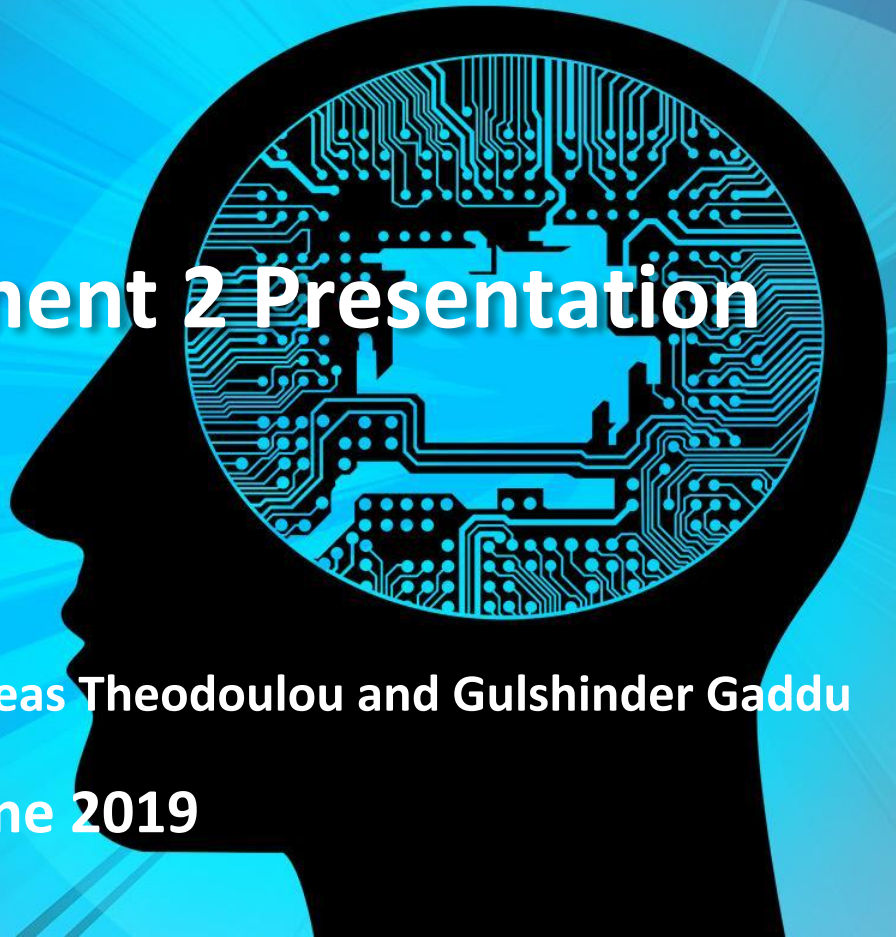
Big Data 2 Assignment 2 Presentation

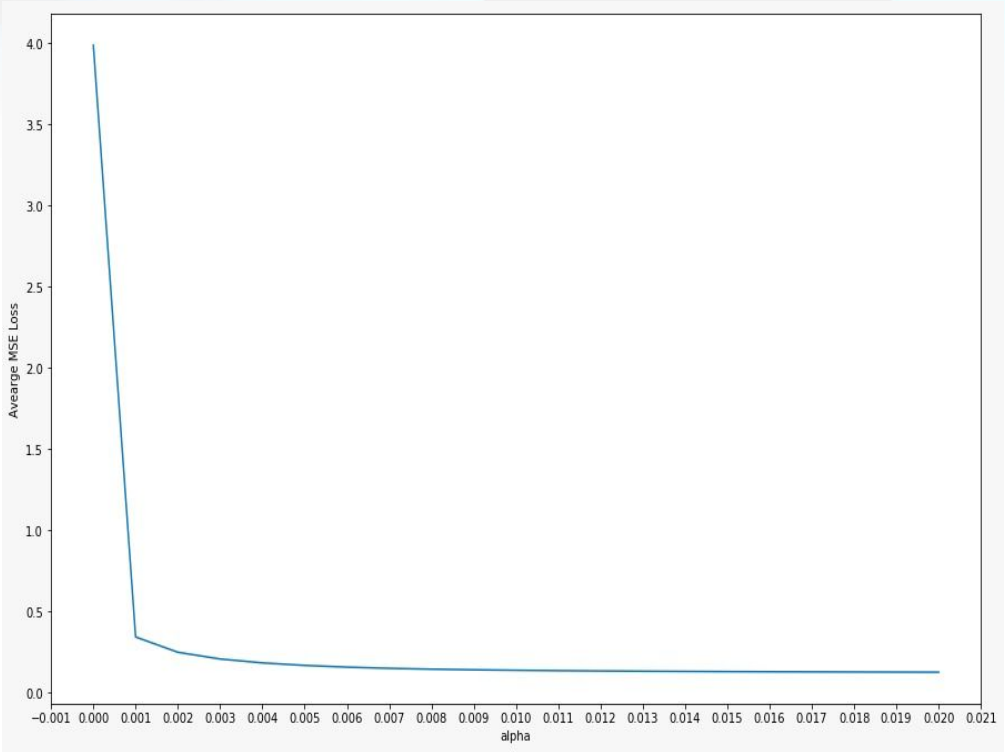


Feiyuan Fang, Edoardo