Analysis on Diabetic Patients' Hospital Admission & Classification of Readmission

A CSCI 271 Data Mining Project by:

AQUINO, Alec
ARROBIO, Anne
VILLAREAL, Rosiel

Outline

Introduction

Methodology

- Pre-processing
- Exploratory Data Analysis
- Data Preparation
- Classification & Evaluation

Insights

Conclusion

In a study of 4,769 patients, patients with *diabetes* were found to have a **40% increased risk of readmission** within 90 days.

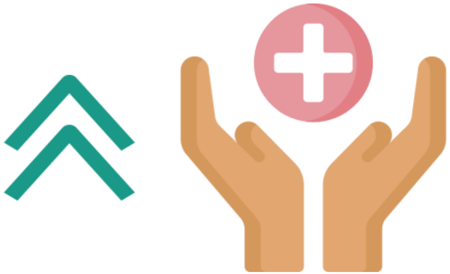
- Source: "Hospital Readmission of Patients with Diabetes" (Rubin, 2015)

$14,400 is the average cost of hospital readmission within 30 days

- Source: Healthcare Cost and Utilization Project (HCUP), which used 2010-2016 nationwide readmissions in US hospitals

Significance of the Study

Hospital readmission as an indicator of quality of patient care



Dataset

10 years of clinical care at 130 US hospitals including over 50 features representing patient and hospital outcomes of inpatients

1. diagnosed with diabetes,
2. admitted from 1-14 days,
3. received medications

patient number

age

gender

race

admission type

discharge disposition

time in hospital

medical specialty of admitting physician

HbA1c test result

Glucose serum test result

number of lab tests performed

diagnoses

number of medications

diabetic medications

number of visits in the year before hospitalization

How might we address hospital readmission to reduce overall health care costs and improve patient care?

We want to know…

What kinds of patients are readmitted within 30 days, more than 30 days?

What are the most important factors that affect their readmission?

…so that we can

Respond with patient-specific interventions

Address the factors that affect readmission

Pre-processing

Exploratory Data Analysis

Data Preparation

Classification & Evaluation

Pre-processing

- Retaining only **unique** patient encounters
- Removing encounters that resulted in discharge to **hospice** or **death**
- Dealing with **missing** values
- Categorizing **diagnosis** codes by type of disease

Pre-processing

- Retaining only **unique** patient encounters
  - First encounter
- Removing encounters that resulted in discharge to **hospice** or **death**
- Dealing with **missing** values
- Categorizing **diagnosis** codes by type of disease

Pre-processing

- Retaining only **unique** patient encounters
- Removing encounters that resulted in discharge to **hospice** or **death**
  - Assumption: these patients will not be readmitted anymore
- Dealing with **missing** values
- Categorizing **diagnosis** codes by type of disease

Pre-processing

- Retaining only **unique** patient encounters
- Removing encounters that resulted in discharge to **hospice** or **death**
- Dealing with **missing** values
    - Weight, Payer code, medical specialty
- Categorizing **diagnosis** codes by type of disease

Pre-processing

- Retaining only **unique** patient encounters
- Removing encounters that resulted in discharge to **hospice** or **death**
- Dealing with **missing** values
- Categorizing **diagnosis** codes by type of disease
    - Diagnosis codes are alphanumeric that lie in ranges

Pre-processing

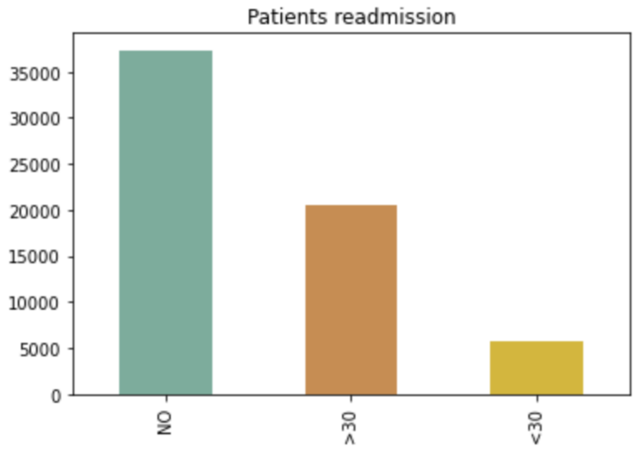
- Retaining only **unique** patient encounters
- Removing encounters that resulted in discharge to **hospice** or **death**
- Dealing with **missing** values
- Categorizing **diagnosis** codes by type of disease

