Using machine learning to predict patent lawsuits

BY Steffen Juranek and Håkon Otneim

DISCUSSION PAPER

NHH

Institutt for foretaksøkonomi
Department of Business and Management Science

FOR 6/2021

ISSN: 1500-4066
June 2021

Using machine learning to predict patent lawsuits

Steffen Juranek[a], Håkon Otneim[a]

[a]*NHH Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway*

Abstract

We use machine learning methods to predict which patents end up at court using the population of US patents granted between 2002 and 2005. We analyze the role of the different dimensions of an empirical analysis for the performance of the prediction - the number of observations, the number of patent characteristics and the model choice. We find that the extending the set of patent characteristics has the biggest impact on the prediction performance. Small samples have not only a low predictive performance, their predictions are also particularly unstable. However, only samples of intermediate size are required for reasonably stable performance. The model choice matters, too, more sophisticated machine learning methods can provide additional value to a simple logistic regression. Our results provide practical advice to everyone building patent litigation models, e.g., for litigation insurance or patent management in more general.

Keywords: patents, litigation, prediction,machine learning

JEL classification. K0, K41, O34

Email addresses: Steffen.Juranek@nhh.no (Steffen Juranek), Hakon.Otneim@nhh.no (Håkon Otneim)

1. Introduction

The use of intellectual property rights is an important part of corporate strategy in an increasingly knowledge-based economy. With the steep increase of the number of patent applications over the last decades (Kortum and Lerner, 1999; Hall, 2005), the management of patent rights became more and more prominent. One important aspect of patent management is that patents have to be enforced through costly litigation by the owner.

Being able to predict which patents will eventually be litigated can be immensely valuable. For example, corporations managing large patent portfolios would be able to focus their resources on high risk patents. Also the patent offices themselves can profit. Litigation is costly and, therefore, a potential burden for innovative activity. This is especially problematic against the background that patents have been criticized for not being perfectly defined property rights. It is argued that they have fuzzy boundaries, partly due to imperfect processes at the patent offices (e.g., Lemley and Shapiro, 2005). Hence, a prediction model could give patent offices guidance on which applications to focus particularly. As a final example there is, as for every risk, a case for insurance. Being better able to understand the risk helps solving the inherent adverse selection problem and allows better pricing. While patent litigation insurance has traditionally been a niche product, it experienced a surge in demand in the recent past (Ganglmair et al., 2018). Furthermore, we believe a patent litigation prediction model is beneficial for a large user group beyond these three examples as it provides information to everyone maneuvering in the technological space. Understanding what the high-profile patents may avoid inadvertent infringement of existing patents and save legal and operational costs.

The aim of our study is twofold. First, we aim to build a classification model for predicting whether a given patent will end up in court. Patent litigation is a relatively rare event, as only around one percent of patents in our sample eventually get litigated. Hence, the prediction task requires a detailed analysis for a proper bias-variance trade-off.

Second, we explore several issues of practical importance. Data on patents are increasingly available in principle, but several patent characteristics are nevertheless difficult to obtain. Hence, the need for larger training samples, which we typically regard as beneficial in empirical analyses, comes with a real cost of data aquisition. As a consequence, there may be a trade-off between prediction power and data costs when adding new observations and/or new chraracteristics to the prediction models. Furthermore, sophisticated machine learning models are computationally more demanding compared to the traditional logistic regression for instance. It is therefore important to understand whether, and by how much, such models outperform the simple logistic regression.

We give practical guidance on which of the following three dimensions matter most for the precision of a prediction model; (i) the number of patents (observations) used in the training data, (ii) the type and number of patent characteristics (independent variables), or (iii) the model type.

Against this background, our task can be seen as a prime example for statistical learning. Relying on the population of US patents from 2002 to 2005, i.e., more than 600000 patents, we find that the performance of model using small training samples show a very unstable performance, meaning that the performance varies drastically across different samples. However, the performance becomes reasonable stable already for medium-sized samples (100000-250000 patents). This means that even though small samples lead to unreliable results, one does not necessarily need to use the full population. In contrast, the predictors play a crucial role for the performance of the classification. The variables do not only have individual predictive power but also complement each other. Finally, we will see that tree-based machine learning models such as the Random Forest (Breiman, 2001) and the xgBoost (Chen and Guestrin, 2016), outperform the logistic regression when predicting patent ligigation by a small margin.

Our results imply that in order to predict patent litigation, it is (i) important to have a moderately large sample in order to be able to learn which characteristics that are associated with high risk of the relatively rare litigation event, (ii) beneficial to spend resources on deriving additional patent characteristics to extend the set of predictors, and (iii) useful to train a fairly complex machine learning model compared to the logistic regression.

Most of the economics, business and law literature on patent litigation focuses on inference, i.e., identifying variables that affect the litigation likelihood (e.g., Lanjouw and Schankerman, 2001, 2004; Cremers, 2004). In contrast, we aim to focus solely on prediction. There are only a few studies on predicting patent lawsuits. Within the legal literature, Chien (2011) classifies litigated patents in a small sample using, similar to Marco and Miller (2019) on the economics side, a matching approach. Campbell et al. (2016) and Wongchaisuwat et al. (2017) propose models for prediction patent lawsuits from an information science perspective using machine learning techniques.

2. Institutional background

A patent grants the patent holder the legal right to exclude others from making, using, offering, selling or importing the invention and products made by the invention throughout the United States (35 USC §154). Hence, it is a legal right to restrict access to the underlying technology. This is supposed to give inventors incentives for innovation in the exchange for disclosing the technology. Furthermore, it strengthens the market of innovation by enhancing the possibility to transfer intellectual property to other organizations and by increasing the effectiveness of licensing (Arora and Ceccagnoli, 2006).

