MODEL CARD

Basic information about the model

a. Overview: AI/ML models for predicting all-cause and cardiovascular mortality during hospitalizations.

b. Name of your team: BC Eagles

c. Participants: Han Zhang, Zhentao Wen, Shiyi Zhou, Yuqi Gao

d. AI/ML model types: Logistic Regression, Gausion Naïve Bayes, Random Forest, and XGBoost

e. Output: The Models, which were trained using the training and validation sets, were initially developed on the test data to evaluate performance prior to being implemented on the quality check data. The model results for test and quality check can be found in the zip file called TeamBCEagles_TestData and TeamBCEagles_QCData.

f. Information about Training algorithms: A mix of Logistic Regression, Gausion Naïve Bayes, Random Forest (RF), and XGBoost

g. Hyperparameters: For feature selection, we set 5-fold cross-validation and an alpha of 0.001 for Lasso in cardiovascular mortality model and 5-fold cross-validation in male and female all-cause mortality models.

h. Fairness constraints: we recreate our dependent variable based on each patient during their age brackets. Since each patient can only experience death once, our dependent variable becomes imbalanced. To address this issue, we employ the undersampling method to balance our dataset before proceeding with further analysis. Moreover, we ensure that all variables are scaled within the range of 0 to 1 to enable the application of a fairness Lasso analysis.

i. Additional approaches and features applied:

Considering computational expense and information mutually exclusive and collectively exhaustive, we have decided to use 8 tables out of 16 tables: (1) condition, (2) demographics_event, (3) demographics_static, (3) ed_visits, (4) inpatient_admissions, (5) measurements_blood_pressure, (6) measurements, (7) medications_administered, and (8) outpatient_visists for further analysis for male all-cause mortality prediction and cardiovascular mortality model. Unfortunately, for female all-cause mortality model we had to exclude emergency room visit table and the measurements table due to missing death data in died_during_ed_visits column in ed_visits table and high proportion of missing date in result_numeric column in measurements table.

Initially, we prioritize the relevant feature “age” instead of using the precise date and time information. We believe age is a more important feature in predicting our dependent variables,

mortality and readmission. To reduce computational expense and data size, we create age brackets (e.g., 50-54, 55-59, etc.) for all selected tables.

In *demographic table*, we employed one-hot encoding for "Ruca_category" column, assuming that urban areas may have have better hospital resources than rural and highly rural areas. Similarly, we applied one-hot encoding for the age bracket, assuming that the older age ranges carry more weight in our predictions compared to the younger ones.

In *conditions table*, we recategorize the condition types into broader categories, considering their frequency of occurrence and relevance to our dependent variables. This reduces the number of condition types from over 6,000 to 36.

For *medication table*, we only include the top 5 most frequently prescribed medicines: paracetamol (acetaminophen), heparin, acetylsalicylic acid (aspirin), metoprolol, and furosemide. Four of these medications are commonly used for heart-related problem, while the left one is frequently used for pain relief. We consider these medicines reasonable for further prediction.

In *measurements table*, we decided to include the measurement and result numeric columns. We generate each type of measurements as an individual column, and we calculate the average result numeric for each patient has during a certain age period.

In *measurements blood pressure* table, since patients tested their blood pressure regularly, we decided to only focus on instances where patients experienced abnormal blood pressure readings.

In *ed* and *inpatient* tables, due to the large number of different diagnosis and discharge service units, we have figured that feature hashing is the most suitable approach to convert the string variables into numeric variables for further analysis.

For the Female_All_Cause_Mortality Model, there are a total of 96 columns including the original provided data pulled from the 8 tables mentioned above. With one hot encoding and feature hashing applied to the conditions, medications, demographics, and measurement features. The Male_All_Cause_Mortality and Cardio_Mortality models included more features due to those models having usable Ed features and measurement features, which contained too many nulls for the female model.

j. Software language used: Google Cloud BigQuery and Python 3.9.13. Standard libraries used:

- numpy==1.24.4
- pandas==1.5.0
- plotly==5.13.0
- scipy==1.9.1
- seaborn==0.12.0
- tensorflow-macos==2.10.0
- tensorflow-metal==0.6.0
- xgboost==1.7.4

- keras==2.10.0
- Keras-Preprocessing==1.1.2
- cloudpickle==2.2.0
- (The full libraries can be found in the zipfile called TeamBCEagles_package_requirement.txt)

k. Run time :
   - BigQuery: Data preprocessing ran on Google Cloud BigQuery approximately 30 seconds for training sets, less than 10 seconds for test and quality check sets.
   - Python: Ran on a 2022 Mac Studio's local environment with 32GB of memory, all notebooks included should run under 10 minutes continuously.

l. We used ChatGPT4 to generate the code for the model comparison plots, and model output structure during the training process.

