

On the relationship of novelty and value in digitalization patents: A machine learning approach

Supplementary Material

1. Descriptive statistics of included variables

Table 1: Patent value, novelty and control variables

Note: Mean and SD values are obtained of the corresponding sample. N refers to the sample size. SD refers to standard deviation.

Variable	Type	Part	N	Mean	SD	Source
<i>Technological value</i>	Value	-	263,960	0.10	0.30	Verhoeven et al. (2016)
<i>Economic value</i>	Value	-	157,523	0.10	0.30	Kogan et al. (2017)
<i>no_bw</i>	Novelty	Backward citations	263,960	1.47	12.02	Ahuja and Lampert (2001)
<i>pairwise_bw</i>	Novelty	Backward citations	263,960	0.05	0.67	Arts and Fleming (2018)
<i>pairwise_cpc</i>	Novelty	Classifications	263,960	0.05	0.54	Fleming (2001)
<i>pairwise_cpc_ratio</i>	Novelty	Classifications	263,960	2.82	12.88	Arts and Veugelers (2015)
<i>pairwise_cpc_bw</i>	Novelty	Backward citations and classifications	263,960	0.07	0.72	Verhoeven et al. (2016)
<i>radicalness</i>	Novelty	Backward citations and classifications	263,960	2.44	3.60	Shane (2001)
<i>min_sem_dist</i>	Novelty	Text	263,960	79.52	18.20	Moehrle and Gerken (2012)
<i>mean_sem_dist</i>	Novelty	Text	263,960	68.37	9.22	Arts et al. (2021)
<i>mean_pat_age</i>	Novelty	Text	263,960	2.48	2.80	Wu et al. (2020)
<i>new_unigrams</i>	Novelty	Text	263,960	0.11	0.59	Arts et al. (2021)
<i>new_bigrams</i>	Novelty	Text	263,960	1.54	2.10	Arts et al. (2021)
<i>new_trigrams</i>	Novelty	Text	263,960	4.19	3.22	Arts et al. (2021)
<i>team_size</i>	Control	-	263,960	2.45	1.78	Lee et al. (2015)
<i>reference_scope</i>	Control	-	263,960	14.50	28.96	Uzzi et al. (2013)
<i>class_scope</i>	Control	-	263,960	2.58	2.34	Barbieri et al. (2020)
<i>text_scope</i>	Control	-	263,960	53.97	24.63	Arts et al. (2021)
<i>claim_scope</i>	Control	-	263,960	19.35	15.92	Galasso and Schankerman (2015)
<i>science</i>	Control	-	263,960	0.53	0.50	Fleming and Sorenson (2004)

<i>filing_year</i>	Control	-	263,960	1998.11	6.34	Huang et al. (2020)
<i>grant_year</i>	Control	-	263,960	2001.31	6.85	Huang et al. (2020)
<i>company_dummy</i>	Control	-	263,960	0.94	0.244	Alcácer et al. (2009)
<i>government_dummy</i>	Control	-	263,960	0.05	0.071	Alcácer et al. (2009)
<i>individual_dummy</i>	Control	-	263,960	0.01	0.064	Alcácer et al. (2009)
<i>US_dummy</i>	Control	-	263,960	0.65	0.490	Gassmann et al. (2021)
<i>foreign_dummy</i>	Control	-	263,960	0.40	0.476	Gassmann et al. (2021)
<i>recent_patenting_activity</i>	Control	-	263,960	778.00	1138.55	Wang et al (2017)
<i>cumulative_patenting_activity</i>	Control	-	263,960	6789.59	10926.80	Wang et al (2017)

2. Correlation tests of including detailed description for novelty variables

To quote the original manuscript in footnote 7:

“Recent literature discusses the value of individual patent parts. While some authors argue that the description of a patent offers poor discrimination power, precision and information overflow (Adams 2010; Denter et al. 2022), others suggest that patent descriptions are more valuable for text processing than patent abstracts or claim sections (Suominen et al. 2017; Kelly et al. 2021). However, which patent part is the most appropriate depends always on the purpose. For robustness checks, we extracted all detailed descriptions and calculated the same text-based novelty variables for the title, abstract, claims and description. Pearson parametric and spearman non-parametric correlation tests are all significant, show large correlations and are available in the supplementary material. Consequently, we argue that including the description is not necessary for our specific novelty measures.”

Table 2: Results from pearson and spearman correlation tests of title+abstract+claims versus title+abstract+claims+description

	Pearson correlation coefficients	Spearman correlation coefficients
<i>min_sem_dist</i>	0.946***	0.965***
<i>mean_sem_dist</i>	0.912***	0.925***
<i>mean_pat_age</i>	0.923***	0.951***
<i>new_unigrams</i>	0.914***	0.920***
<i>new_bigrams</i>	0.897***	0.906***
<i>new_trigrams</i>	0.885***	0.898***

Note: All correlation tests are statistically significant at the 0.001 level.

