Machine Learning Interpretability

Mohammed Quazi

PhD Candidate – Statistics

Translational Informatics Division, Dept. of Internal Medicine, UNM

maquazi@salud.unm.edu

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

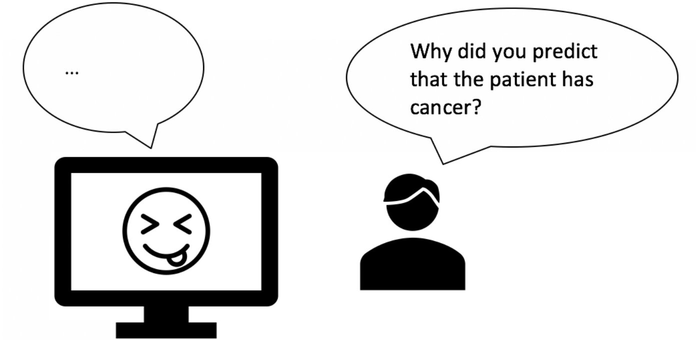
`https://github.com/mquazi`

Overview

1. ML Interpretability
   - Why interpretability?
   - XGBoost
2. Feature Importance – Diabetes – ProteinGraphML
   - Feature Importance
3. Statistical Descriptions
   - Feature Importance – Statistical Descriptions
4. Summary
   - Which features stand out?

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



Consequences

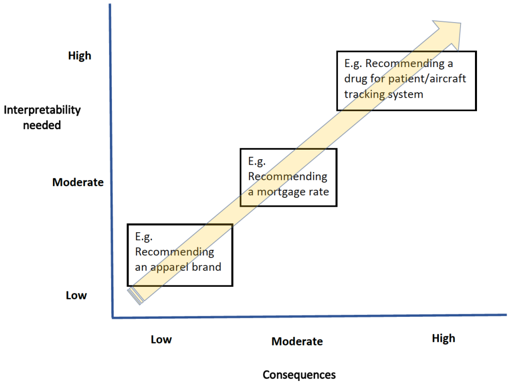

High
Interpretability needed
Moderate
Low
E.g. Recommending a drug for patient/aircraft tracking system
E.g. Recommending a mortgage rate
E.g. Recommending an apparel brand
Low
Moderate
High
Consequences

Figure: Level of interpretability required against consequences of AI/ML outcomes.

Interpreting the XGBoost algorithm

- Single decision trees are easily interpretable – combinations are not
- Popular ML algorithm 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!

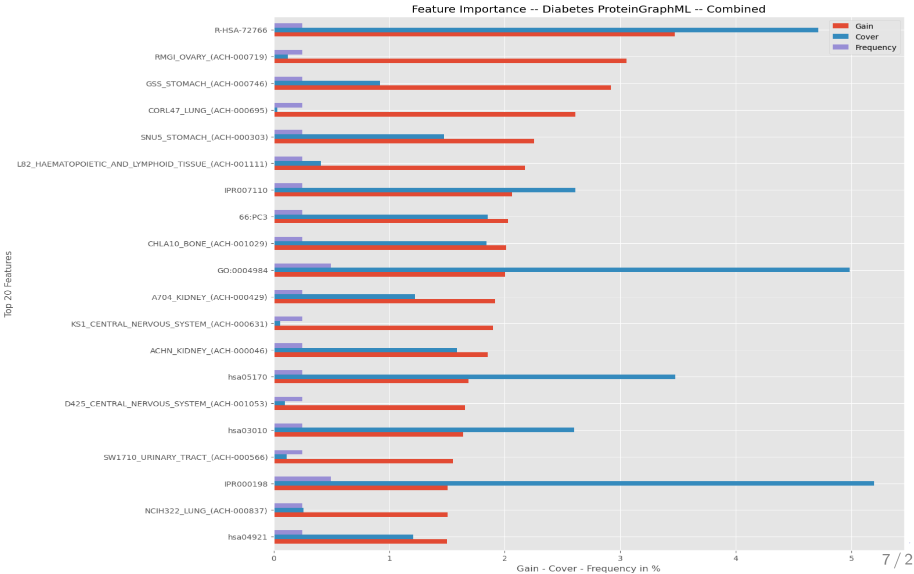
Feature Importance – Diabetes

- Performed FI on the whole training set and on 7 separate feature-classes as well
- Classes are introduced because looking at the whole dataset does not provide useful insights

