UR MEDICINE | WILMOT CANCER INSTITUTE

URMC Ger Oncology

FINAL PRESENTATION

April 20, 2022

Team Members

Olivia Cen
Data Science, Business Analytics

Xinyu Cai
Data Science, Business Analytics

Yaxin Yang
Data Science, Molecular Genetic

Yilin Zhou
Data Science, Biology

AGENDA

I. Overview

II. Data Description & Visualization

III. Predictive model

IV. Results

V. Key Insights

VI. Next Steps

01 PROJECT VISION

Wilmot Cancer Institute aims to increase the effectiveness of chemotherapy in treating older persons with advanced cancer.

PROJECT GOALS

- Feature selection based on understanding and rigid thresholds
- Predictive models to assess the efficiency of medication features in chemotherapy results
- Refine Data Preprocessing Pipeline

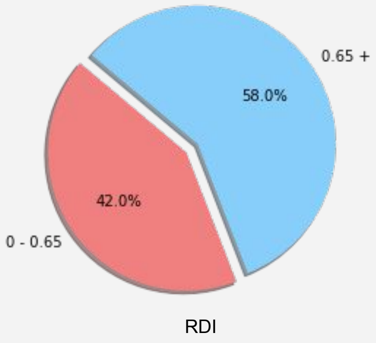
MILESTONES

	Task	Target Date	Actual Date	Status
1	Project Charter Draft	2/23	2/22	Completed
2	Merge Data and Visualization	2/28	2/26	Completed
3	Project Charter	3/1	2/24	Completed
4	Model 1	3/14	3/14	Completed
5	Midterm Presentation	3/20	3/20	Completed
6	Model 2	3/28	3/30	Completed
7	Model Tuning	4/4	4/6	Completed
8	Data Preprocessing Pipeline	4/11	4/11	Completed
9	Final Presentation	4/20	4/19	Completed
10	Final Report, Code, README	5/1		In Progress

02 DATA DESCRIPTION

- **Geriatric Assessment for Patients 70 years and older (GAP-70) Dataset (.csv)**
  - **718** observations, **145** features, **77** missing target variables
  - **Target Variable**
    - **RDI**: Relative dose intensity (RDI) is the ratio of **the delivered dose intensity** to the **standard dose intensity**, reflecting the implementation of the expected dose intensity.

$$\text{RDI} = \frac{\text{Sum of percentage of ideal dose given for each drug}}{\text{Interval between first and last dose}} \times \frac{\text{Expected interval}}{100 \times \text{no. of drugs}}$$



02 DATA DESCRIPTION

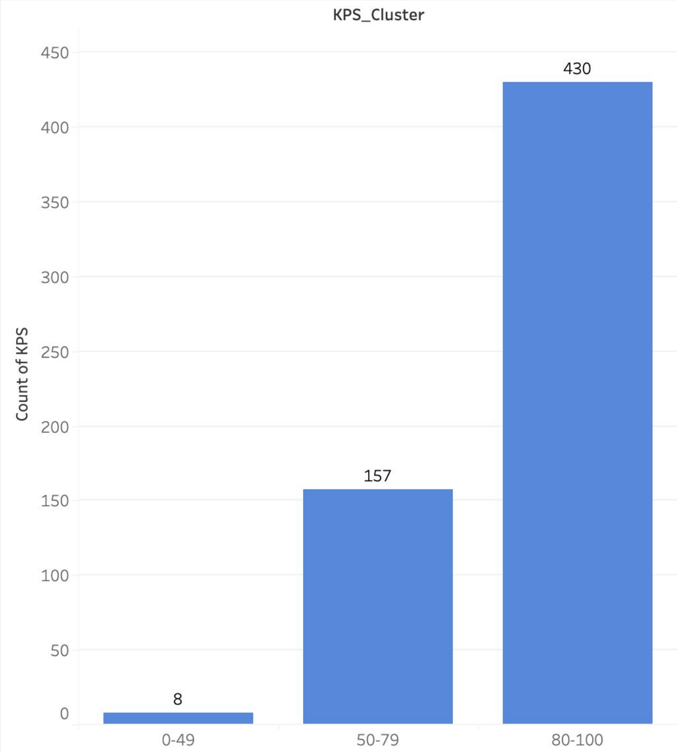
- **Demographic**
  - ➢ Age, etc.
- **Symptoms**
  - ➢ Hair loss, etc.
- **Medical Records**
  - ➢ Cancer type, etc.
- **Psychological Status**
  - ➢ Anxiety and depression tests
- **Cognition Status**
  - ➢ Dementia tests
- **Physical Status**
  - ➢ Weight, body status tests, KPS