3. Supervised Machine Learning Report Cards (SMLR)

We recognize the ongoing trend in research for transparency and reproducibility and the challenge thereof when it comes to machine learning. To enhance both factors, we report our two MLP (Multi-layer perceptron) models – from which we extracted the Permutation Importance and Partial Dependence Plot results – in form of the recently proposed Supervised Machine Learning Report Cards (SMLR) (Kühl, Niklas, Hirt, Robin et al. 2021). As we did not employ our final models in industry, we only report insights into model initiation and performance estimation.

3.1. SMLR of technological value

Table 3: Supervised machine learning report card of the best performing algorithm for technological value

Model initiation			
Problem statement	Predict whether a patent belongs to the top 10% of forward citations received within 7 years since grant on 12 novelty variables and 15 control variables		
Data gathering	All variables are calculated on 263,960 US patents which are classified in this CPC and granted between January 1976 and December 2009		
Data distribution	26,899 patents belong to the top 10% class, 237,061 patents belong to the remaining class		
Sampling	No sampling		
Data quality	No missing values		
Data preprocessing methods	Max-min scaling and standardization		
Feature engineering and vectorizing	No additional features apart from 12 novelty and 15 control variables		
Performance estimation			
Parameter optimization	Yes		
	Search space	Solver	$S \in \{\text{'adam'}\}$, works well on large datasets
		Activation function	$A \in \{\text{'identity', 'logistic', 'tanh', 'relu'}\}$, used all possibilities
		L2 penalty	$L2 \in \{1/100, 1/99, 1/98, \dots, \underline{1/20}, \dots, 1/1\}$, $1/x$ for x ranging from 1 to 100
		Hidden layers	$H \in \{(100,), (100,100,100,100,100), (50,), (\underline{50,50}), (30,30), (60,60), (30,20,10), (20,20), (25,15), (10,40,10), (20,20,20), (30,10), (10,10,10)\}$, used many possibilities
	Search algorithm	Grid search	
Data split	5-fold cross-validation		
Algorithm	Multi-layer perceptron		
Sampling	80% training, 20% test		
Performance evaluation	ROC AUC score on test data: 0.5662		
<i>Note:</i> Bold writing indicates a problem characteristic or choice from the report card. Underlined writing indicates final parameters from optimization.			

3.2. SMLR of economic value

Table 4: Supervised machine learning report card of the best performing algorithm for economic value

Model initiation			
Problem statement	Predict whether a patent belongs to the top 10% of KPSS values on 12 novelty variables and 15 control variables		
Data gathering	All variables are calculated on 263,960 US patents which are classified in this CPC and granted between January 1976 and December 2009		
Data distribution	15,753 patents belong to the top 10% class, 141,770 patents belong to the remaining class		
Sampling	No sampling		
Data quality	106,437 patents are missing KPSS values and therefore are removed from the sample (final sample size: 157,523)		
Data preprocessing methods	Max-min scaling and standardization		
Feature engineering and vectorizing	No additional features apart from 12 novelty and 15 control variables		
Performance estimation			
Parameter optimization	Yes		
	Search space	Solver	$S \in \{\text{'adam'}\}$, works well on large datasets
		Activation function	$A \in \{\text{'identity', 'logistic', 'tanh', 'relu'}\}$, used all possibilities
		L2 penalty	$L2 \in \{1/100, 1/99, 1/98, \dots, 1/17, \dots, 1/1\}$, 1/x for x ranging from of 1 to 100
		Hidden layers	$H \in \{(100,), (100,100,100,100,100), (50,), (50,50), (30,30), (60,60), (30,20,10), (20,20), (25,15), (10,40,10), (20,20,20), (30,10), (10,10,10)\}$, used many possibilities
	Search algorithm	Grid search	
Data split	5-fold cross-validation		
Algorithm	Multi-layer perceptron		
Sampling	80% training, 20% test		
Performance evaluation	ROC AUC score on test data: 0.6767		
<i>Note:</i> Bold writing indicates a problem characteristic or choice from the report card. Underlined writing indicates final parameters from optimization.			

4. Results without control variables

For robustness checks, we conducted the subsequent steps without control variables. Model evaluation results show slightly less fit, however, results from model interpretation remain robust. The results are depicted in the subsequent tables and figures.

Table 5: Evaluation results of classification without control

Note: ROC AUC validation results are based on 5-fold cross validation.

Target	Model	ROC AUC validation data	ROC AUC test data	ROC AUC training data
<i>Technological value</i>	MLP	0.7622	0.5180	0.5187
	RF	0.7435	0.5000	0.5000
	DT	0.7056	0.5000	0.5000
<i>Economic value</i>	MLP	0.6651	0.5006	0.5005
	RF	0.6539	0.5000	0.5000
	DT	0.6181	0.5000	0.5000

