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

Supplementary Material

1 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 finals models in industry, we only report insights into model initiiation and performance estimation.

1.1 SMLR of Technological Value

<table>
<tr><th colspan="4">Model initiation</th></tr>
<tr><td>Problem statement</td><td colspan="3">Predict whether a patent belongs to the top 10% of forward citations received within 7 years since grant on 12 novelty variables and 8 control variables</td></tr>
<tr><td>Data gathering</td><td colspan="3">All variables are calculated on 263,960 US patents which are classified in this CPC and granted between January 1976 and Dezember 2009</td></tr>
<tr><td>Data distribution</td><td colspan="3">26,899 patents belong to the top 10% class,
237,061 patents belong to the remaining class</td></tr>
<tr><td>Sampling</td><td colspan="3">No sampling</td></tr>
<tr><td>Data quality</td><td colspan="3">No missing values</td></tr>
<tr><td>Data preprocessing methods</td><td colspan="3">Max-min scaling and standardization</td></tr>
<tr><td>Feature engineering and vectorizing</td><td colspan="3">No additional features apart from
12 novelty and 8 control variables</td></tr>
<tr><th colspan="4">Performance estimation</th></tr>
<tr><td rowspan="6">Parameter optimization</td><td colspan="3">Yes</td></tr>
<tr><td rowspan="4">Search space</td><td>Solver</td><td>S ϵ {<u>‘adam’</u>},
works well on large datasets</td></tr>
<tr><td>Activation function</td><td>A ϵ {‘identity’, ‘logistic’, ‘tanh’, <u>‘relu’</u>},
used all possibilities</td></tr>
<tr><td>L2 penalty</td><td>L2 ϵ {1/100, 1/99, 1/98, …, <u>1/27</u>, …, 1/1},
1/x for x ranging from of 1 to 100</td></tr>
<tr><td>Hidden layers</td><td>H ϵ {(100,), (100,100,100,100,100), (50,), (50,50), (30,30), (60,60), <u>(30,20,10)</u>, (20,20), (25,15), (10,40,10), (20,20,20), (30,10), (10,10,10)},
used many possibilities</td></tr>
<tr><td>Search algorithm</td><td colspan="2">Grid search</td></tr>
<tr><td>Data split</td><td colspan="3">5-fold cross-validation</td></tr>
<tr><td>Algorithm</td><td colspan="3">Multi-layer perceptron</td></tr>
<tr><td>Sampling</td><td colspan="3">80% training, 20% test</td></tr>
<tr><td>Performance evaluation</td><td colspan="3">ROC AUC score on test data: 0.5552</td></tr>
<tr><td colspan="4">Note:
Bold writing indicates a problem characteristic or choice from the report card.
Underlined writing indicates final parameters from optimization.</td></tr>
</table>

1.2 SMLR of Economical Value

Model initiation			
Problem statement	Predict whether a patent belongs to the top 10% of KPSS values on 12 novelty variables and 8 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 Dezember 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 8 control variables		
Performance estimation			
Parameter optimization	Yes		
	Search space	Solver	S ϵ {‘adam’}, works well on large datasets
		Activation function	A ϵ {‘identity’, ‘logistic’, ‘tanh’, ‘relu’}, used all possibilities
		L2 penalty	L2 ϵ {1/100, 1/99, 1/98, …, 1/17, …, 1/1}, 1/x for x ranging from of 1 to 100
		Hidden layers	H ϵ {(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: 0.5016		
Note: Bold writing indicates a problem characteristic or choice from the report card. Underlined writing indicates final parameters from optimization.			

2 Results without control variables

For robustness checks, I 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 1: Evaluation results of classification without control variables including difference to models with control variables
Note: SD refers to standard deviation.

Target	**Model**	**ROC AUC Validation Data**		**ROC AUC**
		Mean	**SD**	**Test Data**
Technological Value	MLP	0.7622	0.0031	0.5180
	RF	0.7435	0.0039	0.5000
	DT	0.7056	0.0056	0.5000
Economical Value	MLP	0.6651	0.0053	0.5006
	RF	0.6539	0.0058	0.5000
	DT	0.6181	0.0083	0.5000

2.1 Technological Value without control variables

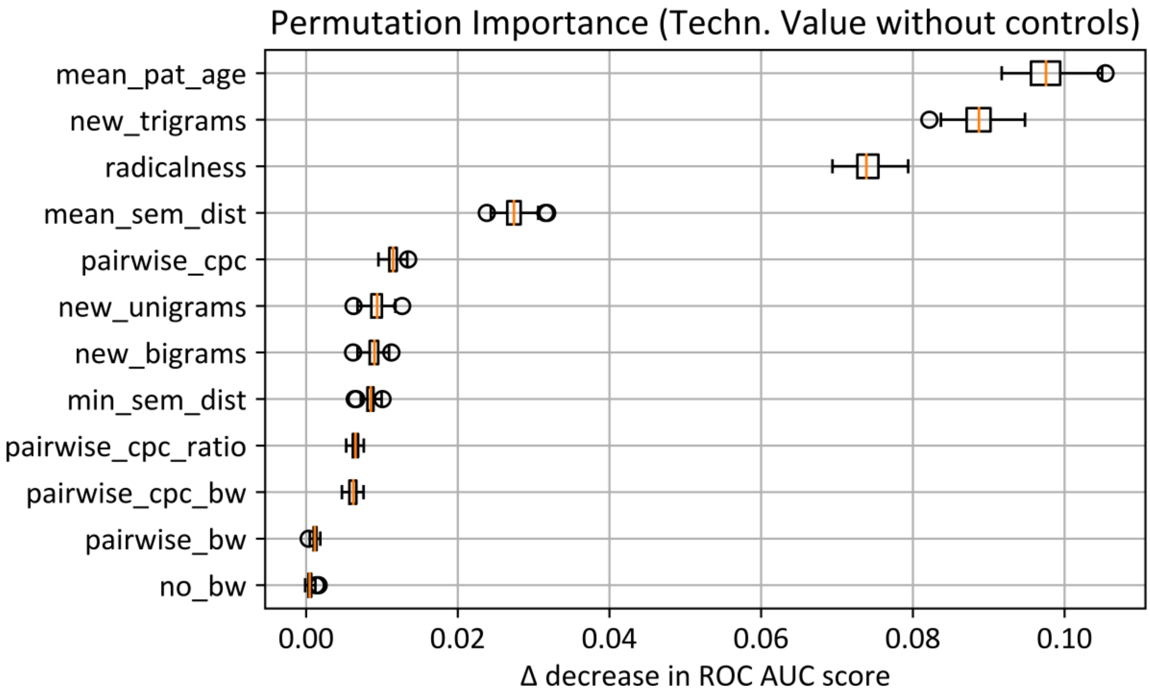

Figure 1: Permutation Importance Results of Technological Value without control variables.
Source: Author.

