



# Predicting Readmission in Patients with Diabetes Mellitus

Vivian Do | Bethany Wang  
University of San Diego  
MS Data Science

# Overview

**01**

**Problem  
Statement**

**02**

**Data  
Wrangling**

**03**

**Exploratory  
Data Analysis**

**04**

**Data  
Modeling**

**05**

**Results &  
Evaluations**

**06**

**Conclusion  
& Discussion**

# Problem Statement

**\$17 billion**

Spent annually by Medicare on hospitalizations that are **avoidable**

**Up to 3%**

of payments can be deducted if hospitals are not able to manage **excess readmissions**

**\$622 billion**

Total predicted **cost** of diabetes by 2030

**14.4 to 22.7%**

Readmission rate of diabetic patients, which is **higher** than the overall readmission rate for all inpatients

# Data Source



## **Health Facts Database**

Clinical patient records at 130 hospitals and integrated healthcare networks across the US



## **Extracted by Strack et al. (2013)**

In their research article "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records"



## **101,766 diabetic inpatient encounters & 50 features**

Additional criteria: Length of stay between 1-14 days, laboratory tests were performed, medications were administered



## **10 years of clinical patient records**

From 1999-2008



# Data Wrangling

# 01

# Data Wrangling

## Feature Deletion

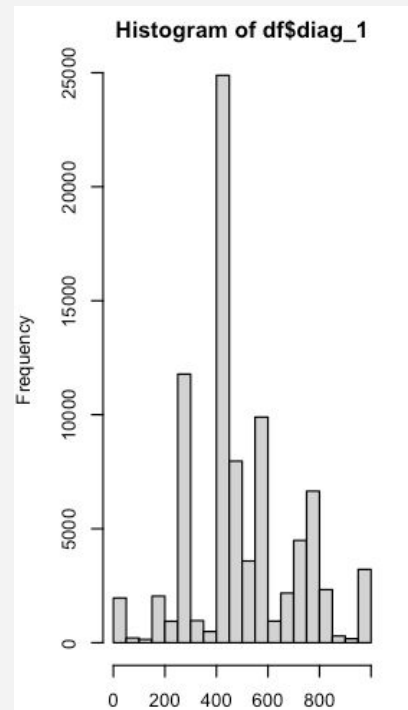
- Encounters associated with a discharge to hospice or death
- Degenerated attributes
- Removed features with a large percentage of null values
  - Weight 96.9%
  - Medical speciality of admitting physician 48.9%

## Feature Extraction

- Primary/Secondary diagnoses categorized into ICD9-CM chapters
- Reduced from 800 levels to 9

## Additional Pre-processing

- Removed outliers  $>/< 1.5IQR$



##	Circulatory	Diabetes	Digestive
##	27207	7185	7873
##	Genitourinary System Injury and Poisoning	Musculoskeletal	
##	4445	6017	4217
##	Neoplasms	Other	Respiratory
##	2762	18242	8618





