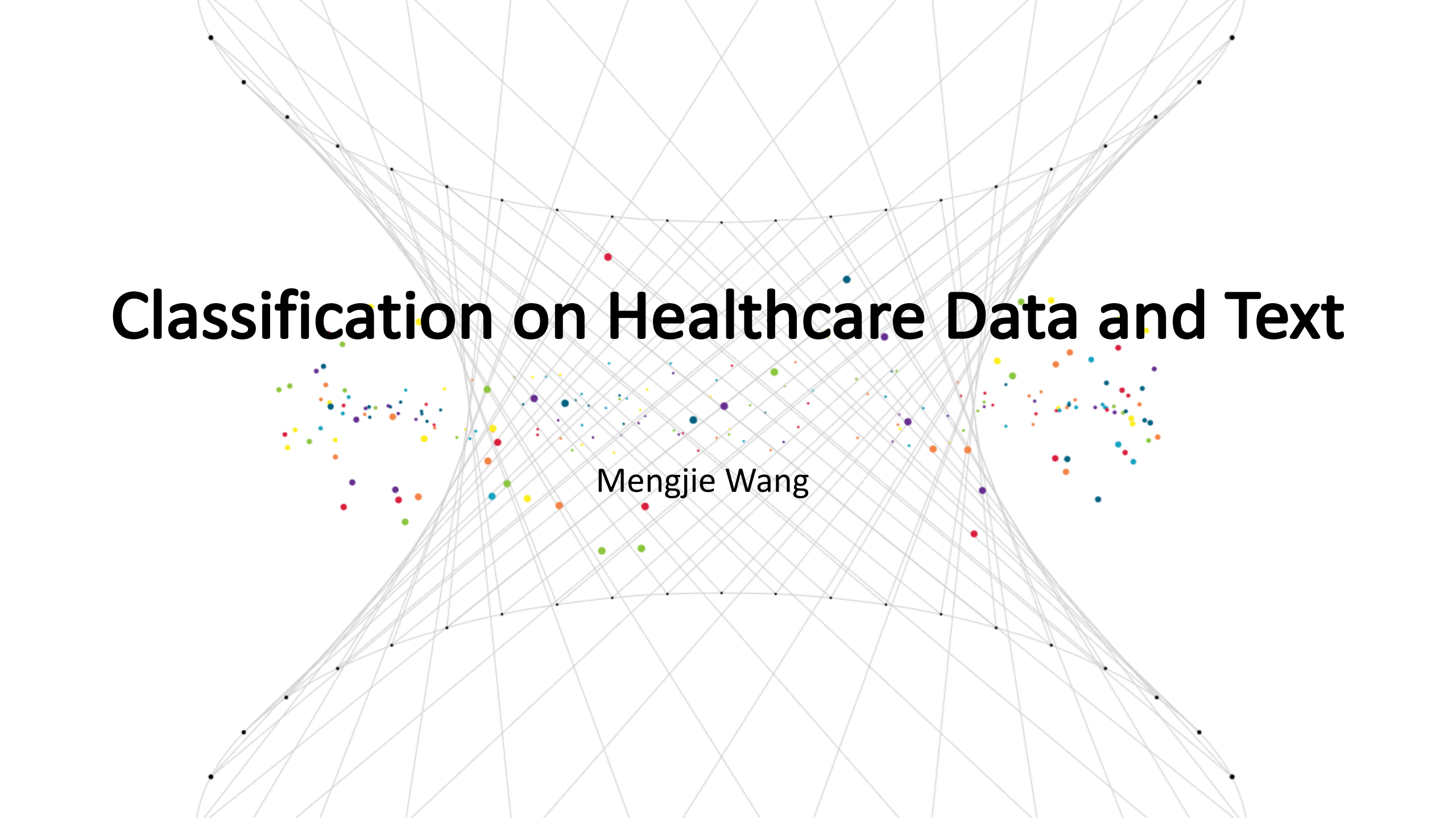


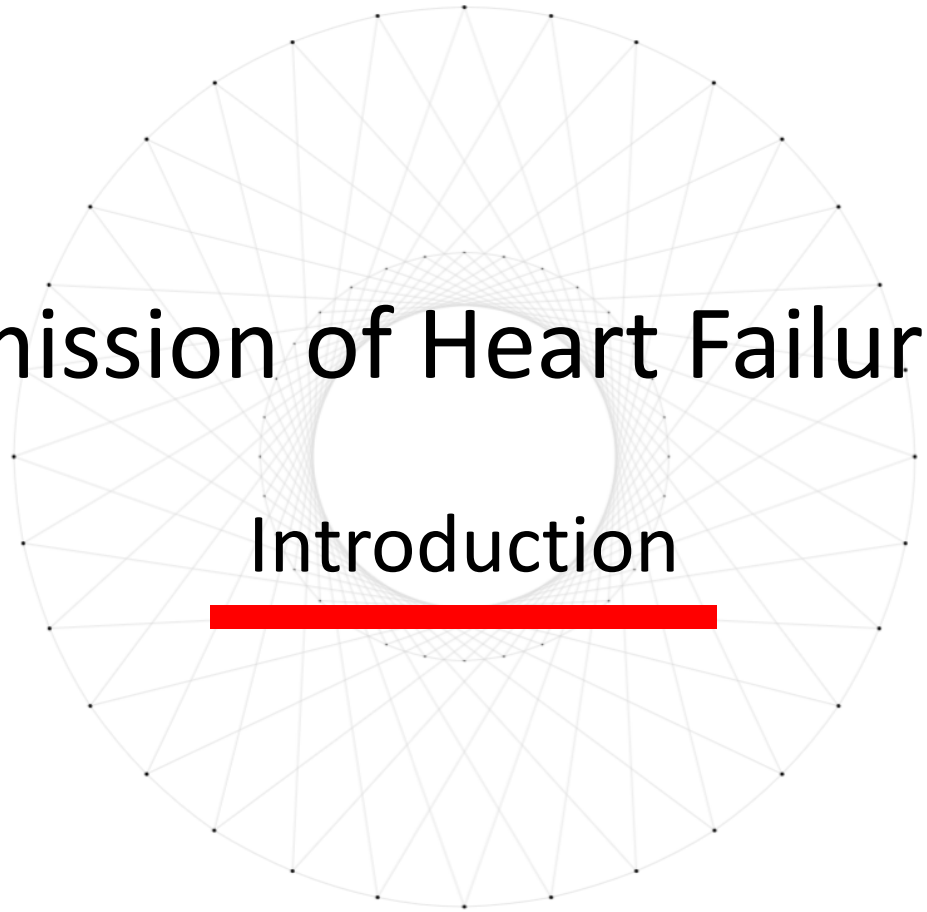
Classification on Healthcare Data and Text