In order to receive a patent, inventors, potentially together with companies to which the ownership will be assigned (the assignees), have to file an application to the USPTO. The application has to describe the invention, to state the scope of a patent in the claims, to cite prior-art and the technology classification. These elements are adjusted in the examination. If the examiner perceives the novelty, non-obviousness and utility requirement of the application to be fulfilled, a patent is granted to the applicants. A patent in the US is valid up to 20 years from the filing of the application, or up to 17 years from the grant, whichever is longer[1].

Unfortunately for a patent holder, patents have to be enforced through costly litigation.

[1] However, patents have to be renewed for a fee up to three times - after 3.5, 7.5 and 11.5 years.

Even worse, patents are not perfectly defined property rights, and the patentee might lose an infringement law suit. Patents can be seen as probabilistic property rights (e.g., Lemley and Shapiro, 2005). A patent infringement lawsuit starts with a formal complaint of the plaintiff at an US district court. Afterward, the pretrial discovery takes place. Defendants answer the complaint, and provide information for their opponents and the court. Once the discovery phase is over, the trial takes place, ending with a judgment. If the patentee prevails, the court may adjudge the patentee a certain amount for compensation of the damage.[2] Additionally, the judgment may deter the infringer(s) from selling any product that is based on technologies protected by the patent without consent of the patentee.[3] If the patentee loses, the court either finds the patent not infringed or even invalid[4]. Usually, all litigants bear their own costs.[5] The costs of a lawsuit can be significant. A survey conducted by the American Intellectual Property Law Association finds median litigation costs for patent infringement suits that last at least until the end of discovery of up to $ 5.0mn, depending on the amount at risk (American Intellectual Property Law Association, 2015). However, the litigants are able to settle their claims at any stage of the lawsuit. In fact, the majority of disputes settle. However, the risk of continuing and incurring the costs affects the decision to litigate.

3. Data and methodology

3.1. Data

The basis for our analysis is the USPTO's Patentsview database that includes the population of patents granted since 1976. The database includes all information printed on the patent. We add additional characteristics from two more datasets from the USPTO, the Patent Examination Research Dataset and the Patent Claims Research Dataset. Furthermore, we add a few additional variables from the OECD Patent Quality Indicators

[2] The main principles to calculate damages are unjust enrichment, lost profits and reasonable royalties.

[3] An alternative remedy is the payment of ongoing royalties, see Shapiro (2016) for a discussion.

[4] This is actually a substantial risk. Allison et al. (2014) finds that roughly 40 percent of all patents litigated in 2008 and 2009 whose validity was decided were found to be invalid. See also Henry and Turner (2016) who provide a long-term perspective on the development of the invalidity rates.

[5] This is the case unless the court finds a case exceptional according to US code Section 285.

database. Finally, we use the USPTO's Patent Litigation Docket Reports Data to identify litigated patents.

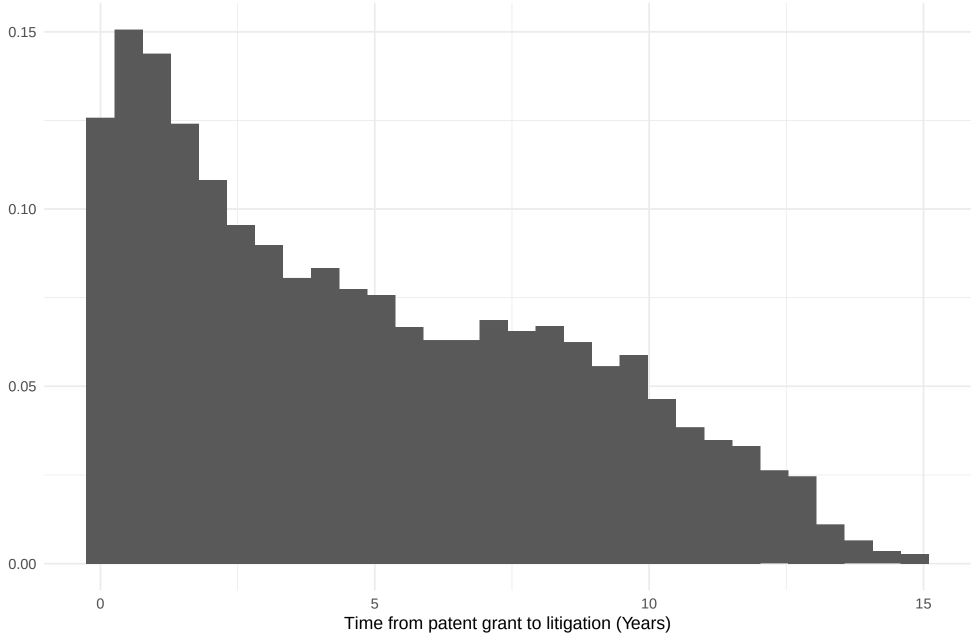
Note that our main aim is to build a prediction model. Therefore, we only use information that is known at the time of the grant and leave out any measure that can be generated from the citations received, ownership transfers and collateral pledges that materialize at a later stage.

We restrict ourself to patents granted between 2002 and 2005. We do so for two reasons. First, publication of patent applications became only mandatory form 2001 onwards in the US. We rely on information from the applications and, therefore, do not consider earlier applications. Second, the patents can be litigated during its entre lifespan, so we need to restrict ourselves to training our prediction models on cases for which the outcome is known. Figure 1 shows the distribution of the time to litigation in years. In total, we use more than 600.000 patents for training and testing the prediction models. We use 27 characteristics derived from different data sets provided by the United States Patent and Trademark Office (USPTO) and the Organisation for Economic Co-operation and Development (OECD). See Table 1 for an overview of the variables and summary statistics.

Our data includes a wide range of information on the patents. We sort the variables into five blocks. Block 1 includes the basic information about the technology such as the number of claims, the technology field, the patent scope (measured by the number of IPC classifications) and the citations to the non-patent literature and other patents, including self-citations. This block includes the typical measures for the value of a patent (e.g., Harhoff et al., 2003).

