ED Admission Case Prioritization using Classification Model

Taesun Yoo

- July 27, 2018 -

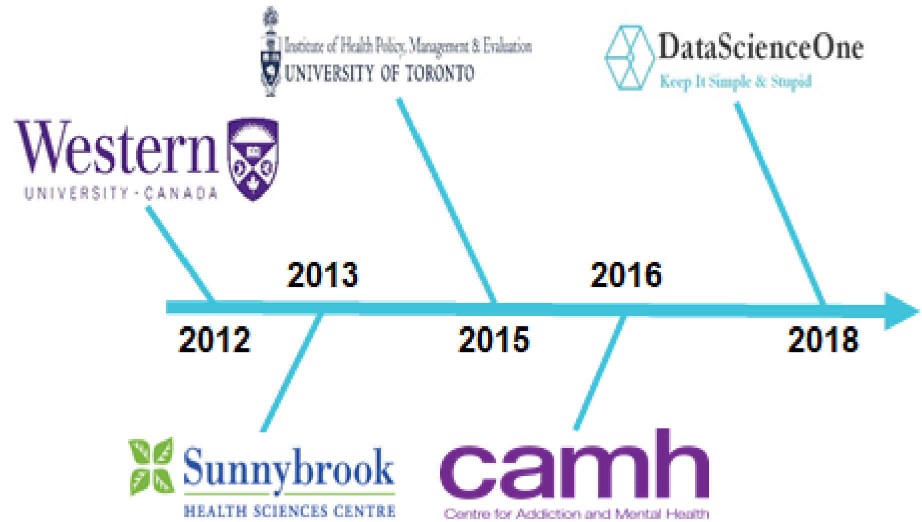
DataScienceOne
Keep It Simple & Stupid

About Myself: Taesun Yoo

- Former BI QA Analyst, ML Enthusiast
- Founder of DataScienceOne (Youtube Channel)
- Completed Master’s in Health Informatics

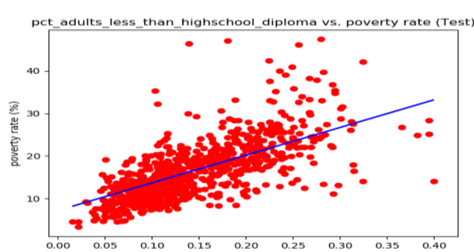
- Research experience: Sunnybrook
  - ❑ Medical Imaging (image processing)
  - ❑ Radiation Physics (cancer treatment)

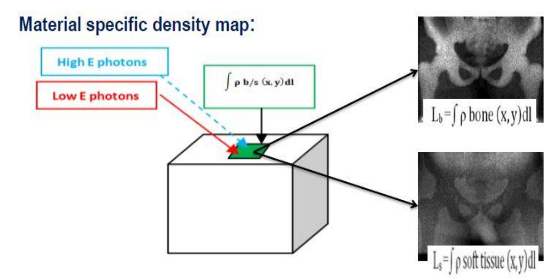
- Work experience: CAMH
  - ❑ Business Intelligence
  - ❑ QA data warehousing
  - ❑ Data visualization/reporting



Kicking some side machine learning projects …

MajorityVote Classifier		
	Predicted Class	
Actual Class	Stroke	Non-stroke
Stroke	41%	9%
Non-stroke	11%	39%





Agenda

1. Problem Overview
2. Data Wrangling: Cleaning and Transforms
3. Exploratory Data Analysis
4. Model Selections and Results
5. Future Work & Recommendations

Problem Overview

ED Admission: Background in Ontario

3 hours

ED wait time for initial assessment

31 hours

Total time spent in ED for admitted patients

618K visits per year

Average ED visits

$260 per visit (2008)

Average cost of ED visit

ED Admission: Problem Statement

Why should we care?

- ↑ incidence of adverse patient outcomes
- ↑ in ED wait time
- ↓ capacity to transfer patients (inpatient beds)

Stakeholders:

- Hospitals: unit managers, clinician groups
- Insurance companies: medical insurers
- Others: caregivers, health policy makers

Goal:

- Improve ED patient case prioritization by classification model(s)

Objective:

- Prioritizing urgent cases over non-urgent cases
- Facilitate management of ED patient flow (volume)

ED Admission: Dataset Overview

Dataset contains **30** input features for predicting an "**admission**" label:

- 16 categorical & 14 numerical features
- Patient demographics and diagnostic measures
- Sample size: 65,000 rows

Observations (rows)

Patient Demographics					Diagnostic Measures						Class Label
Key	Gender	Ethnicity	Avg_Income	Distance	GP_Visits	ED_Visits	....	Test_B	Test_F	Test_G	Admit
1821	Male	C	42247	168	1	0	....	7	N	SAT	1
2018	Female	C	42247	168	7	1	....	4	N	CIP	1
2176	Male	A	70000	200	1	0	....	7	N	LMA	1
2719	Male	C	65000	250	6	0	....	6	N	ACT	1
2734	Male	O	42247	168	1	0	....	5	N	LMA	1

Features (attributes)

Classes (label)

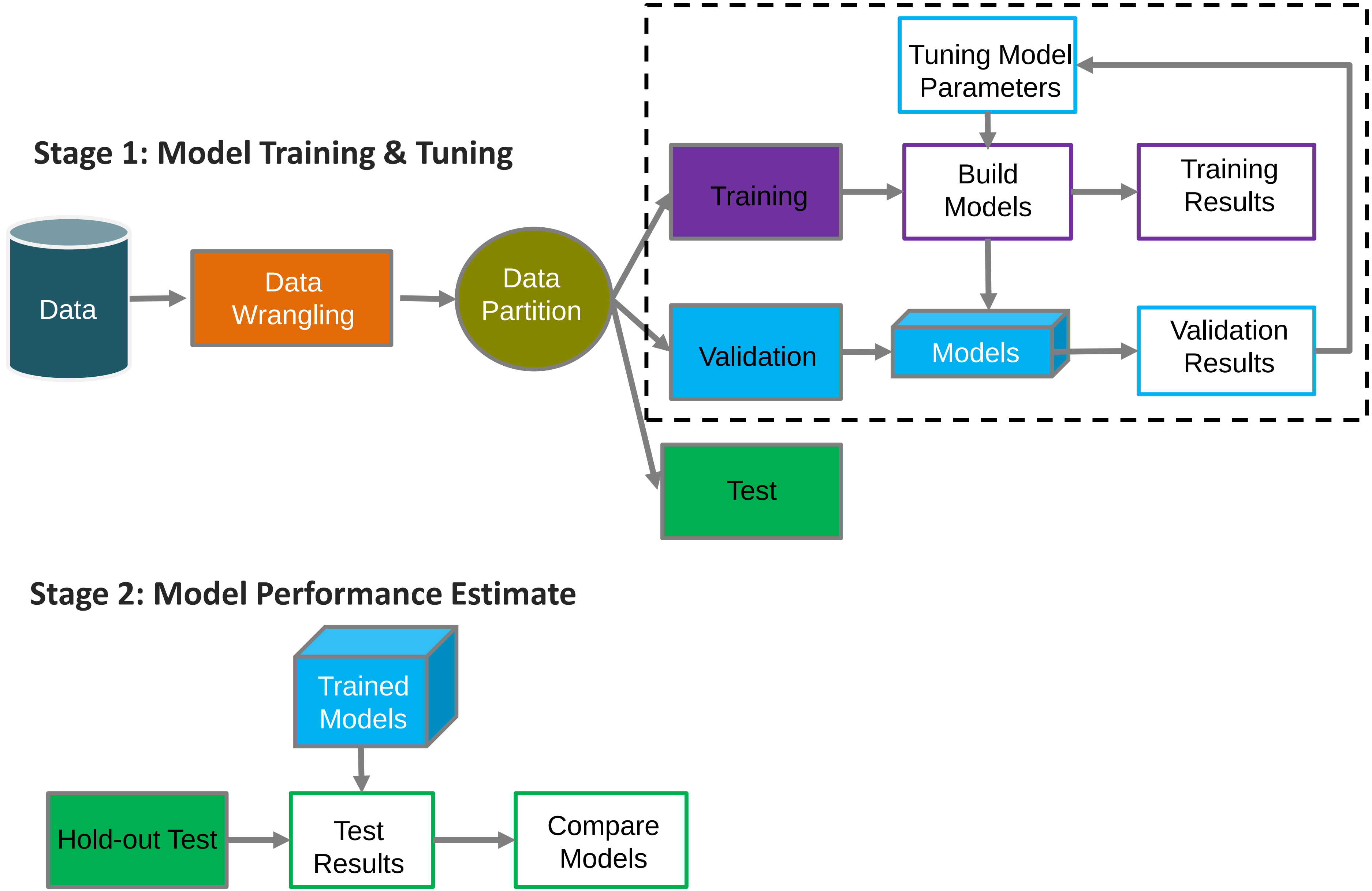
Challenges:

- Class imbalance (96% non-admitted vs. 4% ED admitted)
- Missing values
- Outliers/duplicates

Data Wrangling:
Cleaning & Transforms

DataScienceOne
Keep It Simple & Stupid

Model Workflows



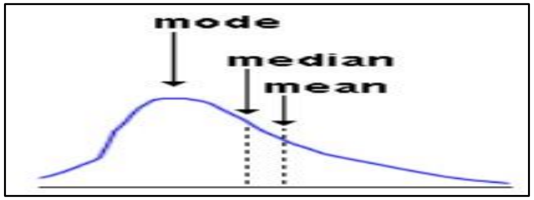
Data Wrangling

Raw Data Loading

↓

Missing Value Replacement

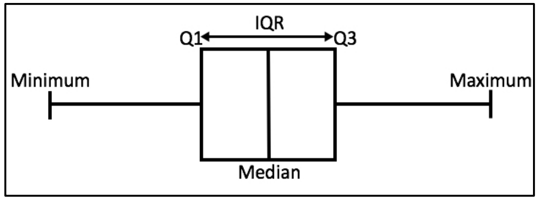
- Feature Drop
- Feature Imputation



↓

Handling Outliers

Interquartile Range:
LB = Q1 – 1.5*IQR
UB = Q3 + 1.5*IQR



↓

Resampling

- Down-sampling

Non-admitted (50%)	ED admitted (50%)

↓

Feature Transform

- Feature Encoding
- Feature Engineering
- Feature Scaling

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

DataScienceOne
Keep It Simple & Stupid

Missing Value Replacement

Method 1: Feature Removal

Feature Names	Missing Value Percent (%)
Nausea_Score	77.4
Referral_Diagnosis_2	72.6
Test_H	70.2
Referral_Diagnosis_1	51.1
Avg_Income	22.5
Ethnicity	20.9
Distance	20.8
GP_Code	17.1
Gender	4.24
Test_G	0.16
Test_A	0

Threshold: X > 50%

Method 2: Feature Imputation

Before: Imputation			
Gender	Ethnicity	Avg_Income	Distance
	C		168
Female		42247	
Male		70000	200
	C		250
Male	O	42247	

After: Imputation			
Gender	Ethnicity	Avg_Income	Distance
Male	C	51498	168
Female	C	42247	206
Male	C	70000	200
Male	C	51498	250
Male	O	42247	206

Handling Outliers

Methods to deal outliers:

- Inter-quartile range (IQR)
- Standard deviation
- Z-score
- Etc. (i.e., linear model)

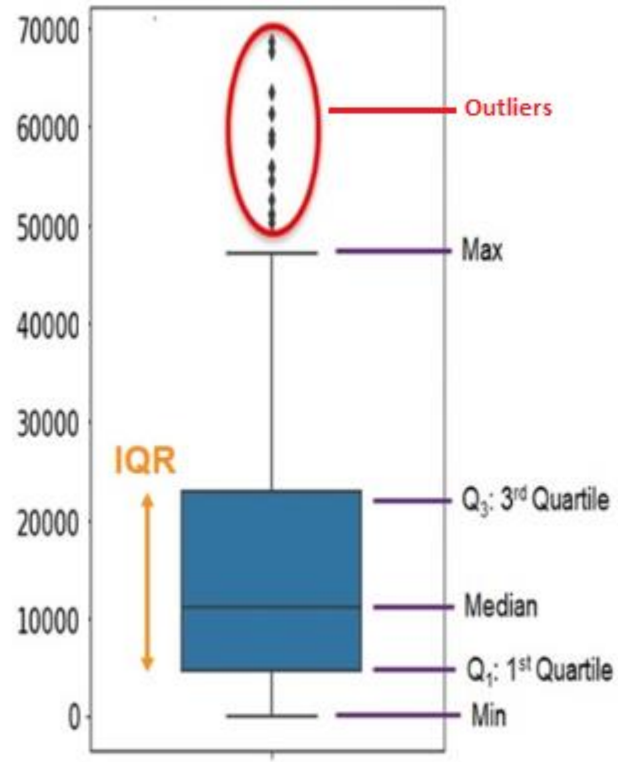
IQR is defined as follow:

- IQR = $Q_3 - Q_1$
- UB: $Q_1 - 1.5 * IQR$
- LB: $Q_3 + 1.5 * IQR$

Observations will be removed:

- Max. value > UB value
- Min. value < LB value

Box Plot: Avg. Income



Resampling: Class Imbalance

Resampling on majority class label:

- Resampled on non ED admitted cases
- Replacement through random selection
- Feasible to test different ratios:
  - [ED admit] vs. [Non ED-admit] = 4:6 ratio
  - [ED admit] vs. [Non ED-admit] = 3:7 ratio
  - [ED admit] vs. [Non ED-admit] = 2:8 ratio

Other resampling strategies:

- Upsampling (minority class label)
- SMOTE
- Etc.