Pre-chemo features: cancer type, number of medicine
Post-chemo features: dose level (stdofcare), treatment type

03 DATA VISUALIZATION

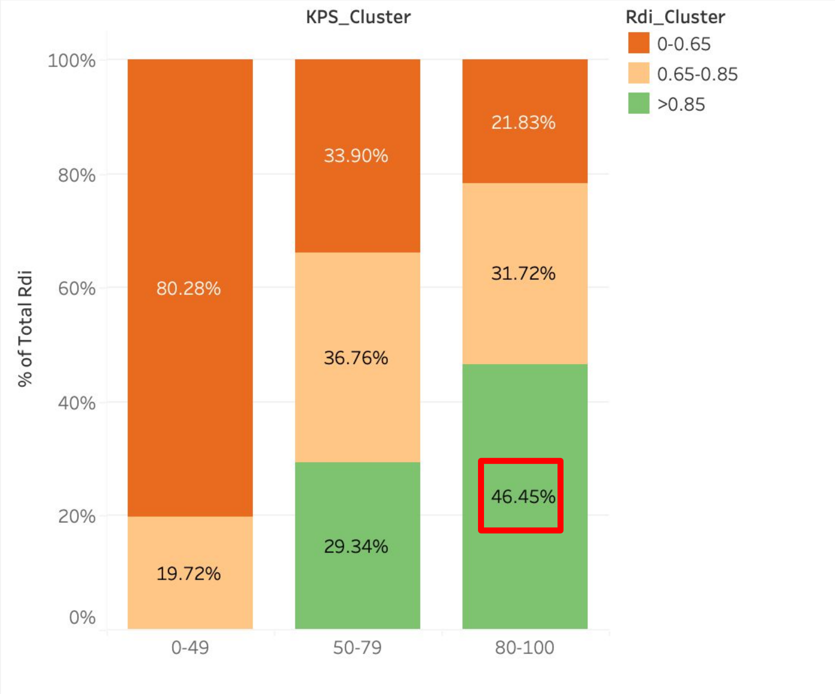
KPS (Karnofsky Performance Status)

KPS	Explanation
0 - 49	Unable to care for self
50 - 79	Unable to work; able to live at home and care for most personal needs
80 - 100	Able to carry on normal activity and to work



03 DATA VISUALIZATION

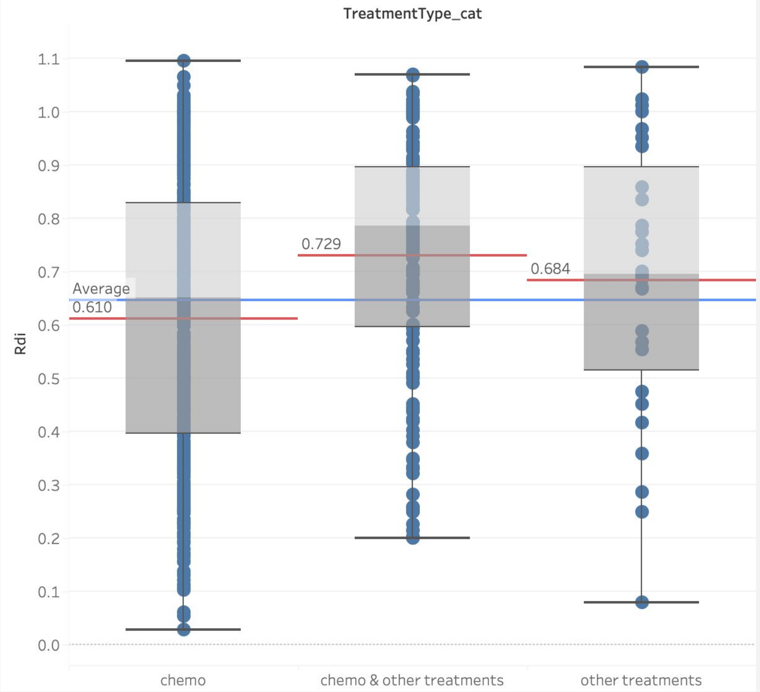
KPS v.s RDI



- **80%** patients in lowest KPS group have RDI **below** 0.65
- Groups with **higher** KPS tend to have more patients with **higher** RDI values