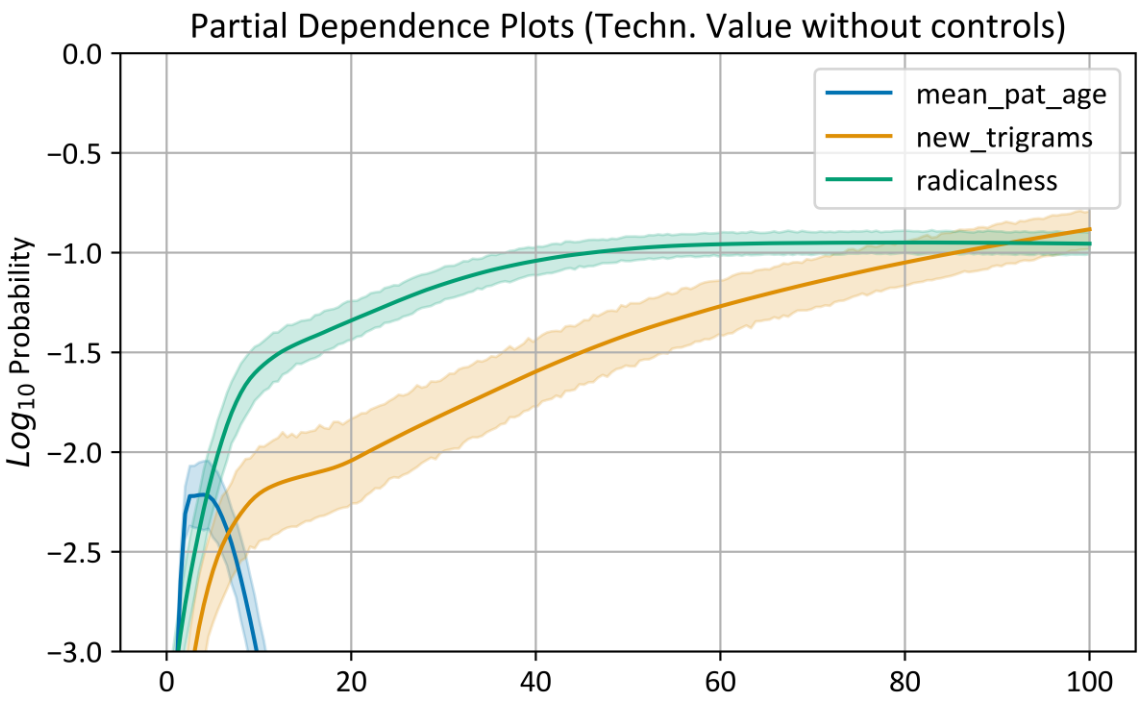

Figure 2: PDP with 95% confidence intervals estimating Technological Value without control variables.
Source: Author.

2.2 Economical Value without control variables

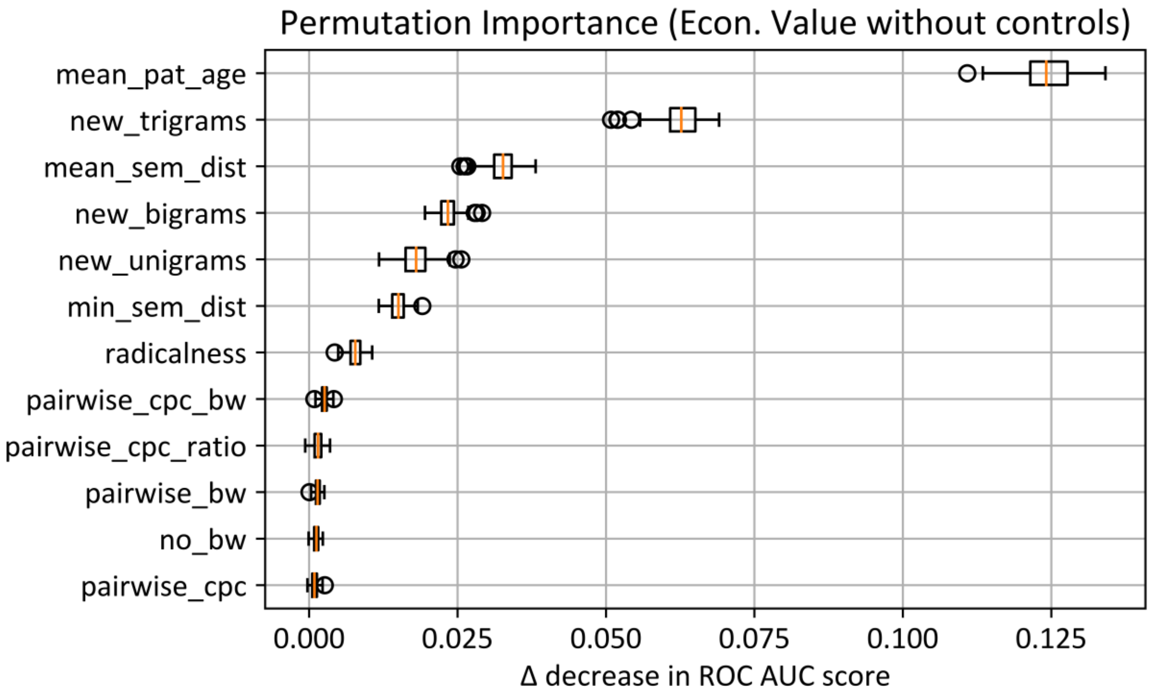

Figure 3: Permutation Importance Results of Economical Value without control variables.
Source: Author.

