VISTA ANALYTICS

1333 H STREET, N.W. SUITE 600W
WASHINGTON, D.C. 20005
(202) 808-2286

Settlement Prediction

The potential of predicting a case outcome in the earliest phases of a litigation (prior to the discovery phase) can provide significant information to attorneys in deciding whether to pursue a defense or how to gain decisive advantage over their opponents in the earliest stages of a dispute. Additionally and potentially self-evident is the financial calculation that a corporation must undertake in the budgeting of the costs related to the litigation vs. the probable outcome of the case. Vista believes that machine learning can be used successfully in courtroom prediction with the resulting benefits.

In the setting of Supreme Court decision rulings, machine learning has predicted more accurate results than seasoned lawyers and law professionals' predictions. In patent litigation, machine learning has been used to predict the success of a patent and its financial impact. In this article, Vista demonstrates the use of machine learning to predict the outcome of a securities fraud class action. Existing settlement prediction analyses in securities fraud focuses primarily on loss causation and economic damage estimation. Those analyses' attempt to address what characteristics of litigation are associated with settlement outcomes. In contrast to the ex post diagnostic explanatory studies, Vista Analytics provides a prescriptive machine learning solution in settlement prediction that can be used for early forecasting.

In the following sections, Vista demonstrates prediction in both the likelihood of settlement and the dollar value of the settlement in securities fraud class action lawsuits. The solution can be applied as early as when a case is filed and therefore can aid corporates in formulating earliest stage strategy. Though securities fraud cases were used for initial research, Vista believes that this machine learning solution can be adapted and applied to most types of court room predictions where sufficient data is available, including cases involving IP, employment law, class action cases regardless of underlying cause and so forth.

Data

The primary data source for the analysis was the Institutional Shareholder Services (ISS) database, previously known as Riskmetrics, which provides information on the specific case and securities. The sample was limited to cases filed after 1995 to remove distorted effects from the initiation of the PLSRA. In addition, only independent variables known at the time of initial filing were used for early forecast. Five types of independent variables were examined in the analysis: 1) case-level variables, for example whether a case alleges Rule 10b-5 violation, 2) securities-level variables, for example, the market capitalization, 3) company-level variables, 4) industry-level variables, and 5) general economic and political trend variables.

VISTA ANALYTICS

1333 H STREET, N.W. SUITE 600W
WASHINGTON, D.C. 20005
(202) 808-2286

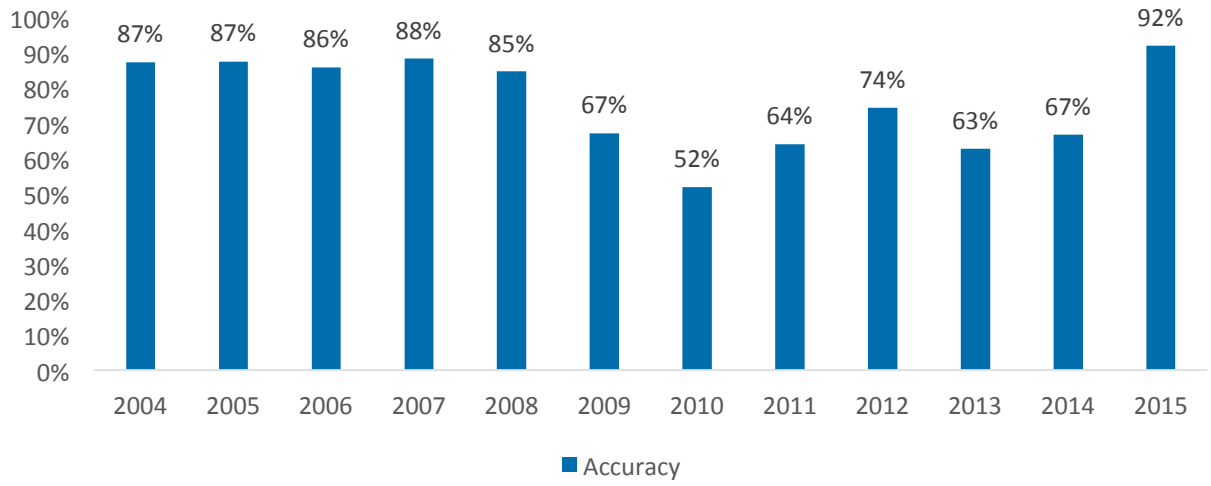
Predicting the Likelihood of Settlement

In predicting the likelihood of settlement, we identified the problem of concept drift, i.e. the underlying pattern in the data changed over time. Particularly, many independent variables became less indicative of the case status or even had the opposite correlation with the case status after the financial crisis compared to pre-crisis data. For example, 87 percent of IPO cases settled before 2009, whereas this number was 57 percent after 2009.