4.1. Technological value without control variables

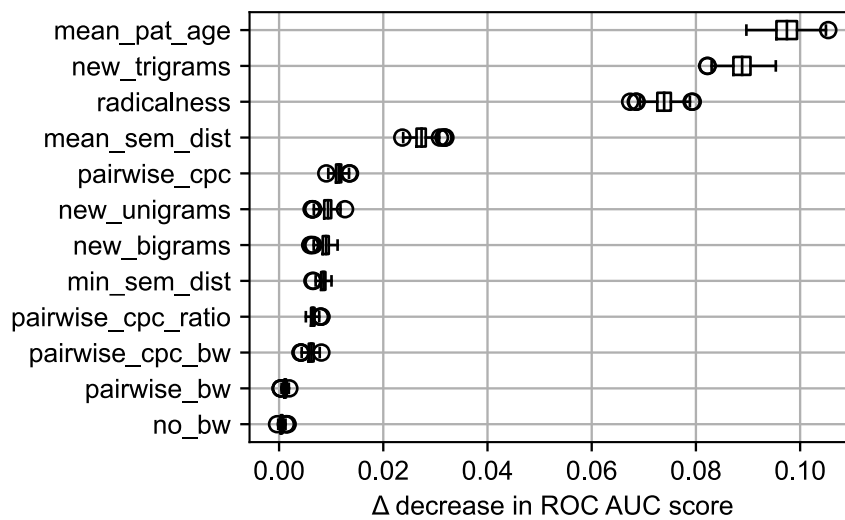


Figure 1: Permutation importance results of technological value without control variables.
Source: Author.

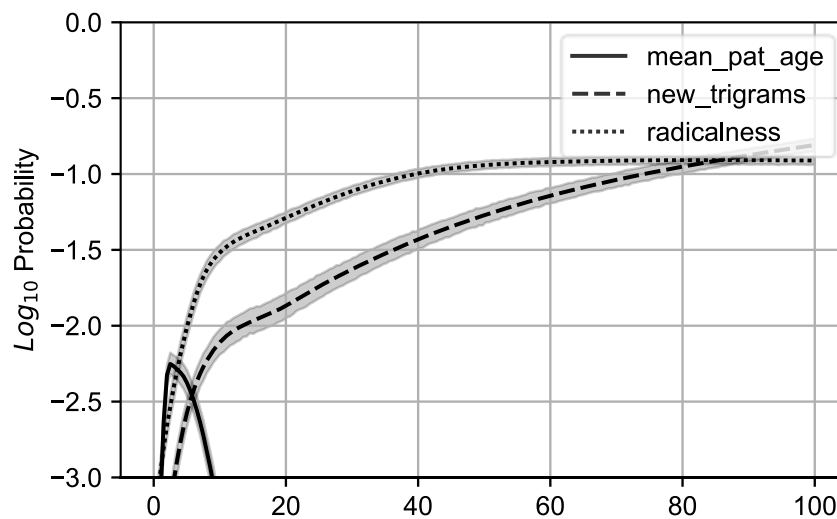


Figure 2: PDP with 95% confidence intervals estimating technological value without control variables.
Source: Author.

4.2. Economic value without control variables

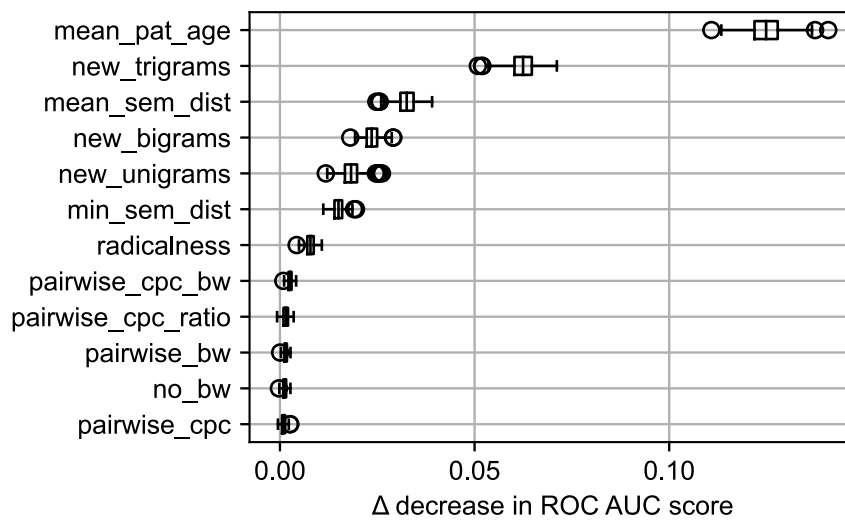


Figure 3: Permutation importance results of economic value without control variables.
Source: Author.