Block 2 includes information about the internationality of a patent - whether the application was processed via the PCT route, whether the application claims a foreign priority, and the family size. This is another set of information that is supposed to correlate with the value of a patent. Filing an application at multiple offices is costly, and shows that the owner value the protection at least above these costs (Harhoff et al., 2003).

Figure 1: Time to litigation



Notes: The time from grant to litigation is calculated by the filing year of the complaint minus the grant year of the patent. In the rare event that a patent is added to an ongoing lawsuit with a filing year before the grant year of the patent, we set the time to litigation zero.

Block 3 consists of information on the owner of the patent - whether the inventors assigned the patent rights to someone else, whether the owner(s) is an individual, corporation or government institution, and whether the owner(s) is foreign, Japanese or from the US. We also identified Fortune 500 companies.Lanjouw and Schankerman (2004) argue that the origin may affect the litigation likelihood because it is more costly to detect and prosecute infringements in the United States for foreign companies. The authors also argue that individuals and small firms are handicapped because they cannot benefit from repeated interaction to resolve a dispute and lack large patent portfolios for trading. Furthermore, they may lack the resources for accessing high-quality legal advice (Ronspies, 2004; Haus and Juranek, 2018; Juranek, 2018).

We include information about the examination of the application in Block 4, the duration of the examination (grant lag), the examiner's experience, and it average grant rates among applications with the same filing year. In their task to examine patent applications, the examiners have some level of discretion. As a consequence, different levels of leniency can be observed (see, e.g., Sampat and Williams, 2019) potentially affecting the

Table 1: Summary statistics for the variables.

Variable	Block	Observations	Mean	Std. Dev	Min	Max
No of claims	1	642977	18.2	15.2	1	887.0
Patent scope	1	642977	1.9	1.2	1	30.0
Backward citations	1	641956	13.5	22.4	0	799.0
BWD self-citations	1	642977	0.9	3.1	0	204.0
NPL citations	1	642977	3.1	10.3	0	199.0
NBER tech fields	1	642977				
Originality	1	638765	0.7	0.2	0	1.0
Family size	2	642977	4.1	4.2	1	57.0
Foreign priority	2	642977	0.4	0.5	0	1.0
PCT	2	642977	0.1	0.3	0	1.0
Assignee	3	642968	0.9	0.3	0	1.0
Number of assignees	3	642977	0.9	0.4	0	14.0
Assignee experience	3	642968	4223.5	7720.0	0	41723.0
Share individual	3	642968	0.1	0.3	0	1.0
Share government	3	642968	0.0	0.1	0	1.0
Share foreign	3	642968	0.3	0.4	0	1.0
Share japan	3	642968	0.2	0.4	0	1.0
Share fortune 500	3	642968	0.2	0.4	0	1.0
Lawyer	3	642977	0.9	0.3	0	1.0
Small entity	4	640560	0.2	0.4	0	1.0
Avg grant rate	4	642872	0.8	0.1	0	1.0
Examiner experience	4	642872	9.7	7.9	0	71.0
Grant lag	4	642977	917.6	437.2	14	10331.0
No of indep claims	5	642897	3.1	2.6	0	130.0
No of dep claims	5	642897	15.1	14.0	0	886.0
Avg word count dep claims	5	622394	37.1	22.2	6	3217.6
Avg word count indep claims	5	642805	155.9	99.8	1	7056.0

litigation likelihood of a patent (Cockburn et al., 2002). Even though one would expect ability to matter and, hence, that a more experienced examiner makes better decisions, Lemley and Sampat (2012) show that more experienced examiners are more likely to grant patents. The reason behind this observation are the incentives to get promoted and the fact that promotions carry with them reductions in examination time allocations (Frakes and Wasserman, 2017). We also add whether an applicant qualifies for the small entity status[6] in this block as this piece of information comes from the same data source. Unfortunately, the information is not available for all patents.

We further differentiate the number of claims into independent and dependent claims in Block 5, and the average number of words per claim.

In total, our data includes 642977 patents, out of which 7225 patents are litigated until 2016. Unfortunately, as shown by Table 1 some of the characteristics are not available for all patents. Table 4 presents an overview of all variables, their definition and source. We have chosen to use the maximum number of observations when analyzing each block or combination of blocks of explanatory variables. This means that the sample varies slightly across the different analyses, depending on the availability of data.

3.2. Methodology

There is a profound difference between *explaining* past observations in a response variable Y by means of a set of explanatory variables X, and *predicting* a not yet seen outcome of Y using a set of predictor variables (see Shmueli, 2010, for a detailed discussion). In the first case we typically set up a regression model that describes the (often linear or log-linear) relationship between the explanatory and response variable, while taking great care to build a model that allows us to identify, isolate certain effects and give specific interpretations to the estimated regression coefficients, which are in turn subject to classical statistical inference such as null hypothesis significance testing.

[6] Applicants with small entity status enjoy a discount of 50 percent for the fees of the USPTO. Small entities are defined as individual inventor applicants, non-profit organizations, or organizations with a "small business concern" with fewer than 500 employees (see the Manual of Patent Examining Procedure 509.02).

The success of a prediction model depends on its ability to predict the outcome of Y for cases that were not used to estimate - or train - the model. When directing our focus towards predictive power, we are less interested in the interpretability of the model or the estimation and isolation of certain, possibly causal, effects. This means that we have much greater freedom to choose the statistical method as we like to construct predictions, as long as they hit the target. Hastie et al. (2009) is the classic reference that contains an introduction and a great number of further references to various methods for *statistical learning*, or *machine learning*, which is a term more widely used among practitioners.[7]

The ability to learn nonlinear patterns and higher order interactions directly from a set of training data does not come for free, but carries the increased possibility of *over*fitting the model, meaning that we to a larger extent than for instance when using ordinary linear regression run the risk of interpreting random noise as systematic patterns. Machine learning methods face the bias-variance trade-off and must therefore be finely tuned in order to avoid such problems, and this task typically gets more complicated for complex prediction methods. In our approach, we randomly split our data into training and test data corresponding respectively to 70% and 30% fractions of the data, meaning that we estimate the models with 70 percent of data and then evaulate their performance on the remaining 30 percent.