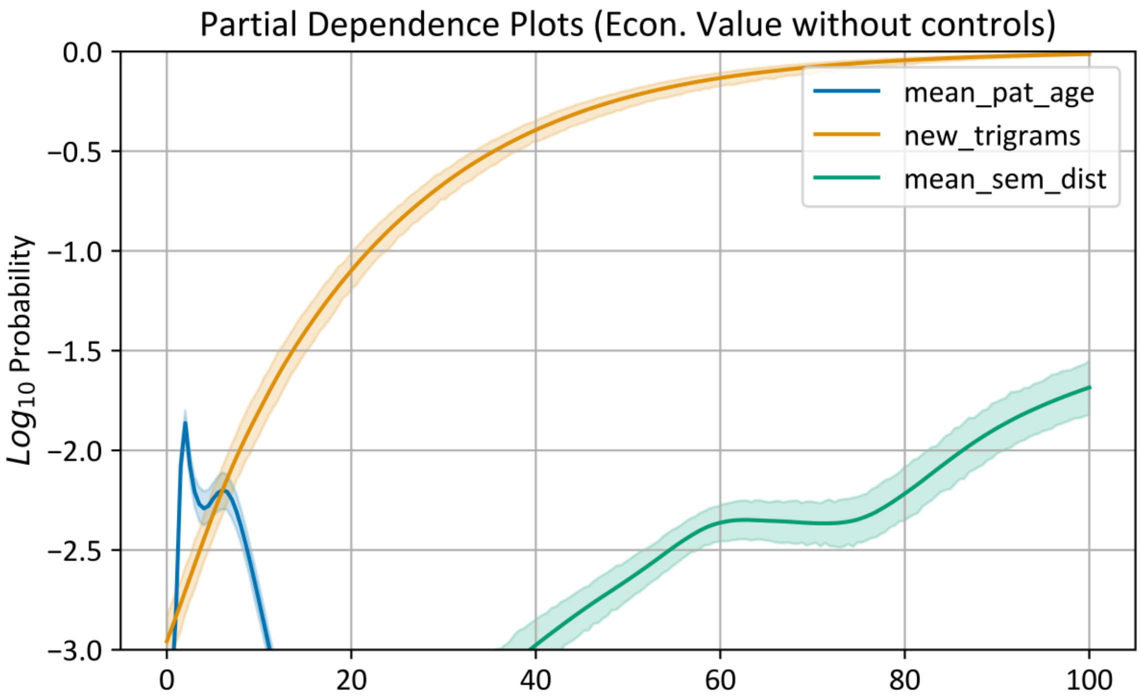

Figure 4: PDP with 95% confidence intervals estimating Economical Value without control variables.
Source: Author.

3 Results from variation of value variables

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

Table 2: Evaluation results of classification with variation in target variables
Note: SD refers to standard deviation.

Target	**Model**	**ROC AUC Validation Data**		**ROC AUC**
		Mean	**SD**	**Test Data**
Technological Value (5 years, top 10%)	MLP	0.8054	0.0009	0.5552
	RF	0.7829	0.0022	0.5020
	DT	0.7485	0.0032	0.5091
Technological Value (10 years, top 10%)	MLP	0.8058	0.0033	0.5562
	RF	0.7844	0.0018	0.5020
	DT	0.7466	0.0018	0.5108
Technological Value (5 years, top 1%)	MLP	0.8726	0.0074	0.5140
	RF	0.8498	0.0149	0.5000
	DT	0.7874	0.0144	0.5000
Technological Value (7 years, top 1%)	MLP	0.8738	0.0116	0.5064
	RF	0.8464	0.0107	0.5000
	DT	0.7820	0.0117	0.5000
Technological Value (10 years, top 1%)	MLP	0.8752	0.0147	0.5038
	RF	0.8500	0.0087	0.5000
	DT	0.7963	0.0121	0.5000
Economical Value (top 1%)	MLP	0.8540	0.0100	0.5032
	RF	0.8409	0.0116	0.5000
	DT	0.7998	0.0108	0.5000

3.1 Technological Value (5 years, top 10%)

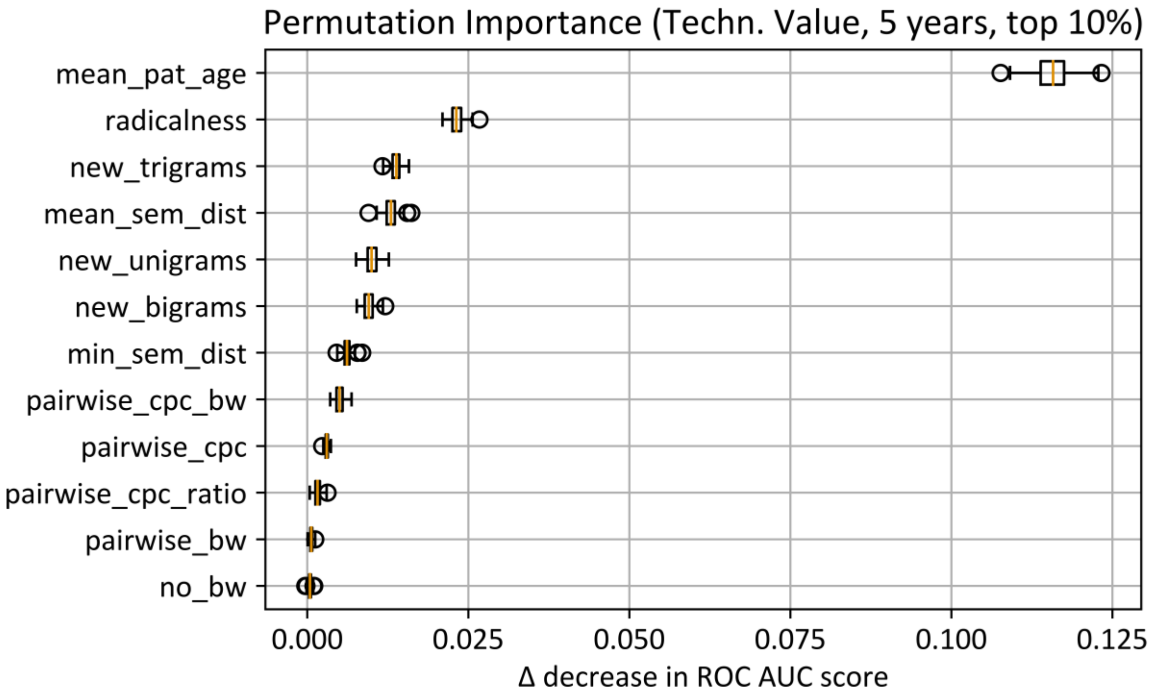

Figure 5: Permutation Importance Results of Technological Value (5 years, top 10%).
Source: Author.