Bonacina, Andreas Theodoulou and Gulshinder Gaddu

7th June 2019

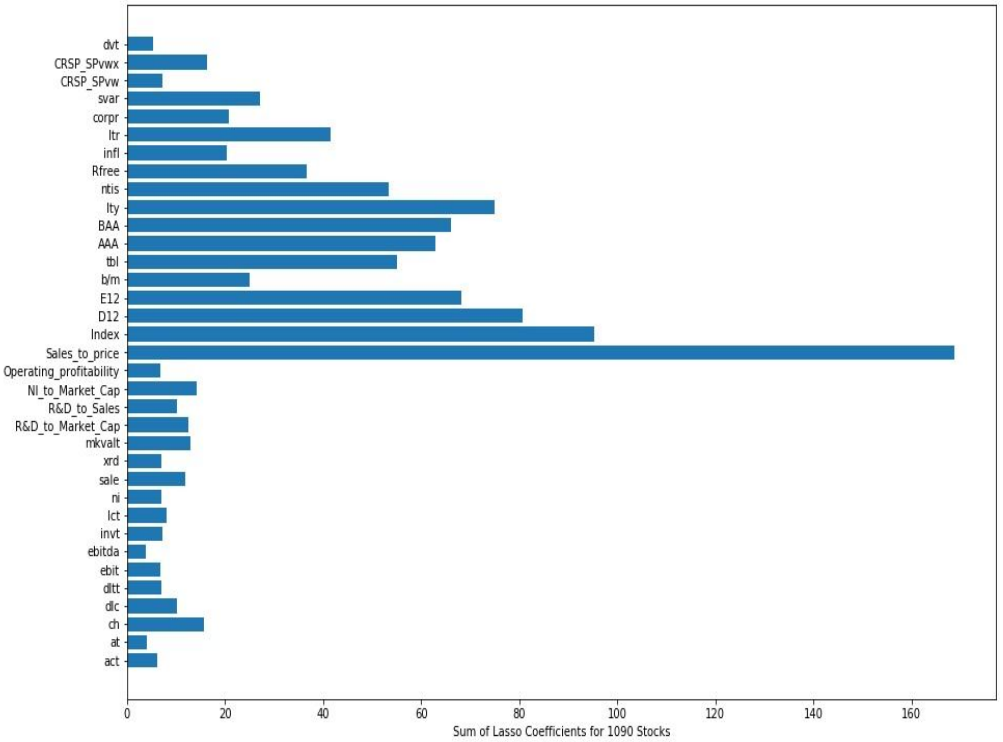
Understanding the Data – Cleaning and LASSO

- **Data Cleaning:**
  1. Drop NaNs and Unwanted Columns
  2. Stocks Retained: 96 Data Points
  3. Cleaned Data: Keeps 1090 of 3219 Stocks
- **Lasso**
  - Tuning - L1 Penalty = 0.001
    - >0.001: All Negative or Close to Zero Predictions
    - MSE Almost Minimal at $\alpha = 0.001$
  - Total Average MSE Loss for 1090 Stocks: 596.826
  - Average MSE Loss: 0.548

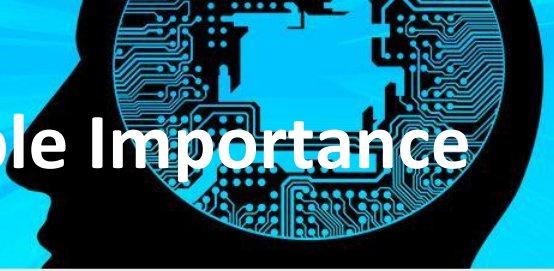


LASSO Coefficients

- **Variable Importance**
  Non-zero Lasso Coefficients - Top 5
  - Stock-specific: Sales to Price Ratio
  - Macro: S&P500,
    Market Dividends (12M),
    Market Earnings (12M),
    Long Term Yield
- **Economic Intuitions:**
  - Market Efficiency
  - Macroeconomic Environment
    More Important than Stock-specific Characteristics



Further Discussions - LASSO Variable Importance



- Are there other asset classes you would/would not expect those relationships to hold?
  - Hold: Bonds

    (Firm-related Assets <-> Firm-related Liabilities)
  - Not Hold: Commodities

    (Can be affected by economic cycles, so maybe only one or two of these variable are important but probably not all 4)
- Are there sample periods (during 2010 - 2017) you would/would not expect those relationships to hold?
  - i.e. An upward trending period (before 2008 crisis)
  - Would the variable importances change during a crisis period?

    e.g. Value, Quality based measures may become more important during crisis

Training Neural Network

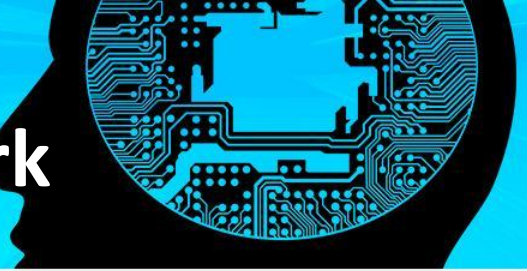
Fixed Hyper Parameters:

- No. of Nodes: Following Geometric Pyramid Rule

Tuning Hyper Parameters

- Epochs: 100, 200, 500
- Activation function: RELU, Sigmoid
  - Tested on Single Layer NN
- Layers: 1-5
  - "Empirical Asset Pricing via Machine Learning" – Gu, Kelly, Xiu
- Penalty: 0.001 – L2 penalty

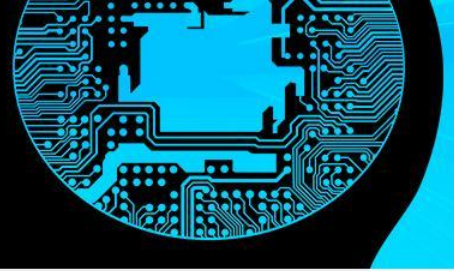
Tuning Neural Network



Average MSE Loss on Validation Set

Lasso: 0.629	1 Layer Model	2 Layer Model	3 Layer Model	4 Layer Model	5 Layer Model
100 Epoch, Sigmoid	0.363	**0.150**	**0.124**	**0.147**	0.166
100 Epoch, ReLu	0.512	0.344	0.285	0.230	0.185
200 Epoch, Sigmoid	0.373	0.180	0.124	0.148	**0.144**
200 Epoch, ReLu	0.588	0.354	0.197	0.237	0.216
500 Epoch, Sigmoid	**0.328**	0.235	0.182	0.175	0.149
500 Epoch, ReLu	0.549	0.387	0.178	0.203	0.195

