Data & Policy (2022), 4: e30
doi:10.1017/dap.2022.21

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A new approach to impact case study analytics

Jiajie Zhang[1], Paul Watson[1,2,*] and Barry Hodgson[1,3]

[1]School of Computing, Newcastle University, Newcastle upon Tyne, United Kingdom
[2]The Alan Turing Institute, London, United Kingdom
[3]National Innovation Centre for Data, Newcastle upon Tyne, United Kingdom
*Corresponding author. E-mail: paul.watson@newcastle.ac.uk

Received: 23 June 2021; **Revised:** 10 March 2022; **Accepted:** 15 August 2022

Key words: impact; infrastructure; knowledge generation; ontology; semantic web

Abstract

The 2014 Research Excellence Framework (REF) assessed the quality of university research in the UK. 20% of the assessment was allocated according to peer review of the impact of research, reflecting the growing importance of impact in UK government policy. Beyond academia, impact is defined as a change or benefit to the economy, society, culture, public policy or services, health, the environment, or quality of life. Each institution submitted a set of four-page impact case studies. These are predominantly free-form descriptions and evidences of the impact of study. Numerous analyses of these case studies have been conducted, but they have utilised either qualitative methods or primary forms of text searching. These approaches have limitations, including the time required to manually analyse the data and the frequently inferior quality of the answers provided by applying computational analysis to unstructured, context-less free text data. This paper describes a new system to address these problems. At its core is a structured, queryable representation of the case study data. We describe the ontology design used to structure the information and how semantic web related technologies are used to store and query the data. Experiments show that this gives two significant advantages over existing techniques: improved accuracy in question answering and the capability to answer a broader range of questions, by integrating data from external sources. Then we investigate whether machine learning can predict each case study's grade using this structured representation. The results provide accurate predictions for computer science impact case studies.

Policy Significance Statement

The 2014 Research Excellence Framework assessed the quality of university research in the UK. A fifth of the assessment was allocated according to peer review of the impact of research, reflecting its growing importance in UK government policy. Each university submitted a set of impact case studies that are mainly free text describing and evidencing the impact of research. Previous analysis of this data has either used qualitative methods or rudimentary forms of text searching and analysis. We present an alternative that enables policymakers and others to extract richer, more specific information from the case studies by creating a structured, queryable knowledge base using semantic web technologies. We then show how machine learning can accurately predict the grade awarded to Impact Case Studies in the Computer Science unit of assessment. These approaches can be applied to other areas of policymaking.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

1. Introduction

Since 1986, all universities in the United Kingdom have been reviewed and ranked according to the quality of their research outputs. These reviews are essential because the results play a significant role in determining the research funding allocated to each university. Exercises were conducted in 1986, 1989, 1992, 1996, 2001, and 2008 as the Research Assessment Exercise, and in 2014 as the Research Excellence Framework (REF). Each subject area ("unit of assessment") is peer-reviewed and ranked by a specialist panel. Until 2014, the review focused mainly on research outputs (e.g., publications) and the research environment. However, a major change was introduced in 2014, when 20% of the overall ranking was allocated according to a review of the research impact. The impact is defined as an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment, or quality of life, beyond academia. In the 2021 exercise, the weight given to impact will rise to 25%, reflecting the growing importance afforded to impact by the UK government.

To allocate a ranking for impact, each university submitted a set of four-page impact case studies for each unit of assessment—approximately one for every seven members of academic staff returned by a university in the unit of assessment. These were then peer-reviewed to determine a ranking. The impact case studies are mainly free text that describes and evidences the impact of the research conducted by one or more academics. After the 2014 exercise was complete, the rankings were published for each unit of assessment, along with all case studies. The rankings give the proportion of case studies that were graded in each of five categories (Higher Education Funding Council for England, 2015):

- world-leading (4*)
- internationally excellent (3*)
- internationally recognized (2*)
- nationally recognized (1*)
- unclassified.

Within the four-page restriction, there are five text sections, which are: "Summary of the impact," "Underpinning research," "References to the research," "Details of the impact," and "Sources to corroborate the impact." Each section of the case study follows the template given by REF criteria and explains the corresponding details within the indicative lengths. The open publication of all the case studies and their rankings has enabled analysis that has attempted to answer questions about them, including "what makes for an excellent case study?" However, the lack of structure in the case studies made this an exercise in free-text analysis.

One of the most impressive analyses comes from King's College London and Digital Science (Grant, 2015). This included visualizing statistical data such as the distribution of the evaluations of case studies, the proportion of the case studies submitted in a different unit of assessment, the most frequent words, and the size and relationships between the impact topics. The primary method they used for topic classification was the Latent Dirichlet Allocation (LDA) topic modeling algorithm. However, there are clear limitations—the contextual information in particular. At the same time, it is possible to extract individuals, institutions, and companies from the case studies; this does not in itself reveal the context in which they appear. For example, knowing that the word "Scotland" frequently appears in case studies does not in itself convey any useful meaning.

In this article, we describe an alternative approach in which the data from the case studies is transferred into a structured format according to specifically designed ontology. This captures the information in a form that gives it context, enabling a rich set of questions to be answered through querying the structured data. Our aim was for this to support a broader range of questions and produce higher-quality answers. This approach also allowed us to explore whether machine learning could be used to predict the ranking of a study.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

In the rest of this article, we first describe the types of questions we wanted to be able to ask about the impact of case studies. We then explain the ontology we designed to structure the information found in the case studies such that these questions could be answered. The success of this approach is then evaluated by creating and querying a database containing the structured representation of the case studies from the 2014 submissions for the Unit of Assessment 11 (Computer Science and Information). We then describe and evaluate an approach to using machine learning to predict the rankings of the case studies, showing that this enables the grades to be predicted with high accuracy. Finally, we cover related work and reflect on the conclusions we can draw from our research.

2. Asking Questions about Impact Case Studies

Before designing the ontology, we had to consider the questions we wanted to answer about the case studies. Our main driver was an interest in how to generate high impact from research projects. We unpacked this into a list of questions about the routes to impact, the quantifiable outcomes of impact, and the underpinning foundations that led to impact (e.g., publications and grants). This motivates us to a list of specific exemplar questions we wanted to answer. Examples are:

- How many impact case studies involved a particular company (e.g., IBM)?
- What was the total value of grants from a specific agency (e.g., EPSRC) supporting impact case studies?
- Which impact case studies were based on research funded by a specific agency (e.g., EPSRC)?
- How many impact case studies were based on research funded by a specific agency (e.g., EPSRC)?
- Which impact case studies were based on software released as open-source?
- How many impact case studies involved patents?
- Which impact case studies involved spin-off companies?
- Which companies acquired spin-off companies created to generate impact from research?
- Which impact case studies included research that influenced standards or government policy?
- How much funding by a specific agency (e.g., the EU) supported highly rated case studies?

3. Literature Review

Two major contributions were made by a project with Grant (2015). One was to make the Impact Case Studies available in the form of a searchable online database that enabled researchers to carry out further analysis. However, this supports only keyword searching for raw text in the database, and so does not allow searching for a term in a specific context. This limitation was identified in the report as an outstanding challenge. For example, the term IBM might appear in the text for very different reasons that are important when analyzing impact: an employee may have coauthored a paper, it may have used the output from the research or it may have bought a company that was spun-off to exploit the research. This limitation provided us with the motivation to design and build a new system that would overcome this problem by providing contexts for entities in the impact case studies.

The second contribution from the King's and Digital Science work was the REF2014 case study database (https://impact.ref.ac.uk/casestudies/) complete with functional APIs which provided the full text and metadata associated with each case study; we used the case studies' similarity estimation and metadata tags (e.g., impact types and research subject areas) in our work.

The Stern report (Stern, 2016) argued that Research Assessments had been influential in driving competition and fostering research excellence. The majority of the report is devoted to impact analysis, and identifies a range of problems and issues with the REF's approach. The first issue is that the requirement to link impact case studies to key research outputs might potentially result in neglecting the impact on industry, public engagement, and policy advice of those cases where the research on which the impact was based was less strong. The second was that research and teaching are usually closely

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

intertwined, and such joint-production is not currently recognized in the REF. Some findings and recommendations to address these problems are given in the report:

- Ensure that the corresponding Teaching Excellent Framework (TEF) is carefully considered during the development of the next REF to capture the impact on university teaching.
- Impact should be based on research of demonstrable quality. However, case studies could be linked to research activity and a body of work, as well as to a broad range of research outputs.
- Guidance on the REF should make it clear that impact case studies should not be narrowly interpreted, need not solely focus on socioeconomic impacts but should also include the impact on government policy, on public engagement and understanding, on cultural life, on academic impacts outside the field, and impacts on teaching.

We designed the system described in this article so that the last two recommendations could be explored.

The work of Meagher and Martin (2017) explored how impact was generated from research by investigating different types of impact and the mechanisms that generate them. In this work, they designed a framework to answer their core questions, each of which could be addressed using quantitative and qualitative methods, including content analysis, surveys, focus groups, and semi-structured interviews. Four fundamental mechanisms that generated impact were found: interdisciplinary, relationship building, knowledge exchange activities, and the recognition that impacts develop dynamically over time.

To tease out the routes to impact in different research contexts, Ross and Morrow (2016) conducted a study of REF2014 Impact Case Studies from leadership, governance, and management (LGM) research which included 2566 case studies. Their report focuses on the research processes and mechanisms that create research impact by conceptualizing topics using text mining and visualizations.

The qualitative analyses of the case studies discussed above showed that the types of impact, and the beneficiaries of impact, varied significantly across disciplines. The studies all revealed some common limitations across the UoAs: their analysis was hampered by a lack of consistent terminology and standard ways of presenting information. In this article, we have addressed this issue by designing a standard ontology to represent the information held in the case studies.

4. The Impact Case Study Ontology

The information held in the REF2014 impact case studies is mainly in the form of unstructured data—an example is shown in Figure 1 with key entities and relationships highlighted by ourselves (the highlighted text in green represents the research output or product provided by other organizations, blue represents organizations or groups of people, yellow means that the highlighted text is part of the "Route to impact"

1. Summary of the impact

New computational analysis methods have been developed to make drug discovery and toxicological analysis much more efficient. These methods have been patented (UK, EU, US) and are employed in e-Therapeutics Plc, a computational drug discovery spin-off company of the University. The company, introduced to the Alternative Investment Market of the London Stock Exchange in 2007, is now the eighth largest company (by market capitalisation - £92.7M (26/6/2013)) in the pharma/biotech sector. The underlying technologies derive from network analysis and workflow research at the University. The company has an anti-cancer drug (ETS2101) in phase I clinical trials in the UK and the US, and an anti-depression drug (ETS6103) planned to enter phase IIb clinical trial shortly. The beneficiaries of this research are e-Therapeutics directly, other drug companies, and ultimately patients.

