STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Deep Learning Applications on SP500 Financial 10-K Reports

Ivan Bakrac

Final Project, BIA 667, Deep Learning and Applications, Fall'20

Table of Contents

Problem Overview

Project Overview

- Various studies have shown importance of natural language processing of published reports (10-K and 10-Q) and its informativeness.[1-3]
- **Goal**: Apply deep learning methods on **10k risk sections** and relate it to fundamental target variable (**PE Ratio**).

Table (1.1) Risk Section of 10k Statements, Oracle Corp, CIK: 0001341439

```
In [16]: print(pull_risk_section(text))
```

```
 Risk Factors We operate in rapidly changing economic and technological environments that present numerous risks, ma
ny of which are driven by factors that we cannot control or predict. The following discussion, as well as our "Critic
al Accounting Policies and Estimates" discussion in Management's Discussion and Analysis of Financial Condition and R
esults of Operations (Item 7), highlights some of these risks. The risks described below are not exhaustive and you s
hould carefully consider these risks and uncertainties before investing in our securities. The COVID-19 pandemic has
affected how we and our customers are operating our respective businesses, and the duration and extent to which this
will impact our future results of operations and our overall financial performance remains uncertain.  A novel strain
of coronavirus (COVID-19) was first identified in late calendar year 2019 and subsequently declared a pandemic by the
World Health Organization in March 2020. The long-term impacts, if any, of the global COVID-19 pandemic on our busine
ss are currently unknown. We are conducting business as usual with modifications to employee travel, employee work lo
cations, and cancellation of certain marketing events, among other modifications. We will continue to actively monito
r the situation and may take further actions that alter our business operations as may be required by federal, state
or local authorities or that we determine are in the best interests of our employees, customers, partners, suppliers
and stockholders. It is not clear what the potential long-term effects of any such alterations or modifications may h
ave on our business, including the effects on our customers and prospects. We have observed other companies, includin
g customers and partners, taking precautionary and preemptive actions to address the COVID-19 pandemic. Such companie
s may take further actions that alter their normal business operations if there are future spikes of COVID-19 infecti
ons resulting in additional government mandated shutdowns. The conditions caused by the COVID-19 pandemic have advers
ely affected our customers' willingness to purchase our products and delayed prospective customers' purchasing decisi
```

Data Description & Preprocessing

Data Preprocessing

- **Step 1**: Construct TF_IDF Scores for each document across time.
- **Step 2**: Calculate Cosine Similarity Score YoY for each SP500 company[2].

