Reverse Engineering the Corporate Credit Ratings Models of Rating Agencies with Neural Networks

Constantine Constantinidis
Department of Engineering
University of Pennsylvania, PA, USA

Abstract

Credit Ratings express a forward-looking opinion about the capacity and willingness of an entity to meet its financial commitments as they come due. This includes the credit quality of an individual debt issue, such as a corporate or municipal bond, and the relative likelihood that the issue may default. There are a few major credit rating agencies S&P, Moody's, Fitch and others that rate most public debt on a scale starting from AAA for the best quality companies to D for the companies in default with a few subcategories in between. The credit rating models used by the credit rating agencies to rate corporate debt are highly proprietary. Following the world-wide financial credit crisis of 2008, tighter regulations were introduced, such as the Dodd Frank Act, that in some cases make sharing such models with financial intermediaries and the public a criminal offense. As such, these models have become black boxes. Hence, there is growing interest in using machine learning techniques to reverse engineer these black box models and predict credit scores. Related studies have shown that neural networks and support vector machines outperform other techniques by providing better prediction accuracy. The purpose of this paper is to develop an appropriate machine learning algorithm to predict accurately the credit rating for long term senior corporate debt listed by US companies as rated by S&P in first instance by training the models on publicly available credit ratings and financial stats obtained from S&P's Capital IQ and CRSP databases available through WRDS (Wharton Research Data Services).

Introduction

Most of the past approaches to predict credit ratings have been based on statistical techniques that employed profit models, linear or logistic regression using mainly company financial ratios as input. However, many of these models may not adequately capture non-linear dynamics that exist in such financial ratio data. Further, these models cannot handle more unstructured data such as sentiment analysis, financial analyst reports, etc. A growing number of machine learning techniques have been deployed to capture non-linear patterns and temporal dependencies among such big financial ratio datasets[1]. Zhao et al. (2015)[2] employed feed forward neural networks in credit corporate rating determination. Mercep at al. (2020)[3] employed deep neural network models for behavioural credit risk assessment. Provenzano (2020)[4] used Moody's dataset, bankruptcy statuses and macroeconomic variables to build three models: a classifier, a default probability model and a rating system.

In the institutional investor world, there is a clear divide in credit ratings between Investment Grade and Sub-Investment Grade (also known as Speculative or Junk).

AAA	Investment Grade: Extremely strong capacity to meet financial commitments
AA	Investment Grade: Very strong capacity to meet financial commitments
A	Investment Grade: Strong capacity to meet financial commitments, but somewhat susceptible to economic conditions and changes in circumstances
BBB	Investment Grade: Adequate capacity to meet financial commitments, but more subject to adverse economic conditions
BB	Speculative Grade: Less vulnerable in the near-term but faces major ongoing uncertainties to adverse business, financial and economic conditions

B	Speculative Grade: More vulnerable to adverse business, financial and economic conditions but currently has the capacity to meet financial commitments
CCC	Speculative Grade: Currently vulnerable and dependent on favourable business, financial and economic conditions to meet financial commitments
CC	Speculative Grade: Highly vulnerable; default has not yet occurred, but is expected to be a virtual certainty
C	Speculative Grade: Currently highly vulnerable to non-payment, and ultimate recovery is expected to be lower than that of higher rated obligations
D	Speculative Grade: Payment default on a financial commitment or breach of an imputed promise; also used when a bankruptcy petition has been filed

Between each classification above the + and – indicators show ratings between two classifications. B+ would be above B and below BB-. So, in fact there are 22 different sub-classifications: AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, CC, C, D. Fitch and Moody's follow similar classifications.

Many investment firms are strictly restricted to investing only in Investment Grade issuances, and for a company dropping or rising from one rating category to another may result in large sales or purchases of their affected debt instruments. Therefore, one of the primary goals of a good credit rating model should be to accurately predict these two broad categories. A secondary goal is to be able to differentiate further between A & higher, BBBs, BBs, and B & lower ratings which are also significant subcategories within these two main categories. Finally, the ultimate goal would be to accurately predict all the ratings classifications down to the smallest sub-classification.

FIGURE 1

This paper is composed of four sections: Section 1 is devoted to describe the input dataset and the data cleaning and pre-processing phase; Section 2 uses certain visualisation techniques and PCA dimension deduction; Section 3 applies k-means clustering and logistic regression techniques to understand the strength of linear relationships between the data; Section 4 explores the development of the core neural network model architecture showing the training and testing results and the accuracy of the model's predictions.

1. Dataset Description

The dataset used was obtained from publicly available credit ratings and financial stats obtained from S&P's Capital IQ and CRSP databases available through WRDS (Wharton Research Data Services).

For training the model, I used S&P credit ratings obtained from the Capital IQ database. This included 122,418 quarterly, semi, or annual S&P credit ratings of 8,635 US companies since 2017. S&P provides different ratings for short-term, long-term, domestic and foreign currency obligations of each company. I only used the 32,000+ standard long-term ratings "STDLONG" for US companies.

With respect to financial ratios, I obtained them from the CRSP database from amongst 480,738 monthly, quarterly, annual observations of 66 financial ratios for 7,103 US companies since 2017 listed

below in Figure 2. These consist of balance-sheet indexes and ratios and key performance ratios calculated from annual financial reports and company filings. They include indicators for operating performance, liquidity (i.e., ratios used to determine how quickly a company can turn its assets into cash if it is experiencing financial distress or impending bankruptcy), debt and solvency (i.e., ratios that depict how much a company relies upon its debt to fund operations).

1	Dividend Yield	23	Pre-tax Return on Total Earning Assets	45	Total Debt/Total Assets
2	Book/Market	24	Gross Profit/Total Assets	46	Total Debt/Capital
3	Enterprise Value Multiple	25	Common Equity/Invested Capital	47	Total Debt/Equity
4	Price/Operating Earnings (Basic, Excl. EI)	26	Long-term Debt/Invested Capital	48	After-tax Interest Coverage
5	P/E (Diluted, Incl. EI)	27	Total Debt/Invested Capital	49	Interest Coverage Ratio
6	Price/Sales	28	Capitalization Ratio	50	Cash Ratio
7	Price/Cash flow	29	Interest/Average Long-term Debt	51	Quick Ratio (Acid Test)
8	Dividend Payout Ratio	30	Interest/Average Total Debt	52	Current Ratio
9	Net Profit Margin	31	Cash Balance/Total Liabilities	53	Cash Conversion Cycle (Days)
10	Operating Profit Margin Before Depreciation	32	Inventory/Current Assets	54	Inventory Turnover
11	Operating Profit Margin After Depreciation	33	Receivables/Current Assets	55	Asset Turnover
12	Gross Profit Margin	34	Total Debt/Total Assets	56	Receivables Turnover
13	Pre-tax Profit Margin	35	Total Debt/EBITDA	57	Payables Turnover
14	Cash Flow Margin	36	Short-Term Debt/Total Debt	58	Sales/Invested Capital
15	Return on Assets	37	Current Liabilities/Total Liabilities	59	Sales/Stockholders Equity
16	Return on Equity	38	Long-term Debt/Total Liabilities	60	Sales/Working Capital
17	Return on Capital Employed	39	Profit Before Depreciation/Current Liabilities	61	Research and Development/Sales
18	Effective Tax Rate	40	Operating CF/Current Liabilities	62	Avertising Expenses/Sales
19	After-tax Return on Average Common Equity	41	Cash Flow/Total Debt	63	Labor Expenses/Sales
20	After-tax Return on Invested Capital	42	Free Cash Flow/Operating Cash Flow	64	Accruals/Average Assets
21	After-tax Return on Total Stockholders Equity	43	Total Liabilities/Total Tangible Assets	65	Price/Book
22	Pre-tax return on Net Operating Assets	44	Long-term Debt/Book Equity	66	Trailing P/E to Growth (PEG) ratio