Mengjie Wang

PROJECTS

Project 1: Readmission of Heart Failure Patients

Project 2: Sentiment Analysis of Yelp Review



Readmission of Heart Failure Patients

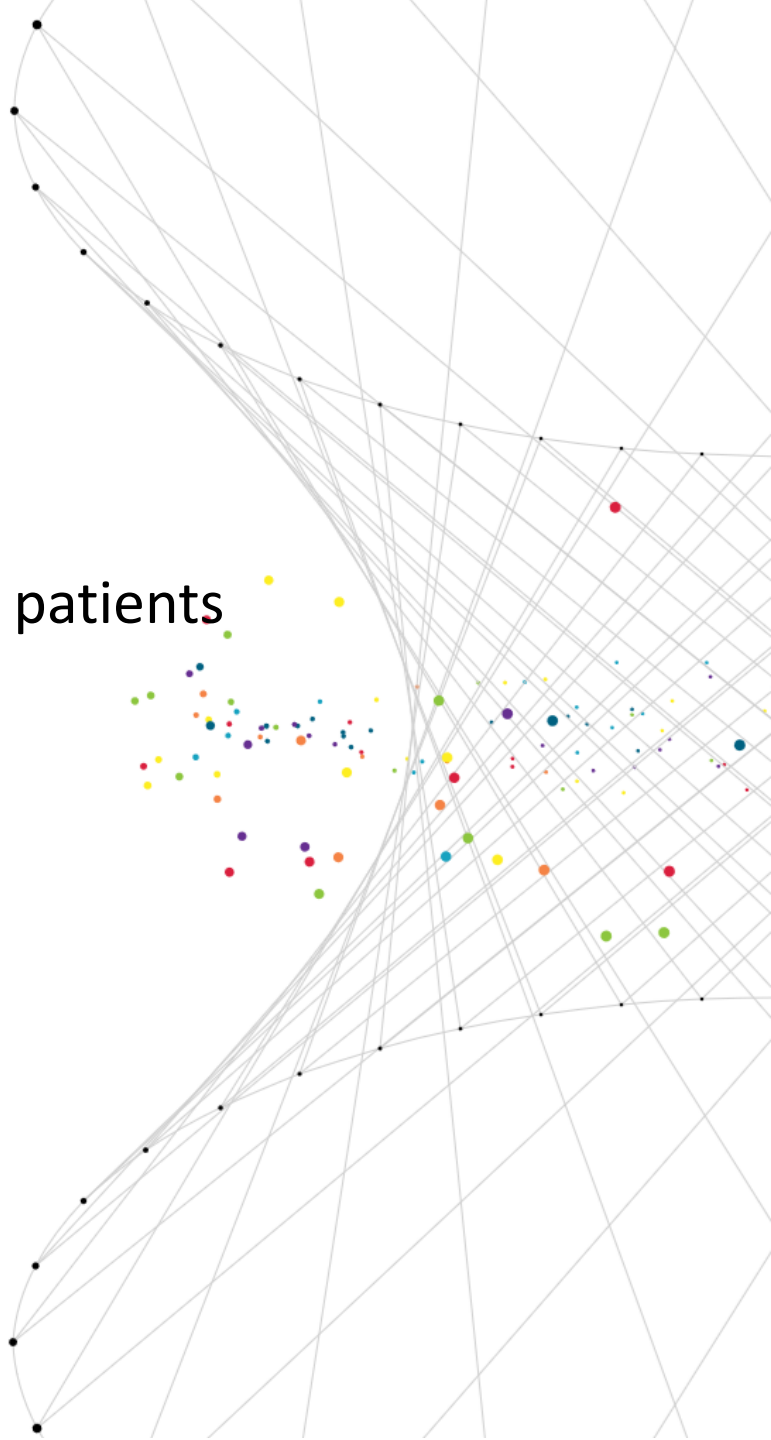
Introduction

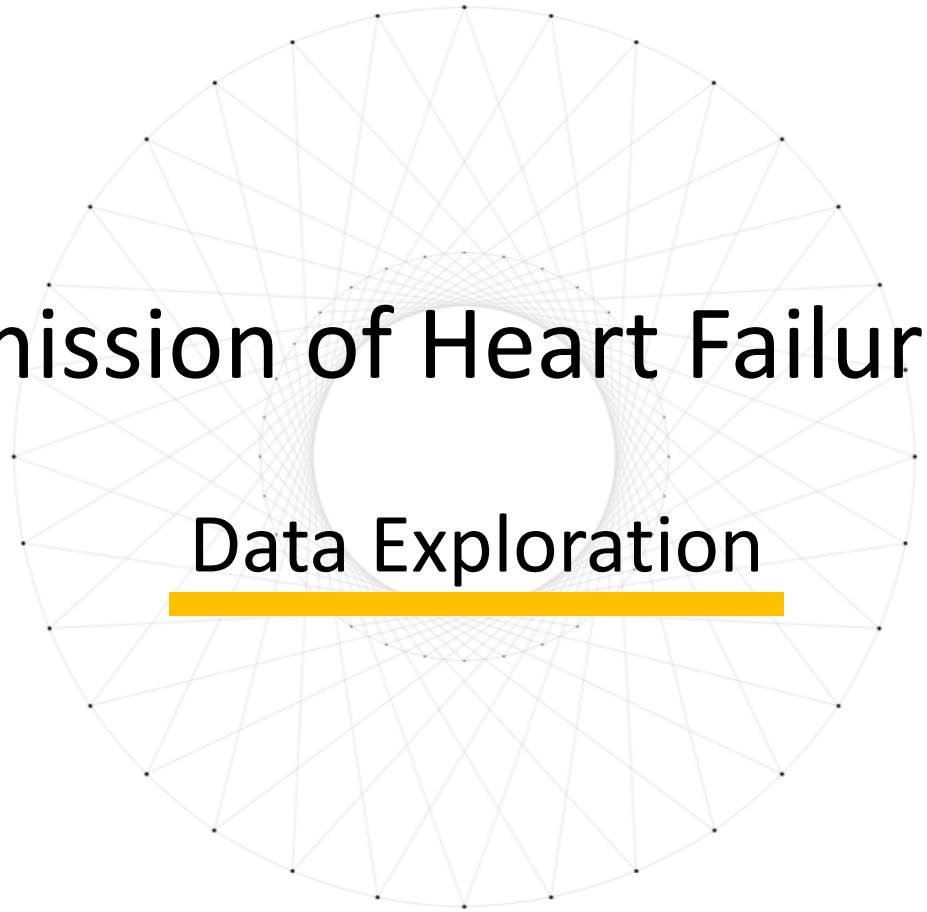
Overarching Goal

- Allocate resources
- Reduce the 30-day readmission rate of heart failure patients

Modeling Goal

- Identify important features
- Predict the risk of readmission





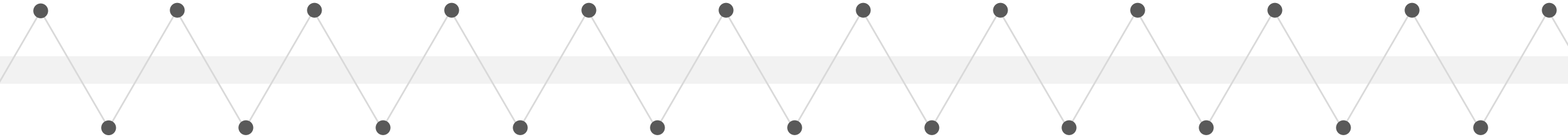
Readmission of Heart Failure Patients

Data Exploration

Data Exploration

Sample size: 4320

Number of readmission: 605 (14%)



Number of features: 832
Categorical features: 331
Numerical features: 501

Dictionary of features is not consistent with the raw data

Based on the understanding of the data

Remove repeated information

Binary indicator of Emergency Department visits in the Prior 30 Days

vs

Count of Emergency Department visits in the Prior 30 Days

Create new features

Merge several original features as a new feature

Count of orders for class Antacids within 12 months prior to admission

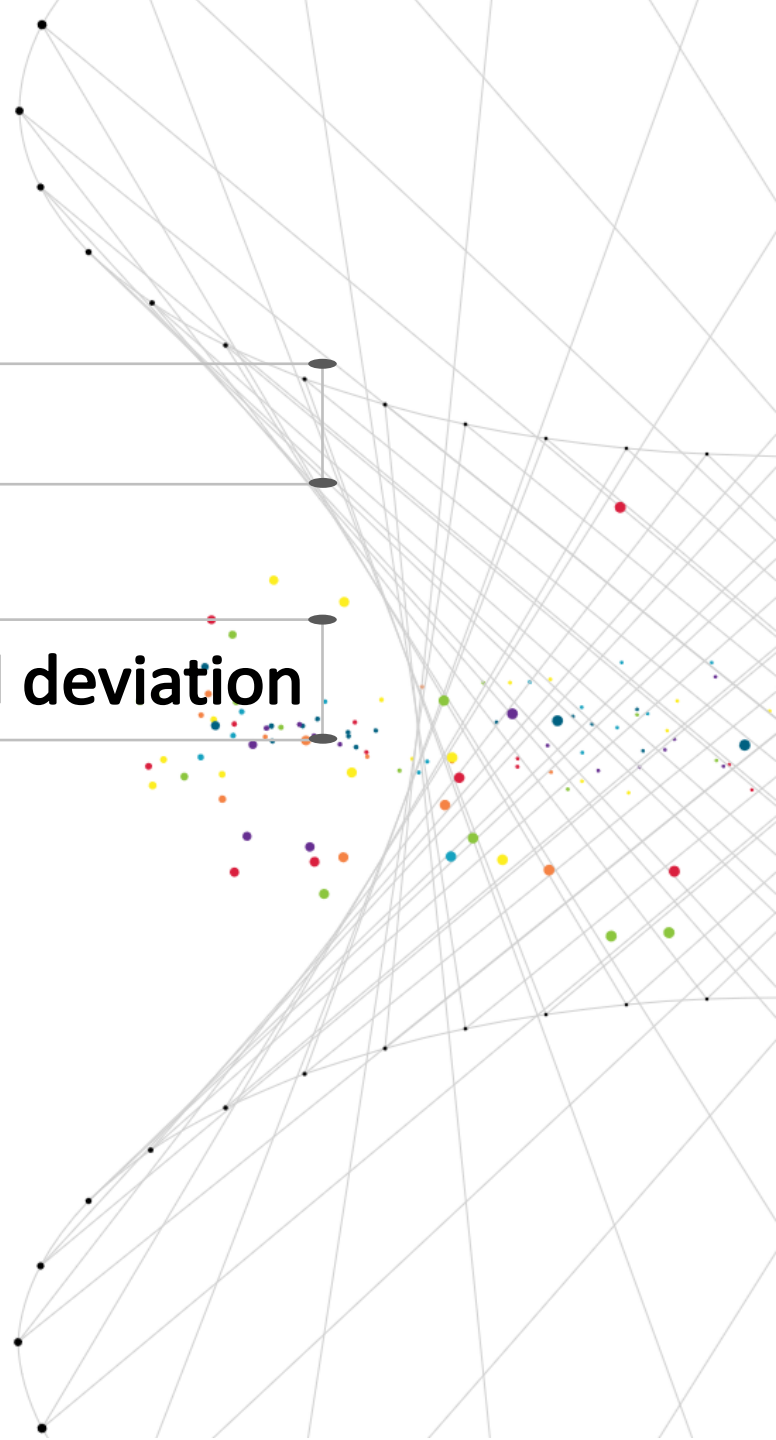
Count of orders for class Antidotes within 12 months prior to admission