Figure 1. *Example of an impact case study's summary of impact section.*

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

feature that is designed into our ontology; underlined text was used as a supplementary description for the corresponding entity).

Our approach requires this data to be turned into a structured form. To achieve this, we designed an ontology that captures the key classes of entities found in the studies and the relationship between them. To structure and query the data, we adopted the knowledge representation technology of the semantic web. Here knowledge is structured in the form of a set of statements, each of which is a triple consisting of a resource, a property, and the value of that property for that resource. These are defined as follows:

- *Resources:* The thing described by the statement, for example, a project, an organization, or a patent; resources are uniquely identified by URIs (Universal Resource Identifiers).
- *Properties:* A specific aspect, characteristic, attribute, or relation used to describe a resource.
- *Value:* This can be a literal value or another resource (effectively creating a graph of statements).

An example triple would be the statement that in one of the case studies, the company e-Therapeutics *(the resource)* used a research output *(the property)* namely the Microbase System *(another resource).*

The semantic web offers a way of defining ontologies that structure a domain, and ensures that all triples conform to that ontology. The ontology is specified in OWL2 (Hitzler et al., 2012). Classes provide an abstraction for grouping resources with similar characteristics. For example, Organization, Project, and Research Output are classes in our ontology. Each resource is then an instance of a class. Each class has a defined set of properties that can then be used in statements; for example, the *Organization* class has properties including subject (e.g., Cloud Computing), *locationCountry*, and *netIncome.* OWL also allows the type of values to be restricted, for example, to a resource from a specific class, or an integer.

Once the ontology has been defined, statements can be created, and a query language—SPARQL (SPARQL Protocol and RDF Query Language)—can be used to retrieve and manipulate sets of triples to answer questions about the data.

Our ontology design was based on the case studies from one of the REF Units of Assessment (UoA-11—Computer Science and Informatics). This was done through an iterative improvement process that first involved a detailed analysis of the 253 impact case studies. Next, we designed a candidate ontology to represent the information in those case studies. Finally, the ontology was populated with triples to capture the information in the case studies. We then reflected on how good the ontology was at representing all the key information contained in the text, before repeating the process in order to improve the representation. The design took into account a set of questions that it was important for the system to answer, given our primary interest in how the impact was being generated from research. These questions covered: routes to impact, the quantifiable outcomes of impact, and the underpinning foundations that led to impact (e.g., *publications* and *grants*). We also considered the need to answer questions found in previous analyses of the unstructured case study text (Grant, 2015). Finally, we also reviewed and took into account existing ways of capturing data on research impact (Researchfish, 2019). The result is the ontology shown in Figure 2. As can be seen, the key entities that are modeled includes *Collaborations*, *Funding*, *Knowledge transfer types*, *Patents*, *Grants*, and *Papers.*

One example of the design refinements we made to the ontology is that in the initial design we created the property *involved* to capture a specific organization's involvement within an impact case study. However, this did not capture key details included in the case studies, and so made it harder to answer specific questions about the nature of the involvement. We, therefore, decided to develop a more expressive way to capture specific relationships. Unlike labeled property graphs, RDF does not support attaching a property to a relation. One possible solution was to add a supplementary description to the target relation using *owl:AnnotationProperty*, which is descriptive and easy to implement. Another is to use *Class* to represent a set of relations/particular events, with each relation/particular event named using an *IRI.* In this way, supplementary descriptions could be transformed into structured relations. We decided to combine both design patterns because they are not mutually exclusive—the structured relations help to connect tagged entities, while the supplementary descriptions provide the context. For the naming scheme that is used for abstracting relations, we firstly used the term "*Involvement*" to

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

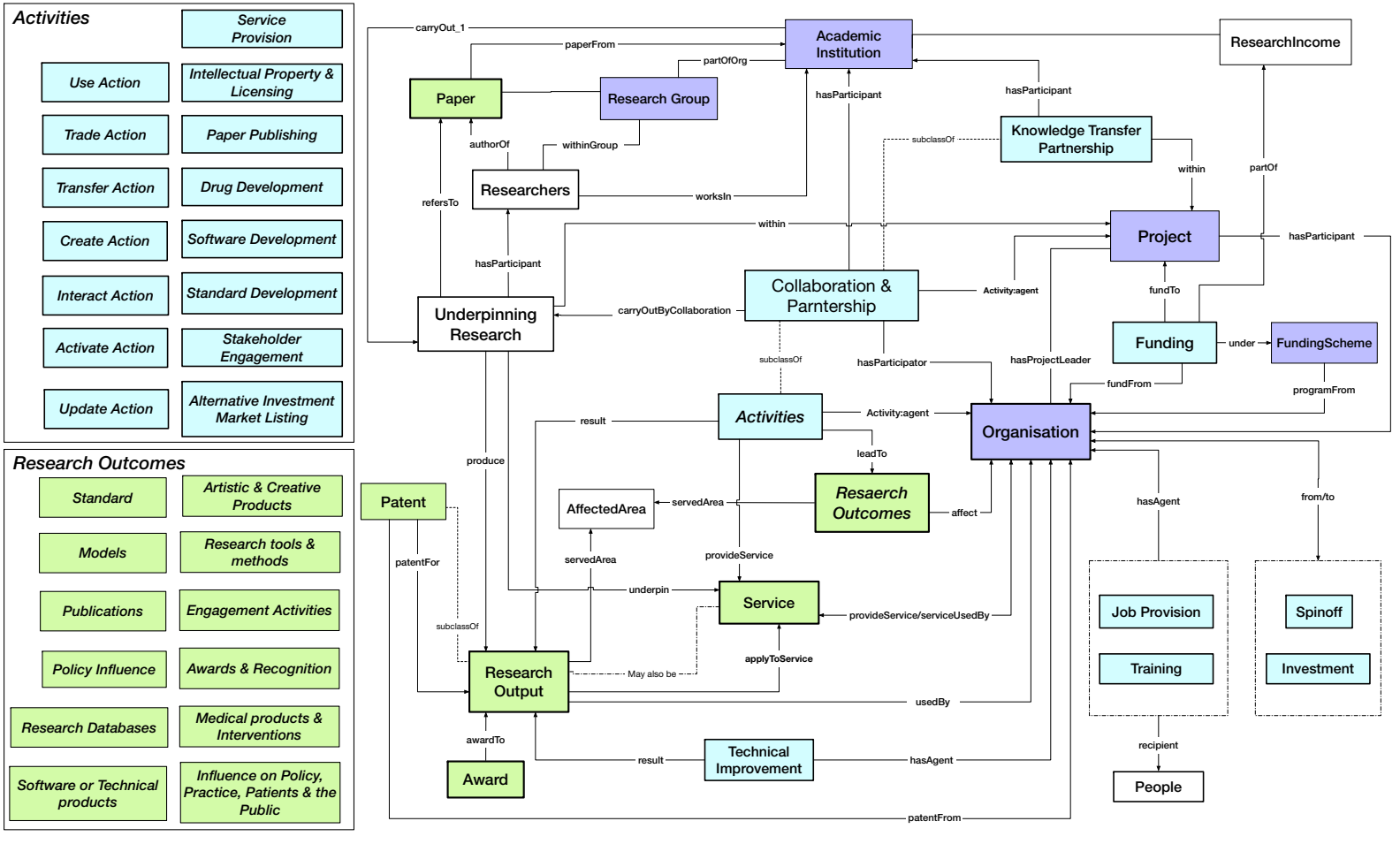


Figure 2. *The research impact ontology. Purple nodes relate to organizations, cyan to activities, and green to research outputs. The list of nodes on the left is subclasses of activities and research outputs.*

include all the activities in which organizations were involved. However, when it comes to representing different kinds of relationships, such as organizations funding research, then multiple domains and ranges would inevitably need to be added to the Class *Involvement.* Therefore, to handle these relations/events, and keep a relatively precise semantics, we replaced the "*Involvement*" class with an "*Activities*" class and further broke this down into subclasses: *Collaboration*, *Funding*, *Spin-off*, *Patent Publishing*, *Knowledge Transfer*, *Contributions to commercial and public software*, *Public engagement activities*, *Standard contributions*, *Policy influence.*

Within the *Funding* class, further information can be included, such as the grant value, currency, and time frame for funding, as illustrated in the example in Figure 3. The object property "*fund_from*" is used to connect *Funding* and *Organization*, specifying which organization is the funder, and "*fund_to*" to connect the *Funding* class and *Underpinning Research* which is funded by the organization. For organizations that have collaborated, the collaboration detail such as duration time will be saved as Datatype property in the class *Collaboration.* If any kind of research output was produced within the collaboration, then the object property "*produce*" is used to connect *Collaboration* and *Research output*, as illustrated in the example in Figure 4.

After defining the ontology, the text from the 253 case studies in UoA-11 was transformed into a semantic web representation—a set of triples conforming to the ontology.

This was mainly done manually, with most of the effort being required to create triples for the "Summary of the impact," "Underpinning research," "Details of the impact," and "Sources to corroborate

