sent sent sent to be rescaled at the sent sent sent sent sent to be rescaled at the sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent sent to be rescaled at the sent sent sent sent sent sent sent sen	1) Comparison of indication to the comparison of the comparison of indication to the comparison of indication of the comparison of	Constanting for some field and the source of the field and the source of the sour
An and a setter apply, for all an an apply and a setter and a setter apply and a setter a	All, 2008. Rowses, Cells (Kapin et al., 1927). Man colorade into a	interpretation These interpretation in patients or very encouraging. They suggest that interpretation of patients of	Sprin, we randomly angle of pericipana with an analysis of the second data and analysis of	Passinder (1) Eventual concerning of Annual of Eventual of Eventua	rability in presarely for sever, in very times is, behavioral, treatment	2 In A small, A large with minimum drawed of down for provident to produce of a property of a pro	In the product of the product of the COM Concession And

Motivation

- In theory, articles published in peer reviewed journals should hold up to scrutiny. When reading a paper, it is often taken for granted that the research is credible and trustworthy.
- However, sometimes bad papers fall through the cracks and end up being retracted.
- Retractions indicate seriously flawed and unreliable research, errors, fraud, ethical issues, or other serious concerns



Can we identify papers at a high risk of retraction?



Starting Small: PLOS One

- We concentrated on the journal PLOS One
- Relatively well known and respected, but has a high number of retractions

Rank	Journal	Retractions*	
1	2011 International Conference on E-Business and E-Government	1280	
2	2011 5th International Conference on Bioinformatics and Biomedical Engineering	1084	
3	PLoS One	944	
4	Journal of Physics: Conference Series	878	
The state of the s		Bit Excce The second of the seco	

The Dataset

- We collected data from the OpenAlex database using the PyAlex API
- Huge amounts of data available (including retractions)
- We used the raw data from OpenAlex to build features of interest
 - Fraction of authors that have previous retractions, whether any authors come from institutions with many retractions, etc.
- Our dataset consists of papers published in 2010-2020: 424,223 papers, with 797 retractions

Challenge: Retracted vs Non-Retracted Classes are massively unbalanced



Our Approach

- → Baseline logistic regression model
- Compare more baseline to nearest neighbour, random forest, and SVC classification methods.

*Key Performance Indicators: F*₁*-score, Precision, and Recall*

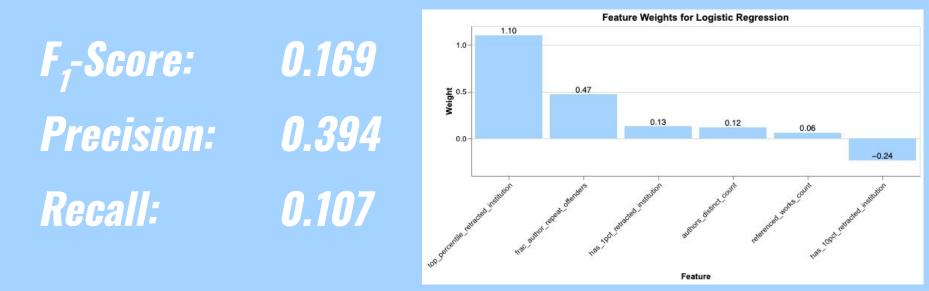


Baseline Model

- Used forward stepwise subset selection to choose features.
- The most informative feature is the proportion of authors on a paper that have been previously retracted
- Second most informative feature is a measure of how many retractions any institution associated to the paper has received.
- Results of subset selection correspond to our intuition on features that correlate to a risk of retraction



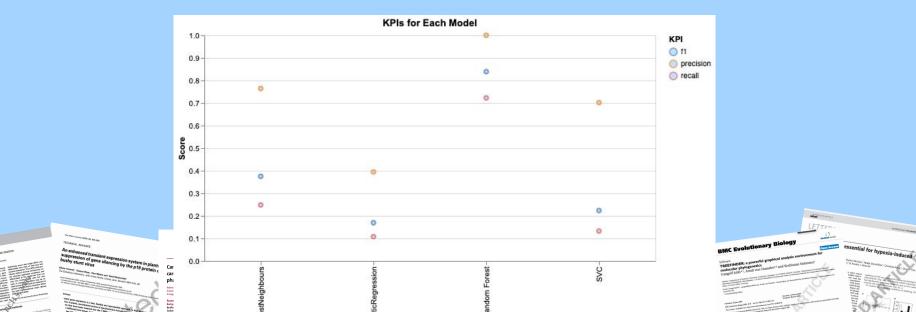
Baseline Model





Model Selection

- Tested k-Nearest Neighbors (k = 1, 2,..., 10), random forests, and support vector classification.
- → We used stratified 10-fold cross-validation to choose hyperparameters.





 Random Forest classification far outperformed our other models in training.

Final Model: Random Forest with 500 estimators, max depth of 20



Future Directions

 Expand dataset to include other journals and features relating to funding sources, journal publishers, etc.

 Investigate papers published in the post ChatGPT-era. Can we use 'Contains AI generated text' as a feature?



Acknowledgements

- Many thanks to our mentor Greg Edwards for his help and expertise throughout this project
- Thank you to the Erdos Institute for providing us with this excellent bootcamp.