If the same static model is used to predict future events, the model becomes less and less relevant and increasingly loses accuracy over time. In response to the evolving pattern in the data, Vista treated the settlement events as streaming data with filing date as the time indicator. We then implemented an incremental learning algorithm to continuously revise and refine the classification model by incorporating new data as it arrives. To improve the model performance further, we built an meta classifier from multiple individual incremental learning classifiers. The result is an ensemble classifier, where the collective decision from all classifiers corrects the error made by a single classifier, thus resulting in better overall accuracy.

To validate our model, we predict the likelihood of settlement in the current period (test) based on a model partially trained from the previous period (train). Figure 1 shows the out-of-sample accuracy over all testing year. The accuracy before 2008 was stable and consistent around 87 percent. The average accuracy across all years was around 80 percent.

Figure 1: Out-of-Sample Accuracy



VISTA ANALYTICS

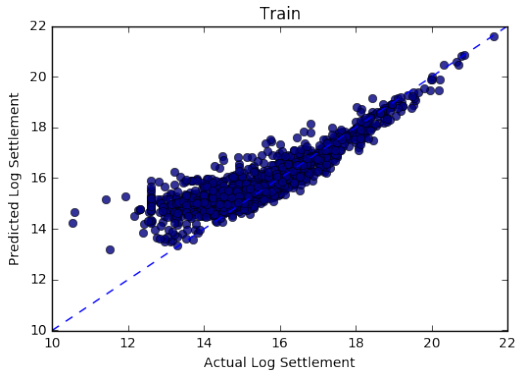
1333 H STREET, N.W. SUITE 600W
WASHINGTON, D.C. 20005
(202) 808-2286

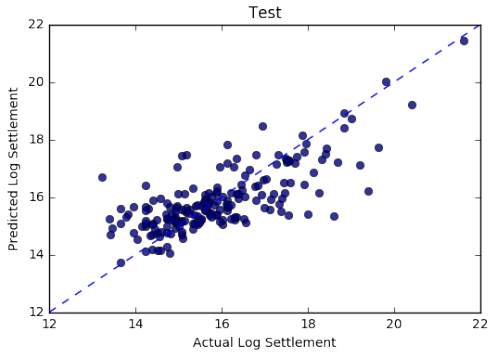
Predicting Settlement Amounts

In predicting settlement amounts, we split the settled cases by filing date. The training set was comprised of settled cases that were filed during 2008 and before. And the testing set was constructed with settled cases that were filed after 2008. The final model is an ensemble model of multiple individual regressors.

Total dollar amount settled is a highly-skewed variable where a small percent of the settled cases had very large settlements. As a result, we used the median absolute percentage error to measure the model performance. Compared to the absolute difference, the percent difference measures the level of magnitude more meaningfully. For example, a $1,000,000 difference in a $2,000,000 case has a much different meaning in a $20,000,000 case, though the absolute difference is the same. From our latest iteration of model tuning, the median training absolute percentage error is 32 percent, and the median testing absolute percentage error is 36 percent.

In comparison with a state-of-the-art settlement prediction in the academic field, our result is nearly 50 percent better, despite our more difficult out-of-sample evaluation from a more robust training and validation split. The graphs below plot the predicted settlement amount against the actual settlement. In both the train and validation set, the points follow the 45-degree diagonal line, i.e. prediction and actual observation are close.





VISTA ANALYTICS

1333 H STREET, N.W. SUITE 600W
WASHINGTON, D.C. 20005
(202) 808-2286

Our machine learning solution can be applied to differentiate high impact cases from low impact cases. Figure 3 showcases the top five high impact cases with a 5-million or plus settlement.

Figure 2: Prediction Highlights

Case Name	Filling Year	Actual Settlement	Predicted Settlement	Percent Error
State Street Corp.	2009	60M	60.2M	0.5%
Covisint Corporation	2014	8M	7.9M	1.1%
Hewlett-Packard Co.	2011	57M	56M	1.6%
Navistar International Corporation	2013	9.1M	8.9M	1.6%
CVB Financial Corp.	2010	6.2M	6M	2.7%

Conclusion

Our analysis shows that with adequate data, machine learning algorithms can potentially provide an 85-90% chance of predicting the likelihood of settlement. As shown in the above commentary certain events (in this case the 2008 financial crisis) can skew results in dramatic fashion. Taking into account Concept Drift can normalize the results much more quickly and accurately than a static view of the prior data.

The question of predicting the dollar value of a settlement will without question always have some degree of error. Settlements occur and are calculated based on a series of decisions by counsel that are not measurable. However, the goal is to calculate the order of magnitude of a settlement. 35% on either side of the median may sound large but if the system were to tell the relevant lawyers that a case had a 87% chance of settling with the settlement amount predicted at between $18,000,000 and $22,000,000 we believe that is actionable data.