03 DATA VISUALIZATION

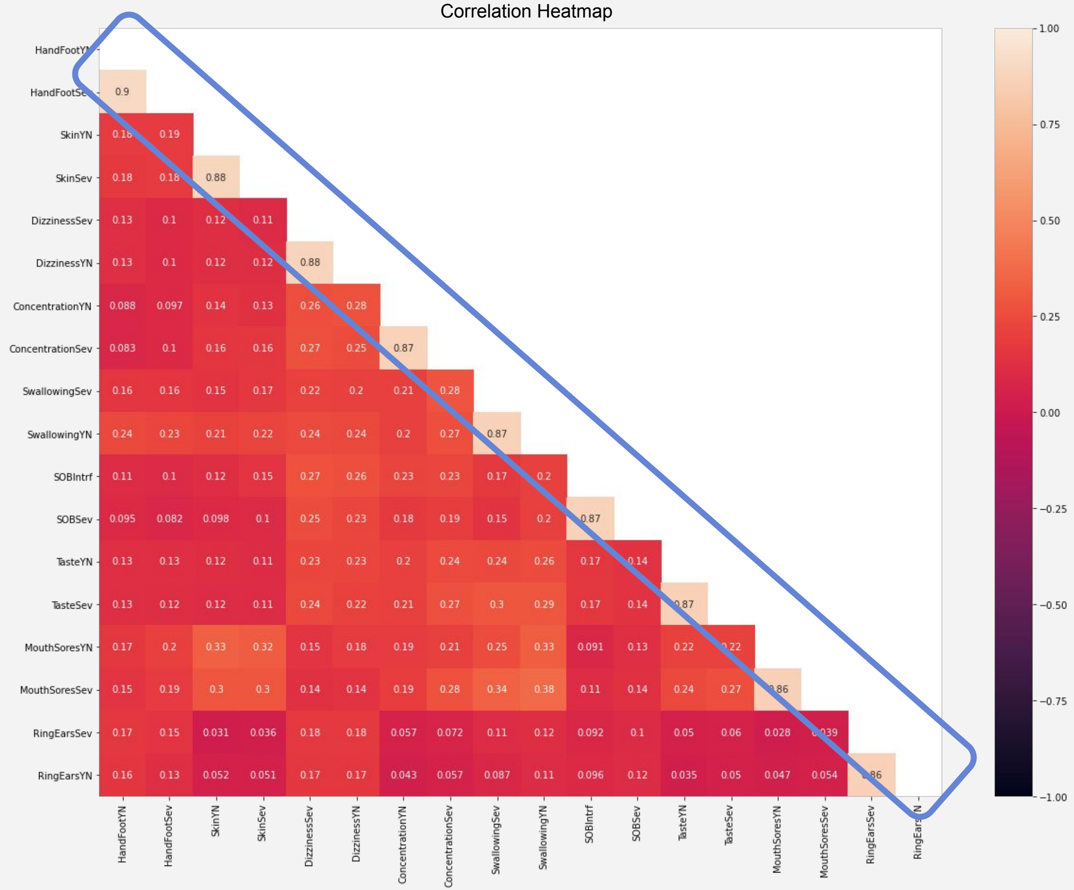
treatment_type v.s RDI



- RDI is **below** the average for the pure chemotherapy group

- Other treatments could largely **increase** the effectiveness of the treatment

04 DESCRIPTIVE ANALYSIS



04 CORRELATION

Weight6MonthsAgo	CurrentWeight	0.952285
HandFootYN	HandFootSev	0.900070
SkinYN	SkinSev	0.883104
DizzinessSev	DizzinessYN	0.876111
ConcentrationYN	ConcentrationSev	0.874495
SwallowingSev	SwallowingYN	0.870621
SOBIntrf	SOBSev	0.869304
TasteYN	TasteSev	0.865022
MouthSoresYN	MouthSoresSev	0.863359
RingEarsSev	RingEarsYN	0.859921