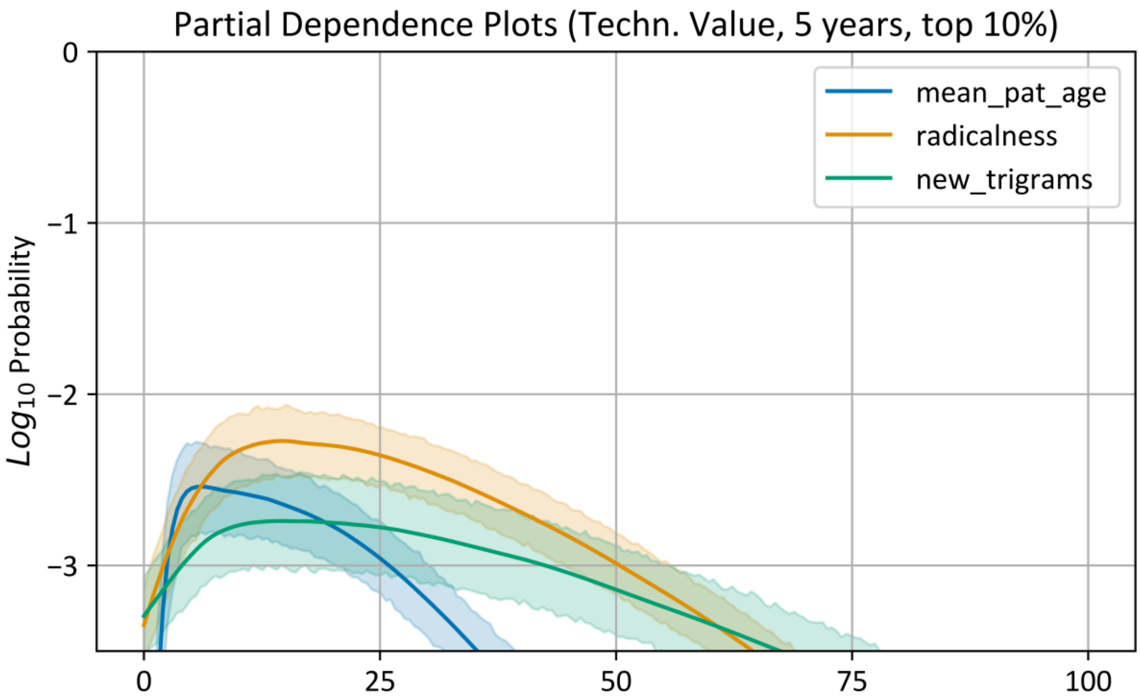

Figure 6: PDP with 95% confidence intervals estimating Technological Value (5 years, top 10%).
Source: Author.

3.2 Technological Value (10 years, top 10%)

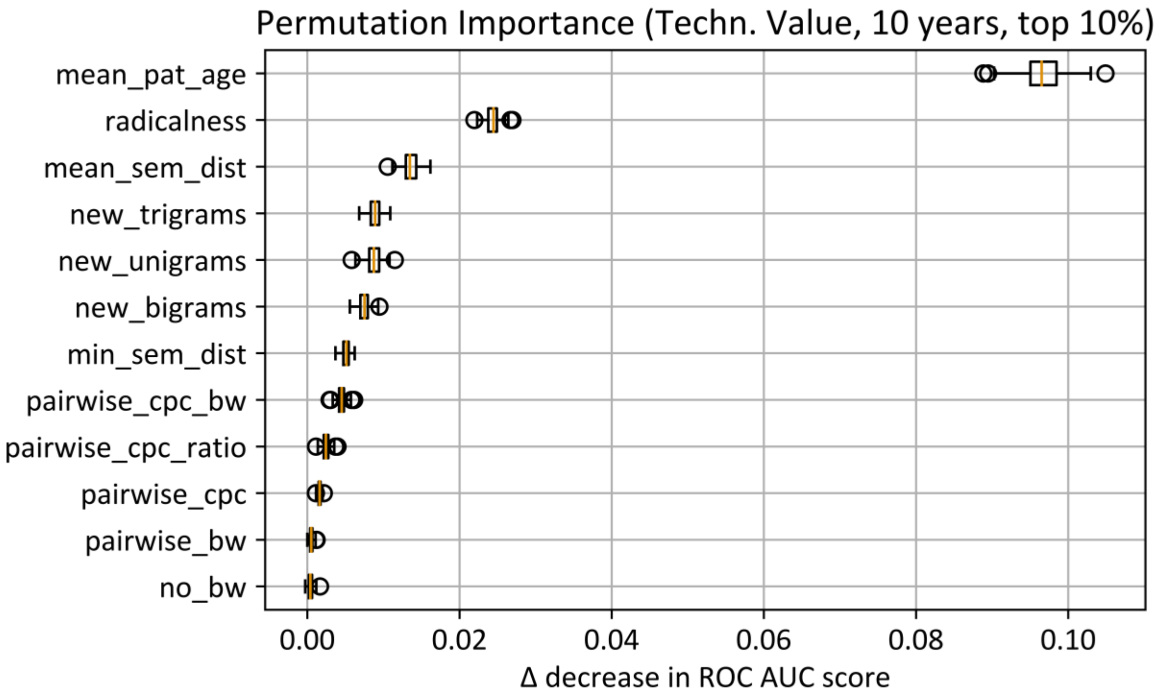

Figure 7: Permutation Importance Results of Technological Value (10 years, top 10%).
Source: Author.

