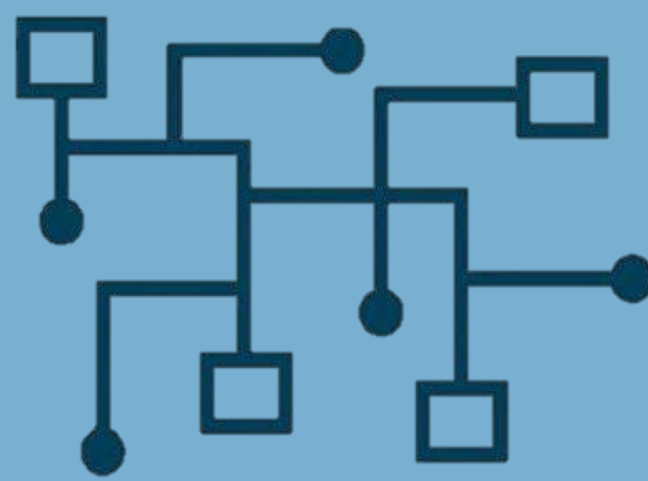# Using Deep Learning to Predict Stock Price Drop

Daniel Min, Garfield Jin, and Prof. Boris Defourny

**Department of Industrial & Systems Engineering, Lehigh University**

Abstract

In this project, we employed deep learning to predict the intraday prices of penny stocks, the common shares of small public companies that trade below $5-per-share in the U.S. market. We focused on penny Stocks because they are too illiquid for Wall Street to scale in, which allows us to avoid competition. In addition, Penny Stocks are very predictable and most of them are called *pump and dump*, which means when the stock price gets pumped up, it is usually only a matter of time before it loses all its gains [1][2]. So, we were able to build a realistic model with the sole parameter being the data of prices. Using models to trade stocks can offer higher success rate than manually trading stocks because it eliminates the emotional aspect of trading such as fear and greed [3]. The models are realistic and offers a good risk/reward ratio [4]. The data set we used are the intraday ticks of roughly 400 penny stocks that ran between May.17th and Nov. 29th, 2018. We then cleaned the data and built different models using the Artificial Neural Network to predict the price action of these stocks. Finally, we used our best models to run simulations and trade stocks using Alpaca.

Data Set and Data Processing

- Data set
  - Imported data from Bloomberg Terminal
  - Approximately 400 stocks that ran between May 17th 2018 and November 29th 2018
  - 42242 set of data with each set representing one daily price action of one stock
  - Each set of data consists of the 390 opening ticks from 9:30 AM to 4:00 PM and a binary decision variable at the end
- Data Processing
  - Cleaned and organized the original 76834 by 1551 raw data into a 42242 by 391 dataframe
  - Missing ticks have been filled by the previous executed trade price

Original Dataset → Cleaned Dataset
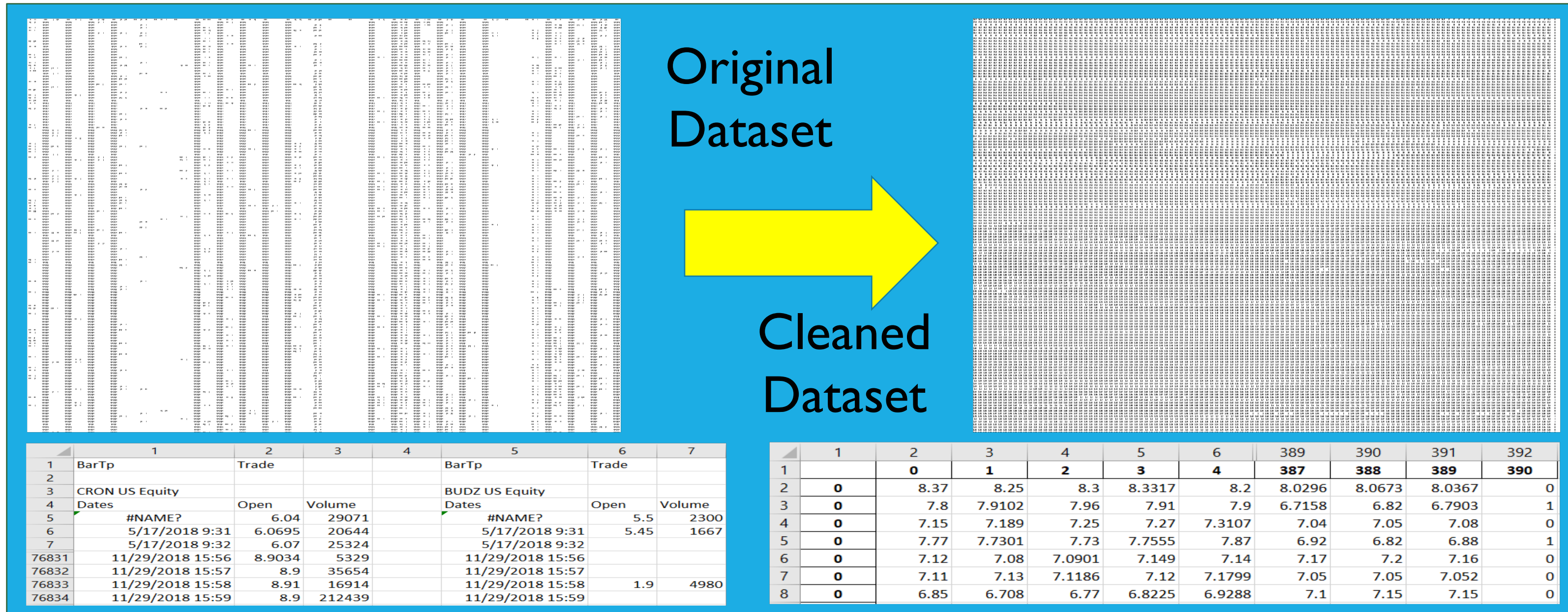
Artificial Neural Network (ANN)