FIGURE 2

After cross referencing the two databases and eliminating duplicates, I ended up with a sample of 6,588 observations of rated company financial ratios for 1,022 companies. However, I had significant gaps in how complete the financial ratio data was for each observation, as can be seen from the blue gaps in the heat map in Figure 3 below.

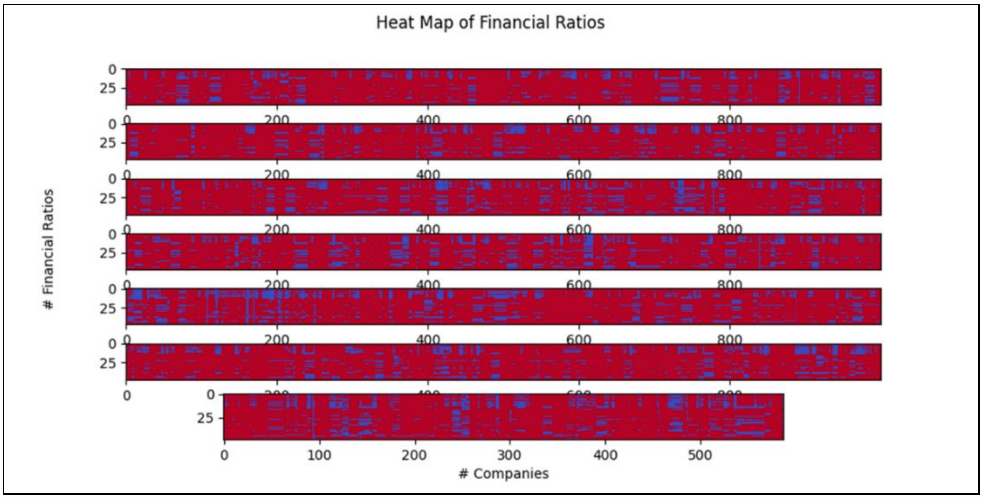

FIGURE 3

If I were to eliminate all observations that had gaps, I would end up with only 855 useful observations. However, many of the financial ratios demonstrated a high degree of covariance with each other (figure 4 below). So, I opted to keep only one financial ratio from each pair (or triplet, etc) that had covariance lower than 0.80, and chose the one with highest number of observations (Figure 5 shows the pairs with highest covariance).

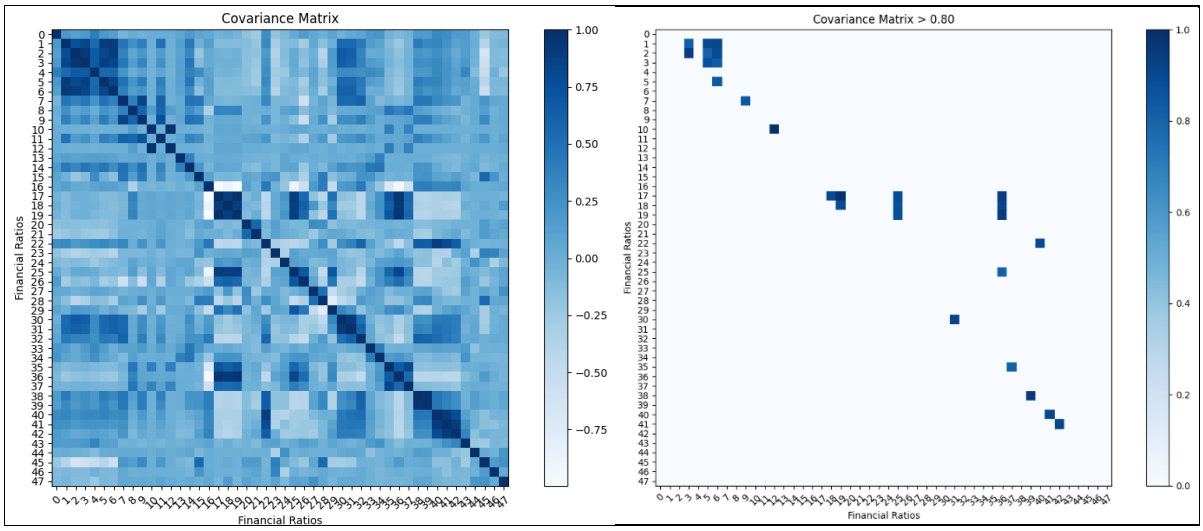

FIGURE 4

FIGURE 5

This enabled me to obtain a useful sample of 2,767 observations with a full set of 32 of the 66 financial ratios (highlighted in yellow in Figure 6 below).

1	Dividend Yield	23	Pre-tax Return on Total Earning Assets	45	Total Debt/Total Assets
2	Book/Market	24	Gross Profit/Total Assets	46	Total Debt/Capital
3	Enterprise Value Multiple	25	Common Equity/Invested Capital	47	Total Debt/Equity
4	Price/Operating Earnings (Basic, Excl. EI)	26	Long-term Debt/Invested Capital	48	After-tax Interest Coverage
5	P/E (Diluted, Incl. EI)	27	Total Debt/Invested Capital	49	Interest Coverage Ratio
6	Price/Sales	28	Capitalization Ratio	50	Cash Ratio
7	Price/Cash flow	29	Interest/Average Long-term Debt	51	Quick Ratio (Acid Test)
8	Dividend Payout Ratio	30	Interest/Average Total Debt	52	Current Ratio
9	Net Profit Margin	31	Cash Balance/Total Liabilities	53	Cash Conversion Cycle (Days)
10	Operating Profit Margin Before Depreciation	32	Inventory/Current Assets	54	Inventory Turnover
11	Operating Profit Margin After Depreciation	33	Receivables/Current Assets	55	Asset Turnover
12	Gross Profit Margin	34	Total Debt/Total Assets	56	Receivables Turnover
13	Pre-tax Profit Margin	35	Total Debt/EBITDA	57	Payables Turnover
14	Cash Flow Margin	36	Short-Term Debt/Total Debt	58	Sales/Invested Capital
15	Return on Assets	37	Current Liabilities/Total Liabilities	59	Sales/Stockholders Equity
16	Return on Equity	38	Long-term Debt/Total Liabilities	60	Sales/Working Capital
17	Return on Capital Employed	39	Profit Before Depreciation/Current Liabilities	61	Research and Development/Sales
18	Effective Tax Rate	40	Operating CF/Current Liabilities	62	Avertising Expenses/Sales
19	After-tax Return on Average Common Equity	41	Cash Flow/Total Debt	63	Labor Expenses/Sales
20	After-tax Return on Invested Capital	42	Free Cash Flow/Operating Cash Flow	64	Accruals/Average Assets
21	After-tax Return on Total Stockholders Equity	43	Total Liabilities/Total Tangible Assets	65	Price/Book
22	Pre-tax return on Net Operating Assets	44	Long-term Debt/Book Equity	66	Trailing P/E to Growth (PEG) ratio

