**Shareholder Class Action Litigation Data Analysis**

*Dolby Laboratories, Inc.*

*Merrimack College – August 2018*

*Patrick Doyle – Dr. Andrew Banasiewicz*

**Abstract**

The purpose of this paper was to utilize market data from different sources to make statistically significant predictions for *Dolby Laboratories Incorporated* with regard to shareholder class action litigation. The chance of litigation within GISC 45 and the settlement amount in the event of a lost lawsuit were predicted using information from five different data sources.

The data sources were explored, transformed and merged to create a unified data set containing the highest possible quality data relevant to the predictions that needed to be executed. Hypothesis testing and predictive modeling techniques were executed on the transformed information with results placed in the results summary section.

**Data Preparation**

The data that was studied, parsed, cleaned and normalized was delivered as five separate filed utilized a tabular structure (comma or tab delimited) spanning the time from 2010 to 2014. Each data set was studied and simplified into data-frames containing information possibly relevant to eventual modeling using a Boolean indicative of litigation as the dependent variable.

At the highest level the data sets were transformed into a form in which each record contained a stock ticker acronym and a fiscal year as a dual attribute unique value so that it could be joined with other data later in the transformation process. A yearly record was chosen because the litigation and SEC fundamentals data were delivered on the yearly scale. All the delivered data with the exception of the class action litigation cases file was specific to GICS 45 – Information Technology Sector. This is the GICS containing Dolby (DLB). The entire group was used even though sub-sector information was provided as using any less than the entire industry would certainly yield a less than appropriate amount of data for analysis.

The *Securities\_DLB.csv* data was used as the base data set. This was chosen as this had the largest number of rows containing unique ticker-year combinations of all the delivered data after aggregation by year. It was discovered that if the *pccrm* field indicating the closing price of a stock on a given month was null, most relevant data was missing. The ticker, year, monthly trading volume and closing price for each month was extracted from the data set and grouped by ticker and year.

The trading volume and closing prices were averaged for a given year. This produced a data set that adhered to the yearly company metric convention decided upon. A final field was calculated and added to the data called *price-trend*. This was calculated by subtracting the last closing price of a given year from the first one. The averages aggregated prior to this calculation left the data with no indication of yearly organizational growth. Finally, null values for monthly trading volumes were replaced with zeros prior to yearly aggregation. Spot checking multiple organizations that had null trading volumes accurately depicted that it was not publicly traded during the period of the record.

*Ratings\_DLB.csv* was a file containing data relevant to Standard & Poor’s various ratings of publicly traded companies on a monthly time scale. The various ratings all had different styles (A+ vs AAA etc..). All ratings fields were converted to a number value as they are categorical, but ranking is still relevant (and averaging). Anything starting with an A was converted to a 4, B’s a 3, C’s a 2, D’s a 1, missing values a 0, and liquidation status a -1. Due to the nature of missing ratings, a correlation was calculated among the various ratings with null and empty values removed. There was a very convincing (over 97%) correlation between the overall quality ranking and the long-term credit rating for an organization. Because of this, missing quality rankings could confidently be substituted with long-term ratings where applicable. The ratings were then aggregated by year, averaging the monthly company ratings and joined to the securities data set.

*Stocks\_DLB.txt* was a tab delimited file containing daily stock data. The useful data from this set was earnings per share and shares outstanding. Both attributes were aggregated and averaged by company and year and added to the securities data set. Any records that contained null values for ESP and for outstanding shares were omitted which eliminated null values from the data.

*Sca\_filings.csv* was a comma delimited listing of corporate litigations with results, dollars settled and the year of the filing. The file had to be parsed to remove the dollar sign symbols. The company listing from the securities file was used to filter the litigation data down to GICS 45. All litigations in an *ONGOING* state were manually researched and modified if the status had been altered and a settlement amount entered where applicable. All remaining *ONGOING* data was removed from the set. The data was then modified to create a Boolean field called *was-litigated* along with filing year and settlement amount. This eliminated the need for a category with more than two states as a true Boolean for *was-litigated* with a zero-dollar settlement amount was indicative of a dismissal. This data was added to the securities data. The dollar amount for settlements was transformed to a percentage of organization value after being added (settlement / [trading value x shares outstanding]).