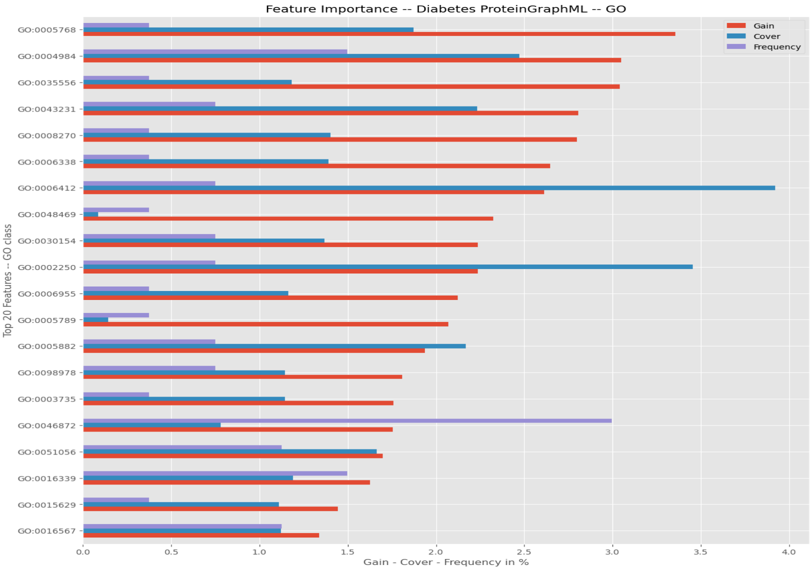
Classes are:

Sl. No.	Class	Source	No. of features
1	GO	Gene Ontology	785
2	ACH – Static	Cancer Cell Line Encyclopedia	1,156
3	R-HSA	Reactome Pathways	340
4	KEGG/HSA	Kyoto Encyclopedia of Genes and Genomes	117
5	IPR	InterPro	362
6	Cell lines – Static	Library of Integrated Network-Based Cellular Signatures	18,997
7	_Cells – Static	The Human Protein Atlas	86
8	*Combined*	All sources	21,843

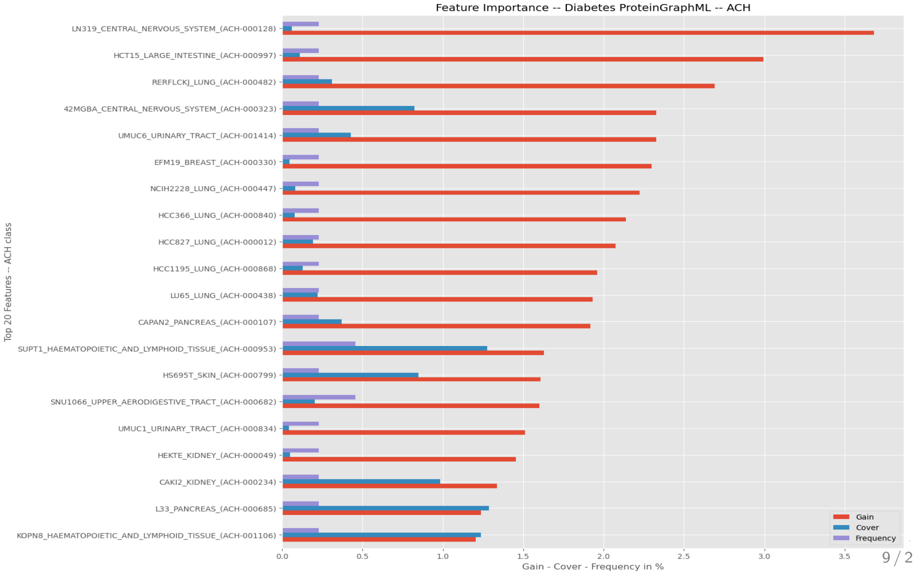
Feature Importance – Combined – 21,843 features