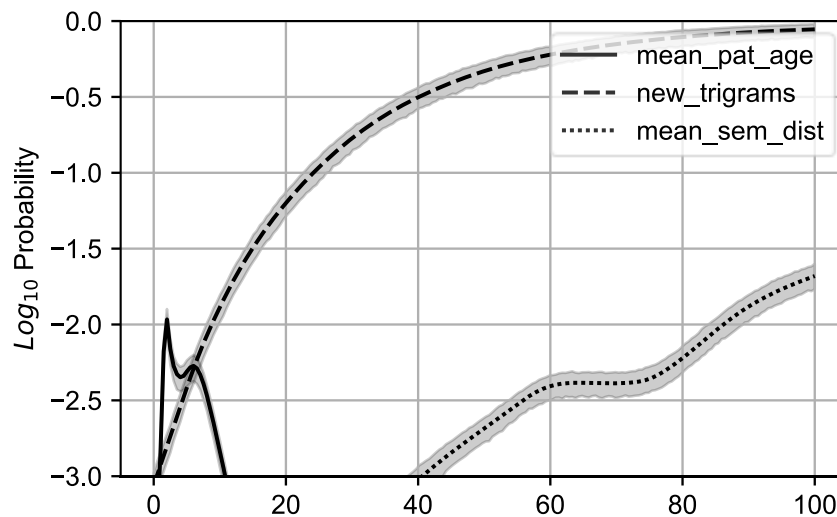


Figure 4: PDP with 95% confidence intervals estimating economic value without control variables.
Source: Author.

5. Results from variation of value variables

For robustness checks, we conducted the subsequent steps with variation of value variables. First, we varied the time span of technological value to five years and ten years after patent publication. Second, we varied the percentile to the top 1 percent for both, technological and economic value. Despite minor changes in magnitude, the results remain robust. The results are depicted in the subsequent tables and figures.

Table 6: Evaluation results of classification with variation in target variables

Note: ROC AUC validation results are based on 5-fold cross validation.

Target	Model	ROC AUC validation data	ROC AUC test data	ROC AUC training data
<i>Technological value (5 years, top 10%)</i>	MLP	0.8169	0.5611	0.5696
	RF	0.7962	0.5032	0.5031
	DT	0.7626	0.5142	0.5148
<i>Technological value (10 years, top 10%)</i>	MLP	0.8178	0.5511	0.5556
	RF	0.7990	0.5023	0.5033
	DT	0.7614	0.5226	0.5195
<i>Technological value (5 years, top 1%)</i>	MLP	0.8843	0.5019	0.5043
	RF	0.8683	0.5000	0.5000
	DT	0.8140	0.5000	0.5000
<i>Technological value (7 years, top 1%)</i>	MLP	0.8906	0.5047	0.5042
	RF	0.8680	0.5000	0.5000
	DT	0.8175	0.5000	0.5000
<i>Technological value (10 years, top 1%)</i>	MLP	0.8895	0.5009	0.5009
	RF	0.8743	0.5000	0.5000
	DT	0.8239	0.5000	0.5000
<i>Economic value (top 1%)</i>	MLP	0.9388	0.5000	0.5000
	RF	0.9304	0.5079	0.5067
	DT	0.9020	0.5000	0.5000

5.1. Technological value (5 years, top 10%)

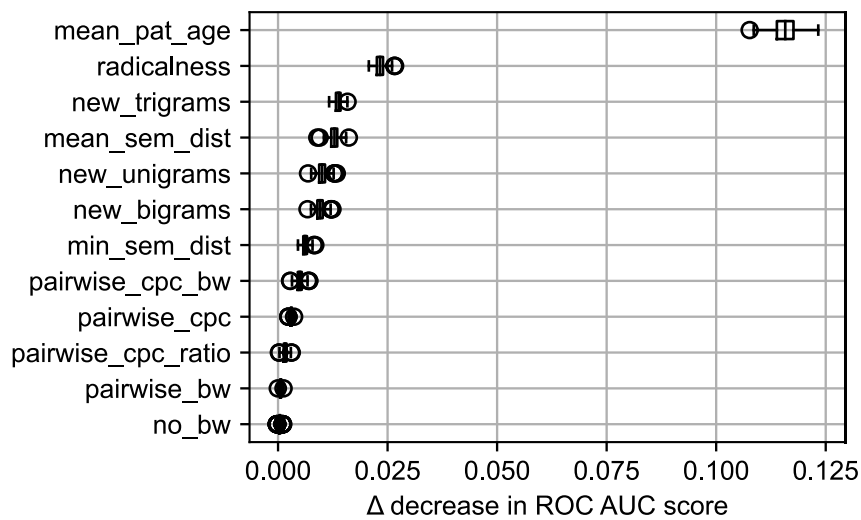


Figure 5: Permutation importance results of technological value (5 years, top 10%).
Source: Author.

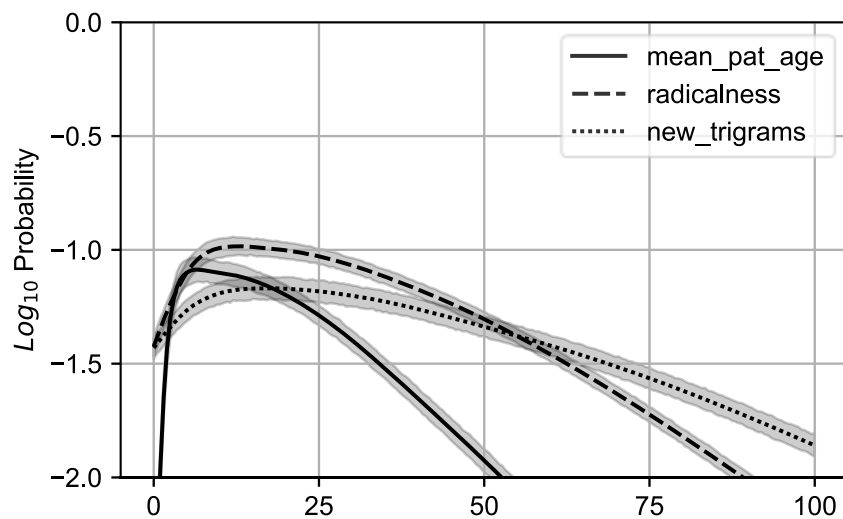


Figure 6: PDP with 95% confidence intervals estimating technological value (5 years, top 10%).
Source: Author.