Samples

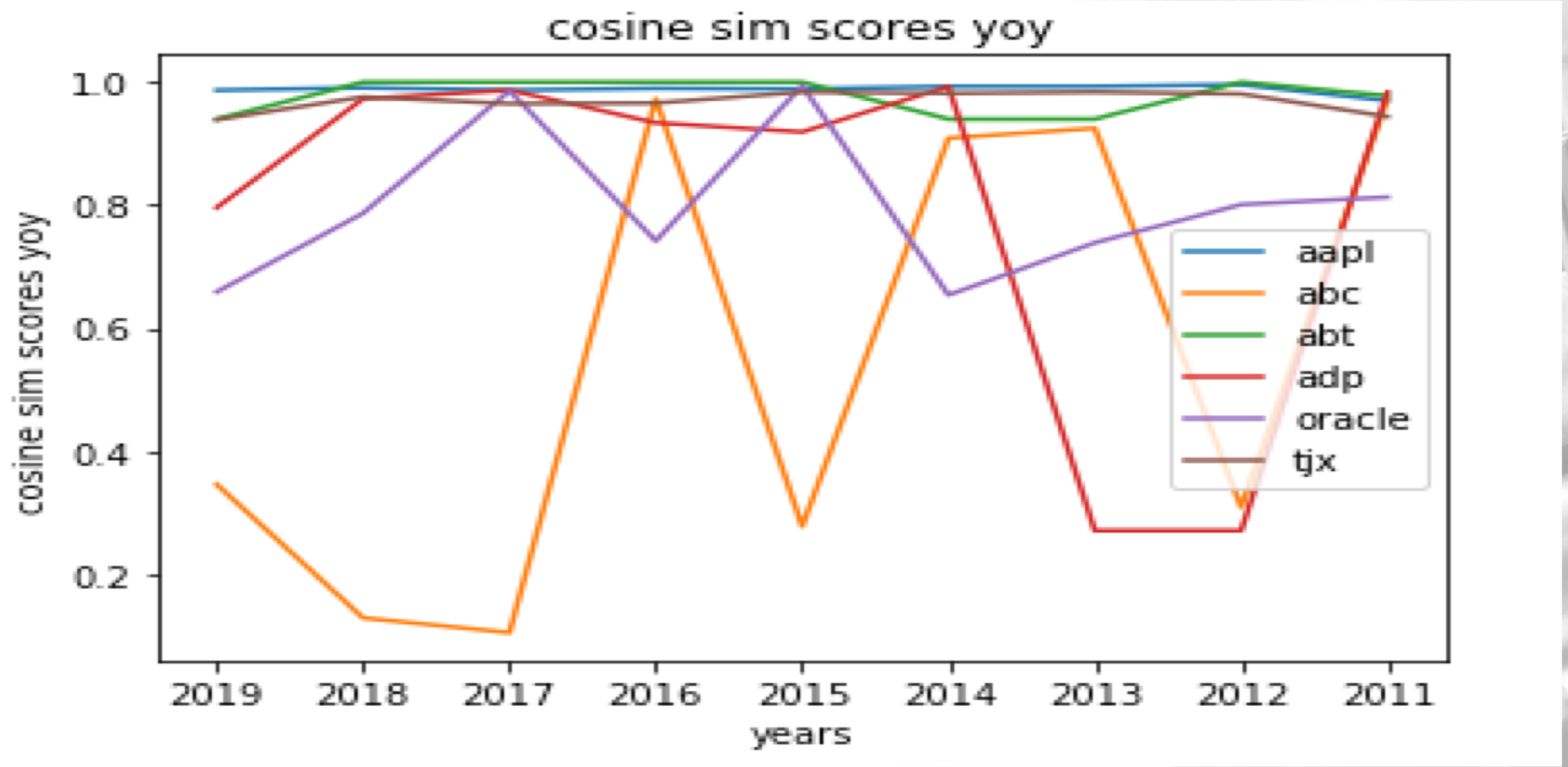
Timestep

	AAPL	ABC	ABT	ADI	ADM	ADP	ADSK	APD	AVP	AZO	BAC	BBY	BDX	BEN	BFA
2019	0.987	0.346	0.940	0.796	0.659	0.704	0.950	0.142	0.854	0.923	0.712	0.955	0.927	0.927	0.775
2018	0.991	0.129	1.000	0.973	0.787	0.879	0.883	0.310	0.987	0.965	0.857	0.970	0.920	0.958	0.740
2017	0.987	0.105	1.000	0.987	0.986	0.896	0.952	0.058	0.979	0.988	0.818	0.964	0.883	0.987	0.176
2016	0.990	0.973	1.000	0.934	0.742	0.922	0.975	0.174	0.942	0.951	0.778	0.973	0.941	0.989	0.820
2015	0.989	0.278	1.000	0.919	0.994	0.847	0.971	0.938	0.833	0.972	0.841	0.903	0.944	0.990	0.884
2014	0.993	0.909	0.940	0.993	0.654	0.917	0.978	0.938	0.903	0.915	0.786	0.957	0.961	0.985	0.860
2013	0.993	0.925	0.940	0.271	0.739	0.887	0.972	0.980	0.936	0.957	0.861	0.956	0.942	0.963	0.882
2012	0.996	0.309	1.000	0.271	0.801	0.883	0.960	0.980	1.000	0.968	0.846	0.863	0.968	0.973	0.871
2011	0.970	0.973	0.978	0.984	0.813	0.857	0.981	0.980	0.890	0.896	0.778	0.937	0.971	0.975	0.727
est_pe_nxt_yr	28.800	11.000	24.600	22.700	13.900	27.500	54.300	25.500	25.500	13.800	13.800	13.500	17.800	8.300	8.300

Target

Data Preprocessing

- **Step 1**: Construct TF_IDF Scores for each document across time.
- **Step 2**: Calculate Cosine Similarity Score YoY for each SP500 company[2].