Remaining data out of all encounters: **62.48%**

Remaining data out of unique patients: **88.9%**

Exploratory Data Analysis

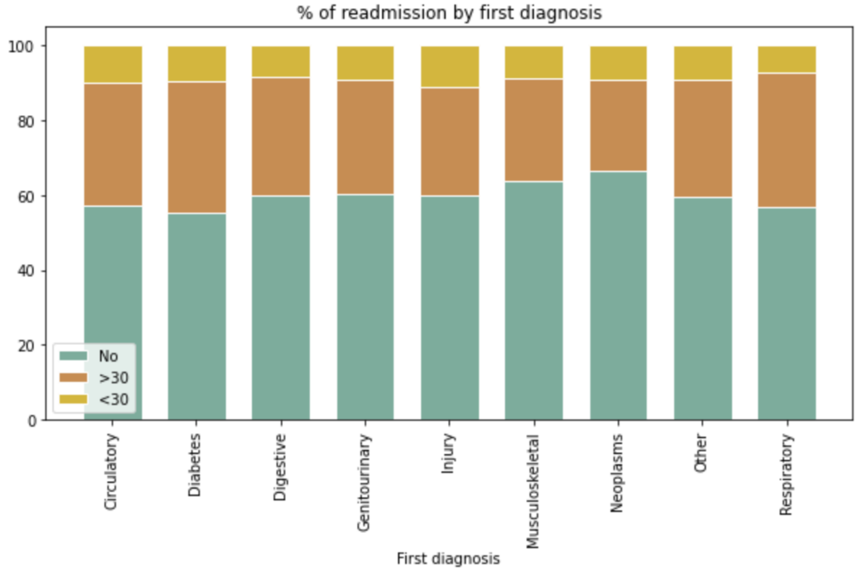
- There is a significant imbalance in the readmission figures
- **58.73%** of patients are not readmitted



	readmitted	counts	pcts
0	<30	5795	9.11
1	>30	20443	32.15
2	NO	37345	58.73

Exploratory Data Analysis

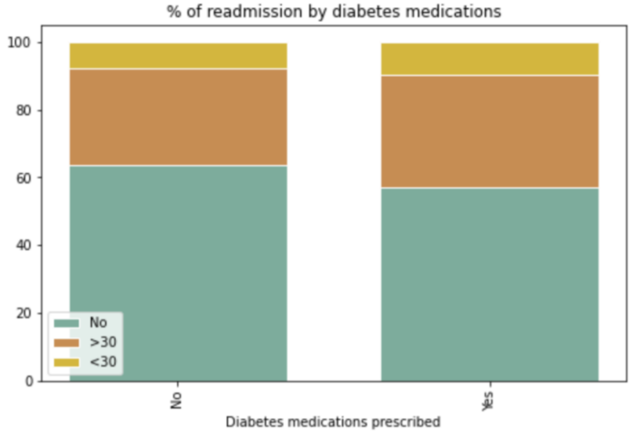
- Patients whose first diagnosis is Diabetes have the highest readmission rate at **44.61%**



Introduction Methodology Insights Conclusion

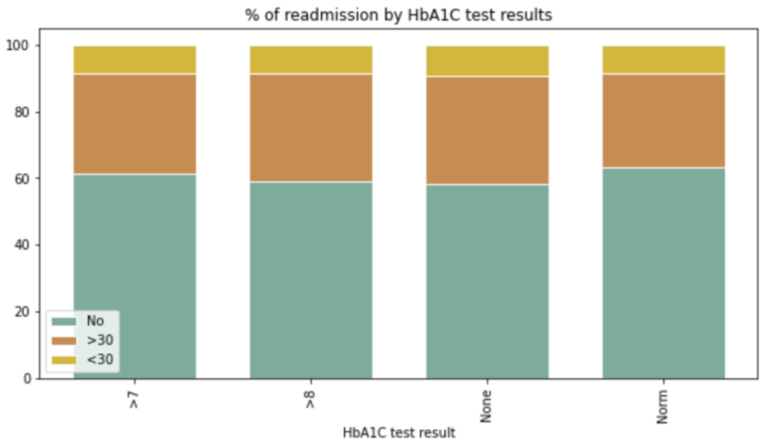
Exploratory Data Analysis

- Patients who were prescribed diabetes medications have **6.73%** higher readmission rate compared to those without