- An artificial neural network is a network of simple elements called artificial neurons, which receive input, change their internal state according to that input, and produce output depending on the input and activation [5].
- Components:
  - Neurons
  - Propagation function
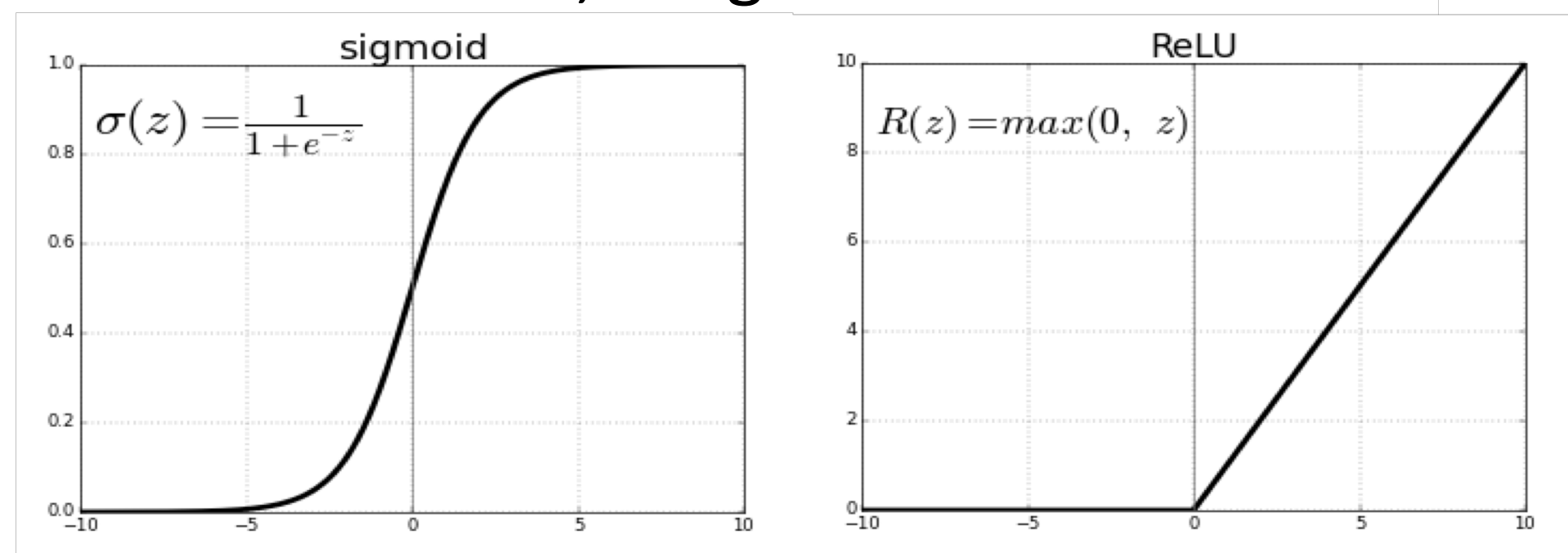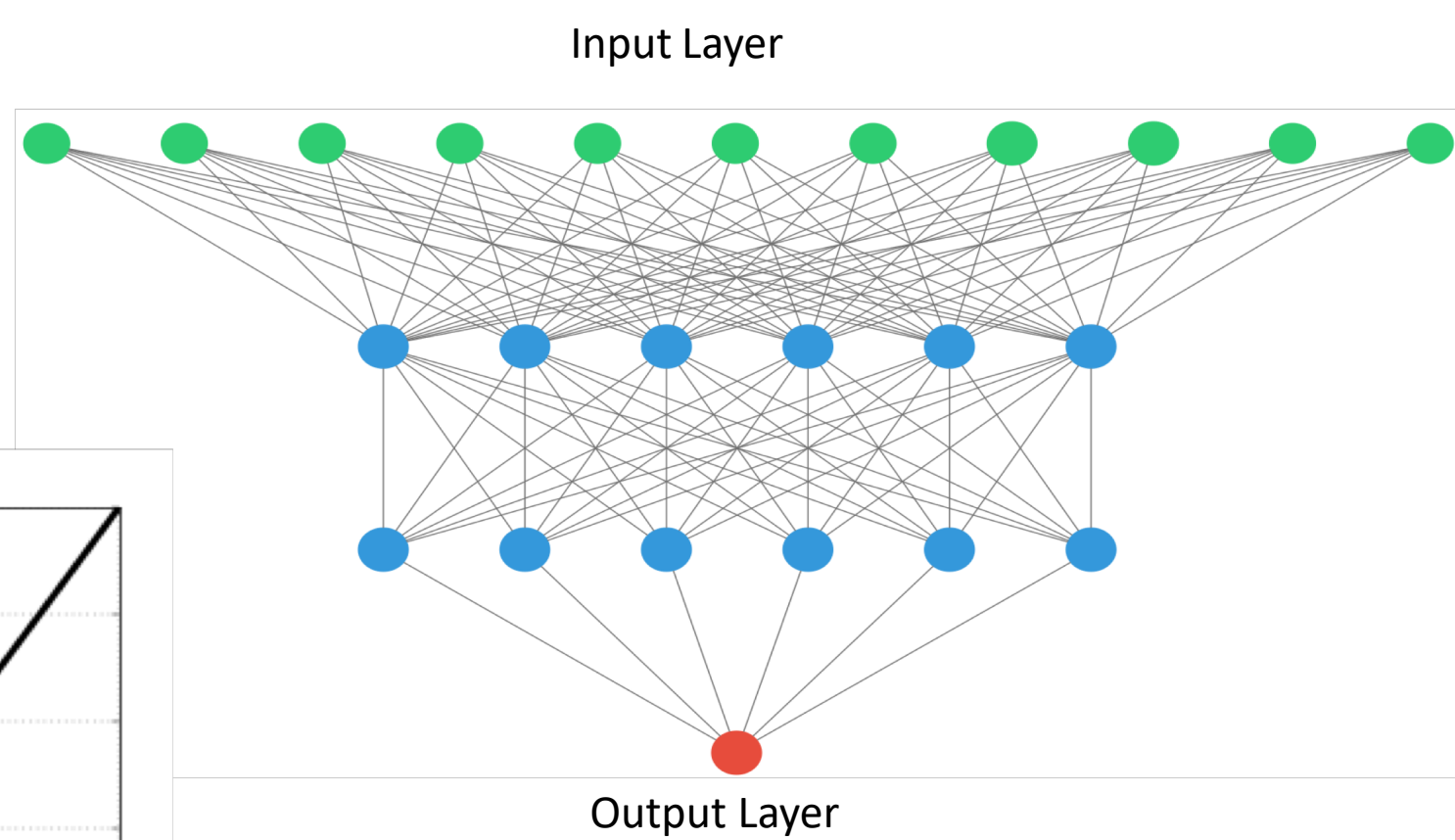 - Connections, weights and biases