02

Exploratory  
Data  
Analysis

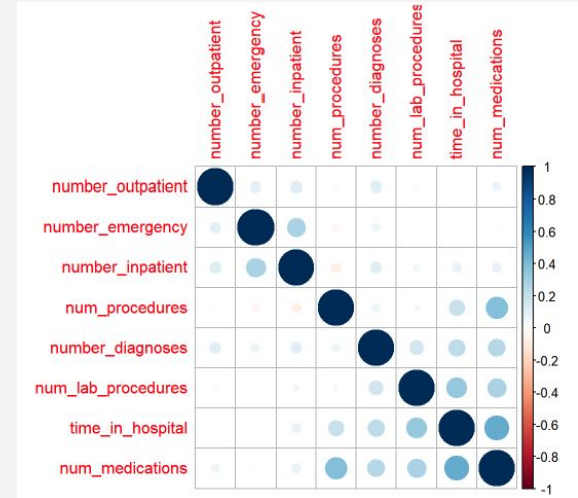
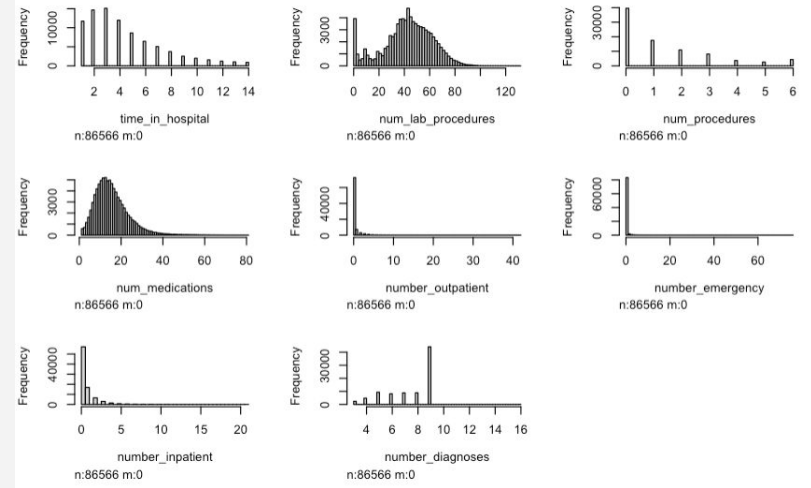
# Numerical Variables

## Distributions

- Number of medications administered and lab procedures performed approximately normal
- All other are heavily right skewed

## Correlations

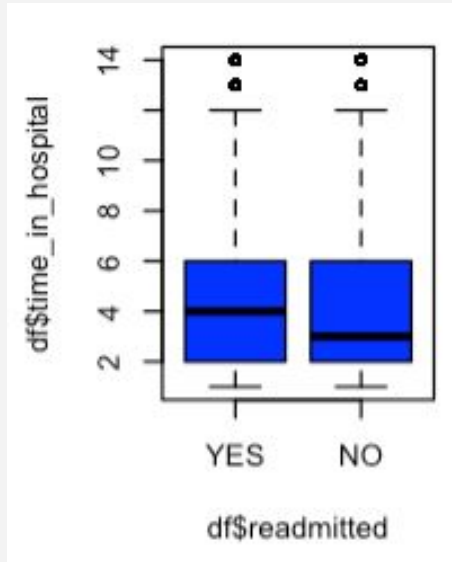
- Moderate correlations between time spent in hospital and number of procedures performed ( $r=.33$ )
- Moderate correlation between number of medications administered and lab procedures performed ( $r=.38$ )
- Moderate correlation between time spent in hospital and number of medications administered ( $r=.46$ )





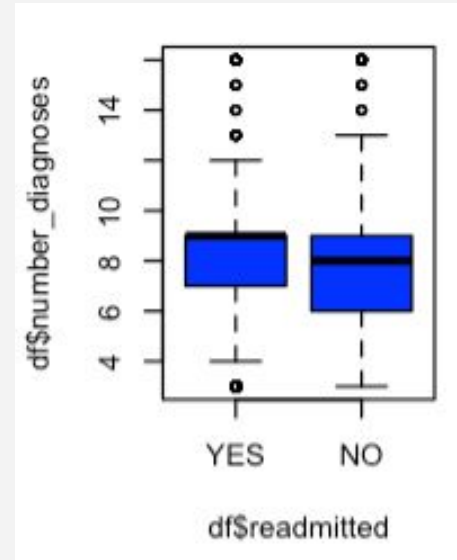
## Time in hospital vs. Readmitted

- Patients who were readmitted spent more time in the hospital, on average (4 days compared to <3).



## Number of diagnosis vs. Readmitted

- Patients who were readmitted have more diagnoses entered into their medical record on average (9 diagnoses compared to 8).



