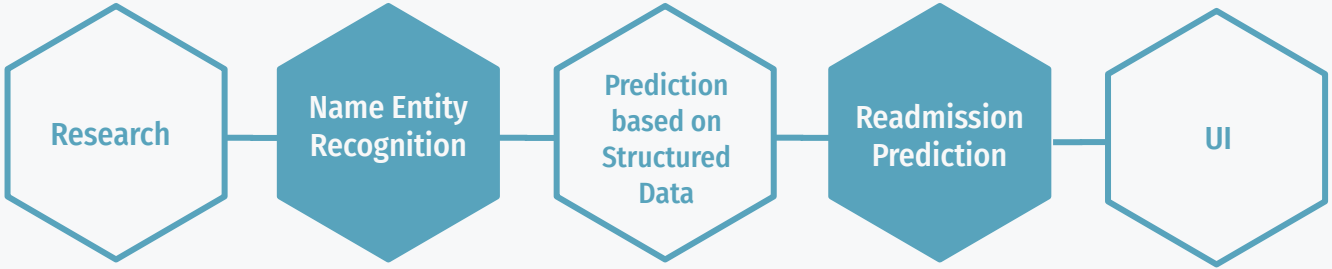
MEDICAL LANGUAGE PROCESSING

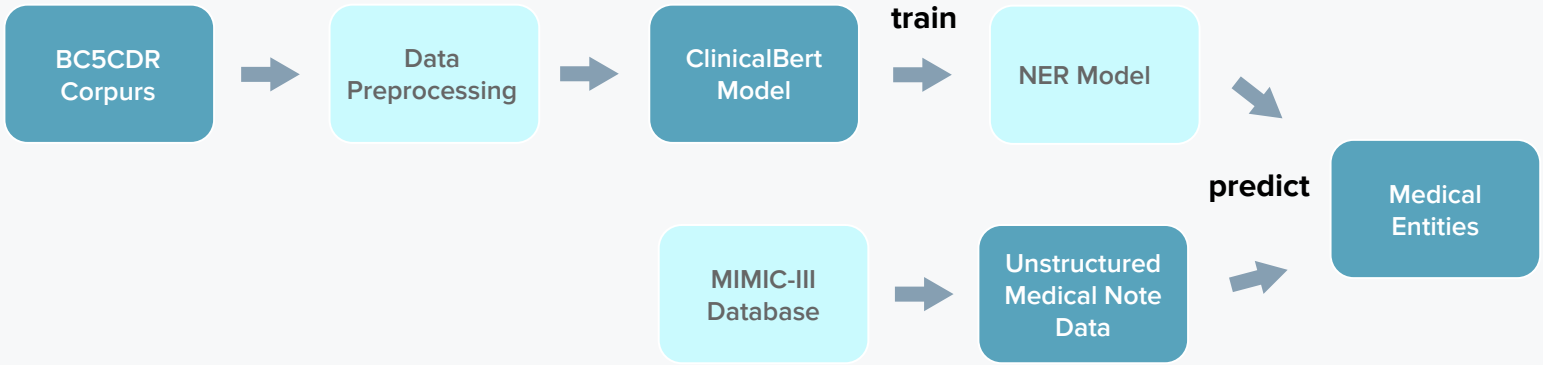
FULLMARK
Lelei Zhang, Muyao Chen, Shiying Tu, Bin Tong, Zhixuan Chi

Introduction



- Use MIMIC 3 dataset to implement **Named Entity Recognition**, **Structured Data Prediction**, **Unstructured Data Prediction** for important clinical indices.
- Integrate trained models to a **Web Application** which can be used to assist medical decisions.

Name Entity Recognition



- BC5CDR Corpus: consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions.
- MIMIC-III: an openly available dataset developed by the MIT Lab for Computational Physiology, including demographics, vital signs, laboratory tests, medications, and more

Name Entity Recognition

Admission Date: [**2118-6-2**] Discharge Date: [**2118-6-14**]

Date of Birth: Sex: F

Service: MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS: This is an 81-year-old female with a history of emphysema (not on home O2), who presents with three days of shortness of breath thought by her primary care doctor to be a COPD flare. Two days prior to admission, she was started on a prednisone taper and one day prior to admission she required oxygen at home in order to maintain oxygen saturation greater than 90%. She has also been on levofloxacin and nebulizers, and was not getting better, and presented to the [**Hospital1 18**] Emergency Room.