Our baseline prediction method is the logistic regression model. The target variable Y, the outcome of which we want to predict, is binary – a patent is either litigated or not. The logistic regression is the classical method for either explaining or predicting a binary outcome by means of a set of covariates or predictors (see, e.g., Lanjouw and Schankerman, 2004). Denote by $Y = 1$ the event that a patent is litigated in the course of its lifetime, and by $Y = 0$ that it is not litigated. Let $X = (X_1, \ldots, X_p)$ be the set of explanatory variables. The logistic regression model defines the following log-linear link

[7] These two terms are in general not equivalent, but the distinction between statistical learning and machine learning does not have any practical interest to us.

between the conditional probability of litigation given X, and the observed values of X:

$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p. \tag{1}$$

We estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ from a training data set using maximum likelihood. We then use the estimated relationship (1) to predict the probability of litigation for all cases in the test data, $\widehat{P}(Y=1|X=x)$. We convert the predicted probability to a predicted *outcome* $\widehat{Y}$ by using a simple cutoff rule

$$\widehat{Y}=\begin{cases}1 \text{ if } \widehat{P}(Y=1|X=x)>c\\ 0 \text{ if } \widehat{P}(Y=1|X=x)\leq c,\end{cases} \tag{2}$$

for some number $0\leq c\leq 1$.

There are at least two important aspects to keep in mind when using the logistic regression as a prediction model for patent lawsuits. The first is that all patent characteristics used for prediction must be observed at the time when the prediction takes place.[8] Second, patent litigation is a relatively rare event, as only 1.1 percent of the patents in our data set eventually gets litigated. Such class imbalance can lead to poorer predictive performance for prediction methods, especially when the sample size is small. Chawla et al. (2002) employ a matching technique in order to balance classes in the data set, and there exist methods for automatic up- and downsampling of data sets for the particular purpose of improving predictive power. We have a large data set, however, and we do not experience any improvement in our results when using such techniques. Hence, we estimate the models in this paper directly on the raw training data.

3.3. Prediction accuracy

We follow the standard methods for classifications, and evaluate the prediction accuracy according to the receiver operating characteristics (ROC) curve, and the area under the curve (AUC). Every classification faces the problem of categorizing litigated patents

[8] That implies that we can not use the number of forward citations as a prediction variable at the grant date of a patent, even though it may be highly significant and carry a lot of predictive power (e.g., Lanjouw and Schankerman, 2004).

as non-litigated, false negatives (FN), and categorizing non-litigated patents as litigated, false positives (FP). The numbers on false negatives and false positives together with the true positives (TP) and true negatives (TN) allows calculating the true positive rate $TP/(TP+FN)$ and the false positive rate $FP/(FP+TN)$. The ROC curve plots the true positive rate on the y-axis against the false positive rate on the x-axis for every cut-off point $0 \leq c \leq 1$ (See Eq. (2)).

The area under the ROC curve can then be interpreted as a performance measure for the prediction accuracy as we aim to achieve a high true positive rate while keeping the false positive rate low. Whereas random guessing corresponds to AUC being equal to 0.5, a perfect classification at all classification thresholds corresponds to AUC equal to 1.

There is an inherent trade-off between the true and false positive rate. Clearly, if one classifies all observations as positive the true positive rate is maximized. Of course, such a model is of little value because the false positive rate is also maximized. In a classification exercise we must therfore balance these two rates against each other when choosing the cut-off value c in a practical problem. However, for applications that do not require a binary classification for each test sample, we will use our prediction models to rather assign litigation probabilities to the patents, conditioned on the values of the predictors.

4. Results

4.1. Patent characteristics:

In the first step, we analyze the role of the predictor variables. In order to do so, we rely on the logistic regression. Table 2 summarizes our results for the different blocks and methods. Focusing on the logistic regression first, we observe that the first 5 blocks all have some predictive power, as the AUC exceeds 0.6 for all of them. Out of the five blocks, the characteristics of the patent owner have the highest predictive power, leading to the highest AUC.

We further observe that the informational value of the different blocks complement each another. Including additional blocks improves the predictive power until block 4, leading to an AUC of around 0.79. Including also block 5 adds only marginal predictive

Table 2: The AUC for the logistic regression.

Variable block	Single block	Cumulative addition of blocks
1	0.673	
2	0.689	0.734
3	0.749	0.786
4	0.601	0.793
5	0.638	0.793

Notes: Area under the curve (AUC) for the five blocks individually and cumulatively using a logistic regression.