Symptoms

- **Delete Y/N columns ⇒ Keep Severity/Interference columns**
- **weight_change = CurrentWeight - Weight6MonthsAgo**

Predictive Modeling: Classification

Response Variable: **RDI**
Explanatory Variables: Selective features

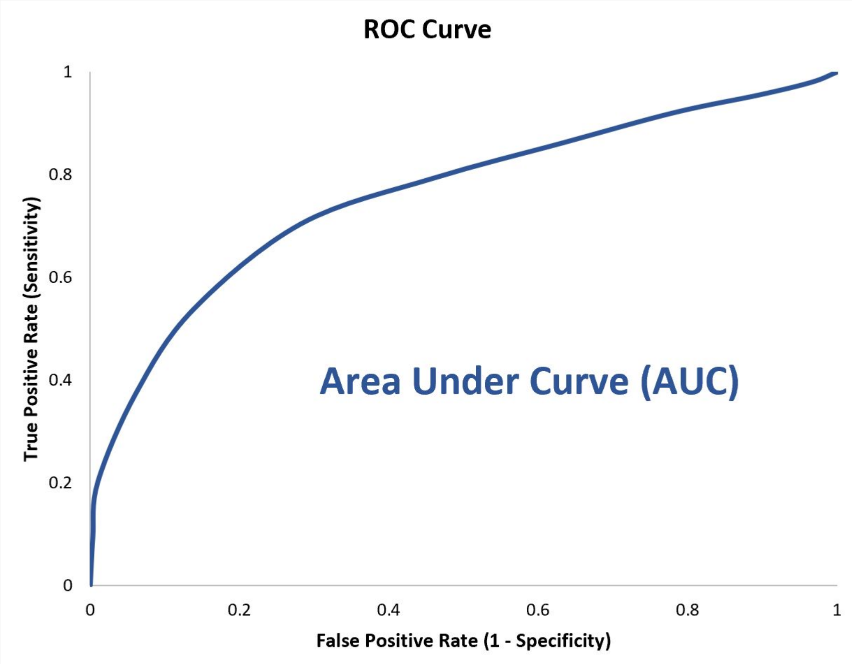
Why Recall?

Predict Class \ True Class	Negative, >0.65	Positive, <=0.65
Negative, >0.65	True Negative	False Negative
Positive, <=0.65	False Positive	True Positive

In reality, patients won't have the **Rdi** value at first. We use models to predict their **Rdi** and decide whether they are able to accept the chemo treatment or not.

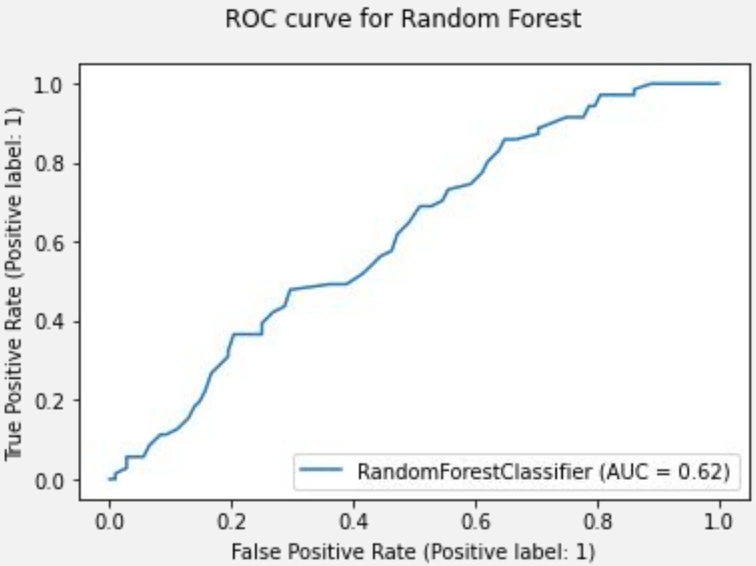
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

