PATENT LITIGATION RISK-SCORING MODEL

Student team: Alex D. Kim, Nicholas W. Partee, Teddy J. Reynolds, Michael A. Santamaria

Faculty Advisor: Peter A. Beling
Department of Systems Engineering

Client Advisors: David V. Ferron, David S. Winer, Douglas F. Woolley
M·CAM, Inc.
207 5th Street, N.E.
Charlottesville, VA 22209
E-mail dsw@m-cam.com


ABSTRACT

We develop a model to predict the probability of a patent being involved in litigation in the near future. Litigation is defined as a legal filing, either offensive or defensive in nature. The model will be used by insurance companies to better evaluate potential clients for intellectual property insurance with a coverage time of two years.

The data for the model are from a random sample of all patent case filings from 1994 through 2000. We added a control group of patents not involved in litigation that are similar to the litigated patents. We gather potentially predictive variables about the patent and its owner to build the data set for regression. A logistic regression model is built in SAS using stepwise regression.

Two models are considered, one with forward citations and one without. The first model finds forward citations and sales from the filing year and two years prior to be significant. In the second model, backward citations, sales from the filing year, and number of employees one year prior to filing are significant. The model can not clearly separate patents into litigated and non-litigated categories, but the model can identify patents with extremely high and low likelihoods of being involved in litigation if very high and low cutoff points are selected.

INTRODUCTION

Intellectual property is arguably the basis of our modern economy. Insurance companies offer intellectual property insurance to protect companies from the high costs associated with patent infringement litigation. Currently, insurance companies have had limited success in this market. The goal of this project is to develop a model that predicts the likelihood of a patent being involved in litigation. An insurance company would use the model to help evaluate potential clients for intellectual property insurance.

DATA ACQUISITION

The data set is crucial in constructing the predictive model. We obtain a random sample of patents involved in litigation from the years 1994 to 2000. We also gather a group of similar non-litigated patents to act as a control group. For each patent, we collect the patent number, patent issue date, patent owner, number of claims, number of backwards citations, number of forwards citations, case filing date, patent age at case filing, patent owner’s net sales, patent owner’s net income, patent owner’s number of employees, and whether the patent is involved in litigation (yes or no). We choose these variables because they are potentially predictive in determining whether a patent will face litigation in the near future.

The initial list of patents comes from a database of all court filings from the mid 1980s through 2000, available through the Inter-University Consortium for Political and Social Research (ICPSR) website. ICPSR provides access to an archive of social science data for research. Next, we find the specific patent involved in a case filing and several of the patent-related data points through Derwent’s LitAlert online database. Derwent is a company that specializes in providing patent information. LitAlert allows us to match a non-litigated patent with a similar international classification code and issue date as a litigated patent from our data set. We collect similar non-litigated and litigated patents so the data points are more comparable.

Next, we gather any available financial data for the patent’s owners at three distinct points in time: the year of filing, one year prior to filing, and two years prior to filing. We determine if these financial values can predict litigation. Therefore, our model predicts
whether a specific patent will be involved in litigation in the next two years. Not all data points are available for many of the patents because private companies are not required to report their financial information. We obtain data for public companies through Hoover’s Online, and data for private companies from Ward’s Business Directory.

**MODEL DESIGN**

The next step is to determine the predictive value of the independent variables. We use logistic regression to develop a model that best represents our data in predicting the probability of litigation. The parameters for a logistic regression are chosen by the maximum likelihood criterion, or chosen such that the probability of the observations in the model is maximized. The general equation for a logistic regression is as follows:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$\pi$ is the probability that a patent will be litigated

$\beta_k$ is the coefficient of regression for the $k$th independent variable

$x_1, x_2, \ldots, x_k$ are quantitative or qualitative independent variables

We use a significance level of 0.10 throughout the modeling process. All modeling is done with SAS and S-plus.

Only the patents where all variables could be obtained are used in our predictive models. We develop two models, one with forward citations and one without forward citations. Ideally, we would like to collect the number of forward citations for a patent at the time of filing. We could not find this particular data, so our forward citations data represent the total number of forward citations as of 2001. Therefore, our forward citations data may be overstated. To gauge the effect of this inaccuracy, we develop a model without forward citations. We believe forward citations are a significant predictor, so we also create a model with forward citations to approximate the predictive nature of forward citations.

