30-Day Readmission Prediction

Tong Shen, Ziyuan Shen

Instructor: Michael Gao, Xiling Shen

Department of Biomedical Engineering
Department of Electrical and Computer Engineering
Duke University

May 3, 2019

Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Background

Motivation

- Early identification of high risk patients
- Prevent early discharging
- Improve managing ICU care and resources

Background

Existing works:

- Feature Extraction
  - Demographic characteristics: age, gender, race etc.
  - Lab results, chart events etc. monitored in ICU[1]
  - Electronic health record data: length of stay, number of admissions, admission type etc.[2][3]
- Techniques
  - Random Forest
  - Artificial Neural Network[4]

---

[1] Yaron Blinder. *Predicting 30-day ICU readmissions from the MIMIC-III database*. https://github.com/YaronBlinder/MIMIC-III_readmission. 2017.

[2] Oanh Kieu Nguyen et al. "Predicting all-cause readmissions using electronic health record data from the entire hospitalization: model development and comparison". In: *Journal of hospital medicine* 11.7 (2016), pp. 473–480.

[3] Frida Kareliusson, Lina De Geer, and Anna Oscarsson Tibblin. "Risk prediction of ICU readmission in a mixed surgical and medical population". In: *Journal of intensive care* 3.1 (2015), p. 30.

[4] Ricardo Bento Afonso. "Feature Extraction and Selection for Prediction of ICU Patient's Readmission Using Artificial Neural Networks". In: (2013).

Background

Existing works:

- Feature Extraction
  - Demographic characteristics: age, gender, race etc.
  - Lab results, chart events etc. monitored in ICU
  - Electronic health record data: length of stay, number of admissions, admission type etc.
- Techniques
  - Random Forest
  - Artificial Neural Network

Overall Performance[5]:

- Accuracy: $0.48 \sim 0.61$
- Sensitivity: $0.72 \sim 0.77$
- Specificity: $0.44 \sim 0.60$

[5] Ricardo Bento Afonso. "Feature Extraction and Selection for Prediction of ICU Patient's Readmission Using Artificial Neural Networks". In: (2013).

Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Feature Generation

- Data source: MIMIC III[6]
- Summary statistics

Info. \ Data	Diagnosis Information	ICU Monitoring Data
Tables Used	ICD_DIAGNOSES	CHARTEVENTS LABEVENTS
Readmission Proportion	5.9%	6.2%
Data Type	Categorical	Categorical & Numerical
# Dims	6776	57
# Samples	58925	42228

[6] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3 (2016), p. 160035.

Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Data Exploration

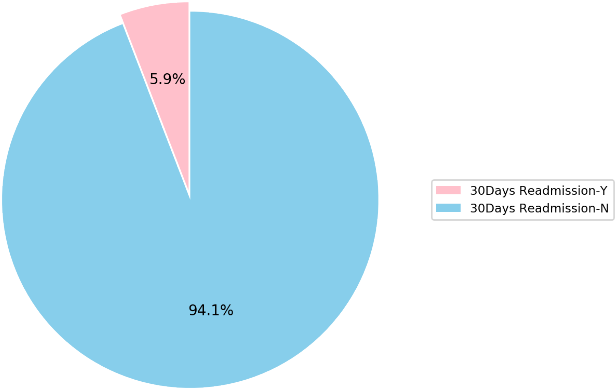


Figure: Readmission proportion of diagnosis dataset.