# Discharge, Admission, and other Patient Demographics

## Discharge Type

- Discharged to Medicare swing bed, outpatient services, and psychiatric unit most likely to be readmitted

## Admission Source and Type

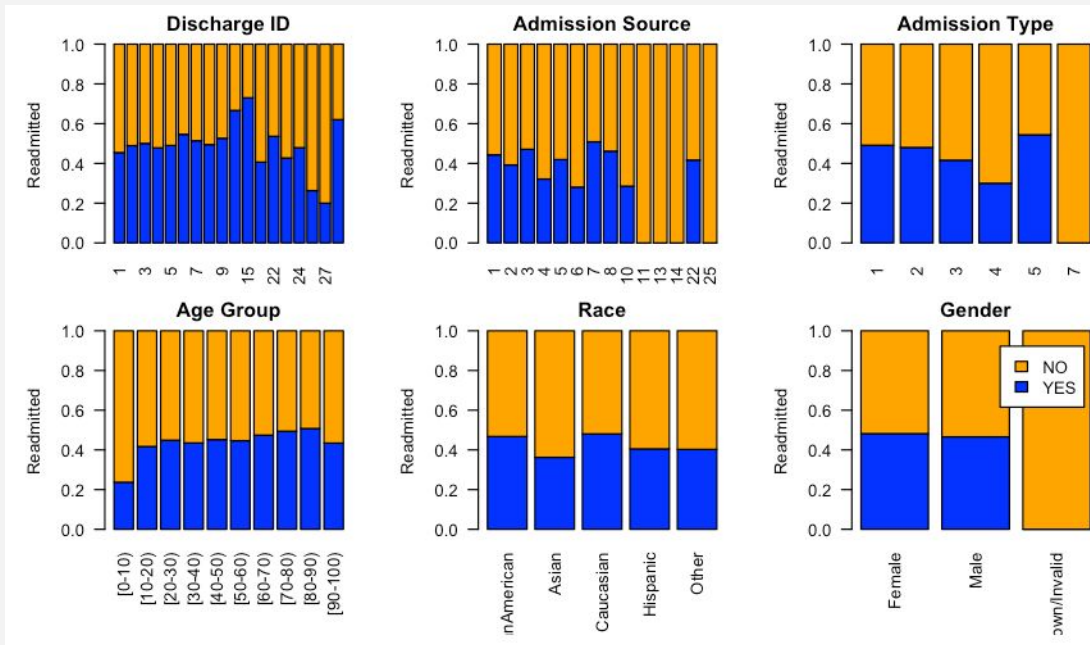
- Least likely to be readmitted if transferred from ambulatory surgery center or admitted through trauma center

## Age

- Increased risk of readmission with age

## Race

- Increased risk of readmission if African-American or Caucasian



# Primary Diagnosis

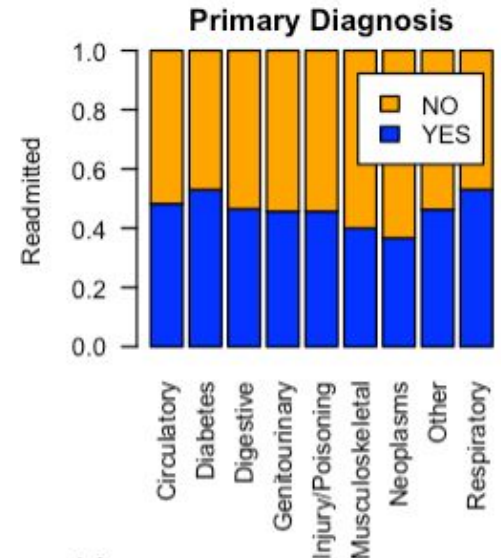
## Most common

- Circulatory (31%)
- Respiratory (10%)
- Digestive (9%)
- Diabetes (8%)
- Suggests that most patients are hospitalized for other reasons besides diabetes

## Readmission Risk

- Patients with a primary diagnosis of diabetes or respiratory diseases were more likely to be readmitted

#	Circulatory	Diabetes	Digestive
#	27207	7185	7873
#	Genitourinary System Injury and Poisoning	Musculoskeletal	
#	4445	6017	4217
#	Neoplasms	Other	Respiratory
#	2762	18242	8618



# Key Findings

## Increased Readmission Risks

Seen in patients being transferred to Medicare swing bed, outpatient services, and psychiatric unit



## Vulnerable Populations

Increased risk seen in elderly and patients who are African-American or Caucasian



## Time in Hospital

Patients who spend more time in the hospital and had more diagnoses entered into the system were more likely to be readmitted



## Comorbidities

of diabetes mellitus include circulatory, respiratory, and digestive diseases. Patients with a primary diagnosis of diabetes or respiratory diseases were more likely to be readmitted.