In the [**Hospital3 **] Emergency Room, her oxygen saturation was 100% on CPAP. She was not able to be weaned off of this despite nebulizer treatment and Solu-Medrol 125 mg IV x2.

Review of systems is negative for the following: Fevers, chills, nausea, vomiting, night sweats, change in weight, gastrointestinal complaints, neurologic changes, rashes,palpitations, orthopnea. Is positive for the following: Chest pressure occasionally with shortness of breath with exertion, some shortness of breath that is positionally related, but is improved with nebulizer treatment.

Name Entity Recognition

B-Disease
B-Chemical
I-Disease
I-Chemical

Service: MICU and then to [**Doctor Last Name **] Medicine

HISTORY OF PRESENT ILLNESS: This is an 81-year-old female with a history of emphysema (not on home O2), who presents with three days of shortness of breath thought by her primary care doctor to be a COPD flare. Two days prior to admission, she was started on a prednisone taper and one day prior to admission she required oxygen at home in order to maintain oxygen saturation greater than 90%. She has also been on levofloxacin and nebulizers, and was not getting better, and presented to the [**Hospital1 18**] Emergency Room.

In the [**Hospital3 **] Emergency Room, her oxygen saturation was 100% on CPAP. She was not able to be weaned off of this despite nebulizer treatment and Solu-Medrol 125 mg IV x2.

Name Entity Recognition

B-Disease
B-Chemical
I-Disease
I-Chemical

Review of systems is negative for the following: Fevers, chills, nausea, vomiting, night sweats, change in weight, gastrointestinal complaints, neurologic changes, rashes, palpitations, orthopnea. Is positive for the following: Chest pressure occasionally with shortness of breath with exertion, some shortness of breath that is positionally related, but is improved with nebulizer treatment.

Prediction on **structured** and **unstructured** Data

Consideration of diverse application scenarios:

- Newly generated clinical notes.
- Existing complicated clinical database(MIMIC 3).

Prediction topics

MORTALITY RATE

- Based on structured data
- Classification problem

LENGTH OF STAY

- Based on structured data
- Regression problem

READMISSION

- Based on unstructured data(text note)

Mortality Rate

Data Preprocessing:

- Integrate MIMIC 3 into one table

  (Medical Reference: "Medical Information Extraction & Analysis: From Zero to Hero with a Bit of SQL and a Real-life Database,")

  - ➢ Summarize the important medical indicators into **numerical values on a daily basis**.
  - ➢ Extract **key categorical data**.

MIMIC 3

- ➢ **26** tables
- ➢ **1 million+** rows
- ➢ **200+** features

Summarized Table

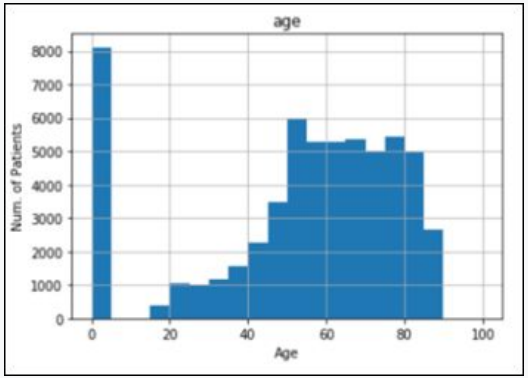
- ➢ **58880** rows(count on admission)
- ➢ **26** medical valuable features
- ➢ **Outlier detection**
- ➢ **Restore the blurred information**
- ➢ **Fill the Missing values**