Figure 1. ANN Structure Visualization (two hidden layers)

Figure 2. Activation Functions

References

[1] "Penny Stock Rules". United States Securities and Exchange Commission. April 4, 2008. Retrieved October 18, 2012.
[2] Tillman, Robert., and Michael Indergaard. Pump and Dump: The Rancid Rules of the New Economy. New Brunswick, N.J.: Rutgers University Press, 2005.
[3] Daniel, Kent, and David Hirshleifer, 2015, Overconfident Investors, Predictable Returns, and Excessive Trading, Journal of Economic Perspectives 29 (4): 61-88.
[4] Bhattacharyya, Ananjan. "The Cross-Section of Expected Returns on Penny Stocks: Are Low-Hanging Fruits Not-So Sweet?" SSRN Electronic Journal, 2016, doi:10.2139/ssrn.2386949.
[5] Patterson, Dan W. Artificial Neural Networks Theory and Applications. Prentice Hall, 1996.
[6] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
[7] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), pp.1145-1159
[8] Chen, Ting, Zhenyu Gao, Wenxi Jiang, and Wei Xiong, 2017, Daily Price Limits and Destructive Market Behavior, NBER Working Paper w24014

Stock Price Drop Prediction

- Assumptions
  - The number of nodes in hidden layer is calculated by taking the average of input nodes and output node (which is 1 for all cases)
  - Activation function is 'relu' for input layer and 'sigmoid' for output layer. Optimizer is 'adam', loss is 'mse' [6].
  - Batch size=10,epochs=50
- Price Drop Deadband
  - Using price from 9:30AM to 11:00AM to predict if there is a more than **2.5%, 5%, 10%** price drop at end of the day
- Results & ROC Curves [7]
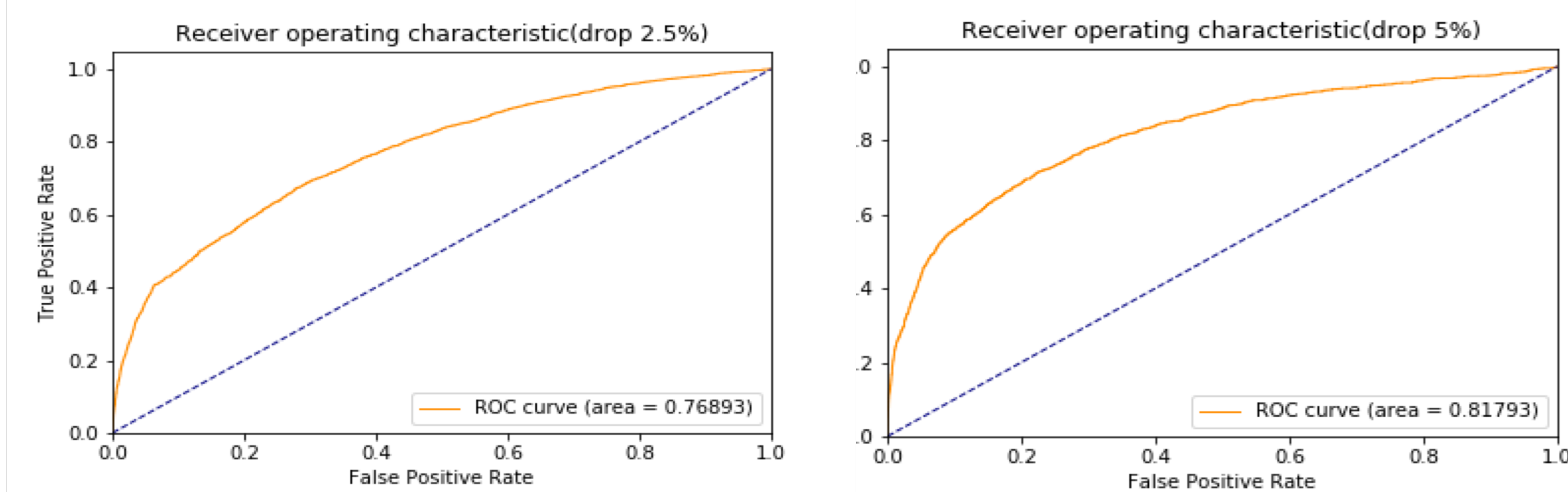
	Predicted No Drop	Predicted Drop
Actual No Drop	7213	132
Actual Drop	2547	**669**

