Statistical Learning & Machine Learning Interpretability

Mohammed Quazi

PhD Candidate – Statistics

TID, DOIM, UNM

maquazi@salud.unm.edu

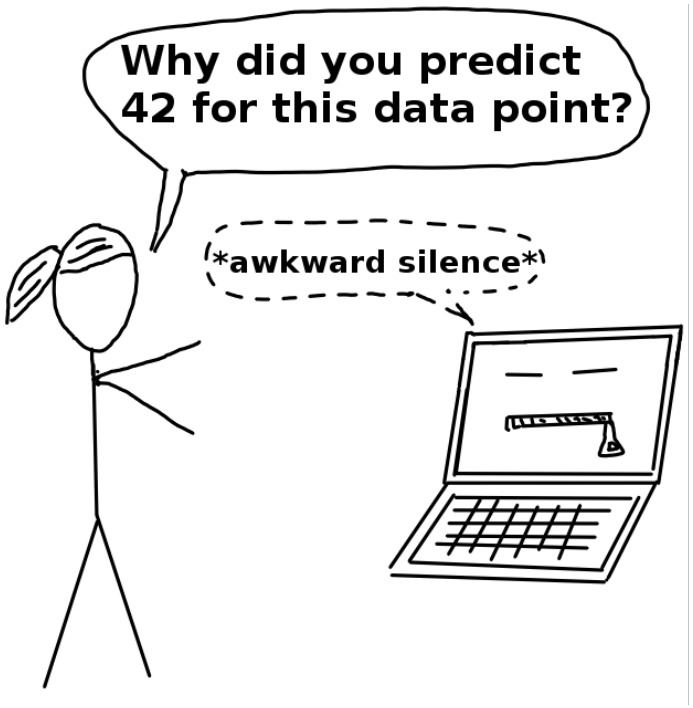
`https://math.unm.edu/~mquazi/`

Overview

1. ML Interpretability
   - Why interpretability?
   - XGBoost
   - Feature importance
   - What drives predictions?
2. Recommendations
   - CARC PI

Why interpretability?

