**A MACHINE LEARNING APPROACH**

**TO PREDICTING MERGERS & ACQUISITIONS**

**BEFORE THEY HAPPEN**

Table of Contents

[INTRODUCTION 3](#_Toc497141495)

[PROBLEM DEFINITION 4](#_Toc497141496)

[FACTOR DATA FOR CLASSIFIERS 5](#_Toc497141497)

[DATA DESCRIPTION 5](#_Toc497141498)

[TIMELINE AND FREQUENCY 7](#_Toc497141499)

[TRAINING DATA AND TESTING DATA 7](#_Toc497141500)

[Training Set 8](#_Toc497141501)

[Testing Set 8](#_Toc497141502)

[Prediction Set 8](#_Toc497141503)

[ALGORITHMS 8](#_Toc497141504)

[Logistic Regression 8](#_Toc497141505)

[SVM 8](#_Toc497141506)

[Decision Tree 8](#_Toc497141507)

[Random Forest 8](#_Toc497141508)

[LSTM Neural Network 8](#_Toc497141509)

# INTRODUCTION

Traditional merger arbitrage works by evaluating the probability of success of an already announced deal. Although this would be an interesting process to apply statistical techniques to, we focus on trying to predict deals before they happen.

We consider several different approaches for this:

* The first, and probably simplest approach, attempts to use a factor driven classification algorithm trained on past announced deals to predict future ones based on the characteristics of the firm and the economy in which they occurred.
* The second attempts to automate the traditional method by which traders have traded on deals historically. We scrape thousands of news articles related to m&a and try to find rumors which we can trade on.
* The third and final method exploits the fact that insiders often trade on deals ahead of announcement. Finding the tell-tale patterns of such trading in stocks which we know may be acquired based on approach 1,2 allows us to trade far less often and with greater certainty.

# PROBLEM DEFINITION

Construct an algorithm which uses non-proprietary information to predict which companies from a basket will be the target of M&A activity over the next quarter. Based on the results of this algorithm, enter trades in a paper portfolio to profit from these predictions while hedging out market exposure.