Based on the descriptive statistics

Remove features containing same value

Remove 10% features with lowest standard deviation

Number of features remained: 421



Data Exploration

Readmit	Gender	Age	9 Level Factor for BMI	Binary Indicator of whether or not the patient had an Advanced Care Plan on Admit	.....	Count of Emergency Department visits in the prior 30 days	Count of Impatient visits in the prior 60 days	Count of Observation visits in the prior 365 days	.....
1	M	90+	2	1		0	1	0	

Readmit Indicator

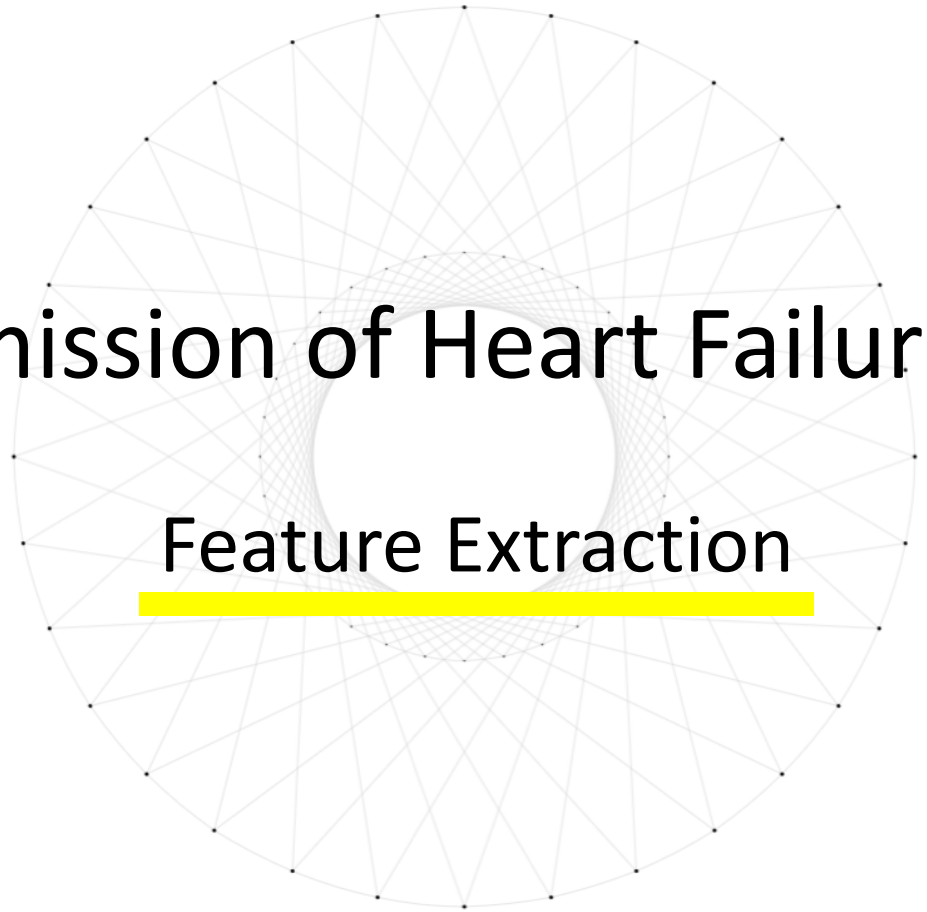
Basic Information

Historical Visiting Hospital Experience

Count of times a diagnosis associated with Acute Renal Failure within 365 days prior to admission or on the problem list (HCC)	Binary indicator of CMI Code Chest pain present in diagnoses or problem list	.....	Raw count of order for Pharmaceutical class Antacids during visit	Raw count of order for Pharmaceutical class Antacids within 12 months prior to admission	.....
0	1		0	0	

Historical and Current Health Status

Historical and Current Medicine Record



Readmission of Heart Failure Patients

Feature Extraction

Feature Importance in Random Forest

Calculated by the sample not selected by bootstrapping (out-of-bag data)

- For each tree m, use the out-of-bag data as the testing set to obtain the prediction error: Err_0^m
- For each variable j, randomly permute its value among the testing samples, and recalculate the prediction error: Err_j^m
- Calculate for each j

$$FI_{m,j} = \frac{Err_j^m}{Err_0^m} - 1$$

- Average $FI_{m,j}$ across all trees to get the importance of each variable $FI_{.j}$

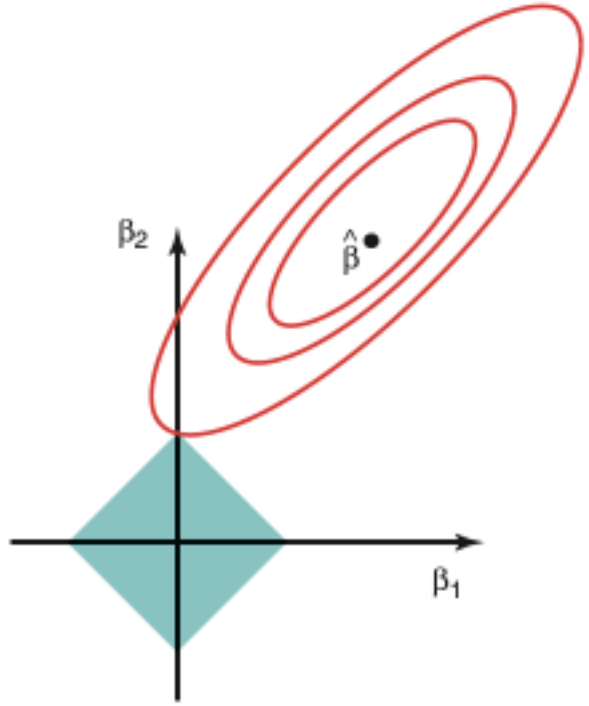
Coefficients in Logistic + LASSO

Objective function:

$$min_{\beta} \quad -\sum_{i=1}^{n}[y^i log\hat{y}^i + (1-y^i)\log(1-\hat{y}^i)] + \lambda\sum|\beta|$$

$$min_{\beta} \quad -\sum_{i=1}^{n}[y^i log\hat{y}^i + (1-y^i)\log(1-\hat{y}^i)]$$