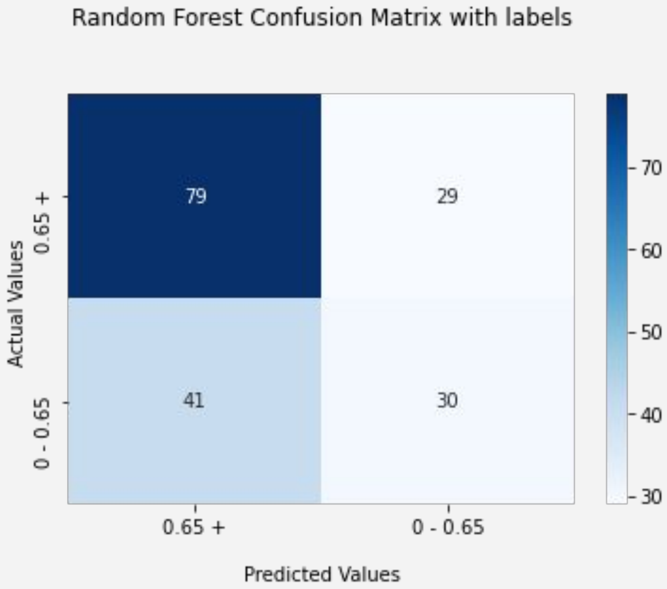
Why ROC Curve AUC?



In clinical epidemiology, ROC analysis is widely used to measure how accurately medical diagnostic tests (or systems) can distinguish between two patient states.

