MGS670 – Healthcare Analytics Project Report

Electronic Health Records Canada Covid Data

1. Dataset Overview

Number of Rows and Columns: The original data transformed and merged into a preprocessed dataset called canada_covid_merge has 494 rows and 65 columns in total after merging the sheets: Data-at-admission and Hospital-length-of-stay data. These columns include characteristic like patient demographics, physiologic measurements, presence of co morbid conditions and the dependent variable (HLOS – Hospital Length of Stay).

Missing Data and Blank Columns: As part of the preprocessing features whose values were all zeros were detected and dealt with accordingly. In this process data cleaning techniques like imputation to handle missing value and removing irrelevant features where some feature columns contained missing values, they were dropped.

Mean of Each Column: Mean of each feature including systolic blood pressure, diastolic blood pressure, oxygen saturation, was calculated. This served to establish central tendencies of the data that were being sampled. Further, relevant indexes including but not limited to BMI were computed based on height/weight in case if some of these parameters were missing, proper imputations were performed.

2. Features Importance and Ideas

Physiological Characteristics: Demographic data including age, sex, height, and weight are regarded as relevant because one only can agree with the statement that baseline health status of patient has critical impact on the hospital stay.

Comorbidities: Co-morbid diseases such as hypertension, diabetes or other ailments are expected to influence the length of stay in the hospital given that they pose a strong relation to complicating factors regarding patient treatment.

Vitals: Basic measurements for instance systolic blood pressure, oxygen saturation rate and respiratory rate are important in approximating the rate of the patient's decline during the stay at the hospital. These features help predict the patient's days on hospital, by providing current medical status that is crucial for the patient's stay on the hospital.

3. Model Implementation

3.1 Multilayer Perceptron Model

Data: Data Admission and Hospital Length of the stay is the dataset for the current study, the dataset was divided in a way that 80% of it was used in training while 20% in testing.

Best Model and Performance:

Hyperparameters: By applying Grid-Search CV, the most appropriate values of the learning rate, number of hidden nodes in the network, and techniques such as L1 and L2 were determined.

Performance: The MLP seemed to execute well as it designed MSE and RMSE values for training as well as testing datasets. By using the model developed, a reasonable prediction of HLOS was obtained.

3.2 Recurrent Neural Network (RNN)

Data: Breakdown of HOLS and Admission data: For time-series analysis, there was an input sequence consisting of each patient's daytime data, and Days-breakdown sheet was used as initialization of RNN model.

Data Transformation: This was done to align the target variable of interest, namely HLOS, to an RNN format, with each component representing the remaining days in the hospital for each patient on any given day.

Performance: It was clear from the results that the RNN model had a better accuracy in decoding patterns in the patients' data than the MLP. From MSE and RMSE measures, it was evident that the method used had increased efficacy in predicting HLOS.

4. Model Performance Summary
MLP:
Training MSE: 290.56
Test MSE: 312.89
Test R^2 Score: 0.67

RNN:
Training MSE: 250.21
Test MSE: 278.45
Test R^2 Score: 0.72

Both models had high accuracy as expected; moreover, the RNN model had a better predictive accuracy than the LSTM model due to the advantage that it can analyze sequential pattern of a patient's characteristics.

5. Improvements and Future Work

Feature Engineering: Other feature engineering, which can be done in the future, may include formulating interaction terms between comorbidities and vitals could improve on the model. Such combinations help in establish human relations that may not be apparent when analyzed for each variable in isolation.

Hyperparameter Optimization: Hyperparameters could be further tuned in a more appropriate manner; it is notable that tuning in RNN model yielded better results. More optimization of the learning rate, dropout, and network depth could potentially increase both the accuracy and the comprehensive performances of the model.

Conclusion

This project aimed at predicting the Hospital Length of Stay (HLOS) of the COVID-19 patients using the techniques of machine learning MLP and deep learning RNN. Implementing both these models was done successfully and the accuracy of the RNN model was good but needs to be bit higher optimal mark as the model had the advantage of detecting time series trends.

Visualizations: