1. Design a 3-layer feedforward neural network consists of an input layer, a hidden-layer of 10 neurons having ReLU activation functions, and a linear output layer. Use mini-batch gradient descent with a batch size = 8, 𝐿2regularization at weight decay parameter 𝛽=10−3 and a learning rate 𝛼=10−3 to train the network.

* 1. a) Use the train dataset to train the model and plot both the train and test errors against epochs.
  2. b) State the approximate number of epochs where the test error is minimum and use it to stop training.
  3. c) Plot the predicted values and target values for any 50 test samples.

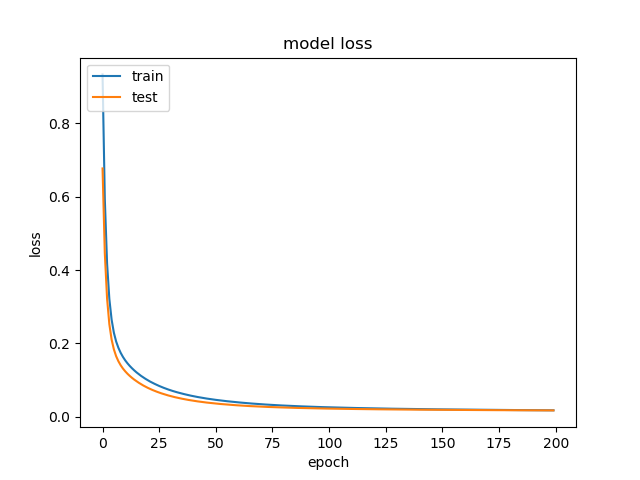


Figure 1a

1. Based on figure 1a, the optimal number of epochs is 100.

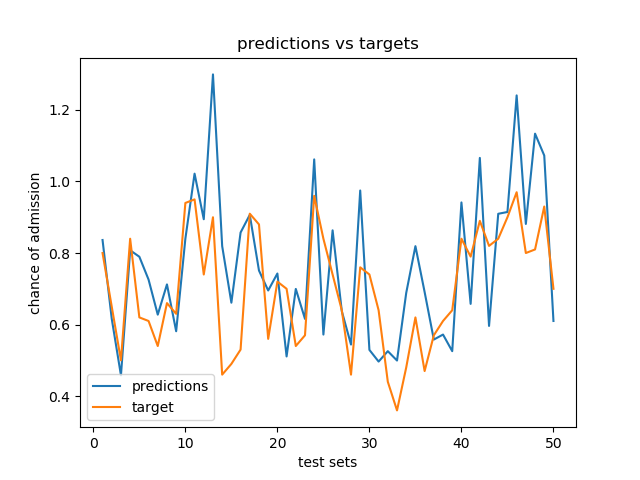
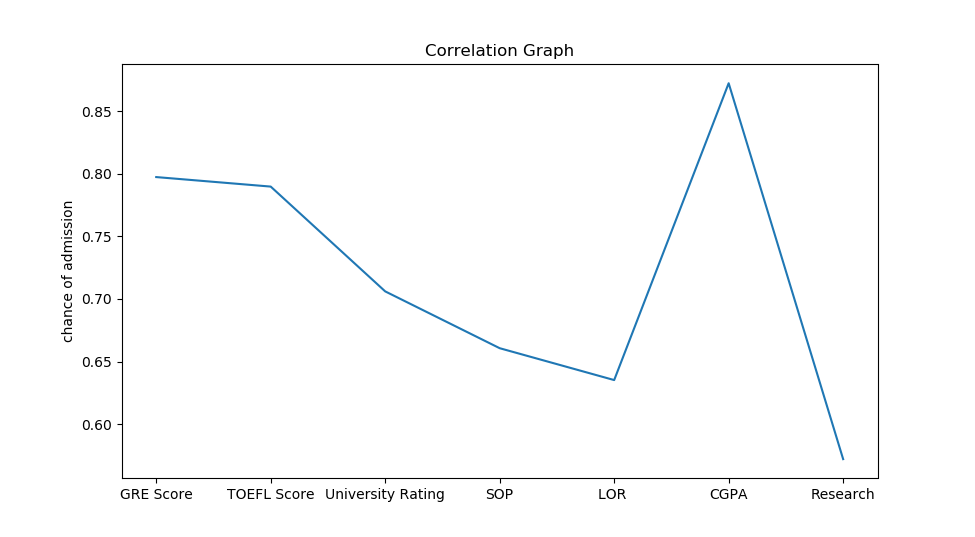


Figure 1c

* 1. 2. Use the train data to compute (and plot) an 8X8 correlation matrix between the different feature scores and the corresponding chances of admit.
  2. a) Which features are most correlated to each other? Is it justifiable?
  3. b) What features have the highest correlations with the chances of admit?