# Data Preparation

# 03



# Data Preparation

## Train/Test Split

- Split with stratification based on proportion of 'readmitted'
- Training set: 70% (59529)
- Test set: 30% (25512)

## Preprocessing

- Centering and scaling
- Conversion of categoricals to  $n-1$  dummy variables

## Final Dataset

- 18 predictors
- Binary target variable 'readmitted'
  - YES = patient was readmitted at any time
  - NO = patient was not readmitted



04

Data  
Modeling

# Model Selection

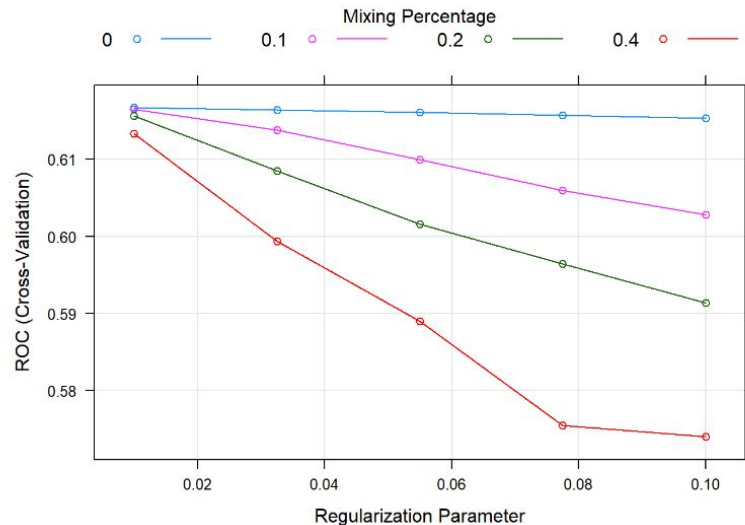
- Suitable for solving binary classification problems
- Capable of computing large dimensional data
- Works with both the numerical and categorical features.
- Linear classification models:
  - logistic regression, penalized logistic regression, and nearest shrunken centroids model.
- Non-linear models:
  - bagged tree, gradient boosted tree, random forest tree, and K-nearest neighbor models.

# 4.1 Logistic Regression

- Linear machine learning algorithm used for binary classification.
- Requires that each observation independent of the other
- Assumes a linear relationship between the independent variables and the log-odds of the dependent variable.
- Predicts the probability of an instance belonging to a class.
- Accuracy: 0.59
- Sensitivity of 0.52.

## 4.2 Penalized Logistic Regression

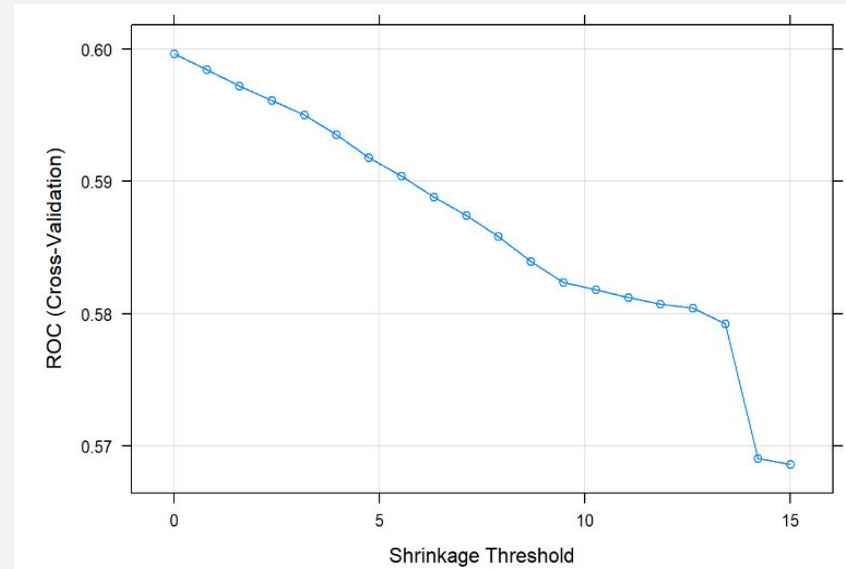
- Applies regularization to the basic logistic regression.
- glmnet adopts ridge and lasso penalties
- Parameter alpha controls the proportion between pure lasso ( $\alpha = 1$ ) and a pure ridge penalty ( $\alpha = 0$ ).
- Parameter lambda controls the amount of penalization.
- Tuned:
  - 4 alpha values (0, 0.1, 0.2, 0.4)
  - 5 lambda values between 0.01 and 0.1.
  - Optimal:  $\alpha = 0$  and  $\lambda = 0.01$
- Accuracy: 0.59
- Sensitivity: 0.52