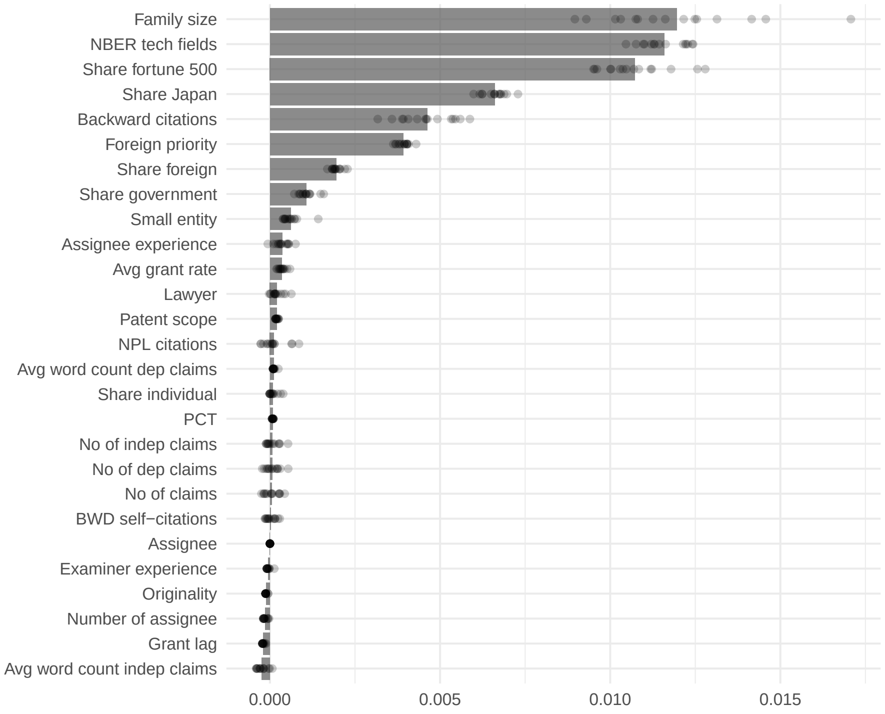
power (on the fourth digit after the comma).

The results for the different blocks already indicate that the variables differ in terms of their explanatory power. We investigate these differences further by relying on a relatively novel approach proposed by Fisher et al. (2019) to measure the variable importance. We measure the contribution from each predictor by fitting the model again on a data set where this particular predictor is randomly permuted, i.e., the elements of the vector are randomly reordered. The permutation removes the contribution of the variable towards predicting the litigation status. We then measure *how much is the AUC reduced by removing the predictive power of that specific variable.* Removing the predictive power of important variables leads to a greater loss in prediction accuracy than removing less important variables. In order to remove the effect of chance due to the random perturbation, we average the results from 15 random perturbations.

Figure 2 presents the result of this approach for the logistic regression model using all five blocks. It is important to note that the variable importance itself does not give an indication of the direction of the influence of a certain variable.

It turns out that the family size of a patent, i.e., the number of jurisdictions at which the underlying technology is protected, has the highest predictive power. It is followed by the technology field dummies showing that the litigation likelihood differs strongly between technology fields. Furthermore, we observe that the assignee characteristics play an important role; six of the assignee characteristics are in the top 10 variables. The most important assignee characteristic seems to be whether the patent's owners are Fortune 500

Figure 2: Variable importance in logistic regression



Notes: The variable importance for the logistic regression model measured using the permutation approach by Fisher et al. (2019). Each dot represents the reduction in AUC after randomly permuting the observations in that variable. The bars represent the average reduction in AUC after 15 such random permutations. Some negative values due to sampling variation are to be expected for variables that are not important.

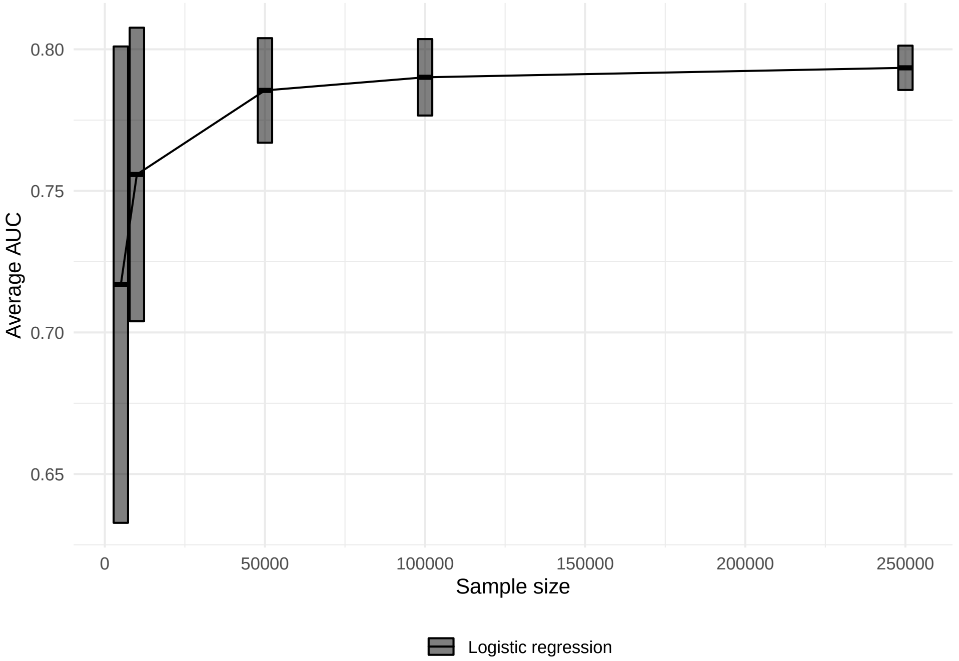
companies. Furthermore, it turns out that patents owned by Japanese companies differ from the other patents. Finally, it is noteworthy that the number of backward citations to previous patents has substantial predictive value.

In sum, our results in this section show that that there is significant predictive power in the characteristics of a patent and its assignees.

4.2. Sampling

In the next step, we analyze the importance of the size of the training data set for the prediction performance. Earlier contributions on patent litigation prediction e.g., Chien (2011), or inference studies (e.g., Lanjouw and Schankerman (2004) relied on fairly small samples. The reason was presumably a lack of data availability. However, due to the increased data transparency at the USPTO, we can rely on the population of patents today.

Figure 3: AUC distribution for different sample sizes



Notes: Average AUC for the logistic regression for 100 random subsamples of variuous sample sizes. The vertical bars represent one standard deviation of the observed AUC on the test data.