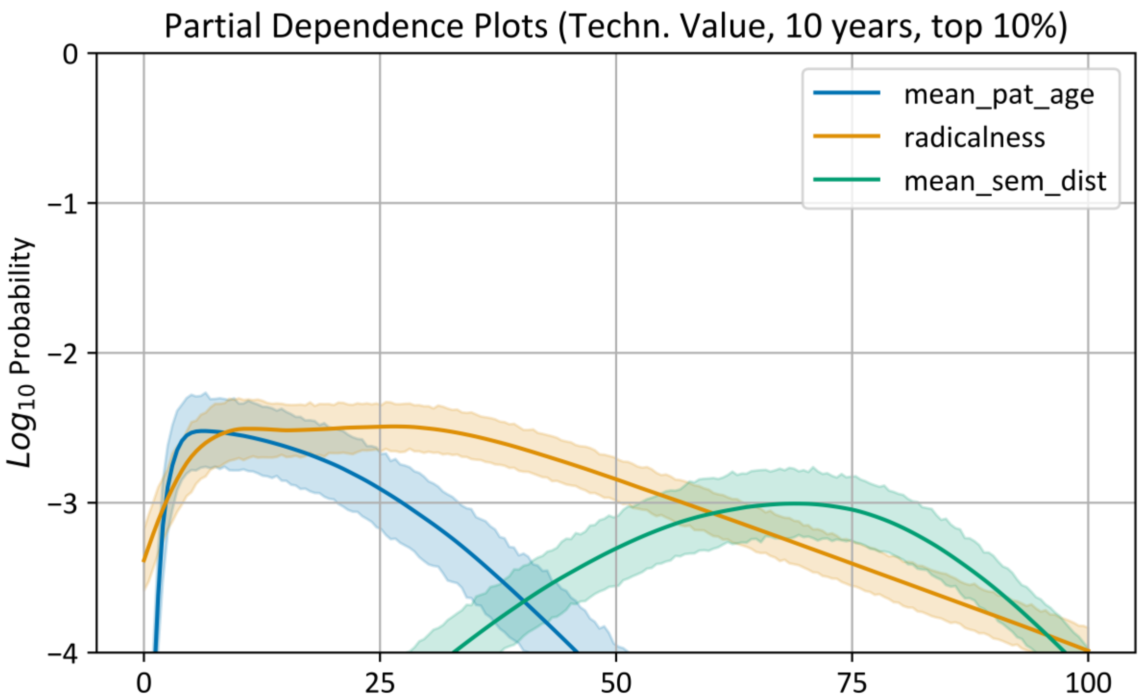

Figure 8: PDP with 95% confidence intervals estimating Technological Value (10 years, top 10%).
Source: Author.

3.3 Technological Value (5 years, top 1%)

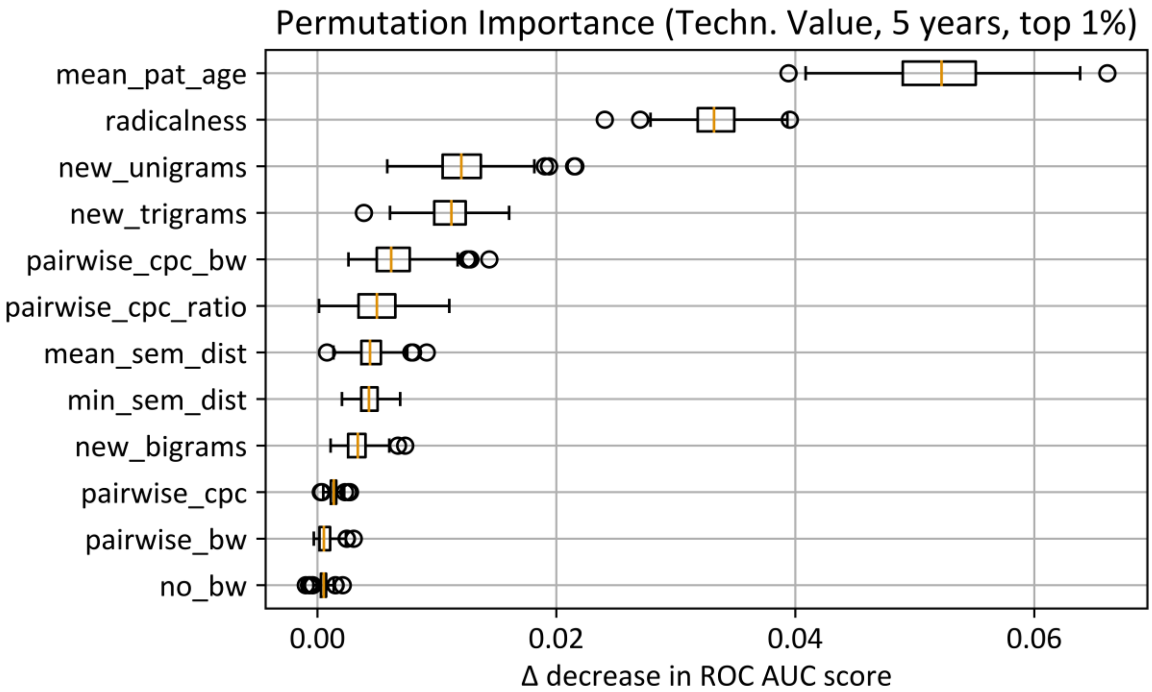

Figure 9: Permutation Importance Results of Technological Value (5 years, top 1%).
Source: Author.