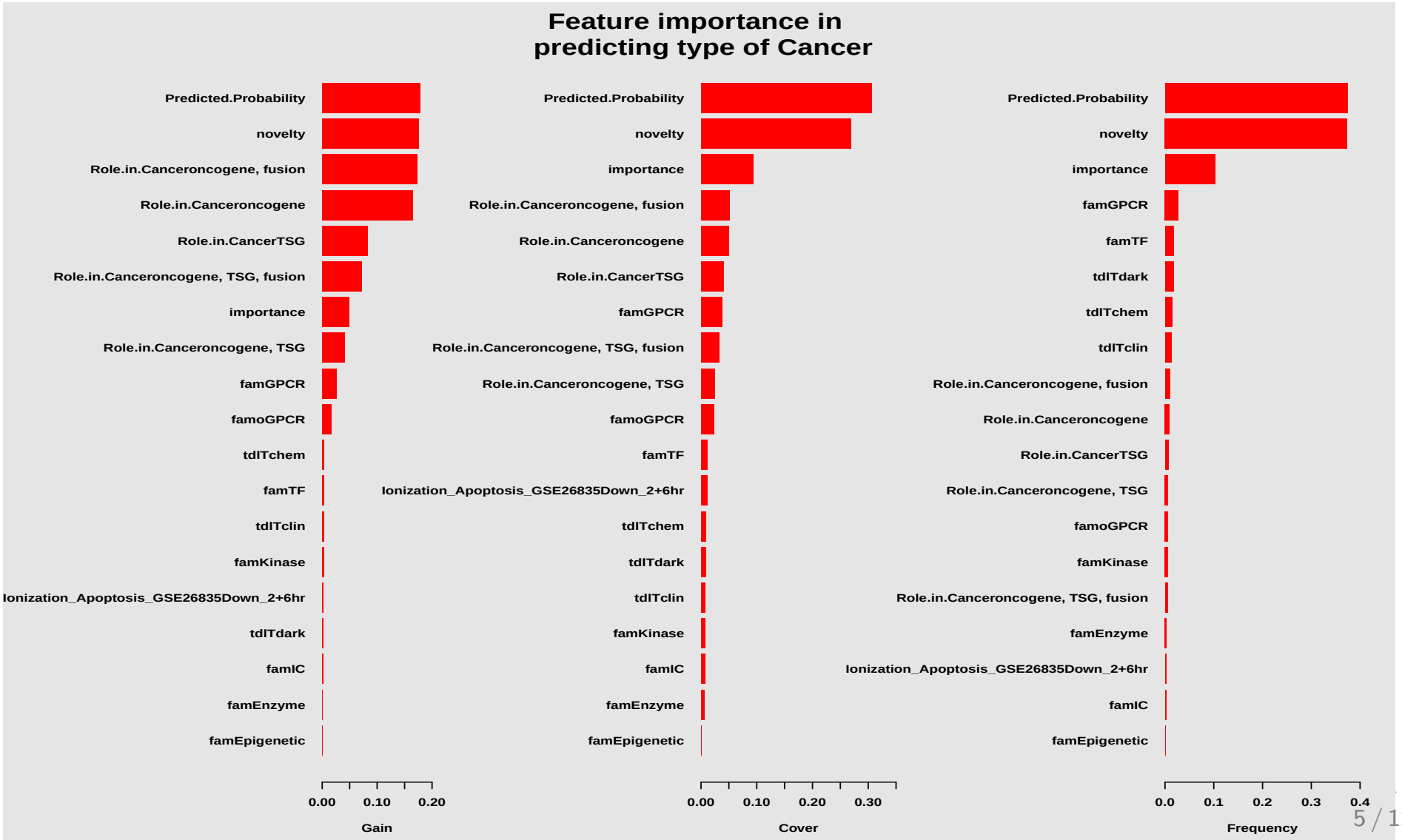
- Definitely not when we are only interested in **what** the model predicts
- We want to know **why** the prediction was made
- Knowing the **'why'** can help us reveal more about the data, the problem, the model and cases where it might fail/improve



Goal: Interpreting the XGBoost algorithm

- Single decision trees are easily interpretable – combinations are not
- By far the most popular ML algorithm today for both regression and classification problems
- Used on MetapathML – disease to gene project
- It's an ensemble – 'black box'
- Used to enhance piezoelectricity in micro-electro-mechanical systems (MEMS)
- It is a statistical model on steroids!