The population allows us to analyze how reliable results from samples are and how much the performance of a prediction depends on the data availability. Therefore, we use the population and draw 100 random samples each of differing sizes: 5000, 10000, 50000 and 1000000, and perform the prediction with the logistic regression model using all five blocks. We record the AUC of each prediction and then create a distribution of AUC for each sample size. Figure 3 summarizes the outcome by reporting the mean AUC plus/minus one standard deviation

We clearly observe better average prediction performance with larger samples, and this is to be expected. It is more important, however, to note that the variation of outcomes decreases for larger samples. Whereas the improvements from samples of size 10000 to samples of size 50000 is very large, increasing the sample size further leads only to marginal additional improvements. This shows that small sample studies should be interpreted cautiously towards predicting future lawsuits. However, even though the population will give the best results, samples of intermediate size will already give reliable results.

4.3. Method choice

Finally, we analyze whether we can improve the prediction quality by moving from the simple logistic regression to more sophisticated statistical learning methods. We focus on *elastic net*, *random forest* and *xgBoost* models, which are used regularly for purposes of machine learning. We shortly introduce each method before presenting the results.

Elastic net: An elastic net aims to increase the predictive power of the logistic regression by adjusting the coefficients. The elastic net method combines two common methods, the lasso and ridge regression. Lasso and ridge regression are based on the same principle: By reducing the magnitude of the coefficient estimates in logistic regression, we lessen the impact of the patent characteristics, resulting in a reduction of the variance of the predicted outcome. In machine learning terminology this is the same as reducing the complexity of the model, which *may* improve predictive power.

Technically, lasso and ridge add each a shrinkage penalty λP to the maximum likelihood model:

$$\left(\widehat{\beta}_0, \ldots, \widehat{\beta}_p\right) = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \left\{ -\sum_{i=1}^{n} \log \mathrm{P}(Y_i = y_i) \right\} + \lambda P, \tag{3}$$

where P is equal to $\sum_{j=0}^{p} |\beta_j|$ for lasso and $\sum_{j=0}^{p} \beta_j^2$ for ridge. λ is a tuning parameter that determines the amount of parameter shrinkage. We estiamte the optimal value of λ in terms of predictive performance by cross validation.

The ridge and the lasso have different statistical properties, and none of them performs better than the other as a general rule. Therefore, we employ a weighted average of the two models, i.e., an *elastic net*. An elastic net includes *both* of the penalty terms in the minimization problem, weighted together with a parameter $\alpha \in [0, 1]$. The parameter α is a tuning parameter, and we choose the combination of λ and α that optimizes predictive power, measured by cross validation.

Random forest: We investigate another line of machine learning methods based on the concept of a classification tree. This approach is more algorithmic in nature, as opposed to the logistic regression that is based on building a probability model for the data generating process. We grow a classification tree for the training data by constructing subsequent partitions that separate patents into “litigated” and not “litigated” using measures of dissimilarity.

A crucial decision when growing a classification tree is when to stop making new splits. In order to avoid overfitting, it is important to identify an optimal threshold, a stopping rule, for which the partitions in the training sample are dissimilar but not too small. We determine this threshold by cross validation.

Because classification trees are purely nonparamteric, they tend to have little bias but high variance. The balance between bias and variance can be achieved by bootstrapping. The *random forest* (Breiman, 2001) is an algorithm that grows many trees, i.e., a forest, based on bootstrapped replications of the training data, and then classify the outcome of a test case based on a majority vote among all such replicated trees.

We implement a random forest with $B = 100$ replicated trees.

xgBoost: Another strategy to improve the predictive power of a simple, and possibly weak, classification rule such as the single classification tree, is *boosting*. The basic idea is to first train a classification tree on the raw data, and then repeat the procedure with a weighted version of the data set in which the observations that were mis-classified in the first step get higher weight in the second step. This procedure is then repeated a number of times, forcing the sequence of classification rules to pay more and more attention to the most difficult data points.

We implement a recent and very computationally effective implementation of this idea, the XGBoost algorithm, that in addition to the basic boosting algorithm takes advantage of sparse matrix computations and recent advances in split-finding algorithms for constructing classification trees. The XGBoost has proven to be one of the most successful machine learning techniques in online prediction contests, for instance on Kaggle (Chen and Guestrin, 2016).

In order to simplify the discussion of our results, we focus on models using all five blocks of variables. Table 3 compares the AUC for the different models. The table shows that the random forest and xgBoost models outperform the logistic regression, and achieve

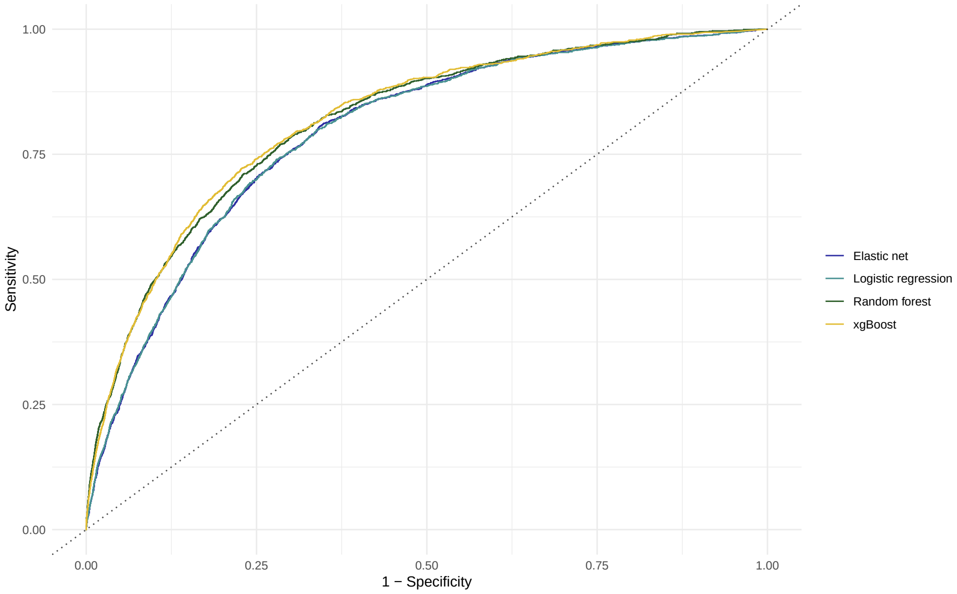
Table 3: The AUC for the different models using blocks 1 to 5

