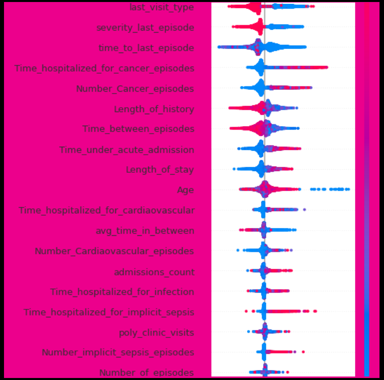
MEDINFO23
8 – 12 JULY 2023 | SYDNEY, AUSTRALIA



Predicting in-hospital death from derived EHR trajectory features

@RajeevBopche

Rajeev Bopche

PhD Student

Norwegian University of Science and Technology, Trondheim, Norway

Introduction

- Electronic Health Records (EHRs)
  - Data driven analyses
    - Disease progression modelling
    - Patient trajectory modelling
    - Disease inference
    - Risk stratification
    - Survival Prediction

- Challenges
  - Sparseness
  - Context-dependency
  - Incompleteness
  - Inconsistency
  - Inaccuracy

Electronic Patient Records (EPRs)

- Patient discharge summaries provide a longitudinal perspective of patients' interactions with hospital service
  - In Norway, with predominantly public specialist healthcare, patients often have long and continuous histories within one hospital's records
- We conducted a retrospective analysis of medical histories with poor outcomes
  - we selectively derived the most relevant count and temporal features and used them to train classifiers for predicting in-hospital mortality for the next episode

Methods

- Data
  - The data for this study includes individual episodes of care from St. Olavs university hospital between 2015-2020 for 35,594 patients that had at least one episode of suspected bloodstream infection (BSI)
  - The episodes range from the introduction of EHR in 1999 until 2020 but do not include primary care or visits to other specialist care
  - The mean age of the complete cohort is 63.6 years, and the gender distribution is 52.5% males to 47.4% females
  - The data contains information on a total of 1.2 million medical episodes

Methods

- Derived Features

Features	Unit
Age at first visit and last visit	Years
Time to recent episodes	Hours
Time between episodes	Hours
Time Hospitalized for implicit sepsis	Hours
Time Hospitalized for explicit sepsis	Hours
Time Hospitalized for cancer disease	Hours
Time Hospitalized for cardiovascular disease	Hours
Total time under acute care	Hours
Total length of stay	Hours
Length of history	Days
Implicit sepsis episodes	Count
Number of cancer episodes	Count
Number of explicit sepsis episodes	Count
Number of cardiovascular episodes	Count
Number of infection episodes	Count
Number of hospital visits	Count

Methods

- Prediction Modelling
  - All derived features are taken as continuous values and empty cell values were imputed to zero
  - The target feature was labeled as 0, if death occurred within 30 days of final episodes, and 1, if the patient was alive
  - The dataset was split into a training set (80%) and a testing set (20%). SMOTE (Synthetic Minority Over-sampling Technique) was used to address the class imbalance in the training set

Results

- Machine Learning (ML) classifiers:
    - Logistic Regression (LR) as the linear model
    - Gaussian Naïve Bayes (GNB) as the probabilistic model
    - K-Nearest Neighbors (KNN) as the non-parametric model
    - Random Forest (RF), Bagging and Boosting decision tree classifiers (BG and ADB), Voting Classifier as the ensemble model
    - Multi-layer perceptron (MLP) as neural network based model
    - eXtreme Gradient Boosting (XGBoost) model
- Model interpretations using SHapley Additive exPlanations (SHAP) values