Exploratory Data Analysis

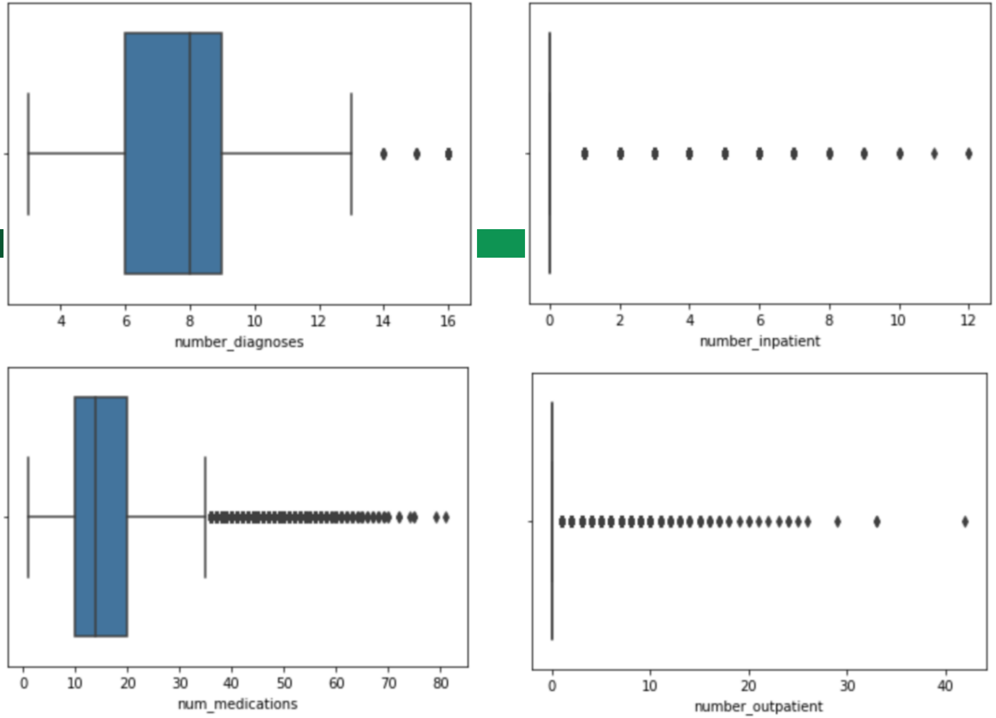
- Patients who were not tested for HbA1C have the highest readmission rate at **41.77%**
    - Only **0.93%** higher than patients with >8 result
    - **3.07%** higher than patients with >7 result
    - **5.19%** higher than patients with Normal result



Introduction Methodology Insights Conclusion

Exploratory Data Analysis

- There are a lot of outliers in the numeric features



Data Preparation

- One-hot encoding categorical features
- Scaling numeric features
- Splitting into train and test sets
- Resampling imbalanced classes*

* if classes for target variable are imbalanced

Data Preparation

- One-hot encoding categorical features

Patient #	Gender
1	Female
2	Male
3	Female

pd.get_dummies()

Patient #	Gender_Female	Gender_Male
1	1	0
2	0	1
3	1	0

Data Preparation

- Scaling numeric features

	time_in_hospital	num_lab_procedures	num_procedures
0	3	59	0
1	2	11	5
2	2	44	1

$$X_{scaled} = \frac{X - median}{Q3 - Q1}$$

preprocessing.RobustScaler()

	time_in_hospital	num_lab_procedures	num_procedures
0	-0.25	0.56	-0.5
1	-0.50	-1.36	2.0
2	-0.50	-0.04	0.0