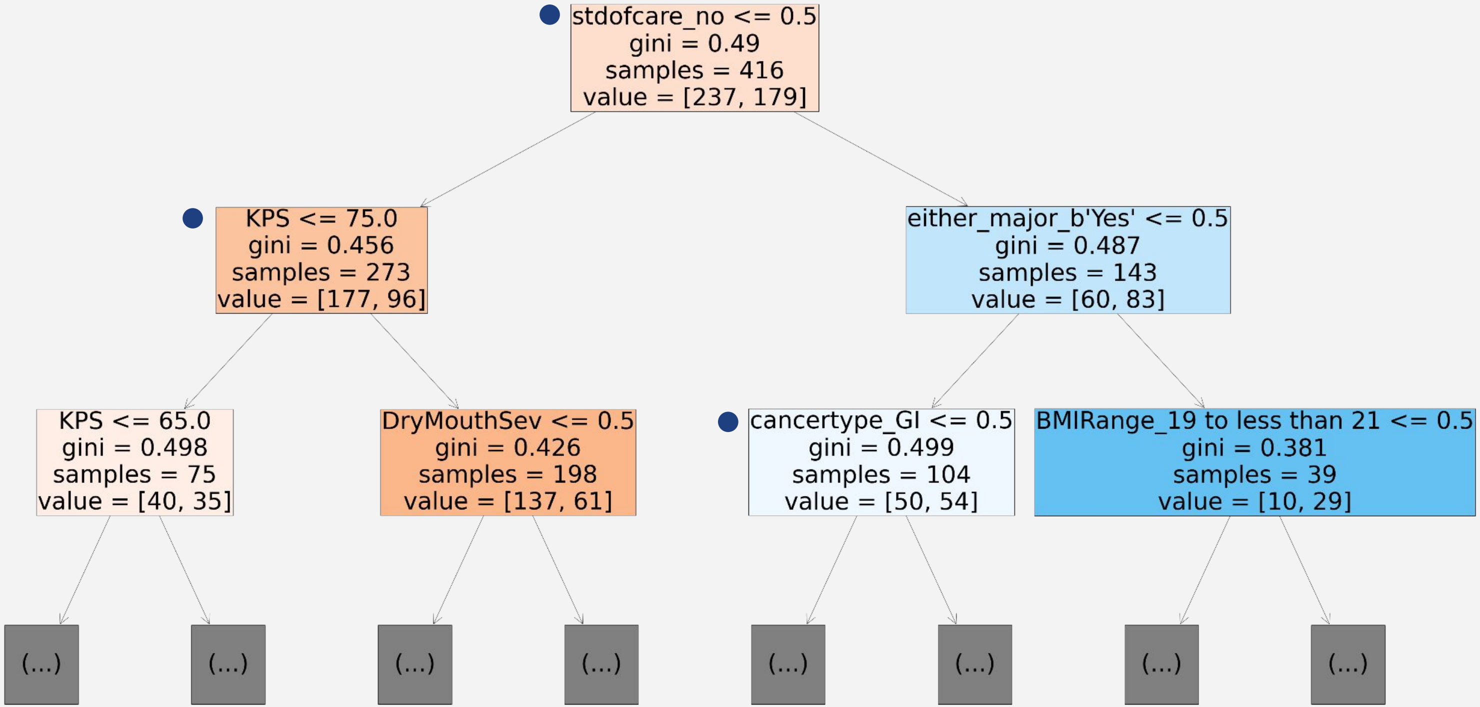
05 RANDOM FOREST



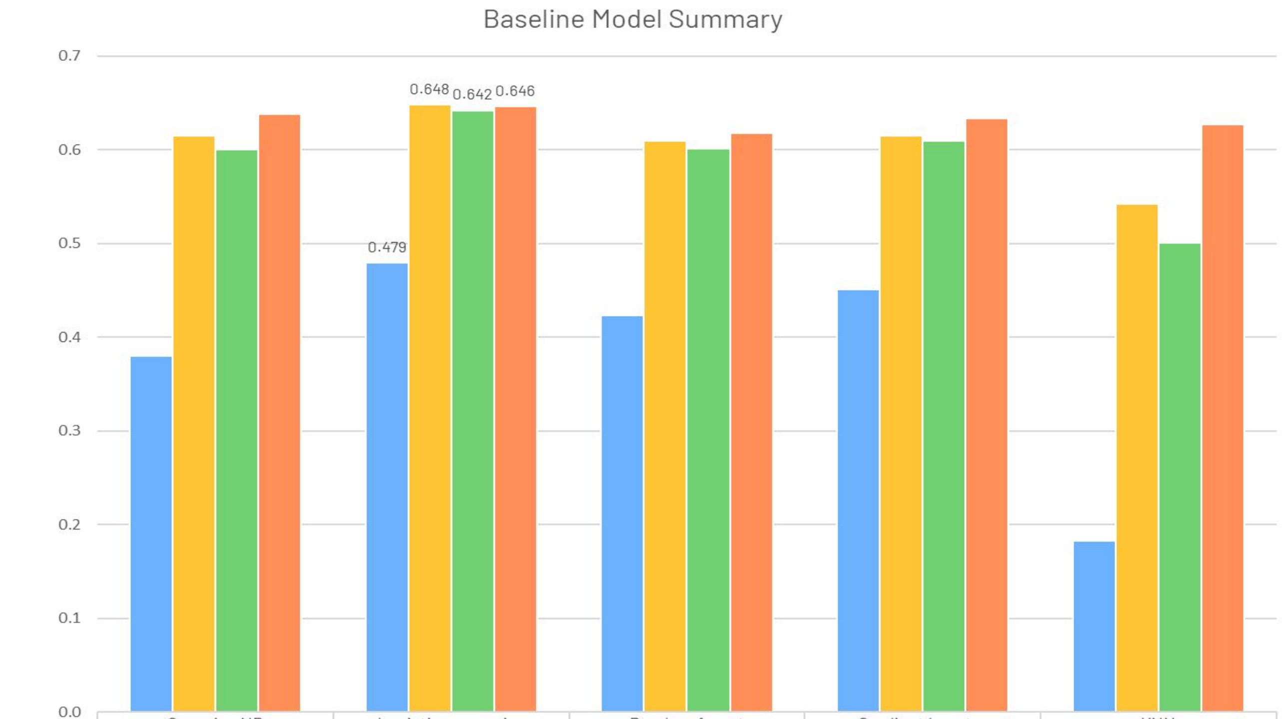


- **Grid Search**
  - Best parameter: {'n_estimators': 1000, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 50, 'bootstrap': False}
- **Accuracy: 0.609; Recall: 0.423; F1: 0.601; AUC: 0.618**