FIGURE 6

The corresponding S&P ratings labels associated with the final dataset obtained showed a significant bias towards BBB and BB ratings over other credit rating categories. This is natural since very few companies are able to maintain over time a very high credit rating, and many that fail to stay in BB/BBB area may quickly deteriorate and go bankrupt. So, there is a degree of selection bias as the rating agencies may stop rating companies that default and/or go bankrupt.

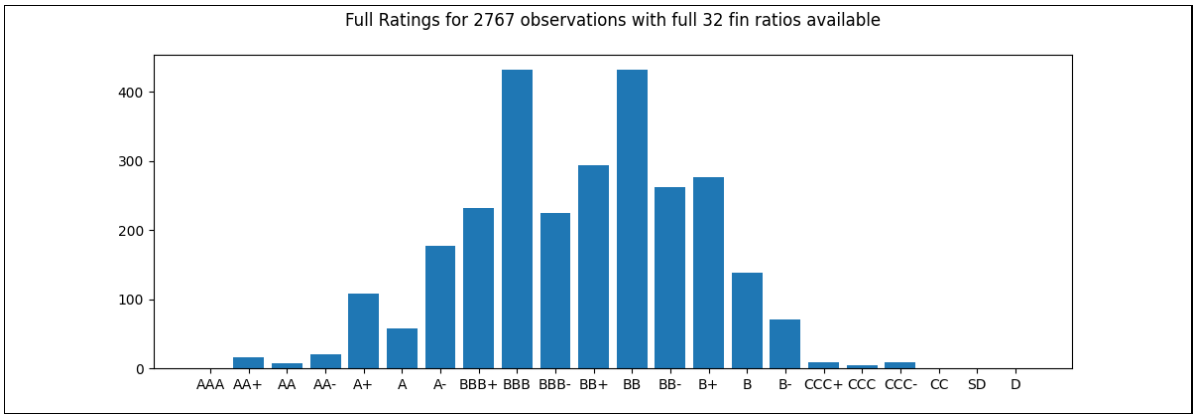

FIGURE 7

Addressed this issue by reducing the classifications to 2 (Investment Grade/Speculative) (Figure 8), which did not have such problem and then tested the model for 3 (Investment Grade, Speculative (BB), Junk (B & lower) (Figure 9) as well as 4 (A & higher, BBB, BB, Junk) (Figure 10). As can be seen from graphs below, significant bias remains for the minority rating classes when consolidating along 3 or 4 classifications. Hajek et al. (2014)[5] addressed this issue by over-sampling the minority classes of ratings in the training dataset. I opted instead to reduce the training set to a size that would allow equal sampling of all consolidated rating classifications and compare the results to the training without adjusting for the sample bias.

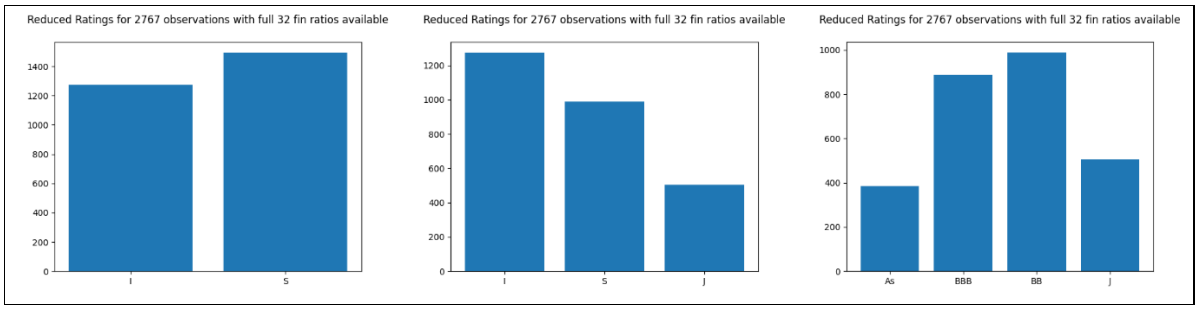

FIGURE 8 FIGURE 9 FIGURE 10

2. Data Visualization & Principal Component Analysis

I performed a principal component analysis to identify the strength of any underlying PCs, try to visualise any clustering in the data, and reduce the dimensionality of the dataset to improve processing time. I then transformed the data by applying the natural log function to the financial ratios and centring them. Figure 11 to the right shows the relative weight of each financial ratio in each PC.

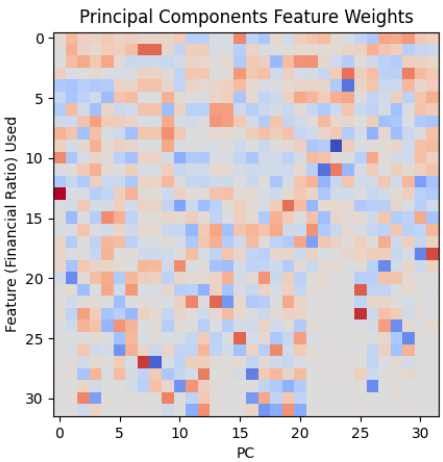

FIGURE 11

The PCA identified that the 1st PC can explain 32.2% of the sample variance, with 8 PCs explaining 90% and **11 PCs explaining 95% of sample variance.**

