Scientometrics (2023) 128:4933–4969
https://doi.org/10.1007/s11192-023-04736-z

Prediction of patent grant and interpreting the key determinants: an application of interpretable machine learning approach

Li Yao[1] · **He Ni**[2]

Received: 26 October 2021 / Accepted: 5 May 2023 / Published online: 22 June 2023

Abstract
Patents are valuable intellectual property only when granted by the governments, and failing to receive an official grant means disclosing valuable technologies and information, which otherwise would be kept as commercial secrets. Yet, a typical patent application process takes years to complete and the outcome is uncertain. This study implements machine learning models to predict patent examination outcomes based on early information disclosed at patent publication and interpret the mechanism of how these models make predictions, highlighting the key determinants to patent grant and delineating the relationships between the patent features and the examination outcome. The predictive models that integrate patent-level variables with textual information accomplish the best prediction performances with a 0.854 ROC-AUC score and 77% accuracy rate. A number of interpretable machine learning methods are applied. The permutation-based feature importance metric identifies key determinants such as applicants' prior experience, page length, backward citation, claim counts, number of patent family, etc. SHAP (SHapley Additive exPlanations), a local interpretability method, describes the marginal contributions to the model prediction of key predictors using two actual patent examples. Our study provides several valuable findings with important theoretical insights and practical applications. Specifically, we show that patent-level information can serve as a predictor of examination outcomes and the relationships between the predictors and outcome variables are complex. Knowledge accumulation and technology complexity positively affect the likelihood of patent grants, albeit with a curvilinear relationship. At lower levels, both factors significantly increase the chance of a grant, but beyond a certain threshold, the marginal effect becomes less pronounced. Additionally, prior experience, patent family size, and engagement with the patent agency have a monotonic and positive relationship with the grant likelihood, whereas the impact of patent scope on patent grants remains uncertain. While a narrower and more specific patent claim is associated with a higher grant rate, the number of claims increases it. Moreover, technology range, inventor team size, and examination duration have little effect on the patent grant results. From a practical standpoint, the accurate prediction of patent grants has significant potential applications. For instance, it could help firms better prioritize resources on the patent applications of high grant potentials to secure the final grant, as failure means a waste of R &D effort and disclosure of technology

Extended author information available on the last page of the article

without IP protection. Additionally, patent examiners could utilize our predictive results as prior knowledge to enhance their judgment and accelerate the examination process.

Keywords Patent grant · Interpretable machine learning · Predictive models

Introduction

A patent legally protects the relevant technology from infringement but the patent application process lasts years and the outcome is uncertain (Tong et al., 2018). When a government grants a patent, on one hand, it creates a short-run social welfare loss due to technology monopoly, on the other, it provides extra incentives for innovators to engage in R & D and to publicize their inventions (Hall & Harhoff, 2012). For innovators, the trade-off lies in the fact that protecting their intellectual property through the patent system requires disclosing the technology which otherwise might be kept as commercial secrets. Thereby, the drawbacks of the patent system for many innovators include not only high monetary and time costs but also the uncertainty of successful grants. Failure to receive a legal patent grant means disclosure of valuable technology without appropriating the intellectual property, as only at the time of grant, injunctive relief and other legal rights are bestowed upon the patentees (Harhoff & Wagner, 2009). Further, the uncertainty of patent grant hinders and delays the commercialization of the patent through licensing or transaction on the market for technology due to the unsettled property right (Gans et al., 2008; de Rassenfosse et al., 2016). Successfully granted patents are not only a valuable intangible asset but also signals of high growth potential for innovative firms at the nascent stage, playing a significant role in attracting venture capital and achieving successful IPO (Mann & Sager, 2007; Useche, 2014). Therefore, on the practical side, early prediction of the patent grant could reduce property rights uncertainty, better advise innovators' strategy on the patent application, guide investors' decision on investment in innovative firms, and help patent examiners prioritize the workloads. In addition, on the theoretical front, identifying the key determinants of patent grants also reveals the characteristics of high-value patents. Though the law specifies novelty, innovativeness, and practical applicability as the patent grant requirement (Liegsalz & Wagner, 2013), it is unclear how examiners interpret this requirement and how various applicant- and technology-level characteristics relate to the likelihood of patent grant. More in-depth theoretical questions need to be evaluated such as what are the roles of knowledge accumulation, technology complexity, and patent scope play in the decision to patent grant?

With such motivation, this paper aims to forecast patent examination outcomes using machine learning methods. More specifically, by applying interpretable machine learning methods, we combine two streams of literature and achieve two goals pursued. First, based on a comprehensive collection of about 400,000 invention patent applications in 2011, we apply machine learning models to predict patent grants using early information of patent publications. Second, using cutting-edge interpretation methods such as SHAP value, we aim to interpret the trained models by emphasizing the key determinants and delineating the relationship between predictors and patent grants. Thus, it closely relates to the two following streams of literature.

Patent grant related literature

The first stream of literature primarily explore variables concerning the patent examination process, such as patent pendency (Yang, 2008; Xie & Giles, 2011; Ahn et al., 2018), grant ratio (Yang, 2008), patent system discrimination (Webster et al., 2014; de Rassenfosse et al., 2020; de Rassenfosse et al., 2020), and patent grant outcomes (Kim & Oh, 2017; Choudhury & Haas, 2018). Yang (2008) noticed that foreign applicants tend to endure longer patent pendency than domestic applicants, highlighting the potential differential treatment based on the nationality by CNIPA (China National Intellectual Property Administration). Liegsalz and Wagner (2013) confirmed Yang's (2008) findings and further emphasized the positive role of prior experience and orientation in the Chinese market in reducing patent delay. As for the patent level characteristics, patents of high value tend to receive early grants, while the number of IPCs assigned and forward reference lengthens the examination for greater complexity of the task (Harhoff & Wagner, 2009; Liegsalz & Wagner, 2013).

While the patent pendency literature concern the duration of the process, the more related studies center on the influencing factors to the examination outcome. Patentability requires the technology to manifest a certain degree of novelty, innovativeness, and practical applicability as specified by the law (Liegsalz & Wagner, 2013). The literature has explored a number of important influencing factors. For instance, studies have uncovered the role of local and international collaboration as positive influencing factors of patent grant (Guellec & van Pottelsberghe, 2000; Drivas & Kaplanis, 2020). Applicant origins and international applications can also be important influencing factors. Webster et al. (2014) illustrated that the patent grant rate significantly varies among applicants of different countries and that domestic applicants were more likely to receive grants. Drivas and Kaplanis (2020) found that USPTO patent applications with greater inventor team and international background were more likely to be granted. How the patent relates to prior art is also an important factor as it reveals the novelty of the patented technology. Bekkers et al. (2020) showed that including standards-related documents for prior art may reduce the patent grant ratio but improve the quality of the granted patents overall. In a similar vein, Sun and Wright (2022) highlighted the non-blocking backward citation to prior art as the key determinants, which are positively correlated with patent grant. Further, during the examination process, it is also plausible that examiners and patent agents would play an important role. For instance, Klincewicz and Szumiał (2022) highlighted the role of patent attorneys and their rich experience as important predictors for successful patenting. Regarding the role of examiners, Lemley and Sampat (2012) uncovered that more experienced examiners tend to cite less prior art and grant more patents, and Frakes and Wasserman (2021) found that examiners' decisions are subject to peer effects through a mechanism of knowledge spillovers; they also showed that time constraints may hamper search and rejection effort, resulting in a higher granting rate.

While the literature has explored these diverse set of influencing factors[1], we still know little about how important each factor compares to each other. There needs to be a systematic analysis to rank these key variables and uncover how these variables predict the patent grant. In addition, a number of key factors, both patent and applicant level, that may potentially affect the outcome need to be thoroughly analyzed. For instance, backward citations,

[1] We have combed out and summarized these influence factors to patent grant in Table 5 in Appendix.

a sign of knowledge accumulation, may correlate with patent grants because technologies that are built on previous inventions tend to have greater practical applicability (Kwon & Geum, 2020). Also, the rich prior experience of the applicant may be positively correlated with the patent grant because the tacit knowledge could help avoid the common mistakes of the patent application. Thus, a more comprehensive approach is necessary to identify multiple key determinants of patent grants.

Machine learning prediction related literature

The second stream pertains to the recent development in machine learning techniques in technological forecasting based on patent information. Resorting to the powerful out-of-sample prediction capacity of machine learning techniques, recent studies have aimed to identify emerging technology or valuable patents (Lee et al., 2018; Chung & Sohn, 2020; Kwon & Geum, 2020; Choi et al., 2021; Kim et al., 2022), predict scientific breakthrough (Wang et al., 2021), forecast technology convergence (Kong et al., 2020; Cho et al., 2021), foresee innovation capacity (Ponta et al., 2020), and predict lawsuits on patents (Juranek & Otneim, 2021).

For example, Lee et al. (2018) proposed a machine learning approach to identifying emerging technologies, which was defined using the number of forward citations, at early stages using patent indicators that can be defined immediately after the patents are issued. Particularly, a feed-forward multi-layer neural network model was employed to capture the complex non-linear relationship between inputs and output variables. Similarly, Ponta et al. (2020) defined innovation capacity as the number of forward citations and applied various machine learning models to predict the innovation capacity of firms. More specifically, they ranked the feature importance of predicting innovation capacity for different countries. Chung and Sohn (2020) went a step further to implement a deep learning multi-modal approach that incorporated both patent features and textual information to better predict emerging technology at an early stage. Their contribution lies in the implementation of the state-of-the-art method such as a model combining the convolution neural network (CNN) with the Bi-LSTM layer, which is a natural language processing method that identifies the order of appearance and structure attributes of text documents. While many of these works consistently used forward citation as a measurement of the value of patents, Choi et al. (2021) overcame this limitation and proposed a hybrid approach considering both expert opinion and patent information to identify emerging technologies. Kim et al. (2022) used the licensing agreement as an indication of technology value, built machine learning models based on early technological characteristics to predict the technology valuation, and applied interpretable machine learning method to investigate the non-linear relationships between the input and output variables. Han et al. (2022) applied graph neural network models to forecast core patents and technology trends and showed that graph-embedding features based on network are superior to traditional patent features in prediction performance. In addition to the foresight of technology value, the machine learning model also exhibits a high capacity in predicting lawsuits associated with patents, which could help alert corporations of risky patents and mitigate litigation risk (Juranek & Otneim, 2021).

These studies sufficiently exploit the powerful out-of-sample prediction capacity of machine learning models to forecast emerging technology, technology convergence, legal events, etc. The logic behind this is that early information and variables are strongly correlated with the predicted outcome (such as technological value), and that machine learning models thoroughly explore all the linear, non-linear, and interaction effects among

these variables to achieve higher accurate prediction than conventional linear regression approach (Lee et al., 2018). Based on the same rationale, as some of the patent and applicant level variables are strongly correlated with patent grant (e.g. Webster et al., 2014, Drivas & Kaplanis, 2020, Schuster et al., 2020, Kim et al., 2022) as summarized in Table 5, we believe the machine learning model could also collect this information from important variables to accurately predict the examination outcome. However, the application of such predictive models has yet to be performed.

While the two literature streams[2] provided valuable insights into how individual factors determine patent pendency and grant outcomes and how to exploit the recent development of machine learning techniques to predict future technologies, there are limitations. First, there needs to be a comprehensive investigation of patent variables, especially at the patent and applicant level, to identify the key determinants of the patent grant. Such analysis could reveal a deeper theoretical mechanism of which variables are the key determinants of patent grants, how they relate to patent grants, and in what ways. It could also offer managerial implications on how to improve the application process to maximize the likelihood of successful patent grants. Second, the machine learning models tend to lack theoretical interpretability of their prediction results (Mullainathan & Spiess, 2017; Ghoddusi et al., 2019), but knowing what determining the evolution of future technology is as important as predicting it. To overcome the limitation, we combine machine learning models with interpretable methods to predict patent grants based on early patent information, identify the key determinants, and delineate the underlying relationship between these key determinants and patent grants.

The remainder of the article is organized as follows. "Data and methodology" section describes the patent data from CNIPA and explains the machine learning workflow. "Prediction results and interpretation" section reports the results and lays out the interpretations of those trained models. "Conclusion" section concludes the paper and discusses the potential practical applications and theoretical implications.