Logistic regression	Elstic net	Random forest	xgBoost
0.793	0.793	0.815	0.818

Notes: Area under the curve (AUC) for the four prediction methods on the full data set containing all the blocks.

better prediction results. However, the improvements are rather marginal compared with the benefits of including more patent characteristics as well as increasing the sample size used for training, as seen earlier in this section.

Figure 4: ROC across different methods



Notes: Receiver Operating Curves (ROC) for the various machine learning methods on the full data set containing all the blocks. The curves that are closer to the upper left corner of the plot are better. We see that the tree-based methods have better predictive power than the regression methods.

This observation is confirmed by Figure 4 that presents the ROC for all four models. The figure shows that we can achieve quite an impressive result; with the xgboost model we can achieve a true positive rate of around 0.75 at a false positive rate of only around 0.25.

5. Discussion

Our results show that all dimensions matter, the number of patent characteristics, the size of the training data and the methods chosen. We achieve the best result by including all characteristics and all observations in the training data, and by using an xgBoost model.

In practice, however, maxing out on all dimensions may be computationally too costly. In that case, our results provide guidance. We show that the strongest improvements can be achieved by extending the number of predictors, i.e., patent characteristics. The performance also increases with the size of the training data. Clearly, more training data improves the performance. However, it turns out that for training data of intermediate size, the prediction becomes already sufficiently stable. Finally, the more complex xgBoost model outperforms the logistic regression. However, relative to the improvements in the other two dimensions the increase in prediction performance appears to be minor. Hence, the model choice depends on the available (computational) resources.

The trade-off between more patent characteristics and the size of the training data matters particularly in practical terms. A number of patent characteristics are costly to derive, for example, network or text analysis (e.g., Arts et al., 2018). Against the background of our results, it seems to be reasonable to focusing on acquiring more characteristics for a smaller sample rather than using the population with a smaller set of predictors.

In general, there are a number of potential variables that could improve the analysis further. That could be, for example, more details from the application process (Haeussler et al., 2013) but also characteristics acquired after the grant such as citations or ownership changes. However, the choice of variables depend on the choice of model. A model used at patent offices in order to identify patents that require particular scrutiny has to rely only on the available characteristics. In the case of litigation insurance, the post-grant information may be valuable and lead to a dynamic pricing. These two examples show that a model has to be specified for its desired purpose.

In terms of the sample size, our results show that one should be careful about predictions formed from small samples. Whereas small samples can give reliable (confidence intervals) of estimates of the influence of a particular variable, the prediction performance will likely be inaccurate. As a consequence, models from inference studies (e.g., Lanjouw and Schankerman, 2001, 2004) should not be directly used for prediction, and prediction models using small samples should be interpreted cautiously.

Our results also show that the usefulness of a prediction model depends on the application. The prediction can be used in two ways. First, the model can be used as a classification of patents. However, it is important to bear in mind that patent litigation is a rare event, as only around 1.1 percent of the patents are litigated. That implies that even with the impressive performance of our most complete model, the majority of patents that are classified as "litigated" will actually not be litigated. With a 1.1 percent litigation probability, our model will deliver 0.75 percent true positives $(0.01 \cdot 0.75)$ but 24.75 percent false positives $(0.99 \cdot 0.25)$. Hence, in that case, only 2.9 percent of the positives will be true positives. Even though this is a strong increase from the unconditional probability of 1.1 percent, it shows that one has to be cautiously interpreting the classification. However, the classification can nonetheless be of value. It can be implemented as a first step before more labor intense analyses take over, as the attention can be focused then only on 25.5 percent of patents that turned out to be positive.

Second, the model allows deriving a litigation probability for a given patent. Assigning probabilities is particularly valuable for the pricing of risk. The obvious example is the case of litigation insurance. Other applications exist as well; for example, personnel planning in in-house legal departments of large corporations.

6. Conclusion

We present a variety of models that allows predicting whether a patent will be litigated. By doing so, we discuss the impact of three dimensions, the size of the training data, the number of patent characteristics as predictors, and the model type. We derive practical advice that shows that whereas the small training data leads to a wide variation of the prediction quality, data of intermediate size leads to reasonably stable results. Furthermore, increasing the model complexity further improves the predictive performance.

References

Allison, J.R., Lemley, M.A., Schwartz, D.L., 2014. Understanding the realities of modern patent litigation. Texas Law Review 92, 1769–1801.

American Intellectual Property Law Association, 2015. 2015 Report of the Economic Survey. Technical Report.

Arora, A., Ceccagnoli, M., 2006. Patent protection, complementary assets, and firms' incentives for technology licensing. Management Science 52, 293–308.

Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strategic Management Journal 39, 62–84.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Campbell, W., Li, L., Dagli, C., Greenfield, K., Wolf, E., Campbell, J., 2016. Predicting and analyzing factors in patent litigation, in: NIPS2016, ML and the Law Workshop.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

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.

Chien, C.V., 2011. Predicting patent litigation. Texas Law Review 90, 283.

Cockburn, I.M., Kortum, S., Stern, S., 2002. Are all patent examiners equal? The impact of examiner characteristics. Technical Report. National Bureau of Economic Research.