In both models, all financial variables are represented using categorical variables because of the large range of the data. Companies are sorted into small, medium, large, and very large bins for their sales and number of employees. For both models, we use stepwise regression to eliminate less meaningful variables.

We analyze the results of the stepwise regression to see whether the variables chosen by the stepwise algorithm in SAS are significant. We check for multicollinearity by analyzing the correlation matrix of predictors and the signs of the parameters. If predictors have correlation greater than 0.7, it is likely that multicollinearity is present. If all the signs of the parameters are logical, it is likely multicollinearity is not present.

**RESULTS**

**Model 1**

Model 1 uses forward citations. The model finds forward citations, sales figures for the filing year, and sales figures two years prior to filing to be meaningful in predicting patent litigation. SAS chose these three predictors at a significance level of 0.10 with the stepwise logistic regression algorithm.

We determine the effectiveness of the model by analyzing the predictions made by the model for the sample data. Density plots graph the predictions made for litigated and non-litigated patents. Separation between the plots indicates that the model is able to separate the two groups of patents. Ideally, there would be little or no overlap between the density curves. Figure 1 shows there is some separation between the density plots for litigated and non-litigated patents, especially between 0.6 and 1.0. The non-litigated patents are more heavily weighted towards zero, which is consistent with our initial assumptions.

![Density Plots for Model 1 (w/ Forward Citations)](image)

**Figure 1:** Density plots for the probabilities produced by Model 1 for both litigated and non-litigated patents.
Figure 2 shows sensitivity, specificity, and the percentage of correct predictions at various cutoff points for Model 1. A cutoff point is a set probability level that determines if the model considers a patent to be litigated or not litigated. The model, by definition, gives probabilities between 1 and 0. A probability above the cutoff point is predicted to be litigated and a probability below the cutoff point is predicted to be not litigated.

Sensitivity is the proportion of actual litigated patents that the model predicts will be litigated. Specificity is the proportion of actual non-litigated patents that the model predicts will be non-litigated. Percent correct simply represents the overall accuracy of the model at each cutoff point.

The graph shows that at high cutoff points, there is high specificity and low sensitivity. At a cutoff point of 0.66, the specificity is 94.6 percent. This means that 94.6 percent of all non-litigated patents produce probabilities below this cutoff value, and are accurately predicted. Approximately 34 percent of all litigated patents produce probabilities above 0.66 with this model. A patent with a predicted probability above 0.66 can therefore be considered a very risky patent to insure. At a lower cutoff point of 0.10, about 90 percent of all litigated patents are correctly predicted. The false positive rate is 61.7 percent, which represents rejecting patents which actually never face litigation. This represents a lost opportunity for insurance companies because they reject a safe patent. The false negative rate is 17.4 percent, meaning that the model predicts a litigated patent will be non-litigated. This will result in a financial loss for the insurance company.

Figure 3 is the Receiver Operating Characteristic (ROC) curve for Model 1. It shows the relationship between sensitivity and one minus the specificity. Sensitivity is the “true positive” rate. It represents the probability that the model predicts filing, given that filing occurs. One minus the specificity is the “false positive” rate. It represents the proportion of patents that our model predicts will go to filing, when in actuality they do not. As the cutoff point increases, both values on these axes increase. This is a tradeoff in predicting outcomes; as sensitivity increases, “false positive” errors increase. The greater the area between the ROC line and the 45 degree line, the better the overall predictive power of the model.

Figure 3: The Receiver Operating Characteristic Curve for Model 1. The ROC curve shows the relationship between the true positive rates (y-axis) and false positive rates (x-axis) as various cutoff points are considered.

Model 2

In this model, we do not include forward citations. The variables deemed significant at a 0.10 significance level are sales in the filing year, backward citations, and the number of employees one year prior to filing.
Figure 4: Density plots for the probabilities produced by Model 2 for both litigated and non-litigated patents.