Data Exploration

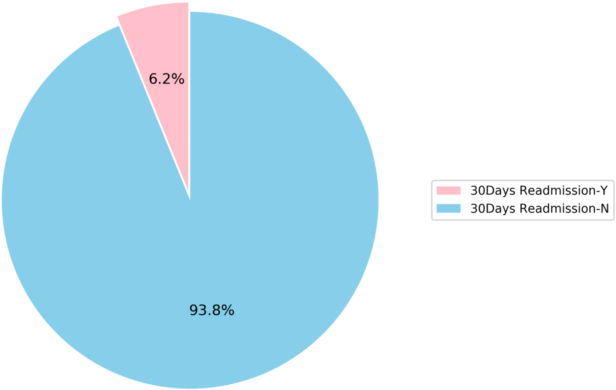


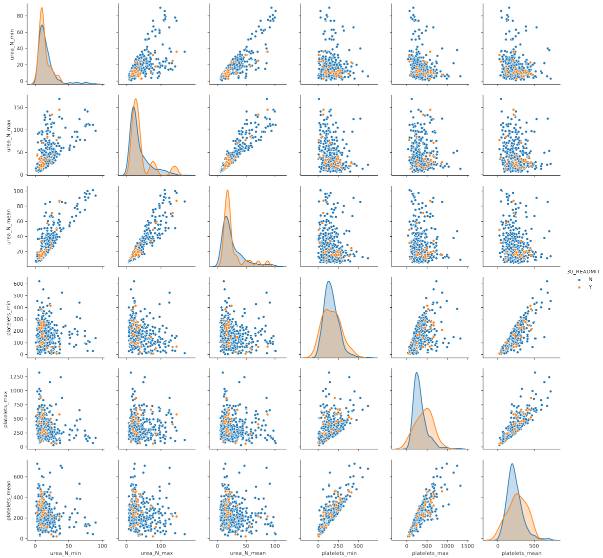
Figure: Readmission proportion of ICU dataset.

30-Day Readmission Prediction

Tong Shen, Ziyuan Shen

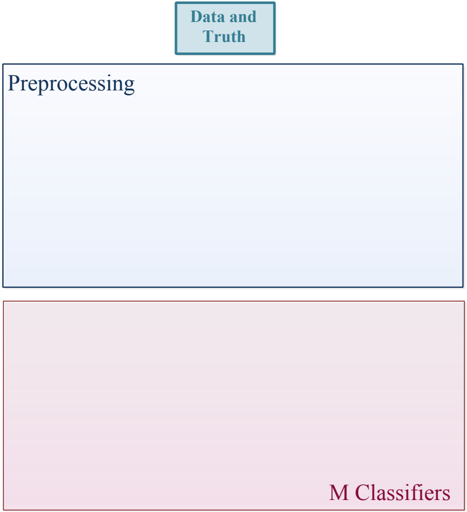
Content
Background
Data Processing
Data Exploration
Algorithmic Pipeline
Experimental Results
Conclusion

