When Machine Learning meets Educational Science...

An abstract of my Master's Thesis

Ramaravind Kommiya Mothilal
Master of Artificial Intelligence

When Machine Learning meets Educational Science...

An abstract of my Master’s Thesis

Clarification of Terms

Learning from data

Statistical Modeling = Machine Learning = Data Mining

Inferences & Explanation

Prediction

Mining Patterns, Big Data

Mathematics, Social Science

Engineering, software systems

When Machine Learning meets Educational Science...

An abstract of my Master’s Thesis

Statistical modeling of students’ performance in an open-admission bachelor program in Flanders

The need to discriminate between explanatory and predictive modeling

An Overview:

How to solve ?

What is expected to complete the study program successfully?

Statistical modeling of **students' performance**

Informed decisions

Increase graduation rate

What's the Problem

Low graduation rate

in an **open-admission** **bachelor program**

in **Flanders**

An Overview:

Statistical modeling of students' performance

Independent variables

- Math, Phy, Chem grades
- Math Level
- Effort Level
- Affective Strategies
- Goal Strategies
- Preference for Time Pressure

Dependent variable

Weighted Average of first-semester course scores

Insights about what is expected from students to perform well in first-semester exams

Early indicator of completing the study program

An Overview:

Statistical modeling of **students' performance**

explanatory and **predictive** modeling

Retrospective
Causal explanations

Prospective
Predictions

Explanatory Modeling - A Deeper Look

Formal language	An illustration
	Theoretical domain
$X \xrightarrow{F} Y$	*Prior Knowledge* $\xrightarrow{\text{positively affects}}$ *First-semester Performance*
Operationalization step	*(mostly based on theories)*
	Measurable domain
$x \xrightarrow{f} y$	$wavg = b_1(math) + b_2(phy) + b_3(chem) + \epsilon$
	math, phy, chem grades $\longrightarrow$ *weighted average (wavg)*

Statistical inference $\longrightarrow$ *Confirm or contradict hypotheses in theoretical domain*

Predictive Modeling - A Deeper Look

Formal language

$x \xleftrightarrow{f} y$

An illustration

Measurable domain

wavg = f^*(*math, phy, chem*)