Feature Importance – from MetapathML workflow



Feature Importance – Statistical Descriptions

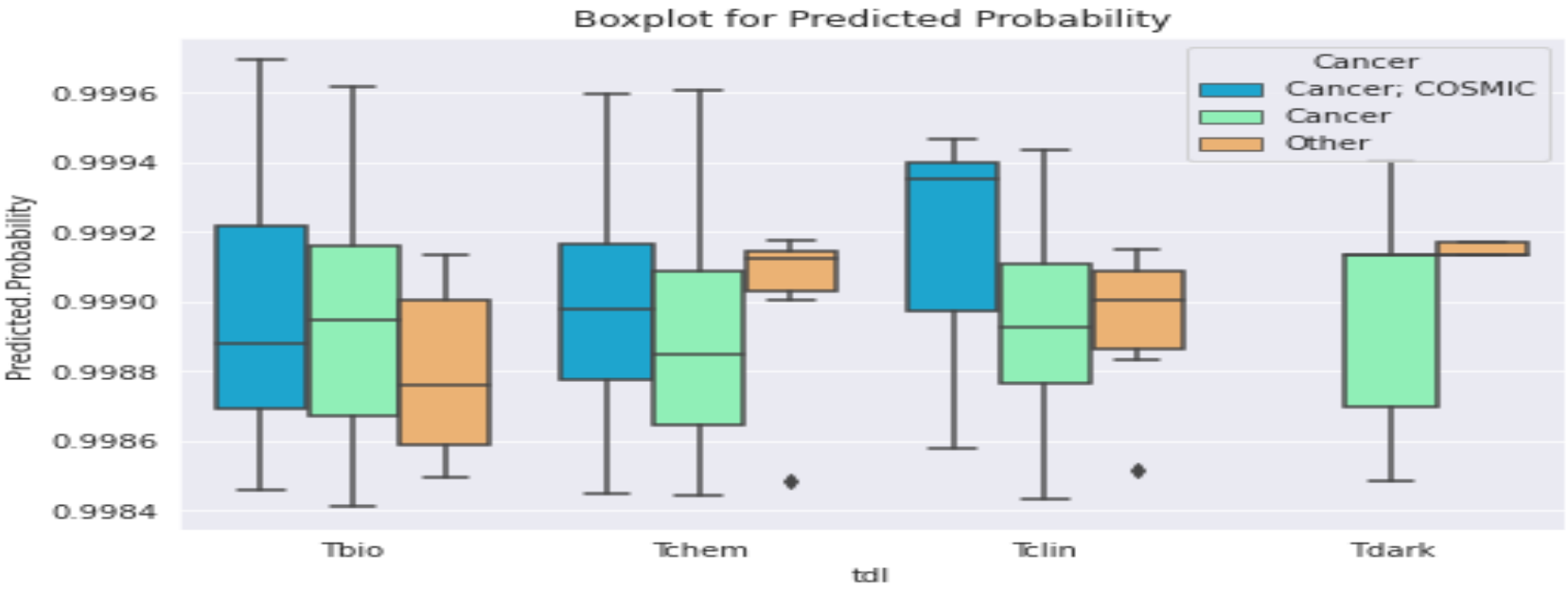
- Gain
  - Naive definition: Average gain across all splits the feature is used in
  - Statistical description: Sum of **squared improvements** over all nodes where that particular variable is chosen as the splitting variable. The reason it is chosen is that it gives maximum estimated improvement in terms of the risk (loss function) over that of a constant fit (simple model). It does not tell us if this feature has to be present or not to get a specific classification.
- Cover
  - Naive definition: Metric of the number of observation related to this feature
  - Statistical description: Percentage of observations (rows) *related* to a particular feature when that feature is selected as the splitting variable.
- Frequency
  - Statistical description: Percentage of relative number of times a feature has been used in all generated trees.

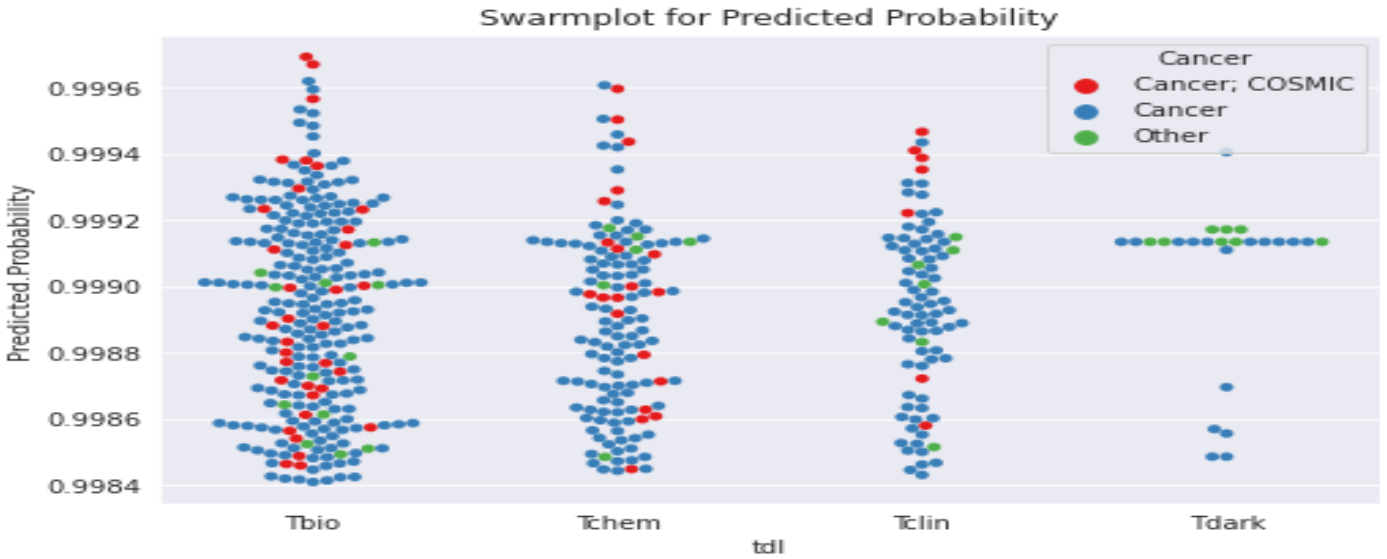
Feature Importance – Statistical Descriptions