Data and methodology

Data and pre-processing

We use the invention application patents maintained by CNIPA and collect the dataset from *incopat.com*[3]. CNIPA follows an international convention for the patent application procedures (Liegsalz & Wagner, 2013). Invention application patents[4] are publicized on the publication dates with early information—such as the number of claims, applicants, backward citings, etc.—before the formal patent examination starts. The examination process, or patent pendency, lasts about two to three years before the final decision is made.

[2] To better present the two streams of related literature, we tabulate these related studies in Table 6 in the appendix.

[3] A paid membership on *incopat.com* allows quick downloading of a large amount of patent data with numerous patent indicators. CNIPA offers a website port for individual patent queries but not for downloading a large number of patent data.

[4] Only invention patent follows the three major stages, namely, application, publication, and final decision. In comparison, utility and design patents are published once and only once when it is officially published and granted, which means the rejected ones are not available to the public. That's why we use invention application patents instead of the other types.

In total, we collected about 400,000 invention applications publicized in 2011, a comprehensive collection of all applied invention patents after deleting samples with some abnormal legal conditions in that year. After data cleaning and deleting patent with abnormal conditions, we are left with 375,512 sample observations. Invention applications of all disciplines and fields with a comprehensive range of international patent classifications (IPCs) are included in the dataset, which provides wider coverage than similar studies that use patents in particular industries (e.g. Chung & Sohn, 2020). The data contains a large number of patent features including numerical, categorical, and textual. We adopt the following pre-processing and data cleaning procedure to tidy up the data and make it readily usable by machine learning models.

The numerical variables are standardized with unit standard deviation and zero mean because machine learning models like neural network and LASSO use gradient descent as an optimization technique that require data to be scaled (Zhang et al., 2019). The feature processing for categorical variables can be more complex since categorical features can not be directly fed into machine learning models. For low cardinality indicator, category feature with distinct values (n), say, less than 10, we apply one-hot encoding, whereas, for high cardinality indicator (i.e, $\mathtt{n} \geq \mathtt{10}$), we apply target encoding by replacing the categorical value with the mean of the target variable. As such, we turn the categorical features into quantifiable variables without creating a sparse data structure with too many column inputs.

Textual data also need to be quantified using textual extraction algorithms before feeding it into machine learning models. A popular and simple method of textual extraction is called the Bag-of-Words (BoWs), an approach to representing documents by fixed-length vectors terms, where each dimension is the numerical value corresponding to a word or a group of words (Li et al., 2020). We use TF-IDF (Term Frequency-Inverse Document Frequency) representation, based on the BoWs philosophy (Kim et al., 2019), to pre-process the textual information, which involves the following steps. First, sentences in patent abstracts are tokenized to single words that appeared and with the non-informative words (or stop words) removed. Second, create a dictionary using a vocabulary of words that uniquely appeared in abstracts and keep the top 1000 most frequent words. Third, we code the BoWs in (TF-IDF), which measures the frequency of a word by how often they appear in all documents and is offset by the number of documents in the corpus that contain the word. This approach is intended to reflect how relevant a term is in a given document, as the higher the TF-IDF score the more relevant that word is in that particular document compared to all others. As we have a very large corpus (i.e., abstract content from 37,5512 applications), each abstract may contain very few of the words in the vocabulary, which results in very sparse vectors. Hence, we apply Principle Component Analysis (PCA) method (Li et al., 2019) to reduce the dimensions of the 1000 TF-IDF scores and keep only the top 10 components. As such, we extract the textual information and compress the rich information into a relatively condensed form of data structure.

Once the three types of predictors are formatted in a quantifiable numerical way, we apply two statistical distance measurements to select relevant features, namely the Kolmogorov–Smirnov test (KS) (Katchanov et al., 2019) and Jensen–Shannon (JS) divergence (Asadi et al., 2018) (details of the KS and JS methods are included in Appendix "Feature selection" section), to filter out non-informative predictors. The informativeness of a predictor is defined by the difference between the variable distributions of the granted and rejected groups. After the selection process, we find a total of 42 informative features, including 18 numerical features, 14 categorical features, and 10 textual features. Figure 1

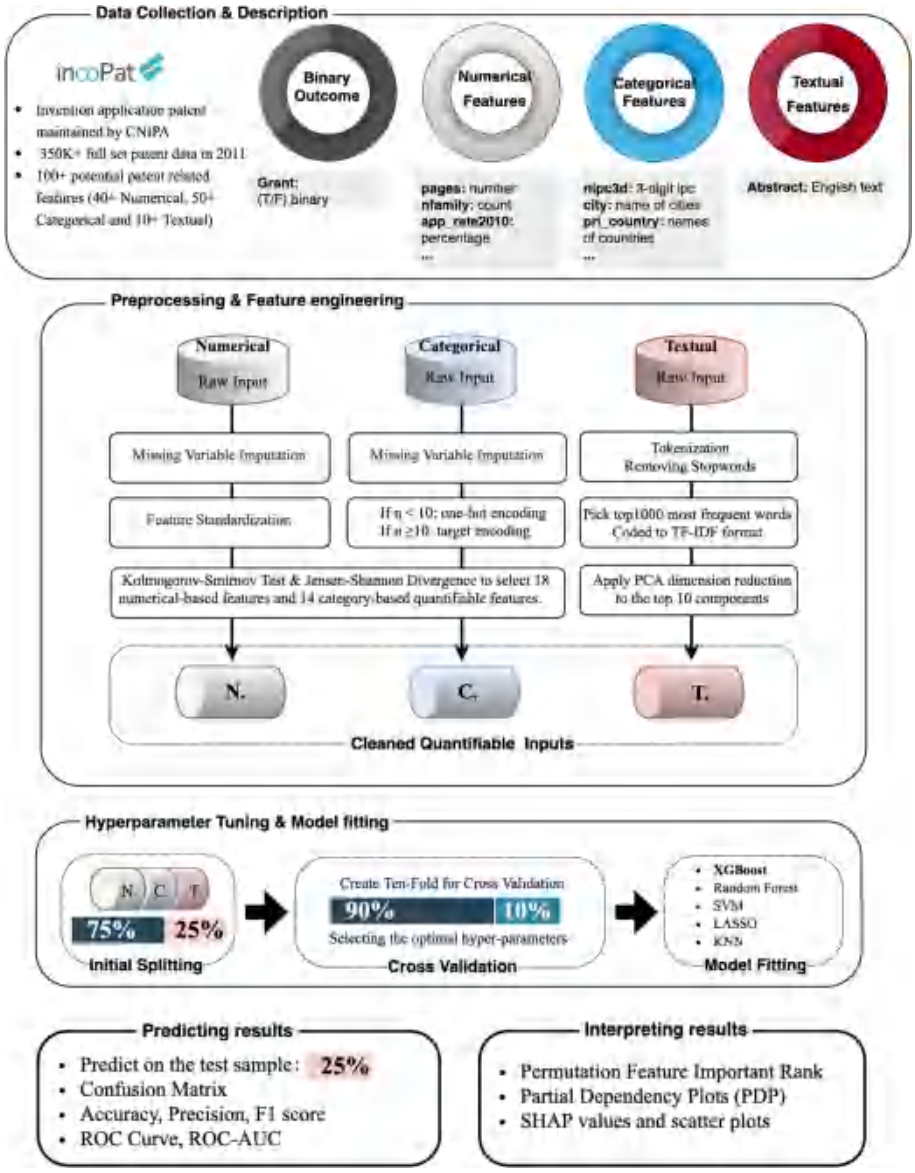

Fig. 1 The work flow and process for fitting machine learning models

summarizes the data pre-processing and feature engineering procedure to generate the cleaned quantifiable inputs as explained in Table 1.

Table 1 details the numerical and categorical predictors used in our models. As one can tell from Table 1, we pay relatively more emphasis on the patent-level variables because the patent-level variables provide the most detailed and idiosyncratic information for the patented technology, which determines patent grant. The examination outcome is determined by the novelty, innovativeness, and practical applicability of the technology embodied in

Table 1 Selected variable definition and statistic summary

Variables	Definition	Mean	SD
Outcome variable			
grant	A binary outcome variable indicating granted or not	0.46	0.50
Numerical features			
app_freq	Number of invetion applications by the same applicant in 2011	555	1434
agent_freq	Number of inventoin applications by the same agent in 2011	4848	7505
app_rate2010	The application success rate of the applicant in previous year	0.30	0.33
agent_rate2010	The application success rate of the agent in previous year	0.36	0.18
pages	Number of pages of the invention application document	13.16	17.11
citing	Number of inventors for a patent	3.59	2.65
cited	Number of forward citations	3.93	6.04
duration	The time span between patent application and publication	387	287
nfamily	Size of patent family	2.32	6.07
nchar_tl	Number of English words of the title	14.71	6.84
nchar_right	Number of English words of the first patent claim	249	256
nchar_ab	Number of English words in abstract	229	69
ntriz	Number of labeled principles out of 40 inventive solutions under the Triz invention theory	3.29	2.18
nipc7d	Number of seven-digit international patent classification (IPC) labeling. If a patent is label with C22C1/10, C22C19/03, and B01F7/18, then nipc7d counts 3	2.46	1.97
nipc3d	Number of three-digit IPC labeling. If a patent is label with C22C1/10, C22C19/03, and B01F7/18, then nipc3d counts 2, because the patent spans category C22 and B01	1.27	0.54
right_no	Number of patent claims	9.01	9.31
app_no	Number of applicants for a patent	1.10	0.36
inventor_no	Number of inventors for a patent	3.00	2.36
Categorical features			
ipc3d	A categorical variable indicating which 3-digit technology class a patent belongs to. In the case of multiply classes, we pick the first	NA	NA
econ	A categorical variable indicating which 3-digit industrial class a patent belongs to	NA	NA

Table 1 (continued)

Variables	Definition	Mean	SD
City	The city of the first applicant	NA	NA
app_country	The country of the applicant	NA	NA
pri_country	The county of prior right	NA	NA
Others	app_type, prior, agent, etc.	NA	NA

the patent, as specified by the patent law (Liegsalz & Wagner, 2013). For example, the variable *citing*—the number of forward citations of the patent—indicates knowledge accumulation and the level of the patent building upon previous technology. This patent-level information may be closely related to the level of practical applicability (more knowledge accumulation relates to high practical applicability) as required by the patent law. In comparison, macro-level variables, such as the city of the applicant or country of origin, may provide some additional information and correlate to potential examination outcomes but tend to be weak predictors because such information may not be closely linked to the patent value per se. Though in this paper, we use all the possible variables we can get to predict patent grant, we emphasize more on the patent-level variables because they are more informative of the potential patent grant.