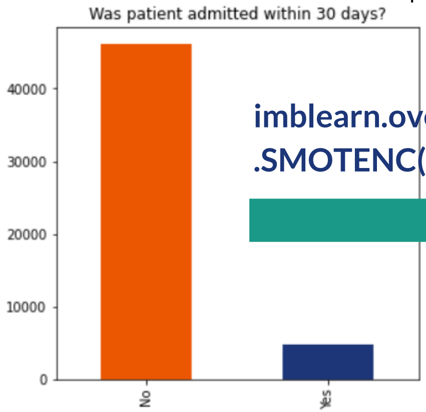
Data Preparation

- Splitting into train and test sets

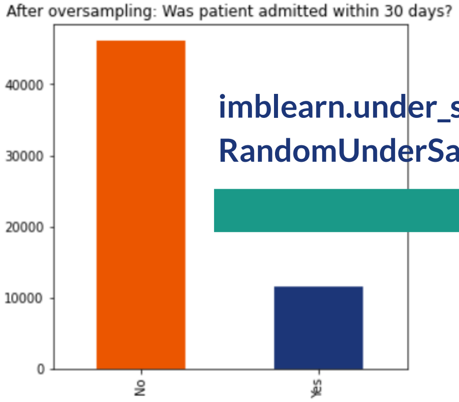
- Train: 80%
- Test: 20%

Data Preparation

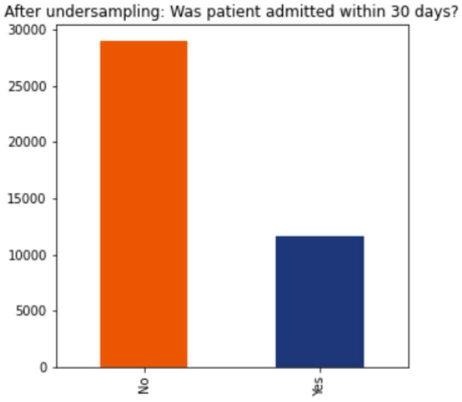
- Resampling imbalanced classes



imblearn.over_sampling
.SMOTENC()



imblearn.under_sampling.
RandomUnderSampler()

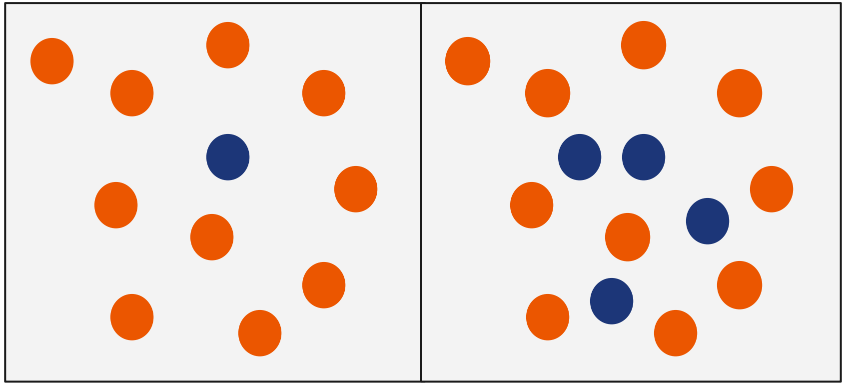


Data Preparation

- Resampling imbalanced classes

Ratio of min/maj =0.10

Ratio of min/maj =0.4

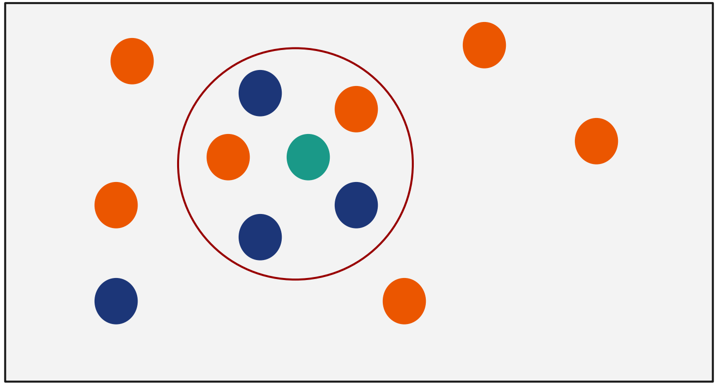


SMOTE oversampling:

Generate synthetic samples for the minority class.

Data Preparation

- Resampling imbalanced classes



SMOTE based on K-nearest neighbors:

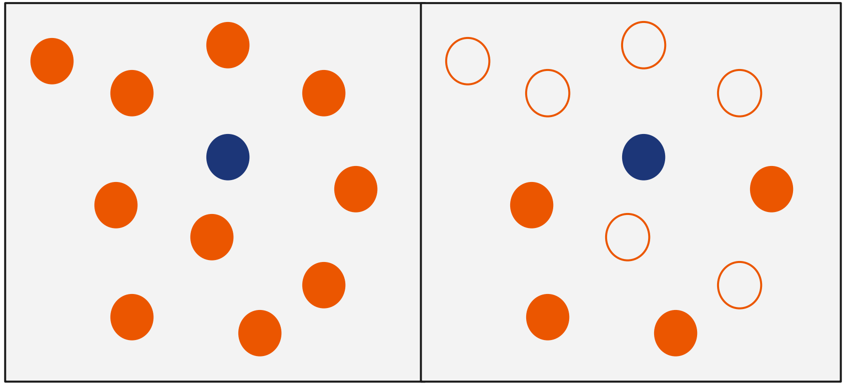
Classify the **point** using the majority of the 5 nearest neighbors (ex. k = 5).

Data Preparation

- **Resampling imbalanced classes**

Ratio of min/maj =0.10

Ratio of min/maj =0.25



Random undersampling:

Randomly pick samples from the majority class.

1. Experiment with different groupings of readmission

- 3 classes (**2**: <30 days, **1**: >30 days, **0**: Not)
- Readmission (**1**: <30 days + >30 days, **0**: Not)
- Early readmission (**1**: <30 days, **0**: >30 days + Not)

Classification & evaluation

2. Experiment with different classifiers

- Logistic regression
- Random forest
- Neural networks

Classification & evaluation

Random Forest Model

Concept

Bagging - bootstrapping the data and aggregating the results to make a decision

Random Forest - uses the bagging method where the model trains a series of decision trees parallel to each other and aggregates the results to get a final decision