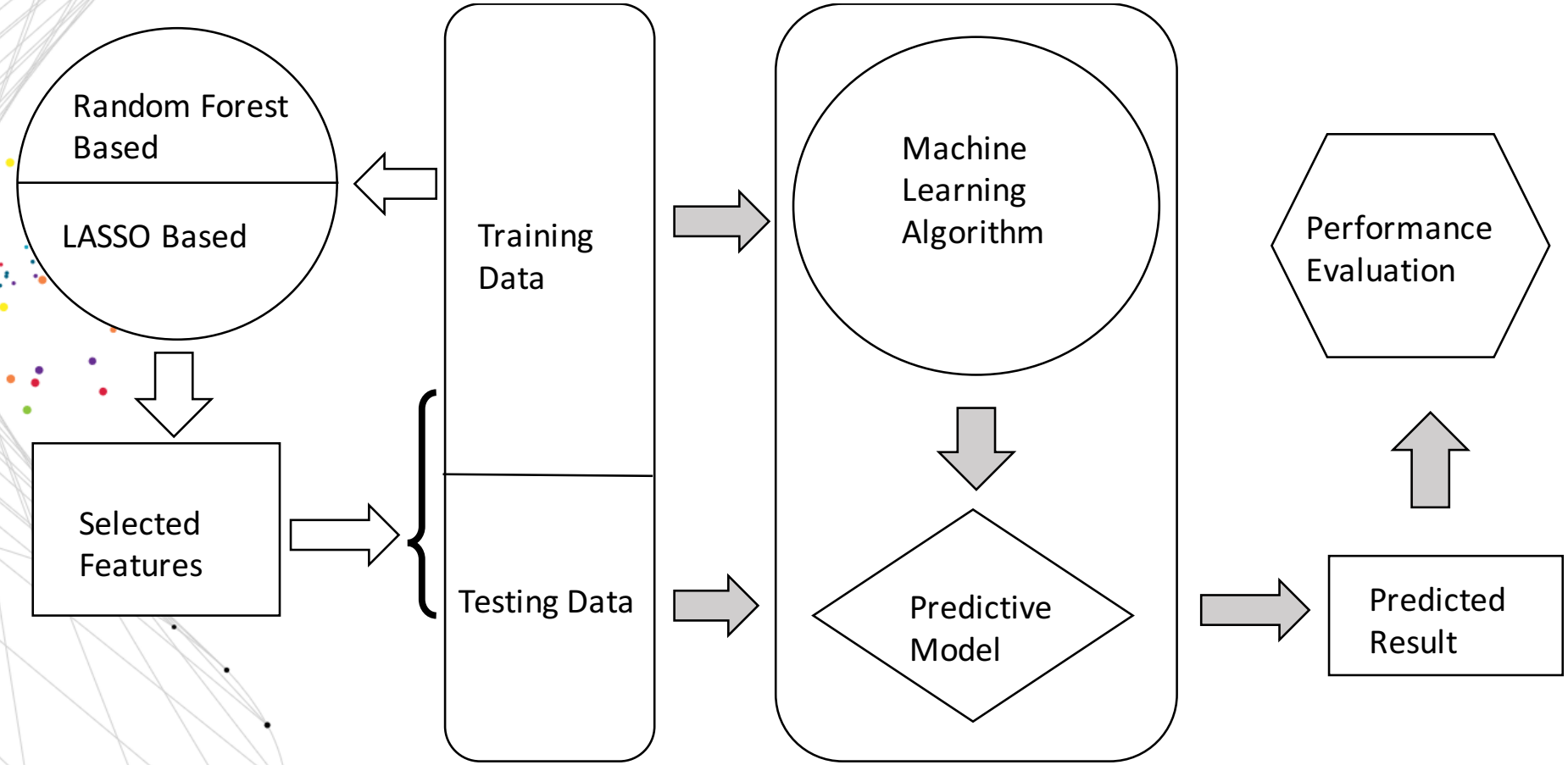
subject to $\lambda\sum|\beta| \leq s$

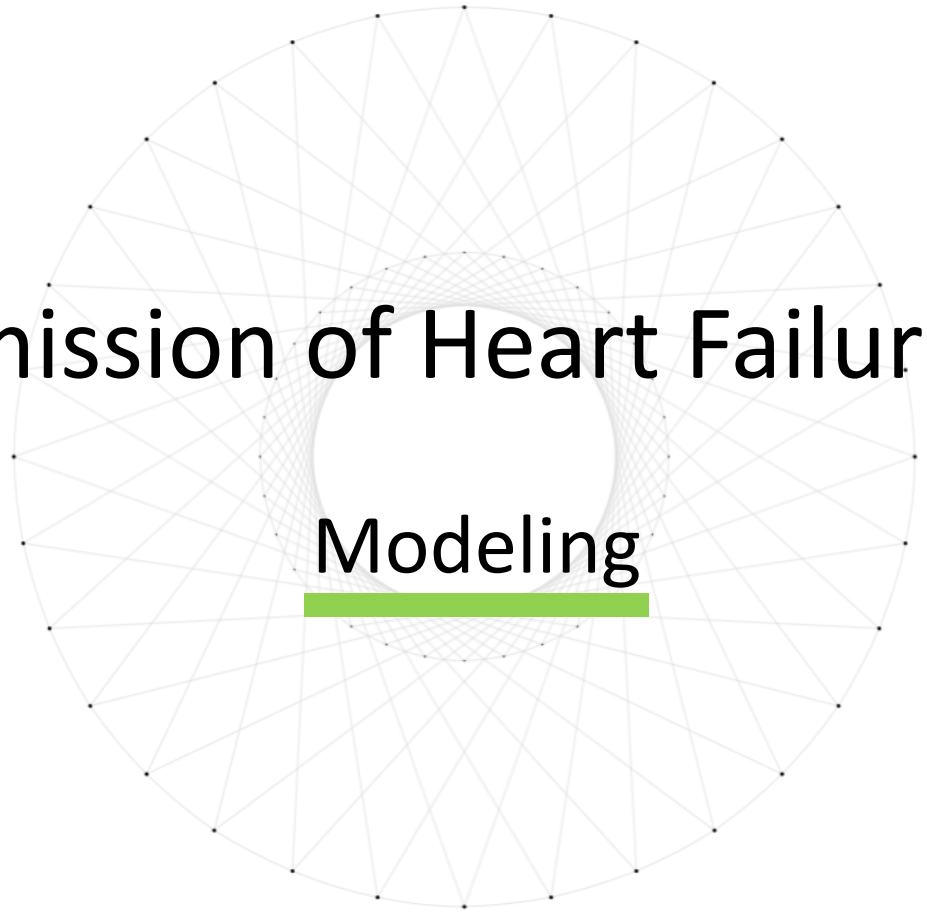


All vs selected features

AUC	With All Features	Random Forest-based with 30 Features	LASSO-based with 30 features
Random Forest	0.72068	0.71594	0.71974
LASSO	0.71753	0.71688	0.71747
Naïve Bayes	0.70193	0.69871	0.70083

Flowchart:


Random Forest Based
LASSO Based
Selected Features
Training Data
Testing Data
Machine Learning Algorithm
Predictive Model
Performance Evaluation
Predicted Result



Readmission of Heart Failure Patients

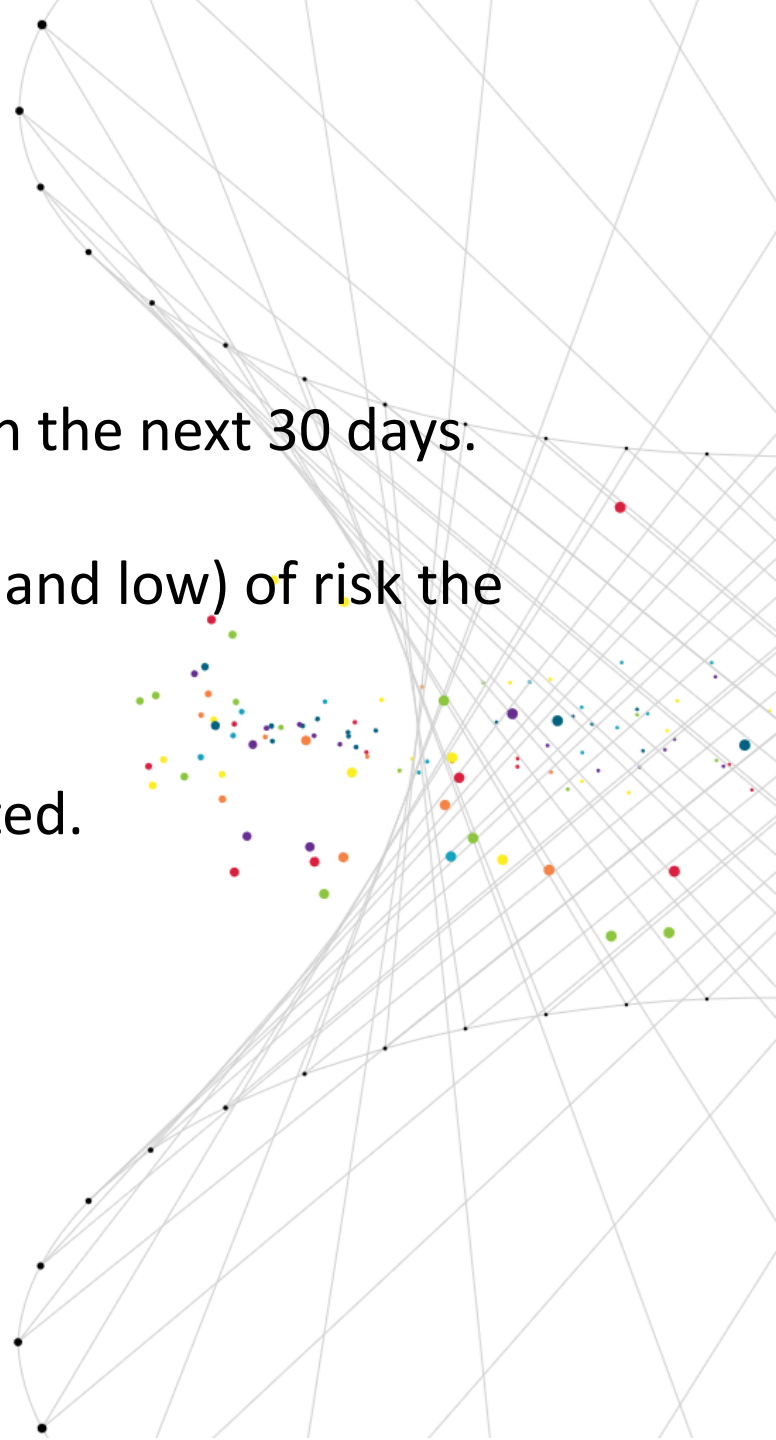
Modeling

Target:

A. Whether or not the patient will be readmitted in the next 30 days.

B. The category (High, medium high, medium low, and low) of risk the patient belongs to?

C. The probability that the patient will be readmitted.

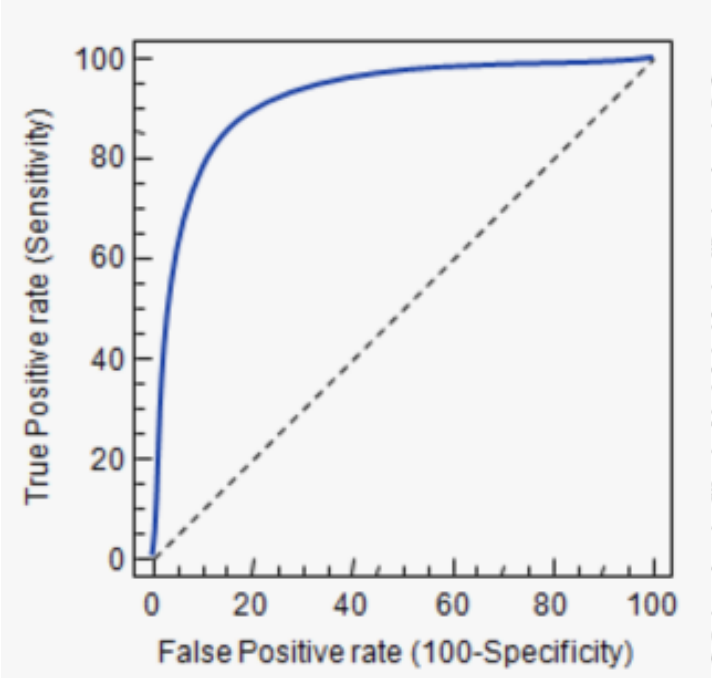


Evaluation:

Imbalanced data: 14%

$$\text{Accuracy} = \frac{\text{Number of Correct Labels}}{\text{Total Number of Labels}}$$

AUC(ROC) = Area Under Curve (of Receiver Operating Characteristic)

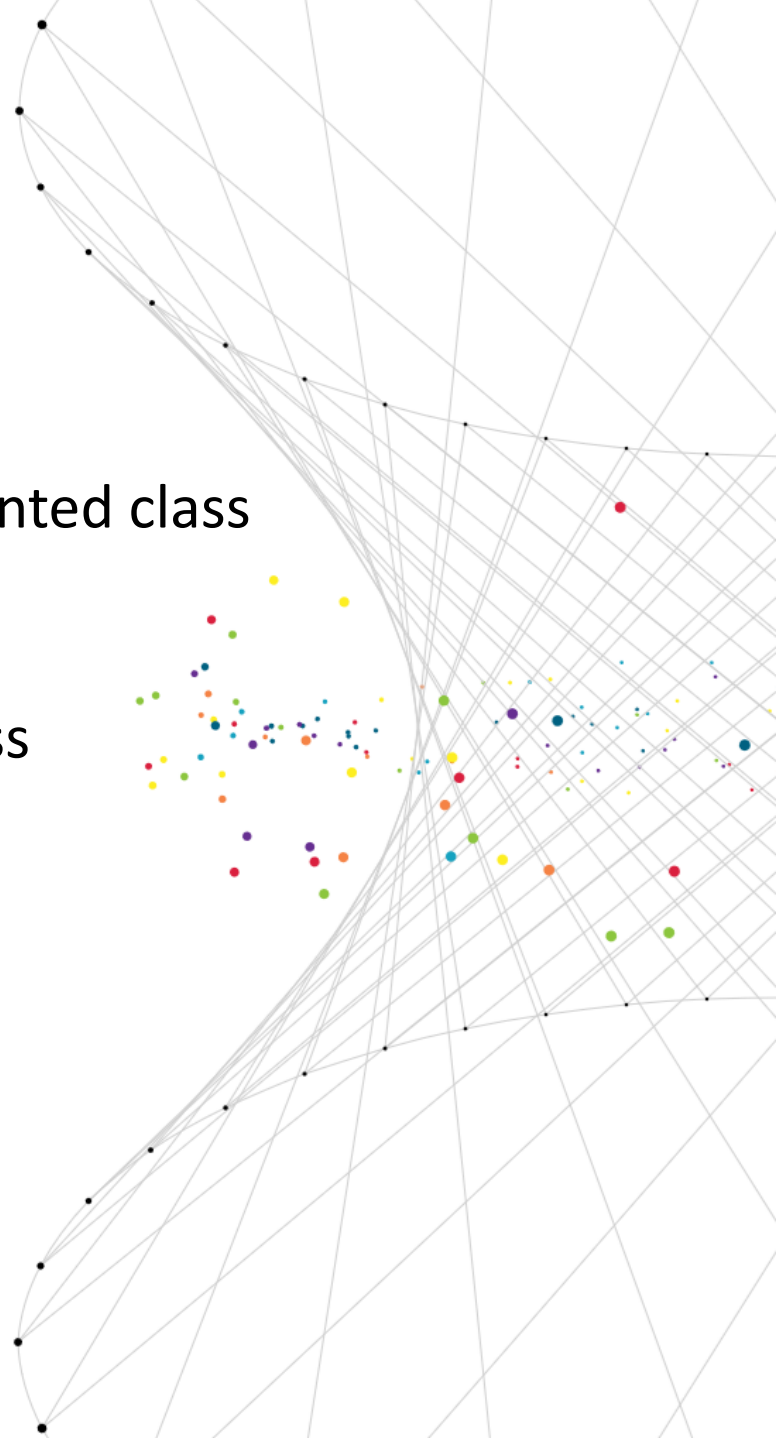


Resampling:

Over-sampling:
Add copies of observations from the under-represented class

Under-sampling:
Delete observations from the over-represented class

Adjusted proportion of readmission: 30%



Modeling

Algorithms:

- Logistic Regression
  - Ridge
  - LASSO
- Naïve Bayes
- Neural Network
- SVM
- K Nearest Neighbors

Generate Single Probability

$$P(\text{readmitted}|\text{Features})$$

- Ensembled Model
  - Bagging with LASSO
  - Stacking with LASSO,RF,KNN
- Ensembled Trees
  - Random Forest
  - Xgboost

Using voting, weighted average or other model to get probability

Hyper-parameters is found via Grid Search method

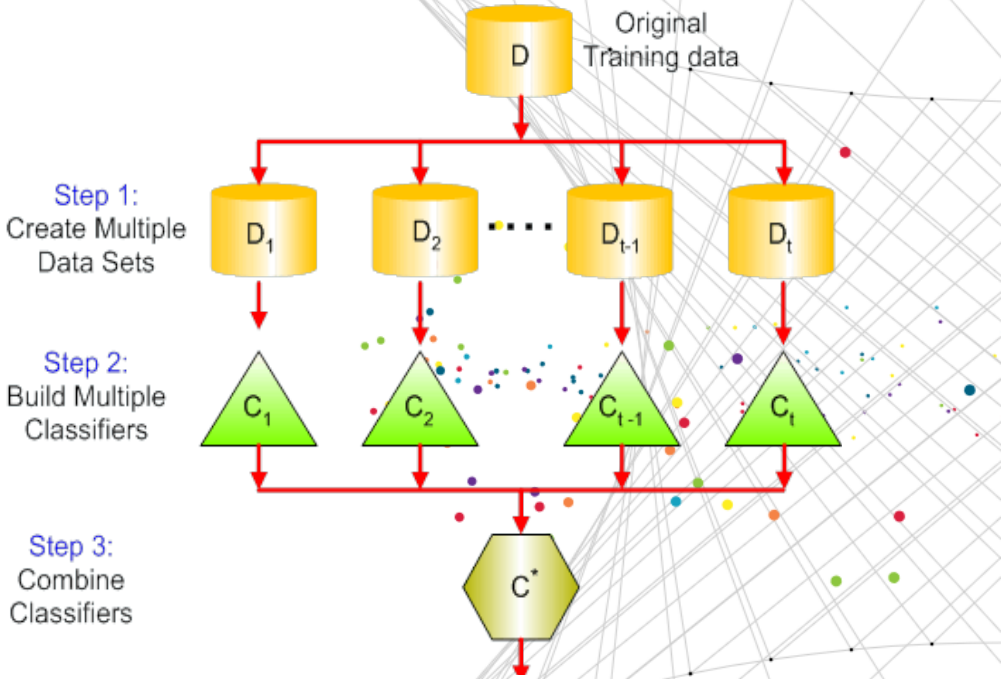
Ensemble Models:

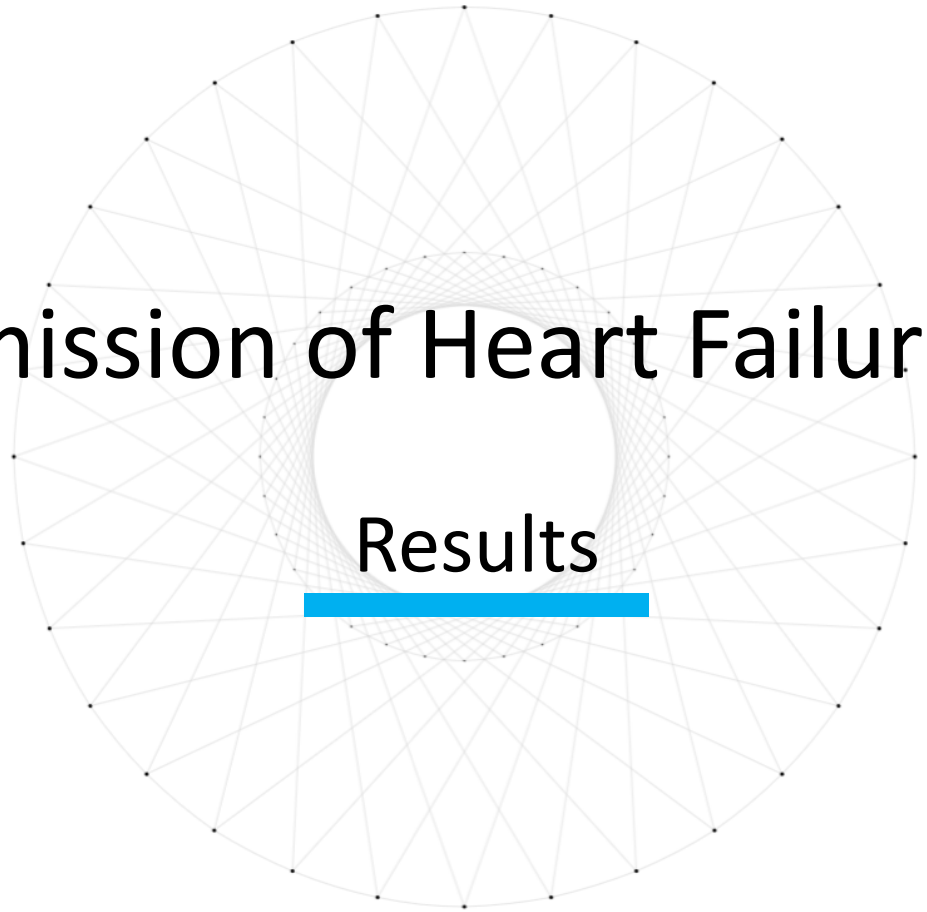
Step 1: Random sampling with replacement

Step 2: Train the multiple classifiers

Step 3: Train the Meta-Classifier

- Bagging: Use weighted Average
  e.g. Bagging with LASSO
- Stacking: Use Logistic Regression
  e.g. Stacking with LASSO,RF,KNN





Readmission of Heart Failure Patients

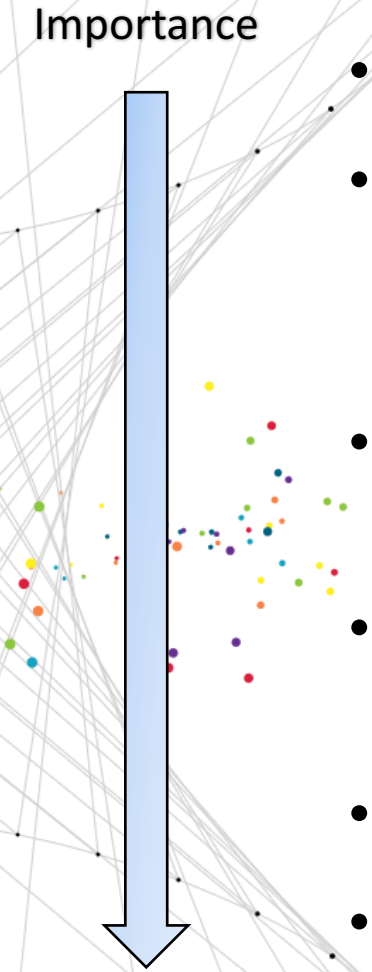
Results

Results

Models	Mean	Standard Deviation	Minimum	Maximum	Running time
Ridge	0.7271632	0.03028427	0.6691973	0.7542462	0.26100 sec
LASSO	0.7476443	0.01864501	0.7211446	0.7686505	0.05900 sec
Naïve Bayes	0.7331464	0.02706958	0.7028235	0.7600289	0.08100 sec
K Nearest Neighbor	0.7427606	0.02595915	0.7015081	0.7661012	9.50399 sec
SVM (Gaussian)	0.6970617	0.01993623	0.6542463	0.717907	38.98800 sec
Neural Network	0.7320775	0.02104812	0.6635383	0.75664068	90.55882 sec
Random Forest	0.7525275	0.01968679	0.7229283	0.7764427	5.05599 sec
Xgboost	0.7423987	0.00793337	0.7355731	0.7487238	8.2000 sec
Bagging LASSO	0.7516213	0.01869591	0.7270531	0.7781743	10.49600 sec
Stacking Algorithm	0.7528694	0.01882238	0.7257153	0.7812234	18.58599 sec

AUC on ten different samples under 10-flold cross validation

Results



- Count of Inpatient visits in the Prior 180 Days
- Count of times a diagnosis associated with Congestive Heart Failure was present within the 365 days prior to admission or active on the problem list
- Count of orders for class Antihyperlipidemic within 12 months prior to admission
- Count of orders for class Antihypertensive within 12 months prior to admission
- Count of orders for class Beta Blockers during visit
- Count of orders for class Analgesics Non Narcotic during visit

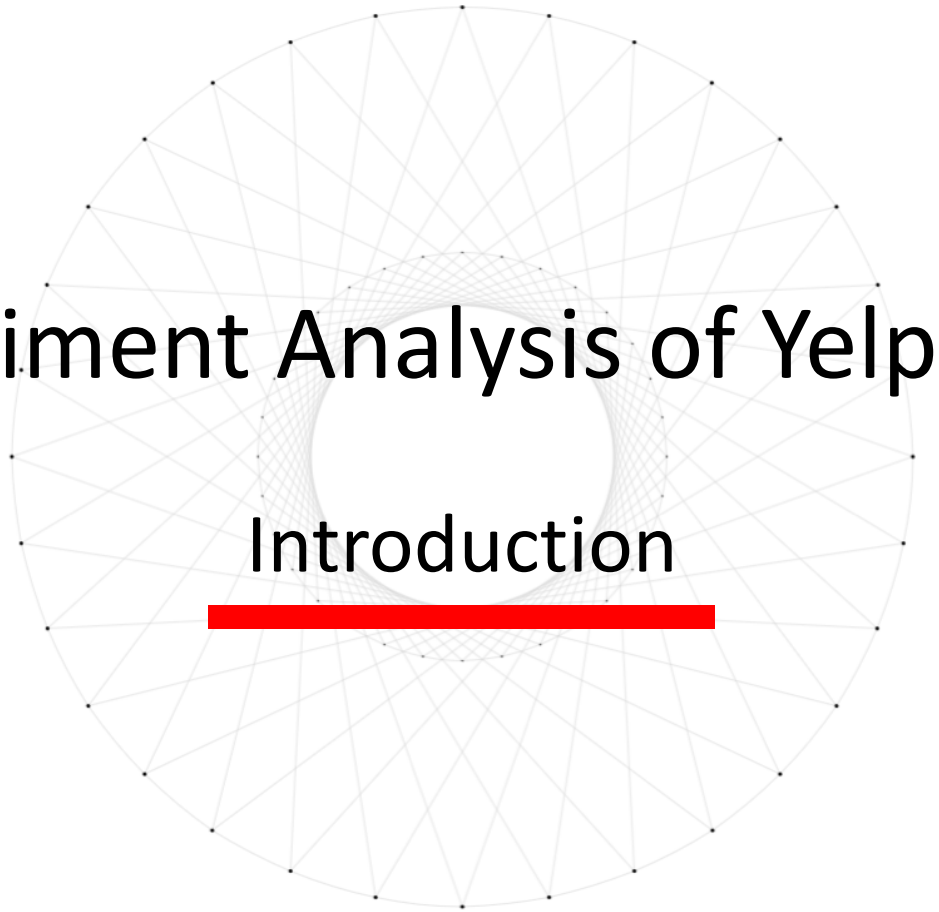
Results

- For the data:
  - ➢ Increase the feature space:

    Add Physical, Clinical Data and follow-up status
  - ➢ Increase the sample space

- For the Model:
  - ➢ Top 3 models:

    Random Forest, Bagging LASSO, Stacking Algorithm
  - ➢ Combine new features and selected features to test the performance



Sentiment Analysis of Yelp Review

Introduction

Dataset

Yelp Dataset Challenge (Round 11)

business, check-in, review, tip, user

The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification

Maybe you've heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?




Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user's relationships define their usage patterns? Where are the trend setters eating before it becomes popular?