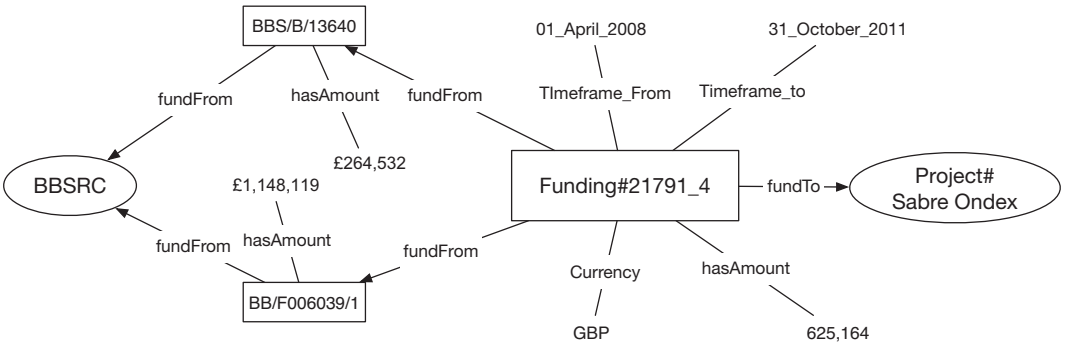


Figure 3. *Funding class (id: 21791).*

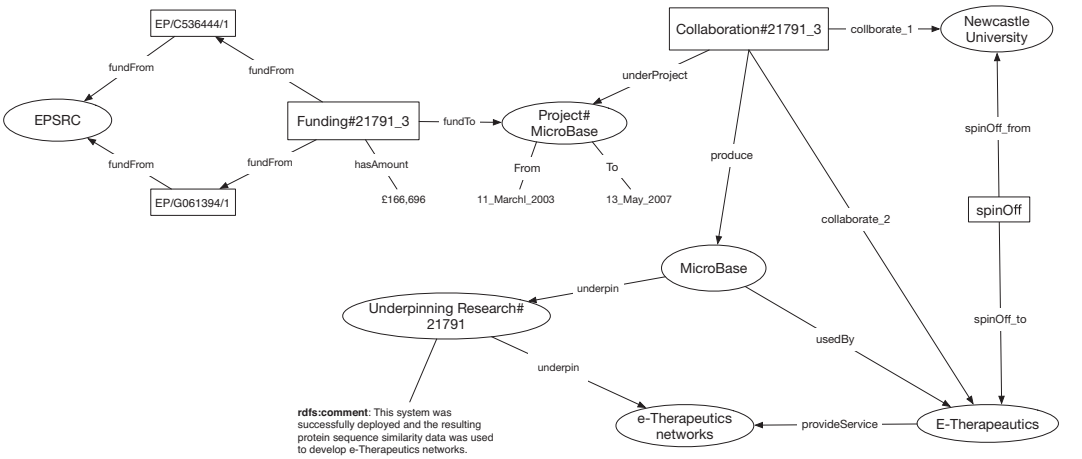


Figure 4. *Collaboration class (id: 21791).*

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

the impact" sections. Some natural language processing tools (e.g., spaCy [Honnibal et al., 2020] and Fast Entity Linker [Pappu et al., 2017]) were used to extract locations, organizations, and people, and to link them to their actual entities in DBpedia so that queries could extract contextual data from DBpedia, as well as from the case studies themselves (e.g., information about the companies mentioned in the case studies). An example of one impact case study structured in this way is shown in Figure 5. In total, this whole process generated around 51,000 triples.

A small number of case studies had sensitive information redacted. In some cases, these could easily be found online (e.g., redacted company product names); in these cases, we incorporated the information into the triple store. If an entire sentence was redacted, this information had to be omitted from the triple store, and so could not be used in question answering or for training the grade prediction model. As the proportion of redacted text was very small across the entire corpus of impact case studies, we would not expect it to affect the prediction model generated from the triple store.

5. Question Answering

Capturing all the case studies in this form allows us to answer questions using the SPARQL query language (Harris et al., 2013). For example, *"How was IBM involved in the Impact Case Studies?"* can be answered by the query shown in Supplementary Appendix 1 and produces the answer shown in Table 1.

In a further example, *"Which case studies were based on research funded by the EPSRC?"* is addressed by the query in Supplementary Appendix 2. It produces the answer shown in Table 2.

We can also exploit the structured data to aggregate information from across the set of case studies, for example, *"What was the total value of EPSRC funding that supported the impact case studies?"* (in Supplementary Appendix 3) returns the answer shown in Table 3.

We have an interest in the role of open-source software, and so could ask *"Which impact case studies were based on software released as open-source?"* (SPARQL query in Supplementary Appendix 4). This returns the answer shown in Table 4.

The role of university spin-offs is also of interest to those researching paths to impact for academic research. The answer to the question *"Which impact case studies involved spin-off companies?"* (in Supplementary Appendix 5) is shown in Table 5.

Because of the structure of the data, we can also investigate other aspects of spin-offs. The question *"Which companies acquired spin-off companies?"* (in Supplementary Appendix 6) returns the results shown in Table 6.

5.1. Integrating external data sources

One of the advantages of adopting this structured approach is that it enables the knowledge base we have created to include links to other external data sources. This enables us to support queries that seamlessly combine information from the impact case studies and these external data sources. For example, when a company appears in an impact case study, we create a link to its entry in DBpedia[1] and to its Companies House records.[2] This provides additional information, including the Standard Industrial Classification Code. An example is that we can answer the question: *"Where are the companies included in impact case studies based?"* (shown in Supplementary Appendix 7). The answer is shown in Table 7.

It would not be possible to answer this question if we could only use information held in the impact case studies. As can be seen from the country-level visualization presented in Figure 6, the research in UoA-11 has industry links to 39 countries, 38% of the companies were in the UK, followed by 24% in the United States, and 3.6% in the BRIC (Brazil, Russia, India, and China) nations.

[1] https://www.dbpedia-spotlight.org/api.
[2] https://beta.companieshouse.gov.uk.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

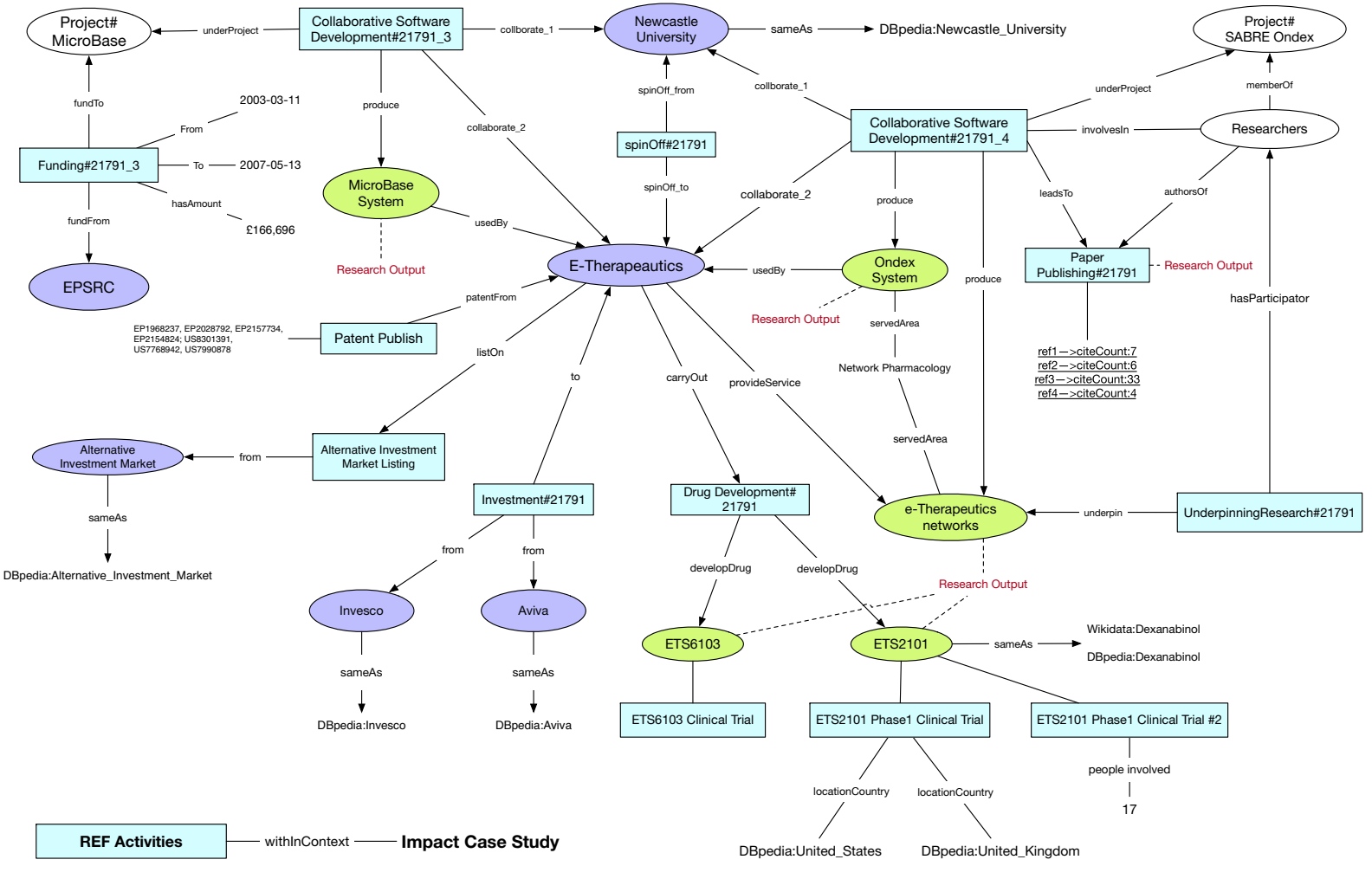


Figure 5. *An example of the graph representation of an Impact Case Study (id: 21791).*

Table 1. *How was IBM involved in the case studies?*

Company	Relation	Object
IBM	Acquired	Transitive
IBM	Collaborate	Collaboration12998
IBM	Collaborate	Collaboration1672
IBM	Collaborate	Collaboration21273
IBM	Collaborate	Collaboration5801
IBM	Uses	TrOWL
IBM	ProvideSevices	WebSphere
…	…	…

Table 2. *Impact case study research funded by EPSRC*

Case ID	Amount	Project leader	Reference	Start date	End date
id12531	£63,429	Nottingham Trent University	GR/R32468/01	01 Mar 2001	28 Feb 2002
id13783	£7,566	Brunel University London	EP/E055141/1	31 Mar 2008	30 Mar 2009
id5800	£5,820,840	University College London	EP/G059063/1	01 Oct 2009	31 Jan 2016
id42146	£6,119,249	Imperial College London	EP/H009744/1	01 Oct 2009	31 Mar 2015
id13830	£283,680	University of Cambridge	GR/S01894/01	01 Jan 2003	31 Dec 2005
id44159	£688,578	University of Southampton	GR/T10664/01	25 Feb 2005	24 Mar 2010
id35118	£887,750	Newcastle University	EP/K006568/1	05 Feb 2013	31 Dec 2016
id2010	£391,850	University of Kent	EP/E049419/1	01 Oct 2007	31 Mar 2012
…	…	…	…	…	…

Table 3. *Total EPSRC funding*

Sum	Currency
200,258,937	GBP

Table 4. *Open source applications*

Certipost, Tufuse, Xen, RASP, AutoLabDB, Festival Speech Synthesis System, ePrints, JCSP, Exposure Fusion,
RealVNC, Greenstone, Apache Cassandra, Moses, Xapian, AmbieSense, tranSMART, MilePost GCC, Camino,
local Laplacian filtering, simplenlg, KRoC, Enfuse, HermiT, Compendium, TapBack, NLTK,
AAC Speech Communicator, Apache Lucene, ChromeVox, Overture, Apache Solr,
GATE (General Architecture for Text Engineering), Ecplise MMT Project, HaRe, PERMIS, Wrangler

6. Grade Prediction

Having shown that the system was able to answer a range of questions about the impact case studies, we next wanted to explore whether the features we had extracted from the impact case studies could be used to accurately predict the grade awarded to each case study. We used machine learning for this.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 5. *Spin-off companies included in case studies*

Case ID	Spinoff from	Spinoff company
id21791	Newcastle University	e-Therapeutics
id21273	Newcastle University	Arjuna
id29895	University College London	Systemwire
id36510	University College London	Sizemic
id29897	University College London	Helicon Health
id36510	University College London	Bodymetrics
…	…	…

Table 6. *Companies that acquired spin-offs*

Spinoff company	Acquired by
XenSource	Citrix Systems
Kevin Connect	Airwave Solutions
Kestra	CyberOptics Corporation
Transitive	IBM
Essential Viewing Systems	Digital Barriers plc
KSS Retail Ltd	Dunnhumby
Imsense Ltd	Apple
Cronto Limited	VASCO Data Security Int
…	…

6.1. Assigning grades to individual case studies

Supervised machine learning algorithms require a training set of examples that include both the input features and the output (labels) that the system is trying to predict—in this case, the grade awarded to the case study. This presented a problem, as the grades for individual impact case studies were not made public. Instead, each case study was given a grade on a 9-point scale comprising integer and half-integer scores from 0 to 4. This was not made public. However, the grades awarded to all the Impact Case Studies in a Unit of Assessment were then combined, together with an impact template (described below) to give a single overall distribution that was published.