Testing Neural Network



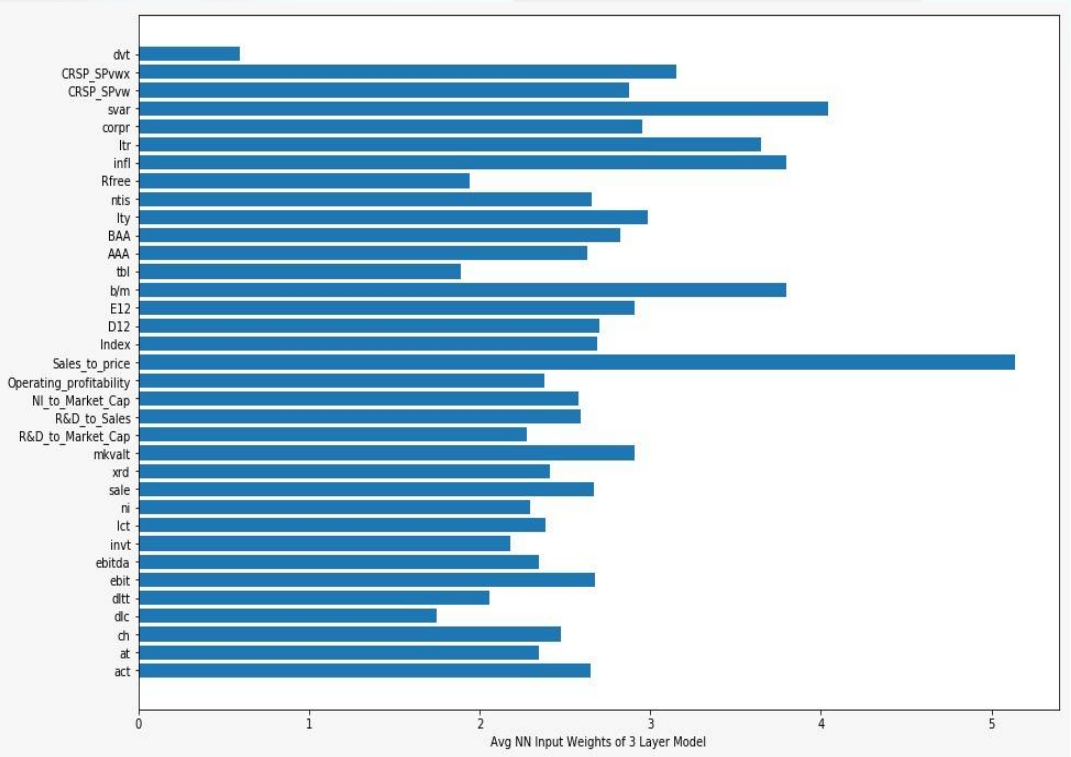
Results after running neural networks with optimal hyperparameter values on testing data (100 stocks):

Models	Cumulative MSE	Mean MSE
2 Hidden Layer, 200 Epochs, Sigmoid	28.24	**0.28**
3 Hidden layer, 200 Epochs, Sigmoid	16.93	**0.17**

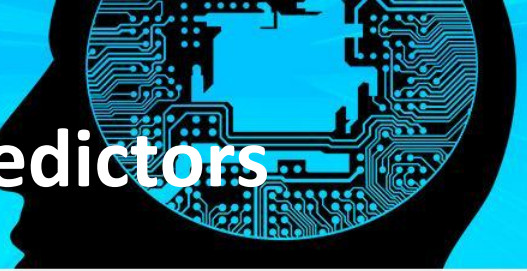
Neural Network Variable Importance

- Important Variables
  - Sales to price
  - Stock Variance
  - Inflation
  - Book to market
  - Long term return

- Primarily stock level characteristics that dominate in contrast to the LASSO
- Suggests that markets are more inefficient than the LASSO implies

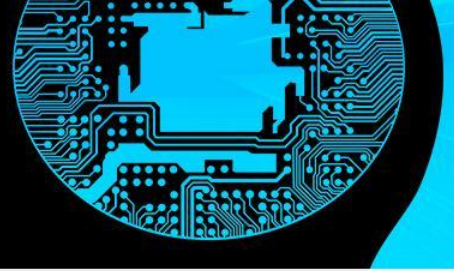


Including Past Returns as Predictors



- Lag 3
  - Big Data 1 showed much more improvement
  - NN are prone to overfitting
  - Computational power
- Hypothesis: Better predictions.
  - Momentum now accounted
- Results
  - Are the predictors better without past returns or with?
  - Economic meaning of result?

