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
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

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

\[
RDI = \frac{\text{Sum of percentage of ideal dose given for each drug}}{\text{Interval between first and last dose}} \times \frac{100 \times \text{no. of drugs}}{\text{Expected interval}}
\]
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
**KPS (Karnofsky Performance Status)**

<table>
<thead>
<tr>
<th>KPS</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 49</td>
<td>Unable to care for self</td>
</tr>
<tr>
<td>50 - 79</td>
<td>Unable to work; able to live at home and care for most personal needs</td>
</tr>
<tr>
<td>80 - 100</td>
<td>Able to carry on normal activity and to work</td>
</tr>
</tbody>
</table>
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
• RDI is **below** the average for the pure chemotherapy group

• Other treatments could largely **increase** the effectiveness of the treatment
DESCRIPTIVE ANALYSIS

Correlation Heatmap
## 04 CORRELATION

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Correlation (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight6MonthsAgo</td>
<td>CurrentWeight</td>
</tr>
<tr>
<td>HandFootYN</td>
<td>HandFootSev</td>
</tr>
<tr>
<td>SkinYN</td>
<td>SkinSev</td>
</tr>
<tr>
<td>DizzinessSev</td>
<td>DizzinessYN</td>
</tr>
<tr>
<td>ConcentrationYN</td>
<td>ConcentrationSev</td>
</tr>
<tr>
<td>SwallowingSev</td>
<td>SwallowingYN</td>
</tr>
<tr>
<td>SOBIntrf</td>
<td>SOBSev</td>
</tr>
<tr>
<td>TasteYN</td>
<td>TasteSev</td>
</tr>
<tr>
<td>MouthSoresYN</td>
<td>MouthSoresSev</td>
</tr>
<tr>
<td>RingEarsSev</td>
<td>RingEarsYN</td>
</tr>
</tbody>
</table>

- **Delete Y/N columns ⇒ Keep Severity/Interference columns**
- **weight_change = CurrentWeight - Weight6MonthsAgo**
Predictive Modeling: Classification

Response Variable: RDI
Explanatory Variables: Selective features
### Why Recall?

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

![Confusion Matrix](https://via.placeholder.com/150)

<table>
<thead>
<tr>
<th>Predict Class</th>
<th>True Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative, &gt;0.65</td>
<td>Negative, &gt;0.65</td>
</tr>
<tr>
<td>Positive, &lt;=0.65</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

Recall = \[ \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \]
In clinical epidemiology, ROC analysis is widely used to measure how accurately medical diagnostic tests (or systems) can distinguish between two patient states.
● **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

- stdofcare_no <= 0.5
  - gini = 0.49
  - samples = 416
  - value = [237, 179]

- either_major_b'Yes' <= 0.5
  - gini = 0.487
  - samples = 143
  - value = [60, 83]

- KPS <= 75.0
  - gini = 0.456
  - samples = 273
  - value = [177, 96]

- KPS <= 65.0
  - gini = 0.498
  - samples = 75
  - value = [40, 35]

- DryMouthSev <= 0.5
  - gini = 0.426
  - samples = 198
  - value = [137, 61]

- cancertype_GI <= 0.5
  - gini = 0.499
  - samples = 104
  - value = [50, 54]

- BMIRange_19 to less than 21 <= 0.5
  - gini = 0.381
  - samples = 39
  - value = [10, 29]
06 PERFORMANCE SUMMARY

Baseline Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian NB</td>
<td>0.380</td>
<td>0.815</td>
<td>0.600</td>
<td>0.838</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.478</td>
<td>0.848</td>
<td>0.642</td>
<td>0.846</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.423</td>
<td>0.809</td>
<td>0.601</td>
<td>0.618</td>
</tr>
<tr>
<td>Gradient booster</td>
<td>0.461</td>
<td>0.815</td>
<td>0.609</td>
<td>0.833</td>
</tr>
<tr>
<td>KNN</td>
<td>0.183</td>
<td>0.542</td>
<td>0.501</td>
<td>0.627</td>
</tr>
</tbody>
</table>
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
SUMMARY WITH SABI

Model Summary with sabi

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian NB</td>
<td>0.451</td>
<td>0.648</td>
<td>0.639</td>
<td>0.648</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.470</td>
<td>0.842</td>
<td>0.637</td>
<td>0.648</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.465</td>
<td>0.820</td>
<td>0.616</td>
<td>0.634</td>
</tr>
<tr>
<td>Gradient booster</td>
<td>0.451</td>
<td>0.592</td>
<td>0.589</td>
<td>0.612</td>
</tr>
<tr>
<td>KNN</td>
<td>0.254</td>
<td>0.820</td>
<td>0.579</td>
<td>0.530</td>
</tr>
</tbody>
</table>
Can we use the features to predict RDI value?

Among all 145 features, which ones are important?
Random Forest Feature Importance

Top 10 important features

- Landuse:
- yes
- DRY_Mouth:
- PTV:
- fatigue:
- Fatigue:
- gdfcare.co:
- Pain:
- Pain:
- Approach1
**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
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

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.
NEXT STEPS...

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