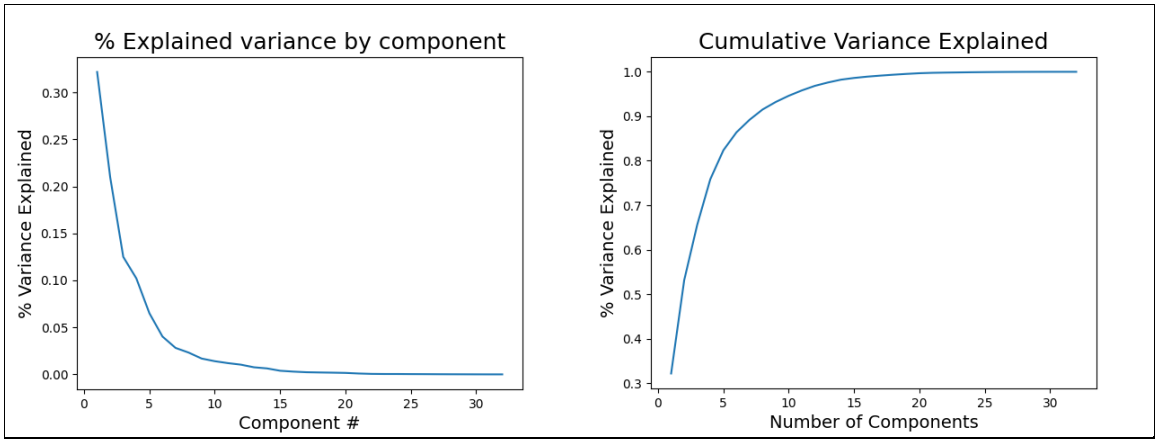

FIGURE 12 FIGURE 13

I tried to visualize the "Ground Truth" of the dataset in Figure 12 below by plotting using different colours for the actual rating labels for 2 classifications (Investment Grade/Speculative) in a scatter plot between the two financial ratios with the highest absolute weights in PC1 and PC2. No clustering patterns exists. I did the same for PC1 vs. PC2 in Figure 14. Again, no obvious clustering pattern emerges.

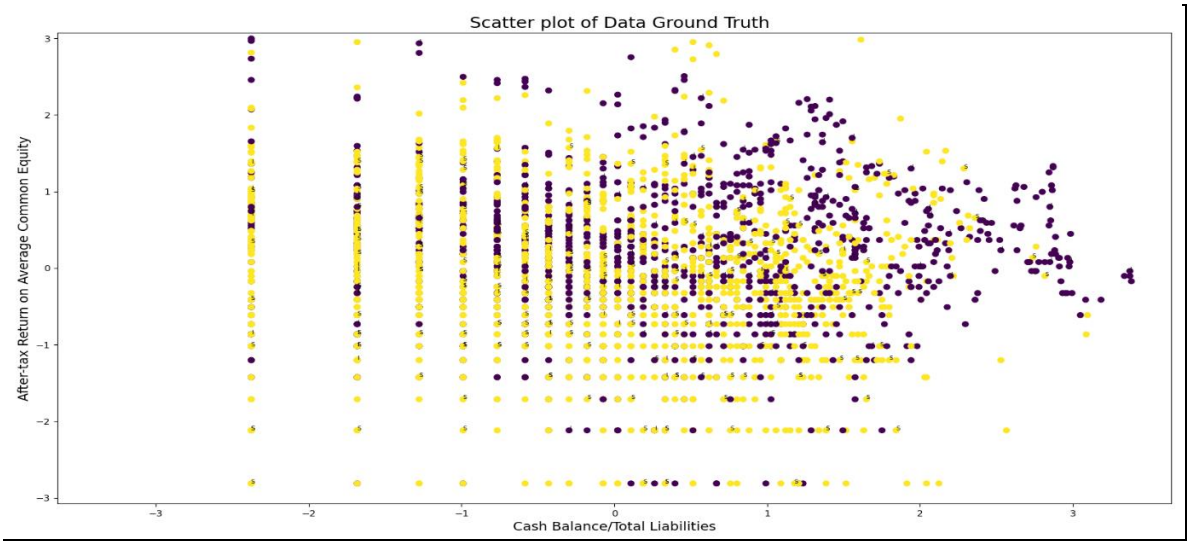

FIGURE 14 (above)

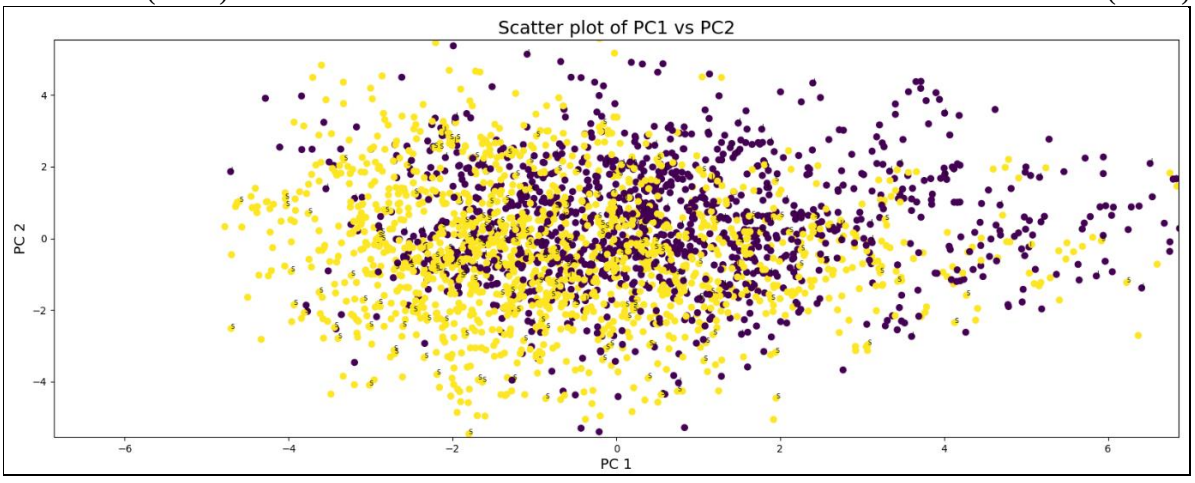

FIGURE 15 (below)

I also attempted to see if t-SNE analysis on the 11 PCs would help visualize any clustering patterns better. t-SNE is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. t-SNE has a # cost function that is not convex, i.e., with different initializations we can get different results. As seen in Figure 16 below, some clustering seems to emerge, but the two classes still overlap significantly.