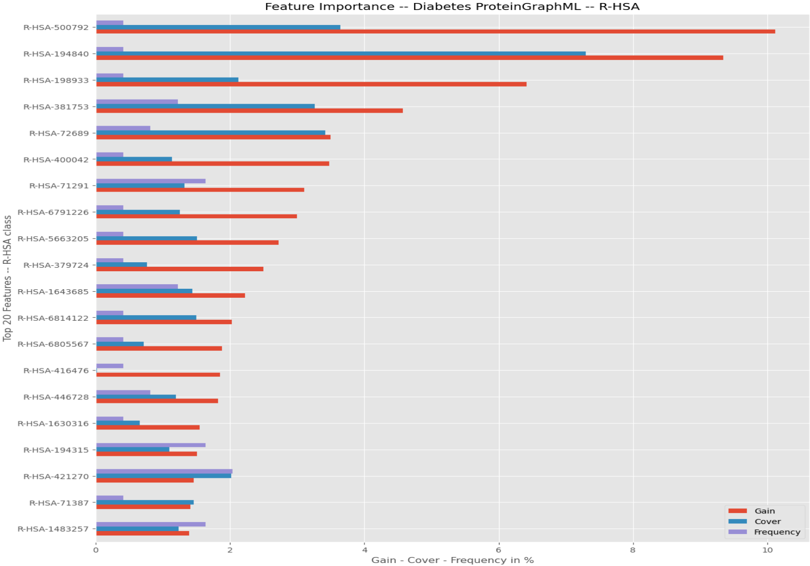
Feature Importance – GO – 785 features



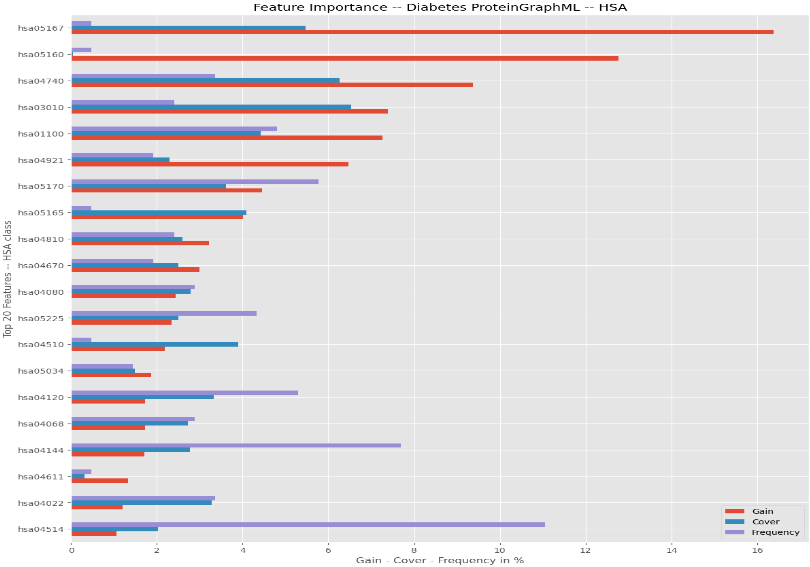
Feature Importance – ACH – 1,156 features