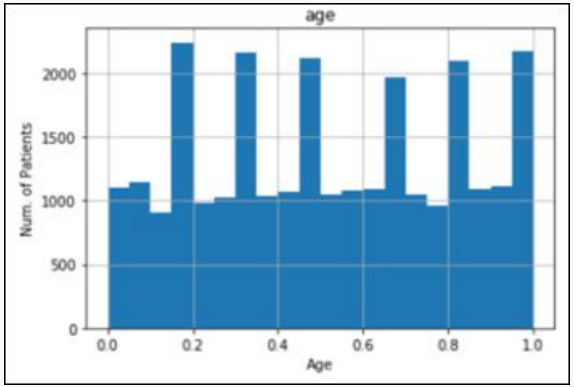
```
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   admit_type        58976 non-null  object
 1   admit_location    58976 non-null  object
 2   insurance         58976 non-null  object
 3   religion          58518 non-null  object
 4   marital_status    48848 non-null  object
 5   ethnicity         58976 non-null  object
 6   AdmitDiagnosis    58951 non-null  object
 7   ExpiredHospital   58976 non-null  int64
 8   LOSdays           58976 non-null  float64
 9   age               58976 non-null  float64
 10  gender            58976 non-null  object
 11  NumCallouts       58976 non-null  float64
 12  NumDiagnosis      58976 non-null  float64
 13  NumProcs          58976 non-null  float64
 14  NumCPTevents      58976 non-null  float64
 15  NumInput          58976 non-null  float64
 16  NumOutput         58976 non-null  float64
 17  NumLabs           58976 non-null  float64
 18  NumMicroLabs      58976 non-null  float64
 19  NumNotes          58976 non-null  float64
 20  NumProcEvents     58976 non-null  float64
 21  NumTransfers      58976 non-null  float64
 22  NumChartEvents    58976 non-null  float64
 23  NumRx             58976 non-null  float64
 24  AdmitProcedure    50944 non-null  object
 25  TotalNumInteract  58976 non-null  float64
```

Mortality Rate

Feature Engineering:

- **Feature Selection**: Drop the irrelevant features and deal with ‘Leaking Data from the Future ’ problem. Eg: drop the LOSdays.
- **Feature Transformation**: Categorical features to Numerical features using one-hot encoding
- **Feature Scaling:** Normalization on skewed data(X).

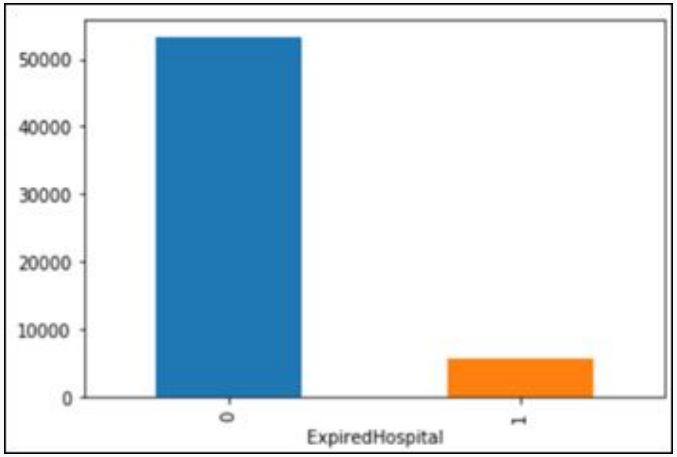


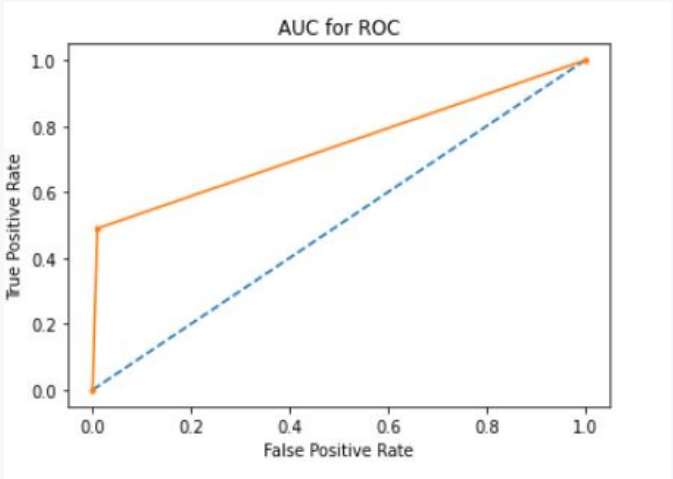


Mortality Rate

Imbalanced Dataset(1:9):

- **Trivial Accuracy rate**.
- Recall, Precision, F1 score, ROC AUC . etc.
- Recall and Precision **tradeoff concern** for medical projects.