Figure 4 shows the model without forward citations has significantly less separation between the predicted probabilities for litigated and non-litigated patents than in Model 1. There is still some separation, and the non-litigated patents are slightly skewed to the left.

Figure 5: Graphs of the sensitivity, specificity, and percent correct rates for varying cutoff points in Model 2.

Figure 5 shows sensitivity, specificity, and correct percentage rates for Model 2. These values are similar to those found in Model 1. At a cutoff point of 0.50, the specificity is 94.6 percent. The model accurately predicts 94.6 percent of all non-litigated patents to be non-litigated. With a sensitivity of 38 percent, the model also correctly rejects 38 percent of all litigated patents. At a lower cutoff point of 0.24, around 96 percent of litigated patents are predicted correctly. The false positive rate is a lofty 60 percent with a 0.24 cutoff point. The more severe false negative rate is 8.7 percent.

The ROC curve for Model 2 is illustrated in Figure 6. The curve does not appear to be as successful as the one produced by Model 1. The true positive rates (x-axis) do not increase as drastically in the Model 2 ROC curve without sacrificing false positive rates (y-axis).

CONCLUSIONS

Summary

In this project, we created two models to predict the probability that a patent will face litigation in the next two years. We defined litigation as the point at which a case filing occurs. We gathered two groups of patents, litigated and non-litigated and gathered data on both of these patent groups. The data points included characteristics about the patent and its owners. We then used logistic regression to build a model from our final patent data set.

In Model 1, we included the number of forward citations. Forward citations and sales from the case filing year and two years prior are significant in Model 1. In Model 2, forward citations were not included. Backward citations, sales in the filing year, and number of employees one year prior to filing were deemed significant.

Interpretation

The model cannot clearly separate patents into litigated and non-litigated categories at a single cutoff point, but the model still retains some predictive power. The model can identify patents with extremely high and
low likelihoods of being involved in litigation if different cutoff points are selected. With Model 2, a cutoff point of 0.50 correctly predicts 94.6 percent of all non-litigated patents. Thus, probabilities produced for patents above 0.50 are seemingly very risky to insure. A low cutoff point of 0.24 correctly predicts 96 percent of all litigated patents. However, false positive rates would also need to be considered with such low cutoff points. Insurance companies need to weigh the relative costs of insuring a patent which faces litigation against rejecting a patent which actually turns out to be safe. Our model may be helpful in helping them obtain an optimal tolerance level.

Recommendations

The biggest area of improvement for developing a model would come through data acquisition. A single database with a complete listing of patents involved in case filings and their patent characteristics would cut down on errors during the data gathering process and hopefully lead to a better model.

The model is not accurate enough to be the only factor in determining whether a patent is a good candidate for intellectual property insurance. The process to evaluate potential clients still requires expert analysis, but the model could provide some assistance during the underwriting portion of the insurance process by identifying risky and safe patents.

REFERENCES


BIOGRAPHIES

Alex Kim is a fourth-year Systems Engineering major from Alexandria, VA, concentrating in Management Information Systems. His main contributions to the project were finding online databases that provided the group with hours of enjoyment. Alex has an employment offer to stay in Charlottesville next year, where his title will likely be that of a “professional townie”.

Nick Partee is a fourth-year Systems Engineering major from Great Falls, VA, concentrating in Management Information Systems. His main contribution to the project was transporting the team in a speedy fashion. While unfortunately the po-po caught him on one occasion, they let us off because we told them we were in a hurry to go “gather data for a school project.”

Teddy Reynolds is a fourth-year Systems Engineering major from Kingsport, TN, concentrating in Economics. His main contribution to the project was his excellent liaison skills between project team and the B-dog. He will be moving to New York where he hopes to have an impact on the city similar to that of John Rocker.

Michael Santamaria is a fourth-year Systems Engineering major from Houston, TX, concentrating in Economics. His main contribution to the project was keeping us up to date on useful and exciting events such as crew, bad jokes, and techno. His plan next year is to work in a Taiwanese nightclub dancing in a cage for pennies.
Patent Litigation Risk-Scoring Model