Methods & Results

Methods & Results

- Deep Learning Business Application Design

Step 1:

Text

- Extract Risk Section of 10k's for SP500 Companies

Step 2:

Cosine Similarity

- Apply Unsupervised Learning
- Derive Time Series of Document Similarity
- Dimensionality Reduction

Step 3:

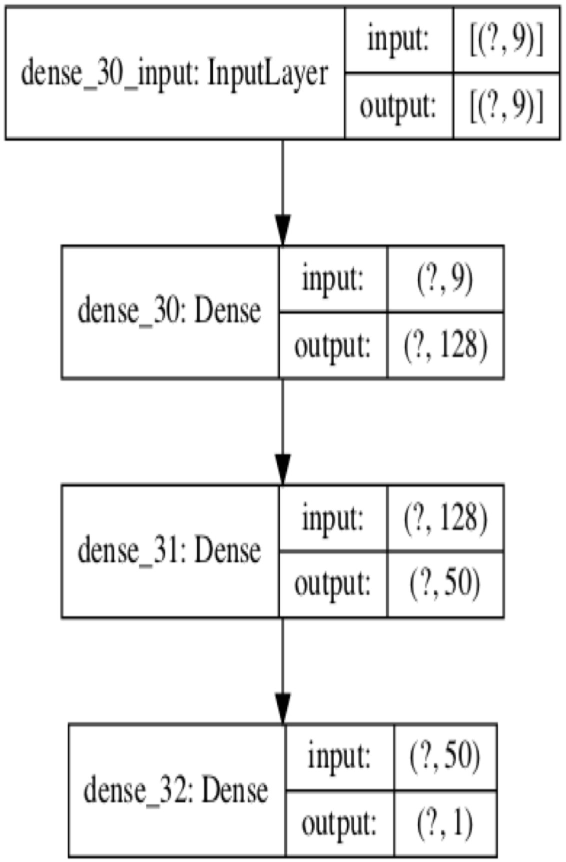