Random Forest Model

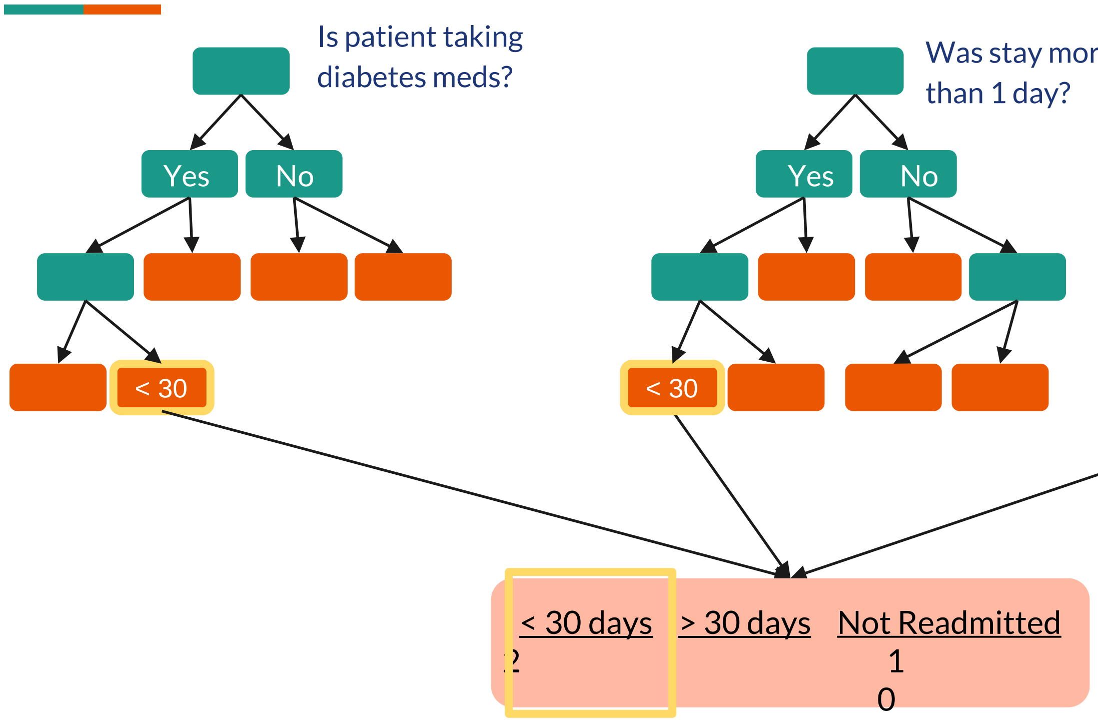
Concept

Step 1:
Bootstrap resampling

Step 2:
Create a decision tree using the bootstrapped data set

Step 3:
Repeat Steps 1 & 2

Result:
The majority decision of the trees is the final decision of the model.


Is patient taking diabetes meds?
Yes
No
< 30
Was stay more than 1 day?
Yes
No
< 30
> 30
< 30 days
> 30 days
Not Readmitted
2
1
0