Before Resampling: Original

Admit Label	Observations	Proportion (%)
Non-admit	58240	96.4%
ED admit	2144	3.6%

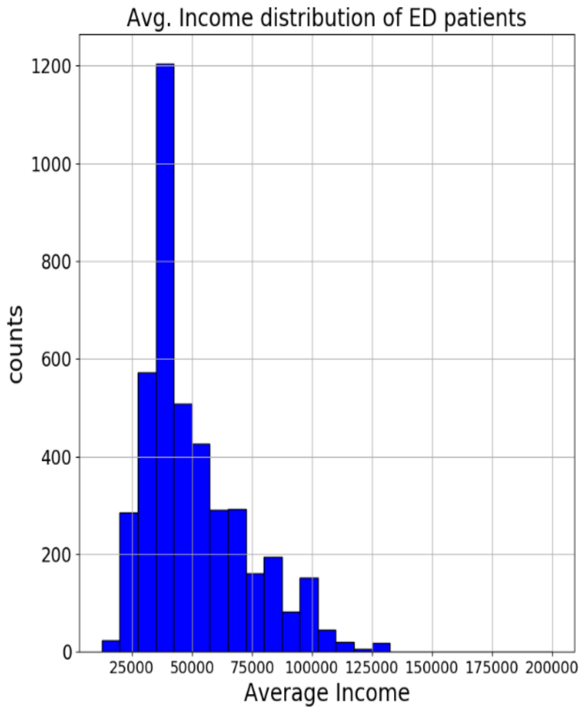
After Resampling: Downsampled

Admit Label	Observations	Proportion (%)
Non-admit	2144	50.0%
ED admit	2144	50.0%

Feature Transformation

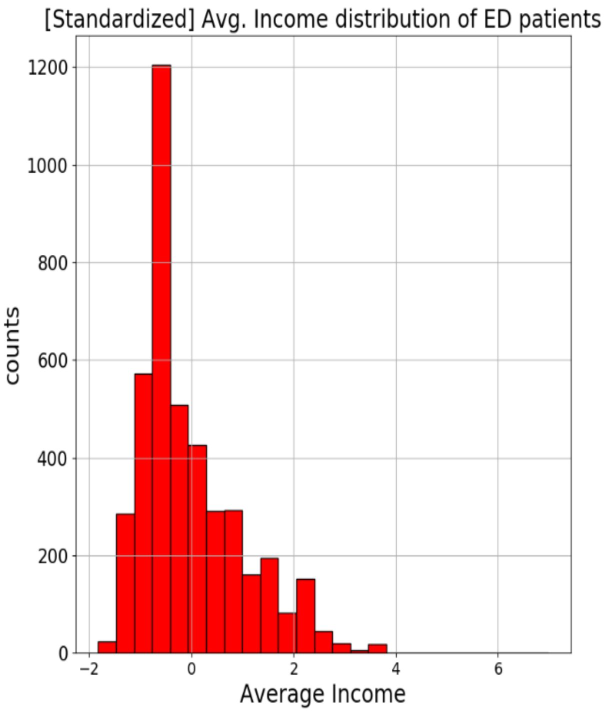
Feature Scaling: standardize range

Range: 0 ~ 150,000



f(x)

Range: -2 ~ 4



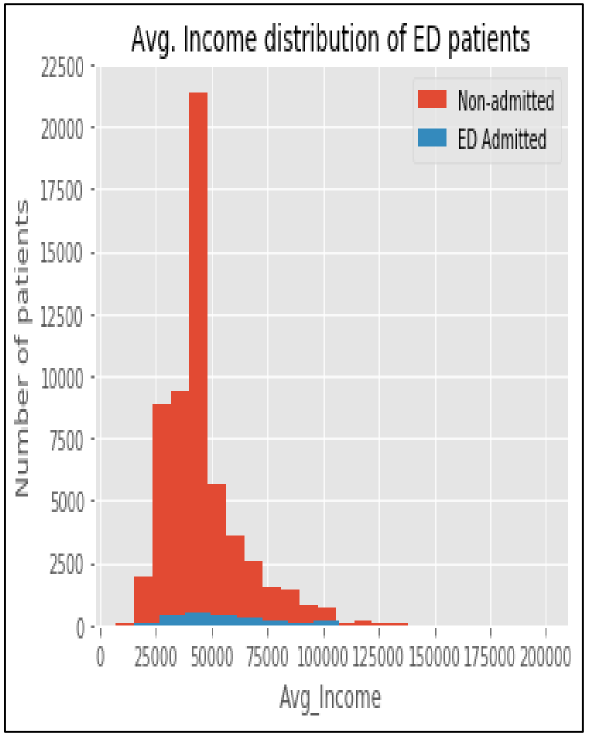
Feature Encoding: conversion into numerical format

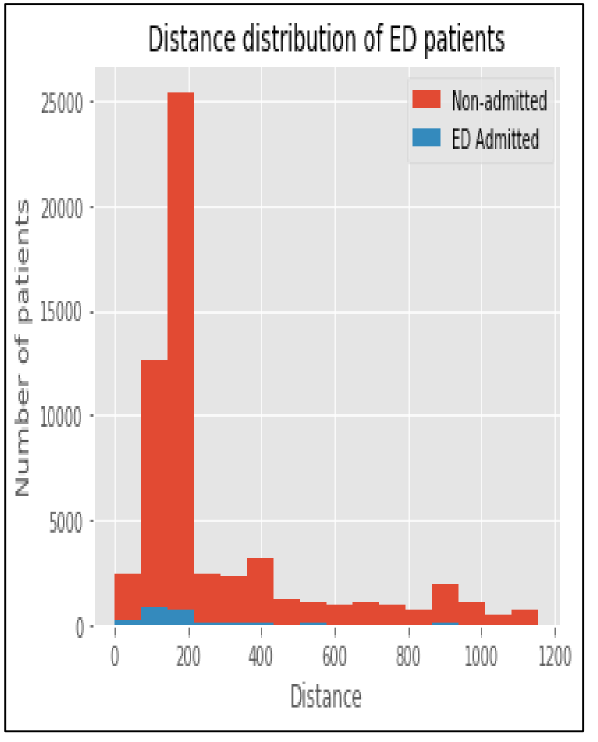
Ethnicity		Ethnicity_C	Ethnicity_A	Ethnicity_B	Ethnicity_N
C	→	1	0	0	0
A	→	0	1	0	0
B	→	0	0	1	0
N	→	0	0	0	1