i. MLP

ii. LSTM

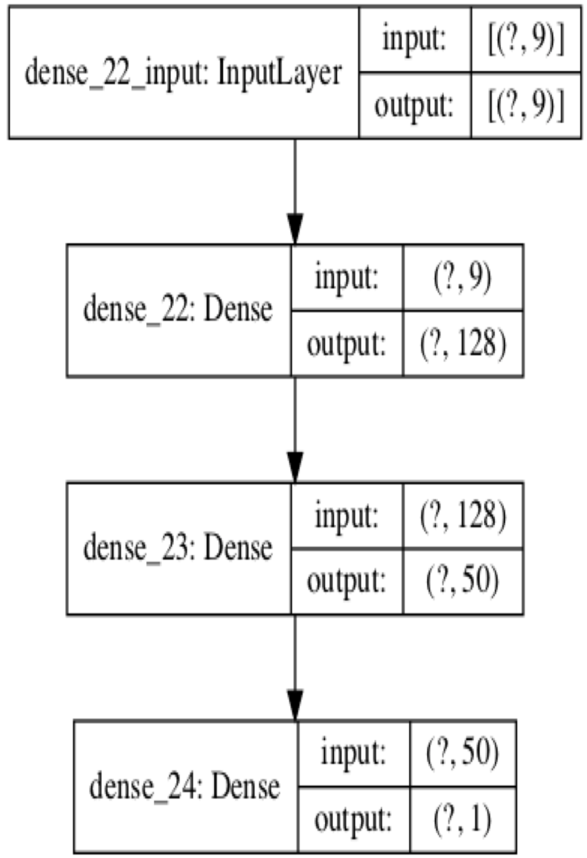
- Apply LSTM
- Predict PE

Method 1: Multilayer Perceptron Architecture

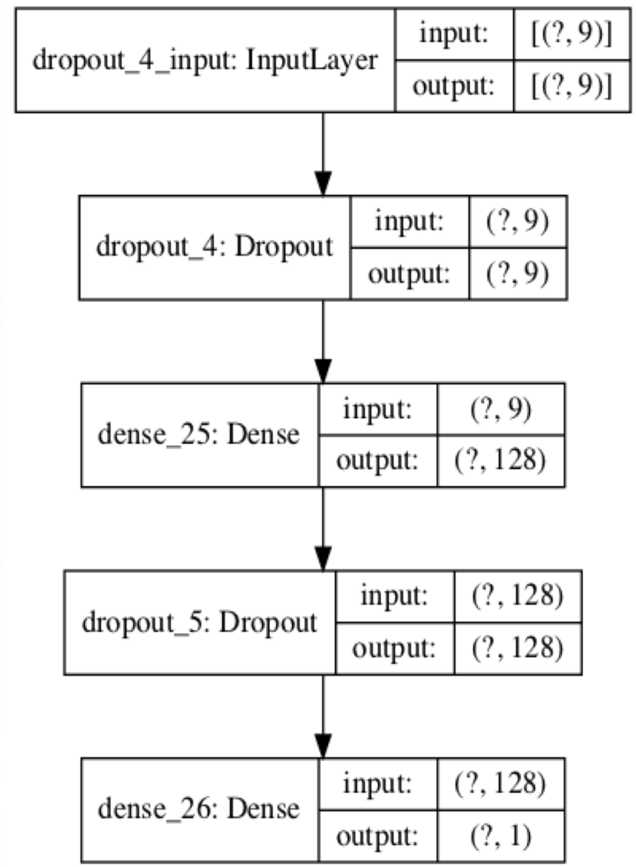
Model A



Model B: Model A + L2

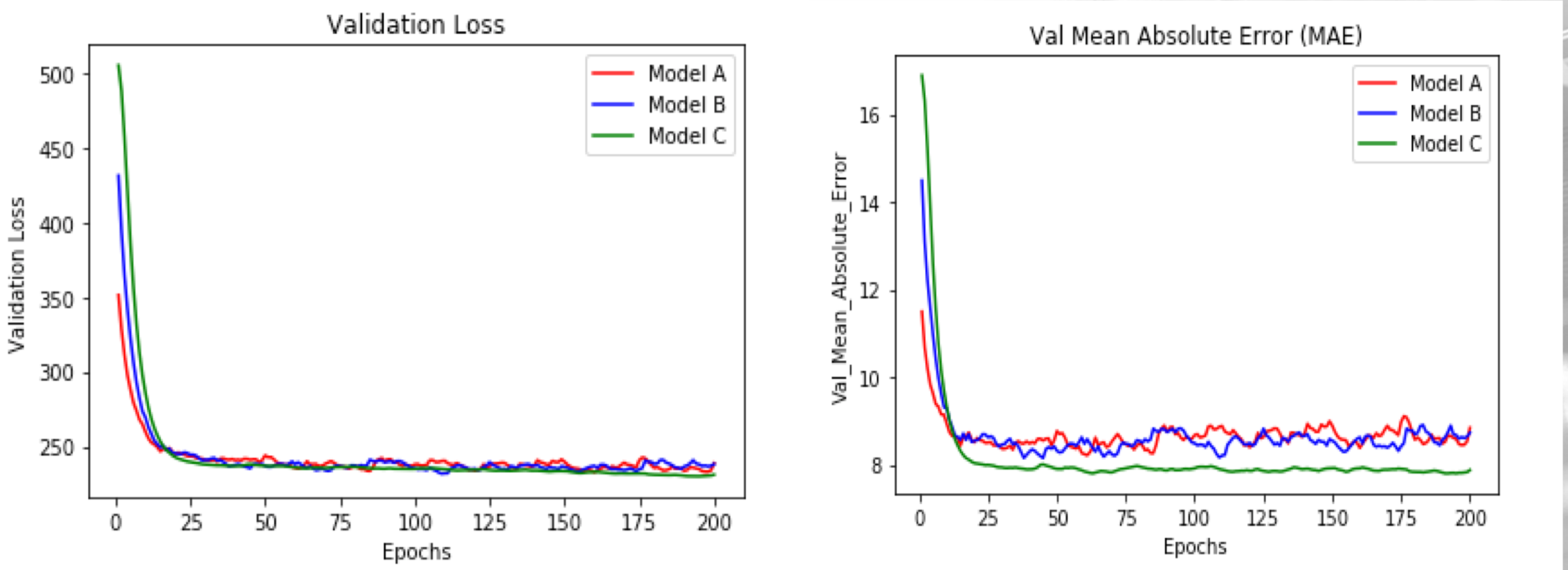


Model C, Dropout



Method 1: Multilayer Perceptron (MLP)