math, phy, chem grades ⟷ *weighted average (wavg)*

→ *Accurate Predictions on holdout set*

🚫 *Not interested in explaining causal relation*

Explanatory and Predictive Modeling - A Deeper Look

Statistical modeling of **students' performance**

The **need** to **discriminate** between

explanatory and **predictive** **modeling**

Retrospective
Causal explanations

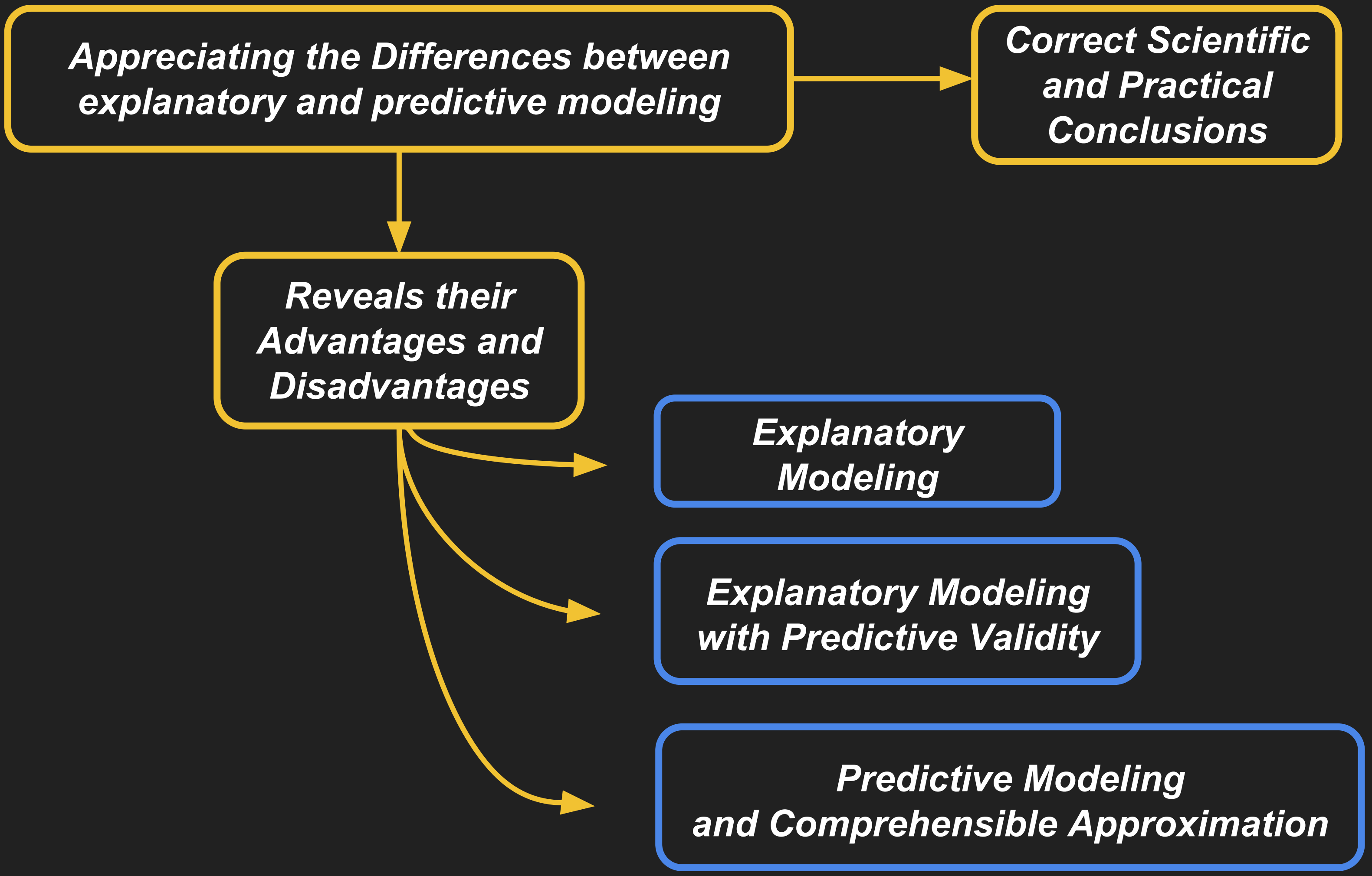
Different goals through the entire modeling process