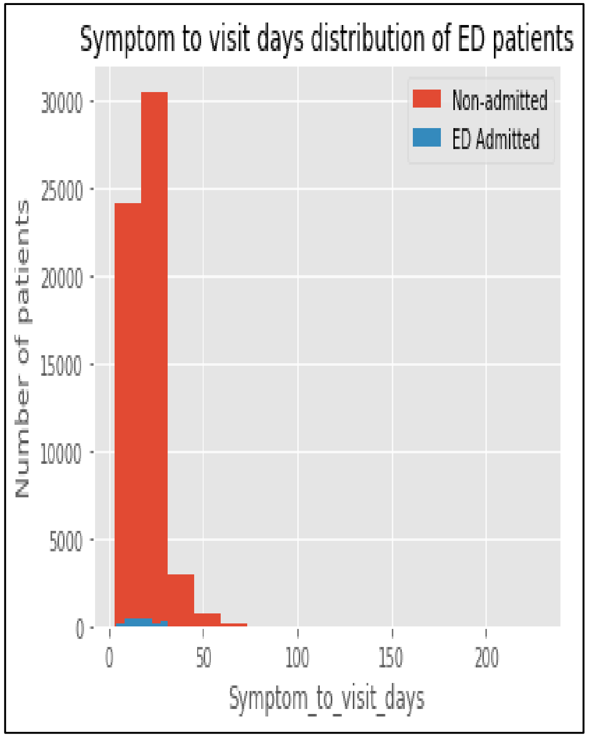
DataScienceOne
Keep It Simple & Stupid

Exploratory Data Analysis

ED Population: Distributions







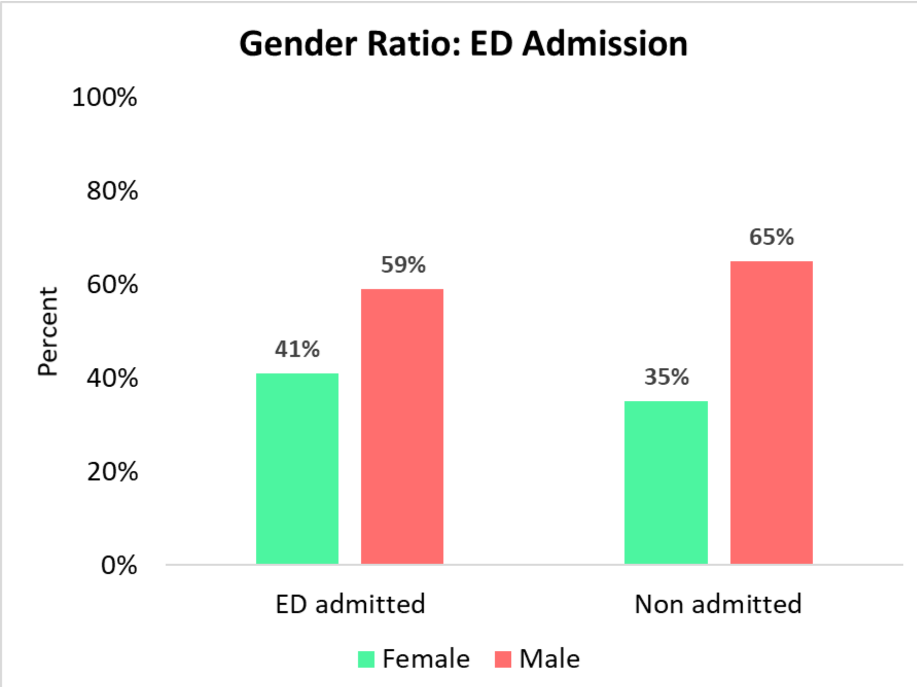
Avg. Income:
- majority from low to middle income class (skewed to right)

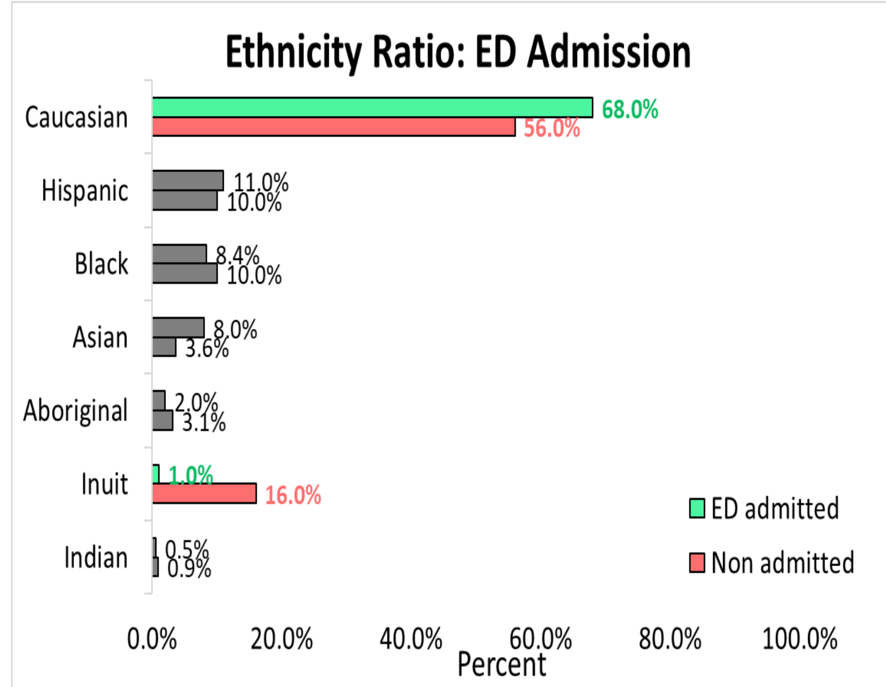
Distance:
- majority live nearby hospitals (100 to 200)

Symptom to visit days:
- relatively short symptom to visit days (skewed to right)

DataScienceOne
Keep It Simple & Stupid

ED Population: Patient Demographic





Gender:

- Higher % male >> female patients in both population

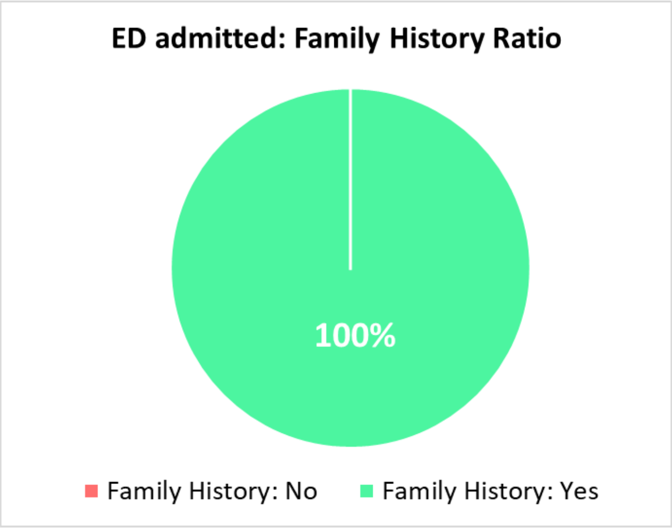
Ethnicity:

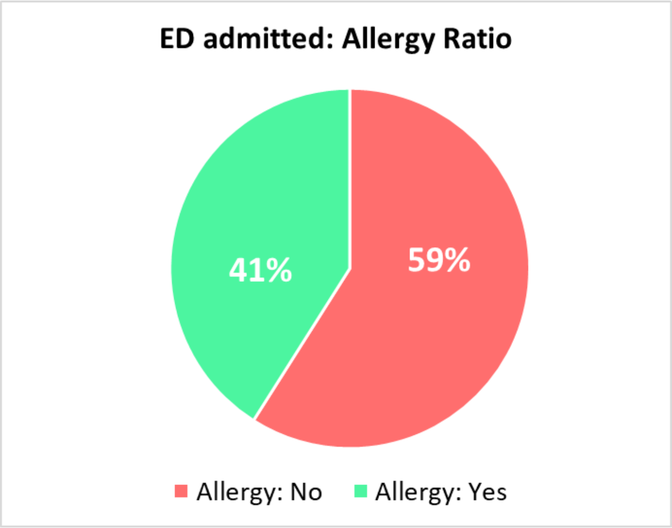
- Higher % of Caucasian in both populations
- Higher % of Inuit patients within non-admit population

DataScienceOne
Keep It Simple & Stupid

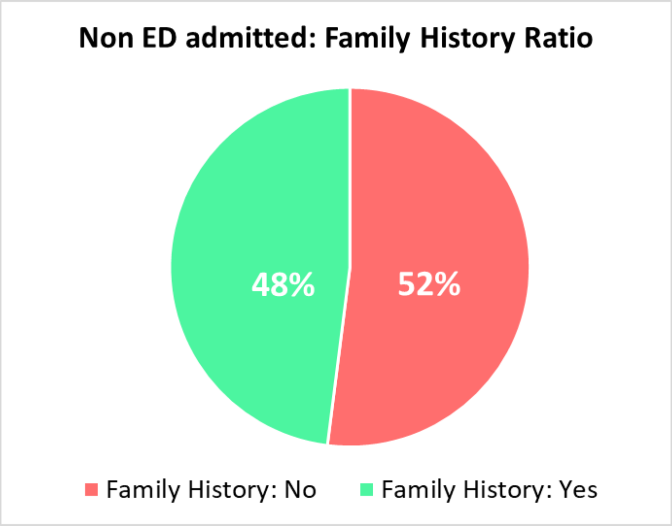
ED Population: Diagnostic Factor I

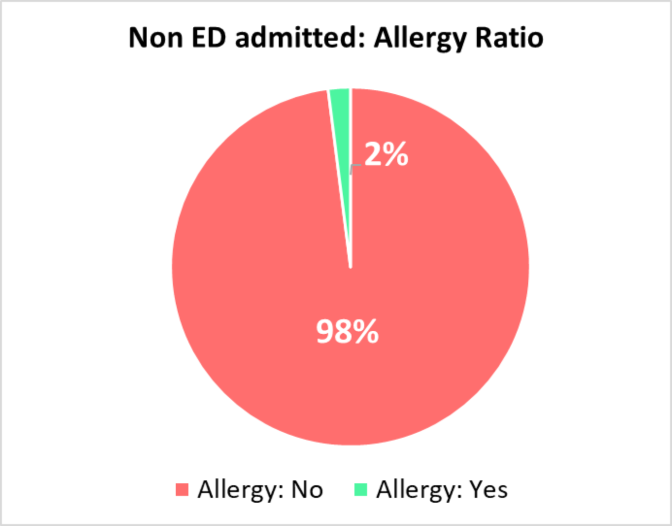
ED admitted





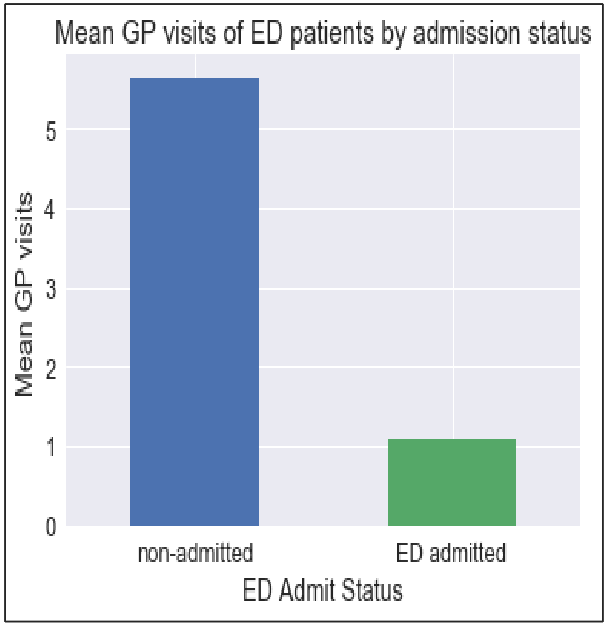
Non-admitted

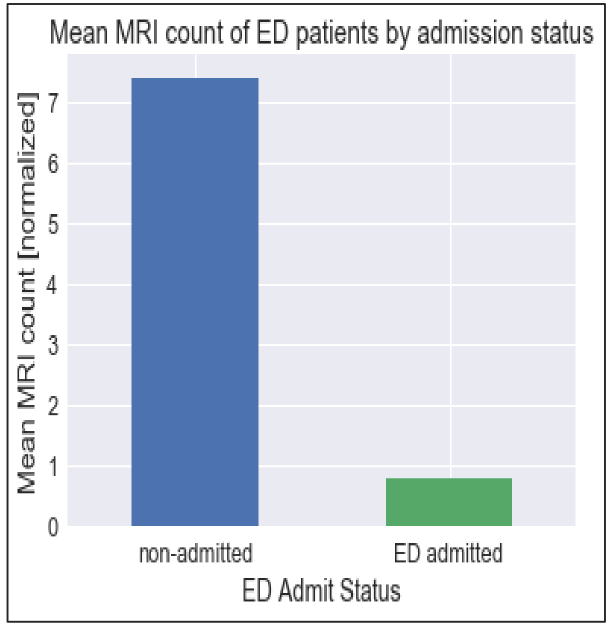


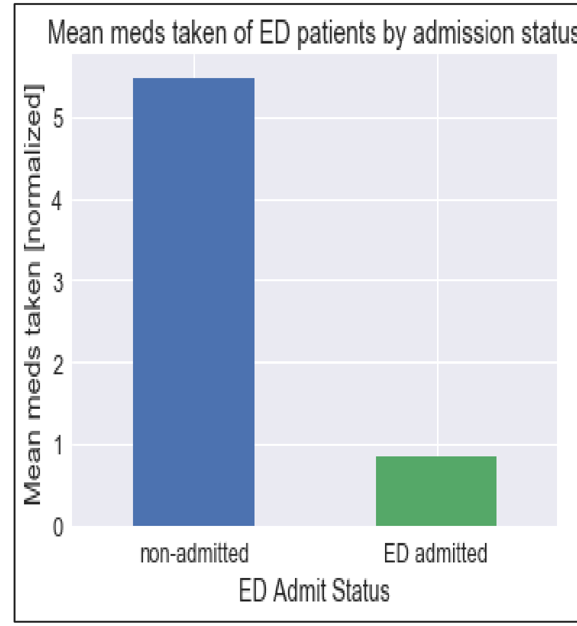


DataScienceOne
Keep It Simple & Stupid

ED Population: Diagnostic Factor II







GP Visits:
- Non-admitted patients had a mean of 5 GP visits

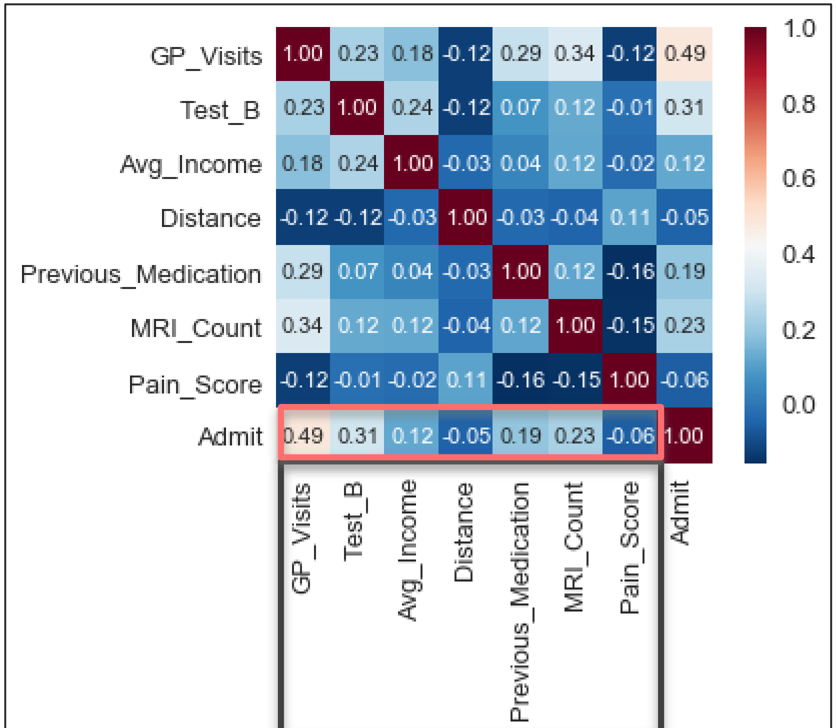
MRI Count:
- Non-admitted patients likely done MRI exam 7 times more

Previous Medication:
- Non-admitted patients likely taken medications 5 times more

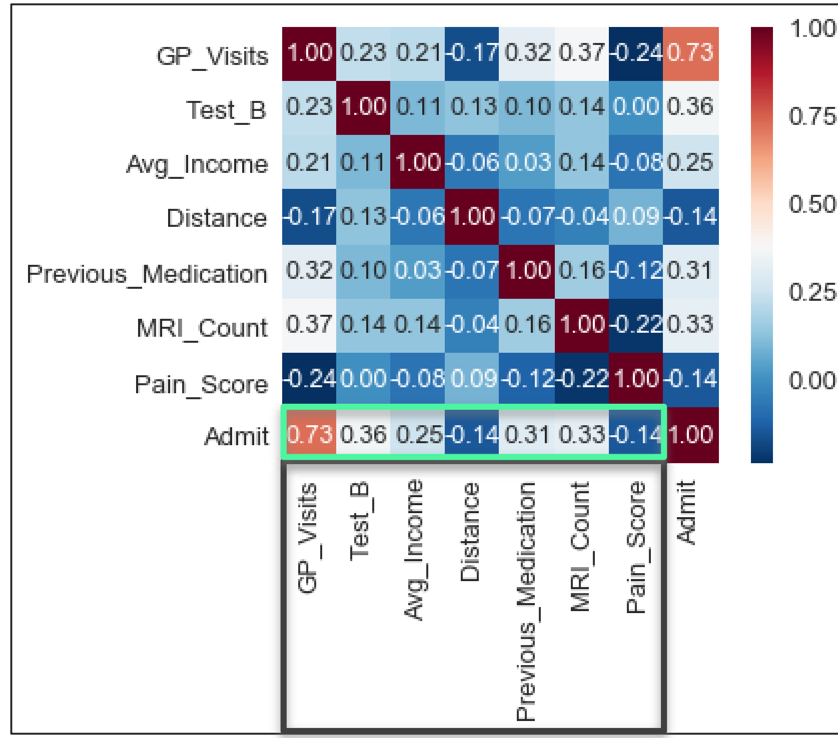
DataScienceOne
Keep It Simple & Stupid

Correlation Matrix

Before resampling



After resampling

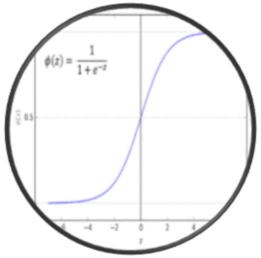


Highlights:

- ↑ in correlation among all numerical features