1. Correlation Matrix

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of admission |
| GRE Score | 1 | 0.820288 | 0.63166 | 0.565743 | 0.499807 | 0.811529 | 0.578914 | 0.797415 |
| TOEFL Score | 0.820288 | 1 | 0.657618 | 0.622333 | 0.53029 | 0.807247 | 0.474875 | 0.789813 |
| University Rating | 0.63166 | 0.657618 | 1 | 0.729212 | 0.604471 | 0.722389 | 0.459666 | 0.706071 |
| SOP | 0.565743 | 0.622333 | 0.729212 | 1 | 0.660102 | 0.706401 | 0.400011 | 0.660664 |
| LOR | 0.499807 | 0.53029 | 0.604471 | 0.660102 | 1 | 0.63742 | 0.382613 | 0.635199 |
| CGPA | 0.811529 | 0.807247 | 0.722389 | 0.706401 | 0.63742 | 1 | 0.506941 | 0.872403 |
| Research | 0.578914 | 0.474875 | 0.459666 | 0.400011 | 0.382613 | 0.506941 | 1 | 0.572018 |
| Chance of admission | 0.797415 | 0.789813 | 0.706071 | 0.660664 | 0.635199 | 0.872403 | 0.572018 | 1 |



1. Examinations-based features and chance of admission have high correlation (>0.78). It is justifiable because examinations require a lot of effort & are very competitive. Therefore, applications with high examination scores are selected.

Examinations-based features have a high correlation to each other (>0.80). It is justifiable because if a student is good at one examination, he should also be good at other examinations.

1. Examinations-based features like GRE Score, TOEFL Score & CGPA.

3. Recursive feature elimination (RFE) is a feature selection method that removes unnecessary features from the inputs. Start by removing one input feature the causes the minimum drop (or maximum improvement) in performance. Repeat the procedure recursively on the reduced input set until the optimal number of input features is reached. Remove the features one at a time. Compare the accuracy of the model with all input features, with models using 6 input features and 5 input features selected using RFE. Comment on the observations.

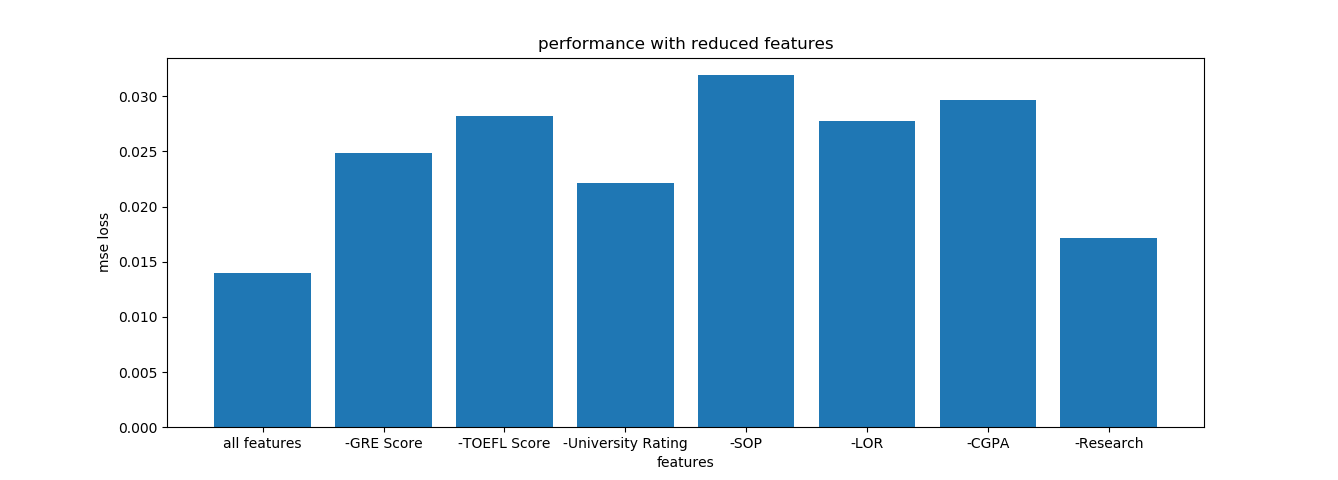
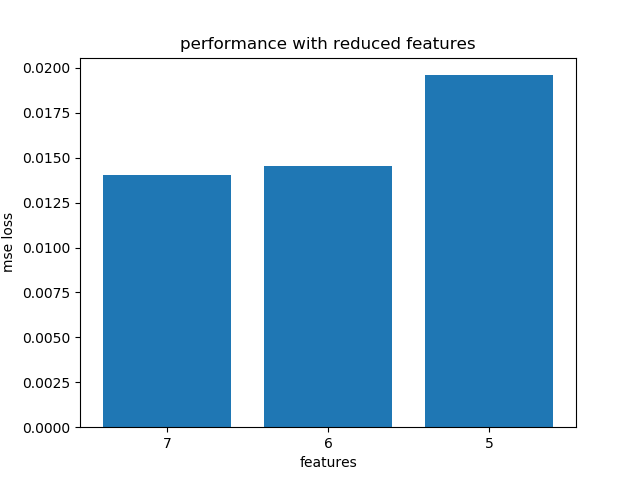


