Predictable Delivery of Clinical Trials

Problem Statement

Illustrative Clinical Trial Pipeline

Selected segment

In navigating the complexity of clinical trials our goal is to characterize the performance of our selected clinical trial segment, as well as provide pathways for improved operations.

Impact

6% faster cycle times

Upon leveraging numerous machine learning models, we quantified the impact of two statistically significant operational components. These were found to affect both trial delivery and net present value (NPV).

Deep Dive

Data

Few observations but numerous features to screen
Varying degrees of sparsity associated for each dataset
Complex data structures from a range of sources

Descriptive

To model our dependent variable, we leveraged several Machine Learning families of approaches:

- Interpretable regressions
- Count and ensemble models
- Binary and multiclass classification

Predictive

We were able to achieve strong Out-of-Sample performance predicting on-time delivery of clinical trials aligned with internal time expectations. We used a wide range of models, such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Categorical Boosting.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time predictions</th>
<th>Trial delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Out-of-sample Area Under Curve</td>
<td>0.82</td>
<td></td>
</tr>
</tbody>
</table>

Since we care about interpretability of our models, we found statistically significant parameters in two specification:
- Time predictions of specific trial segment
- On-time delivery with the specified segment

Leveraging our results is going to be the next step in Takeda analysis.

Impact

Upon leveraging numerous machine learning models, we quantified the impact of two statistically significant operational components. These were found to affect both trial delivery and net present value (NPV).

Impact

6% faster cycle times

Upon leveraging numerous machine learning models, we quantified the impact of two statistically significant operational components. These were found to affect both trial delivery and net present value (NPV).

Deep Dive

Data

Few observations but numerous features to screen
Varying degrees of sparsity associated for each dataset
Complex data structures from a range of sources

Descriptive

To model our dependent variable, we leveraged several Machine Learning families of approaches:

- Interpretable regressions
- Count and ensemble models
- Binary and multiclass classification

Predictive

We were able to achieve strong Out-of-Sample performance predicting on-time delivery of clinical trials aligned with internal time expectations. We used a wide range of models, such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Categorical Boosting.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time predictions</th>
<th>Trial delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Out-of-sample Area Under Curve</td>
<td>0.82</td>
<td></td>
</tr>
</tbody>
</table>

Since we care about interpretability of our models, we found statistically significant parameters in two specification:
- Time predictions of specific trial segment
- On-time delivery with the specified segment

Leveraging our results is going to be the next step in Takeda analysis.