Student Name: Yifei Yu
Student ID: 3342446
Stream: MAM2020

Customer Churn

1. **Why did we remove "customerID"?**
customerID column is a unique classifier that the neural network might use to perfectly fit the training data without using any other column, if the model has enough capacity. It needs to be removed for the model to actually generalise patterns with all information available.

2. **Handle missing values**

a. **Why and how deleting observations with null values affect the model**
Deleting observations with null values removes all information about these observations from the model. The benefit is not to have to deal with missing values but the cost is to suffer from potential information loss with fewer observations. Another risk is to biase the model because the cause of missing data is systematic other than random.

b. **Explore different techniques of handling missing values**
Missing values, in the case of some K-nearest neighbours and tree-based models, don't have to be handled at all because these methods are robust to missing values.
In case of otherwise, generally missing values are either deleted or imputed depending on circumstances. When data is big enough and the cause of missing data is random, we can delete missing data without biasing the model.

Imputation is available when data is small or the cause of missing data is unknown. A few different ways of imputing data exist. We can simply replace missing values with a summary statistic (i.e. mean, mode or median) from their columns. Alternatively, we can draw random values from other non-missing values in the same column as the missing data as proxies for the missing data. Additionally, missing data can be replaced by output from another predictive algorithms.

c. **Implement one of these techniques and determine if it has an effect on the overall accuracy of your model**
After imputing missing values in 'TotalCharges' with its mean, the test accuracy seems to have improved, however, this is because the mean includes information about test data. The test accuracy is falsely inflated.

3. **Why did we log transform the 'TotalCharges' variable?**
We log transform the variable to put it into a smaller scale.

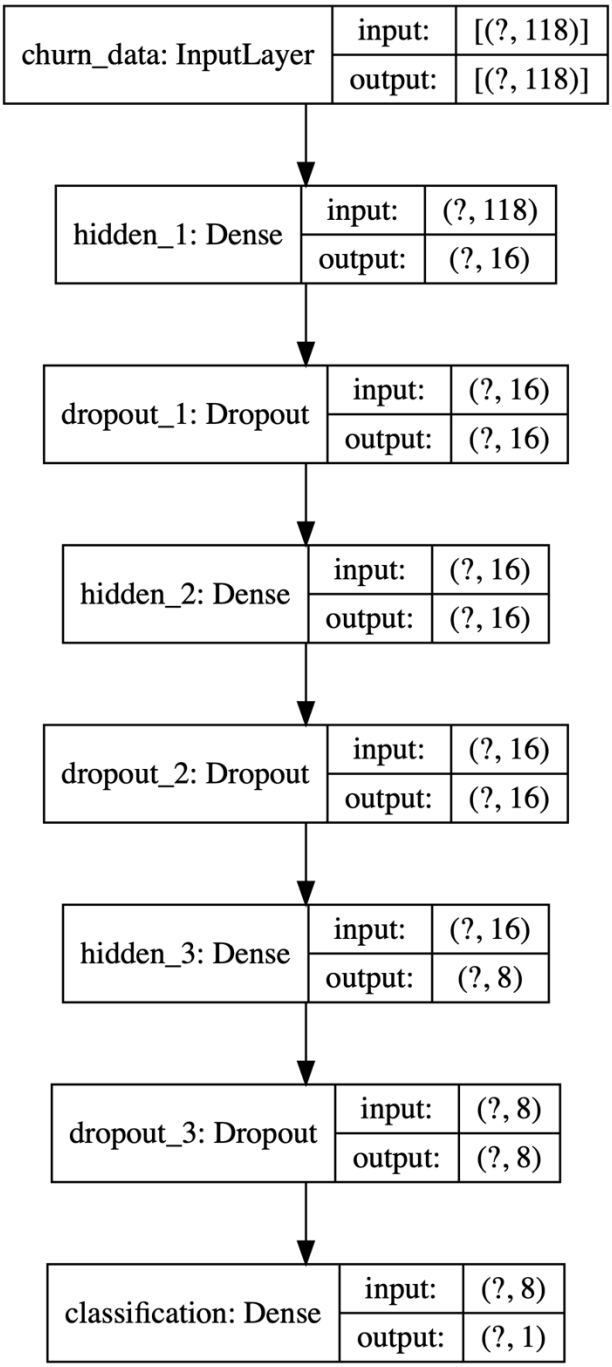
4. **Why did we normalise the data?**
The data is centered and scaled for faster computations. Otherwise the model has to estimate a different distribution for every layer as the input data change.