5.2. Technological value (10 years, top 10%)

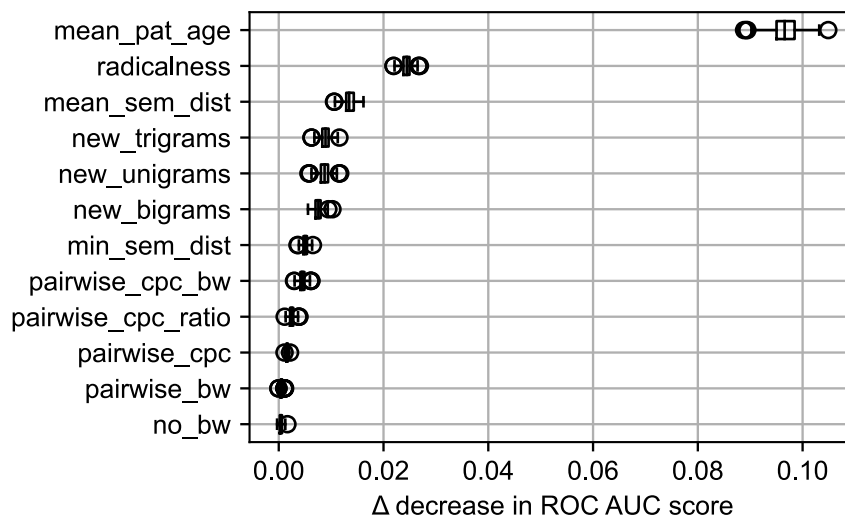


Figure 7: Permutation importance results of technological value (10 years, top 10%).
Source: Author.

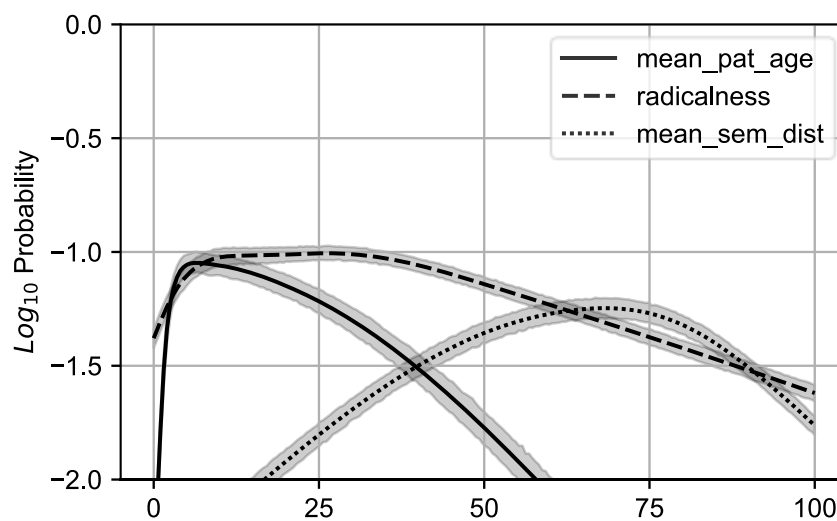


Figure 8: PDP with 95% confidence intervals estimating technological value (10 years, top 10%).
Source: Author.

5.3. Technological value (5 years, top 1%)

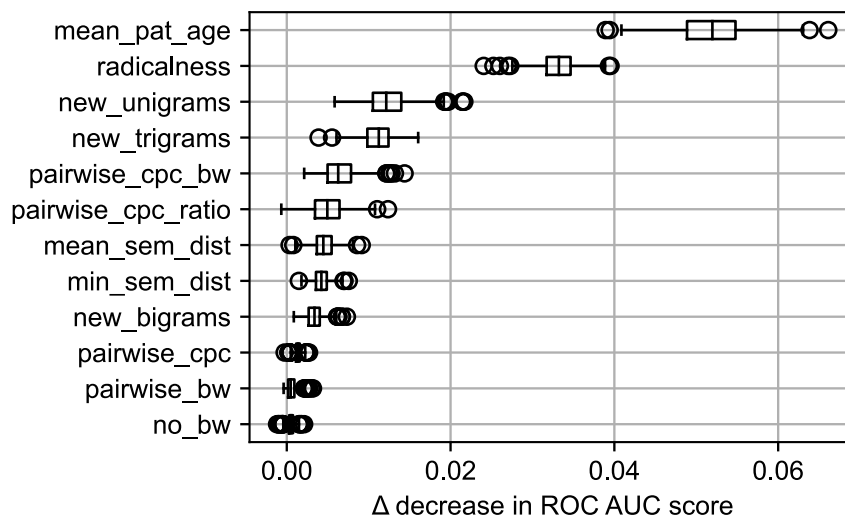


Figure 9: Permutation importance results of technological value (5 years, top 1%).
Source: Author.