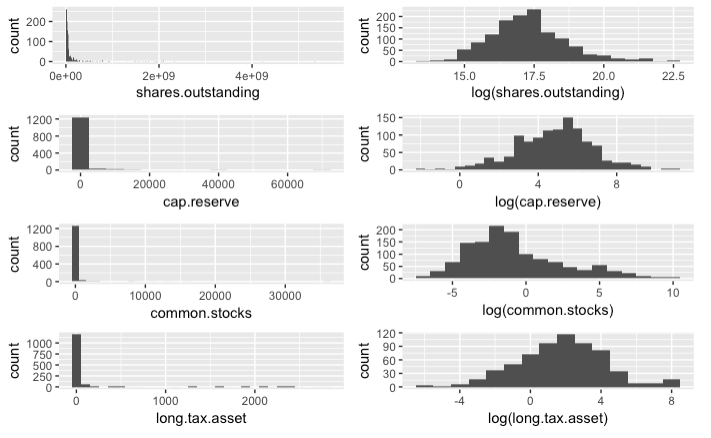
*Fundamentals\_DLB.csv* was a large file containing over 1800 attributes for each record. Mostly null records labeled *SUMM\_STD* were eliminated as the first step in the cleaning process, followed by records that were over 80% null. These two steps narrowed the data down to 418 columns. The data was then manually analyzed in batches that made sense according to the data dictionary that was provided. The batches were subjected to multi-variate correlation analysis and common-sense relevance of the attribute. The batches of non-correlating data were then tested for correlation with each other and further reduced (A correlation of .65 or -.65 was deemed significant and reduced based on sensibility and amount of null values). The result was a clean data-set with null values replaced with zeroes. There were 35 attributes that represented the data for 1600 company-years. The data was then joined to the securities data to complete the data frame.

It is worth noting that the fundamentals data also contained an S&P rating for each company. This was used to supplement any null values from the ratings set to further increase quality.

The unified data set was then tested again for correlation analysis and further reduced. A final record was created called *market-valuation*. This was derived from multiplying the shares outstanding by the closing price. This aided in creating information more valuable for analysis as closing price and shares outstanding both only tell part of a picture. Closing price also had high correlations with earnings per share and long-term taxes. Those were eliminated as a side-effect of creating higher quality information. Market-Valuation also produced over a 75% correlation with EPS, retained earnings, liability and accounts receivable fields, allowing for further reduction. Records with many null entries as results of full joins were eliminated. The resulting data set contained the following information:

* Settlement as Percent of Market Valuation
* Accumulated Loss
* Capital Reserves
* Cash
* Common Stock
* Finance (Other/Broad)
* Finance (Cash Flow)
* Other Funds
* Inventory Differential
* Options Granted
* Risk Free Rate
* Assumed Volatility
* Sale of Investments
* Long-Term Taxable Assets
* Price Trend
* Closing Price (Mean)
* EPS (Mean)
* Shares Outstanding (Mean)
* Was Litigated (Boolean)
* Company Rating
* Market Valuation
* Net Tax Liability
* Year of Data Captured
* Ticker Acronym

***The resulting 24 attribute set contained 1297 records representing 394 organizations from 2010-2014***. Outliers in the data had value but, in some cases, caused extremely skewed data, notably in data that contained values exclusively greater than zero. These were perfect cases for using the logarithm of the data to normalize:

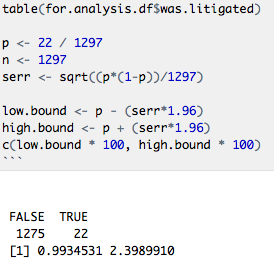


Left: Raw Attribute. Right: Log-Transformed Attribute.

The rest of the data with exception to *settlement-percent* and *was-litigated* was quite normal. In some cases, the kurtosis was a little higher than normal but not beyond usability of the information.