The distribution gives the percentage of case studies that were "unclassified," 1*, 2*, 3*, and 4* (a case study or impact template given half-integer score had half of its grade assigned to both of the grades that it fell between). Each Unit of Assessment also submitted one impact template that described its approach to supporting and enabling impact from research conducted within the unit. The impact template was also graded on the same scale as the case studies, and contributed 20% to the overall distribution.

For example, if there were two case studies in the submission, each would contribute 40% to the submission, and the impact template provided the remaining 20%. If one case study scored 2.5, it would contribute 20% to both 2* and 3*. If another case study was graded as 3* and the impact template was graded as 1*, this would result in the overall published distribution: 1*: 20%, 2*: 20%, 3*: 60%, and 4*: 0%.

This scoring scheme makes it impossible to recreate the actual grade for every individual impact case study, which is needed for machine learning. To overcome this, the learning phase only used case studies from institutions where the variance in the grades awarded across all the case studies was low. This approach allowed us to use the weighted score (equation (1)) of the Unit of Assessment's overall distribution as a reasonable estimate of each case study's grade in a Unit of Assessment.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 7. *Countries in which companies included in impact case studies are based, as measured by the count of unique company entities in "Unnderpinning research," "Details of the impact," and "Sources to corroborate the impact" found in Sections 2, 4, and 5 of the case studies*

Country	Companies	% of total	Country	Companies	% of total
United Kingdom	212	37.9	Israel	3	0.5
United States	135	24.1	Mexico	3	0.5
Australia	34	6.1	Egypt	2	0.4
Germany	15	2.7	Iran	2	0.4
Netherlands	14	2.5	Malaysia	2	0.4
France	13	2.3	Russia	2	0.4
Sweden	13	2.3	Slovenia	2	0.4
Japan	12	2.2	Switzerland	2	0.4
China	11	2.0	Turkey	2	0.4
South Korea	10	1.8	Argentina	1	0.2
Italy	8	1.4	Croatia	1	0.2
New Zealand	6	1.1	Denmark	1	0.2
Belgium	5	0.9	Ireland	1	0.2
Singapore	5	0.9	Luxembourg	1	0.2
South Africa	5	0.9	Norway	1	0.2
Spain	5	0.9	Portugal	1	0.2
Brazil	4	0.7	Qatar	1	0.2
Czech Republic	4	0.7	Romania	1	0.2
Canada	3	0.5	Saudi Arabia	1	0.2
Finland	3	0.5	Slovakia	1	0.2
Greece	3	0.5	Sri Lanka	1	0.2
Hungary	3	0.5	Thailand	1	0.2
India	3	0.5	United Arab Emirates	1	0.2

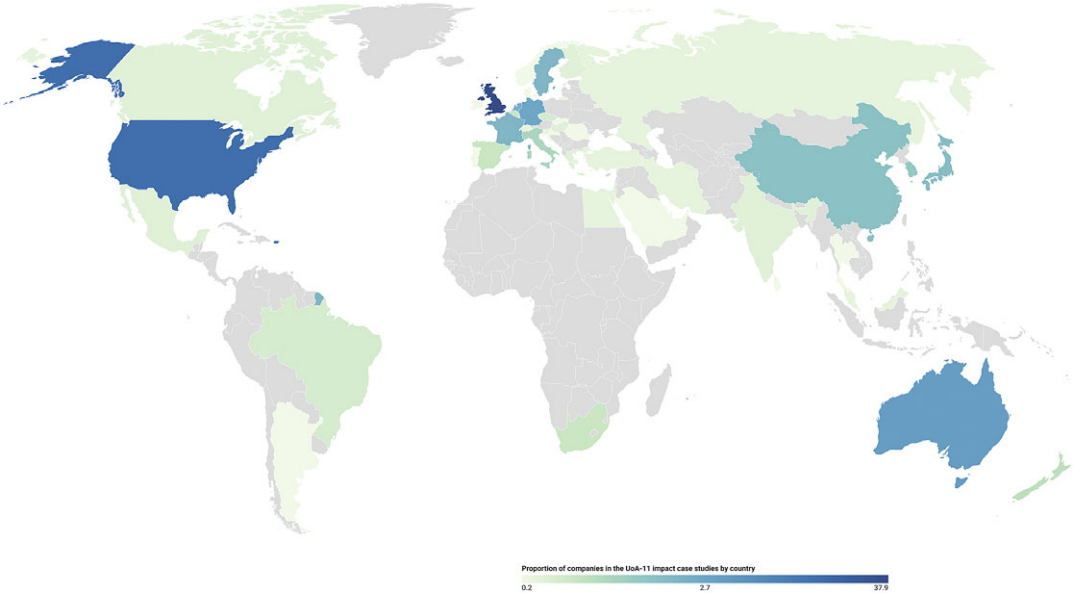


Figure 6. *The global reach of industrial impacts arising from research undertaken in UoA-11.*

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

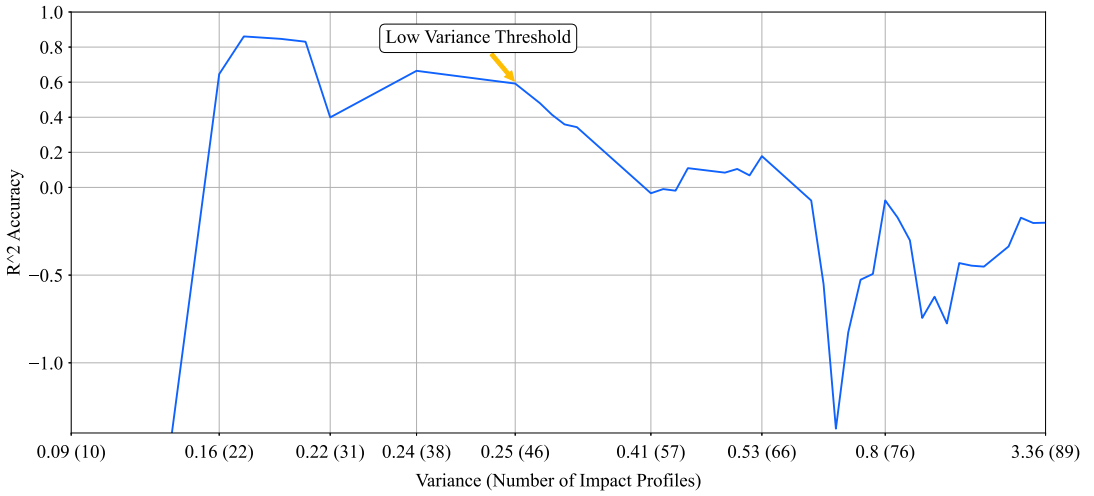


Figure 7. *The impact of using different variance thresholds to predict scores.*

$$\text{Weighted score} = 0.04.(4*) + 0.03.(3*) + 0.02.(2*) + 0.01.(1*). \quad (1)$$

6.2. Setting the variance threshold

This approach raises the question of how to select the level of variance below which a unit of assessment's case studies should be used by the machine learning algorithm. If the threshold is too low, then too few case studies will be eligible to train a robust machine learning model, whereas if it is too high, then the grades assigned to the individual case studies will not be accurate enough to train a reliable predictor.

In order to set a threshold, experiments were conducted across a range of 42 different variance thresholds. For each experiment, only the case studies from those institutions whose variance was below the threshold were used to train the machine learning model.

Figure 7 shows each predictor's cross-validated accuracy (*y*-axis) using different numbers of training sets (*x*-axis) depending on the variance threshold. The labels on the *x*-axis show that as the variance threshold rises, the impact profiles from a greater number of institutions can be included in the machine learning. The graph shows that, after an initial sharp rise, there is a downward trend in the accuracy as the size of the training set increases—the accuracy ranges from a maximum of about 85% at 24 to a minimum of −140% at 72. A rapid decline in model performance starts at 47, where the predictor can still achieve acceptable accuracy. This was therefore the threshold we selected; these 47 impact profiles contain a total of 125 case studies.

We then explored the distribution of grades of the impact case studies in the selected low variance profile. Almost all classification learning methods share an underlying assumption that the number of training samples in each different category is roughly equal. Imbalanced data may cause biases in the training process as it will be possible to achieve high accuracy by biasing predictions toward the major classes (Chawla et al., 2004).

Figure 8 shows the imbalance across different grades for the impact case studies in the low variance profile (in orange).

