**Shareholder Class Action Litigation Data Analysis**

*Dolby Laboratories, Inc.*

*Merrimack College – August 2018*

*Patrick Doyle – Dr. Andrew Banasiewicz*

Abstract

* Purpose of paper
* Target industry
* Target Company

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.

Overall, 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 were 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. Records with many null records as results of full joins were eliminated. The resulting data set contained the following information:

* Accumulated Loss
* Asset and Liability Differential
* Capital Reserved
* Cash
* Common Stock
* Finance (Other)
* Finance (Cash Flow)
* Other Funds
* Inventory Differential
* Investments
* Options Granted
* Risk Free Rate
* Assumed Volatility
* Retained Earnings
* Accounts Receivable Differential
* Sales of Investments
* Long-Term Tax Assets
* Price Trend
* Closing Price (Mean)
* EPS (Mean)
* Shares Outstanding (Mean)
* Was Litigated (Boolean)
* Company Rating
* Net Tax Liability
* Year of Data Captured
* Ticker Acronym

The resulting 28 attribute set contained 1297 records. Many of the attributes were very skewed due to many zero values or series of outliers. The nature of this data is such that this is the case of the true data, outliers should not be eliminated. The following data was transformed by the logarithm of it for a cleaner distribution prior to analysis.

[ LOG DATA XFORMS HERE ]

[ SHOW GRAPHICAL PROOF OF NORMALITY ]

[ ANALYIZE HUGE OUTLIERS POSSIBLY ]

Chance of litigation

[ DO TOMORROW ]

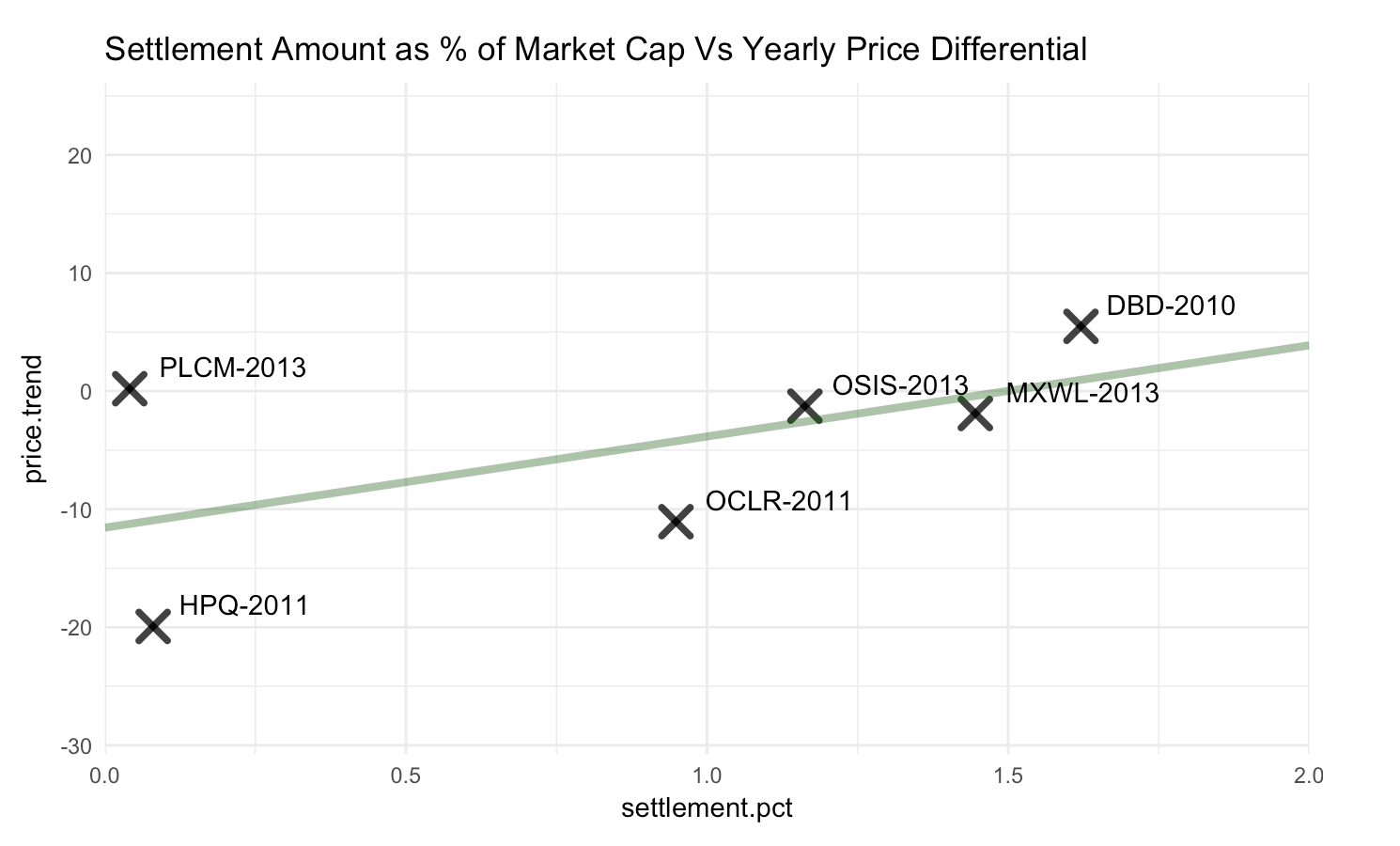
[ CALCULATE BASED ON SIMPLE MATH ]

[ ATTEMPT A MODEL ]

[ ATTEMPT HYPOTHESIS TEST on important variables. Is there a significant difference between any variables in non-litigated companies and those that are? ]

Important predictors of litigation

Monetary Impact



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

Limitations

* Data to link to litigations too small of a date range. Very limiting
* Litigation data is tied to time of litigation and not time of settlement. Further analysis could be done on the time during litigation upto and after the verdict to explore the monetary impact beyond a litigation settlement. Maybe litigation insurance could be built to cover the potential side-effects as well

Results Summary

Conclusion