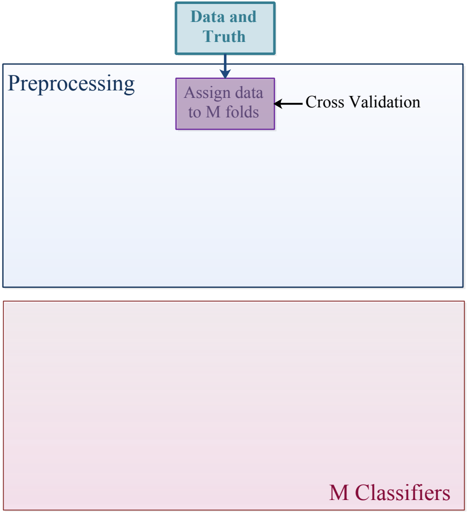
Scatter plot of numerical data

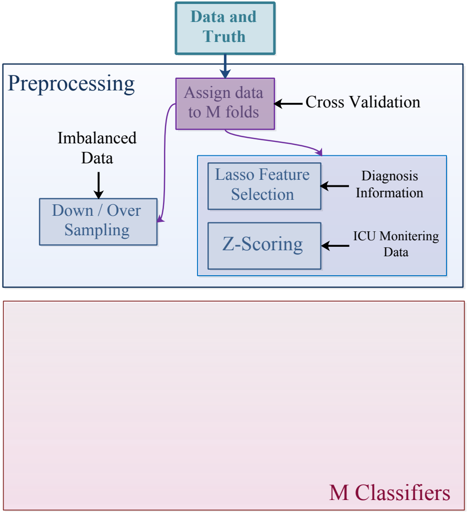


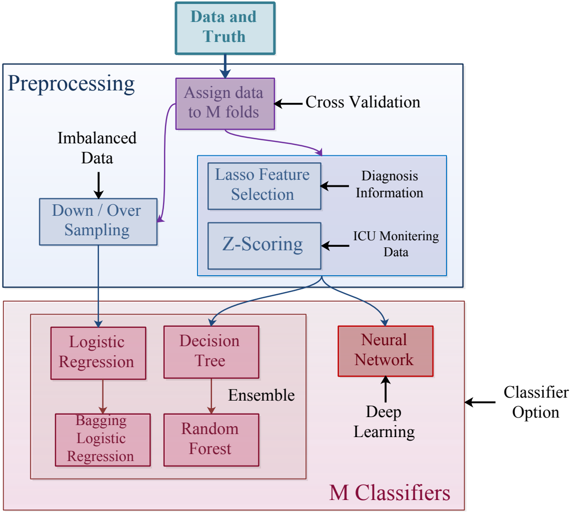
Content

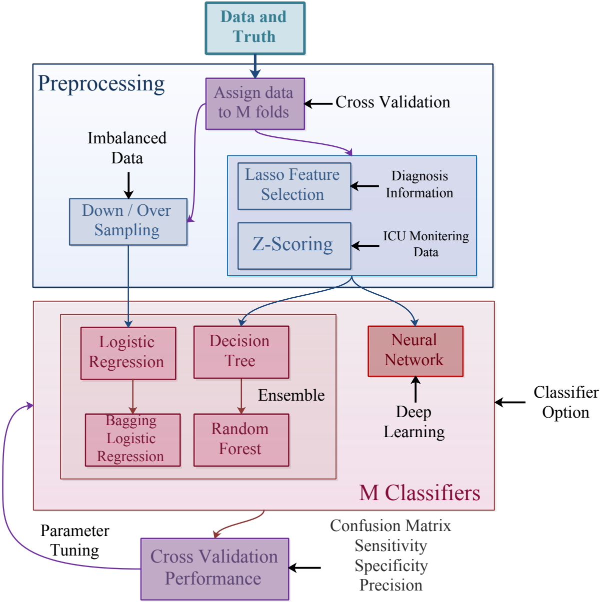
- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion


Data and Truth
Preprocessing
M Classifiers

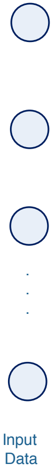

Data and Truth
Preprocessing
Assign data to M folds
Cross Validation
M Classifiers


Data and Truth
Preprocessing
Assign data to M folds
Cross Validation
Imbalanced Data
Down / Over Sampling
Lasso Feature Selection
Diagnosis Information
Z-Scoring
ICU Monitering Data
M Classifiers

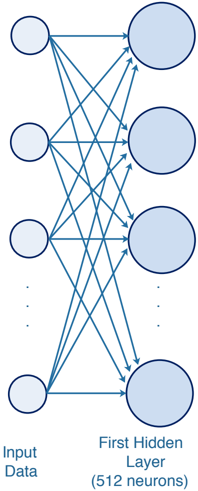

Data and Truth
Preprocessing
Assign data to M folds
Cross Validation
Imbalanced Data
Down / Over Sampling
Lasso Feature Selection
Diagnosis Information
Z-Scoring
ICU Monitering Data
Logistic Regression
Decision Tree
Ensemble
Bagging Logistic Regression
Random Forest
Neural Network
Deep Learning
Classifier Option
M Classifiers


Data and Truth
Preprocessing
Assign data to M folds
Cross Validation
Imbalanced Data
Down / Over Sampling
Lasso Feature Selection
Diagnosis Information
Z-Scoring
ICU Monitering Data
Logistic Regression
Decision Tree
Ensemble
Bagging Logistic Regression
Random Forest
Neural Network
Deep Learning
Classifier Option
M Classifiers
Parameter Tuning
Cross Validation Performance
Confusion Matrix
Sensitivity
Specificity
Precision