We used resampling methods to minimize the negative effects of imbalance. These methods modify the training dataset so that standard learning algorithms can be effectively trained on it. We, therefore, used ADASYN (He et al., 2008), an adaptive synthetic sampling method for imbalanced data, to increase the size of the minority classes in our training set, choosing the sampling strategy based on the distribution given by the REF overview report (Higher Education Funding Council for England, 2015) (the blue line in

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

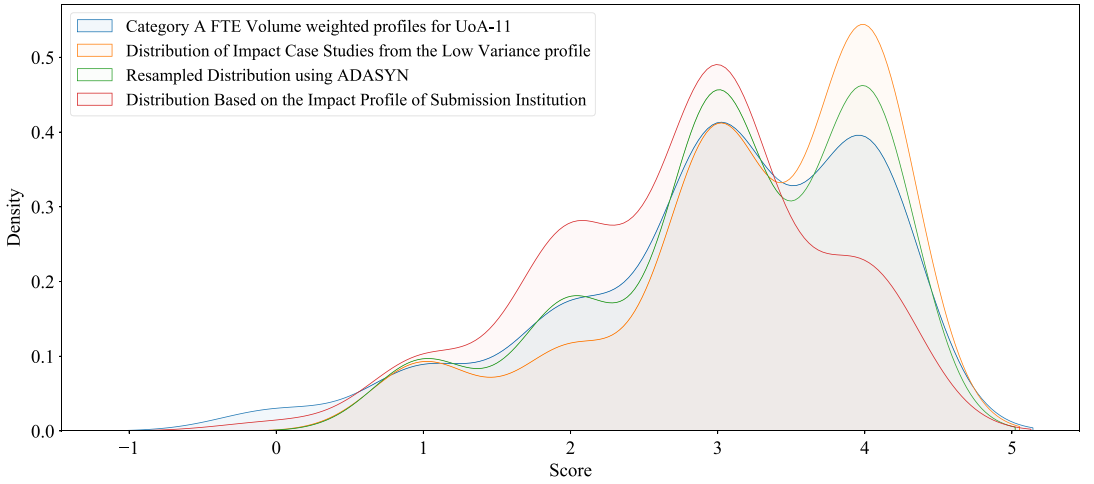


Figure 8. *Grade distribution of different parts of the impact case studies.*

Figure 8). We also applied "RepeatedStratifiedKfold" cross-validation during the oversampling, as (Santos et al., 2018) suggest. The distribution of data after resampling is shown by the green line in the same figure.

Out of a total of 253 case studies submitted in Unit of Assessment 11, our approach allowed us to use 125 case studies for machine learning. The labels produced in this way are real numbers, which are only suitable for regression-based machine learning. To open the opportunity to explore classification-based machine learning algorithms we also created two other sets of integer labels:

- 5-class labels—integers ranging from 0 to 4 (achieved by rounding the real number labels to the nearest integer)
- 9-class labels—integers ranging from 0 to 8 (achieved by doubling the real number labels before rounding them to the nearest integer)

6.3. Feature selection

To accommodate the needs of machine learning, the key parts of the data are flattened from the RDF graph into a tabular format. The conversion from the graph to flattened data used Protégé (Musen and Protégé Team, 2015) first in order to reduce human effort as much as possible. However, human input was applied to remove some very rare features that only exist in one or two specific case studies, as those features are too sparse for traditional machine learning tasks. The 42 flattened features used as inputs to the machine learning model fall into two categories: data from the whole of the unit of assessment (Table 8) and data from the impact case study itself (Table 9).

Table 8. *Unit of assessment features*

Term	Meaning
FTE	Full-time Equivalent staff in post on 31 October 2013
Research Income	Total research income during the REF period
Number of cases†	Number of impact case studies submitted
Number of patents†	Number of granted patents
Outputs†	Number of outputs
CitedByCount†	Total citations of the outputs
Degrees awarded†	Number of Doctoral Degrees awarded in the REF period

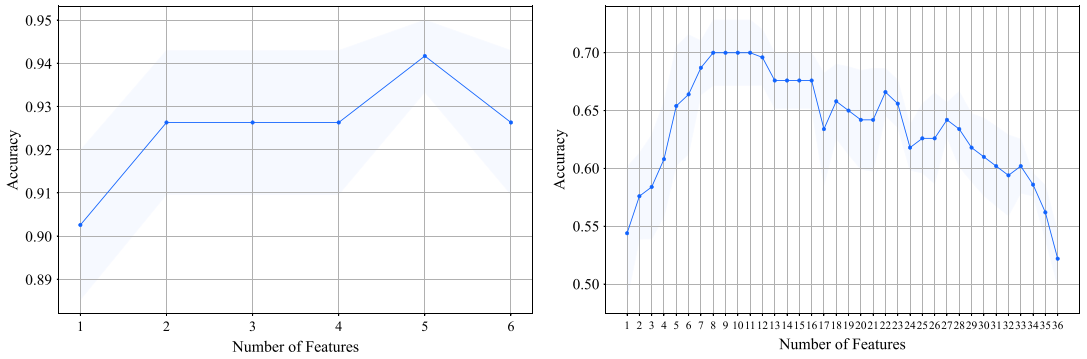
https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 9. *Case study features*

Term	Meaning
Total Funding	Total of all funding (converted to GBP)
Currencies	Number of different currencies referenced
Economic Benefit	Monetary amount generated from the impact (e.g., through investment, cost-savings, or revenue rise)
SpinOffs	Number of spin-offs created
Acquisition	Number of spin-offs that were acquired
Merged	Number of spin-offs that were merged
KTP	Knowledge Transfer Partnerships included in the case study
Collaborations	Number of external collaborations
Industrial use	Number of industrial uses cases in the case study
Academic use	Number of use cases in the academic sector
Government use	Number of use cases in government
Public use	Boolean indicating if research outputs were used by the public
Open source	Number of open-source applications released
Policies	Number of policies created or influenced
Standards	Number of standards created or influenced
Patents	Number of patents (applications and published)
Awards	Number of awards won
Clinical trials	Number of clinical trials conducted
Citations	Total citations (in 2014) of all papers included in "References to the research"
Duration	Duration of the research programme
Funders	Number of funding bodies
Economic impact	Boolean indicating if the case study has a direct or potential economic impact
Impact type	Main impact theme: cultural, economic, environmental, health, legal, political, societal or technological
Public well-being	Boolean indicating impact on public well-being
Public service delivery	Boolean indicating impact on public service delivery
Improved accessibility	Boolean indicating impact on improving the accessibility of public services
Improved workforce	Boolean indicating impact on improving the workforce, for example, provide training
Changing public attitude	Boolean indicating impact on changing public attitudes, for example, enhancing public awareness
Solutions to societal probs	Boolean indicating if the case study includes impact on societal problems, for example, reducing crime
Improve env. sustainability	Boolean indicating if the case study includes impact on environmental sustainability, for example, reducing energy usage
Public engagement	Boolean indicating if the case study includes the provision of public engagement activities

Having a large number of features increases the computational cost, and the risk of overfitting, while the number of observations required to produce a convincing prediction grows exponentially with the number of features. To prevent this, feature selection and extraction techniques are used to avoid this curse of dimensionality (McLachlan, 2004). We employed only feature selection in this experiment since it preserves the original data characteristics when a subset of the features is selected. While feature extraction techniques such as Principal Component Analysis (PCA) are highly effective at reducing dimensionality, interpretability is necessarily reduced because the retrieved features are transferred to a

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press



(a) Feature Selection on UoA features

(b) Feature Selection on Case study features

Figure 9. *Feature selection using the deterministic wrappers method.*

new space. Interpretability was critical for this study since we wanted to determine which characteristics are critical for grade prediction.

Calculating all possible feature combinations prior to deleting unnecessary and redundant features is computationally impractical. As a result, we used the Deterministic Wrappers approach in this experiment, which follows the sequential forward selection (SFS) principle by greedily evaluating all alternatives, to match the 5-classes label based on an XGBoost classifier. The experiments were conducted individually for two categories, with the results displayed in Figure 9. Figure 9a illustrates the high accuracy achieved when five UoA features are employed; Figure 9b illustrates the initial increase in performance to 11 Case Study features before the accuracy begins to fall due to the number of dimensions.

To replicate the environment in which the REF panel made their decisions, we excluded those features ("number of degrees awarded," "number of outputs," and "GPA of outputs") that were not available to the panel when they chose the grades.

Apart from the features described above, the symbol † is used in the glossary tables to denote other features not used in machine learning. The feature selection procedure indicated that these features only slightly improved the model's performance, while increasing its dimensionality. The reason why some of these features do not contribute to the overall performance as much as might have been expected is that either some of them are highly correlated with each other, or they are Boolean features that have low variance distributions.

A clustered heat map (Figure 10) produced by the Seaborn package (Waskom, 2021) was used to visualize correlation analysis. As the features comprise both continuous and discrete variables, we used Spearman's rank correlation coefficient to correlate the remaining features after feature selection and the estimated scores. Each intersecting cell shows the correlations between the variables on each axis. The strength of the color within its color gradient is proportional to the correlation coefficient. The dendrogram reveals the similarities between the variables. For example, the features "GovUse" and "Policy Influence," the "SpinOff" and "Acquisition" were close in the case study context.

Rich information can be drawn from both positive and negative correlations between two variables in the heatmap. The main message is that "Research Income" is the variable most correlated with the score. The rest of the variables mostly follow the pattern: the higher the correlation with "Research Income," the higher the correlation with the score.

6.4. Model performance metrics

The ability of the regressor and classifier to generalize from the test set and make accurate predictions of the grade are evaluated using the following commonly used evaluation criteria (Bishop, 2006) for supervised machine learning regression and classification models:

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

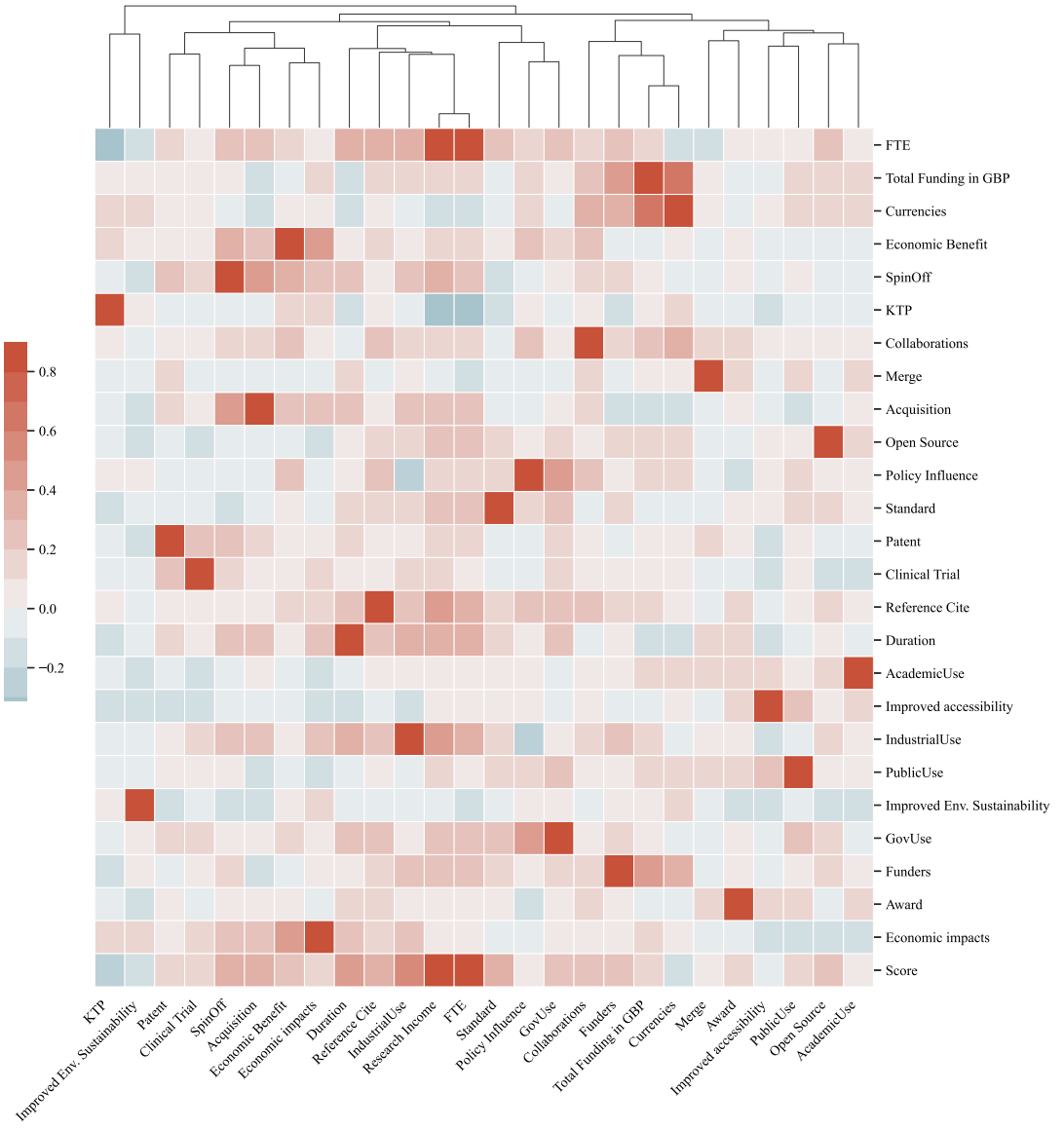


Figure 10. *Heatmap visualization for selected features.*

- R-squared (R^2): The coefficient of determination measures a regression model with two sums of squares: one is the total sum of squares (denoted as $\mathrm{SS}_{\mathrm{TOT}}$) and the residual sum of squares (denoted as $\mathrm{SS}_{\mathrm{RES}}$). The definition is:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{RES}}}{\mathrm{SS}_{\mathrm{TOT}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}. \tag{2}$$

- Mean absolute error (MAE): used to show the average absolute error between the true value x_i and predicted value y_i with the same scale.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 10. *The confusion matrix of classification*

	Prediction value	
True value	Positive class	Negative class
Positive class	TP (True Positive)	FN (False Negative)
Negative class	FP (False Positive)	TN (True Negative)

The confusion matrix shown in Table 10 helps in understanding the classification metrics below.

- Precision is the probability of the sample actually being positive out of all the samples that were predicted to be positive.
- Recall is the probability of a positive sample being predicted in a sample that is actually positive.
- The F1-Score: In general, precision and recall are inversely proportional, we can either use the area under the Precision-Recall curve or F1-Score, which is the harmonic mean of Precision and Recall, to evaluate the classifiers from both perspectives. The formula is:

$$\mathrm{F1} = \frac{2}{\mathrm{Recall}^{-1} + \mathrm{Precision}^{-1}} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}. \tag{3}$$

- Hamming Loss: a loss function that indicates the fraction of the wrong class to the total number of classes, the optimal value is zero, a smaller value means better performance of the classifier (Tsoumakas and Katakis, 2007).
- Area under ROC (AUC): defined as the area under the receiver operating curve (ROC) (Hand and Till, 2001). A ROC curve plots two parameters: sensitivity (a synonym for recall) and specificity.

The metrics above could be used directly in binary classification problems, whereas in multiclass classification problems, precision and recall should be averaged in the manner of Macro, Weight, and so forth (Sokolova and Lapalme, 2009). Macro averages the metrics (Precision/Recall/F1-score) of the different categories, giving the same weight to all categories. It allows each category to be treated equally. The weight method gives different weights to different classes (the weights are determined according to the proportion of class distribution), and each class is multiplied by the weights and then summed. This method considers the imbalance of the classes and its values are more likely to be influenced by the majority class, which are grade 3, 3.5, and 4 in our case.

6.5. Predicting grades of impact profiles

We mainly used XGBoost (Chen and Guestrin, 2016) and the built-in algorithms in scikit-learn (Pedregosa et al., 2011) for machine learning predictions.

XGBoost is a decision-tree-based ensemble machine learning algorithm that uses the gradient boosting framework. We chose it because it gives excellent results with small data sets, which we have in this situation. Additionally, we compared its performance to that of a Support Vector Machine (SVM) model.

The first experiment was conducted using the average impact grades of each institution in UoA-11 as the target label, with each university's research income, and full-time equivalent (FTE) staff used as input features for training the model. This is to explore whether it is possible to use UoA features to make grade predictions over the impact profiles, which have a much smaller size compared to the number of impact case studies.

Comparison is made between two different subsets of the impact profiles. One is the full profile of all 89 institutions (the red legend in Figure 8), while the other is the 46 profiles that fall within the low variance threshold (the orange legend in Figure 8).

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 11. *Results for the full and low variance profiles using only UoA features*

Profile	MAE	R^2
Full (89 institutions)	0.57	0.35
Low variance (46 institutions)	0.43	0.47

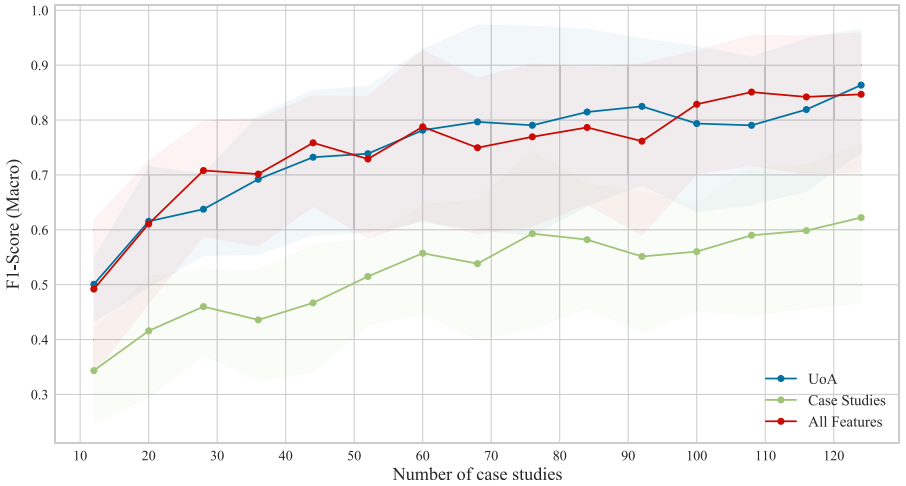


Figure 11. *The learning curve of three different feature sets using the case studies within the low variance threshold.*

The experiments used 75% of the observations for training, and the remainder for testing. The results are shown in Table 11. In the XGBoost based regression model, the larger dataset only achieves 35% accuracy, while the smaller one achieved a higher accuracy (47%). This experiment shows the limits of using only the limited UoA features for impact profile grade prediction. It also shows the advantage of training on the low variance dataset, and so that is used for the rest of our experiments. To validate the performance of the case studies from the selected low variance impact profiles, we used the learning curves of three classifiers in Figure 11 to show the increase in macro-averaged F1-score performance of three feature sets—case study features, UoA features, and both of them when using different proportions of training data.

6.6. Predicting the grade of a case study within the low variance profile

In the second experiment, two algorithms are applied to predict the scores on three different subsets of features using regression and classification methods. For the classification problem, the "5-classes" and "9-classes" labels defined above are used as the target for prediction.

For the classification experiment, we used 60% of the whole dataset, which is 75 out of 125 observations, for training and validation; the remaining 50 are for testing. We chose 60% over the commonly used 75% for the training set because there are only six samples for minor classes from 1 to 2. Increasing the test set to 40% allows the minor classes sample to increase to 18, making the minor classes evaluation less likely to be affected by having only a few samples. The learning curve in Figure 11 also shows the reasonable performance of three different feature sets when using 75 case studies.

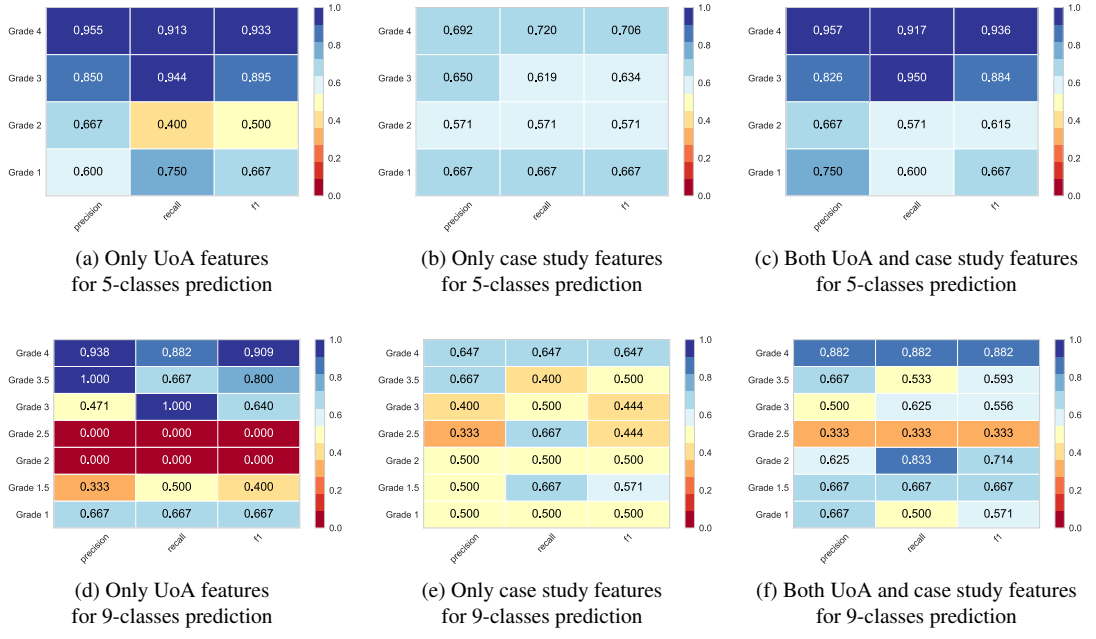
https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

Table 12. *XGBoost and SVM's classification performance (Precision Recall and F1-Score) with 5 and 9 classes. The bold numbers indicate the best result for each feature set in weighted and macro average.*