Model	MSE Test	MAE[1]	Architecture: Layers, Nodes, Activation, O.Activation, L1/2
Model A	257	12.3	2,128,50, Rectified Linear Unit (RELU) ,Regression (None),None
Model B	242	11.3	2,128, 50, Rectified Linear Unit (RELU) ,Regression (None), L2 (l=0.001)
Model C	**230**	**10.8**	**1,128, Rectified Linear Unit (RELU), Regression (None), Dropout (0.5)**

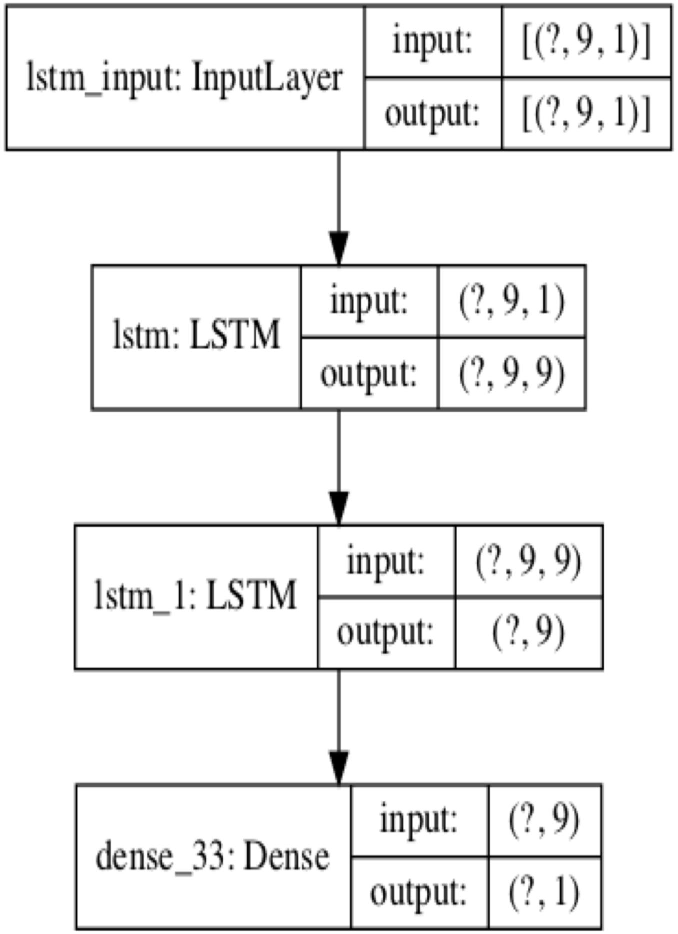


Reference : [4]

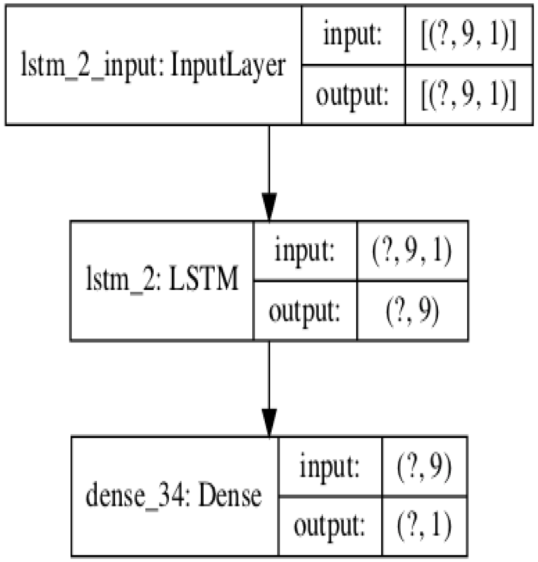
[1] MAE= Mean Absolute Error , Loss Function = MSE, Mean Squared Error