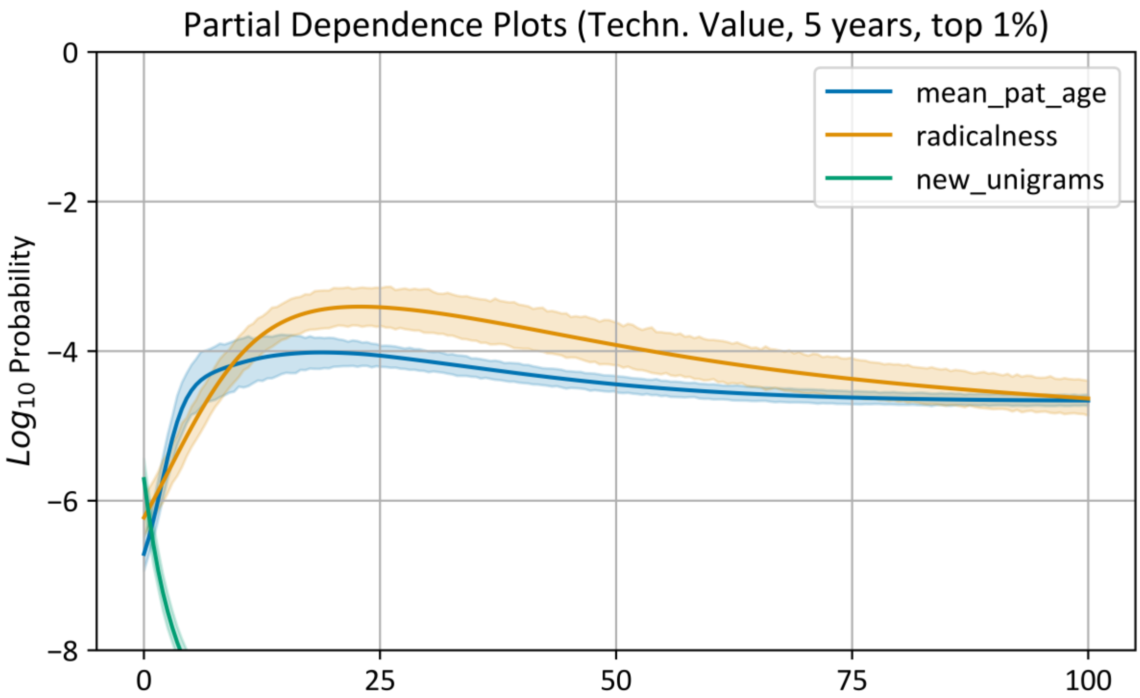

Figure 10: PDP with 95% confidence intervals estimating Technological Value (5 years, top 1%).
Source: Author.

3.4 Technological Value (7 years, top 1%)

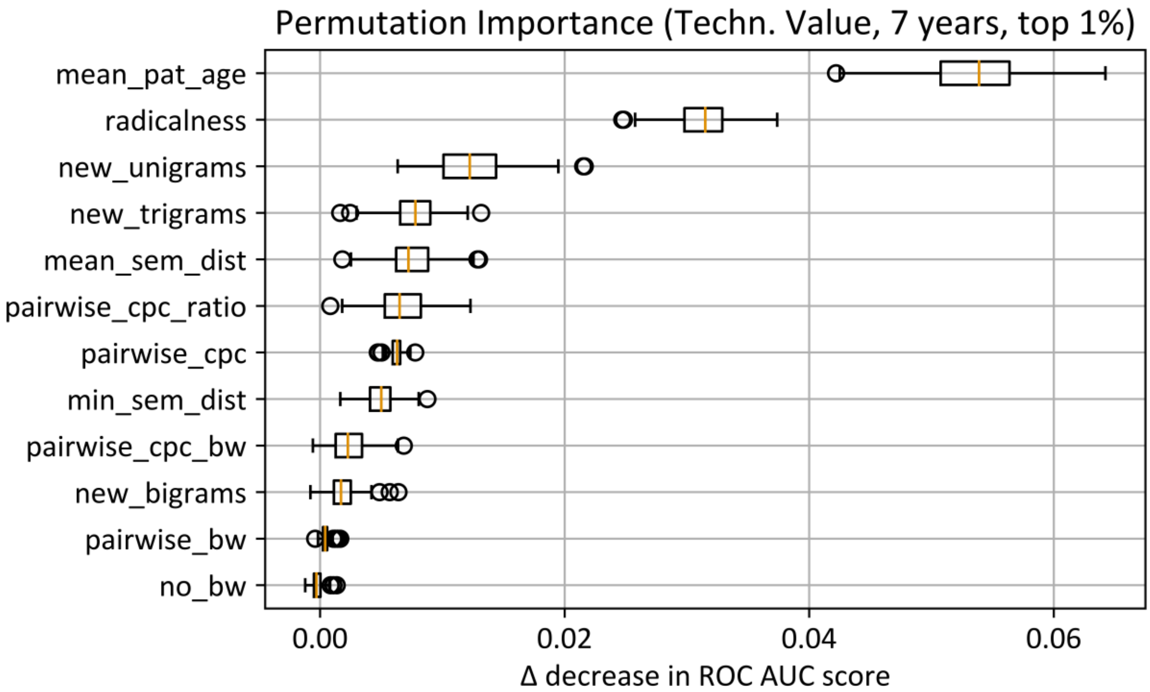

Figure 11: Permutation Importance Results of Technological Value (7 years, top 1%).
Source: Author.