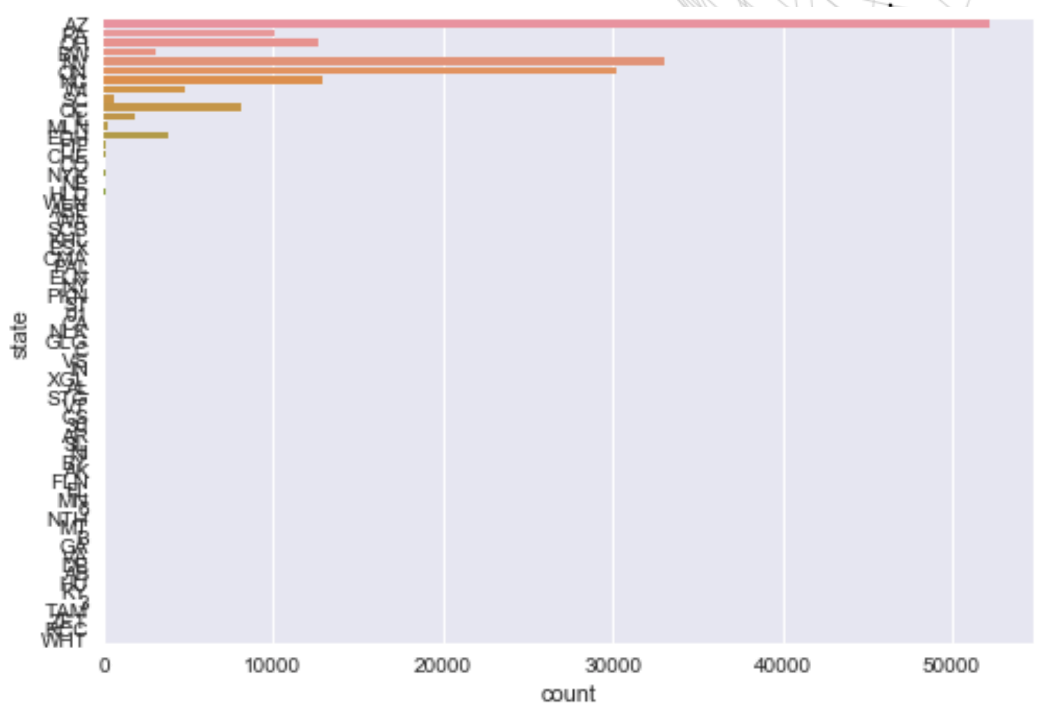
Number of reviews: 5.26 million

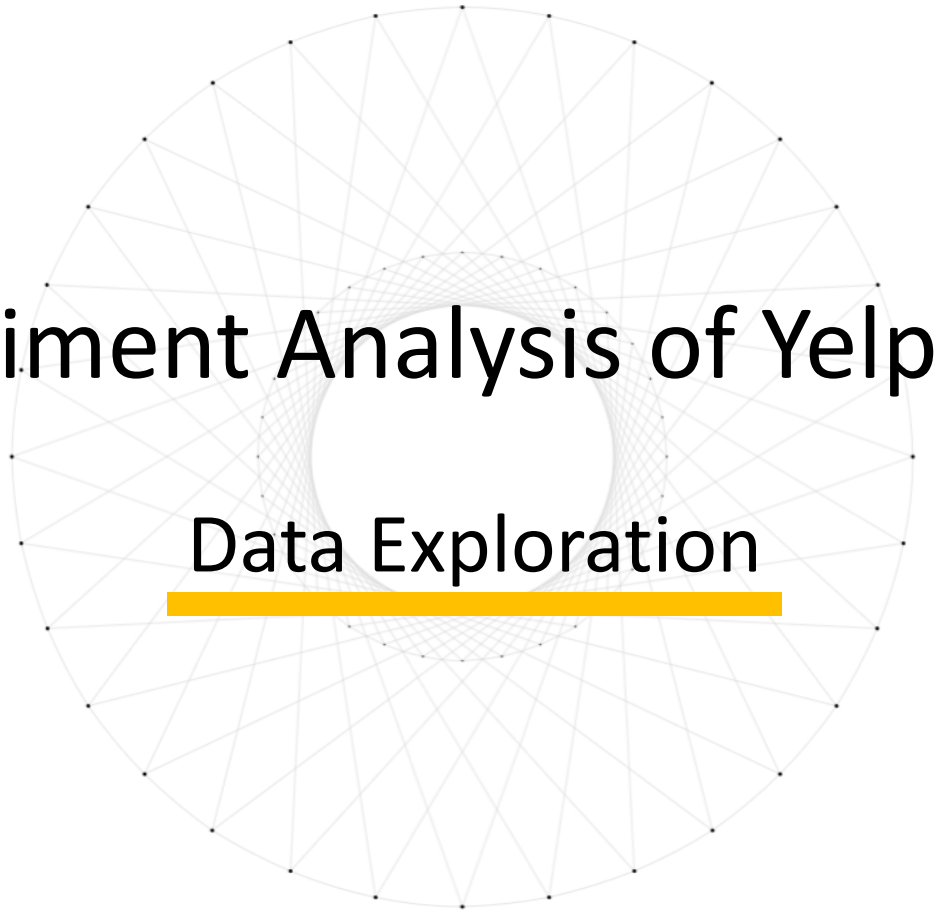
Sample subset

0.866 of all business are in states:
['OH', 'NC', 'ON', 'AZ', 'PA', 'NV']