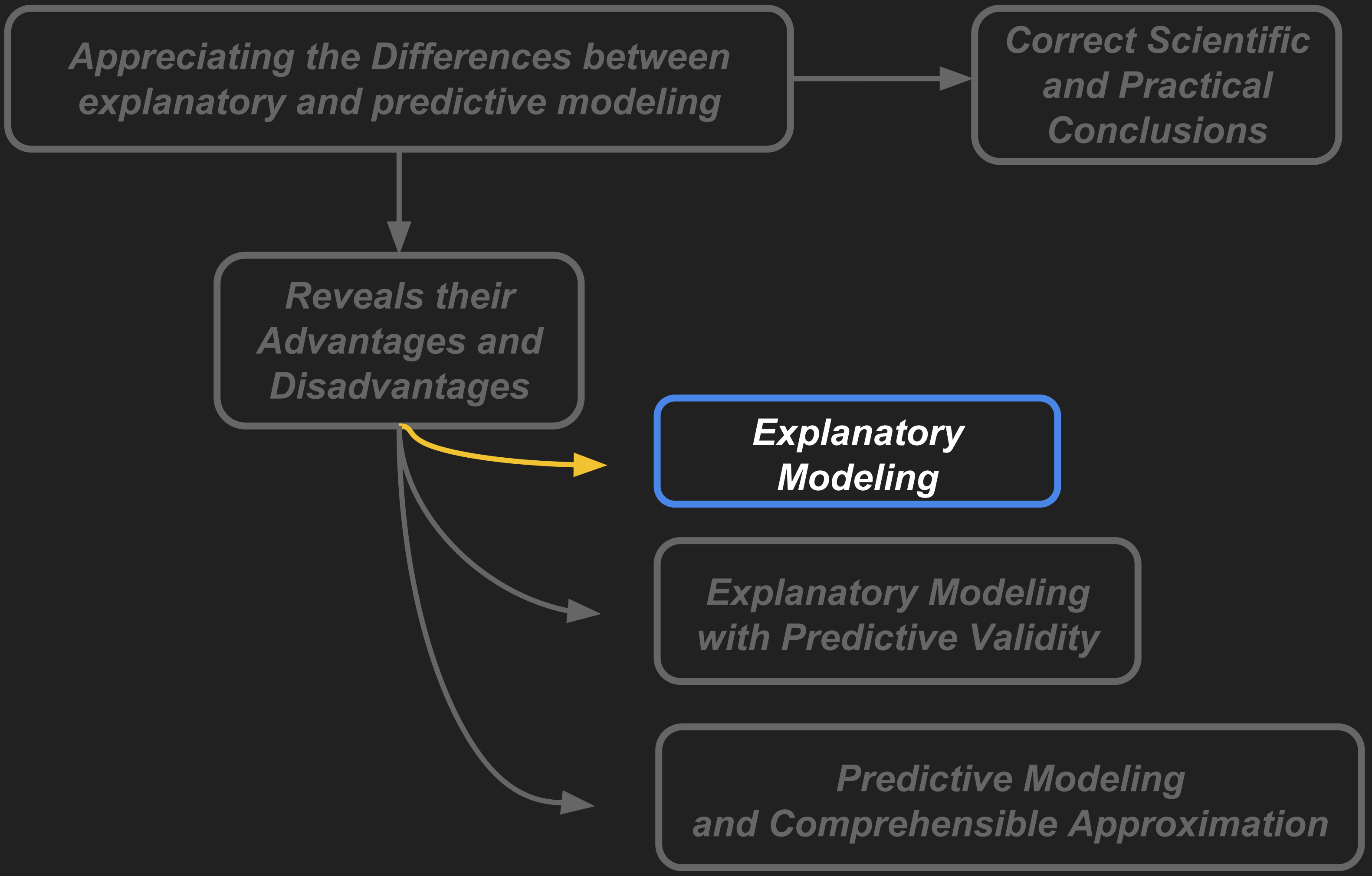
Prospective
Predictions

But, confusion in related literature...

Statistical goals**, choice of methods, **Evaluation criteria

(R^2 for predictive power)


Appreciating the Differences between explanatory and predictive modeling
Correct Scientific and Practical Conclusions
Reveals their Advantages and Disadvantages
Explanatory Modeling
Explanatory Modeling with Predictive Validity
Predictive Modeling and Comprehensible Approximation


Appreciating the Differences between explanatory and predictive modeling
Correct Scientific and Practical Conclusions
Reveals their Advantages and Disadvantages
Explanatory Modeling
Explanatory Modeling with Predictive Validity
Predictive Modeling and Comprehensible Approximation

Explanatory Modeling	Explanatory Modeling with Predictive Validity	Predictive Modeling and Comprehensible Approximation

Outcomes:

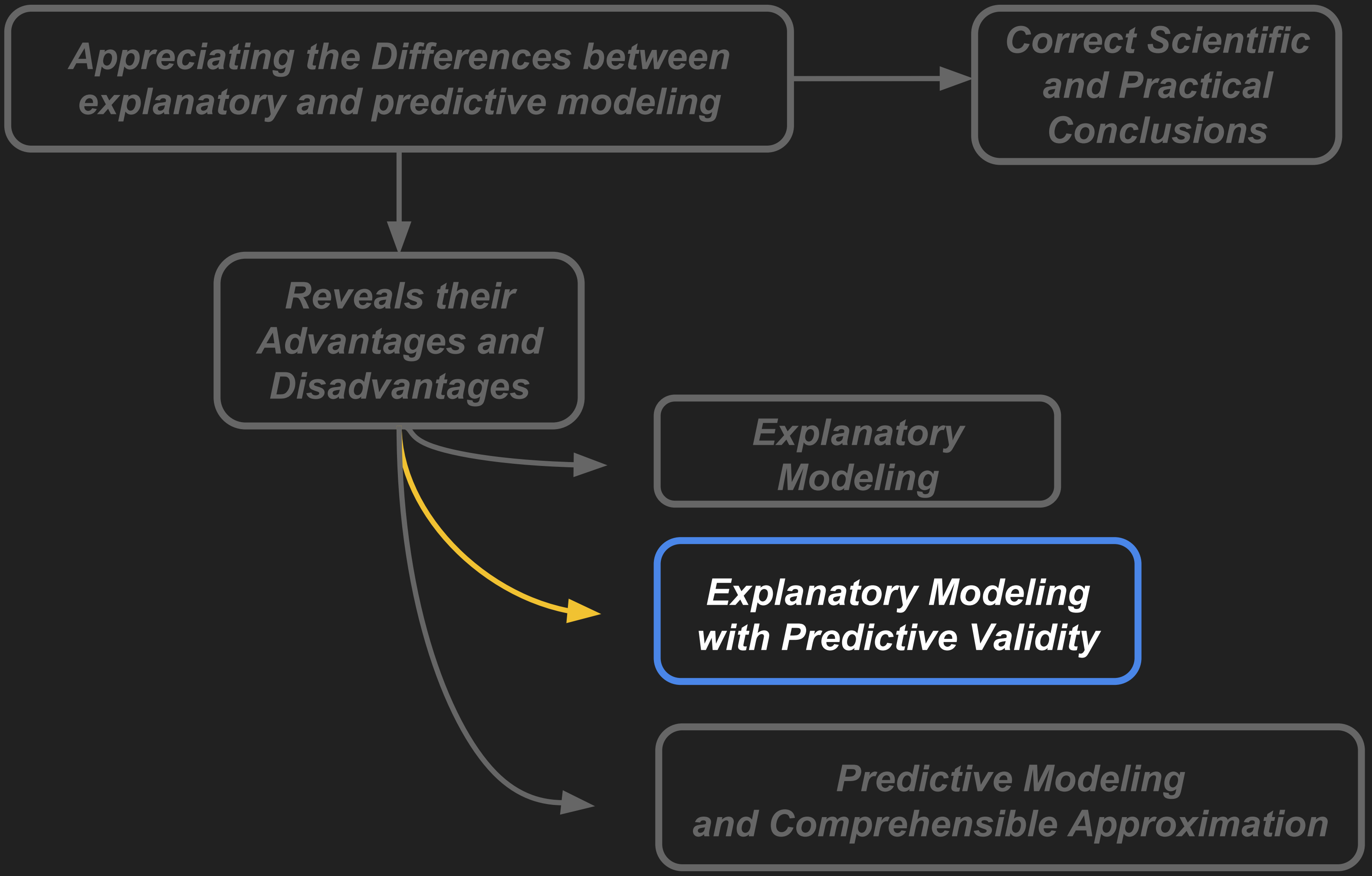
***Model:* Multiple Linear Regression**

Prior academic knowledge (*math, phy, chem* and *hrs*) **positively affects** → 36% **First-semester performance (*wavg*)**

Affective and Goal strategies (*affe* and *goal*) **positively affects** → 4% **First-semester performance (*wavg*)**

Preference for Time pressure (*press*) **does not influence** **First-semester performance (*wavg*)**

Average-case statistical patterns


Appreciating the Differences between explanatory and predictive modeling
Correct Scientific and Practical Conclusions
Reveals their Advantages and Disadvantages
Explanatory Modeling
Explanatory Modeling with Predictive Validity
Predictive Modeling and Comprehensible Approximation

Explanatory Modeling	**Explanatory Modeling with Predictive Validity**	Predictive Modeling and Comprehensible Approximation

Why Optimize both explanatory and predictive goals?