## 4.3 Nearest Shrunk Centroids Model

- Assumes the centroids in the feature space are different for each target label.
- Summarizes a set of centroids for each class.
- Use the distance between a given data instance and each centroid to find the closest centroid for classifying the query.
- Tuning hyperparameter. shrinkage
- Tuned 20 shrinkage between 0 and 15.
- The optimal model: shrinkage threshold is 0.
- Accuracy: 0.57
- Sensitivity of 0.50.

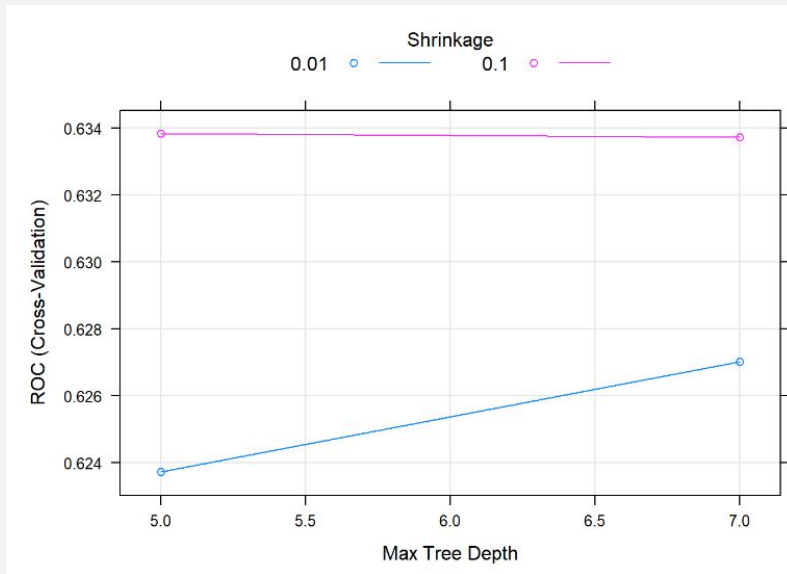


## 4.4 Bagged Trees

- An ensemble machine learning method to strengthen the single decision tree model.
- Weak learners learn from each other independently in parallel
- Are combined to produce a powerful tree model.
- Bagging decreases variance, and solves over-fitting issues
- With a number of bags parameter 30
- Accuracy: 0.57
- Sensitivity: 0.54.