Cremers, K., 2004. Determinants of patent litigation in germany .

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, 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, 550–563.

Ganglmair, B., Helmers, C., Love, B., 2018. The effect of patent litigation insurance: Theory and evidence from npes. unpublished working paper .

Haeussler, C., Harhoff, D., Mueller, E., 2013. Signaling and certification-the role of patents for venture capital-financing, in: Academy of Management Proceedings, Academy of Management Briarcliff Manor, NY 10510. p. 12206.

Hall, B.H., 2005. Exploring the patent explosion. The Journal of Technology Transfer 30, 35–48.

Harhoff, D., Scherer, F.M., Vopel, K., 2003. Citations, family size, opposition and the value of patent rights. Research Policy 32, 1343–1363.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Haus, A., Juranek, S., 2018. Non-practicing entities: Enforcement specialists? International Review of Law and Economics 53, 38–49.

Henry, M.D., Turner, J.L., 2016. Across five eras: Patent validity and infringement rates in us courts, 1929–2006. Journal of Empirical Legal Studies 13, 454–486.

Juranek, S., 2018. Investing in legal advice. Information Economics and Policy 44, 28–46.

Kortum, S., Lerner, J., 1999. What is behind the recent surge in patenting? Research policy 28, 1–22.

Lanjouw, J., Schankerman, M., 2001. Characteristics of patent litigation: A window on competition. The RAND Journal of Economics 32, 129–151.

Lanjouw, J., Schankerman, M., 2004. Protecting intellectual property rights: Are small firms handicapped? The Journal of Law and Economics 47, 45–74.

Lemley, M.A., Sampat, B., 2012. Examiner characteristics and patent office outcomes. Review of Economics and Statistics 94, 817–827.

Lemley, M.A., Shapiro, C., 2005. Probabilistic patents. The Journal of Economic Perspectives 19, 75–98.

Marco, A.C., Miller, R.D., 2019. Patent examination quality and litigation: Is there a link? International Journal of the economics of business 26, 65–91.

Ronspies, J.A., 2004. Does david need a new sling? small entities face a costly barrier to patent protection. The John Marshall Review of Intellectual Properrty Law 4, 184—-211.

Sampat, B., Williams, H.L., 2019. How do patents affect follow-on innovation? evidence from the human genome. American Economic Review 109, 203–36.

Shapiro, C., 2016. Patent remedies. American Economic Review 106, 198–202.

Shmueli, G., 2010. To explain or to predict? Statistical Science 25, 289–310.

Trajtenberg, M., Henderson, R., Jaffe, A., 1997. University versus corporate patents: A window on the basicness of invention. Economics of Innovation and New Technology 5, 19–50.

Wongchaisuwat, P., Klabjan, D., McGinnis, J.O., 2017. Predicting litigation likelihood and time to litigation for patents, in: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, pp. 257–260.

Block	Variable	Definition	Source
1	**No of claims**	Number of claims stated in a patent	Patentsview
1	**Patent scope**	Number of 4 digit IPC classes assigned to a patent	Patentsview
1	**Backward citations**	Number of patents that are cited by the asserted patent, including self-citations	Patentsview
1	**BWD self-citations**	Number of citations to patents with the same applicant	Patentsview
1	**NBER tech fields**	NBER technology categorization	Patentsview
1	**NPL citations**	Number of citations to non-patent literature	OECD patent quality indicators
1	**Originality**	The ”Originality” measure proposed by Trajtenberg et al. (1997). It measures the range of technology classes of patents that are cited by the asserted patent. See the Appendix **??** for the definition.	OECD patent quality indicators
2	**Family size**	The number of patent offices at which a given invention has been protected	OECD patent quality indicators
2	**PCT**	Dummy variable equal to 1 if the application is filed via the PCT route	Patentsview
2	**Foreign priority**	Dummy variable equal to 1 if the application claims foreign priority	Patentsview
3	**Assignee**	Dummy variable equal to 1 if the patent is assigned to an assignee	Patentsview
3	**Number of assignees**	Number of assignees to which the patent is assigned to	Patentsview
3	**Assignee experience**	Number of granted patents assigned to an assignee before (the maximum if multiple assignees)	Patentsview
3	**Share individual**	Share of assignees that are individuals	Patentsview
3	**Share government**	Share of assignees that are government institutions	Patentsview
3	**Share foreign**	Share of non-US/non-JP assignees (individuals if patent has no assignee) that are from abroad	Patentsview
3	**Share japan**	Share of Japanese assignees (individuals if patent has no assignee) that are from abroad	Patentsview
3	**Share fortune 500**	Share of assignees that are part of the Fortune 500 list	own calculations
3	**Lawyer**	Dummy variable equal to 1 if the application was processed by a lawyer	Patentsview
4	**Small entity**	Dummy variable equal to 1 if the applicant has small entity status	USPTO examination dataset
4	**Avg grant rate**	The examiners average grant rate for applications with the same filing year (excluding pending applications)	USPTO examination dataset
4	**Examiner experience**	The examiner’s experience in years at the filing date of the application	USPTO examination dataset
4	**Grant lag**	Time between filing of the application and its grant.	USPTO examination dataset
5	**No of indep claims**	Number of independent claims stated in a patent	USPTO patent claims dataset
5	**No of dep claims**	Number of dependent claims stated in a patent	USPTO patent claims dataset
5	**Avg word count dep claims**	Average number of words in the dependent claims of a patent	USPTO patent claims dataset
5	**Avg word count indep claims**	Average number of words in the independent claims of a patent	USPTO patent claims dataset

Table 4: Patent characteristics

NHH

NORGES HANDELSHØYSKOLE
Norwegian School of Economics

Helleveien 30
NO-5045 Bergen
Norway

T +47 55 95 90 00
E nhh.postmottak@nhh.no
W www.nhh.no