- Explanations are retrospective
  - Need to Generalize statistical inferences to "unseen" incoming students

Explanatory Path:

Similar statistical inferences

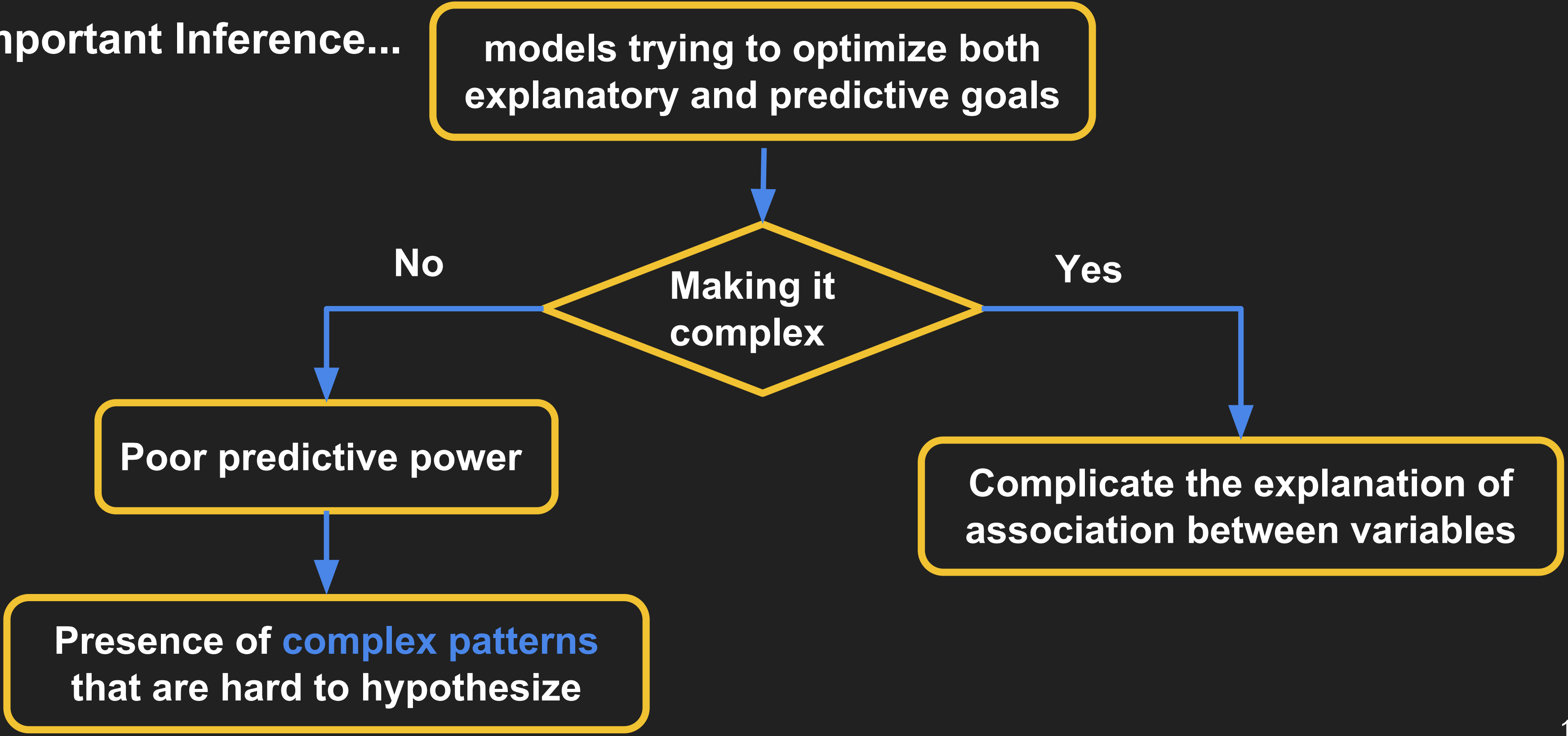
***Model:* Ordinal Logistic Regression**