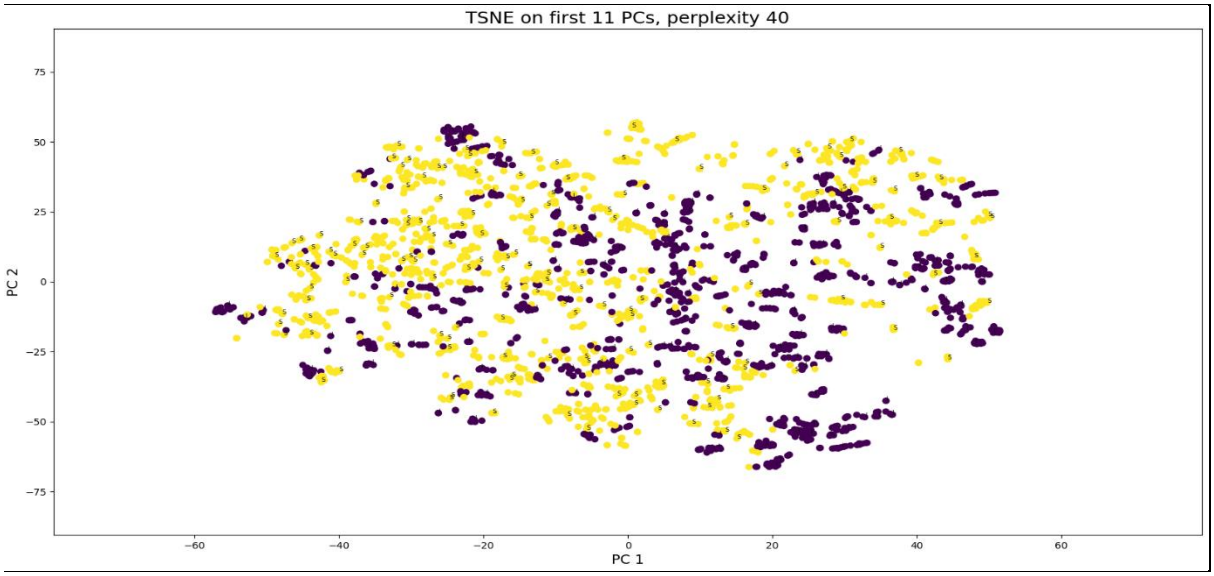

FIGURE 16

I applied Multidimensional Scaling (MDS) to the centred dataset which seeks a low-dimensional representation of the data in which the distances respect well the distances in the original high-dimensional space. In general, MDS is a technique used for analysing similarity or dissimilarity data. It attempts to model similarity or dissimilarity data as distances in a geometric space. As seen in Figure 17, I could not identify a higher degree of clustering.

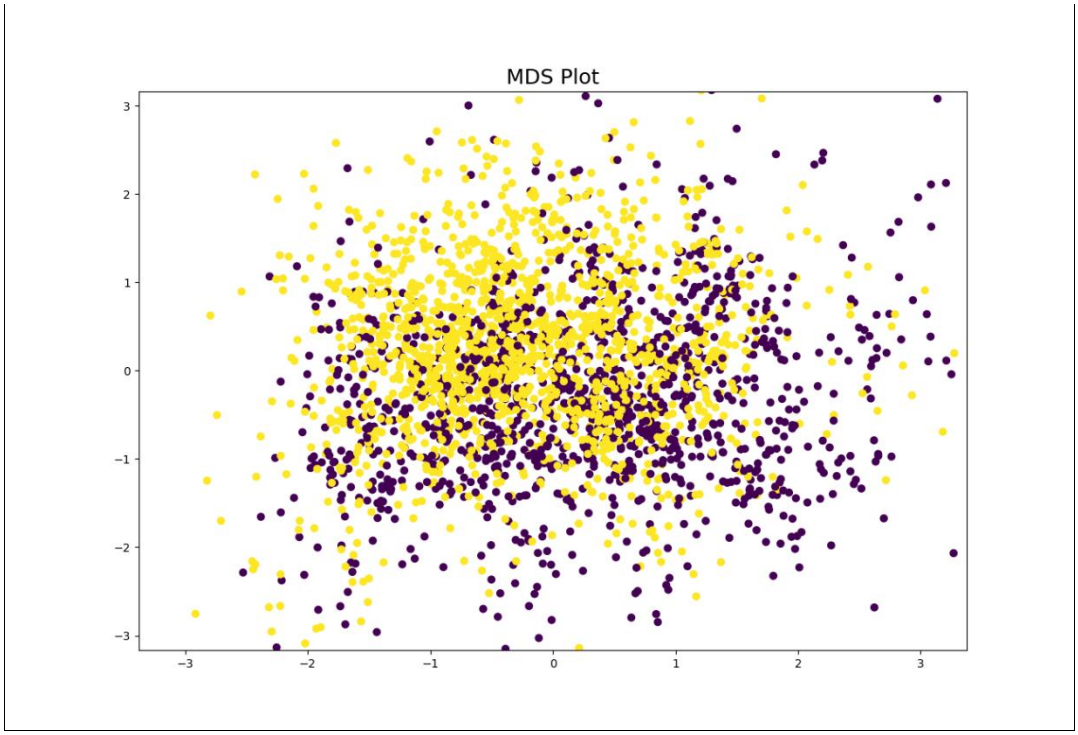

FIGURE 17

3. k-Means Clustering and Logistic Regression Analysis

k-Means Clustering

The k-Means algorithm clusters data by trying to separate samples in n groups of equal variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares where one specifies the number of target clusters. I used the k-means clustering method on both the PCA transformed dataset on full dimension as well as restricting the analysis to the top 11 PCs and could not obtain any significant degree of clustering (Figure 18) or predictability of results in either training or testing sets (Figures 19). The prediction accuracy of k-Means is only 0.61, as can be seen in Figure 19.