Figure 3

The reduced models with the best performance are “-Research” followed by “-University Rating”. This coincides with Research and University Rating having low correlation as mentioned in Question 2. However, removing SOP and LOR also caused a big increase in loss despite having one of the least correlations.

Instead, what is common between Research & University Rating is that they have the fewest possible input space. Research is a binary input & University Rating is an integer between 1 to 5. Meanwhile SOP and LOR have twice as much input space as University Rating as they are a multiple of 0.5 between 1 to 5. This suggests that useful features have bigger input space and that input space is more important than correlation.

Therefore, our 6-feature model will remove Research feature & our 5-feature model will remove both Research and University Rating feature. After training the models, the performance for each model is shown below.



From the above figure, the model with the best performance has 7 features. However, 6-features model has comparable results. This suggests that Research is not important for admission.

Finally, the optimal feature set includes all 7 features.

4. Design a four-layer neural network and a five-layer neural network, with the hidden layers having 50 neurons each. Use a learning rate of 10-3 for all layers and optimal feature set selected in part (3).

Introduce dropouts (with a keep probability of 0.8) to the layers and report the accuracies. Compare the performances of all the networks (with and without dropouts) with each other and with the 3-layer network.

We have trained 4 models for this question, where each model has 7 features. The figure below shows their performance according to ‘mean square error loss’. Note that the 3-layer network have 10 neurons in the hidden layer.

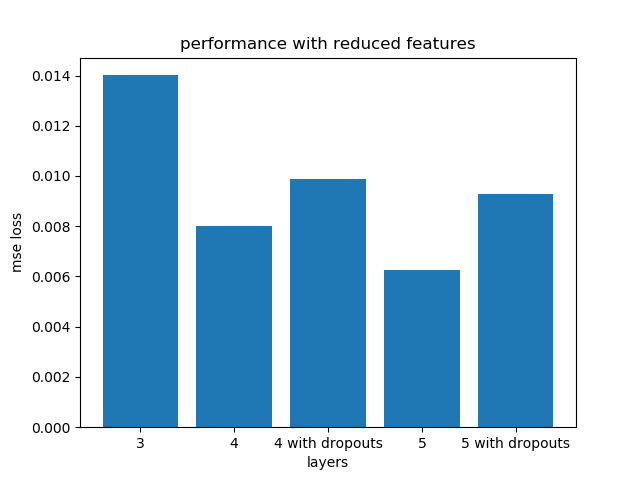


Figure 4a

The above figure shows the performance of models with different number of layers. Models with more layers have lower ‘mse loss’ and better performance. Unfortunately, models with dropouts have a lower performance despite reducing overfitting. We will attempt to explain why in the next section.

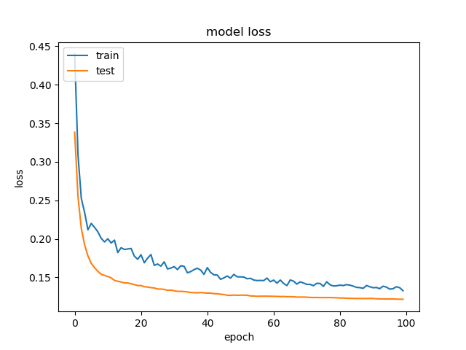
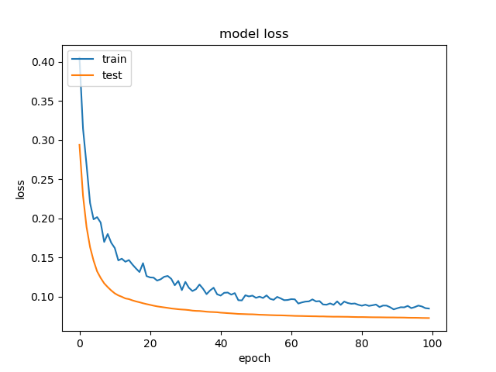