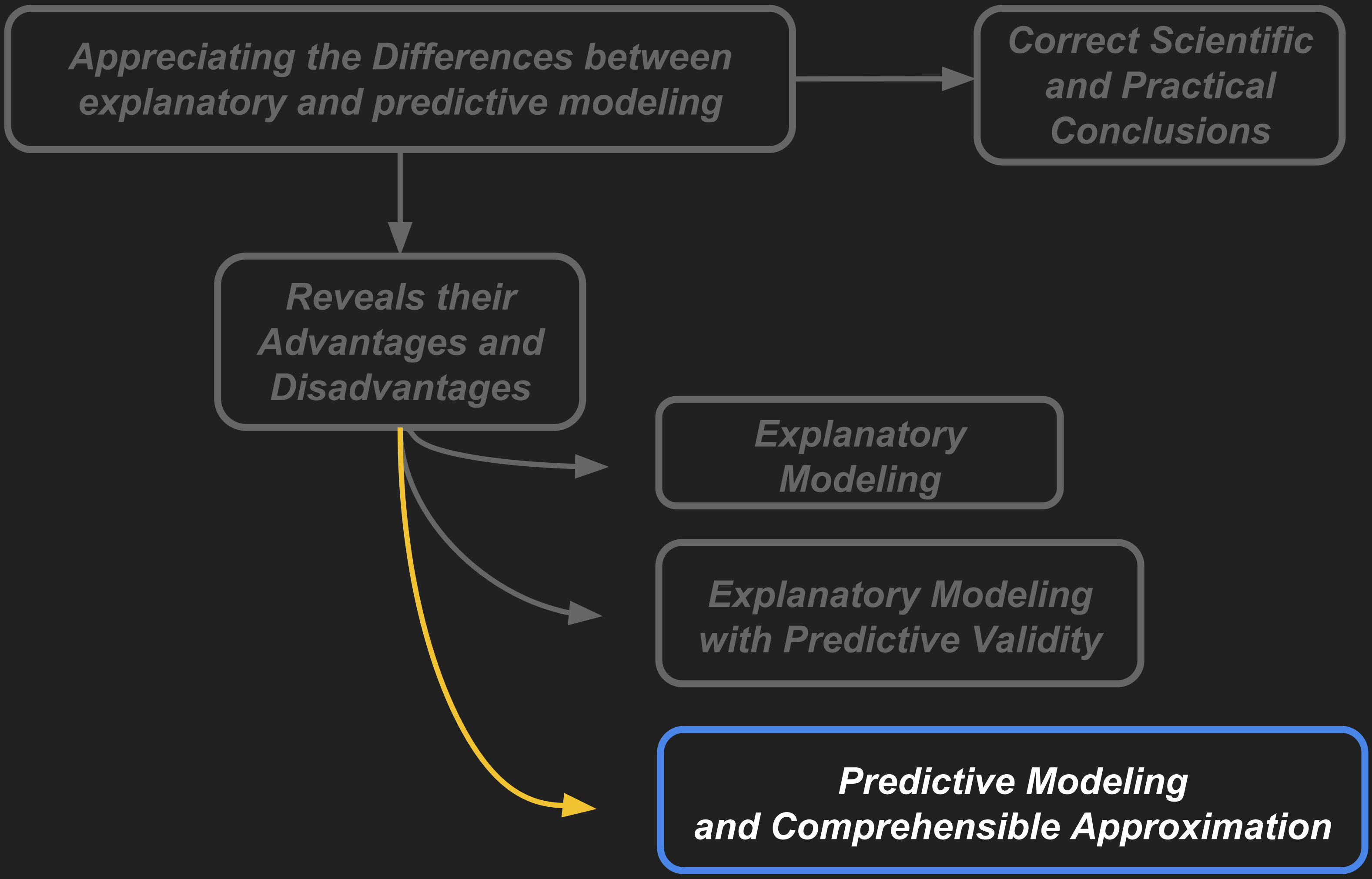
Predictive Validity:

Low predictive power → Cannot generalize statistical inferences to "unseen" incoming students

- Predicts only 60% students in the "at-risk" zones
- Wrongly predicts too many students (around 60%) as "no-risk"