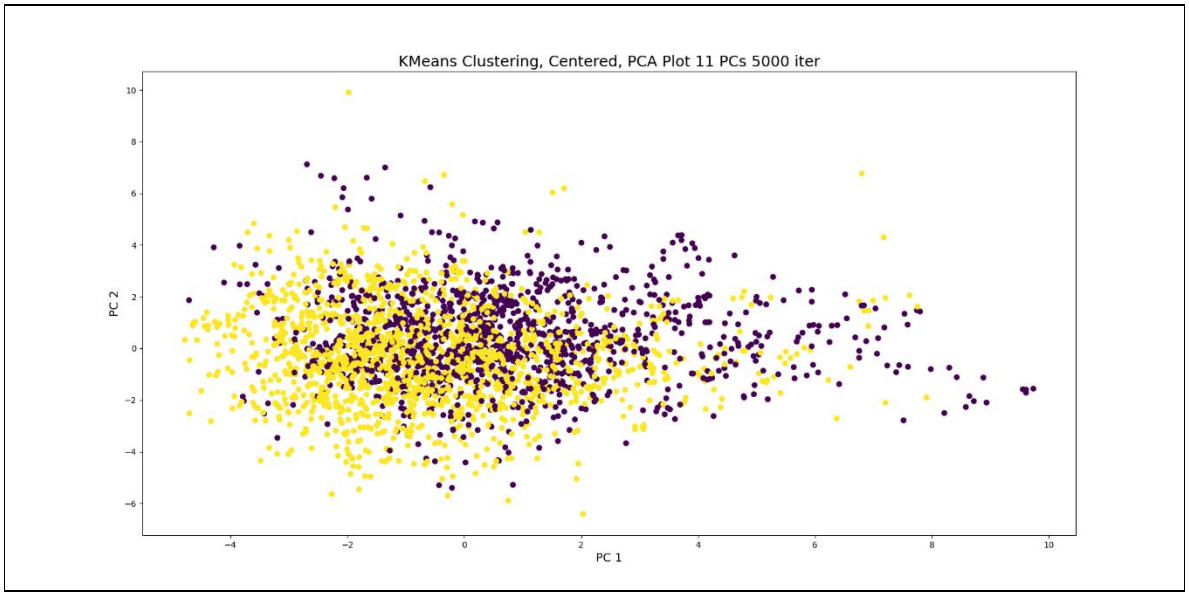

FIGURE 18

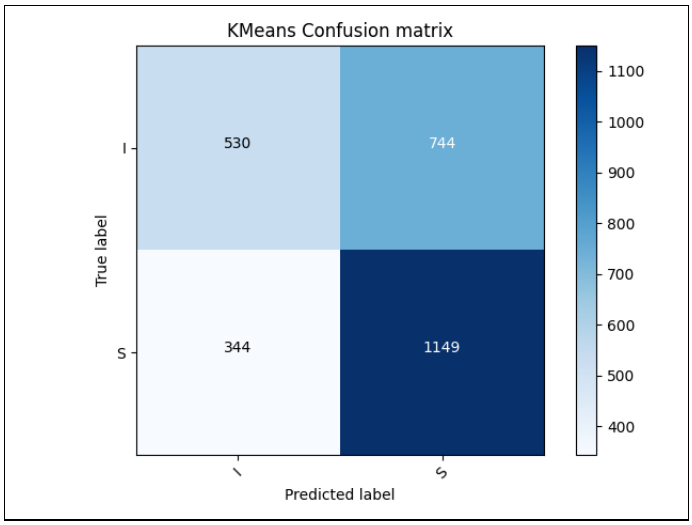

FIGURE 19

Logistic Regression

I used Logistic Regression with a gradient descent algorithm to minimize the minimization of the multi-class cross-entropy with learning rate=0.2 and 2000 maximum iterations. I got Quick Convergence (Figure 20) with a strong 0.814 Training Score (Figure 21a) and 0.801 Testing Score (see Figure 21b)

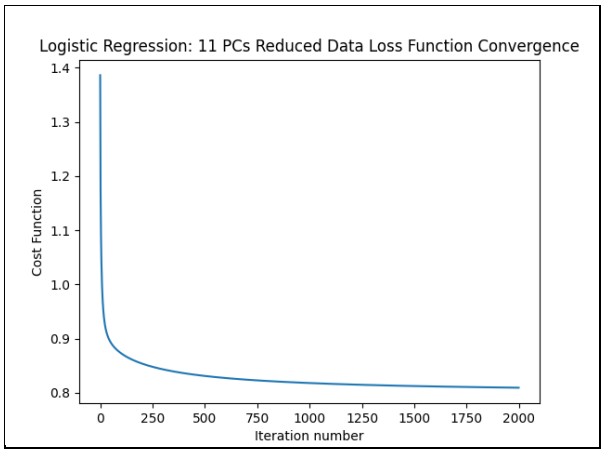

FIGURE 20

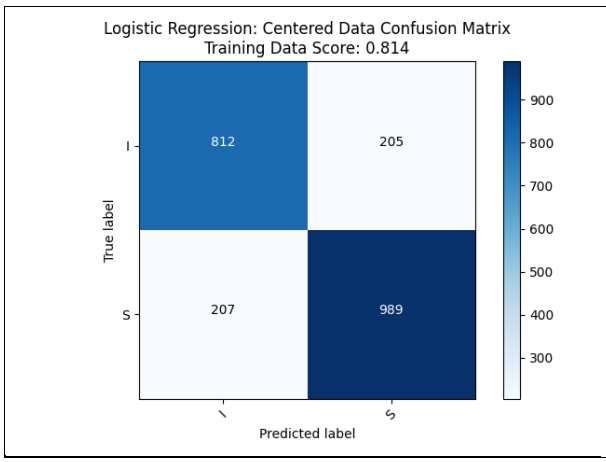

FIGURE 21a

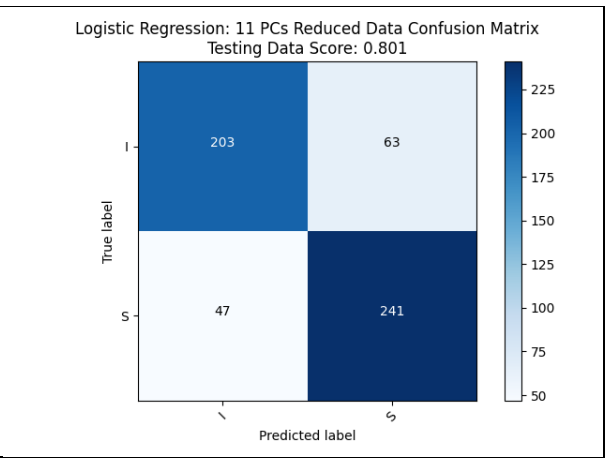

FIGURE 21b

I applied the method to predict 3 ratings classes by (A) ensuring an equal amount of minority classes were sampled in the training set and (B) by using the original sample set without adjusting. I got a decent result with prediction scores but nowhere close to the 0.80 score for 2 rating classes (Figure 21).

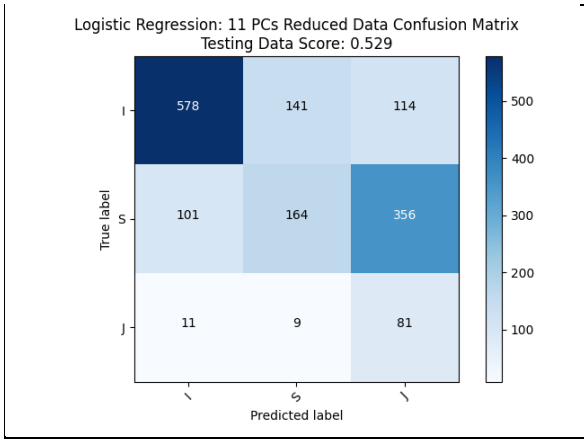

FIGURE 22

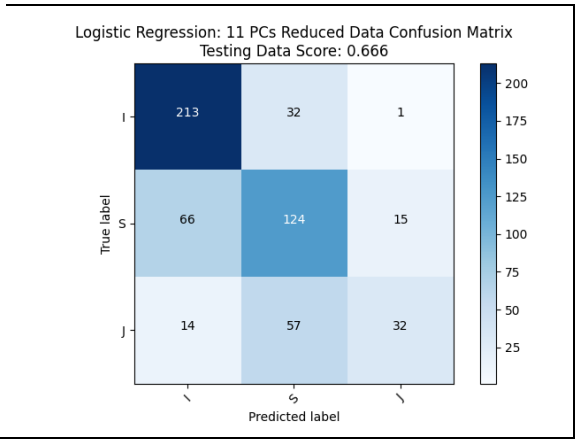

FIGURE 23

It is interesting that the unbiased sampling (A) produces a lower score. I evidenced the same dynamic with neural networks as well. I am inclined to believe that the reduction of the size of the training data set from 2,208 (80% of 2,767) to 840 observations, as a result of ensuring equal numbers of minority classes are sampled, has a much more pronounced effect on the model results than the implied biased when using the full training set. Perhaps also ratings are biased towards the middle rating classes, so using the original data set without such adjustment may be preferable.