5. **What should we use for multiclass classification problem?**

For problems where the label has multiple levels, we should use categorical crossentropy loss function.

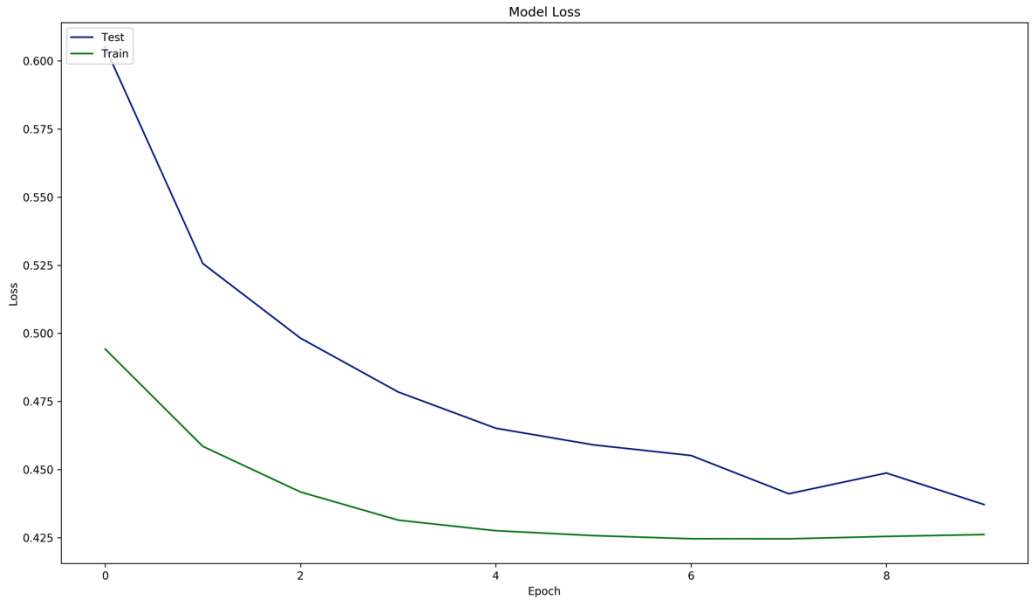
6. **Perform hyperparameter tuning to better understand the performance of your model**

a. **Present the architecture and parameters of your best possible model**



The overall architecture consists of three dense layers and three dropout layers as well as one final dense layer with one sigmoid computational unit to estimate probability.
The hyperparameters are 20% dropout rate, 64 batch size and 10 epochs.

b. **Does your model suffer from underfitting or overfitting? Discuss using plots.**



Neither. Both training loss and testing loss dropped to a low level. The complexity and the regularisation effort paid off.

c. **Inspect the performance with all the metrics used in class.**
Accuracy on test data is around 78% as calculated by the code below.
ROC-AUC score is about 82%.

7. **Check the performance with other classifiers you know**
Logistic regression achieves a higher classification accuracy than the neural network. This suggests limited complexity of the relationship within the data which requires only simpler methods such as logistic regression.

GE Stock Price Prediction Case

8. **Stock price prediction**

a. **Use any combination (at least three) of features to predict closing value of the upcoming day**

The performance did improve after adding more features as demonstrated in the code. The test loss with three features is 0.2485. The test loss is lower after adding more features.

b. **Use RNN with one or any combinations of features to predict the closing value of one upcoming day. Did the performance increase after adding more features?**

The performance also improved after adding more features.

c. **Perform hyperparameter tuning to improve the performance of your LSTM model**

I increased the number of timesteps (previous days) and added another LSTM layer with 16 units before the linear Dense layer in the end. The model performance increased as a result.

I increased the number of timesteps from 4 to 10. The model performance also increased as a result.

d. **Perform hyperparameter tuning to improve the performance of your RNN model**

The performance also improved after adding one more layer or increasing the number of timesteps.

e. **Predict the closing values for 2 upcoming days using feature values from 8 previous days**

Multiple timesteps can be predicted but the further away from the range of historical data, the worse the prediction accuracy is.

f. **Discuss in half A4 page the limitations of deep learning on predicting stock market**

Although neural network is powerful at approximating the statistical relationshiop between historical financial data and future financial data, usually future stock prices are correlated with non-financial factors that are not included in the dataset.
Another issue is the availibility of data. Predicting daily stock prices is much easier than predicting monthly or annual stock prices because there is much less monthly and yearly financial data.