- Correlation of GP visits: ↑ from 0.49 to 0.73
- Order of increase in 'Pearson r': previous meds taken >> MRI count, etc.

Model Selection & Results

Model Selections

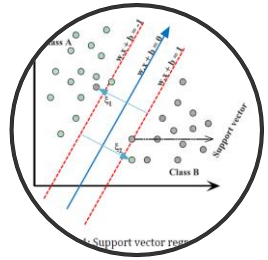


Logistic Regression

Sigmoid logit function:
log(p/(1-p))

Transforms:
Input values → estimated into prob. range (0, 1)

Works well on linearly separable classes.

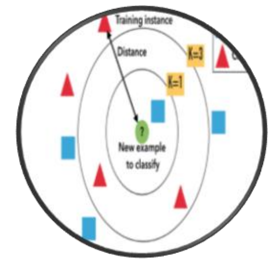


Support Vector Machine

Marginal classifier

Draw best decision boundary
- Compute max. margin between two hyper planes

Works well on linearly separable classes.

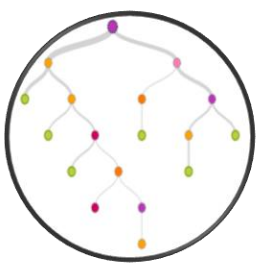


K-Nearest Neighbors

Choose k neighbors and count #

Computed by Euclidean distance

Assign new data point to category

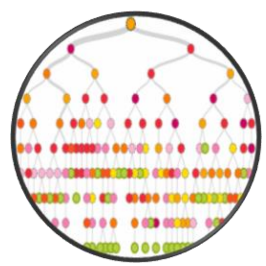


Decision Tree

Split data on features.

Repetitive splitting procedure.

Continue split until each node left with same class label.

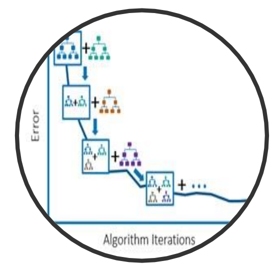


Random Forest

Ensemble learning.

Creates many decision trees.

Average performance of trees.