**Chance of Litigation**

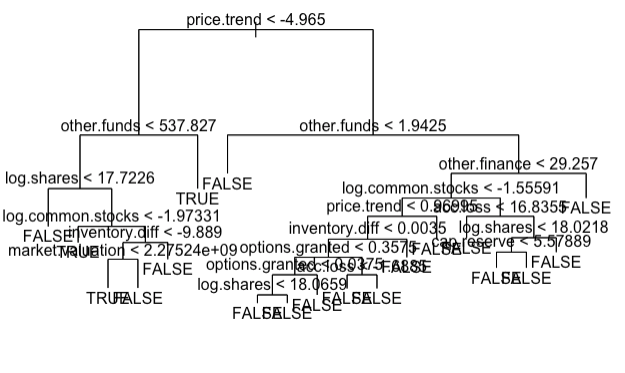
In the dataset cleaned for analysis there were 22 organizations that had litigations filed against them and 1275 that did not of the 1297 records. For the time span covered in the data from 2010 until 2014, GICS 45 only was litigated against 1.7% of time. After computing the standard error for a simply Boolean sample a 95% confidence interval was created:



Because the samples for both true and false are greater than ten and the sample size is sufficiently large (using the entire population), a 95% confidence interval can be confidently relied upon ***that the chance of litigation in GISC 45 is between 0.99% and 2.40%.***

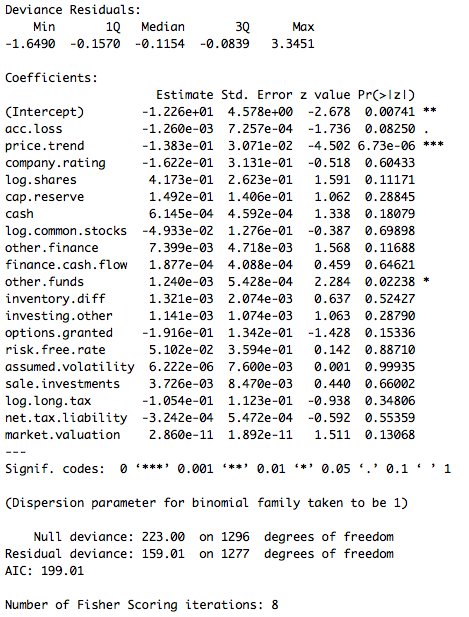
A classification tree utilizing leave-one-out cross validation was utilized in order to uncover what attributes may contribute to the chance of being litigated. Many convenient methods of modeling large data sets were not available for the given problem due to the amount of litigated organizations being so small. Splitting a set into testing and training data with 98% of the data being *was-litigated=FALSE* would yield a model that would be quite successful by just giving every record a *FALSE* prediction. Many machine learning methods would require a data set with a much larger amount of *TRUE* state organizations in order to properly train.

Models utilizing cross-validation techniques however are still available to us. A tree can be over-fit and then pruned and cross validated. Advantages of this measure would be a model that reflects the data and a great visual decision tree for uncovering what factors may forecast litigation.



Cross-Validated Regression Tree

The over-fit tree model shows that the data is not sufficient to create a model that is interpretable or accurate based on the data. What it does display is important attributes. Price trend, funds (other), finance (other), shares and stocks appear to be important legs in the tree. A cross-validated classification-regression binomial linear model may aid in visualizing the importance of these attributes.