I trained the dataset for 4 rating classes, as well, and obtained a weaker testing score of 0.572 (using biased dataset) with most of the divergence is happening around the main convergence axis as can be seen is confusion matrix in Figure 24.

These were promising results and a significant improvement over k-means clustering but still short of the degree of predictability that would make such model practical to use.

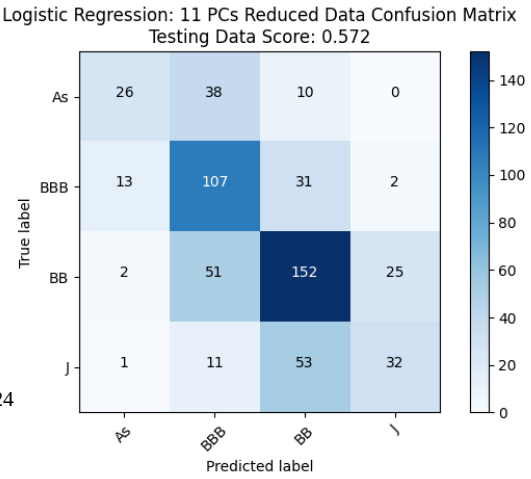

FIGURE 24

4. Developing a Neural Network model

After normalizing the PC adjusted centered data, I applied a simple neural net with 1 hidden dense layer of 16 weights, a tanh() activation function, and a SoftMax activation function for the output layer, so that the result could be interpreted as a probability distribution in order to select the most probable class. I used Adam optimizer and run 2000 epochs for batch size equal to 32. The model produced its highest testing scores if trained for 250 epochs at a score of 0.92 (Figure 25) but with clear signs of overfitting (Figure 26).

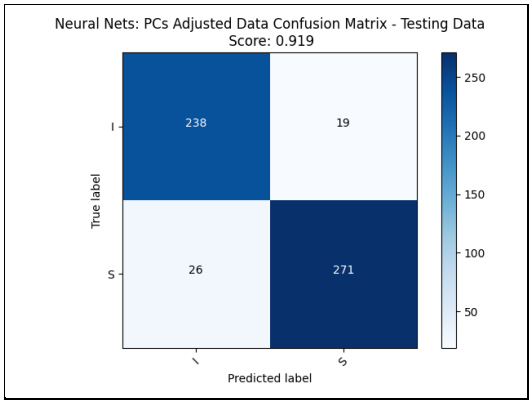

FIGURE 25

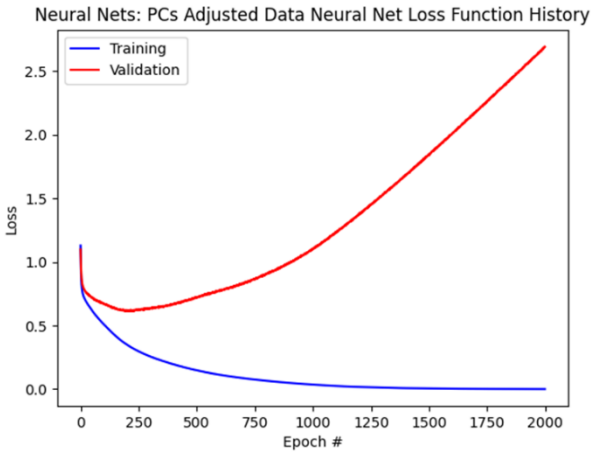

FIGURE 26

When I applied L1 or L2 regularization to the loss function and increased batch size to 64, I was able to overcome the overfitting problem with training converging within 1000 epochs or less (Figure 27) and high prediction score of close to 0.93 (Figure 28). L2 regularization faired a bit better than L1. When using only 11 PCs the testing score may drop to 0.874 (Figure 30) but convergence seem to occur much faster (Figure 29). This is to be expected as the model is trained on a subset of the full dataset gaining speed without losing much accuracy.

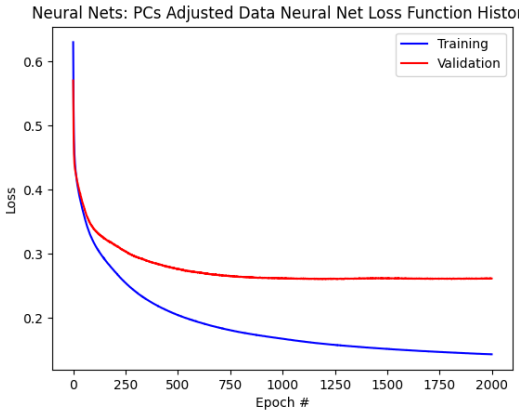

FIGURE 27

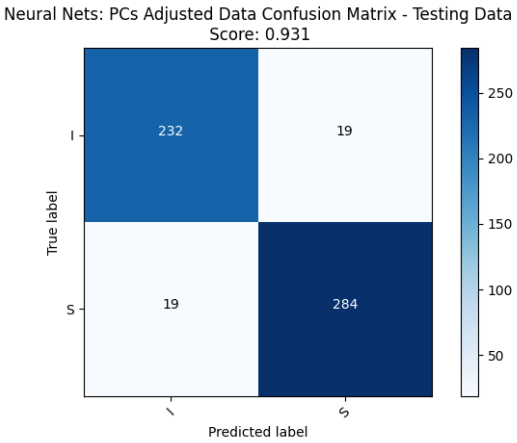

FIGURE 28

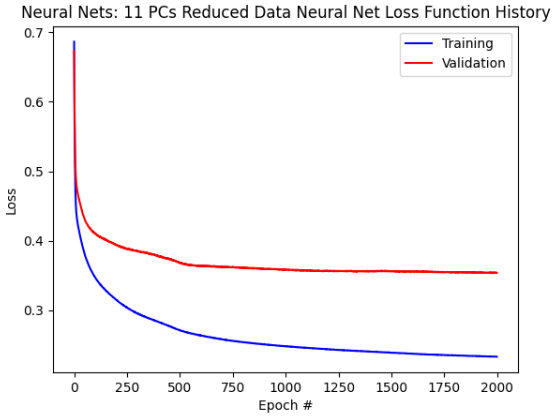

FIGURE 29

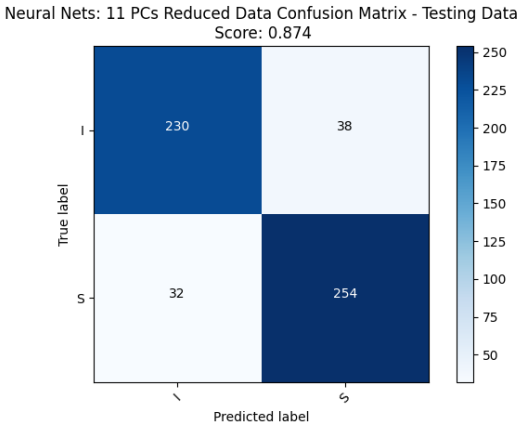

FIGURE 30

I tried improving the prediction scores by including an additional hidden dense layer with eight (8) weights and a dropout layer with an 20% drop-out rate (to avoid overfitting). The model's increased complexity did not improve the accuracy. Tried to increase/decrease the number of weights in the first layers but also this did not improve accuracy. So, the best accuracy remained this relatively simple neural network model with one hidden dense layer.