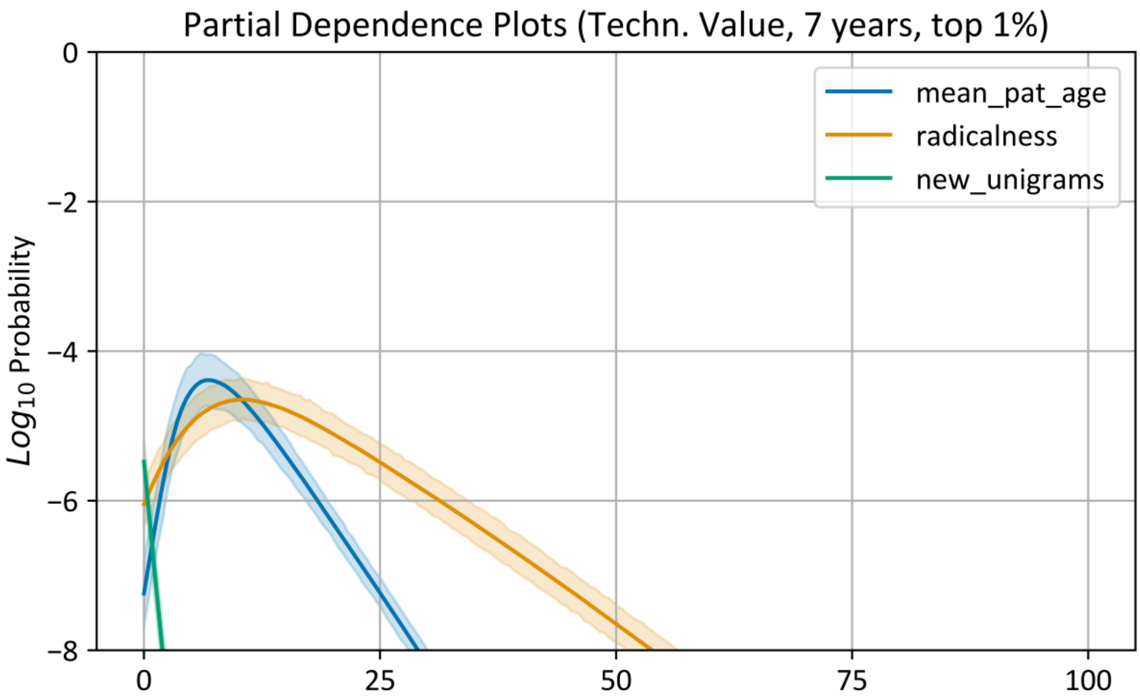

Figure 12: PDP with 95% confidence intervals estimating Technological Value (7 years, top 1%).
Source: Author.

3.5 Technological Value (10 years, 1%)

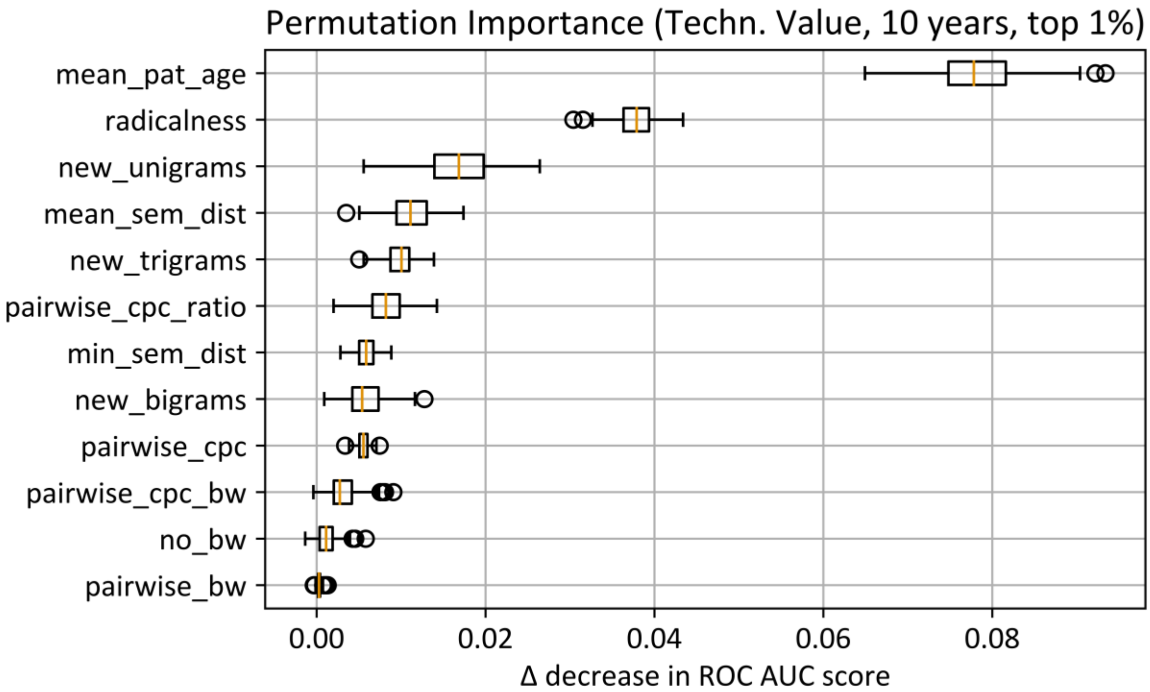

Figure 13: Permutation Importance Results of Technological Value (10 years, top 1%).
Source: Author.

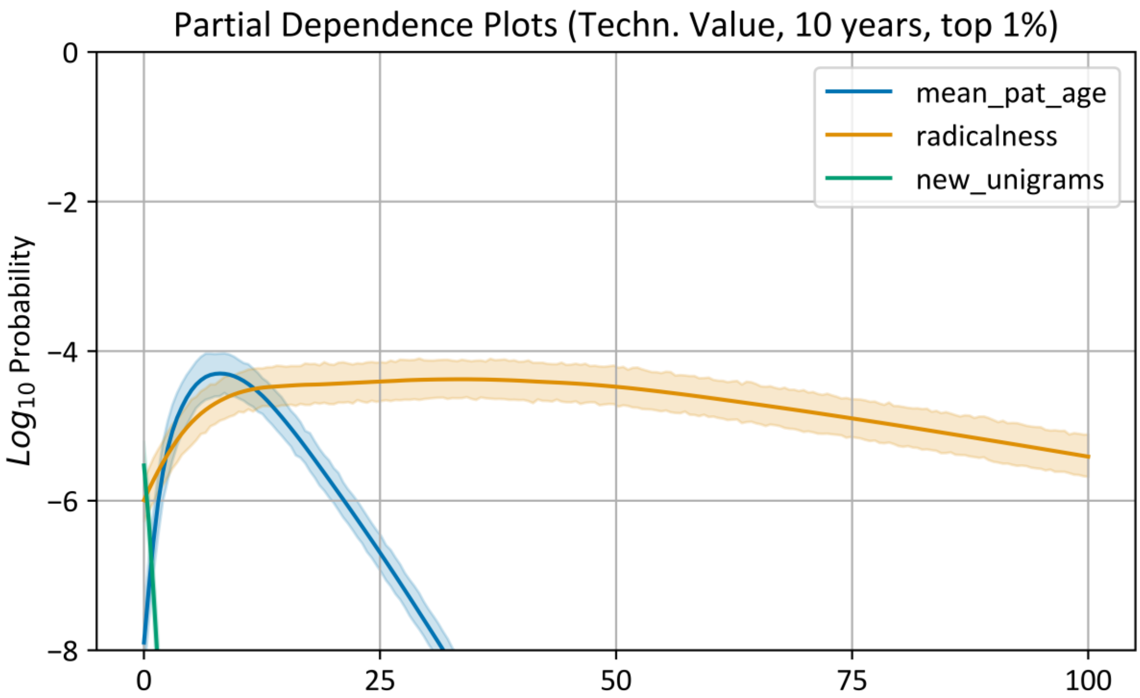

Figure 14: PDP with 95% confidence intervals estimating Technological Value (10 years, top 1%).
Source: Author.