Randomly select 1% of reviews in these states: 48360 reviews

Variables: text, stars



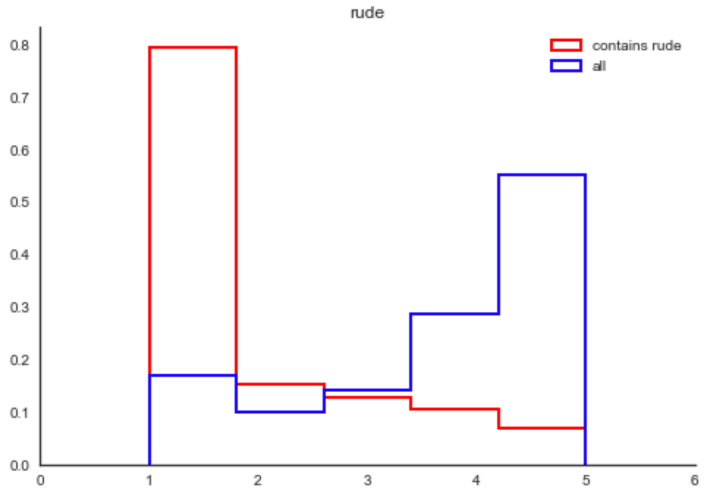


Sentiment Analysis of Yelp Review

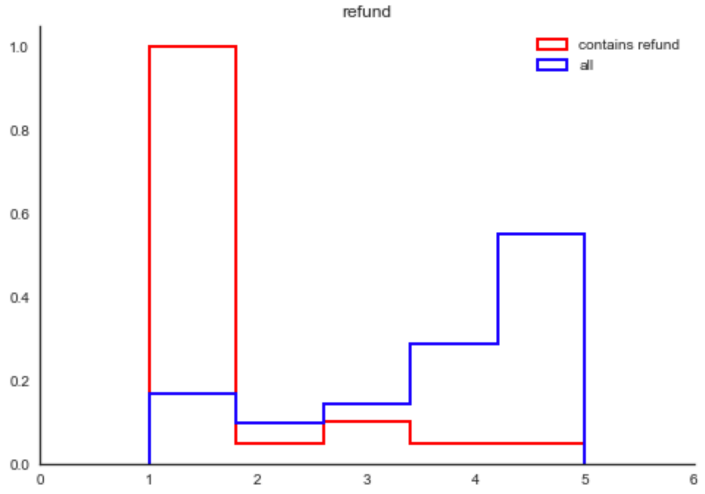
Data Exploration

Data Exploration

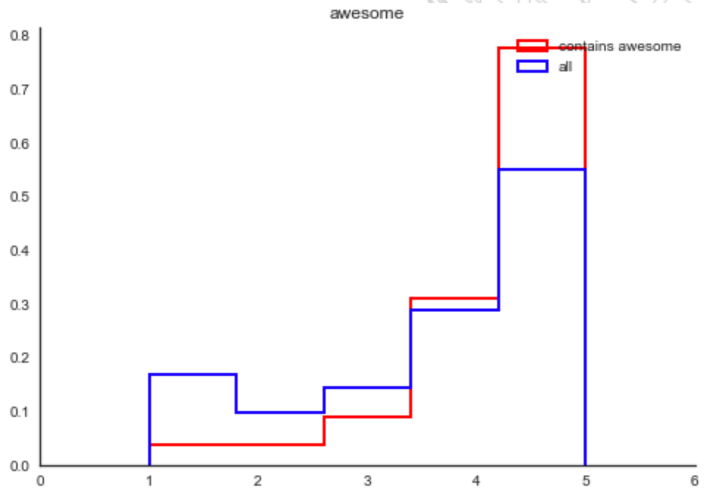
rude



refund



awesome



Goal

Classify a review as positive or negative

positive: stars 4 - 5

negative: stars 1 - 3

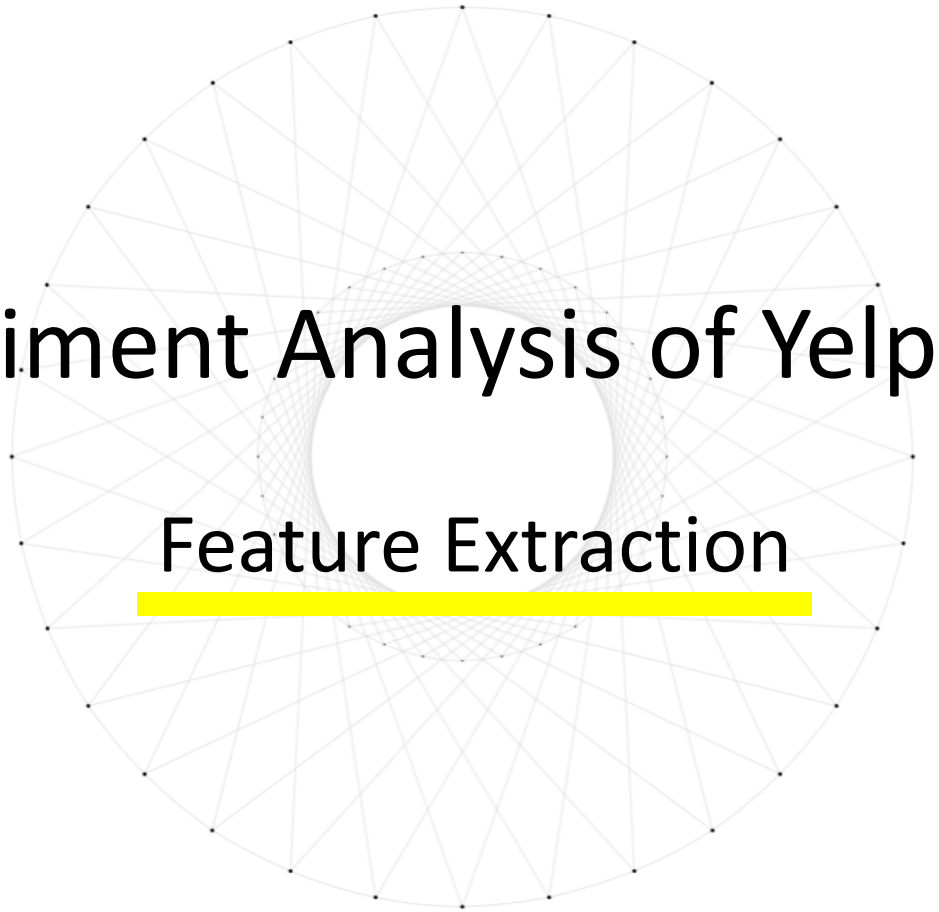
Excellent service and I love the chefs, very polite and take the time to talk with you. I've been at this location 3 times and have loved the food and it's been very good. The AYCE is great and they have an excellent happy hour for spirits and beer and sake.
This last visit my friend had some type of beef served on a small dish and I had a bite, but it tasted like rubber. Other than that great sushi and rolls!

Data Exploration

Words Clouds

positive

negative



Sentiment Analysis of Yelp Review

Feature Extraction

Data Preprocessing

- Change to lower case
- Remove punctuations
- Remove numbers
- Remove stop words
- Stem words

```
['classify', 'classifies', 'classifying', 'classified', 'classification']
['classifi', 'classifi', 'classifi', 'classifi', 'classif']
```

Bag of N-grams

- **Tokenize** strings and give an integer id for each possible token
- **Count** the occurrences of tokens in each document
- **Normalize** with diminishing importance tokens that occur in the majority of documents.

Choice of N:

e.g. “not good”

n = 1, “not” “good”

n = 2, “not good”

mixed 1&2, “not” “good” “not good”

Term Frequency

- Doc1: Theory is when you know everything but nothing works.
- Doc2: Practice is when everything works but no one knows why.
- Doc3: In our lab, theory and practice are combined, nothing works and no one knows why.

	theory	practice	everything	nothing	work	…
Doc1	1	0	1	1	1	…
Doc2	0	1	1	0	1	…
Doc3	1	1	0	1	1	…

TF–IDF

TF-IDF = $TF(t,d) * IDF(t)$

TF: Term Frequency

High TF means a term is common in a single document

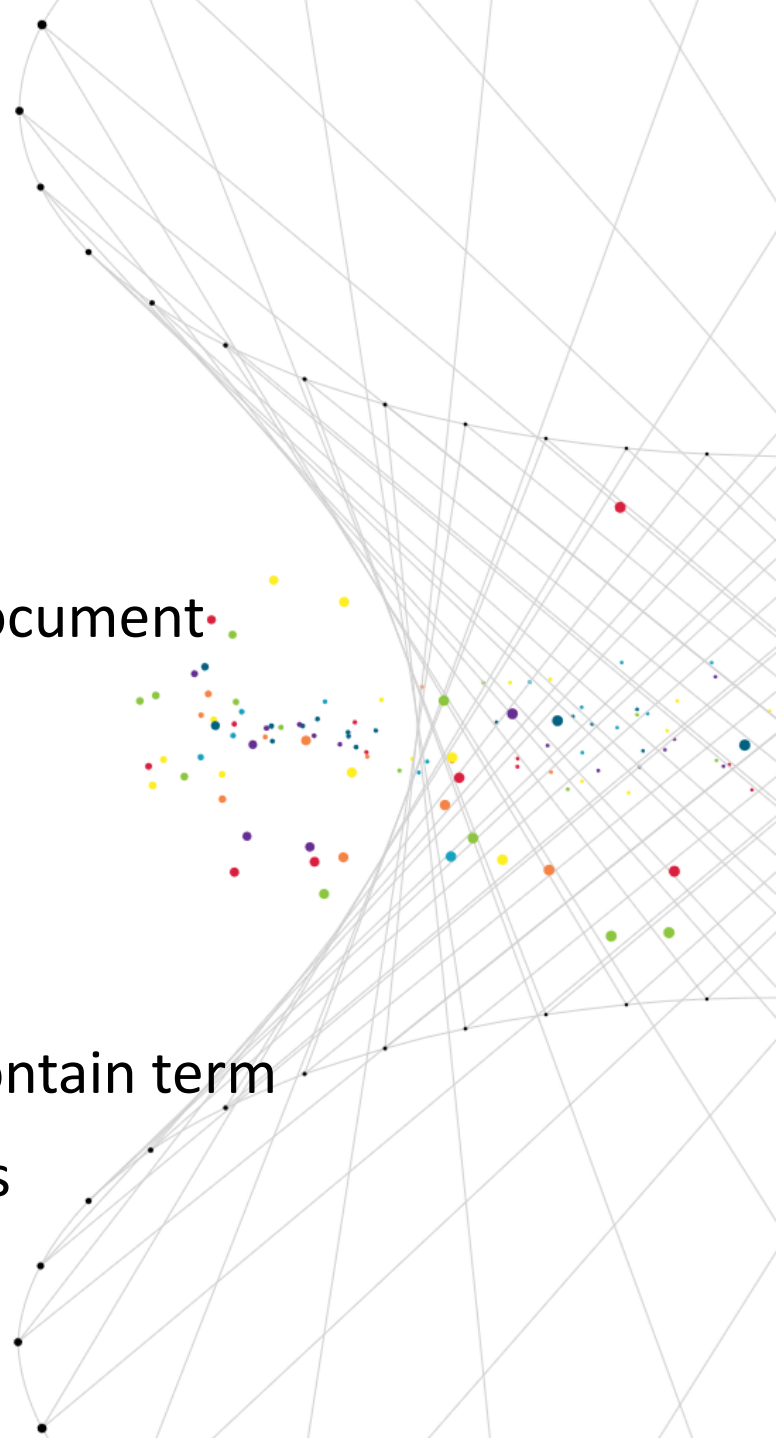
IDF: Inverse Document Frequency

$$IDF(t) = log\frac{1+n_d}{1+df(d,t)} + 1$$

n_d: total number of documents

$df(d,t)$: number of documents that contain term

High IDF means a term is rare in corpus



Sentiment Analysis of Yelp Review

Results

Results

Bag of N-grams vs TF - IDF

AUC	Logistic	Linear SVM	RF
Unigram	0.835	0.820	0.760
Bigram	0.798	0.773	0.704
Mix of Unigram & Bigram	0.859	0.842	0.747

AUC	Logistic	Linear SVM	RF
Unigram	0.864	0.848	0.729
Bigram	0.801	0.790	0.688
Mix of Unigram & Bigram	0.867	0.859	0.745

All vectorizers have max_df = .95 and min_df = 3.

Linear SVM is trained with stochastic gradient descent.

AUC above is the average AUC of test data under 10 different training (0.8)– test(0.2) splits.


Q & A
THANK YOU