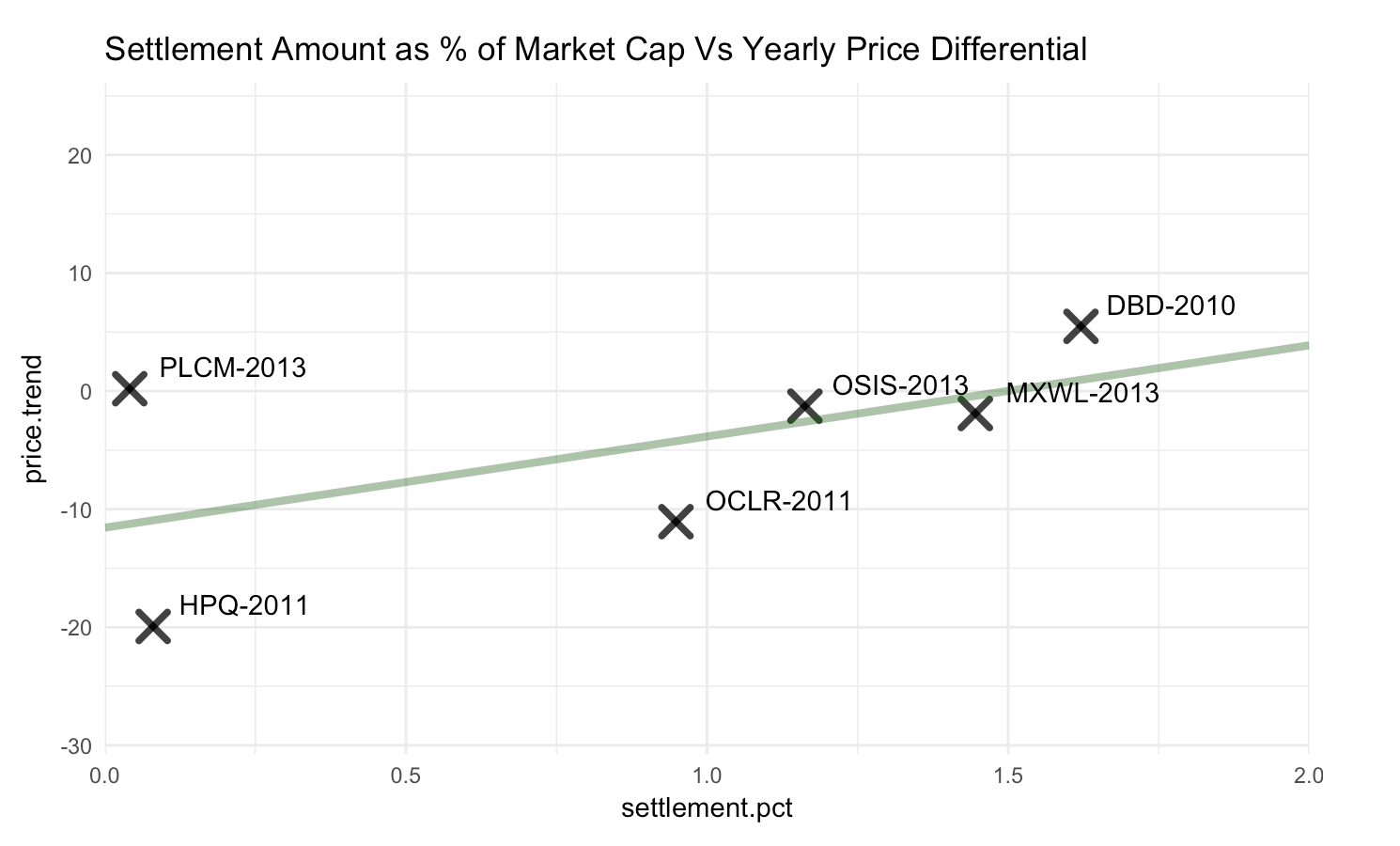


The above linear model calculated p-hypothesis values for the attributes in search of significance. ***Price-trend, other-funds and accumulated-loss showed significant difference between the litigated and the not litigated. While the data capturing litigated organizations is thin, these three attributes may be indicative of a greater chance of class action litigation.***

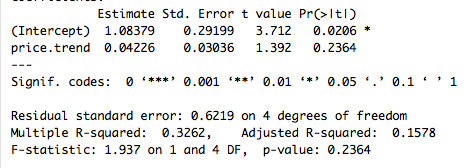
**Monetary Impact**

Monetary impact was assessed using the six records of litigation settlements in the compiled data. Settlement percent was modeled as a function of every attribute in order to build a simple linear model for estimation given the size of the available information. Settlement percent (dollar amount of a settlement expressed as a percentage of the organizations market cap) was plotted as a function of price trend (closing price on the last month of a fiscal year less than the closing price on the first month).

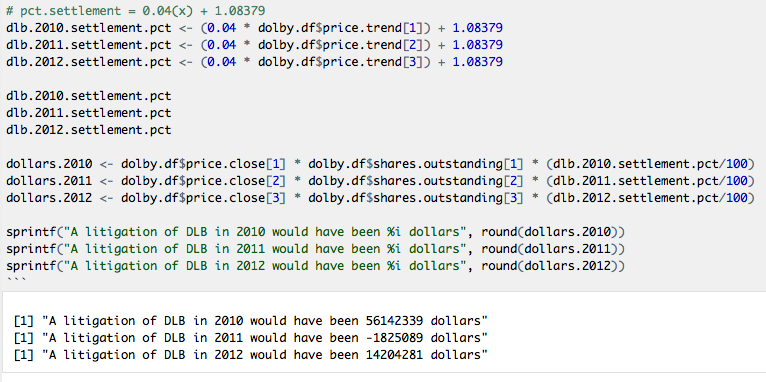
The simple model yielded an interesting result:



*\*\*\* left PLCM in despite appearing to be an outlier because there is not enough data to be confident.*



The interesting part about the yearly price trend being the largest factor in a litigation settlement is the relationship to *Maximum Probable Loss* and it appears the accepted relationship between MPL and litigation settlement holds true in this market sector. Using this linear model, a dollar value could be very roughly estimated should a lawsuit end in settlement.



Of the three records available for DLB, one of them was a very big losing year for the organization. Because none of the six companies used for the simple model ended a year worse-off, the prediction fails as a settlement cannot end in dollars paid out. What it does indicate is that a company is much less likely to be litigated if it is doing badly (it probably is not lying on SEC filings).

Applying the formula to the 2017 DLB data yielded a dollar settlement of $111,609,543 or 1.649% of market-cap.

**Limitations**

The underlying data quality was the largest limitation toward the goals of predicting the chance of litigation and the possible settlement amount. The number of companies litigated against during this window in GICS 45 that had data to match with across the delivered sets was far from sufficient. Increasing the time window across the entire data set would probably remedy the situation. Rather than a four-year window for data analysis, a ten or twenty-year span would yield far more litigations within GICS 45.

More detailed information regarding the litigations and event timing would allow for a deeper investigation than simply possible settlement amounts. Certainly, there are effects to public sentiment and employee intangibles that a lawsuit has an effect on. With greater details regarding lawsuit settlement/dismissal time and the financial data spanning the duration, an investigation to quantify the broad spanning effects of litigations during and after for settlements and for dismissals could be modeled and predicted. This type of detailed investigation could result in a negotiation with the insurance company to provide insurance to cover side-effect monetary losses while a lawsuit is taking place.

**Results Summary**

|  |  |
| --- | --- |
| **Results Summary** | |
| Unified Data Set Description | |
| *tic* | ticker symbol |
| *year* | year of data |
| *acc.loss* | accumulated losses |
| *price.trend* | closing price of last month of a year less the first month |
| *company.rating* | S&P's quality rating |
| *eps* | earnings per share |
| *log.shares* | logarithm of shares outstanding |
| *cap.reserve* | logarithm of reserve capital |
| *cash* | company cash |
| *log.common.stocks* | logarithm of common stock |
| *other.finance* | finance (other) |
| *finance.cash.flow* | cash finance |
| *other.funds* | broad/other funds |
| *inventory.diff* | inventory differential |
| *investing.other* | investments (other) |
| *options.granted* | options granted |
| *risk.free.rate* | risk free rate |
| *assumed.volatility* | assumed volatility |
| *sale.investements* | investment sales |
| *log.long.tax* | logarithm of long term taxable assets |
| *net.tax.liability* | net tax liability |
| *settlement.pct* | settlement amount as a percent of market-cap |
| *market.valuation* | closing price multiplied by outstanding shares |
| *was.litigated* | Boolean TRUE if lawsuit was filed |
|  |  |

|  |  |  |
| --- | --- | --- |
| **Litigation Findings** | | |
|  | Count | Percent |
| Total Cases | 1297 |  |
| Litigated | 22 | 1.70% |
| Not Litigated | 1275 | 98.30% |
|  |  |  |
| **Prediction Function for Settlement Amount** | | |
| *pct market-cap = 0.4 x (high-price - low-price) + 1.08379* | | |
|
| Prediction for Dolby Labratories Settlement | | |
| Year | Predicted ($) | Predicted (%) |
| 2010 | $ 56,142,339 | 1.740% |
| 2012 | $ 14,204,281 | 0.800% |
| 2017 | $ 111,609,543 | 1.649% |
|  |  |  |
| **Chance of Litigation** | |  |
| Upper Bound | 2.40% |  |
| Lower Bound | 0.99% |  |
|  |  |  |
| **Factors that may alter chance of litigation** | | |
| Price-trend or MPL | | |
| Funds From Operations (Other) | | |
| Accumumlated Other Comprehensive Income (Loss) | | |