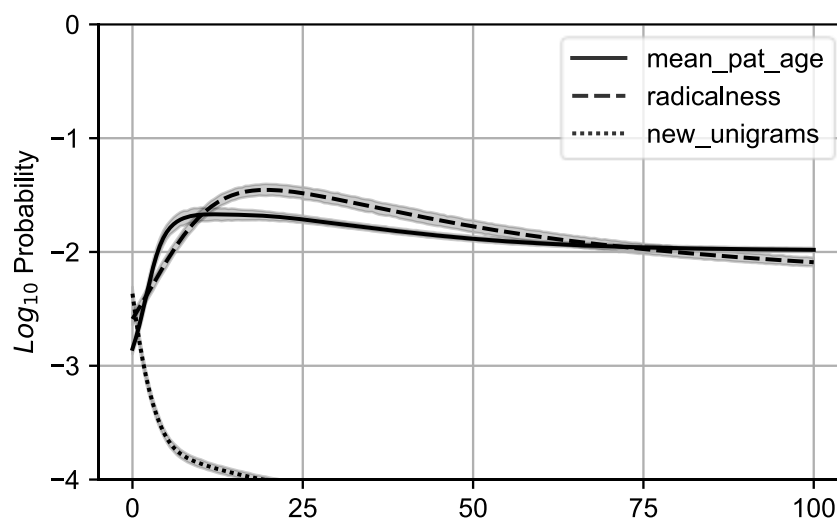


Figure 10: PDP with 95% confidence intervals estimating technological value (5 years, top 1%).
Source: Author.

5.4. Technological value (7 years, top 1%)

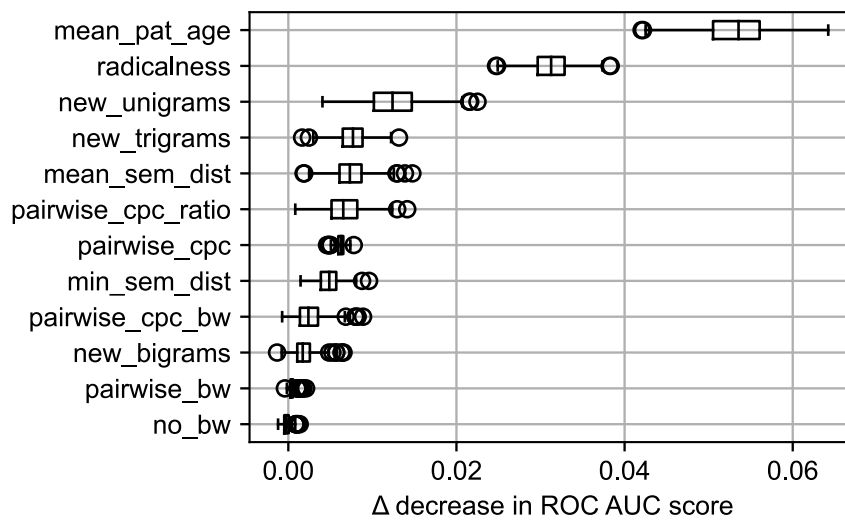


Figure 11: Permutation importance results of technological value (7 years, top 1%).
Source: Author.

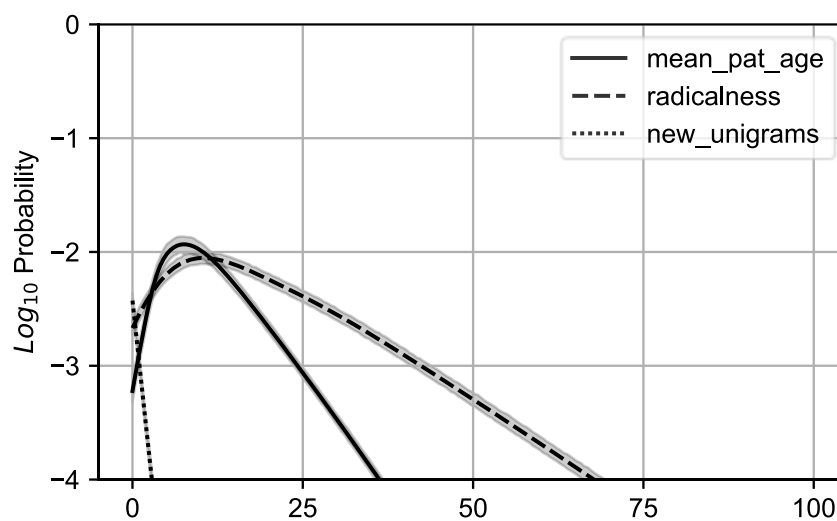


Figure 12: PDP with 95% confidence intervals estimating technological value (7 years, top 1%).
Source: Author.