05 RANDOM FOREST



06 PERFORMANCE SUMMARY

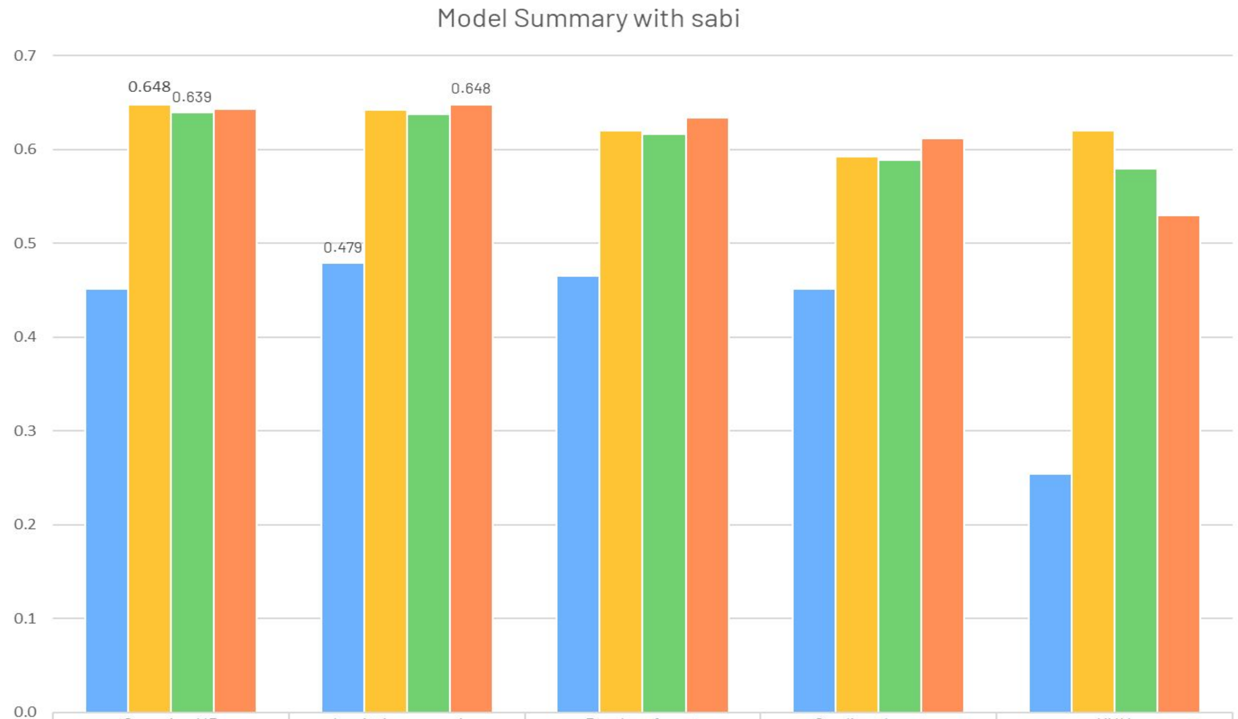


	Gaussian NB	Logistic regression	Random forest	Gradient booster	KNN
Recall	0.380	0.479	0.423	0.451	0.183
Accuracy	0.615	0.648	0.609	0.615	0.542
F1	0.600	0.642	0.601	0.609	0.501
AUC	0.638	0.646	0.618	0.633	0.627

SABI Column

- Integrated numerical column for cancer symptoms
    - Combined all binary results, pain level and interference for a symptom
    - Assigned different weights for different symptoms
- Still under development