Random Forest Model

Implementation

n_estimators: 50

max_depth: 10

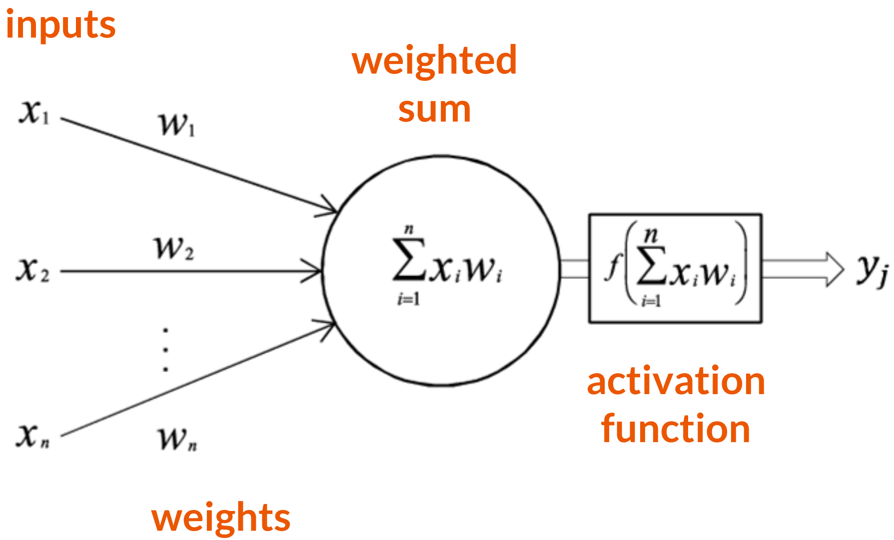
max_features: sqrt

min_samples_split: 5

min_samples_leaf: 5

Neural Networks

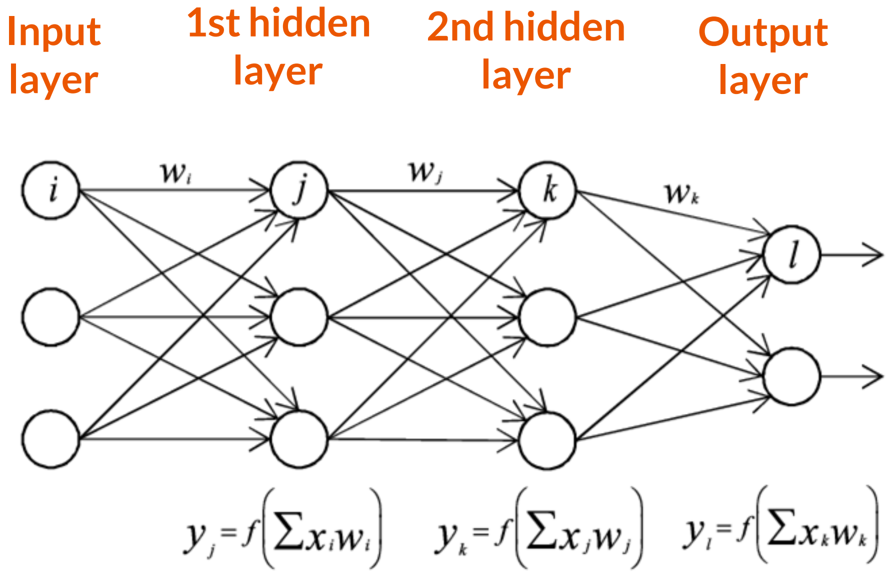
Concept



Perceptron

Neural Networks

Concept



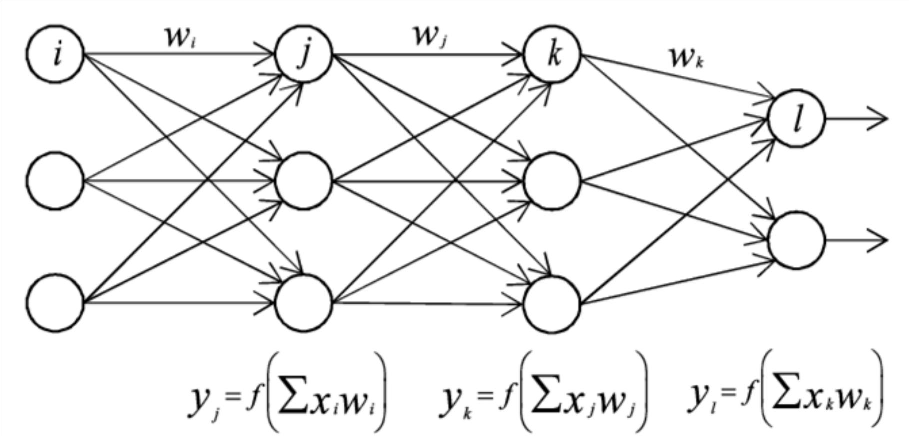
Neural Network

Input layer

1st hidden layer

2nd hidden layer

Output layer



Neural Network

Step 1: Random initialization

Step 2: Forward Propagate

- Inputs transformed by weights
- Weighted Sum transferred by an activation function
- Final Activation as last hidden layer

Step 3: Backpropagate

- Loss is calculated
- Results are used to update the weights

Step 4: Iterate until convergence

3 Labels

Implementation

Loss Function: Sparse Categorical Crossentropy

Activation Function: ReLu, Softmax

Optimizer: Adam

Batch Size: 32

Validation Split: 0.2

Epochs: 1000

Early Stop Patience: 50

3 Labels

Input Layer :173
Layer1 : 512
Layer2 : 256
Layer3 : 128
Layer4 : 64
Layer 5 : 32
Output Layer : 3
Drop Out : 0.4

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	89088
dropout (Dropout)	(None, 512)	0
batch_normalization (BatchNo	(None, 512)	2048
dense_1 (Dense)	(None, 256)	131328
activation (Activation)	(None, 256)	0
dropout_1 (Dropout)	(None, 256)	0
batch_normalization_1 (Batch	(None, 256)	1024
dense_2 (Dense)	(None, 128)	32896
activation_1 (Activation)	(None, 128)	0
dropout_2 (Dropout)	(None, 128)	0
batch_normalization_2 (Batch	(None, 128)	512
dense_3 (Dense)	(None, 64)	8256
activation_2 (Activation)	(None, 64)	0
dropout_3 (Dropout)	(None, 64)	0
batch_normalization_3 (Batch	(None, 64)	256
dense_4 (Dense)	(None, 32)	2080
activation_3 (Activation)	(None, 32)	0
dropout_4 (Dropout)	(None, 32)	0
batch_normalization_4 (Batch	(None, 32)	128
dense_5 (Dense)	(None, 3)	99
activation_4 (Activation)	(None, 3)	0