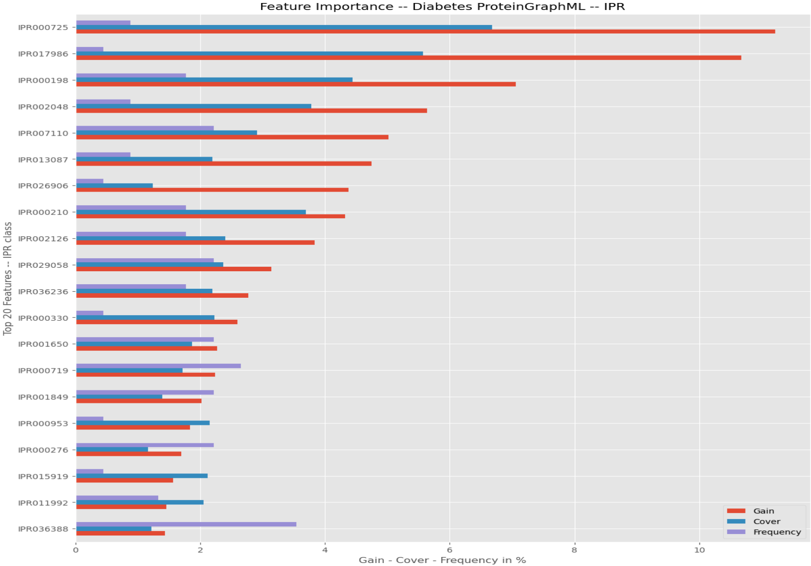
Feature Importance – RHSA – 340 features



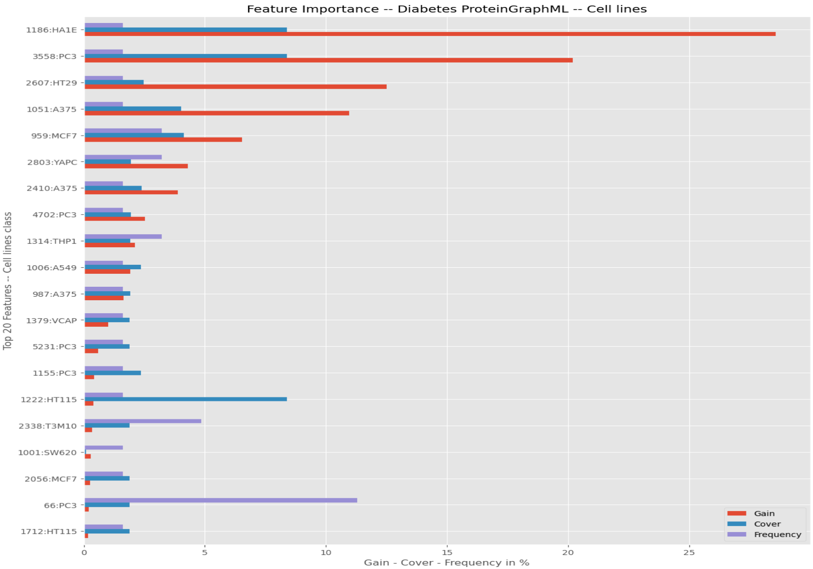
Feature Importance – KEGG/HSA – 117 features