Figure 4b & 4c

Figure 4b and 4c shows the training and test loss during the training of the model. Compared to figure 1a, we can see that train loss has not reached its minimum. As such, this suggests that training is incomplete, and we must increase the number of epochs during training. We decided to try an epoch of 1000 instead of 100

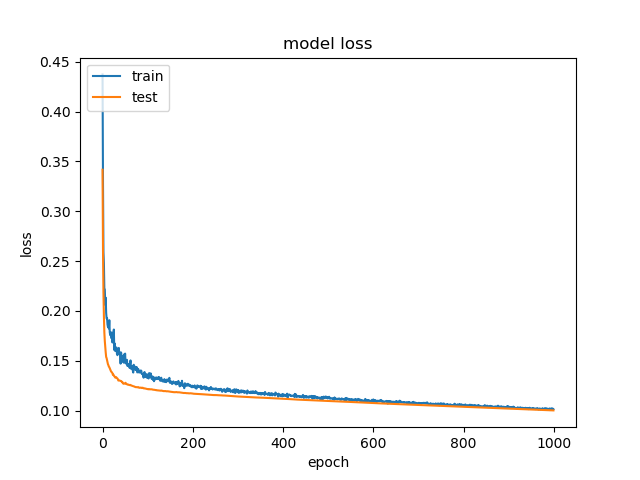
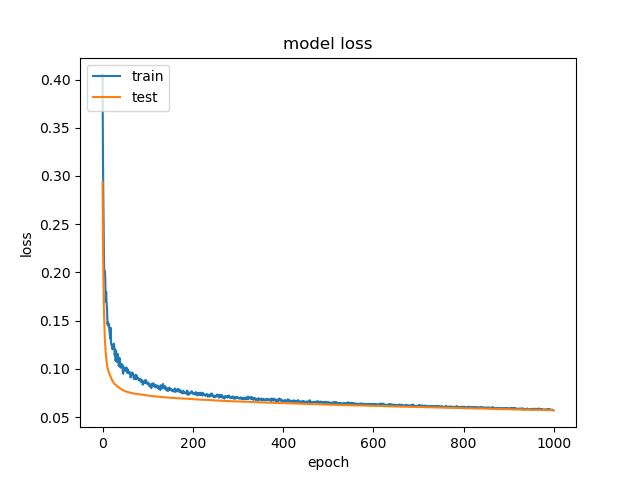


Figure 4d & 4c

Figure 4d & 4d shows the new training and test loss during training. We can see that loss has reached a minimum.

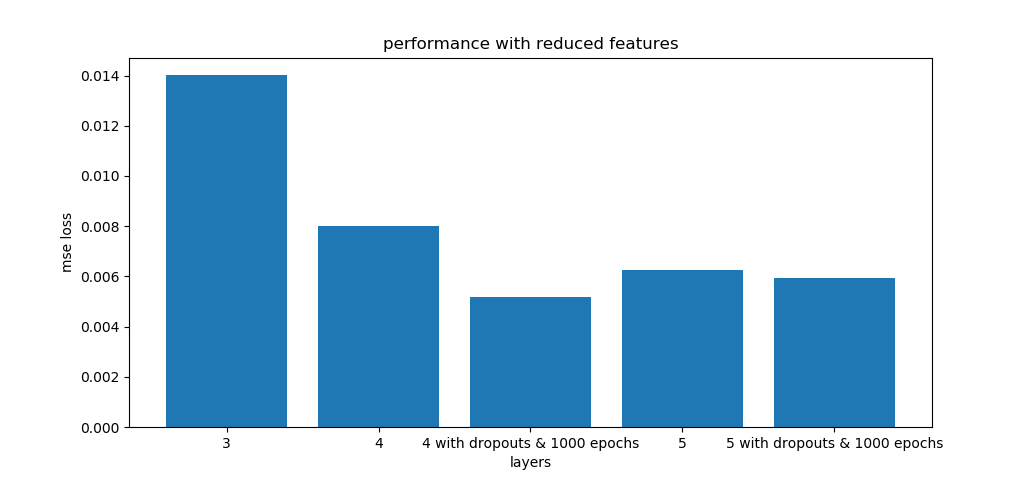


Figure 4e

Finally, Figure 4e shows the new comparison graph between models. Our 3,4,5-layer models without dropouts are unchanged from figure 4a. We can see that 4-layer model with dropouts and finished training has the best performance. Therefore, we must increase the number of epochs when using dropouts to ensure a thorough training. We should also note that trained models with 1000 epochs have slightly better performance than trained models with 100 epochs, but we decided to overlook this difference.

In conclusion, we should use a 4-layer model with 7 features & dropouts to predict admission chances.