For example, famGPCR variable:

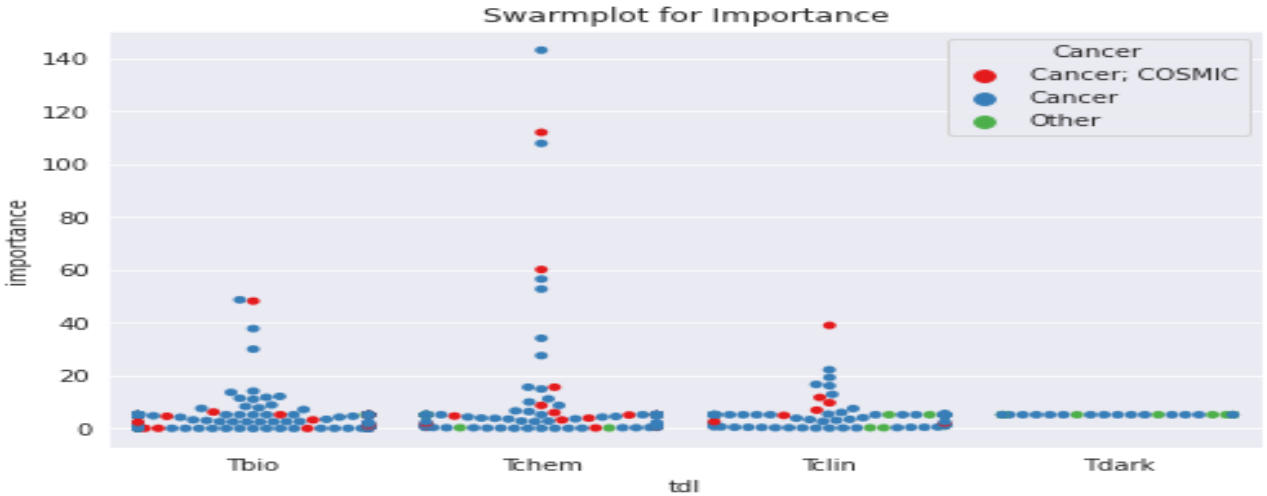
- Is used in 9.2% **(frequency)** of all generated trees
- But when it was used, it produced the maximum estimated improvement 16.44% of the time **(gain)**, hence it was the splitting variable
- And out of those 16.44% of the time, on average, 5% **(cover)** of the whole dataset was concerned with this variable
- Feature importance algorithm favors continuous predictor variables over categorical predictor variables