Gradient Boost

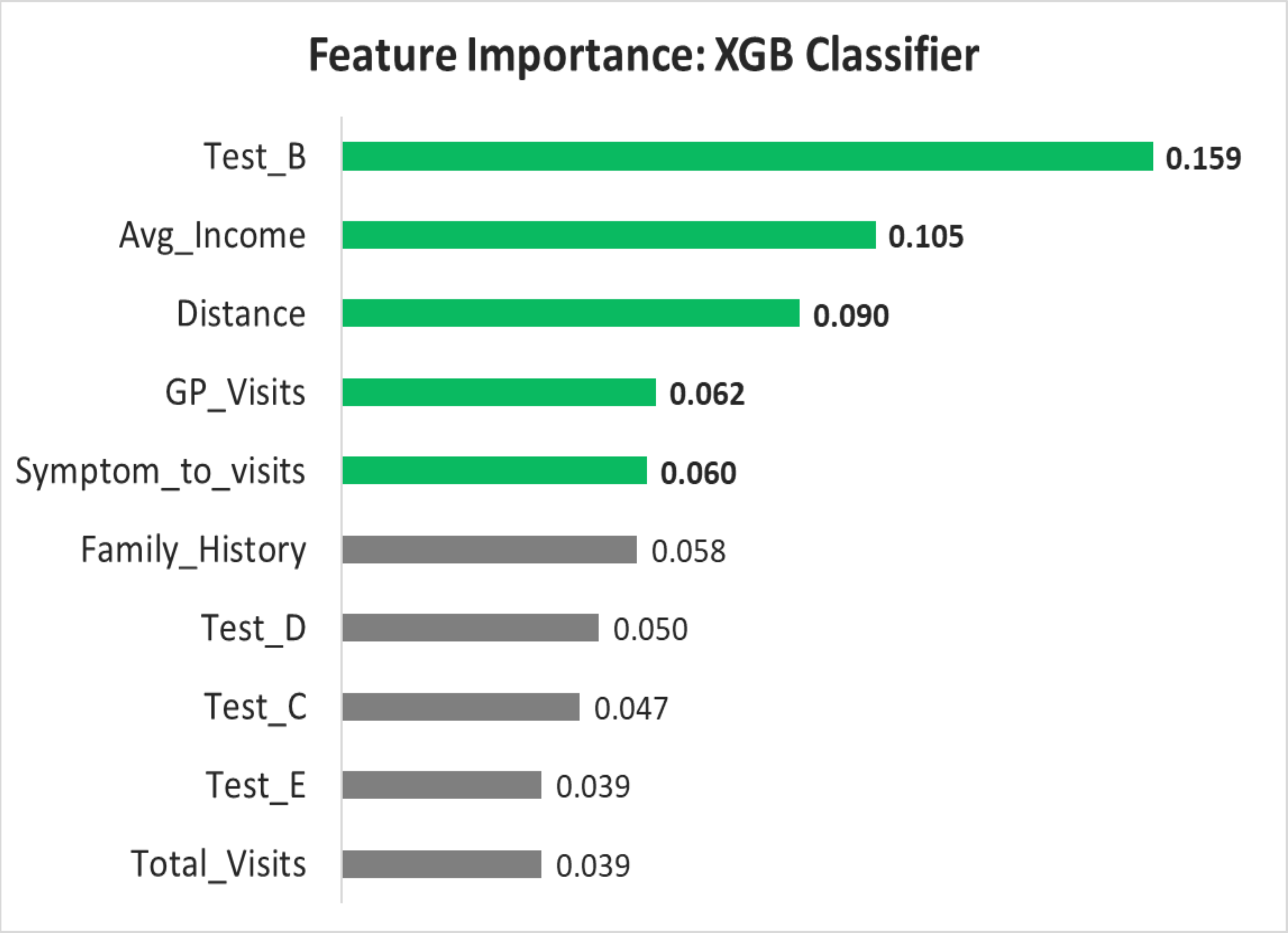
Sequential training.

Learn from residual errors.

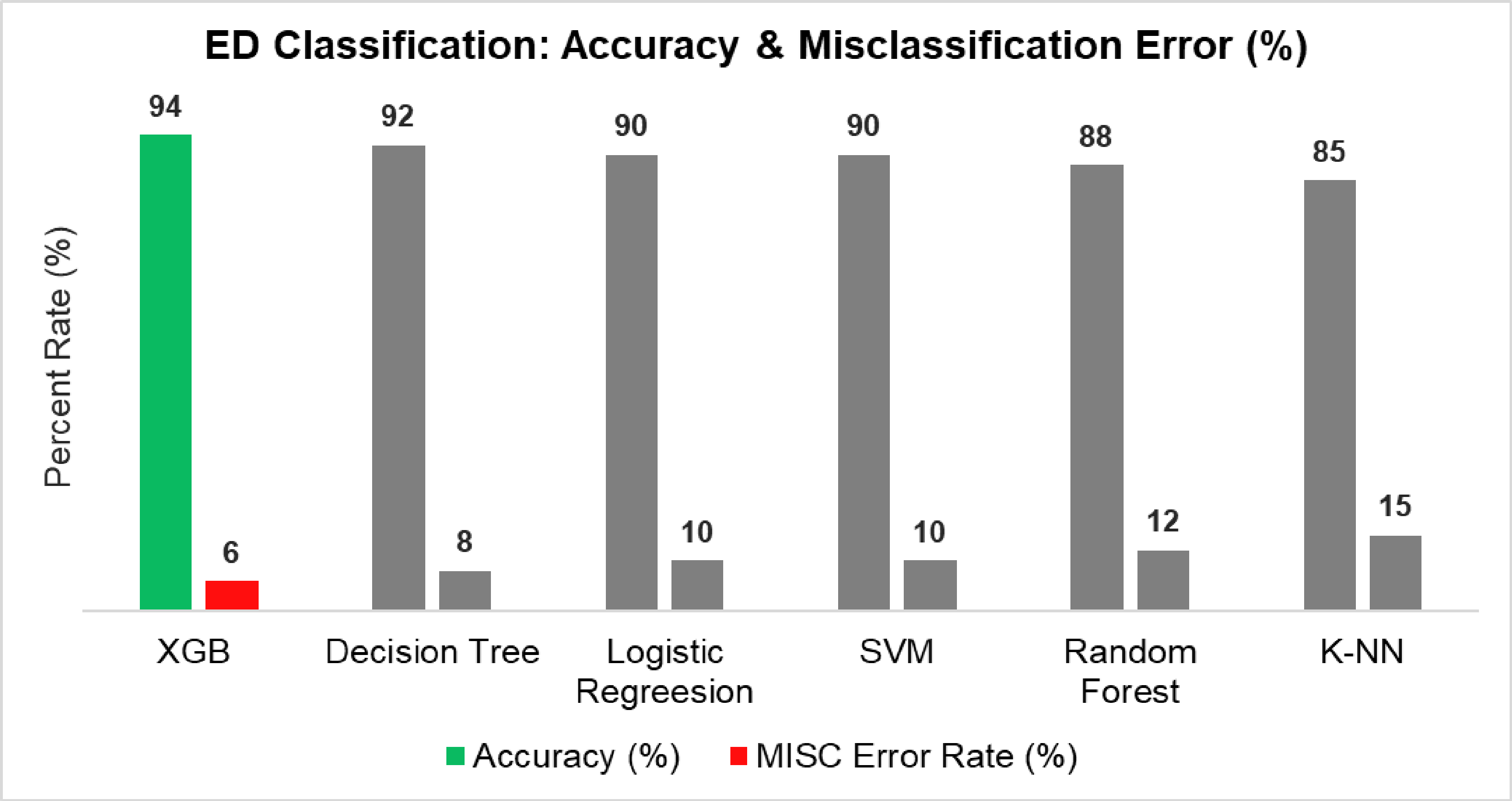
Step-wise forward

DataScienceOne
Keep It Simple & Stupid

Feature Selections



Model Comparison



In terms of accuracy & error rate:

- Best performing model was “XGB classifier!”