3.6 Economical Value (1%)

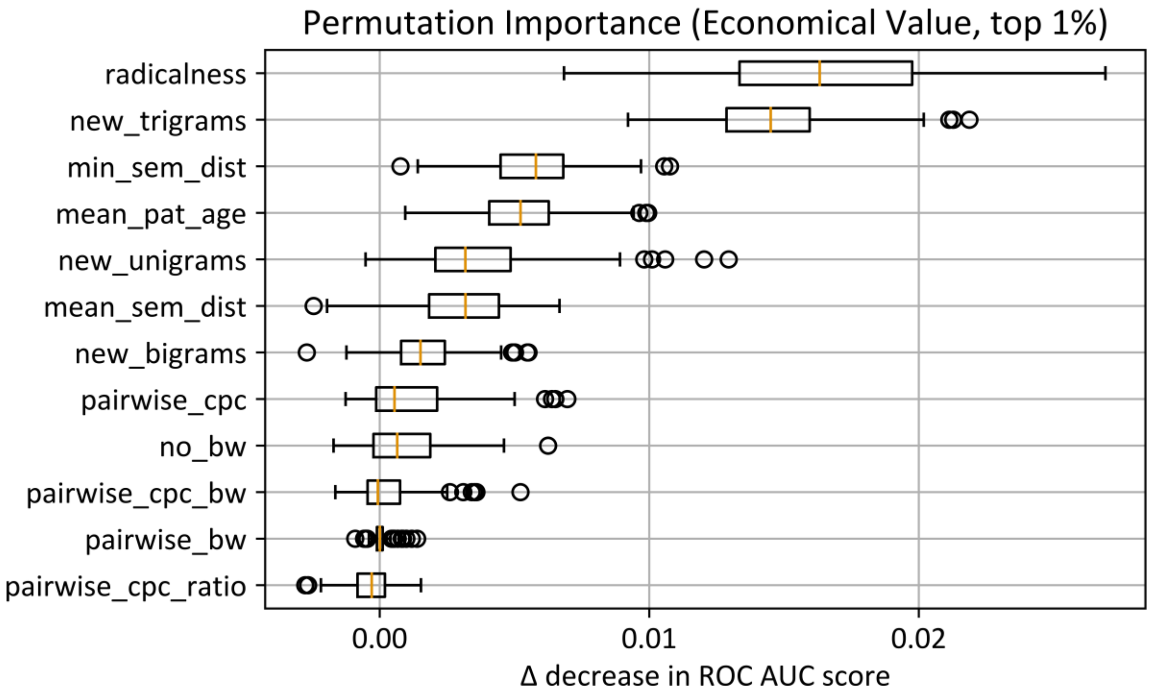

Figure 15: Permutation Importance Results of Economical Value (top 1%).
Source: Author.

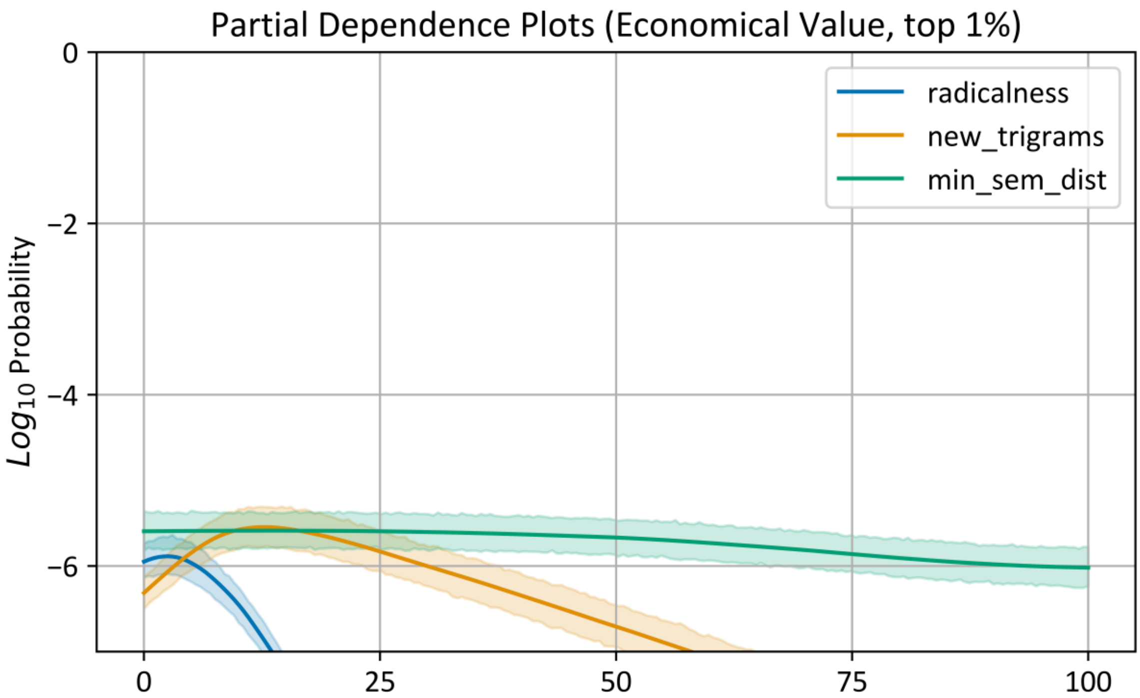

Figure 16: PDP with 95% confidence intervals estimating Economical Value (top 1%).
Source: Author.

References

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.