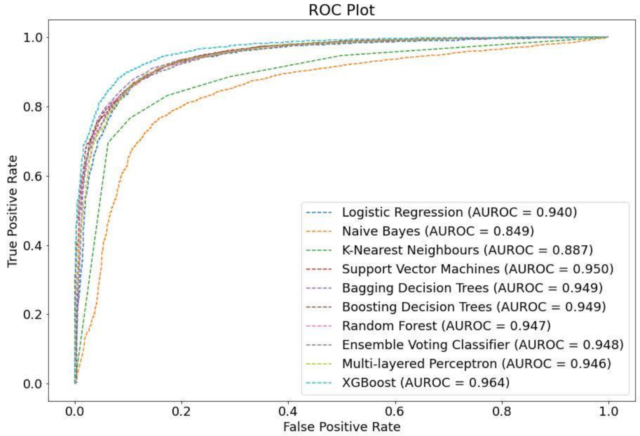
Results

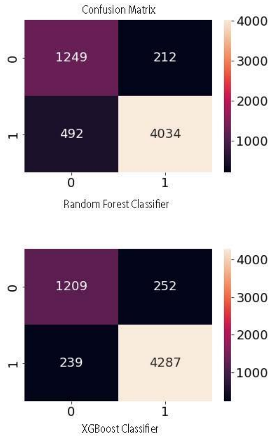
- Model performance comparison

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUROC
Logistic Regression	0.883	0.868	0.887	0.851	0.940
Naïve Bayes	0.823	0.631	0.885	0.760	0.849
K-Nearest Neighbors	0.830	0.826	0.831	0.792	0.887
Support Vector Machine	0.885	0.859	0.893	0.853	0.950
Bagging Decision Trees	0.900	0.774	0.941	0.863	0.949
Boosting Decision Trees	0.902	0.776	0.942	0.863	0.949
Random Forest	0.882	0.854	0.891	0.865	0.947
Ensemble	0.895	0.702	0.958	0.852	0.948
Neural Network	0.899	0.768	0.942	0.861	0.946
XGBoost	0.918	0.827	0.947	0.888	0.964

Results

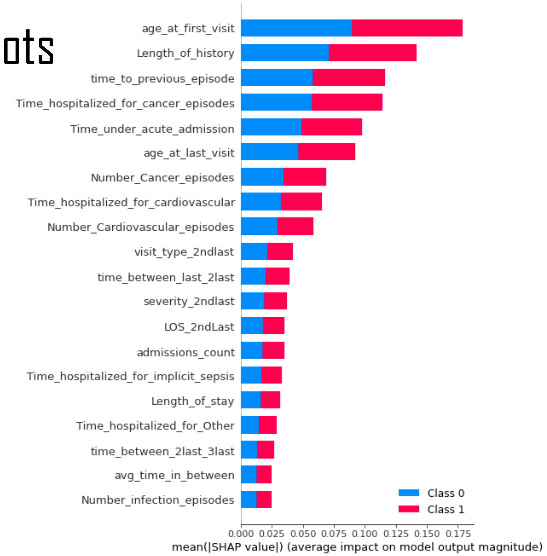
- ROC curve for all models and Confusion matrices Random Forest model (top right) and XGBoost model (bottom right)

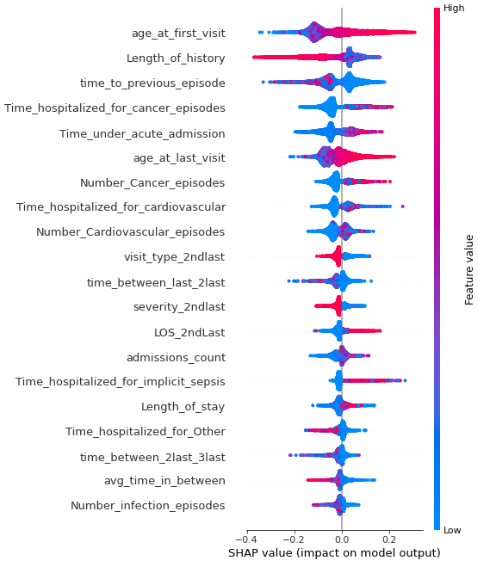




Results

- Feature importance and summary plots using SHAP values





Discussions

- The last part of the medical history of the patients who died in hospital usually consisted of not one single episode but a series of episodes
- The age at the first visit is the most important feature indicating that patients arriving at the hospital for the first time very late in life are at very high risk
- The length of history is the second most important feature with a shorter length of history contributing towards a higher risk of in-hospital death

Discussions

- One limitation of this study is that live patient histories were taken as complete instead of trimming histories up to some critical episodes
- The major contribution of this study is that these early warning models can be easily implemented in the current and developing digital health platforms to predict adverse outcomes enabling proactive and precautionary care

Conclusions

- Accurate prediction of mortality can assist in better hospital resource allocation, enabling healthcare providers to prioritize patients who are at higher risk of adverse events
- We were able to predict if an impending disease episode entailed a risk of death