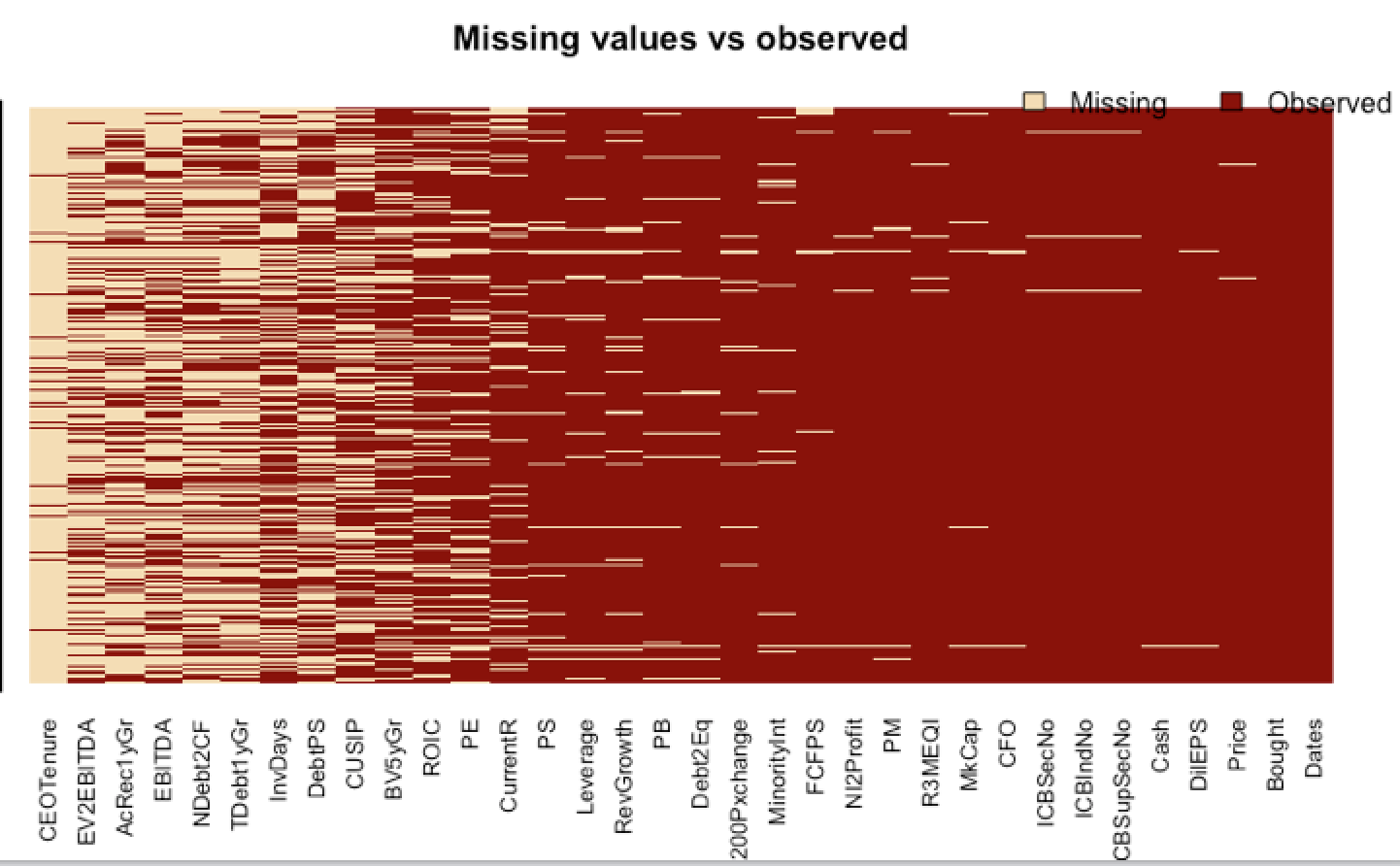
# FACTOR DATA FOR CLASSIFIERS

## DATA DESCRIPTION

|  |  |  |
| --- | --- | --- |
| **Factor** | **Description** | **Rationale** |
| 1y Growth in accounts receivable |  |  |
| 1y Growth in total debt |  |  |
| 200d price change |  | Companies are likely to be targeted when experiencing a short run adverse price shock |
| 3m return on equity |  |  |
| 5y Growth in book value |  |  |
| Cashflow from operations |  |  |
| CEO Tenure | How long the current CEO has been in the position | Only available for latest CEO, not historical |
| Current ratio |  |  |
| Current stock price |  |  |
| CUSIP |  |  |
| Debt per share |  |  |
| Debt to equity |  |  |
| Diluted earnings per share |  |  |
| Enterprise value to EBITDA |  |  |
| Free cashflow per share |  |  |
| Industry | The industry to which the company belongs | Different industries have different merger drivers |
| Inventory days |  |  |
| Leverage |  |  |
| Market Capitalization |  |  |
| Minority Interests |  |  |
| Net debt to cashflow |  |  |
| Price to Book value ratio | Latest PB ratio |  |
| Price to Earnings ratio | Latest PE ratio |  |
| Price to Sales ratio | Latest PS ratio |  |
| Return on invested capital | 1 year return on invested capital |  |
| Revenue Growth | Growth in income statement revenue over the past year |  |
| Sector | Sector to which the company belongs | Different sectors have different merger drivers |
| Sub-sector | Sub-sector to which the company belongs | Different sub-sectors have different merger drivers |
| Total Cash | Total cash on balance sheet |  |
| Bought in next quarter |  | Variable to predict |