In Table 1, the outcome variable *grant* indicates whether or not the invention application is legally granted by the end of 2020, the time of data collection. Since the legal information of each patent is updated in real time, the legal condition for the patent dataset is registered at the time of data collection, which is at the end of 2020 in our case. Due to long patent pendency, we cannot determine the examination outcome of the most recent patents—for example, a 2019 applied patent is likely still in the examination process. Thus, we use invention applications in the year 2011 for the consideration that it is early enough to have the examination decision made for most patents and also is recent enough to reflect the latest condition of patent prosecution in China. Overall, on average 46% of the invention applications are granted by the end of 2020. Although all variables potentially contribute to the likelihood of patent grants, the main variables of interest are the ones that are closely related to the applicants' prior experience, patent value, and whether or not the applicant resorts to professional patent agents.

app_rate2010 indicates the applicants' successful patent grant rate in the previous year. If an applicant did not apply in the last year, the variable is coded zero. This feature suggests applicants' prior experience. *pages* measures the length of the invention application document with a mean of 13.2 pages and an extensive standard deviation of 17.1. The length relates to the technological complexity of the patent. *citing* counts the patent backward citations. As inventing is a recombinant process that searches and recombines different streams of knowledge, a higher number of citations suggests the new invention is based on more previous innovations. Recent studies have highlighted the role of backward citation on patent value (Hur & Junbyoung, 2021). On average, a patent builds on 3.59 previous inventions. Forward citation, coded as *cited*, counts the future inventions that are based on the focal patent and is a widely accepted measurement of patent value (Jaffe et al., 1993; Hall & Trajtenberg, 2005; Moser et al., 2017). *nfamily*—the number of jurisdictions in which a single invention seeks protection—tends to indicate higher patent value, as innovators have the incentive to protect the highly valued inventions internationally so as to fully appropriate the technology (Harhoff et al., 2003). *nchar_right* measures the number of Chinese characters in the first independent patent claim.[5] The character length of the first patent claim, as argued by Sampat and Williams (2019), serves as a good proxy for patent scope because the first claim usually is the broadest and a longer length suggests a narrower scope. *right_n* counts the number of patent claims. A typical patent records 9 claims, and the standard deviation is 9.01. *app_n* is the number of applicants. More than one applicant indicates co-patent of inventive collaboration, which can be among individuals, firms,

[5] Independent claims are ones that are contingent of other claims. The first patent claim is always independent by law.

universities, and research institutes. As R & D collaboration has been shown an effective method for boosting innovation quality (Faems et al., 2005; Zhang & Tang, 2017), we believe the number of applicants could predict the prospect of patent prosecution to a certain degree. *inventor_n* records the inventor numbers, indicating internal collaboration effort. *nipc3d* and *nipc7d* count the number of IPCs each patent is assigned, which is another often used measurement of patent scope (Lerner, 1994; Novelli, 2015). The logic is that patents labeled in many classifications reveal a broader technology range. For instance, a patent labeled C22C1/10, C22C19/03, and B01F7/18 spans three 7-digit and two 3-digit classifications. The mean for *nipc3d* and *nipc7d* are 2.46 and 1.27. Among the categorical feature, *ipc3d* indicates the three-digit technology class the patent belongs to, which is interpreted as the technology field. *econ* indicates the three-digit level industrial classification for national economic activities, indicating the industry. *app_country* is country origin of the patent applicant, and *pri_country* is the country where the patent is initially applied. Other variables include *app_type* (such as individual, firm, research institute, or university), *prior* (whether the patent is applied based on a former application), *agent* (whether the patent application relies on the patent agent), etc.

Methodology

Machine learning models

As described in Fig. 1, model tuning and fitting come after data pre-processing. This study features a number of major machine learning models, including XGBoost (eXtreme Gradient Boost), random forest, K-nearest neighbors (KNN), and support vector machine (SVM). Among these models, we specifically demonstrate the details of the hyper-parameters tuning procedures for XGBoost model.

XGBoost is based on the original boosting models but more efficient (Friedman, 2001). In boosting, the trees are built sequentially such that each subsequent tree learns from its predecessors, updates the residual errors, and aims to reduce the errors of the previous tree. Each of these weak learners contributes some vital information for prediction, enabling the boosting technique to produce a strong learner by effectively combining these weak learners. XGBoost introduces optimal cache access patterns and data compression to build a scalable tree boosting system, and regularizing methods are used to control over-fitting and obtain better predicting performance. XGBoost thus is a highly efficient, flexible, and widely used application of gradient boosting system (Chen & Guestrin, 2016) and intensively used by the data scientists to achieve state-of-the-art prediction performance. It has become the most frequently used algorithm among data scientists in online machine learning competitions (Chen & Guestrin, 2016). More recently, scholars have applied XGBoost in predicting banking stress and failure in finance studies, and a set of key variables are identified to anticipate bank failures and defaults (Climent et al., 2019; Carmona et al., 2019).

The random forest algorithm is an extension of the bagging method which creates an uncorrelated forest of decision trees (Denisko & Hoffman, 2018). Different from decision trees that consider all the possible feature splits, random forests only use a subset of those features. Random forest algorithms have three main hyper-parameters, which include node size, the number of trees, and the number of features sampled. KNN, or the K-nearest neighbors algorithm, is one of the supervised learning techniques that assumes that similar things exist in close proximity (Friedman, 2017). In other words, similar things exist

Table 2 Hyer-parameters for XGBoost

Paramter	Explanation	Tuned value
trees	The number of trees contained in the ensemble.	1000(pre-set)
mtry	The number of predictors randomly sampled at each split.	33 (tuned)
min_n	The minimum data number required for a node to be split further.	13 (tuned)
tree_depth	The maximum tree depth	5 (tuned)
learn_rate	The rate the boosting algorithm adapts from iteration-to-iteration.	0.3 (default)
loss_reduction	The reduction in the loss function required to split further.	0 (default)
sample_size	The amount of data exposed to the fitting routine.	1 (default)
stop_iter	The number of iterations without improvement before stopping.	inf (default)

in close proximity, or, in our case, the granted or rejected patent applications should be approximate to each other in high dimensions of input features. We predict whether a patent application is granted based on how its neighbors are classified. The hyper-parameter of KNN is k which refers to the number of nearest neighbors to include. There is no structured method to find the best value for k but through cross-validation. SVM, or Support Vector Machine, can be used for both regression and classification and can produce significant accuracy with less computation power. It is a non-parametric method that finds a hyperplane in a high dimensional feature space that distinctly classifies data samples into grant group and reject group. SVM finds the hyper-plane with a maximum margin which provides reinforcement so that testing data can be classified with confidence. LASSO is a shrinkage method that imposes a $L1$ penalty on the parameter estimation (Zhang et al., 2019). The penalty encourages sparsity in the learned parameters and can drive many parameters (i.e., the least useful ones) to exactly zero. LASSO is therefore a continuous feature selection method. It is particularly well suited as we want to minimize the number of indicators our prediction model shall use.

Tuning and fitting XGBoost[6]

Following the data pre-processing as described in Fig. 1, once data has been pre-processed, the data is split into 75% training and 25% testing sets, and the training set is then used to create 10-folds for cross-validation in the next step of hyper-parameters tuning. XGBoost model uses a number of hyper-parameters, as reported in Table 2, of which some intend as a regularization mechanism during the fitting process to prevent over-fitting. For example, **min_n** stands for the minimum data number required for a node to be split further. That is, each node needs to contain at least **min_n** amount of samples, preventing overfitting by restricting too detailed tree splitting. Higher **min_n** means a more conservative model. It is essential to tune a set of hyper-parameters to reach an optimal model that delivers the highest prediction accuracy. This process of finding the optimal set of parameters is usually conducted through cross-validation, a statistic practice to estimate the performance of models with different sets of hyper-parameters. More specifically, we partition the training data sample into 10 folds, use

[6] In this section, we only detail the tuning and fitting process for the XGBoost model as an illustration and exemplification, since most machine learning models are tuned and trained in similar ways.

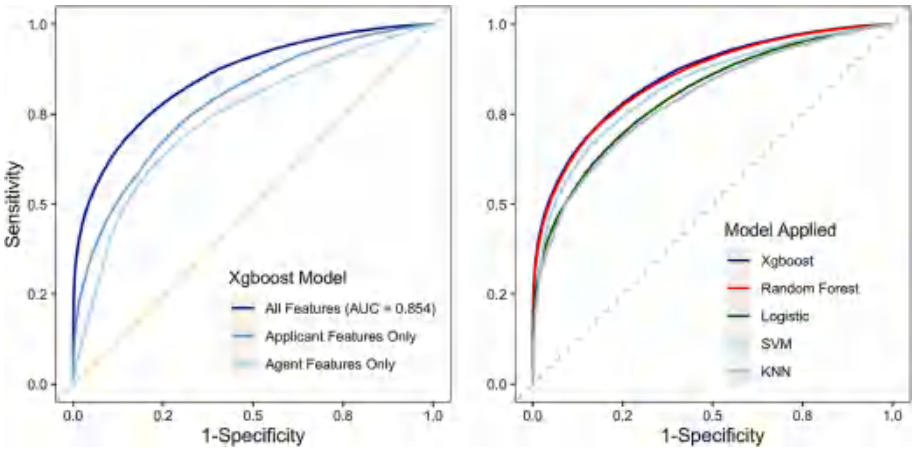

Fig. 2 The ROC curves of different models and specifications

the 9 folds to train the model, test it on the hold-out fold, then pick another fold as the hold-out, and repeat the process 10 times. Table 2 illustrates the parameter definitions and tune parameter values. We pre-set the parameter *trees* to be 1000 and set other non-essential hyper-parameters to default values. For the key hyper-parameters, such as *mtry*, **min_n**, and **tree_depth** (described in Table 2), we generate random hyper-parameter combinations through R command and feed into the tuning process. Once the model is tuned, it needs to finalize the XGBoost model by fitting the whole training set one last time and performing prediction on the test set. The optimal set of hyper-parameters is reported in Table 2.

In a similar way, we also tune other machine learning models and select the optimal hyper-parameters to achieve the best prediction capacity using the cross-validation method.

Prediction results and interpretation

Prediction results

The XGBoost model was tuned to maximize the AUC (Area under the curve) of receiver operating characteristics (ROC). The ROC-AUC is a metric to measure how well the model predicts the result using the whole range of thresholds. The ROC shows the relationship between sensitivity and (1 - specificity) as shown in Fig. 2.

Sensitivity is the proportion of positive that are correctly identified, i.e., the proportion of correctly predicted patent grants among all granted patents, which is indicated by the following equation:

$$\textbf{Sensitivity} = \frac{\textbf{TruePositive}}{\textbf{Positive}} = \frac{\textbf{TruePositive}}{\textbf{TruePositive} + \textbf{FalseNegative}} \tag{1}$$

Specificity is the proportion of correctly identified rejected patents among all rejected patents, which is expressed as:

Table 3 The Confusion Matrix of the XGBoost Prediction Results

		Actual grant	
		No	Yes
Predict	No	**33,993**	8,714
	Yes	12,986	**38,185**

The overall accuracy is 76.9% at a threshold of 0.5. Bold values indicate the number of patent applications predicted correctly and incorrectly according to actual results

Table 4 The our-of-sample prediction performance of machine learning models

	XGBoost	RandForest	KNN	SVM	LASSO	Logistic
Accuracy	0.769	0.761	0.720	0.753	0.721	0.721
Sensitivity	0.814	0.793	0.734	0.781	0.733	0.734
F1 Score	0.767	0.751	0.716	0.741	0.718	0.717
ROC-AUC	0.854	0.845	0.797	0.832	0.802	0.801

$$\textbf{Specificity} = \frac{\textbf{TrueNegative}}{\textbf{Negative}} = \frac{\textbf{TrueNegative}}{\textbf{TrueNegative} + \textbf{FalsePositive}} \tag{2}$$

(1-specificity) is the percentage of false negatives among all negatives, and there is a strong correlation between sensitivity and (1-specificity). The intuition is that, in order to identify the true positive as much as possible, one needs to take some risk of falsely predicting the negative as positive. Hence, there's an upward relationship between sensitivity and (1-specificity) as illustrated in Fig. 2. A random classification yields a ROC curve of 45-degree straight line because under a random guess sensitivity equals (1-specificity), i.e., the predicted positive could equally be actual false or positive. The closer the curve is to the upper-left corner, the better the model performs in overall prediction. People may choose different thresholds to predict positive or negative, because they may face different costs of two types of errors. Hence, the area under the ROC curve (ROC-AUC) can be used to indicate overall model prediction performance weighted with a whole range of thresholds.

In the left panel of Fig. 2, the dark blue line indicates the AOC curve of the fitting XGBoost model, with a ROC-AUC score of 0.854. It also shows that as more features the model used, it delivers better performance as the curve approaches the upper-left corner. On the right panel, the graph shows the ROC curves of different models specification. The random forest model delivers similar but slightly worse performance as the XGBoost model, while the baseline logistic model is noticeably worse in overall prediction performance.