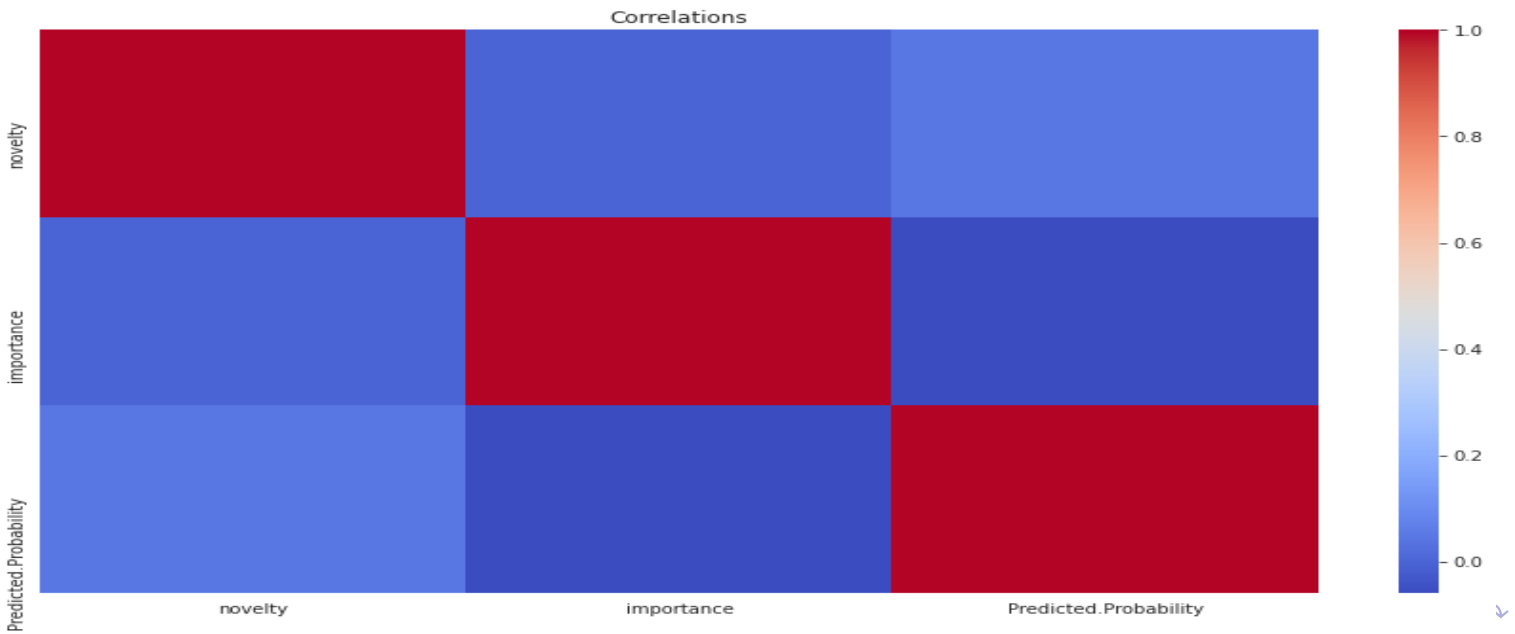
What drives predictions?





What drives predictions?





What drives predictions?

- Highly distiguishable difference in means of groups – similar to treatment means in ANOVA and DOE
- Continuous predictor variables – feature importance favors continuous predictors
- Number of observations in each group – unequal samples/unbalanced designs should be avoided – damages the robustness of the model
- XGBoost nullifies multicollinearity unlike linear regression or GLMs – so high correlation should not affect predictions – this is also a reason for its popularity among all statistical modeling techniques

Recommendations – CARC PI setup

- This will help researchers at TID who use notebooks for computations or are just starting programming
- ISBDS Fall 2021 course students
- Easy to setup R/Python environments (we currently don't have this facility on Hatch)
- Great tech support
- 4 clusters at UNM CARC – Wheeler, Xena, Gibbs and Taos
- But we need a PI under whom other TID researchers/students will setup their accounts
  `https://carc.unm.edu/new-users/request-a-project1.html`

Thank You!