2 Labels

Implementation

Loss Function: Binary Cross Entropy

Activation Function: ReLu, Sigmoid

Optimizer: Adam

Batch Size: 32

Validation Split: 0.2

Epochs: 1000

Early Stop Patience: 50

2 Labels

Input Layer :173
Layer1 : 512
Layer2 : 256
Layer3 : 128
Layer4 : 64
Layer 5 : 32
Output Layer : 1
Drop Out : 0.4

Layer (type)	Output Shape	Param #
dense_12 (Dense)	(None, 512)	89088
dropout_10 (Dropout)	(None, 512)	0
batch_normalization_10 (Batc	(None, 512)	2048
dense_13 (Dense)	(None, 256)	131328
activation_10 (Activation)	(None, 256)	0
dropout_11 (Dropout)	(None, 256)	0
batch_normalization_11 (Batc	(None, 256)	1024
dense_14 (Dense)	(None, 128)	32896
activation_11 (Activation)	(None, 128)	0
dropout_12 (Dropout)	(None, 128)	0
batch_normalization_12 (Batc	(None, 128)	512
dense_15 (Dense)	(None, 64)	8256
activation_12 (Activation)	(None, 64)	0
dropout_13 (Dropout)	(None, 64)	0
batch_normalization_13 (Batc	(None, 64)	256
dense_16 (Dense)	(None, 32)	2080
activation_13 (Activation)	(None, 32)	0
dropout_14 (Dropout)	(None, 32)	0
batch_normalization_14 (Batc	(None, 32)	128
dense_17 (Dense)	(None, 1)	33
activation_14 (Activation)	(None, 1)	0

- 3 classes (2: <30 days, 1: >30 days, 0: Not)
- Readmission (1: <30 days + >30 days, 0: Not)
- Early readmission (1: <30 days, 0: >30 days + Not)

	Logistic Regression	Random Forest	Neural Networks
Accuracy	58.69%	57.09%	59.53%
F1-score	56.10%	53.84%	47.00%

Classification & evaluation

- 3 classes (2: <30 days, 1: >30 days, 0: Not)
- **Readmission (1: <30 days + >30 days, 0: Not)**
- Early readmission (1: <30 days, 0: >30 days + Not)

	Logistic Regression	Random Forest	Neural Networks
Accuracy	62.16%	62.2%	59.76%
F1-score	60.55%	60.75%	60.00%

Classification & evaluation

- 3 classes (2: <30 days, 1: >30 days, 0: Not)
- Readmission (1: <30 days + >30 days, 0: Not)
- **Early readmission (1: <30 days, 0: >30 days + Not)**

	Logistic Regression	Random Forest	Neural Networks
Accuracy	90.33%	90.77%	88.68%
F1-score	86.56%	86.73%	89.00%

Classification & evaluation

Most Significant Features: Early Readmission

Logistic Regression	Random Forest	Neural Network
chlorpropamide_No	*time_in_hospital*	*diag_1_desc_Circulatory*
tolazamide_No	*age_[50-60)*	*discharge_disp_Discharged/transferred to another short term hospital*
tolbutamide_No	*number_inpatient*	*diag_3_desc_Circulatory*
acarbose_No	*diabetesMed_Yes*	*discharge_disp_Discharged/transferred to ICF*
glipizide-metformin_No	*discharge_disp_Discharged/transferred to another short term hospital*	*diag_2_desc_Circulatory*

Insights and Observations

The kinds of patients that were readmitted early (within 30 days) were:

- admitted longer and had more inpatient visits in the past year
- taking diabetes meds and had a change in meds
- not taking specific diabetes meds
- diagnosed with circulatory diseases
- discharged to another short-term hospital or to an Intermediate Care Facility

Recommendations to hospitals based on most significant features

- Prepare bed and room allocation in advance
- Prepare resources needed in advance
- Double check whether the patient actually needs certain medications
- Take note of discharge disposition

Recommendations to improve modeling

- More Features and/or More Complete Data
  - Weight
  - Medical Specialty
  - Payer Code
- Feature Engineering
- Other methods to address imbalance in the dataset

Thank you for listening!

Questions?