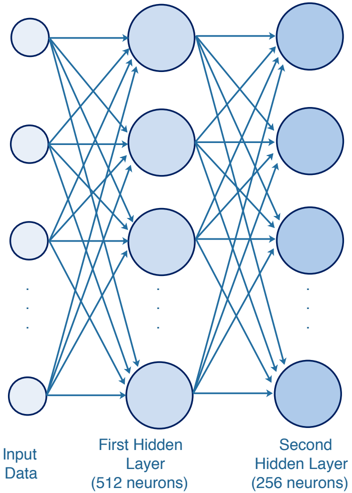
Neural Network



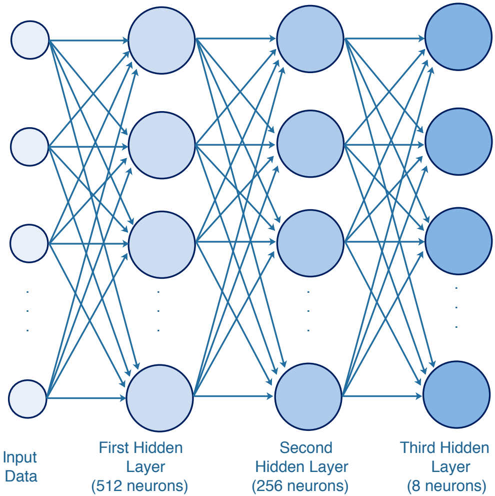
Neural Network



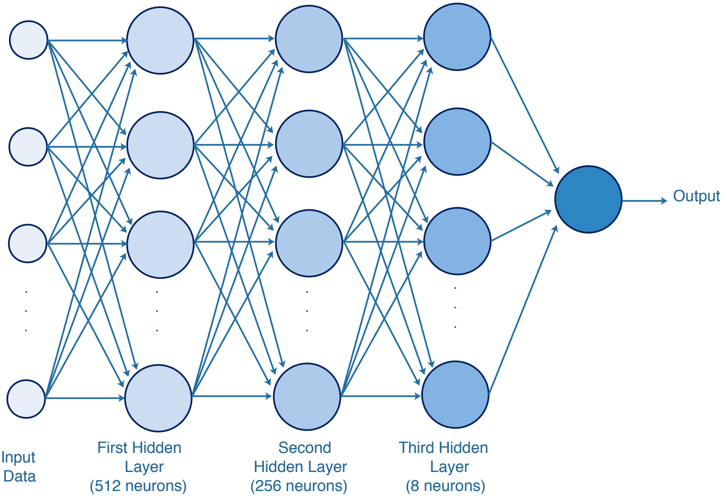
Neural Network



Neural Network



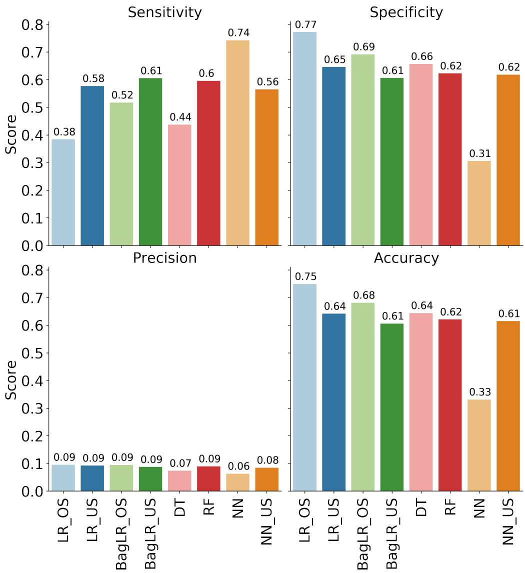
Neural Network



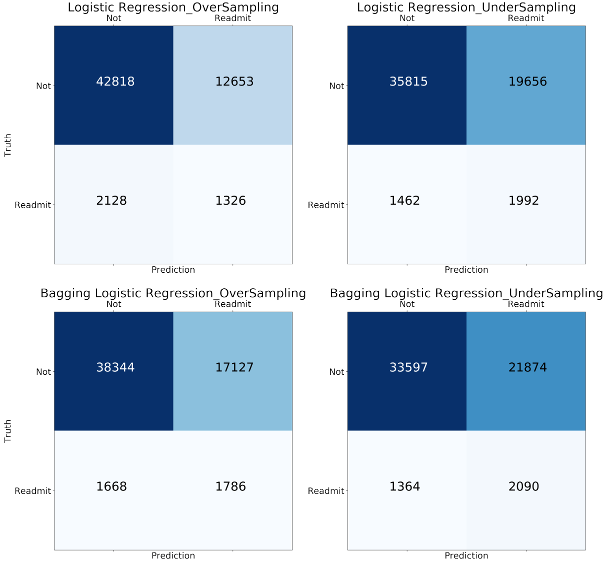
Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Performance Scores-Diagnosis



Confusion Matrix-Diagnosis



Content

Background

Data Processing

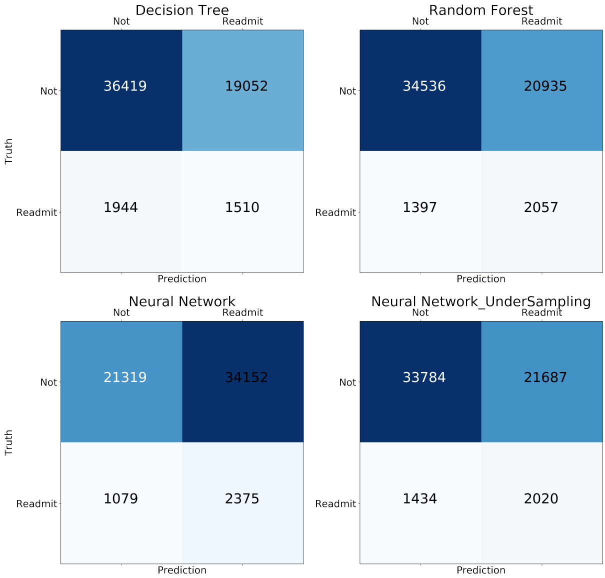
Data Exploration

Algorithmic Pipeline

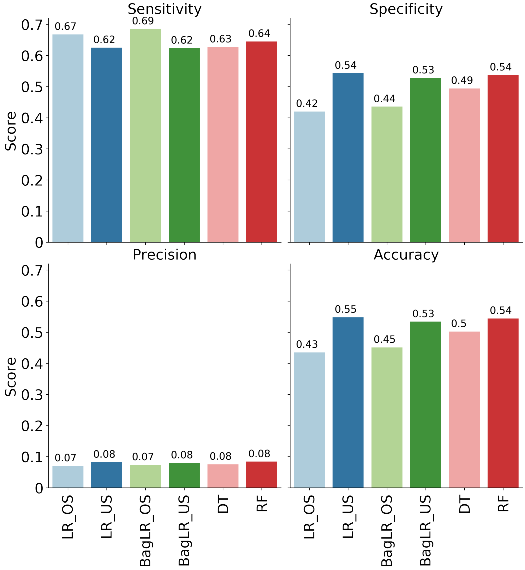
Experimental Results

Conclusion

Confusion Matrix-Diagnosis



Performance Scores-ICU



Confusion Matrix-ICU

Content

- Background
- Data Processing
- Data Exploration
- Algorithmic Pipeline
- Experimental Results
- Conclusion

Conclusion

- Fair performance in general
  - feature extraction and data cleaning
  - class imbalance
  - high dimensional data
- Ensemble methods
  - can increase performance
  - usually perform well with structured data unless we have a lot of data
- Neural Network
  - achieve average performance with careful design
  - not as efficient as traditional classifiers
  - usually apply to unstructured data such as images or natural text

Thank you for listening!

Q&A