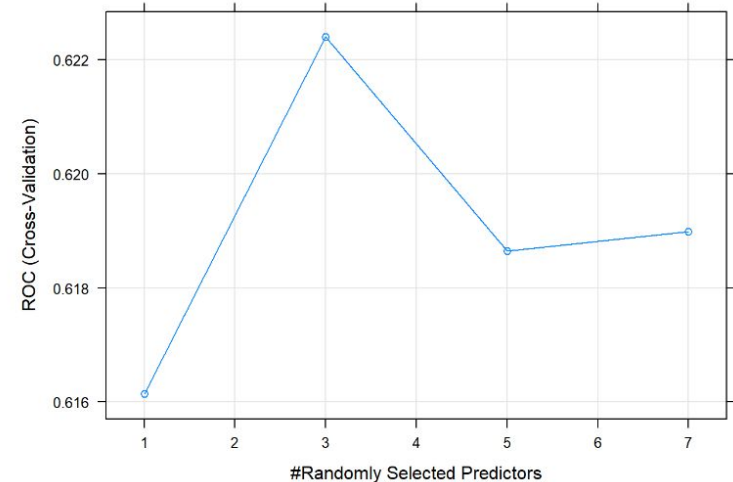
## 4.5 Boosted Trees

- Ensemble method
- Starts with a base/weak learner tree
- Constructs multiple tree models in sequence.
- Each tree corrects the previous one's errors.
- The final tree model works as a strong learner that shows the weighted mean of all the tree models.
- Decreases the bias error.
- We created a tuning grid with the following perimeters:
  - interaction depth (5, 7),
  - number of trees (500),
  - and shrinkage (0.01, 0.1).
- The optimal model: shrinkage is 0.1 and depth is 5.
- Accuracy: 0.6, Sensitivity: 0.56



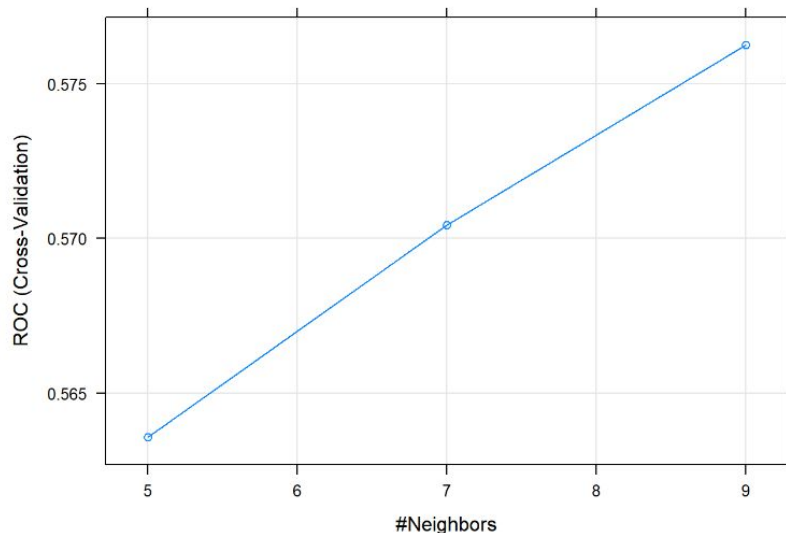
## 4.6 Random Forest

- Builds a series of decision trees with a random sample of the training dataset and combines them to decide the final classification.
- Tuning parameters: the number of trees and the number of randomly selected predictors.
- We tuned the model with 100 trees and the number of predictors of 1, 3, 5, and 7.
- The optimal model: the number of predictors 3.
- Accuracy: 0.59 and sensitivity: 0.53..




## 4.7 K Nearest Neighbors (KNN)

- A lazy learner algorithm
- Stores the available data without training them.
- It classifies a new data sample based on its similarity to its K nearest neighbors.
- Tuning parameter: K, the number of neighbors.
- Tuned K with values of 5, 7, and 9
- 9 is the optimal value for K.
- Accuracy: 0.56
- Sensitivity of 0.53.







# Results and Evaluations

# 05

## 5.1 Baseline Model

- Model's ability to predict the positive (readmitted-Yes) accurately is important.
- Use the all positive model as the baseline model.
- 48% of cases have readmitted label 'Yes'.
- The accuracy for this baseline model to predict the positive is 0.48.