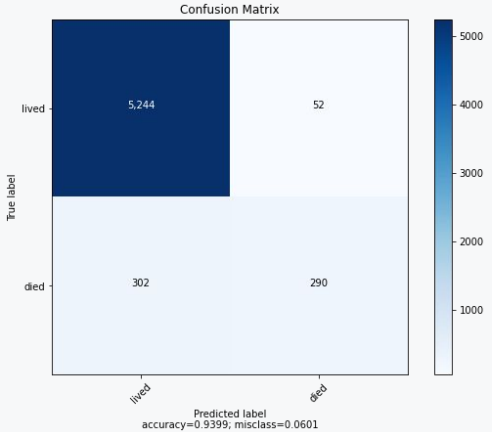


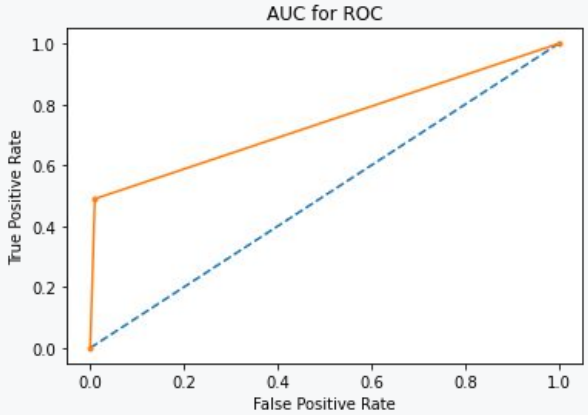


Mortality Rate

Classification models(RF, SGD, LogReg, Knearest, SVM, GaussianNB, Decision tree):

- Model selection using **cross validation** with evaluation metrics mentioned.
- Fine tune the RF model using **Grid Search**.
- Visualize the model performance with ROC-AUC curve, confusion matrix .etc
- Visualize the **feature importances**.



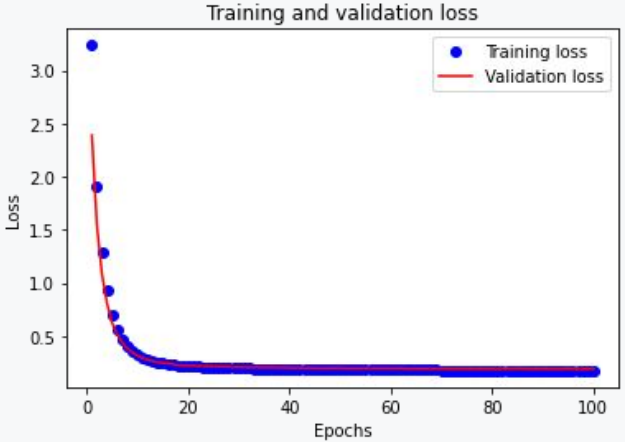


precision	0.848
recall	0.4899
accuracy	0.9399
F1 score	0.621

Mortality Rate

Neural Network:

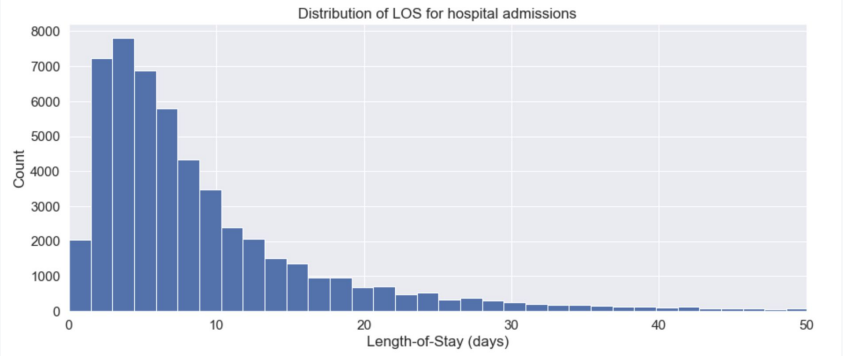
- Four hidden layers(2048 units) with ReLU activation function.
- Use **dropout** and **batch normalization** layers to avoid overfitting.
- **Sigmoid** output layer.
- Similar performance as RF model after 100 epochs.