Tuning Neural Network
(Incl Lag1, Lag2, Lag3 Returns)



Average MSE Loss on Validation Set

Lasso: 0.629	1 Layer Model	2 Layer Model	3 Layer Model	4 Layer Model	5 Layer Model
100 Epoch, Sigmoid	0.412	**0.171**	**0.120**	**0.120**	0.133
100 Epoch, ReLu	0.704	0.453	0.227	0.183	0.145
200 Epoch, Sigmoid	0.397	0.175	0.128	0.123	**0.117**
200 Epoch, ReLu	0.662	0.421	0.226	0.173	0.144
500 Epoch, Sigmoid	**0.359**	0.278	0.242	0.203	0.135
500 Epoch, ReLu	0.587	0.339	0.257	0.158	0.172

Testing Neural Network with Lag Returns

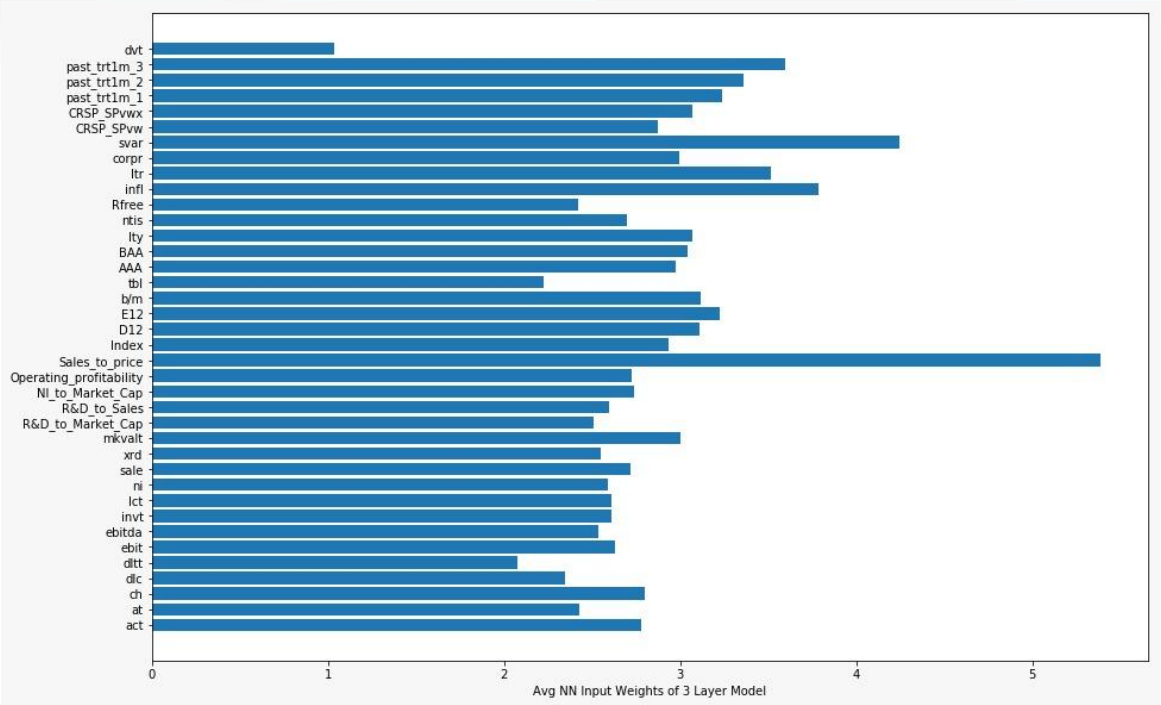
Results obtained after applying neural networks with optimal hyperparameter values on testing data including previous 3 month returns:

Models	Cumulative MSE	Mean MSE
2 Hidden Layer, 200 Epochs, Sigmoid	29.61	**0.30**
3 Hidden Layer, 200 Epochs, Sigmoid	16.86	**0.17**

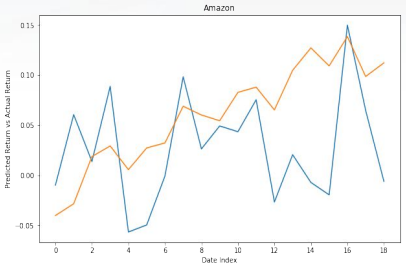
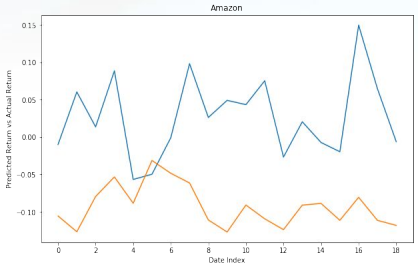
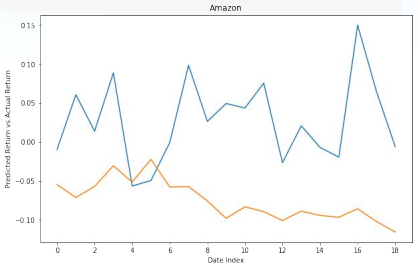
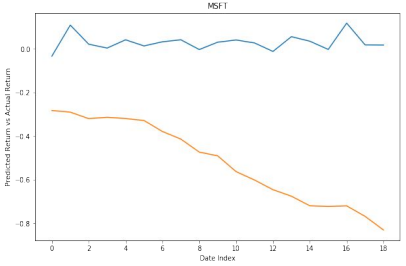
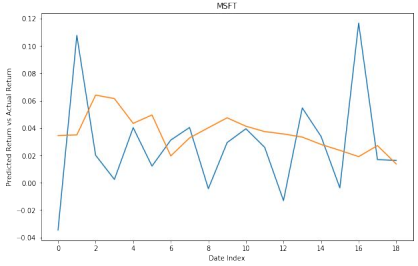
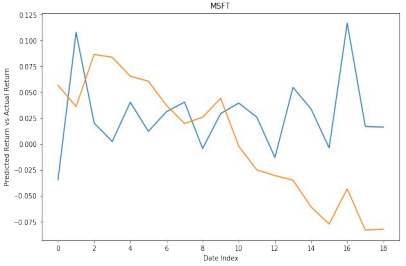
Neural Network Variable Importance (with Past Returns)

- Important variables
  - Sales to Price
  - Stock Variance
  - Inflation
  - 3 month Mom
  - long Term Yield

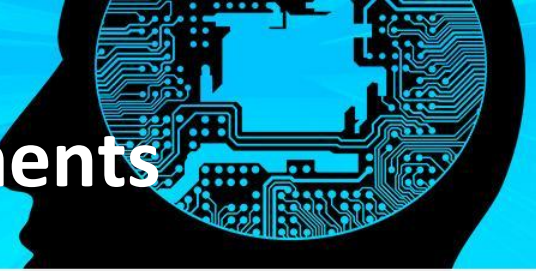
- Again stock level characteristics dominant over macroeconomic variables
- Inclusion of 3 month momentum suggests that momentum over a longer time frame more important than over a shorter one



Visualising Actual vs Predicted Returns

	Lasso	NN3	NN3 Lag Ret
Amazon			
Microsoft			

Summary and Improvements



- The Neural Network greatly outperformed Lasso but adding lagged returns had at best no negative impact
  - Importance of past returns is not relevant over a small horizon and should be expanded
- The impact of depth of the NN diminished after 3 layers
  - Worth exploring other architectures to see if this is robust
- Stock level characteristics seem to dominate
  - Potential to add more stock level characteristics
  - Perhaps the market is only looking at linear relationships with these variables
- Try different types of cross validation (eg Nested, Combinatorial purged)
- Try Recurrent Neural Network e.g. Long Short Term Memory NN