## 5.2 Evaluation

Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp + fn}{tp + fp + tn + fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp + fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn + fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp + fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp + fn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2 * p * r}{p + r}$	This metric represents the harmonic mean between recall and precision values

## 5.3 Comparisons of Models on Cross-Validation Metrics

	Metric.Train	LR	GLMN	NSC	GBM	TRBAG	RF	KNN
1	Accuracy	0.585	0.585	0.57	0.596	0.572	0.586	0.555
2	Sensitivity	0.524	0.521	0.504	0.561	0.541	0.532	0.527
3	Specificity	0.64	0.642	0.629	0.628	0.601	0.635	0.581
4	Precision	0.57	0.57	0.553	0.578	0.552	0.57	0.534
5	Recall	0.524	0.521	0.504	0.561	0.541	0.532	0.527
6	F-Measure	0.546	0.544	0.527	0.569	0.546	0.55	0.53
7	ROC	0.617	0.617	0.6	0.627	0.603	0.622	0.576
8	AUC	0.503	0.51	0.502	0.502	0.499	0.503	0.495

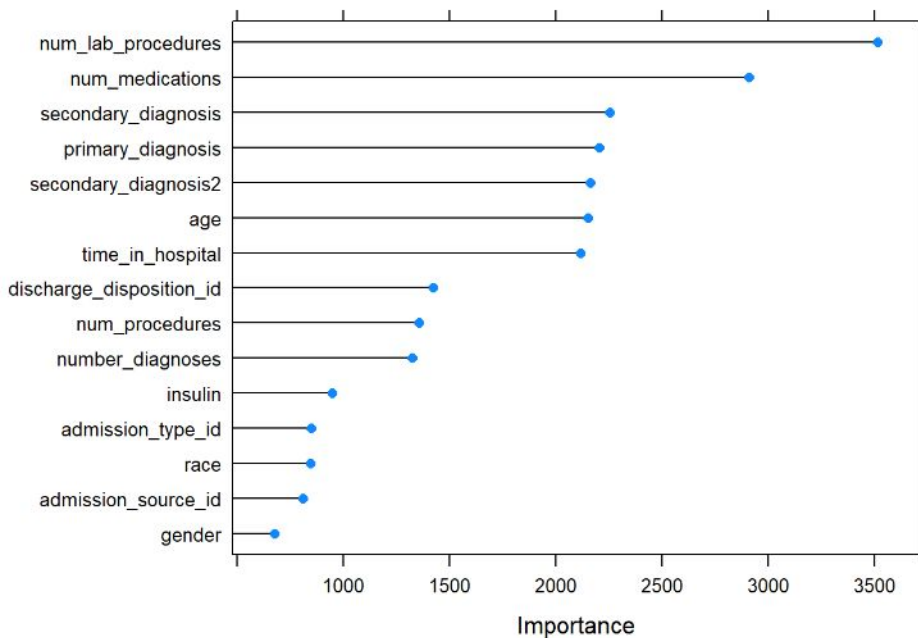
## 5.3 Comparisons of Models on Testing

	Metric.Test	LR	GLMN	NSC	GBM	TRBAG	RF	KNN
1	Accuracy	0.586	0.586	0.575	0.601	0.569	0.587	0.557
2	Sensitivity	0.524	0.520	0.507	0.565	0.535	0.503	0.528
3	Specificity	0.643	0.646	0.637	0.633	0.601	0.663	0.583
4	Precision	0.572	0.572	0.559	0.583	0.549	0.576	0.535
5	Recall	0.524	0.520	0.507	0.565	0.535	0.503	0.528
6	F-Measure	0.547	0.545	0.532	0.574	0.542	0.537	0.531



## 5.5 Feature Importance

- Checked the first 15 important features from logistic regression, random forest, and KNN model.
- Some common important features:
  - Number\_diagnoses
  - Num\_procedures
  - Num\_lab\_procedures
  - Num\_medication
  - time\_in\_hospital. etc.





Conclusions

06

## 6. Conclusion and Discussion

- Project recap
  - Help hospitals reduce readmissions of patients with diabetes
  - Gradient Boosted Tree is selected as the final model
  - It improves the baseline model's ability to predict the readmission yes by 10%
- Strength
  - Large amount of data
  - Compared multiple models
- Weakness
  - Extracted dataset-not original
  - Limited computing power
- Future work
  - Use original data
  - Run on cloud computing, AWS