	Predicted No Drop	Predicted Drop
Actual No Drop	8944	74
Actual Drop	1232	**311**
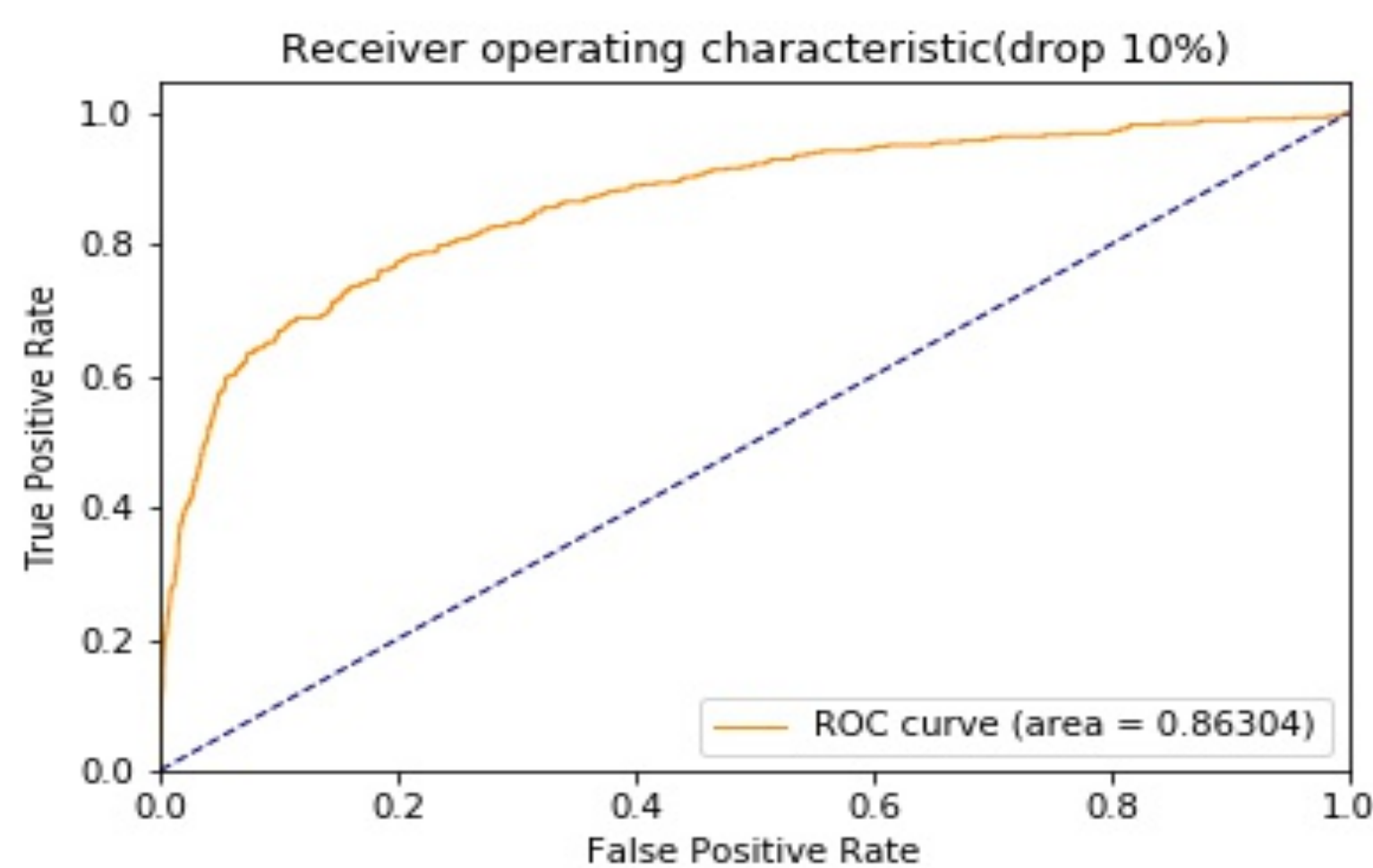
	Predicted No Drop	Predicted Drop
Actual No Drop	10155	1
Actual Drop	383	**22**

	Precision	Recall	F1-score
0	0.87893	0.99179	0.93196
1	0.80779	0.20156	0.32261
Micro avg	0.87634	0.87634	0.87634
Macro avg	0.84336	0.59667	0.62729
Weighted avg	0.86854	0.87634	0.84293

*Micro Average: averaging the total true positives, false negatives and false positives
*Macro Average: averaging the unweighted mean per label
*Weighted Average averaging the support-weighted mean per label
*Precision: the ability of a classifier not to label an instance positive that is actually negative
*Recall: the ability of a classifier to find all positive instances
*F1 score: a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0

