Predict Stock Market Delisting in India

Anjali Sebastian - October 14, 2018

Motivation

On the 4th of July, 2018 the Bombay Stock Exchange (BSE) delisted over 200 stocks from its platform. Stocks often get delisted in emerging market economies like India because the market is very volatile and the regulatory scrutiny for a company to get listed on the exchange is low. In the last three years alone over a 1000 companies have delisted their stock from either the National Stock Exchange (NSE) or the Bombay Stock Exchange (BSE). This is extremely distressing and can be a huge loss for investors. My motivation behind this project is to see if something can be done to assist investors. Can we use Data Science to predict if a stock will delist and use this information to help investors choose safer and better companies to invest in?

Background

Before trying to design a model that predicts if stock will get delisted or not from an exchange, it is important to understand the main reasons behind why delisting happens. Delisting typically happens due to one of the following reasons.

1. Bankruptcy or financial distress
2. Failure to comply with regulations
3. Fraud (company maybe a shell company)
4. Privatization for profit retention/more control

Other than the 4th reason which is privatization (very few companies fall in this group) all the other three reasons are negative reasons. Therefore there must be some prior indication of distress, fraud, etc. in the financial statements and market data that can be used to identify these companies. Based on this observation we can try to use supervised machine learning to try and classify companies into two categories - those that we expect to stay listed and those that could delist.

Design

I used data from Prowess Dx which is a database of the financial performance of companies in India. This database is maintained by CMIE (Centre for Monitoring Indian Economy Pvt. Ltd.) and is considered to be the best database on companies in India. This database is provided only for academic use to certain academic institutions and I was lucky to get access to it through my father.

In order to create a dataset that I could use for modeling, I combined financial statement information along with stock market data. I then selected the features which were to be included in the model using domain experience. For each company I looked up if there was a delisting date and used this to make my target (y=0 - company is listed, y=1 company delisted). I used 2017 data for companies that were listed. My delisted companies were those that delisted from 2015 onwards and I used the last available data before delisting. Financial information for 2018 was available for around 1600 companies and I kept this aside to use in my flask app for predicting delisting outcomes in 2019.

Before I started modeling I split my dataset into train set and a test set. My dataset had null values and was also imbalanced. In order to address this I used the sklearn imputer. I first fit and transform on my training set (X_train) and used the same fit to transform the testing set (X_test). In order to handle the imbalance in my data I oversampled the minority class using SMOTE only on my training set (X_train and y_train). I used a StandardScaler transformation for my logistic regression model. I also tried to use PCA (Principal Component Analysis) in order to reduce my feature space but as it did not make much of a difference and this was not used for my final model. I tried several supervised learning classification algorithms such as

1. K- nearest neighbors
2. Logistic Regression
3. Support Vector Machine
4. Decision Trees
5. Random Forest
6. Gradient Boosting

My best model was a Random Forest. Gradient boosting had good results as well but I went ahead with the random forest model as it had fewer False Negatives . I used confusion matrices, classification reports and ROC curves as metrics to determine the best model to use. After selecting the model I tuned the threshold for the model based on the metric that was most important which in this case was precision on class 0 (listing). I also made a flask app with 2018 data to predict the future.

Data

I used data from Prowess Dx database which I cleaned and appropriately restructured. As part of my cleaning process I looked for null and error values. I had to impute missing values and used a mean as my method of imputation. After performing cleaning and restructuring, my dataset had 4770 data points. 4295 of these were for listed companies and 475 were for delisted companies. I also had a very large number of features 92 in total that I used in the final model. All my features were continuous numeric data. I had no categorical data that I used in the model.