DataScienceOne
Keep It Simple & Stupid

ED Classification: Evaluation

ED Admission: XGB Classifier		
	Predicted Class	
Actual Class	ED admitted	Non-ED admitted
ED admitted	48%	2%
Non-ED admitted	4%	46%

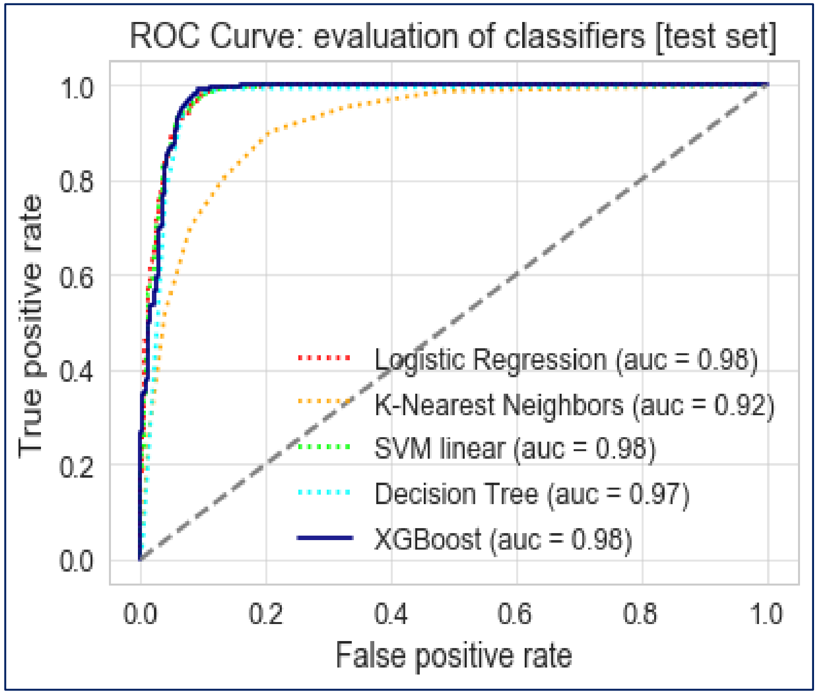
Model Interpretation:

- 94% of correct predictions
- 6% of mis-classification errors

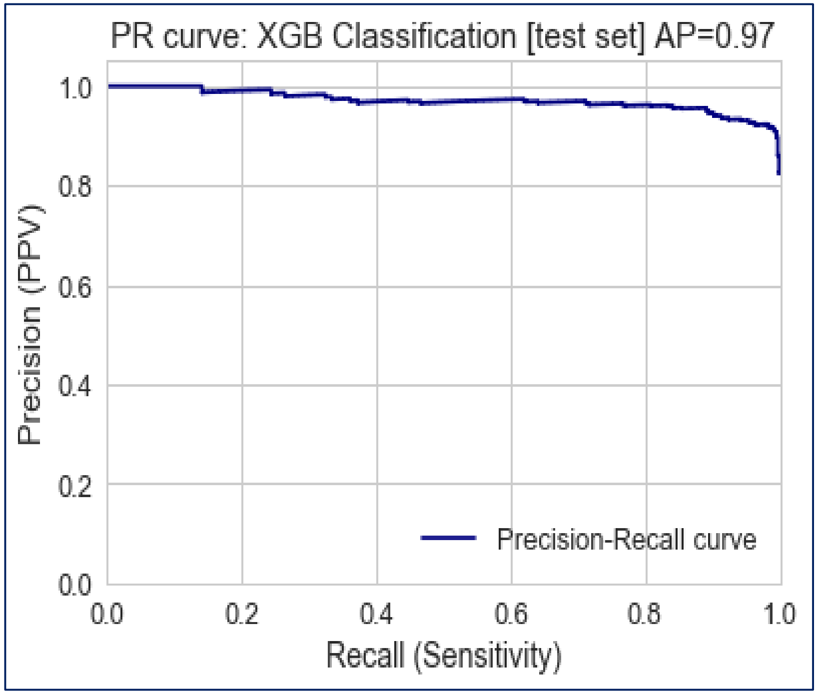
Balanced between model & human intervention:

- Type II error = 2% (*false negatives*)
- This can lead to adverse outcome (↑ *mortality rate*)

ED Classification: ROC vs Precision Curve



ROC curve



Precision-Recall curve

DataScienceOne
Keep It Simple & Stupid

ED Classification: Summary

Goal

- Improve ED patient case prioritization by classification model(s)

Results

- Model was able to predict whether or not a case required ED admission
- 94% of accurate predictions were made on a test set

Risks & Mitigation

Risks:
Model incorrectly identified with 2% of error as likely do not need admissions when they needed

Mitigation:
Review identified error cases with SMEs before decision making

Next Steps

- Conduct a pilot study with real data set
- Model improvement with tuning, sample size, different algorithms

DataScienceOne
Keep It Simple & Stupid

ED Classification: Investment Returns

$260 per single ED visit on average in Ontario hospitals (CIHI 2008)

618K ED visits across Ontario in FY16/17 (NACRS)

$161M annual spending on ED visits in Ontario (estimated)

1 in 5 ED visits in Canada can be treated at Doctor's Office (HQO 2017)

Estimated Annual R.O.I:

= 1/5 x 618,000 ED visits x $260/ED visit

= $32M SAVED!!

Future Work & Recommendations

Limitations & Future Work

Limitation:

- Simulation study (not real dataset!)
- Validity of study is questionable

Future Work:

- Conduct pilot study at institutional level (real patient data)

- Model improvement:
  - Stacking
  - Boosting/Bagging

- Resampling strategies:
  - SMOTE
  - Upsampling (i.e., minority class: ED admitted cases)

- GP visit counts stratified classifiers:
  - Low GP visits patients cohort (GP visits < 5)
  - High GP patients cohort (GP visits > 5)

Recommendations

Implement case prioritization guideline

- Non-admitted vs. ED admitted patients likely have:
  - 5 times higher GP visits
  - 7 times higher MRI exams done
  - 5 times higher medications taken

Conduct pilot study

Develop a P.O.C ML model and pipeline for model deployment

Data collection & integration

- Need for integration EMR, administrative data
- Inclusion of meaningful features like
  1. Triage category (severity of condition)
  2. Hours after admission in past
  3. Others

Thank You!

Questions?