While ROC-AUC measures the overall performance of the statistical model at different thresholds, a more intuitive and straightforward metric is accuracy, the total correct prediction over the total predictions, which is reported in the confusion matrix in Table 3. The confusion matrix is a summary of results from predicting the test sample. The two columns indicate the actual patent grant, and the two rows report the binary prediction made by the model. Table 3 shows that the overall accuracy of XGBoost

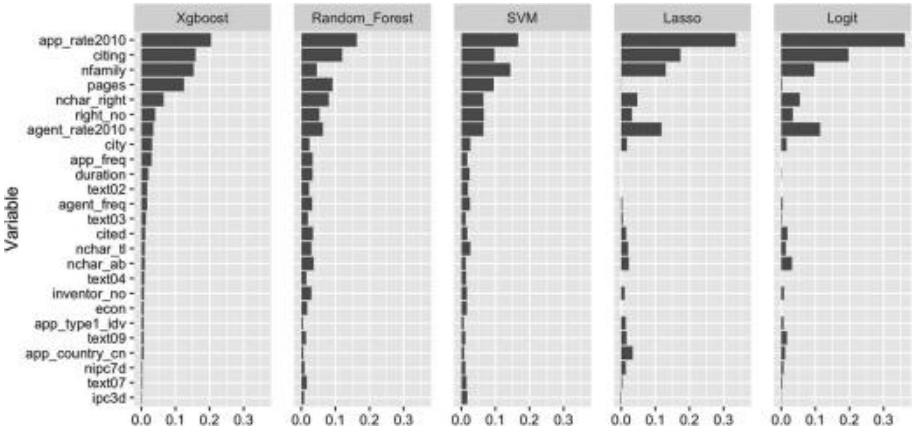

Fig. 3 The permutation-based feature importance for different models

model can reach 76.9%. Considering a random guess could get correct about half of the time, 76.9% is a large improvement in forecasting whether a patent will eventually get granted or not, hence eliminating the uncertainty of intellectual property rights. Note that the sensitivity, $38,185/(8,714+38,185) = 81.4\%$, is higher than the specificity, $33,993/(12,986+33,993) = 72.4\%$, meaning it is easier to have the actual granted patents correctly identified than the other way around.

We also train other machine learning models that yield the prediction results presented in Table 4, showing four different prediction performance metrics. In comparison, XGBoost exceeds all other methods in out-of-sample prediction capacity, achieving state-of-the-art performance. The random forest comes close to the second, and SVM also delivers a decent prediction performance. LASSO achieves a very similar result as logistic regression, suggesting overfitting may not be a serious issue for linear models. KNN model obtains the worse prediction performance compared to the rest. Overall, the comparison suggests that tree-based and non-linear models such as XGBoost and random forest are best at prediction performance relative to other models.

Feature importance

While the ROC curve and the confusion matrix suggest the overall predicting accuracy of the model, these two metrics measure the prediction performance but do not answer which features are more useful at prediction and how each predictor relates to the outcome variable. In this section, we focus on the interpretation of the models and open up the black box by highlighting which predictors impact the likelihood of patent grant and in what ways.

One such method is the feature importance metric, a technique that assigns each variable a score based on how relatively useful they are at the prediction task. An importance value is an assigned score to a feature of a trained model that indicates the relative importance of that feature when making a prediction. There are different ways to measure how important a feature is. These methods can be categorized into two types—model-specific and model-agnostic based approaches. The model-specific feature importance measurement only works for a certain type of machine learning model. For instance, tree-specific feature importance

computes the average reduction in impurity across all trees in the forest due to each feature. Variables that tend to split nodes closer to the tree root result in a larger importance value. One advantage of tree-based models is that the feature importance can be easily calculated. A variable that has frequently been chosen to split nodes, and leads to more impurity reduction will be assigned by a larger importance value (Hastie et al., 2017). However, the model-specific feature importance is not comparable among different models. Hence, we adopt a model-agnostic approach to measure feature importance consistently among various models.

Permutation-based feature importance, a model-agnostic method, initially introduced by Breiman (2001) based on the idea that a feature is more "important" if shuffling its value magnifies the prediction error, and the opposite means low feature importance. The detailed algorithm is described in Appendix "The permutation feature importance" section. Figure 3 illustrates the permutation-based feature importance of the top variables for these trained models. The order based on the relative importance shows a fairly consistent ranking. Exceptions come from LASSO and the Logit classification. LASSO regression, due to the built-in feature selection mechanism of $L1$ regularization, forces some of the features to be zero in importance, such as "*pages*" and "**app_freq**", possibly due to a non-linear relationship with the target variable. Yet, this does not rule out the importance of these variables in reality, because there could be a more complicated non-linear relationship among the variables. Linear models tend to pick up the features that have a linear relationship with the outcome variable while neglecting the non-linear features. Overall, the various models generate consistent estimations of feature importance ranks. The leading variable is the "**app_rate**2010", the applicants' success rate in the previous year, suggesting the importance of prior experience and innovation capacity of the applicant in the successful patent application. Other variables such as *citing*, *nfamily*, *pages*, **nchar_right**, **agent_rate**2010, **right_no**, etc., also rank consistently high on the importance metrics. In the following section, we center on these highly ranked variables and attempt to interpret how these variables impact the patent grant decision as predicted by these machine learning models.

SHAP values and interpretation of XGBoost

While the feature importance plot compares which variables are relatively more important in prediction, it is unclear what are the marginal effects of predictors on the outcome variable and by how much. For linear parametric models, such as linear regression models, the marginal effects can be simply interpreted using the coefficients. However, marginal effects of predictors are not obvious in non-linear, non-parametric, and complex models such as XGBoost, Random forest, and SVM. Even though we know from the feature importance metric that the number of backward citation significantly relates to the likelihood of patent grant, is there a positive or negative correlation between the two? is the correlation manifested in a linear or curvilinear fashion?—these questions need to be addressed using methods such as SHAP value and PDP. Thus, to better interpret the models, we resort to a method called SHAP (SHapley Additive exPlanations), which was proposed by Lundberg and Lee (2017) and use the more traditional method PDP (Partial dependence plot) to ensure the robustness of our results. SHAP borrows the concept of Shapley Value in the economic game theory (Winter, 2002), which was proven to be the unique fair-allocation solution of marginal contributions by each player in a game. SHAP, like Shapley value determining how merits are distributed fairly among players in a game, aims to allocate marginal contributions among predictors in machine learning models. The goal of SHAP is to explain the prediction of an instance X by computing the contribution of each feature

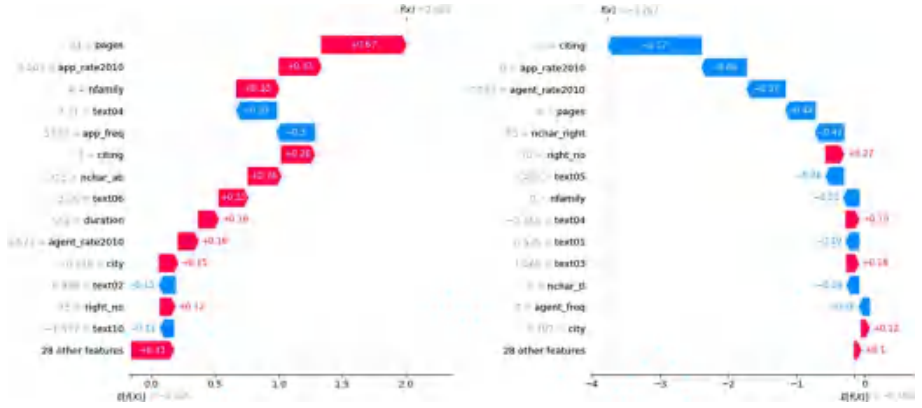

Fig. 4 Local interpretation for individual patents using SHAP values. *Note* The left panel presents the SHAP value waterfall-plot for the patent (number: *CN102843739A*) titled "Method, device and system for handover among femtocells" and applied by Huawei Device Co Ltd. On the right, the panel illustrates for the patent (number: *CN103174477A*) titled "Absorption type heat energy expansion machine" applied by Jiangsu Hezheng Energy Technology Co Ltd. The X-axis of both plots indicate the log odd ratio. $\mathsf{E}[\mathsf{f}(\mathsf{x})] = -0.155$ suggests the log odd ratio for the whole is $^{-}0.155$, which translates to 46.2% base rate of patent grant. The first patent was granted, whereas the second failed to acquire legal grant. Both patents are correctly predicted using our XGBoost model

to the prediction. Thus, SHAP is a method mainly for local interpretation—although SHAP is also capable of global interpretation of models when aggregated (Lundberg et al., 2020). The calculation of SHAP values can be model-agnostic as well as model-specific. We apply an algorithm called treeSHAP proposed by Lundberg et al. (2020) to calculate the SHAP of the XGBoost trained model. Figure 4 showcases two sample patents of how SHAP values are used to interpret the marginal contribution of features.

Figure 4 illustrates the waterfall-plot of SHAP value for two individual patents. Both plots illustrate how the difference—between global average $\mathbf{E}[\mathbf{f}(\mathbf{x})] = -0.155$ and predicted log odd ratios $f(x)$—decomposed and allocated between variables. The plot ranks feature using the absolute SHAP value for each patent. The red arrow indicates a positive marginal contribution to the likelihood of a patent grant, while the blue arrow indicates a negative contribution. Due to the addictive nature of the SHAP value, the summation of SHAP values for all variables exactly equals the difference between the average and predicted log odds ratio.

This type of plot in Fig. 4 elegantly illustrates the marginal contributions of features for each sample observation. For instance, the left panel in Fig. 4 illustrates the SHAP value decomposition for a patent (publication number: *CN102843739A*) applied by Huawei Device Co Ltd. The XGBoost model predicts this patent to be granted with a log odds ratio of 2.005, which translates to 88.1% chance of patent grant (it turns out the prediction is correct). The difference between average and predicted log odds ratio $\mathbf{f}(\mathbf{x}) - \mathbf{E}[\mathbf{f}(\mathbf{x})] = 2.005 + 0.155 = 2.16$, which is then attributed to each feature. The first variable "*pages* = 24" increases the log odds ratio by 0.67, the largest increment among these variables. The second "**app_rate**2010" suggests that the same applicant had about 0.507 successful patent grant rates in the previous year, which increases the likelihood of grants according to the model. The third variable, having 4 family patents also increases the likelihood of a patent grant by 0.33 units. The next two variables *text*04 and **app_freq** moderately reduce the grant likelihood. The model estimates that the fourth principle component reduces the grant rate by 0.31 and that "**app_freq** =1122" reduces the grant change

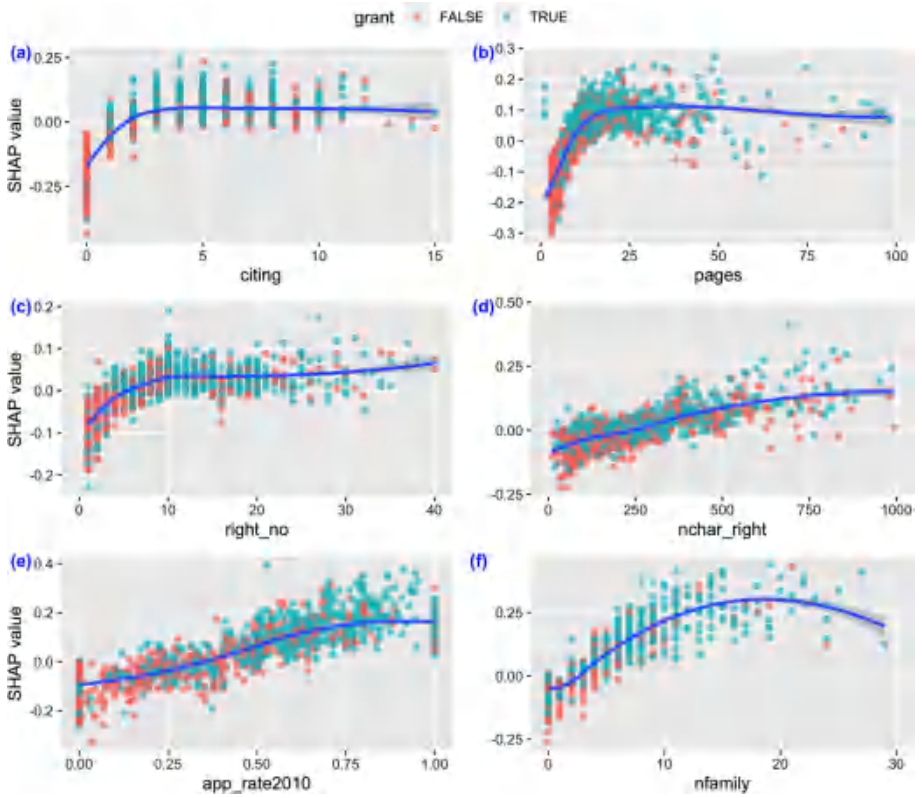

Fig. 5 The scatter plot of SHAP value and the focal variable

by 0.3—applying for 1122 invention patents in a year could be too many for a high grant ratio. The other notable variables such as backward citation (*citing*), number of abstract words (**nchar_ab**), duration between application and publication (*duration*), and the city of the applicant (*city*) all contribute positively to the likelihood of patent grant. The rest 28 other features contribute 0.33 to the likelihood of a patent grant.