I then trained the model for 3 rating classifications and obtained good convergence and 0.83 accuracy score (Figure 32).

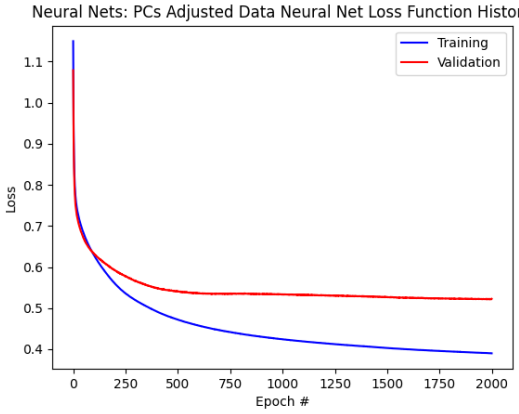

FIGURE 31

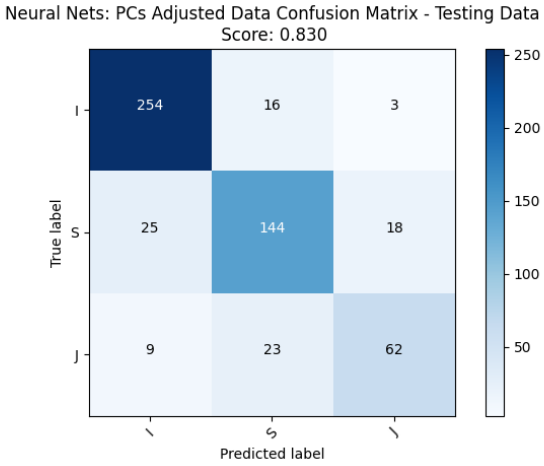

FIGURE 32

Finally, I trained the model for 4 rating classifications and obtained good convergence and 0.767 accuracy score (Figure 34).

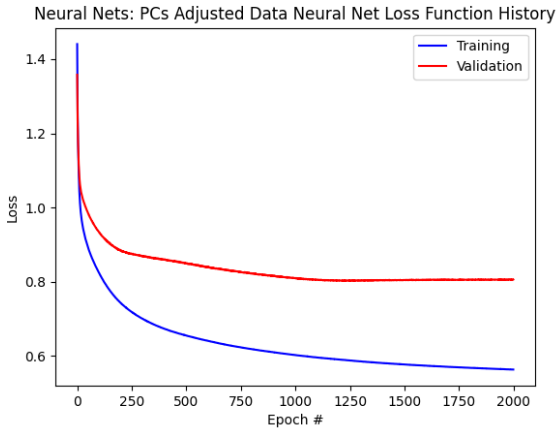

FIGURE 33

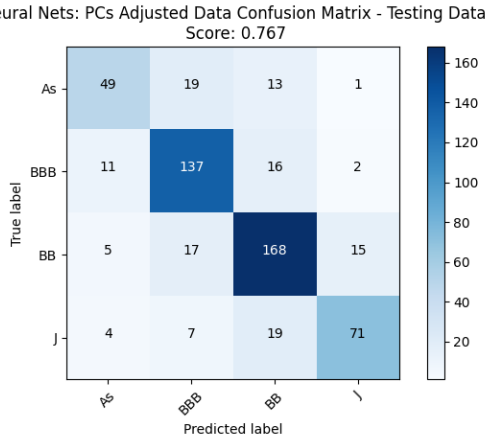

FIGURE 34

Conclusion

Starting with relatively low-quality datasets of financial ratios for 1,022 US companies, and using the corresponding S&P long-term credit ratings for each company and financial ratio reporting date, I was able to create a neural net-based model that predicts up to four broad categories of S&P long-term credit rating with a high degree of accuracy.

Despite being a relatively simple neural network, the model is able to:

- Differentiate whether a company is investment grade or speculative grade with close to 0.93 accuracy, and
- Distinguish within the Speculative class, the BBs from B or lower with 0.83 accuracy score, and
- Overall differentiate between As (AAA/AA/A), BBB, BB, and B or lower with 0.77 accuracy.

The quantity and quality of data would be a key area of improvement in order to obtain better results. Perhaps purchasing full sets of financial ratio data from professional financial data providers would be appropriate. Finally, I would recommend testing whether any COVID-19 effect should be accounted for years 2019 and 2020 where companies' financial ratios may have deteriorated without the rating agencies downgrading them due to the perceived temporary effect of this crisis. Perhaps a future improvement is to introduce a COVID-19 flag parameter for the respective reporting dates.

An advantage of neural networks is that we can extend the model to include macro-economic information as well as alternative non-financial data, even unstructured data such as sentiment data, twitter/Bloomberg chats, customer/analyst reports and ratings, etc... that may predict better the credit quality of borrowers than the credit rating agencies. Such models could enable investors to extend credit to a wider universe of borrowers that currently may not have access to credit due to lack of traditional financial data as well as reduce exposure to riskier borrowers for where traditional credit models fail to capture their deteriorating condition.

Endnotes

[1] Parisa Golbayania, Ionuţ Florescub, Rupak Chatterjeec "A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees North American Journal of Economics and Finance 54 (2020) 101251, pp.

[2] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring," Expert Systems with Applications, vol. 42, no. 7, pp. 3508-3516, 2015.

[3] Mercep, A.; Mrcela, L.;Birov, M.; Kostanjcar, Z. Deep Neural Networks for Behavioral Credit Rating. Entropy 2021, 23, 27. https://doi.org/10.3390/e23010027

[4] A. R. Provenzano, D. Triro, A. Datteo, L. Giada, N. Jean, A. Riciputi,G. Le Pera, M. Spadaccino, L. Massaron and C. Nordio, "Machine Learning approach for Credit Scoring", Working paper, arXiv:2008.01687v1 [q-fin.ST] 20 Jul 2020

[5] Petr Hájek and Vladimír Olej, "Predicting Firms' Credit Ratings Using Ensembles of Artificial Immune Systems and Machine Learning – An Over-Sampling Approach", AIAI 2014, IFIP AICT 436, pp. 29–38, 201