Where available, the data as of the latest preceding quarter is used

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Min** | **Median** | **Mean** | **Max** | **NaNs** |
| **Dates** | 1999-03-31 | 2008-09-30 | 2008-08-17 | 2017-09-30 | 0.00 |
| **Price** | 0.021 | 22.530 | 32.251 | 4197.950 | 1331 |
| **DebtPS** | 0 | 6 | 22048 | 741184145 | 117792 |
| **Debt2Eq** | -6006895 | 22 | 120 | 7341116 | 12117 |
| **CurrentR** | 0.00 | 2.03 | 3.07 | 415.00 | 53812 |
| **FCFPS** | -1412.531 | 0.160 | 0.493 | 9699.439 | 7267 |
| **PE** | 0.01 | 18.92 | 46.53 | 199894.73 | 66746 |
| **PS** | 0.00 | 1.66 | 29.17 | 300274.69 | 17499 |
| **PB** | 0.0 | 2.2 | 8.8 | 679255.3 | 14634 |
| **EV2EBITDA** | 0.06 | 10.15 | 22.22 | 33538.62 | 188855 |
| **PM** | -1.091e+09 | 5.000e+00 | -4.484e+03 | 1.817e+06 | 6053 |
| **EBITDA** | -1.090e+09 | 3.426e+07 | 2.062e+08 | 2.682e+10 | 158781 |
| **Cash** | -2.600e+07 | 7.026e+07 | 4.653e+08 | 1.796e+11 | 3335 |
| **MinorityInt** | -1.060e+09 | 0.000e+00 | 7.101e+07 | 5.658e+10 | 9500 |
| **MkCap** | 0.000e+00 | 9.856e+08 | 5.374e+09 | 2.137e+12 | 5133 |
| **Leverage** | -46.29 | 2.28 | 5.40 | 38822.22 | 16685 |
| **RevGrowth** | -3353 | 8 | 178 | 25144000 | 16452 |
| **ROIC** | -19952.29 | 7.30 | 5.68 | 83973.29 | 70788 |
| **CFO** | -1.995e+11 | 1.711e+07 | 1.204e+08 | 1.964e+11 | 4040 |
| **DilEPS** | -7820000 | 0 | -29 | 5737 | 3314 |
| **R3MEQI** | -98.002 | 0.557 | 2.112 | 1374.164 | 5137 |
| **ICBSupSecNo** |  |  |  |  | 3677 |
| **ICBIndNo** |  |  |  |  | 3677 |
| **ICBSecNo** |  |  |  |  | 3677 |
| **CUSIP** |  |  |  |  | 104019 |
| **NDebt2CF** | -14781.78 | 1.78 | 4.63 | 58027.90 | 147929 |
| **NI2Profit** | -6.166e+10 | 7.882e+06 | 5.342e+07 | 1.980e+10 | 6506 |
| **AcRec1yGr** | -100.0 | 7.6 | 84.0 | 932820.0 | 165232 |
| **InvDays** | -1152.9 | 64.0 | 104.0 | 812382.8 | 126534 |
| **TDebt1yGr** | -100 | 2 | 1375 | 12307399 | 147785 |
| **d200Pxchange** | -99.63 | 5.74 | 13.52 | 225604.90 | 10232 |
| **BV5yGr** | -100.00 | 7.34 | 8.29 | 1427.65 | 85154 |
| **CEOTenure** | 0.00 | 5.58 | 7.89 | 2007.92 | 253082 |
| **Bought** | 0.000e+00 |  |  | 1.00 | 0.00 |



The main issue we confront in the dataset is that it is unbalanced. Of the 271,757 observations in the complete dataset there are only 3,906 (1.4%) takeover instances. The rarity of the positive case results in problems in training with most algorithms.

## TIMELINE AND FREQUENCY

Data was downloaded for all 31 factors for dates from 1999 March 30 to 2017 September 30 at a quarterly frequency. Since the algorithm attempts to find targets in the next quarter and most accounting data is only updated quarterly, we think this frequency makes sense.



## 

## TRAINING DATA AND TESTING DATA

The total data set was split into 3 groups:

* The first from 1999 to 2009 was used as the training set
* The second from 2010 to 2017 June was used as the testing set
* The third for Q3 2017 was used as the prediction set for constructing the paper portfolios.

Since there are a lot of NaNs and most classifiers have problems when there are NaNs, we replace NaNs with the mean of each data series so as not to introduce bias.

### Training Set

Contains 171,621 observations with 2,159 positive case.

### Testing Set

Contains 100,136 observations with 2,159 positive case.

### Prediction Set

# CLASSIFICATION ALGORITHMS

## Logistic Regression

We run the logistic regression classifier on two types of training data. The first is the standard training dataset and the second is an oversampled version of the training set.

The oversampled version is constructed from the normal version by random sampling with replacement to get an approximate balance of positive and negative buyout cases. We up-sample by ~10% to get 300,000 observations using this approach.

The logit model is trained on both the regular and oversampled versions of the training data.

In sample fits are shown below by way of confusion matrices & ROC:

Regular

|  |  |  |
| --- | --- | --- |
| Actual/Model | 0 | 1 |
| 0 | 169,462 | 0 |
| 1 | 2,159 | 0 |

Oversampled

|  |  |  |
| --- | --- | --- |
| Actual/Model | 0 | 1 |
| 0 | 92,628 | 76,834 |
| 1 | 63,269 | 67,269 |

Although the model trained on the oversampled version has a higher false positive rate, the regular version is unable to predict even a single positive case. Thus, we are forced to compromise on the false positive rate to get even a slightly viable model.

Out of sample fits are shown below by way of confusion matrices & ROC:

Regular

|  |  |  |
| --- | --- | --- |
| Actual/Model | 0 | 1 |
| 0 | 98,377 | 12 |
| 1 | 1,747 | 0 |



Oversampled

|  |  |  |
| --- | --- | --- |
| Actual/Model | 0 | 1 |
| 0 | 56,376 | 42,013 |
| 1 | 888 | 859 |



We made an internal decision to accept a higher false positive rate in order to get some predictability even though it was only slightly better than 50%.

## SVM

## Decision Tree

In sample fits are shown below by way of confusion matrices & ROC:

## Random Forest

## LSTM Neural Network