On the right in Fig. 4, the panel illustrates a patent (publication number: CN103174477A) with a low predicted chance of grant (It is correctly classified as being rejected). The predicted log odds ratio of the patent is −3.767, or a 2.25% chance of being granted in probability. The most noteworthy aspects include zero backward citation (*citing*), low applicant grant rate in 2010 (**app_rate2010**), low agent grant rate in 2010 (**agent_rate2010**), short document page length (*pages*), and a low number of words in the first right claim (**nchar_right**), all contributing negatively to the likelihood of patent grant. Due to the nature of non-linearity and interaction effect, SHAP values for the same variable differ significantly from case to case. Hence local interpretability, as illustrated in Fig. 4, clarifies the underlying mechanisms of how black-box machine learning models predict the outcomes, and, in our case, explains which variables are important in contributing to successful patent grants.

SHAP value is not only useful in local interpretability, but the aggregation of individual SHAP values also indicates the global relationship between predictors and the outcome variable (Lundberg et al., 2020). In Fig. 5, we draw SHAP dependence plots, essentially

scatter plots with fitted curves relating the SHAP values to the actual feature values, and try to understand the complex relationship between actual feature value and the SHAP values, or likelihood of patent grant. We pay close attention to these features ranked high in the feature importance metric.

The SHAP dependence plot in Fig. 5 presents rich and condensed information. The y-axis of each plot represents the SHAP values and the x-axis measures the actual value of each feature. Each dot is one sample observation and the color indicates whether the patent is granted—blue indicates grant and red indicates rejection (to make the plot more readable, we randomly sample only 2000 patents). The cluster of dots also indicates the sample distribution over the x-axis—as one could see from the plot, most of the distributions are left-skewed.

The blue line goes through a scatter plot, indicating the smooth average relationship between the two axes. For instance, Panel (a) in Fig. 5 illustrates the relationship between the backward citation and the SHAP values, where most samples are concentrated on low citation numbers. When there is zero backward citation, the SHAP values are mostly negative and the dots are red (hence rejected), indicating having no backward citation decreases the likelihood of a patent grant. On average, having zero backward citation reduce about 0.15 probability of patent grant (as indicated by the blue line). It is also reflected in the colors that the patents with low backward citations are more likely to be red, i.e., rejected. As backward citation increases, the SHAP values and likelihood of patent grant increase, and once above four, the SHAP values stay positive and level off. The backward citation has been used as an indicator of knowledge flow from existing technology in the literature (Jaffe et al., 1993), and can be a sign of patent value (Harhoff et al., 2003). In recent works, Hur and Junbyoung (2021) uncovered the role of the network configuration of backward citation on the value of invention because innovation is essentially a recombinant process that involves searching and recombining different streams of knowledge. Hence, as modern technologies are becoming increasingly complex, innovation building on existing technology is an indispensable step to creating patentable and successful technology, suggesting the strategy of reinventing the wheel is unlikely to succeed.

In Panel (b) of Fig. 5, the patent document page length (*pages*) also increases the likelihood of patent grant. Patent length, as van Zeebroeck et al. (2009) noted, has risen steadily over the years because as technology becomes more complex, more words are required to describe and claim it. Patent page length, thus, can be interpreted as technology complexity. Having short document length, thus low technological complexity is associated with low SHAP value and also a low rate of patent grant, implying that a certain level of technological complexity is necessary for successful innovation and patent grant. The threshold occurs at around 20. That is, below 20 pages, patent document length drastically increases the likelihood of patent grant, and once beyond that threshold, the effect disappears.

Patent length differs from patent claim counts (**right_no**) in that the former relates to technology complexity while the latter is often used as patent cope (Novelli, 2015). The number of patent claims in Panel (c) shows that it increases the likelihood of a patent grant and the threshold occurs at about 10. Thus, like technology complexity, a certain degree of patent scope increases the likelihood of patent grants. While the other measure of patent scope indicates the opposite. The length (character count) of the first patent claim (**nchar_right**) is another commonly used measurement of patent scope—more words mean less scope because a longer claim implies more conditions that must be met for the patent to be violated, and vice versa (Kuhn & Thompson, 2019). It is shown in Panel (e) that the length of the first claim monotonically increases the patent grant probability, suggesting narrower patent scope increases the grant rate and thus resonating with Marco et al. (2019)

for similar findings. Theoretically, the patent scope should be negatively correlated with the likelihood of a patent grant because a greater scope translates to a larger set of technologies from which the owner can exclude others from using (Marco et al., 2019). By this logic, we argue that the length of the first claim is a better representation of the patent scope, as it shows a monotonic relationship with the likelihood of patent grant in that narrower scope indicates a higher possibility of patent grant. Our study echoes with Kuhn and Thompson (2019) who argued the number of claim rights is a weak indicator of patent scope, because, in our case, it contradicts that logic—a narrower patent is more likely to be granted.

Panel (e) presents the applicants' previous success rate (**app_rate**2010) on the SHAP values. There is almost a monotonic upward relationship between previous success rate and SHAP value, implying the importance of prior experience in obtaining successful patent grants. Panel (f) shows that the number of patent family (*nfamily*) increases the likelihood of patent grants since applicants are more likely to seek international patent protection for valuable patents. Moreover, for robustness, we explore the whole variables in Appendix "The partial dependence plots" section using a similar method called partial dependence plots (PDP), a similar technique that relates the feature value to the average predicted patent grant rate. In addition to the six variables, we also lay out the relationships with outcome variables for many other predictors.

Conclusion

Invention patent examinations take years to complete and the outcome is uncertain. Thus, the patent application process creates a predicament where applicants need to publicize valuable technologies while the outcome is unclear. An early prediction of whether a patent will eventually get granted, hence, helps reduce the uncertainty of intellectual property rights. This study applies a number of appropriate (i.e. eligible in making a binary prediction based on inputs of various data types) machine learning approaches for forecasting patent grants at the early stage using patent and applicant level information. The central tenet is that early information contained in the invention patent publication documents—such as abstract, document length, backward citations, applicant's prior experience, and patent family—can provide evidence for the likelihood of a future patent grant. To this end, numerous patent features were extracted from a comprehensive collection of invention patent applications publicized in 2011 from the CNIPA, and various machine learning models and interpretable methods were applied to achieve two major goals pursued. The contributions of this paper thus come in two folds—using machine learning methods to predict patent grants and interpreting the results, which respectively correspond to the practical and theoretical implications of our study.

The first contribution is to implement machine learning models that integrate several patent-level variables, including numerical, categorical, and textual data, under one single machine learning framework to predict patent examination outcomes. The cross-validation method is used to tune the model hyper-parameters. Once tuned, the XGBoost model, accomplishing the best prediction performance among the models, is able to achieve 76.9% accuracy of binary classification and 0.845 of the ROC-AUC value. The machine learning approach delivers a major improvement in out-of-sample prediction accuracy over the baseline logistic classification model.

The practical implications of our accurate prediction model are substantial. For patent examiners, our model can help optimize the examination process by stratifying patent

applications into groups of different predictive results. By doing so, examiners can make better or faster judgments regarding the filed patent. For patent applicants, an accurate prediction of the patent grant enables them to conduct a better cost-benefit analysis of whether the benefit of a certain probability of protecting intellectual property outweighs its opportunity cost. Venture capitalists can also benefit from our model by evaluating the patent portfolios of start-up firms whose patents are at the application stage, and then determining the investment potential of the start-up firms. Additionally, in the technology market, early prediction of examination outcomes may facilitate potential technology transactions even before the final grant decision, since patents with property rights settled down are more likely to be licensed and transferred (Gans et al., 2008). Therefore, our study represents a significant step in predicting patent grants and reducing IPR uncertainty, and has important practical implications for patent examiners, applicants, venture capitalists, and technology markets.

Our second contribution goes beyond just building of prediction model, which usually lacks interpretability (Loyola-Gonzalez, 2019; Choi et al., 2021), but probes into the black-box model to uncover the key determinants of patent grant and identify the marginal contributes of input variables. Using the permutation-based feature importance metric, we rank the key features in terms of their relative impact on the models. Moreover, we explore a locally interpretable method called SHAP values and uncover the marginal contribution of each feature to the model prediction for individual patents. Using the calculated SHAP values, we create scatter plots that relate the real feature values to the respective SHAP values. The fitted loess curve in these scatter plots, known as SHAP dependence plot (Molnar, 2020), illustrates the relationship between the SHAP values and the actual feature values, showing whether it increases or decreases the target and whether such a relationship is in a linear, monotonic, or more complex way. Then, we also conduct a robustness study using the more traditional partial dependence plot (PDP) to confirm our findings (Friedman, 2001).

Based on the interpretation, our study thus makes a number of theoretical contributions to the literature on the patent examination process. The feature importance metric suggests that only a handful of variables are the key determinants of patent grants. Variables such as page length, backward citations, previous year success rate (prior experience), and the number of patent family are some of the topmost prominent determinants of patent grants. The feature importance metric consistently ranks *pages*, *citing*, **app_rate**`2010` as the top determinants by different machine learning models. In addition to individual experience, where the applicant comes from, i.e., cities and countries of the applicant, though less significant, also contribute to the prediction of patent grants. At the applicant level, resorting to professional patent agent service also increases the likelihood of patent grants, especially those agents with high service quality, as suggested by **agent_rate**`2010`.

At the technology level, we uncover that a successful patent grant is closely related to knowledge accumulation, technology complexity, and patent scope, which are embodied in the backward citation, page length, and the number of claims. Backward citation, a patent makes to related earlier patents and other sources of knowledge contain not only the information of the cited patents but also how the cited resources relate to the citing patent (Higham et al., 2021). It is often regarded as a sign of knowledge flow (Jaffe et al., 1993) and knowledge accumulation (Schoenmakers & Duysters, 2010; Kwon & Geum, 2020). In our prediction models, backward citations is one of the major determinants of a successful patent grant, suggesting that a certain level of reliance on existing technologies, or knowledge accumulation, is essential for the patent grant and that successful modern technology needs to stand on the shoulder of giants. Even though a patent grant requires a certain level of novelty and inventiveness

(Liegsalz & Wagner, 2013), we show that equally important is knowledge accumulation based on pre-existing technologies. Patent length indicates the technological complexity in that the more complex the technology is more words are required to describe and claim it (van Zeebroeck et al., 2009). A high level of patent length, hence technology complexity, increases the likelihood of patent grants. The same applies to the number of patent claims, indicating patent scope. Notably, all three, backward citation, patent length, and the number of claims have curvilinear relationships with the outcome variable—there exists a cutoff effect, once beyond a certain level of backward citation (4 citations), patent length (20 pages), and the number of patent claims (12 claims), the marginal contribution of these variables stay roughly constant. Our result also highlights the other measurement of patent scope, length of the first claim, and finds that narrower patent scope is associated with a higher likelihood of a patent grant, resonating with the previous work by Marco et al. (2019). It thus creates a conflicted implication for the effects of patent scope—while measured in the length of the first claim, the patent scope is negatively correlated with the patent grant; measured in the number of claims, it however suggests the opposite. Such contradiction, we believe, comes from the inconsistent measurement of patent scope, which should need to be solved in the future study.

A larger patent family contributes positively to the likelihood of patent grants because more valuable patents tend to seek international protection. In comparison, many other patent variables are not as important. For instance, technology range (the number of IPC labels), inventor team size, and examination duration all have little effect on the likelihood of a patent grant as suggested in the result interpretation.