Metrics / Features	Average	XGBoost						SVM					
		Classification-5			Classification-9			Classification-5			Classification-9		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
UoA	Weighted	0.86	0.86	0.85	**0.74**	**0.70**	**0.69**	**0.77**	**0.82**	**0.79**	**0.67**	**0.56**	0.49
	Macro	0.77	0.75	0.75	0.48	0.52	0.48	0.55	0.65	0.59	0.41	0.51	0.37
Case study	Weighted	0.66	0.66	0.66	0.57	0.54	0.54	0.57	0.57	0.54	0.38	0.42	0.38
	Macro	0.65	0.64	0.64	0.51	0.55	0.52	0.60	0.54	0.54	0.41	0.37	0.37
All features	Weighted.	**0.86**	**0.86**	**0.85**	0.69	0.68	0.68	0.74	0.71	0.71	0.60	0.55	**0.56**
	Macro	**0.80**	**0.76**	**0.78**	**0.62**	**0.62**	**0.62**	**0.63**	**0.61**	**0.59**	**0.53**	**0.59**	**0.54**

Table 13. *XGBoost and SVM regression performance. The bold numbers indicate the best result for each feature set.*

Features \ Metrics	XGBoost		SVM	
	R^2	MAE	R^2	MAE
UoA	0.85	**0.11**	0.71	**0.29**
Case study	0.46	0.45	0.37	0.46
All features	**0.88**	0.18	**0.73**	0.30



(a) Only UoA features for 5-classes prediction

(b) Only case study features for 5-classes prediction

(c) Both UoA and case study features for 5-classes prediction

(d) Only UoA features for 9-classes prediction

(e) Only case study features for 9-classes prediction

(f) Both UoA and case study features for 9-classes prediction

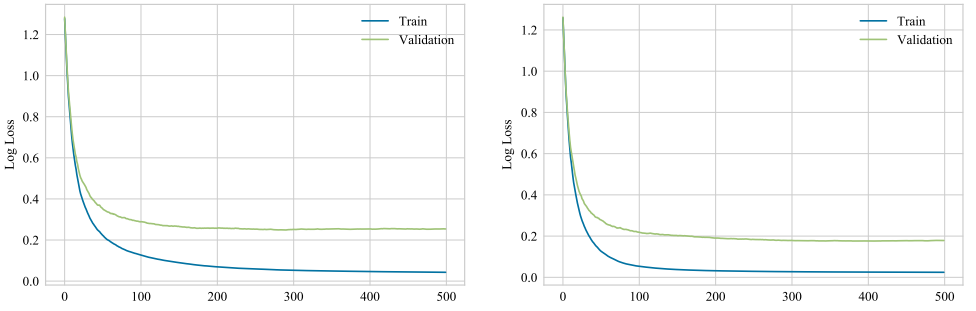
Figure 12. *Classification report using XGBoost with three different features.*

Tables 12 and 13 show results from the nine experiments using XGBoost and SVM for regression and classification. We noted that the figures only show four classes (Figure 12); this is due to the lack of case studies with label 0 using our estimation method. Similarly, in the 9-classes classification, the figure only shows seven classes because of the lack of data with labels 0 and 0.5.

Table 14. *XGBoost AUC and Hamming loss with 5 and 9 classes. The bold numbers indicate the best result for each feature set.*

Features \ Metrics	Classification-5		Classification-9	
	AUC	Hamming loss	AUC	Hamming loss
UoA	**0.98**	**0.18**	0.85	**0.27**
Case study	0.81	0.33	0.78	0.50
All features	0.97	0.19	**0.95**	0.28

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press



(a) Validation curve of classifier based on UoA features (b) Validation curve of classifier based on all features

Figure 13. *Validation curves of classifiers using UoA and all features in 9-classes classification.*

6.7. Model validation

We used classification reports to observe the model performance for each category through a confusion matrix, which comprises True Positives, False Positives, True Negatives, and False Negatives. Each classification report (Figure 12) shows the model's performance using the metrics (Precision and Recall and F1-Score) corresponding to each category. Two other metrics (AUC (Area Under Curve) and Hamming loss) for XGBoost classifiers are provided in Table 14.

From those results listed above, we can see the XGBoost model outperforms the SVM in our dataset, so we will mainly focus on interpreting the results from XGBoost models. The UoA features achieve satisfactory performance in Grade 3–4, but much lower performance in Grade 1.5–2.5. The case study features achieve relatively lower performance compared to the UoA features in Grade 3–4, but have a more stable results across all seven grades. This may explain why the UoA features outperform case study features in the 5-classes classification but have similar performance with case study features in the 9-classes classification. When we use the combination of the two features, the performance in Grade 3–4 remains as high as the UoA features and outperforms UoA and case study features separately. Different metrics using AUC and Hamming loss in Table 14 also tell the same story. We also used the validation curves to show the generalization behavior of the classifiers using UoA and all features in the 9-classes classification (Figure 13). Regardless of the two small gaps between the two validation curves, which indicate the need for a larger data set, neither the UoA classifier nor the all features classifier has under-fitted or over-fitted.

6.8. Result interpretation

Several methods are provided in the XGBoost package to interpret the model that has been created. Figure 14 provides a visualization of part of one of the trees in the model that split the data using the conditions. However, it is not practical to visualize the full trees given their size.

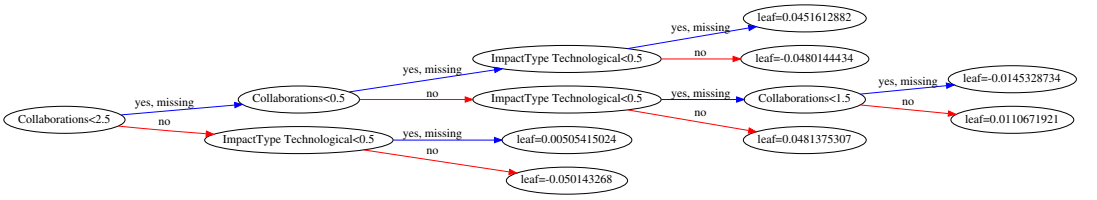
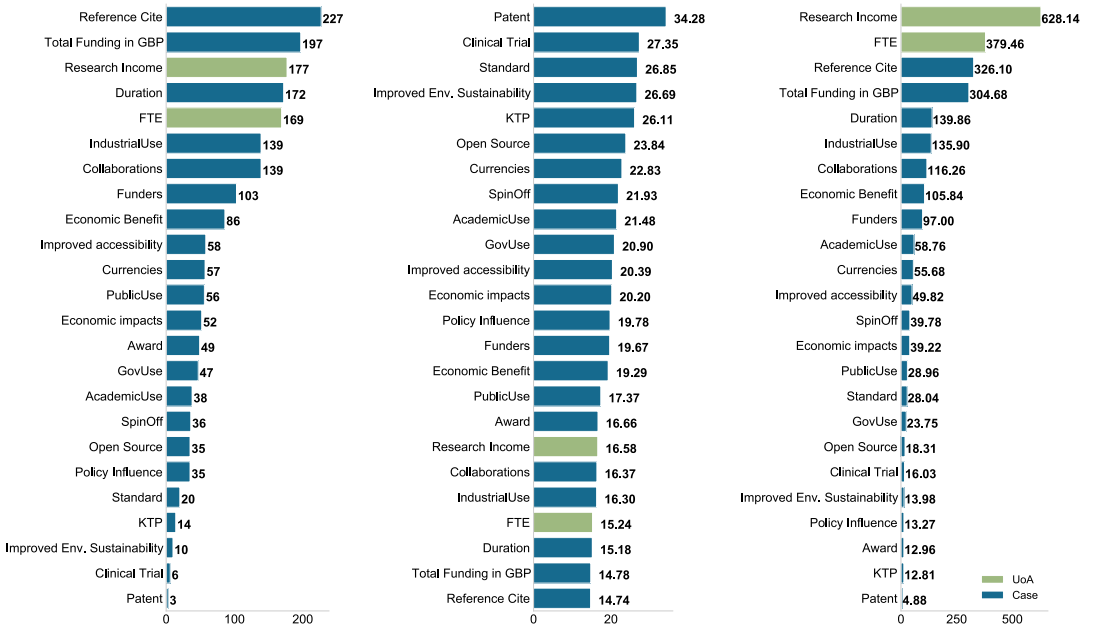


Figure 14. *Tree visualization of the decision tree for grading case studies.*

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press



(a) Weight (b) Coverage (c) Total Gain

Figure 15. *Feature importance (Weight, coverage, and total gain).*

Therefore, a different method was used to measure the feature importance for the model shown in Figure 12c, which uses both UoA and case study features as input and 5-classes as output. Figure 15 shows feature-importance ranked by weight, coverage, and total gain:

- *Weight* is the frequency of a specific feature used for splitting all trees across all weak learners. This shows that the "Reference Cite" feature has been used the most in splitting the trees.
- *Coverage* is the number of observations that are split, weighted by the frequencies of the same feature used in splitting across all the trees. This indicates that the feature "Patent" has split the most observations.
- *Total gain* is the total gain in accuracy brought by the specific feature when splitting the trees. This shows that the feature "Research Income" contributed the most to the model prediction accuracy.

It should be noted that binary features like "Improved Environmental Sustainability" can only be used once in each tree. As a result, binary features are scored much lower in metrics where *Weight (frequency)* is used.