Length of Stay

ADMISSIONS.csv, PATIENTS.csv, DIAGNOSES_ICD.csv, ICUSTAYS.csv, CPTEVENTS.csv, INPUTEVENTS_CV, OUTPUTEVENTS.csv, MICROBIOLOGYEVENTS.csv

1. Ensure that there are no admissions resulting in death.
2. Remove null values.
3. Drop columns that are unused.



Length of Stay

Structure Data → Data Preprocessing → Feature Engineering → Supervised Learning → Evaluation

List of ICD-9 codes

From Wikipedia, the free encyclopedia

The following is a list of codes for International Statistical Classification of Diseases and Related Health Problems.

- List of ICD-9 codes 001–139: infectious and parasitic diseases
- List of ICD-9 codes 140–239: neoplasms
- List of ICD-9 codes 240–279: endocrine, nutritional and metabolic diseases, and immunity disorders
- List of ICD-9 codes 280–289: diseases of the blood and blood-forming organs
- List of ICD-9 codes 290–319: mental disorders
- List of ICD-9 codes 320–389: diseases of the nervous system and sense organs
- List of ICD-9 codes 390–459: diseases of the circulatory system
- List of ICD-9 codes 460–519: diseases of the respiratory system
- List of ICD-9 codes 520–579: diseases of the digestive system
- List of ICD-9 codes 580–629: diseases of the genitourinary system
- List of ICD-9 codes 630–679: complications of pregnancy, childbirth, and the puerperium
- List of ICD-9 codes 680–709: diseases of the skin and subcutaneous tissue
- List of ICD-9 codes 710–739: diseases of the musculoskeletal system and connective tissue
- List of ICD-9 codes 740–759: congenital anomalies
- List of ICD-9 codes 760–779: certain conditions originating in the perinatal period
- List of ICD-9 codes 780–799: symptoms, signs, and ill-defined conditions
- List of ICD-9 codes 800–999: injury and poisoning
- List of ICD-9 codes E and V codes: external causes of injury and supplemental classification

DIAGNOSES_ICD.csv. ----- ICD9-Code
The ICD is designed as a health care classification system, providing a system of diagnostic codes for classifying diseases.

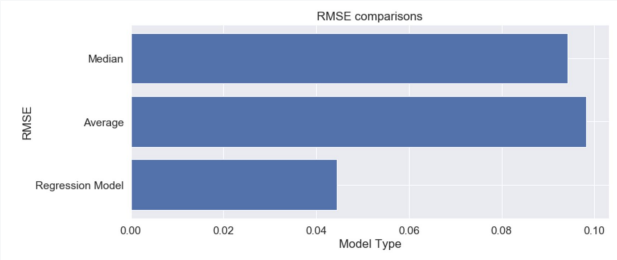
ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE
1297	109	172335	1	40301
1298	109	172335	2	486
1299	109	172335	3	58281
1300	109	172335	4	5855
1301	109	172335	5	4254
1302	109	172335	6	2762
1303	109	172335	7	7100

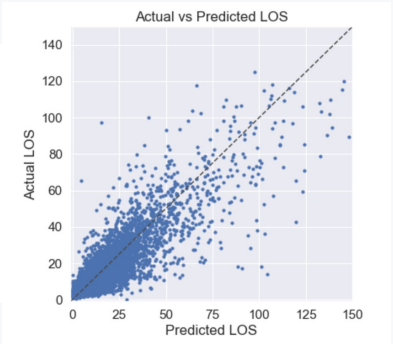
congenital anomalies	diseases of the blood	diseases of the circulatory system	diseases of the digestive system	diseases of the genitourinary system	diseases of the musculoskeletal system and connective tissue	diseases of the nervous system and sense organs	r
0	0	2	2	2	0	2	
0	1	2	4	0	0	0	
0	0	0	0	0	0	0	
0	0	1	2	0	0	0	
0	1	7	0	0	0	0	