In a nutshell, our interpretable machine learning approach, complementing the traditional regression method that focuses on one or two key variables in the patent examination process, takes a comprehensive approach, investigates all the patent-related variables we can get, ranks the features in terms of its important contribution to patent grant, delineates the correlation between each variable and the likelihood of patent grant, and uncovers that knowledge accumulation, technology complexity, and patent scope are the major determinants to patent grant. Most of the key determinants positively correlate with the likelihood of patent grants in a curvilinear fashion.

Limitation and future extension

Our study is not without limitation. Though XGBoost and other machine models with the BoWs approach are effective ways to predict patent grants in our case, it however may not be the optimal way to extract information from textual data. The BoWs approach disregards the word orders and hence ignores the semantic meaning of textual information. It is possible that recurrent neural network models (RNN) with deep learning methods could do a better job extracting textual information and improve the prediction accuracy (Chung & Sohn, 2020). Also, although we include as many patent features as possible, due to the data limitation, we did not include patent examiner characteristics and did not consider different patent systems other than the CNIPA. It is thus unclear how some of our theoretical conclusion extends to other countries. One future extension is to investigate how international patent offices differ in terms of influencing factors to patent grant.

For future research, our study exemplifies the application of interpretable machine learning methods, such as permutation-based feature importance, SHAP values, and partial dependence plots in the field of technology forecasting (Molnar, 2020). We believe these methods are powerful tools to explain what happens in black-box models and reveal the underlying mechanisms that merit theoretical explanations. Such

application could be potentially used to explain what determines emerging technology (Lee et al., 2018; Choi et al., 2021), scientific breakthrough (Kwon & Geum, 2020), technology convergence (Cho et al., 2021), etc.

On the theoretical front, our results point to the importance of knowledge accumulation in determining the success of patent grants. We notice that there still exists limited research on how knowledge accumulation affects the success of emerging technologies. Numerous concepts of knowledge accumulation could be extracted from patent backward citation information—especially variants such as average quality, country origins, and technology cycle time of knowledge accumulation—on the future performance of the technology. Also, we find conflicting results on how patent scope affects the patent grant. On one hand, we show that smaller patent scope, as indicated by a longer first claim, is associated with a higher likelihood of patent grant, thus confirming Marco et al.'s (2019) findings. On the other, our results also suggest that the number of claims, another common measure of patent scope, is positively associated with the likelihood of a patent grant. The conflicting result calls for further investigation to answer: which variable is the more appropriate measurement of patent scope and what is the real effect of patent scope on patent grant?

Appendix

Table of literature review

Feature selection

While machine learning models are good at handling large dimensions of inputs variables, there is a consequence of using non-informative predictors in that they could add uncertainty to the predictions and reduce the overall effectiveness of the model (Kuhn & Johnson, 2013). Hence, the logic behind feature selection is that removing non-informative features could improve model prediction performance. We measure the informativeness using the difference between variable distributions of the granted and rejected groups. The intuition is that, the more different distributions of the focal variable between the granted and rejected are, the more informative that variable is. Thus, we calculate the differences between the pair for each feature and we believe the difference is proportional to the usefulness of features.

The distribution distance are illustrated in Fig. 6. The blue density distribution indicates the granted sample while the yellow distribution indicates the rejected sample. The extent to which the two distributions differ indicates is the evidence of informativeness of that focal predictor.

We apply two well-established approaches to measure the distributional difference.

Kolmogorov–Smirnov test (KS test)

KS test is a form of minimum distance estimation used to compare two datasets. The KS statistic quantifies a distance between the empirical distribution functions of two groups of samples. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and

Table 5 Summary of influencing factors to patent grant

Article	Variables	Research and findings
Guellec and van Pottelsberghe (2000), Drivas and Kaplanis (2020)	Collaboration; Inventor team	Guellec and van Pottelsberghe (2000) showed that patenting strategy and local and international collaborations are key positive determinants to patent grant. Drivas and Kaplanis (2020) uncovered that USPTO patent applications with greater sizes of inventor teams and international collaboration are more likely to be granted
Webster et al. (2014), Zhao (2022), Guellec and van Pottelsberghe (2000)	Country origins; International application; Patent family	Webster et al. (2014) show international background is one key positive determinant to patent grant. Zhao (2022) show that patent applications through PCT (Patent Cooperation Treaty) system are more likely to be granted than those direct applications to the Canadian patent office. Guellec and van Pottelsberghe (2000) show that family size is a major determinant of patent grant but the relationship is likely to be a non-linear one
Schuster et al. (2020)	Applicant characteristics	Schuster et al. (2020) demonstrated that applicants of minority or females received unfavorable treatment in terms of patent ratios
Klincewicz and Szumiał (2022)	Patent agent/attorney	Klincewicz and Szumiał (2022) highlighted the role of patent attorney and its rich experience as important predictors for successful patenting
Lemley and Sampat (2012), Frakes and Wasserman (2021)	Patent examiners	Lemley and Sampat (2012) showed that more experienced examiners tend to cite fewer prior art and grant more patents. Frakes and Wasserman (2021) found that examiners' decisions are subject to peer effects through a mechanism of knowledge spillovers
Bekkers et al. (2020), Sun and Wright (2022)	Backward citations	Bekkers et al. (2020) showed that including standards-related documents for prior may reduce the patent grant ratio but improve the quality of the granted. Sun and Wright (2022) illustrated that non-blocking backward citations to prior arts positively predict patent grant while the blocking citations do the opposite

Table 6 Literature review summary

Article	Method	Data	Research and findings
Patent grant related			
Guellec and van Pottelsberghe (2000)	Probit regression	391,440 patent application at EPO between 1985 and 1992	Using whether patents get granted or not as the dependent variable, this study finds that the determinants of patent value are not only confined to citation, patent family, and renewal but also patent strategy and international cooperation
Yang (2008)	Lagged regression	Secondary data of WIPO statistics from 1985 to 2002	Compares and contrasts the invention patents of the US and China focusing on the application and granting practices. It shows that both countries make efforts to provide equal treatment to domestic and foreign applicants for patents in terms of pendency, but domestic applicants enjoy more certainty of the patent being granted within the pendency period than foreign applicants in both countries
Harhoff and Wagner (2009)	Survival analysis, Cox proportional hazards (PH) model.	215,265 EPO patent applications filed between 1982 and 1998	Valuable patents are granted significantly earlier than less valuable ones, controlling for important variables such as claims, number of patent references, originality, etc
Lemley and Sampat (2012)	OLS, Probit regression.	9,960 original patent applications from 2001 to 2006 from Delphion	Finds that more experienced examiners cite less prior art, are more likely to grant patents and are more likely to grant patents without any rejections
Liegsalz and Wagner (2013)	Survival Analysis, Cox proportional hazards (PH) model.	443,533 patent applications filed at the CNIPA between 1990 and 2002	Finds that the average grant lag between 1990 and 2002 was 4.71 years, with variation across 30 different technology areas. Chinese applicants achieved faster patent grants than their non-Chinese counterparts

Table 6 (continued)

Article	Method	Data	Research and findings
Webster et al. (2014)	Fixed effect probit and logit models.	70,473 patent applications from 10 different patent offices	Finds that domestic inventors have a higher likelihood of obtaining a patent grant than foreign ones and that the positive domestic inventor effect is stronger in areas of technological specialization in the domestic economy
Frakes and Wasserman (2017)	Binary regression	1.4 million utility patent applications filed from 2001 to 2012 and 9,000 examiners	Insufficient examination time may hamper examiner search and rejection efforts, leaving examiners more inclined to grant invalid applications, and these additional patents being granted on the margin are of below-average quality
Tong et al. (2018)	Weibull distribution with the parametric AFT model, Cox model.	About 1.1 million invention patent applications to CNIPA from 1993 to 2006	Analyze how various applicant, application, and environmental characteristics affect the duration of patent examination for all three competing examination outcomes (grant, withdrawal, and refusal)
Drivas and Kaplanis (2020)	Logit and probit regression models.	475,857 patent applications filed between 2001 and 2009 that disclose at least one inventor from one of the 28 EU countries	Finds that a US entity, either as inventor or owner, plays an important role in securing the grant. Low innovative countries, international collaborations matter more for the likelihood of securing the patent grant than high innovative countries
Schuster et al. (2020)	Logit regression	3,924,889 USPTO patent applications filed between 2000 and 2015	Finds that race and gender affect the patent grant rate, and minority and women applicants are significantly less likely to secure a patent relative to others
Bekkers et al. (2020)	Diff-in-Diff regression (DID)	78,194 pairs of patents applied at both USPTO and EPO	When examiners are given systematic access to documents by SSO members, they are more likely to reject patents and improve the quality of granted patents

Table 6 (continued)

Article	Method	Data	Research and findings
Zhu et al. (2022)	Mathematical modeling and Cox proportional hazard models.	258,104 pharmaceutical patent applications filed between 1985 and 2017 at CNIPA	Finds a U-shaped relationship between the level of technology diversity of a patent application and the pendency time for patent office decisions, including both grants and refusals
Machine learning related			
Kyebambe et al. (2017)	Naive Bayes (NB), Aritificial Neural Networks (ANN), Logistic Regression, Support Vector Machine.	USPTO utility patents from 1979 to 2010	Using supervised machine learning method to predict emerging technologies
Lee et al. (2018)	ANN, Random Forest, Support Vector Machine (SVM).	35,356 USPTO patents belonging to 424 classes from 2000 to 2009	Patent forward citation as the outcome variables, discretized into four levels. The models achieve a relatively high accuracy rate
Chung and Sohn (2020)	A combination of CNN (convolutional neural network) with Bi-LSTM.	Semiconductor patents granted by the USPTO from 2000 to 2015	Patent forward citation as the outcome variables, discretized into three levels, grades A, B, and C. The multi-modal model combines patent indices and textual information and achieved higher prediction accuracy than models without textual information
Kwon and Geum (2020)	LR, SVM, RF, Adaptive Boosting (ADA), Gini feature importance metric.	363,620 G06F (Electric digital data processing) patents collected from Google Patent websites	Using forward citations as the outcome variable, machine learning methods achieve a high accuracy rate. The feature importance suggests that knowledge accumulation quality is the most important feature that helps predict the promising technology

Table 6 (continued)

Article	Method	Data	Research and findings
Ponta et al. (2020)	Regularized Least Square, ANN, RF.	A total of 1,635,734 patent family and 8,093,138 individual patents of 32 states	Number of forward citation patents as a proxy to innovation capacity (IC). The machine learning methods approach extracts the significant features in predicting IC with a low percentage of error
Zhou et al. (2020)	Generative adversarial network (GAN) combined with deep neural network (GAN-DNN), SVM, RF.	Thomson Innovation patent database from 2000 to 2016	Use GAN to expand the scale of sample data and to label whether a patent is emergent or non-emergent. Then DNN is used to train on the augmented trained sample, and achieve more accurate prediction than regular machine learning methods without data augmentation
Cho et al. (2021)	RF, SVM, Latent Dirichlet Allocation (LDA).	USPTO patent from 2012 to 2014	Construct various link prediction indicators among patents and use the information to predict technology convergence. Then LDA is applied to identify the top keywords of high technology convergence
Choi et al. (2021)	Semi-supervised learning, RF, SVM, Extreme Gradient Boosting(XGBoost).	6,287 USPTO patents related to electric vehicle batteries	Combining an expert evaluation with a machine learning method, the semi-supervised technique, in which a small amount of labeled data (patents evaluated by experts) are used with a large amount of unlabelled data, was applied to achieve accurate prediction of emerging technologies

Table 6 (continued)

Article	Method	Data	Research and findings
Wang et al. (2021)	SVM, TextCNN, BERT, BERT-KeyPos-Last sentence.	12 million journal articles, 135,526 citances, and 467 abstracts tied to the high-quality breakthrough papers from PubMed	First, identify 237 potential breakthrough articles using both the “self" and “others" evaluation process. Based on the breakthrough articles identified, using SVM, TextCNN, and BERT to train the models to identify abstracts with breakthrough evaluations
Kim et al. (2022)	Random Forest, XGBoost, SVM, Interpretable methods	2,288 university licensed patents matched to a large volume of university patent data	They investigate the importance of the features of technological characteristics (and their interactions) in technology valuation and offer theoretical and practical implications of the analysis results

