Final Project Report **CS 4275: Machine Learning Foundation**

Objective

As discussed in the proposal for the final project, the main objective was to build and train LSTM-based models in analysis and pattern prediction of health care datasets. The model is regression based on LSTM to predict billing amount from a health care dataset. The life cycle for machine learning is the same in this project as well. The dataset was preprocessed so it can handle categorical data as well as date and numerical data were scaled for compatibility with LSTM models.

Dataset Preprocessing

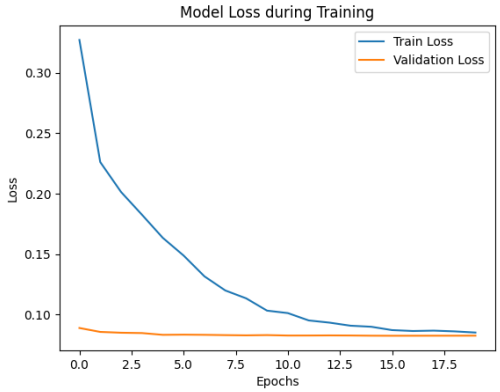
Dataset was preprocessed to handle categorical, numerical and date related data which were then scaled for LSTM models compatibility. First thing was to handle missing values, if there were any missing values in the dataset, they were replaced with column-wise means. Categorical columns were then label encoded. The date columns were then converted to year, month and day components to fit the float type and original date columns were dropped as well. The numerical data were scaled using Min-Max normalization. Few columns were dropped from the dataset which were excluded from the model training. The target column is billing amount. The remaining features were then shaped into a sequence of 20 to feed into LSTM model and target values were also adjusted for the sequence alignment.

Configuration

The model was configured with an input layer of 64 units. Dropout was set to 20% to control overfitting. Dense layer was set to 1 unit for regression output. Adam optimizer was used with the learning rate of 0.0005.

Training Process

```
Epoch 19/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 11ms/step - loss: 0.0858 - val_loss: 0.0825
Epoch 20/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 11ms/step - loss: 0.0848 - val_loss: 0.0825
63/63 ━━━━━━━━━━━━━━━━━━━━ 1s 6ms/step
Mean Absolute Error (MAE): 0.24859581887722015
Mean Squared Error (MSE): 0.08252758532762527
R-squared (R²): -0.0007128098640547531
63/63 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 0.0832
Test Loss (MSE): 0.08252757042646408
```



The training epochs are shown in the picture above. The validation loss and training loss are used for the performance evaluation of the model. The validation loss stabilizes around

epoch 8. And the loss decreases consistently meaning the model generalizes well when learning on the training set.

Test Results

The model was tested on a test set, which yielded the following results:

Mean Absolute Error: 0.248

Mean Squared Error: 0.0825

R-squared: -0.00071

The test loss matches the validation loss, which could indicate that the model generalizes on the test set similarly to the training set. The R-squared is negative, which means the model performs slightly worse than the expecting the mean of the target variable.

Observations

The consistent decrease in training and validation loss indicates effective learning. The Mean Absolute Error of 0.248 indicates that model's average prediction error is around 25% of the target variable's normalized scale. The negative R-squared, even though is closer to 0, shoes that the model doesn't capture the variability in the target variable effectively. This indicates that the model is slightly worse than predicting the meaning of the target variable. Adding epochs to 50 and learning rate to 0.01, R^2 tends to 0.

Conclusion

The model demonstrates decent learning through the convergence of training and validation loss. The prediction power is limited, as indicated by the negative variance score. Increasing model complexity, enhancing feature engineering, and doing thorough hyperparameter optimization could all lead to better results.