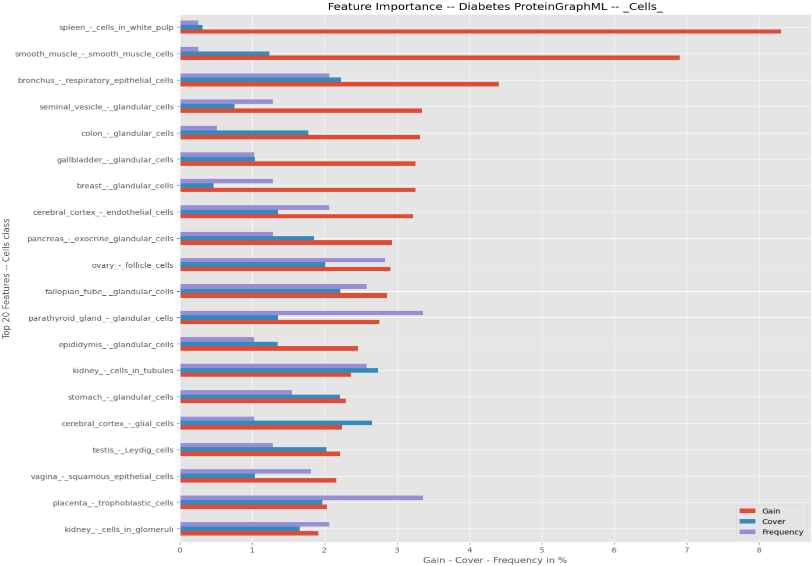
Feature Importance – IPR – 362 features



Feature Importance – Cell lines – 18,997 features



Feature Importance – _Cells – 86 features



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

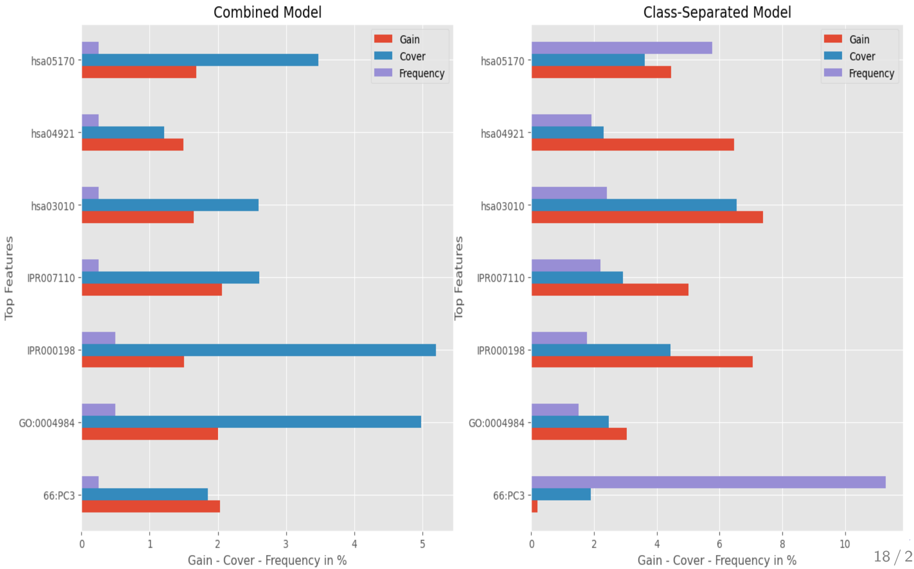
For example, 1186:HA1E variable in Cell lines class:

- Is used in 2% **(frequency)** of all generated trees
- But when it was used, it produced the maximum estimated improvement 29% of the time **(gain)**, hence it was the splitting variable
- And out of those 29% of the time, on average, 8% **(cover)** of the cell lines class dataset was concerned with this variable

What drives predictions?

- Highly distiguishable difference in means of groups – similar to treatment means in ANOVA and DOE
- Feature importance algorithm favors continuous predictor variables over categorical predictor variables
- 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 statistical modeling techniques

Top – *Top* Features



Top – *Top* Features

Sl. No.	Feature	Source Page	Class
1	hsa05170	Human immunodeficiency virus 1 Infection	KEGG/HSA
2	hsa04921	Oxytocin signaling pathway	KEGG/HSA
3	hsa03010	Ribosome	KEGG/HSA
4	IPR007110	Immunoglobulin-like domain	IPR
5	IPR000198	Rho GTPase-activating protein domain	IPR
6	GO:0004984	Olfactory receptor activity	GO
7	66:PC3	Perturbagen ID: 66, LINCS: PC3 Cell line	Cell lines – Static

Notable literature

- Danny Byrd et. als' SHAP: https://github.com/slundberg/shap

Thank You!