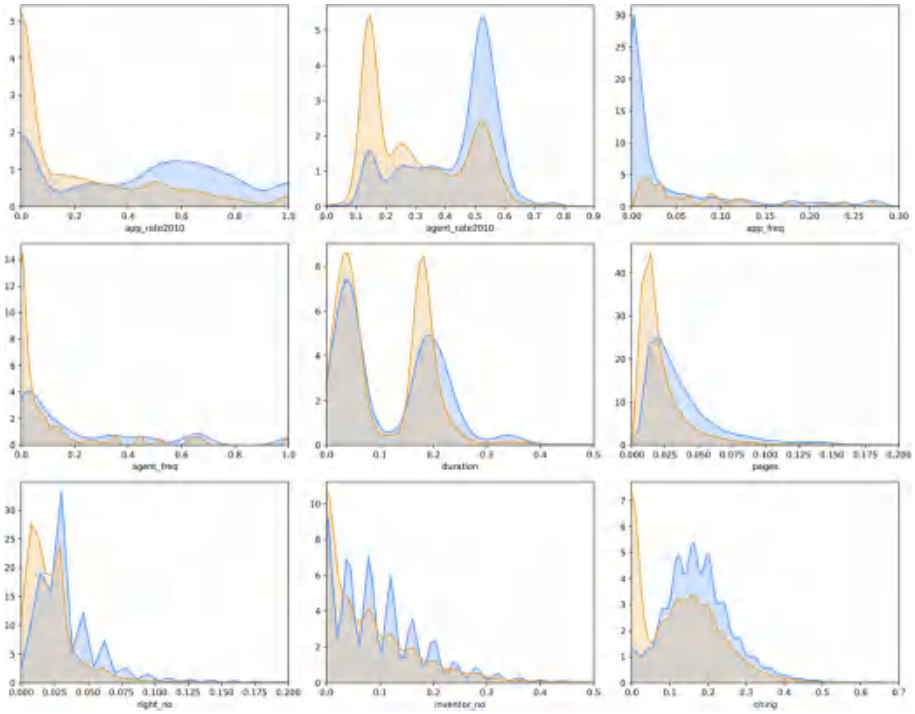

Fig. 6 The different between distribution of the granted and th rejected

shape of the empirical cumulative distribution functions of the two samples. The KS test is based on the maximum distance between these two curves produced by two empirical cumulative distribution functions. The KS test statistic is defined as:

$$\mathbf{D_n} = \max_{1 \le x \le N} (\mathbf{F_a}(\mathbf{Y_x}) = \mathbf{F_b}(\mathbf{Y_x})) \tag{3}$$

where F is a theoretical cumulative distribution of the distribution being tested which must be a continuous distribution. The statistic evaluates the greatest separation between the cumulative distribution $\mathbf{F_a}$ and $\mathbf{F_b}$. When $\mathbf{F_a} \neq \mathbf{F_b}$, larger values of D are expected.

Jensen–Shannon divergence

JS divergence is based on Kullback–Leibler (KL) Divergence divergence, a method of measuring statistical distance sometimes referred to as relative entropy. KL calculates a score that measures the divergence of one probability distribution from another. It can be used as a distance metric that quantifies the difference between two probability distributions. If $\mathbf{F_a}$ and $\mathbf{F_b}$ represent the probability distribution of a discrete random variable, the KL divergence is calculated as a summation:

$$\mathbf{KL}(\mathbf{F_a}||\mathbf{F_b}) = \sum \mathbf{F_a}(\mathbf{x}) \log \left(\frac{\mathbf{F_a}(\mathbf{x})}{\mathbf{F_b}(\mathbf{x})} \right) \tag{4}$$

The intuition for the KL divergence score is that when the probability for an event is large, but the probability for the same event is small, there is a large divergence. When the probability from is small and the probability from is large, there is also a large divergence, but not as large as the first case.

JS divergence extends KL divergence to calculate a symmetrical score and distance measure of one probability distribution from another. The JS divergence can be calculated as follows:

$$\mathbf{JS}(\mathbf{F_a}||\mathbf{F_b}) = \frac{\mathbf{KL}(\mathbf{F_a}||\mathbf{F_M})}{2} + \frac{\mathbf{KL}(\mathbf{F_b}||\mathbf{F_M})}{2} \tag{5}$$

where $\mathbf{F_M}$ is calculated as:

$$\mathbf{F_M} = \frac{\mathbf{F_a} + \mathbf{F_b}}{2} \tag{6}$$

Based on both information from the KS and JS statistic measurement of the distributional difference between the granted and rejected, we select the relatively more informative features as predictors of our machine learning models.

The permutation feature importance

Permutation feature importance measures the increase in the prediction error of the model after the features' values are permuted. The idea is intuitive—if shuffling a feature value increases the model error more, then the model relied on this feature for the prediction to a greater degree. A feature is hence "unimportant" if shuffling its values leaves the model error unchanged because, as we can interpret, the model ignored the feature for the prediction. This method was initially introduced Breiman (2001) for random forest, and extended to a model-agnostic version by Fisher et al. (2019). The permutation feature importance algorithm we use was based on Fisher et al. (2019):

Input: trained model $\hat{\mathbf{f}}$, feature matrix X, target vector y, error measure $\mathbf{L}(\mathbf{y}, \hat{\mathbf{f}})$.

1. Estimate the original model error $\mathbf{e_{orig}} = \mathbf{L}(\mathbf{y}, \hat{\mathbf{f}}(\mathbf{X}))$(e.g. R-square or mean squared error)
2. For each feature $\mathbf{j} \in \{1, \ldots, \mathbf{p}\}$ do:
   - Generate feature matrix $\mathbf{X_{perm}}$ by permuting feature *j* in the data *X*. This breaks the association between feature *j* and true outcome *y*.
   - Estimate error $\mathbf{e_{perm}} = \mathbf{L}(\mathbf{Y}, \hat{\mathbf{f}}(\mathbf{X_{perm}}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance as quotient $\mathbf{FI_j} = \frac{\mathbf{e_{perm}}}{\mathbf{e_{orig}}}$ or difference $\mathbf{FI_j} = \mathbf{e_{perm}} - \mathbf{e_{orig}}$
3. Sort features by descending *FI*

In practice, we use the **vi_permute** command in the *R* package *vip* to generate feature importance for XGBoost, random forest, SVM, LASSO, and Logit.

The partial dependence plots

The SHAP dependence plot shows a fitted curve for a scatter plot of the SHAP values and the actual feature values, illustrating the relationship between the two, whether it is a linear, monotonic, or more complex relationship.

The more traditional way to present the interrelation as a visualization tool is the Partial Dependence Plot (PDP) proposed by Friedman (2001), which is able to illustrate how each variable affects the chance of patent grant. The difference between the SHAP dependence plot and PDP is that first, the y-axis of the SHAP dependence plot is the SHAP value (or marginal contribution) while the y-axis of PDP is the average prediction. Thus, the y-axis in the SHAP dependence plot is centered on zero, while it is centered on 0.46 (the average grant ratio) for PDP. Second, SHAP dependence is able to present more information, such as the distribution of individual samples. Yet, PDP is a more well-accept and applied tool for model interpretation. We apply PDP for the robustness of our results.

Given the output function $f(x)$ of machine learning algorithm, the partial dependence of f on variable $\mathbf{X_s}$ is defined as (let c be the complement set of s)

$$\mathbf{f^s(x_s) = E_{X_c}[f(x_s, X_c)] = \int f(x_s, x_c) dP(x_c)} \tag{7}$$

where the PDP $\mathbf{f^s}$ is the expectation of f over the marginal distribution of all variables except for $\mathbf{x_s}$. In practice, PDP is estimated by taking average over the training data with fixed $\mathbf{X_s}$

$$\mathbf{\bar{f}^s(x_s) = \frac{1}{n} \sum_i = 1f(x_s, X_c^i))} \tag{8}$$

PDP plots indicate direct causal effect of a variable s on the outcome variable if the back-door condition is satisfied (Zhao and Hastie, 2021). We admit that some of the variables are likely endogenous or descendants of other variables in the terminology of the directed acyclic graph (DAG). For instance, the variable duration, the time duration between application and publication, can be affected by other confounding variables. Though may not be explained in a causal manner, nonetheless many of the high correlation nature are worth pointing out.

Figure 7 illustrates 12 PDP plots for the major predictors. To keep the plots comparable, we limit the y-axis ranging from 0.4 to 0.6 and the x-axis according to the distribution of each predictor. Note that from Panel (a) to (f), the PDP generates similar relational patterns between the predictors and output variable as ones from Fig. 5. Backward citation, patent length, number of claims, length of the first claim, grant rate of the previous year, and patent family all show up a similar pattern of positive correlation with the outcome variable. For backward citation, patent length, and the number of claims, there exists a threshold effect. Once beyond 4 for *citing*, 20 for *pages*, and 10 for **right_no**, the marginal benefits of these variables stay positive and constant. While **nchar_right**, **app_rate**2010, and *nfamily* all show up a monotonic increasing relationship with the outcome variable. Hence, PDP confirms the relationships of the predictors and outcome variables as suggested and discussed in our main paper.

Patent length or *pages* increase the likelihood of a patent grant. There are notable data outliers with one or two pages, and these are patents with an above-average

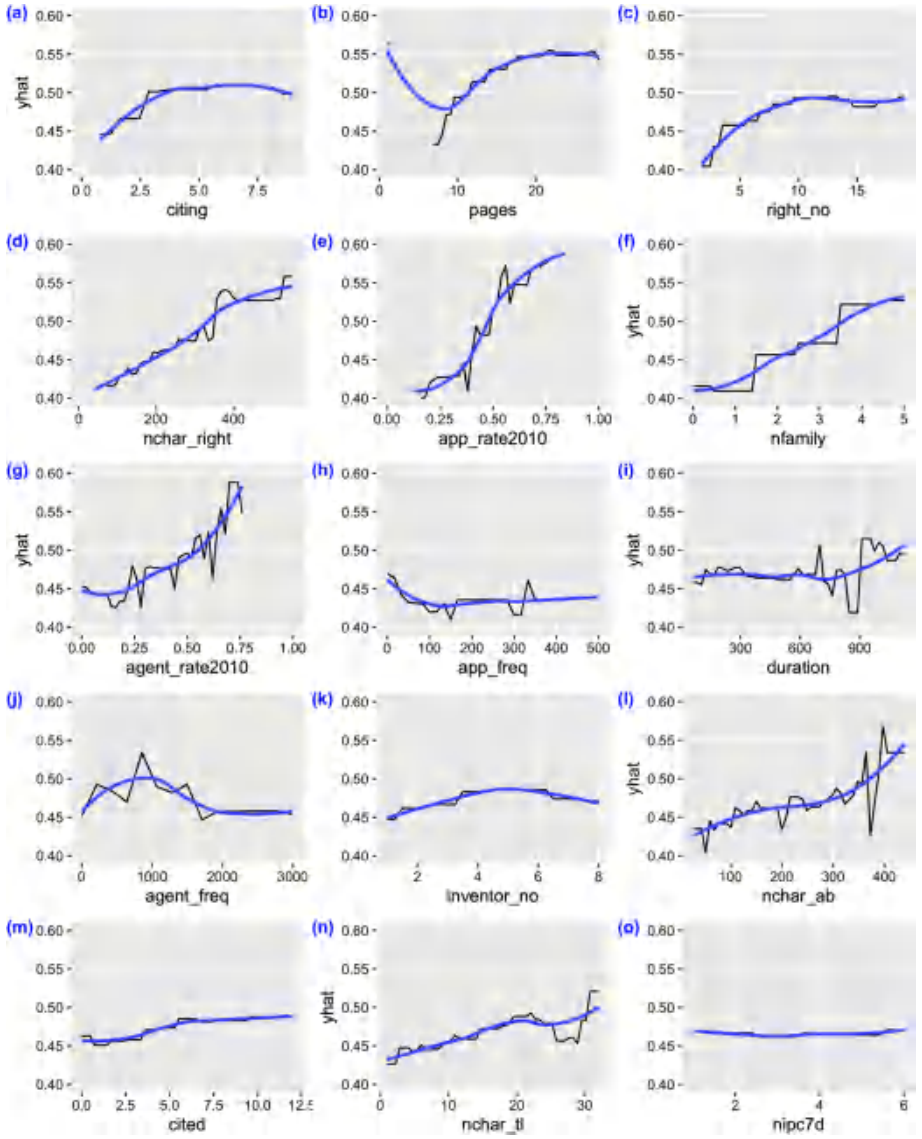

Fig. 7 The partial dependence plots of major variables

grant rate. It is also present in Panel (b) of Fig. 5 that there are four blues dots on the upper left corner. There are possible data anomalies but it does not interfere the model interpretation.