Industry Sector Specialization

- We specialize the model to the emerging industry sector of legal Marijuana
- Selected roughly 25 stocks out of the 400 stocks that are in the Marijuana sector.
- Using price from 9:30AM to 11:00AM to predict if there is a more than 5% price drop at end of the day
- Batch_size=10, epochs=400
- Result
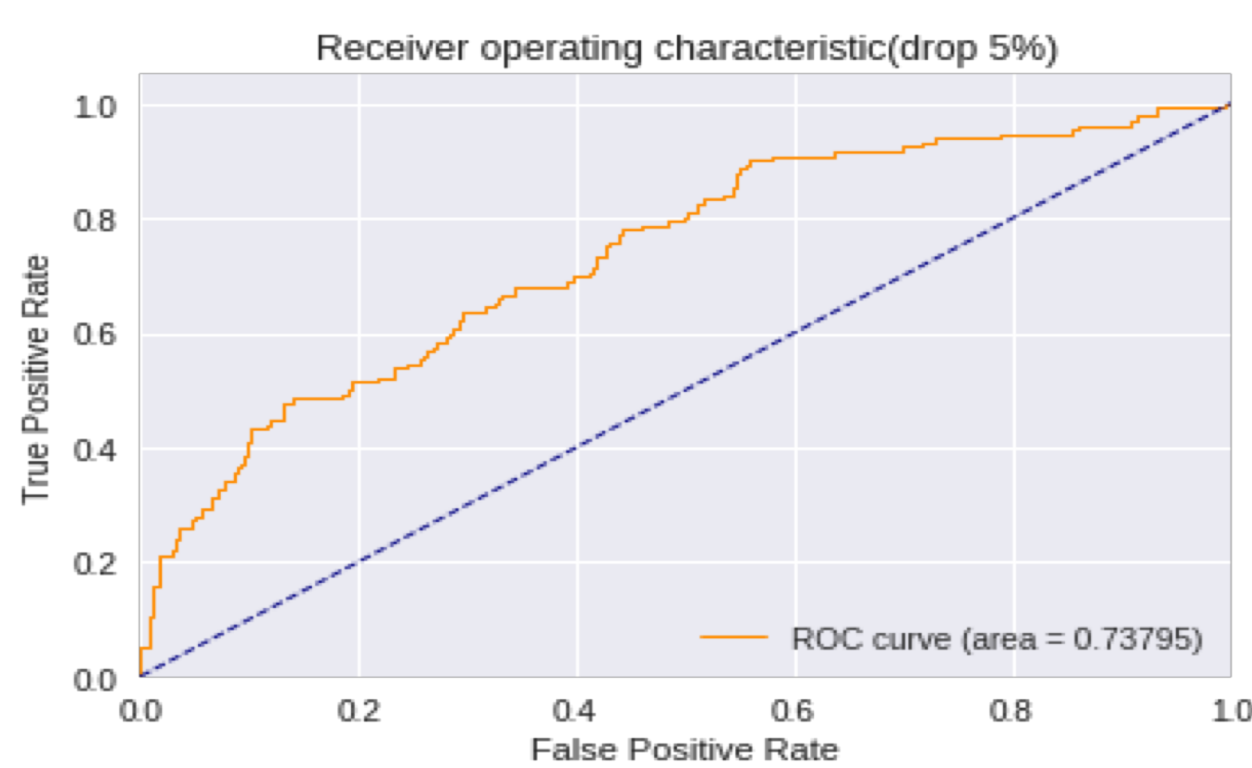
	Predicted No Drop	Predicted Drop
Actual No Drop	427	5
Actual Drop	115	**17**

	Precision	Recall	F1-score
0	0.78782	0.98843	0.87680
1	0.77273	0.12879	0.22078
Micro avg	0.78723	0.78723	0.78723
Macro avg	0.78028	0.55861	0.54879
Weighted avg	0.78429	0.78723	0.72326

Alpaca &n Heroku

- Alpaca is a simulation API designed for algorithmic trading.
- Heroku is a cloud platform that allows us to continuously run our algorithms and place trades.
- Our trading algorithm based on our model in a JSON file are being run during market trading hours continuously to back test our strategy.

Summary of Results

- 0 indicates that a decision has been made to not trade.
- 1 indicates that a decision has been made to short the stock at 11:00 AM and cover the stock at 4:00 PM expecting that the stock drop in price.
- A 80% accuracy means that for all the trades made, 80% of them would have earned a profit of larger than or equal to 10%.
- The rest of the 20% doesn't necessarily mean losses, but rather the profit earned is less than 10%.
- Our model is very conservative as indicated by the low Recall score, it offers a high win rate, but captures only about 20% of the potential profits.