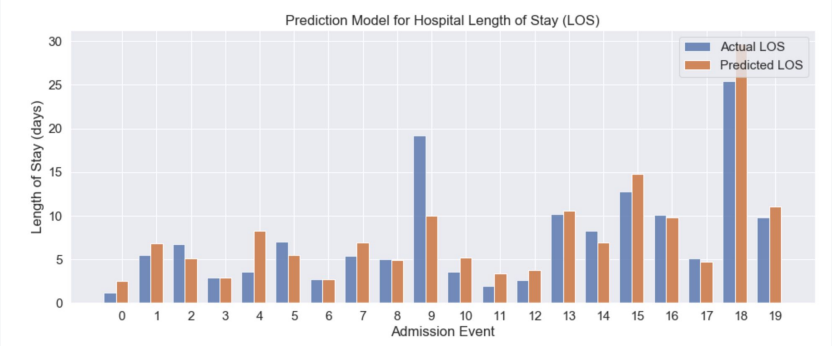
Length of Stay

Structure Data → Data Preprocessing → Feature Engineering → Supervised Learning → Evaluation

Supervised Learning → Regression



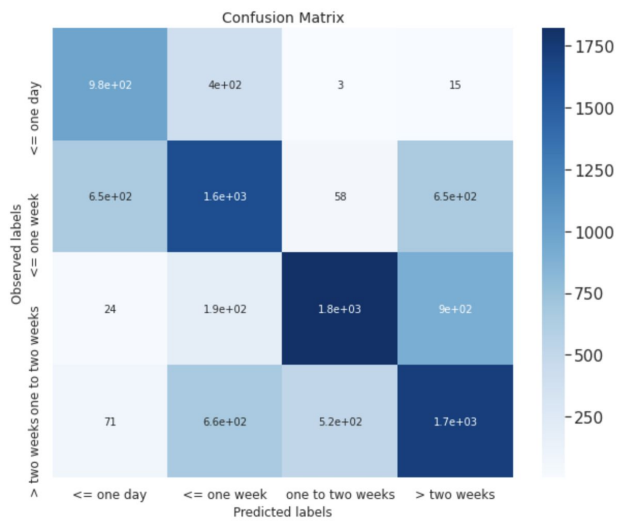


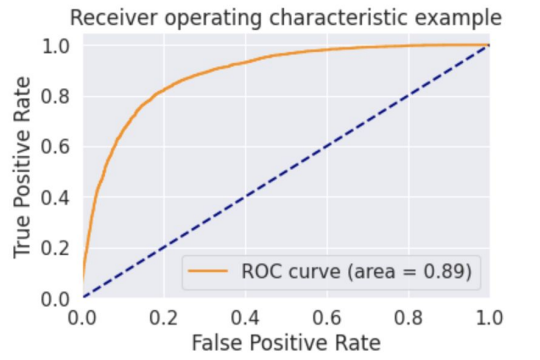


Length of Stay

Structure Data → Data Preprocessing → Feature Engineering → Supervised Learning → Evaluation

Supervised Learning → Classification → RNN LSTM

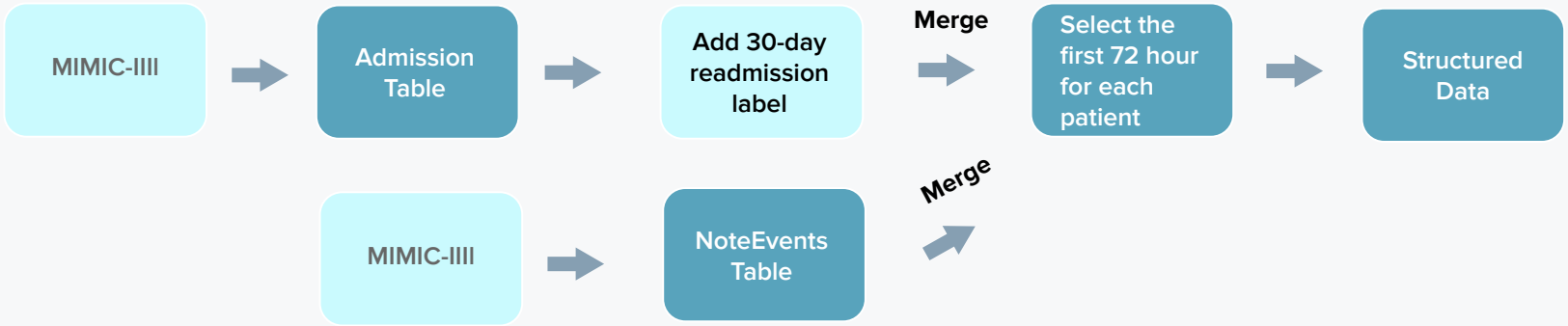




Readmission Using ClincalBERT

Importance:

- Trusted measure of effective and responsible care
- Help the hospitals to make plans in various areas: cost, bed occupancy



Readmission Using ClincalBERT

Clinical note text with 3-day readmission label

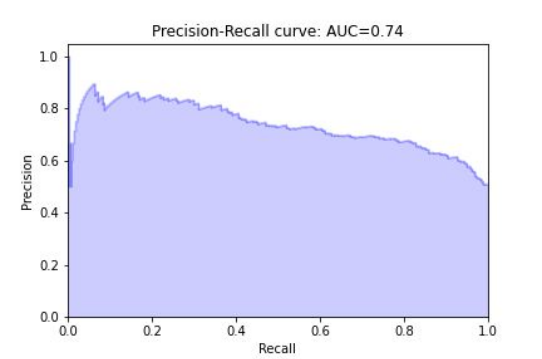
Accuracy

0.621760705752619

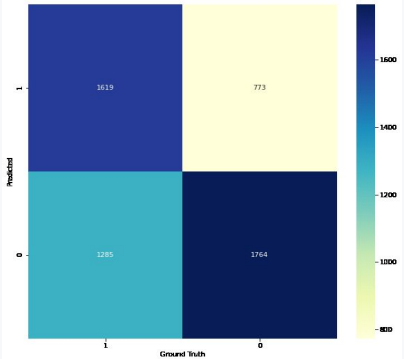
Label 1 vs Label 0

3:1

Precision-Recall Curve



Confusion Matrix



Why NLP? *Using pure text to predict readmission*

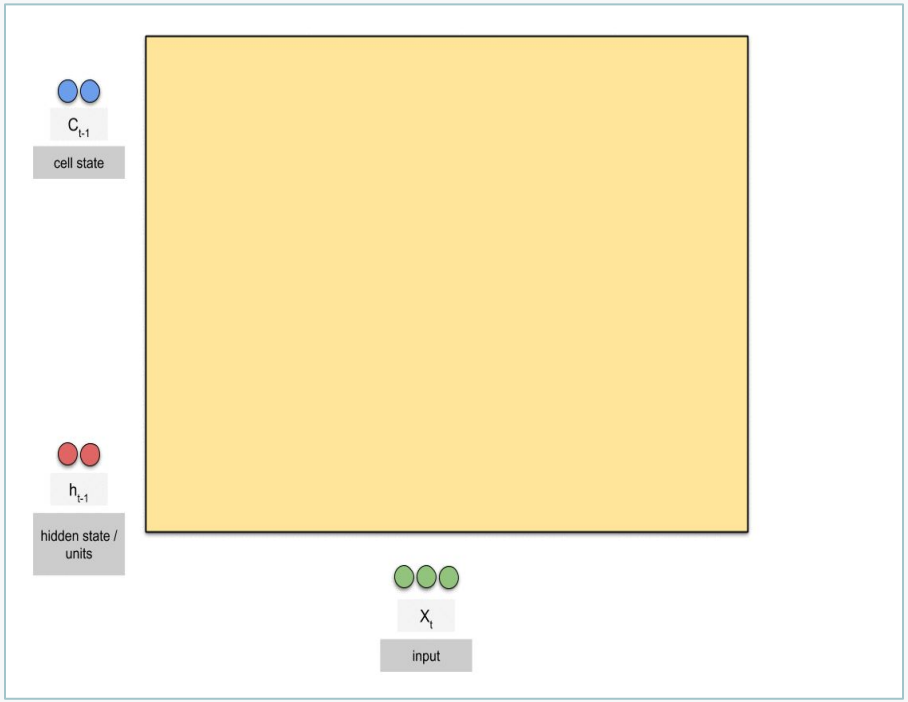
Social benefit to help with COVID-19 situation. *Eg. Extract information, Find complications, Predict ICU occupancy*

Embedding is a universal idea not only limited in NLP. *Eg. Compression, Dimensionality Reduction, Signal Processing*

Increasing possible cases we do not have any structural data. *Eg. Trump's talk, Opec oil meeting*