Besides the six major variables, we also explore other less important yet noteworthy predictors. The agent success rate in the previous year, **agent_rate**`2010`, increases the likelihood of patent grant, suggesting relying on a high-quality patent agent could help the examination outcome. Agent application load, $\mathbf{agent_f req}$, initially increases but eventually decreases the patent grant rate. Application frequency in the same year, **app_freq**, could

lower the outcome variable because too many applications can be distracting. Panel (i) shows that the duration between application and publication has a complex relationship with the outcome variable. Panel (k) shows that Inventor numbers have little effect on the likelihood of patent grant. There is a slight increase for a higher number of inventors. Panel (l) shows the length of the abstract, **nchar_ab**, which increases the patent grant because it may indicate the technology complexity similar to patent length.

It is also interesting to note, Panel (m) and (o) show forward citation and the number of IPC assigned have little predicting power of whether a patent gets granted or not.

Funding The authors are grateful for financial support from Philosophy and Social Science Foundation of Zhejiang Province (Grant No. 22NDQN246YB), National Natural Science Foundation of China (Grant No. 72204222), the Characteristic and Preponderant Discipline of Key Construction Universities in Zhejiang Province (Zhejiang Gongshang University-Statistics), Zhejiang Gongshang University "Digital+" Disciplinary Construction Management Project (Grant No. SZJ2022B001), and the Tailong Finance School.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahn, J. M., Mortara, L., & Minshall, T. (2018). Dynamic capabilities and economic crises: Has openness enhanced a firm's performance in an economic downturn? *Industrial and Corporate Change, 27*(1), 49–63.

Asadi, M., Ebrahimi, N., Kharazmi, O., & Soofi, E. S. (2018). Mixture models, Bayes Fisher information, and divergence measures. *IEEE Transactions on Information Theory, 65*(4), 2316–2321.

Bekkers, R., Martinelli, A., & Tamagni, F. (2020). The impact of including standards-related documentation in patent prior art: Evidence from an EPO policy change. *Research Policy, 49*(7), 104007.

Breiman, L. (2001). Random forests, machine learning 45. *Journal of Clinical Microbiology, 2*(30), 199–228.

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance, 61*(MAY), 304–323.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

Cho, J. H., Lee, J., & Sohn, S. Y. (2021). Predicting future technological convergence patterns based on machine learning using link prediction. *Scientometrics, 126*(7), 5413–5429.

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics, 126*(7), 5431–5476.

Choudhury, P., & Haas, M. R. (2018). Scope versus speed: Team diversity, leader experience, and patenting outcomes for firms. *Strategic Management Journal, 39*(4), 977–1002.

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change, 158*, 120146.

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. *Journal of Business Research, 101*(AUG.), 885–896.

de Rassenfosse, G., Palangkaraya, A., & Webster, E. (2016). Why do patents facilitate trade in technology? Testing the disclosure and appropriation effects. *Research Policy, 45*(7), 1326–1336.

de Rassenfosse, G., Palangkaraya, A., & Raiteri, E. (2020). Technology protectionism and the patent system: Strategic technologies in China. *The Journal of Industrial Economics*.

de Rassenfosse, G., Palangkaraya, A., & Hosseini, R. (2020). Discrimination against foreigners in the US patent system. *Journal of International Business Policy, 3*(4), 349–366.

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences, 115*(8), 1690–1692.

Drivas, K., & Kaplanis, I. (2020). The role of international collaborations in securing the patent grant. *Journal of Informetrics, 14*(4), 101093.

Faems, D., Van Looy, B., & Debackere, K. (2005). Interorganizational collaboration and innovation: Toward a portfolio approach. *Journal of Product Innovation Management, 22*(3), 238–250.
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.
Frakes, M. D., & Wasserman, M. F. (2017). Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. *Review of Economics and Statistics, 99*(3), 550–563.
Frakes, M. D., & Wasserman, M. F. (2021). Knowledge spillovers, peer effects, and telecommuting: Evidence from the U.S. patent office. *Journal of Public Economics, 198*, 104425.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. London: Springer.
Gans, J. S., Hsu, D. H., & Stern, S. (2008). The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays. *Management Science, 54*(5), 982–997.
Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics, 81*, 709–727.
Guellec, D., & van Pottelsberghe, B. (2000). Applications, grants and the value of patent. *Economics Letters, 69*(1), 109–114.
Hall, B. H., & Harhoff, D. (2012). Recent research on the economics of patents. *Annual Review of Economics, 4*(1), 541–565.
Hall, B. H., & Trajtenberg, J. M. (2005). Market value and patent citations on. *Rand Journal of Economics, 36*(1), 16–38.
Han, S., Huang, H., Huang, X., Li, Y., Ruihua, Y., & Zhang, J. (2022). Core patent forecasting based on graph neural networks with an application in stock markets. *Technology Analysis & Strategic Management, 34*, 1–15.
Harhoff, D., & Wagner, S. (2009). The duration of patent examination at the European patent office. *Management Science, 55*(12), 1969–1984.
Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. London: Springer.
Higham, K., De Rassenfosse, G., & Jaffe, A. B. (2021). Patent quality: Towards a systematic framework for analysis and measurement. *Research Policy, 50*(4), 104215. Publisher: Elsevier.
Hur, W., & Junbyoung, O. (2021). A man is known by the company he keeps?: A structural relationship between backward citation and forward citation of patents. *Research Policy, 50*(1), 104117. Publisher: Elsevier.
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly journal of Economics, 108*(3), 577–598.
Juranek, S., & Otneim, H. (2021). Using machine learning to predict patent lawsuits. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3871701
Katchanov, Y. L., Markova, Y. V., & Shmatko, N. A. (2019). The distinction machine: Physics journals from the perspective of the Kolmogorov-Smirnov statistic. *Journal of Informetrics, 13*(4), 100982.
Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Information Sciences, 477*, 15–29.
Kim, J., Lee, G., Lee, S., & Lee, C. (2022). Towards expert-machine collaborations for technology valuation: An interpretable machine learning approach. *Technological Forecasting and Social Change, 183*, 121940.
Kim, Y. K., & Oh, J. B. (2017). Examination workloads, grant decision bias and examination quality of patent office. *Research Policy, 46*(5), 1005–1019.
Klincewicz, K., & Szumiał, S. (2022). Successful patenting–not only how, but with whom: The importance of patent attorneys. *Scientometrics, 127*(9), 5111–5137.
Kong, D., Yang, J., & Li, L. (2020). Early identification of technological convergence in numerical control machine tool: A deep learning approach. *Scientometrics, 125*(3), 1983–2009.
Kuhn, J. M., & Thompson, N. C. (2019). How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business, 26*(1), 5–38.
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
Kwon, U., & Geum, Y. (2020). Identification of promising inventions considering the quality of knowledge accumulation: A machine learning approach. *Scientometrics, 125*(3), 1877–1897.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change, 125*, 236–244.
Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.
Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. *Review of Economics and Statistics, 94*(3), 817–827. Publisher: The MIT Press.
Lerner, J. (1994). The importance of patent scope: An empirical analysis. *The RAND Journal of Economics, 25*(2), 319.
Li, K., Cursio, J. D., Sun, Y., & Zhu, Z. (2019). Determinants of price fluctuations in the electricity market: A study with PCA and NARDL models. *Economic research-Ekonomska istraživanja, 32*(1), 2404–2421.
Li, P., Mao, K., Yuecong, X., Li, Q., & Zhang, J. (2020). Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Knowledge-Based Systems, 193*, 105436.
Liegsalz, J., & Wagner, S. (2013). Patent examination at the State Intellectual Property Office in China. *Research Policy, 42*(2), 552–563.
Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access, 7*, 154096–154113.
Lundberg, S. M., Lee, S-I. (2017). A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems (pp. 4768–4777).
Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. Publisher: Nature Publishing Group.
Mann, R. J., & Sager, T. W. (2007). Patents, venture capital, and software start-ups. *Research Policy, 36*(2), 193–208.
Marco, A. C., Sarnoff, J. D., & deGrazia, C. A. W. (2019). Patent claims and patent scope. *Research Policy, 48*(9), 103790.
Molnar, C. (2020). Interpretable machine learning, Lulu. com.
Moser, P., Ohmstedt, J., Rhode, P. W. (2017). Patent citations—An analysis of quality differences and citing practices in hybrid corn. Management Science mnsc.2016.2688
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.
Novelli, E. (2015). An examination of the antecedents and implications of patent scope. *Research Policy, 44*(2), 493–507.
Ponta, L., Puliga, G., Oneto, L., & Manzini, R. (2020). Identifying the determinants of innovation capability with machine learning and patents. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2020.3004237
Sampat, B., & Williams, H. L. (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review, 109*(1), 203–236.
Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy, 39*(8), 1051–1059.
Schuster, W. M., Evan Davis, R., Schley, K., & Ravenscraft, J. (2020). An empirical study of patent grant rates as a function of race and gender. *American Business Law Journal, 57*, 39.
Sun, Z., & Wright, B. D. (2022). Citations backward and forward: Insights into the patent examiner's role. *Research Policy, 51*(7), 104517.
Tong, T. W., Zhang, K., He, Z.-L., & Zhang, Y. (2018). What determines the duration of patent examination in China? An outcome-specific duration analysis of invention patent applications at SIPO. *Research Policy, 47*(3), 583–591.
Useche, D. (2014). Are patents signals for the IPO market? An EU-US comparison for the software industry. *Research Policy, 43*(8), 1299–1311.
van Zeebroeck, N., de la Potterie, B. P., & Guellec, D. (2009). Claiming more: The Increased voluminosity of patent applications and its determinants. *Research Policy, 38*(6), 1006–1020.
Wang, X., Yang, X., Jian, D., Wang, X., Li, J., & Tang, X. (2021). A deep learning approach for identifying biomedical breakthrough discoveries using context analysis. *Scientometrics, 126*(7), 5531–5549.
Webster, E., Jensen, P. H., & Palangkaraya, A. (2014). Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics, 45*(2), 449–469.
Winter, E. (2002). *The shapley value. Handbook of game theory with economic applications* (Vol. 3, pp. 2025–2054). London: North-Holland.

Xie, Y., & Giles, D. E. (2011). A survival analysis of the approval of US patent applications. *Applied Economics, 43*(11), 1375–1384.
Yang, D. (2008). Pendency and grant ratios of invention patents: A comparative study of the US and China. *Research Policy, 37*(6–7), 1035–1046.
Zhang, Guiyang, & Tang, Chaoying. (2017). How could firm's internal R &D collaboration bring more innovation? *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2017.07.007
Zhang, Y., Ma, F., & Wang, Y. (2019). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? *Journal of Empirical Finance, 54*, 97–117.
Zhao, L. (2022). On the grant rate of Patent Cooperation Treaty applications: Theory and evidence. *Economic Modelling, 117*, 106051.
Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics, 39*(1), 272–281.
Zhou, Y., Dong, F., Liu, Y., Li, Z., JunFei, D., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics, 123*(1), 1–29.
Zhu, K., Malhotra, S., & Li, Y. (2022). Technological diversity of patent applications and decision pendency. *Research Policy, 51*(1), 104364.

Authors and Affiliations

Li Yao[1] · He Ni[2]

✉ He Ni
nihe@zjgsu.edu.cn

[1] School of Business Administration, Zhejiang Gongshang University, Hangzhou, China

[2] Tailong Finance School, Zhejiang Gongshang University, Hangzhou, China