Sample of Data Features used for the Model

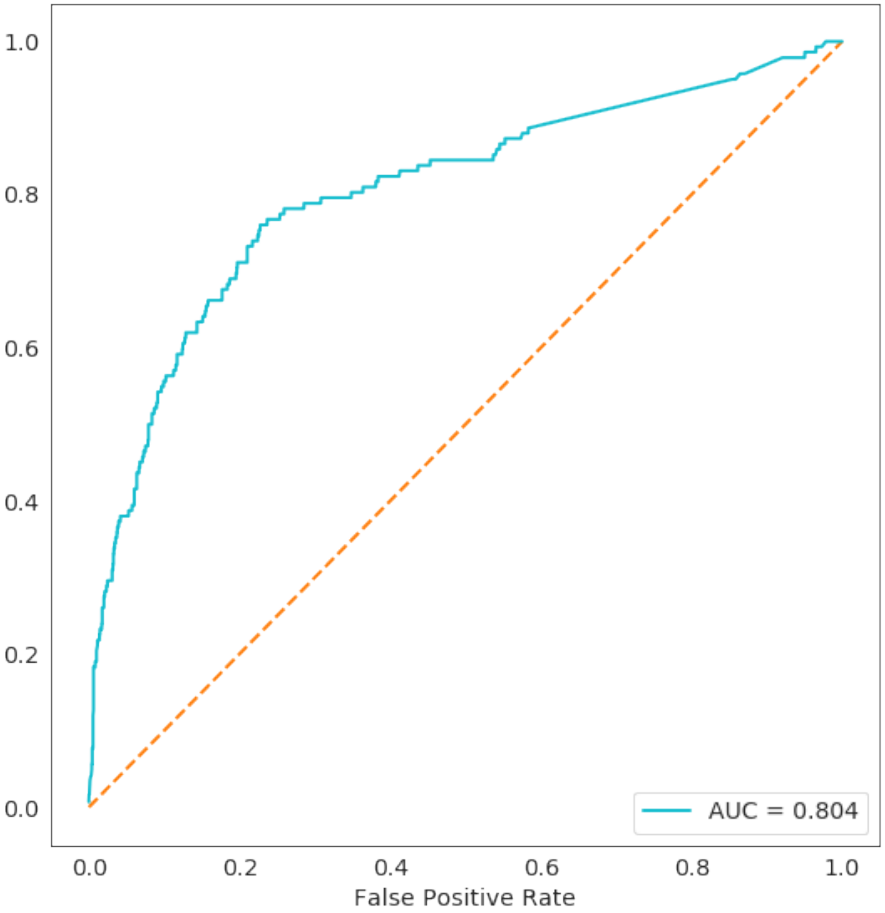
Financial Statement Features	Market Data Features
Sales	Earnings per share
Total Income	Highest Price of stock in a year
Total Assets	Market Capitalization
Profit After Tax	P/E
Total Liabilities	Stock Returns
Working Capital	Traded Quantity
….	…
X = 92 Features	
Y = Delisted (1-yes 0-no)	

Tools

IDE	Jupyter Notebooks	Atom	
Data handling and cleaning	Numpy	Pandas	
Data Visualization	Matplotlib	Seaborn	Tableau
API development	Flask	Javascript	HTML
Modeling	Statsmodel	Sklearn	
Documentation	Powerpoint	Typora	Pages
Version Management	Github		

Results

On the train set the model performed well and had an accuracy of 92%. Accuracy in this case can be used to judge the fit of the train set since I had used SMOTE to oversample the minority class.The results on the test set were also good. I used a confusion matrix, precision, recall and the ROC curve to assess the model. The figure below shows the ROC curve for the test and the AUC = 0.80 which is reasonably good.



The idea behind the project is to develop a model that can assist people by suggesting stocks that are safe ie. have a very low chance of delisting. Since this is my main requirement I used a very low threshold of 0.15 so that my precision on class 0 is high (97%). This indicates that of all the companies the model predicted as safe 97% of them stayed listed. This lower threshold results in lower false negatives (21%) - companies my model predicted would stay listed but actually delisted. This comes at a cost of increase the false positives which after changing the threshold became 429 from 136.

The tables below show the impact of changing the threshold on my model in order to increase its precision for predicting safe (stay listed) companies.

Classification and Confusion Matrix (Threshold = 0.5)

	Precision	Recall	F1-score	Support
Listed (0)	**95%**	89%	92%	1289
Delisted (1)	36%	**55%**	44%	1431

	Predicted (0)	Predicted (1)
Actual (0)	**1153**	136
Actual (1)	64	**78**

Classification and Confusion Matrix (Threshold = 0.15)

	Precision	Recall	F1-score	Support
Listed (0)	**97%**	67%	79%	1289
Delisted (1)	21%	**80%**	33%	1431

	Predicted (0)	Predicted(1)
Actual (0)	**860**	429
Actual(1)	28	**114**

Real World Use

I developed a flask app where I used data from 2018 of companies that are currently listed. Around 1613 companies out of the 4295 that are currently listed have made their 2018 financial data available. I have used this data in a flask app to predict which companies will stay listed and which may delist. My model predicts that around 270 of companies are not safe. Some of these are listed below.

Company Name	
A F ENTERPRISES LTD.	ADHARSHILA CAPITAL SERVICES LTD.
A V I POLYMERS LTD.	ARIS INTERNATIONAL LTD.
ACE SOFTWARE EXPORTS LTD.	ADVENT COMPUTER SERVICES LTD.
ADANI POWER LTD.	HUBTOWN LTD.

What I would do differently

I would have liked to have added macro-economic information to my model. Macro economic trends can help predict the risk level for the future. Additionally adding industry / sector related information could have improved the model. I spent quite a bit of time trying to create ratios (feature engineering) but later on due to lack of time I was unable to select the best features I had engineered to add into the model. I spend a lot of time gathering the data

and running the models multiple times and I wish I could have spent more time doing visualizations especially on Tableau. I wanted to try out the Naive Bayes classifier but could not at the last minute

Future Work

- Look at stock exchanges in other emerging economies.
- Add macro-economic and industry specific features.
- Try to model different reasons for delisting separately.
- Look at other indicators of financial health such as Altman's Z-score to see if they can be used.