Method 2: Long Short Term Memory Architecture

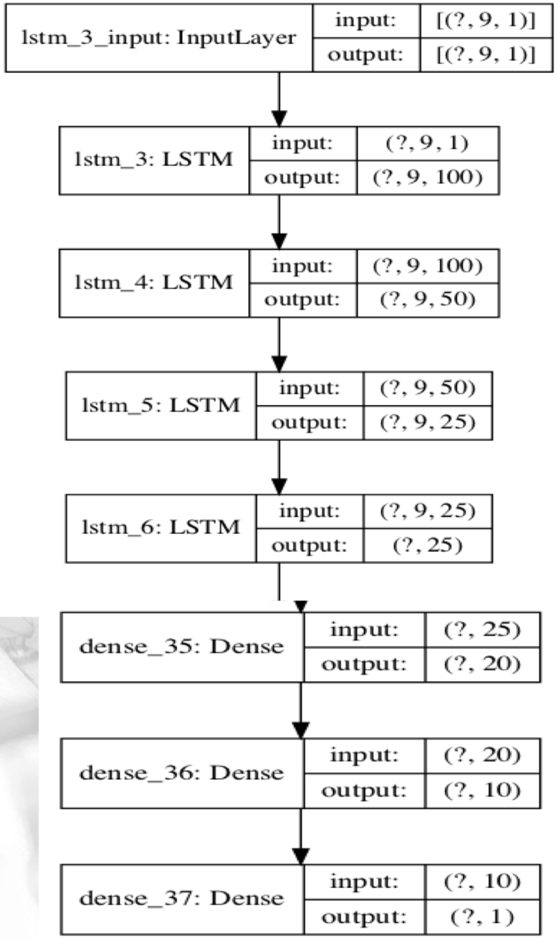
Model A



Model B

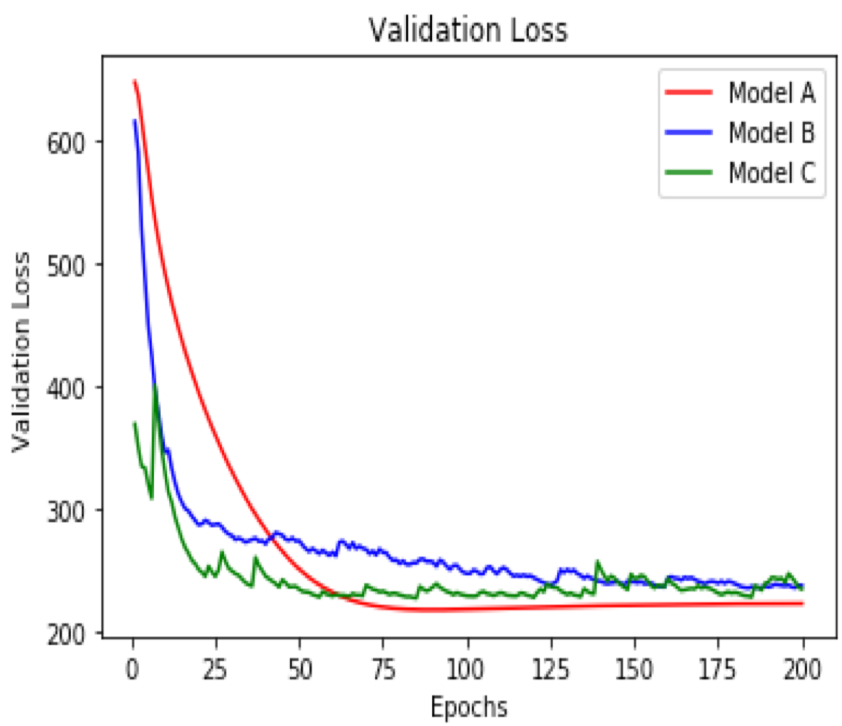


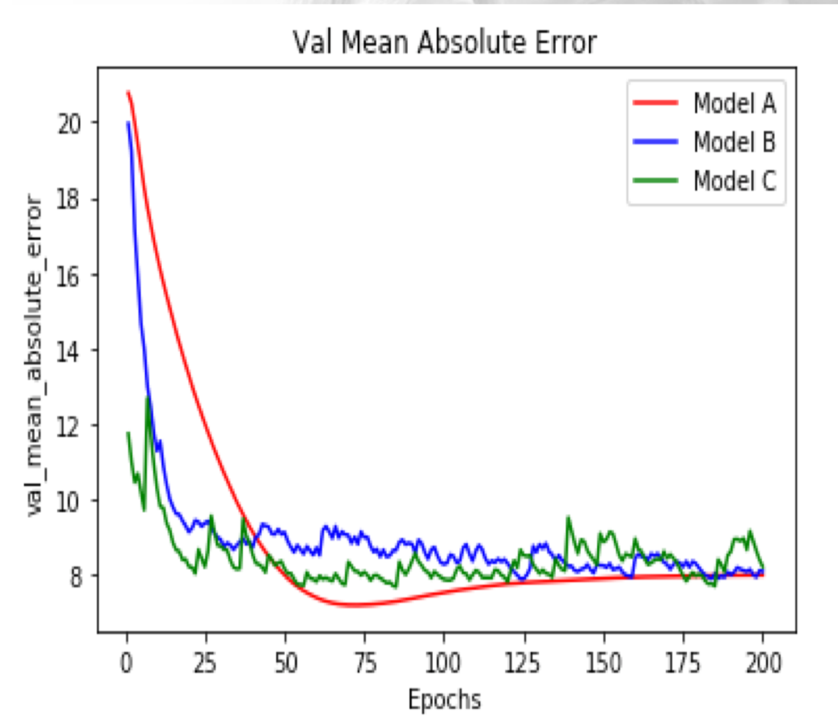
Model C



Method 2: Long Short Term Memory (LSTM)

Model	MSE Test	MAE[1] Test	Architecture: Layers, Timesteps, Nodes, Activation, O.Activation, L1/2
Model A	**238**	**7.5**	**2,9,9 Rectified Linear Unit (RELU) ,Regression (None), Dropout (0.5)**
Model B	256	11.2	1,9, 9, Rectified Linear Unit (RELU) ,Regression (None), None
Model C	255	10.8	4, 100,150,25,25, Rectified Linear Unit (RELU), Regression (None), None





[1] Metric= MAE, Mean Absolute Error , Loss Function = MSE, Mean Squared Error

Insights & Analysis

Insights & Conclusions

- Both models are able to **converge with LSTM outperforming MLP** slightly and intuitively this makes sense given “**temporal” relationships**.
- MSE (loss function) and MAE (metric) are still **relatively high** and can be further improved with more data or model experiments.
- Simpler ”**shallow”** learners models tend to do better than **“deep”** learners” for MLP structure but this might be function of tensor size.
- Simple **LSTM stacked structure** (2 hidden layers) delivers the best results.
- **Regularization tends to improve results for both MLP and LSTM** model with ideal fit around 65 epochs and dropout applied for LSTM and 45-50 epoch for MLP with dropout applied as well.