Natural Language Processing

Using RNNs(LSTM, GRU, Bi-directional) to capture sequential information

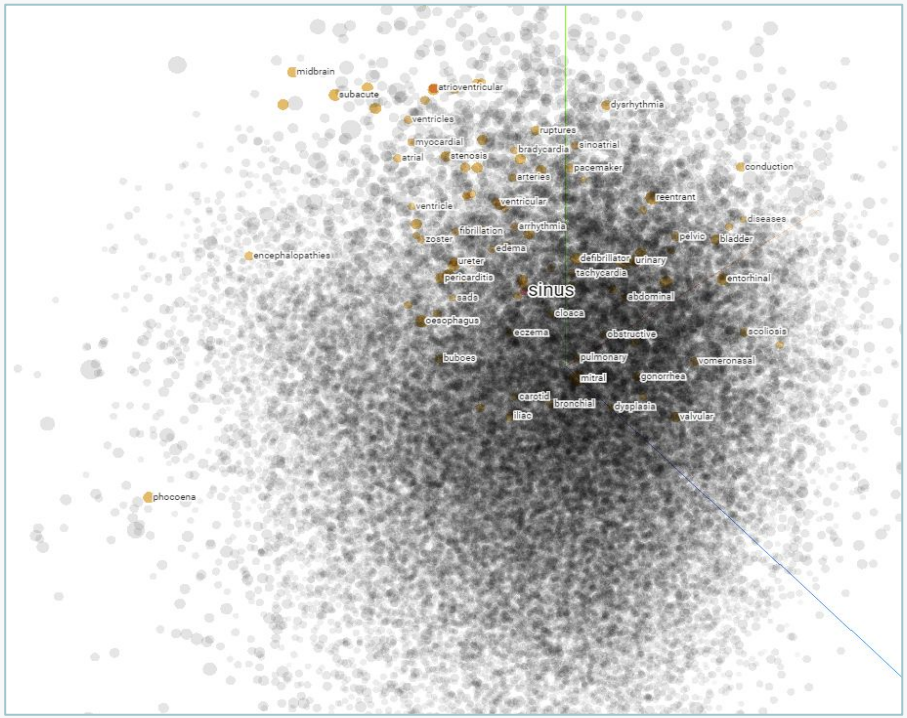


① Load a pre-trained embedding model

② Convert a text to numerical vector presentation

③ Create a machine learning/deep learning model as usual

Natural Language Processing



Sinus is a kind of nasal disease. All these words below are strongly associated with it (they are all clinical terminologies):

-eczema

-cloaca

-tachycardia

-abdominal

-edema

.....

U

Patient Information

Medical ID

RETRIEVE

Please enther your note:

Date

Patient is infectious by XXX, final diagnose is still in process

Drag and Drop or Select Files

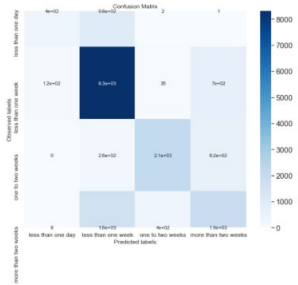
SUBMIT

Medical Volcabulary

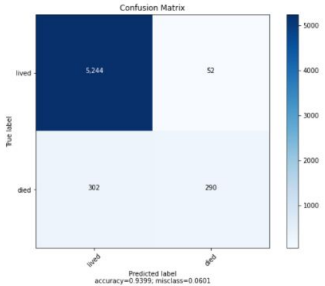
Date Patient is infectious by XXX , final diagnose is still in process

PREDICTION

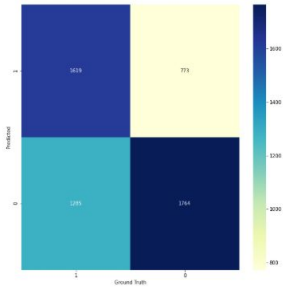
Length of Stay: Less than 1 week



Mortily Rate: Low



Readmission: None



Achievement

- Successfully integrate in-class knowledge to the project
- Explored large corpus such as MIMIC-III and BC5CDR
- Applied various NLP models
- Handle imbalanced data

Future Works

1. Build search engine for doctors to search patients by medical entities
2. Add NER results to structured data for prediction
3. Add longitude data for further prediction

THANKS

Q / A