5.5. Technological value (10 years, top 1%)

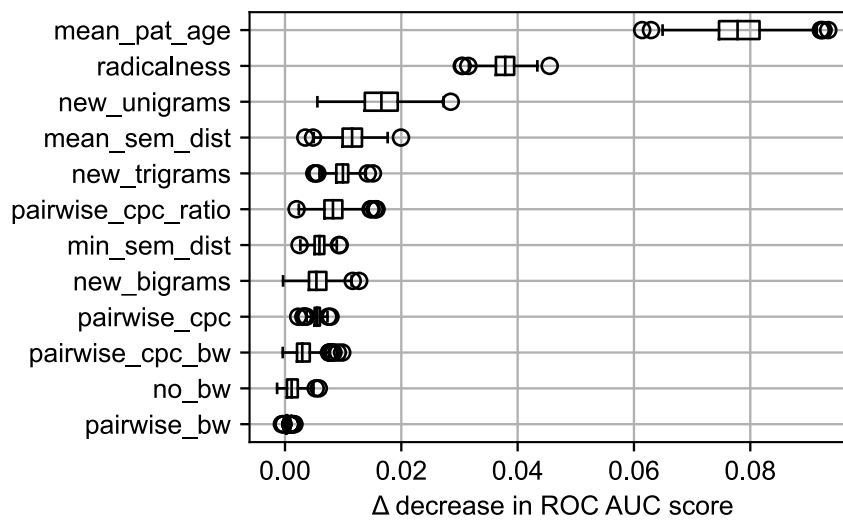


Figure 13: Permutation importance results of technological value (10 years, top 1%).
Source: Author.

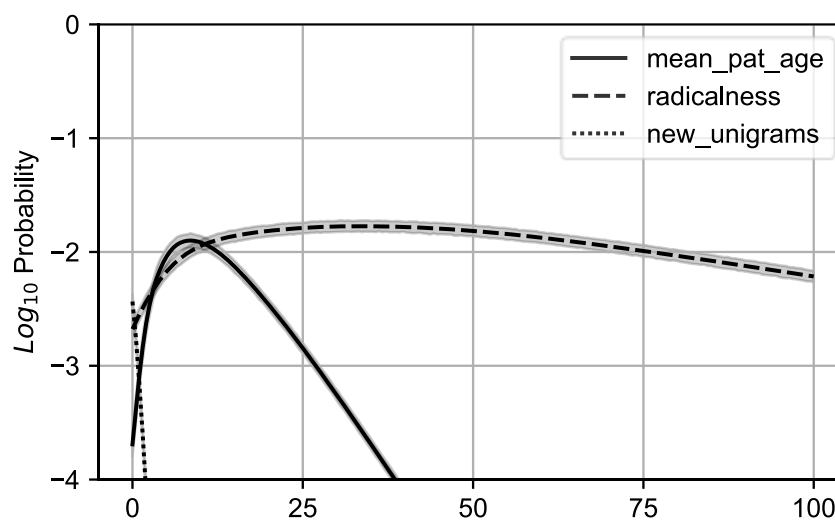


Figure 14: PDP with 95% confidence intervals estimating technological value (10 years, top 1%).
Source: Author.

5.6. Economic value (top 1%)

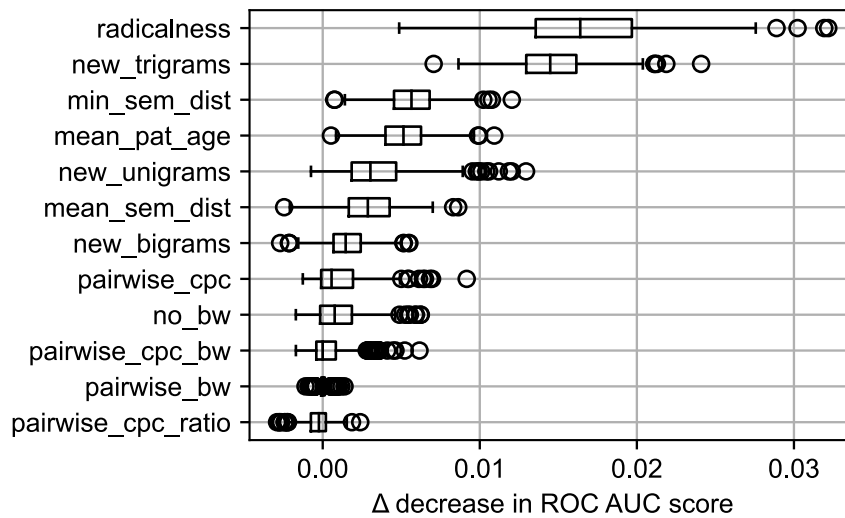


Figure 15: Permutation importance results of economic value (top 1%).
Source: Author.