Additional model information (if applicable)

a. If any additional data used (besides the data provided in Phase 1 of the challenge), please provide information on this additional data:

   We did not use any additional data.

b. Peer-reviewed publication or other resource for more information [If applicable, please citeany relevant publications for your model]:

   We did not use any additional peer-reviewed publications.

Exploratory Features (please list any non-traditional factors investigated for the model development, including demographic or phenotypic groups, environmental conditions, and technical attributes)

a. Non-traditional factors: Let us know which non-traditional factors influenced your model. Please provide a brief (1-2 sentences) explanation of the logic of using these factors:
   Some non-traiditional factors we applied specific in our models including (1) creating age brackets, (2) one-hot encoding of "Ruca_category" and age brackets, (3) recategorize condition types, (4) using different aggregate methods in measurements and blood pressure, and (5) feature hashing for emergency room and inpatient visits diagnosis as we mentioned in the **Basic Information section**. The reasons are provided below:
   (1) Based on our considerations for selecting mutually exclusive and collectively exhaustive information, we have decided to include only one time-relevant column. In this case, age provides more insights than date and time, as older individuals are more likely to experience mortality and hospital visits;
   (2) For "Ruca_category" and "age bracket", we specifically chose one-hot encoding instead of ordinal encoding. Our assumption is that urban areas provide greater hospital resources, and we believe that older age groups carry more significance in predicting outcomes;

(3) We recategorized condition types and more specifically, we look at that are the conditions each patient had and assessed whether these conditions could impact the patient's survival rate;
(4) We count the occurrences of abnormal diastolic and systolic values per patient and calculated the average values for different measurements. Regular health check-ups at the hospital make it more meaningful to consider abnormal values experienced by patients;
(5) Lastly, we specifically chose feature hashing to convert the diagnosis and discharge service units information from emergency room and inpatient visits table into numeric values. This approach enables us to retain all the relevant information, meanwhile process map reduction to combine the similar categories and to assign appropriate weights to each, thereby preserving their importance.

b. Evaluation factors: How were the factors evaluated?

Through the implementation of these tailored feature engineering techniques, we successfully addressed two primary challenges encountered during our analysis: (1) the computation running time, and (2) the overfitting problem. These data preprocessing methods substantially decreased the computation running time, resulting in increased the overall team productivity. Moreover, they led to improved test model accuracy, rising from 77 percent to 84 percent by applying our best model, effectively mitigating the overfitting issue.

c. Additional notes[Optional]: Provide any further information relevant to the exploratory features:

No further information

Caveats and Recommendations: (Please indicate the potential weaknesses of your model, and also provide recommendations on optimal deployment of the model):

For the mortality models, the potential weakness lies in the overfitting problem. To address this, we propose some recommendations for future work that could potentially further improve our test accuracy:

- Creating three features for the first date of each condition and medication per patient, the longevity per condition and medication, and the stop reasons. This will provide valuable insights into which conditions might lead to mortality and which medication positively impact survival rate;
- Creating a feature for extra medication doses taken, as it could significantly influence the survival rate or the probability of hospital readmission;
- Explore the use of three dimensional LSTM models with three inputs: the number of patients, patient's age, and core features. This approach may potentially yield higher accuracy predicitons.

Due to the limited time range, we were unable to conduct a thorough analysis of readmission. Our dependent variable for readmission was equal to 1 for all the patients. This happened could due to ignoring the time range and diagnosis type. For future development of a readmission model, it is essential to:

- Clearly define the time range for readmission (e.g., readmission refers within 3-9 moths) to create a more precise and meaningful target variable;
- Use date and time as the time feature instead of age, as this is more relevant to readmission predicitonsl
- Count the occurrence of revisits based on each diagnosis type to build a more accurate and reliable readmission model.