06 SUMMARY WITH SABI

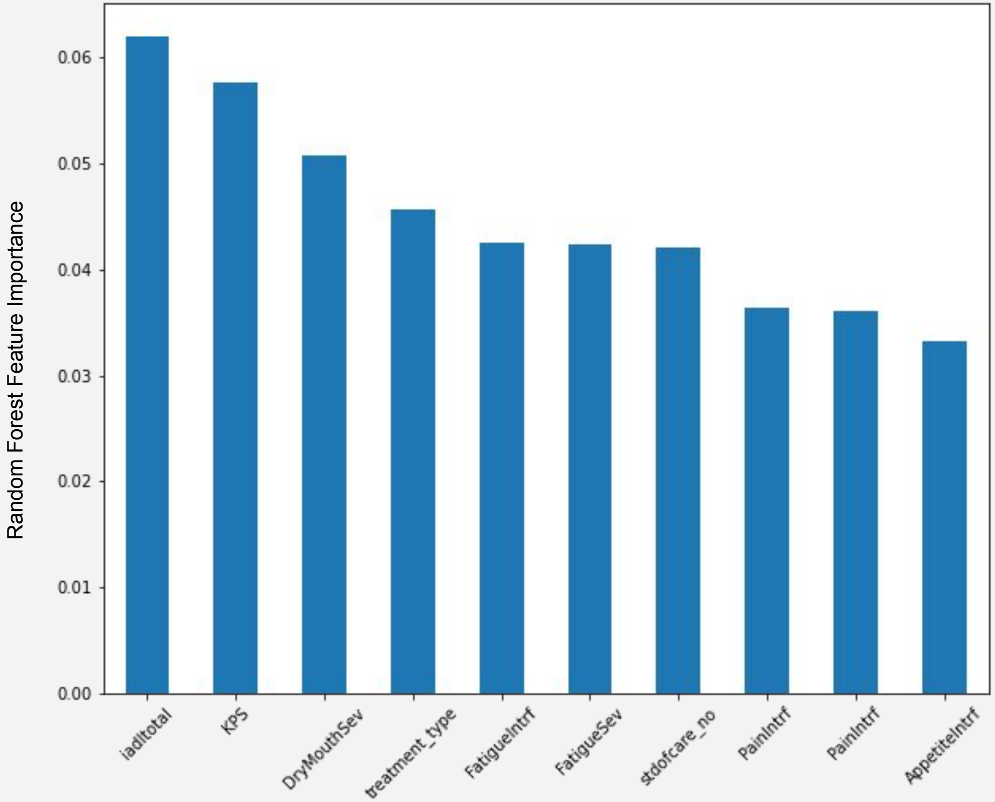


	Gaussian NB	Logistic regression	Random forest	Gradient booster	KNN
Recall	0.451	0.479	0.465	0.451	0.254
Accuracy	0.648	0.642	0.620	0.592	0.620
F1	0.639	0.637	0.616	0.589	0.579
AUC	0.643	0.648	0.634	0.612	0.530

07 PRIMARY INQUIRIES

- Can we use the features to predict RDI value?
- Among all 145 features, which ones are important?

07 FEATURE SELECTION



Random Forest Feature Importance

Top 10 important features

07 FEATURE SELECTION

- **Elastic Net (30 features with lowest MSE 0.063)**
- **Forward/ Backward Feature Selection (top 50 features)**
    - LogisticRegression; scoring: AUC;
- **Random Forest Feature Importance (top 50 features)**

Overlapping features (8 features):

- **Physical Status:** CalcTUG**,** KPS
- **Symptoms:** DizzinessIntrf, FatigueSev, PainIntrf,
- **Medical Record:** cancertype, **stdofcare**, treatment_type

08 PIPELINE REFINEMENT

Aim to process raw data for physicians, could choose various models

- Dimension reduction: PCA, NMF, ICA
- K-fold
- Encoder: Label, OneHot
- Imputer: KNN, DropNA, Mean, Median
- **Feature selection: Ridge, Lasso, Elastic Net**

09 CHALLENGES

Complex Dataset

Feature Selection

Model Improvement

10 KEY INSIGHTS

Can we use the features to predict RDI value?

- **Logistic regression** works the best to predict the **RDI** value.
- Not ideal metric performance

DATA COLLECTION

Among all 145 features, which ones are important?

- Features related to **physical status** are insightful for prediction.
- Information such as **demographic**, **psychological status**, and **cognition status** is not critical.

11 NEXT STEPS...

- Continue on insightful suggestions
- Organize charts, graphs, and codes
- Finish report paper

12 ACKNOWLEDGEMENT

We appreciate the help from **Dr. Ramsdale** as our sponsor, providing detailed guidance on the phrases explanations and feature selection process.

We appreciate the help from **Professor Anand** as our advisor, providing constructive suggestions towards our model building process.

ANY QUESTIONS?

THANKS!