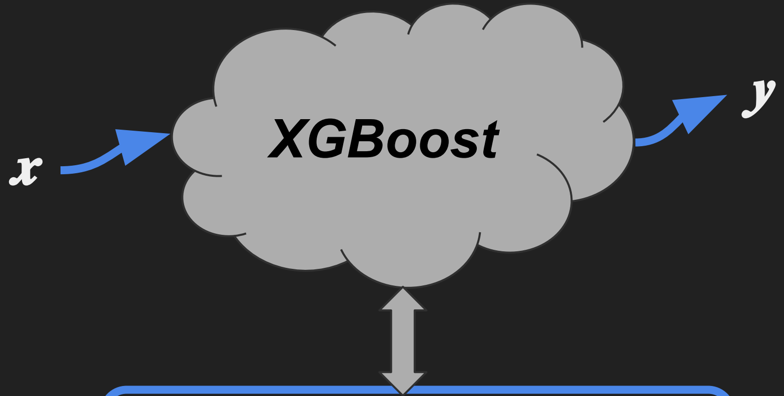
Important Inference...




Appreciating the Differences between explanatory and predictive modeling
Correct Scientific and Practical Conclusions
Reveals their Advantages and Disadvantages
Explanatory Modeling
Explanatory Modeling with Predictive Validity
Predictive Modeling and Comprehensible Approximation

Why Predictive Modeling and Comprehensible Approximation?

Generalize ~~statistical inferences to "unseen" incoming students~~ predictive performance



- At-risk
- Moderate-risk
- No-risk

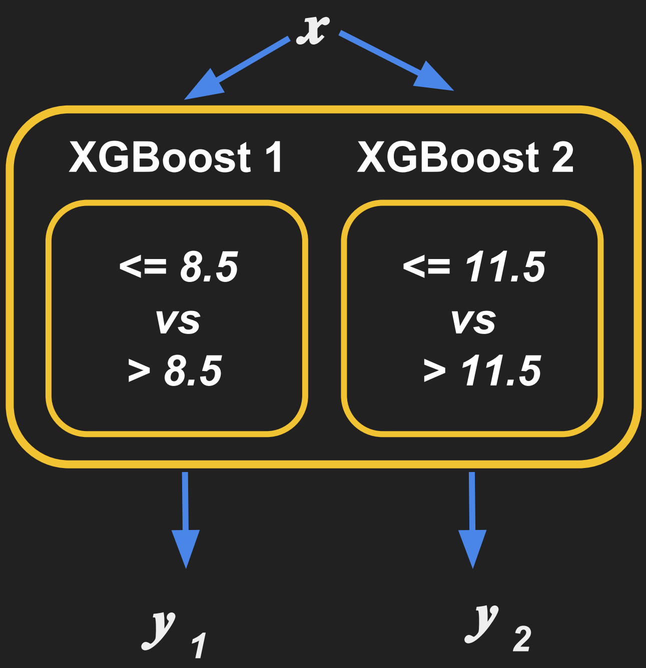
Explanations by LIME
(qualitative understanding)

Comprehensible Approximation

Build trust in predictions

Identify complex patterns that are hard to hypothesize

Model Design :



y_1	y_2	*Interpretation*
<= 8.5	<= 11.5	at-risk
> 8.5	<= 11.5	moderate-risk
> 8.5	> 11.5	no-risk

Why two-binary-classifiers model ?

- difficult to identify students in the "moderate-risk" zone
- easy to train and optimize

Explanatory Modeling	Explanatory Modeling with Predictive Validity	**Predictive Modeling and Comprehensible Approximation**