For a more in-depth analysis of the results, we use SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to provide a more consistent and accurate way of measuring features. SHAP is an

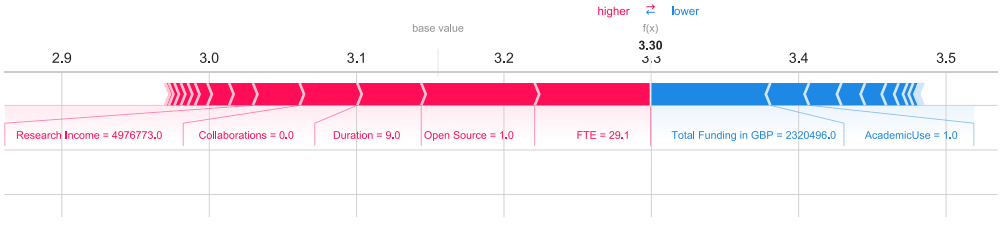


Figure 16. *SHAP force plot.*

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

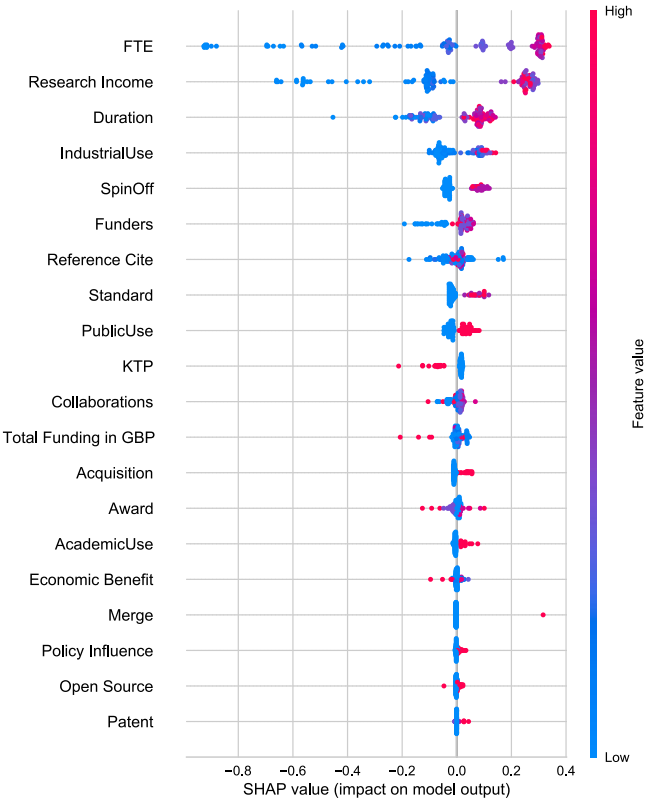


Figure 17. *SHAP summary plot.*

innovative approach to improving the predictions from a complex model and exploring relationships between the features and the individual observations.

Figure 16 shows a force plot produced by SHAP values. The blue side indicates a negative contribution to the prediction, while the red side indicates a positive contribution. As shown on the right-hand side, the features "Research Income," "Collaboration," "Duration," "Open Source," and "FTE" have a positive contribution, while the "Total Funding in GBP" and "Academic Use" have a negative contribution. This may be due to the fact that a significant proportion of the impact case studies do not mention financial information or only mention very vague figures. Of the 125 observations, 24 out of 31 case studies had blank funding data and scored between 3 and 4. Interestingly, the remaining seven cases only scored between 1 and 1.5. This lack of information may have led to features such as "Total Funding in GBP" being ranked low in terms of feature importance.

To get an overview of which features are most important for a model we can plot the SHAP values of every feature for every sample. The summary plot in Figure 17 sorts features by the sum of SHAP value importance over the case studies and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the score value (red high, blue low). This reveals that high "Research Income," "FTE," "Duration," "IndustrialUse," "SpinOff," "Standard," "PublicUse," "Acquistion," and "AcademicUse" increases the predicted score; higher "KTP" lowers the predicted score.

7. Conclusions

The article describes how creating a structured knowledge base makes it possible for policymakers to answer a broader range of questions about evidence than is possible with free text data. This is because the structured

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

representation gives context to the terms included in the text. We have also shown that it is possible to link to external data sources, so supporting questions that rely on this additional information, for example on companies, grants, and publications. The focus of our work has been on supporting those seeking to extract value from the REF2014 Impact Case Studies, and a significant part of our work was designing the ontology to structure this data. However, once this was done, and the information transferred into a structured form, the results show that it is possible to provide quantitative answers to a set of questions—this goes well beyond what was possible with previous research that had to work with free text data.

We then used features extracted from the knowledge base to explore the use of machine learning to predict the grades that were assigned to the case studies. The results show that this can be achieved with good accuracy. Further, they show that it is the features of the units of assessment (Research Income and FTEs—the number of Full-Time Equivalent staff) that are of great importance in grade prediction, especially for an impact case study with a grade in the range from three to four. Perhaps surprisingly, the specific features of individual case studies proved less important for grade prediction, though they still contribute to the overall accuracy across all grades. This work has focused on the REF2014 Impact Case Studies in Unit of Assessment 11, and it would be good to see, in future studies, the same methods applied to other Units of Assessment so as to explore the impact of variability in the way different assessment panels implemented the guidance under which they were operating.

The overall approach we have taken is applicable to other areas where policymakers wish to analyze both the evidence gathered in an evaluation exercise, and the most important factors behind the awarding of grades based on that evidence. In order to facilitate this, we would recommend that those designing evidence-gathering exercises—whether or not for grading—capture information in a structured form where possible, rather than just in free text. In the example that is the focus of this article, this would have considerably reduced the time and effort needed to transform the free text data into RDF triples. One novel outcome of our study was the method we devised that used low-variance aggregated data from a Unit of Assessment to create a dataset that could be used to train the machine learning models, but this would have been significantly simplified if the individual grade awarded to each case study had been published. Another benefit of this work is that it can guide those selecting case studies for submission to the REF. For example, those based on Knowledge Transfer Partnerships (KTPs) performed less well than others in the 2014 UK exercise. Studying those features that contributed most to high scores can also be used to focus efforts on which evidence to try to collect—this consumes a significant amount of time for the teams set up in UK universities to prepare Impact Case Studies.

Acknowledgment. This work was presented at the data for policy conference in September 2020: https://zenodo.org/record/3967667#.YMyZAi1Q0eY.

Funding Statement. This work was not supported directly by any grant.

Competing Interests. The authors declare no competing interests exist.

Author Contributions. Conceptualization: P.W., B.H., J.Z; Resources and Funding acquisition: B.H.; Methodology, Formal analysis, Investigation, Data curation, Visualization, Validation, Original draft, and Software: J.Z.; Project administration and supervision: P.W.; Reviewing and editing of the draft: P.W., J.Z.

Data Availability Statement. The impact case study text data used in the study are available at: https://impact.ref.ac.uk/casestudies/APIhelp.aspx. The ontology and data are available at: https://github.com/articoder/REF-UoA11-Ontology.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Reproducibility Statement. Methods and settings to reproduce the paper's main results in grade prediction: imbalanced-learn (Lemaître et al., 2017) implementation of ADASYN, XGBoost, Scikitlearn implementations of SVM and grid searches with cross-validation of the following parameter spaces:

- Use of ADASYN in 5-classes classification: Sampling strategy: {0: 12, 1: 18, 2: 60, 3: 55}, Neighbor = 4.
- XGBoost regressors: Maximum depth: 1, 2, 3, 4; Minimum child weight:1, 2; Subsample: 0.6, 0.7, 0.8, 0.9; Colsample_bytree: 0.6, 0.7, 0.8, 0.9; Number of estimators: 100, 200, 300, 400, 500; Learning rate: 0.05, 0.06, 0.07, 0.08, 0.09.
- XGBoost classifiers: Maximum depth: 2, 3, 4; Minimum child weight:1, 2; Subsample: 0.6, 0.7, 0.8, 0.9; Colsample_bytree: 0.6, 0.7, 0.8, 0.9; Number of estimators: 100, 200, 300, 400, 500; Learning rate: 0.05, 0.06, 0.07, 0.08, 0.09.

https://doi.org/10.1017/dap.2022.21 Published online by Cambridge University Press

- SVM regressors: Kernel: linear, rbf, sigmoid; coef0: 0, 0.01, 0.03, 0.05; Degree: 1, 2, 3, 4, 5; Gamma: 1, 2, 3; C: 0.1, 1, 3, 5, 7, 10.
- SVM classifiers: Kernel: linear, rbf, sigmoid; coef0: 0, 0.01, 0.03, 0.05; Degree: 1, 2, 3, 4, 5; Gamma: 1, 2, 3; C: 0.1, 1, 3, 5, 7, 10.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/dap.2022.21.

References

Bishop CM (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer-Verlag.

Chawla NV, **Japkowicz N and Kotcz A** (2004) Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter 6*(1), 1–6.

Chen T and Guestrin C (2016) XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery.

Grant J (2015) *The Nature, Scale and Beneficiaries of Research Impact: An Initial Analysis of Research Excellence Framework (Ref) 2014 Impact Case Studies*, Technical Report. London: King's College London and Digital Science.

Hand DJ and Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning 45*(2), 171–186.

Harris S, **Seaborne A and Prud'hommeaux E** (2013) SPARQL 1.1 query language. *W3C Recommendation 21*(10), 778.

He H, **Bai Y**, **Garcia EAand Li S** (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328.

Higher Education Funding Council for England (2015) *Research Excellence Framework 2014: Overview Report by Main Panel b and Sub-Panels 7 to 15*, Technical Report. Higher Education Funding Council for England. 47–64. https://ref.ac.uk/2014/media/ref/content/expanel/member/Main_Panel_B_overview_report.pdf

Hitzler P, **Krötzsch M**, **Parsia B**, **Patel-Schneider PF and Rudolph S** (2012) Owl 2 web ontology language primer (Second Edition). *W3C Recommendation*. https://www.w3.org/TR/owl-primer/

Honnibal M, **Montani I**, **Van Landeghem L and Boyd A** (2020) *spaCy: Industrial-Strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303 (2020).

Lemaître G, **Nogueira F and Aridas CK** (2017) Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*(17), 1–5.

Lundberg SM and Lee S-I (2017) A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

McLachlan GJ (2004) *Discriminant Analysis and Statistical Pattern Recognition*, vol. 544. John Wiley & Sons.

Meagher LR and Martin U (2017) Slightly dirty maths: The richly textured mechanisms of impact. *Research Evaluation 26*(1), 15–27.

Musen MA and **Protégé Team** (2015) The protégé project: A look back and a look forward. *AI Matters 1*(4), 4–12.

Pappu A, **Blanco R**, **Mehdad Y**, **Stent A and Thadani K** (2017) Lightweight multilingual entity extraction and linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. New York, NY: Association for Computing Machinery.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Researchfish (2019) *Research Outcomes Common Question Set*, Technical Report. Researchfish.

Ross F and Morrow E (2016) *Mining the ref impact case studies for lessons on leadership, governance and management in higher education.* Impact of Social Sciences Blog.

Santos MS, **Soares JP**, **Abreu PH**, **Araujo H and Santos J** (2018) Cross-validation for imbalanced datasets: Avoiding over-optimistic and overfitting approaches [Research Frontier.]. *IEEE Computational Intelligence Magazine 13*(4), 59–76.

Sokolova M and Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Information Processing and Management 45*(4), 427–437.

Stern N (2016) *Research Excellence Framework Review: Building on Success and Learning from Experience*, Technical Report. London: Department for Business, Energy & Industrial Strategy.

Tsoumakas G and Katakis I (2007) Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM) 3*, 1–13.

Waskom ML (2021) Seaborn: Statistical data visualization. *Journal of Open Source Software 6*(60), 3021.

Cite this article: Zhang J, Watson P and Hodgson B (2022). A new approach to impact case study analytics. *Data & Policy*, 4: e30. doi:10.1017/dap.2022.21