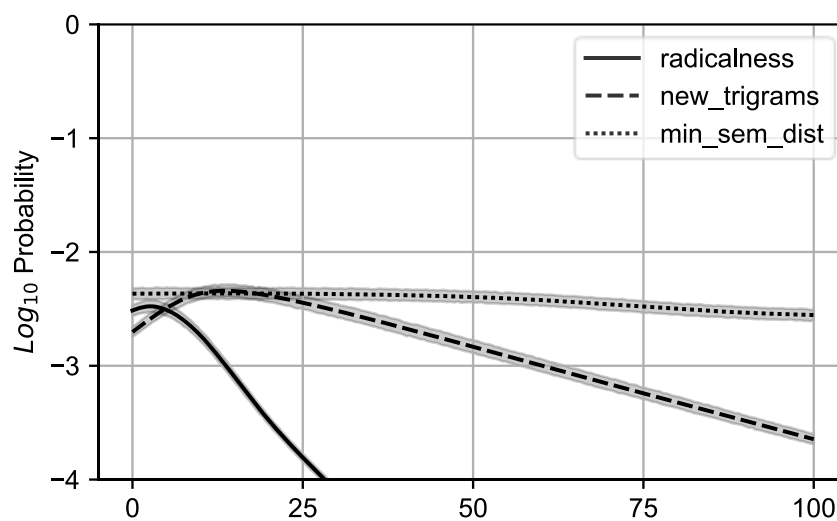


Figure 16: PDP with 95% confidence intervals estimating economic value (top 1%).
Source: Author.

References

- Adams, S. (2010). The text, the full text and nothing but the text: Part 1 - Standards for creating textual information in patent documents and general search implications. *World Patent Information*, 32, 22–29. doi:10.1016/j.wpi.2009.06.001.
- Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22, 521–543. doi:10.1002/SMJ.176.
- Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, 38, 415–427. doi:10.1016/j.respol.2008.12.001.
- Arts, S., & Fleming, L. (2018). Paradise of Novelty—Or Loss of Human Capital? Exploring New Fields and Inventive Output. *Organization Science*, 29, 1074–1092. doi:10.1287/orsc.2018.1216.
- Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*. doi:10.1016/j.respol.2020.104144.
- Arts, S., & Veugelers, R. (2015). Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change*, 24, 1215–1246. doi:10.1093/icc/dtu029.
- Barbieri, N., Marzucchi, A., & Rizzo, U. (2020). Knowledge sources and impacts on subsequent inventions: Do green technologies differ from non-green ones? *Research Policy*. doi:10.1016/j.respol.2019.103901.
- Denter, N. M., Aaldering, L. J., & Caferoglu, H. (2022). Forecasting future bigrams and promising patents: Introducing text-based link prediction. *Foresight*. doi:10.1108/FS-03-2021-0078.
- Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science*, 47, 117–132. doi:10.1287/mnsc.47.1.117.10671.
- Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, 25, 909–928. doi:10.1002/smj.384.
- Galasso, A., & Schankerman, M. (2015). Patents and Cumulative Innovation: Causal Evidence from the Courts. *The Quarterly Journal of Economics*, 130, 317–369. doi:10.1093/qje/qju029.
- Gassmann, O., Bader, M. A., & Thompson, M. (2021). *Patent Management: Protecting Intellectual Property and Innovation* (Management for Professionals). Cham: Springer International Publishing.
- Huang, Y., Chen, L., & Zhang, L. (2020). Patent citation inflation: The phenomenon, its measurement, and relative indicators to temper its effects. *Journal of Informetrics*. doi:10.1016/j.joi.2020.101015.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, 3, 303–320. doi:10.1257/aeri.20190499.
- Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics*, 132, 665–712. doi:10.1093/qje/qjw040.

- Kühl, Niklas, Hirt, Robin, Baier, L., Schmitz, B., & Satzger, G. (2021). How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine Learning Report Card. *Commun. ACM*, 48, 589–615. doi:10.17705/1CAIS.04845.
- Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, 44, 684–697. doi:10.1016/j.respol.2014.10.007.
- Moehrl, M. G., & Gerken, J. M. (2012). Measuring textual patent similarity on the basis of combined concepts: Design decisions and their consequences. *Scientometrics*, 91, 805–826. doi:10.1007/s11192-012-0682-0.
- Shane, S. (2001). Technological Opportunities and New Firm Creation. *Management Science*, 47, 205–220. doi:10.1287/mnsc.47.2.205.9837.
- Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, 115, 131–142. doi:10.1016/j.techfore.2016.09.028.
- Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342, 468–471. doi:10.1126/science.1240474.
- Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, 45, 707–723. doi:10.1016/j.respol.2015.11.010.
- Wang, T., Libaers, D., & Park, H. D. (2017). The Paradox of Openness: How Product and Patenting Experience Affect R&D Sourcing in China? *Journal of Product Innovation Management*, 34, 250–268. doi:10.1111/jpim.12359.
- Wu, L., Hitt, L., & Lou, B. (2020). Data Analytics, Innovation, and Firm Productivity. *Management Science*, 66, 2017–2039. doi:10.1287/mnsc.2018.3281.