Challenges & Future Work

- Models would more likely improve if one includes **more data** outside of SP500 (large cap, small cap etc..) which would result in **sparse matrix**.
- **Refinement of target variable** by extracting common market factors.
- Experiment with **multiple timesteps and features** in LSTM by stacking multiple time steps along with rolling target (sequence of shape [t1,t2,t3| target at t +4] is alternative model to explore.
- Model architectures such as **Seq2Seq model** where encoder would represent LSTM model in this project and decoder would be current stock variables while decoder target variable p/e ratio at t+1 (output).

References

References

[1] Eric Sorensen. 2019. The Golden Age of Quants. *The Journal of Portfolio Management* 46(1)

[2] Mustafa A. Sakarwala. 2019. Use Advances in Data Science and Computing Power to Invest in Stock Market. *SMU Data Science Review* 2(1), Article 17

[3] Cohen, Lauren and Malloy, Christopher J. and Nguyen, Quoc. 2019. Lazy Prices. *Academic Research Colloquium for Financial Planning and Related Discipline.*

[4] Azimi, Mehran and Agrawal, Anup. 2019. Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning. http://dx.doi.org/10.2139/ssrn.3258821

[5] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882

[6] https://www.coursera.org/lecture/machine-learning-asset-management-alternative-data

[7] https://www.youtube.com/watch?v=_eN7D0IKu4Q.