✓ *Overall Good Predictive Performance*

XGBoost 1	Precision	Recall
<= 8.5	0.64	0.80
> 8.5	0.88	0.77
w. avg	0.80	0.78

Low misclassification of "at-risk" students

Accurately predicts "at-risk" and "moderate-risk" students

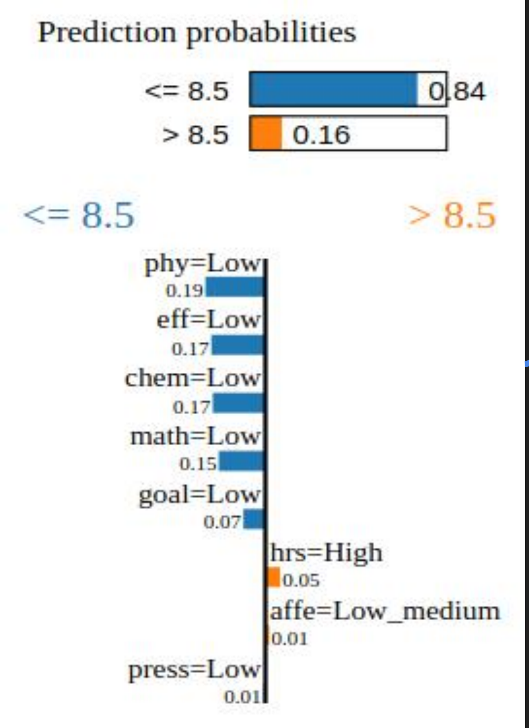
XGBoost 2	Precision	Recall
<= 11.5	0.87	0.85
> 11.5	0.68	0.70
w. avg	0.81	0.81

Significant misclassification of "moderate-risk" students

Explanatory Modeling	Explanatory Modeling with Predictive Validity	Predictive Modeling and Comprehensible Approximation

Trusting correct identification of "at-risk" students - LIME Explanations

Example 1



A Low background in *math, phy, chem* contributes 19 + 17 + 15 = 51% of P(<= 8.5)

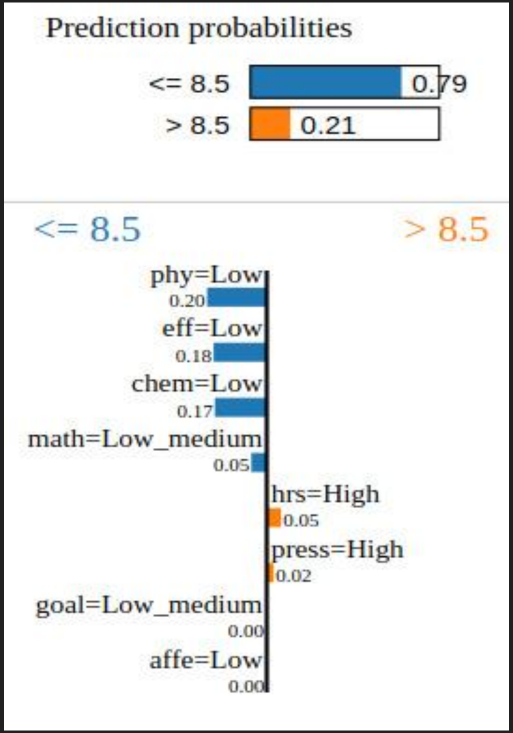
Observation:

"Low" value for most of the IVs contributes heavily to class "<= 8.5" (at-risk students)

Practical Implication:

Individual counselling

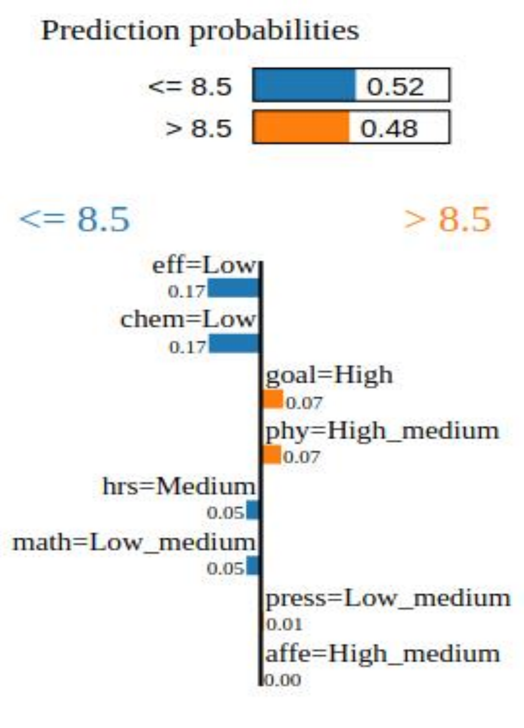
Example 2

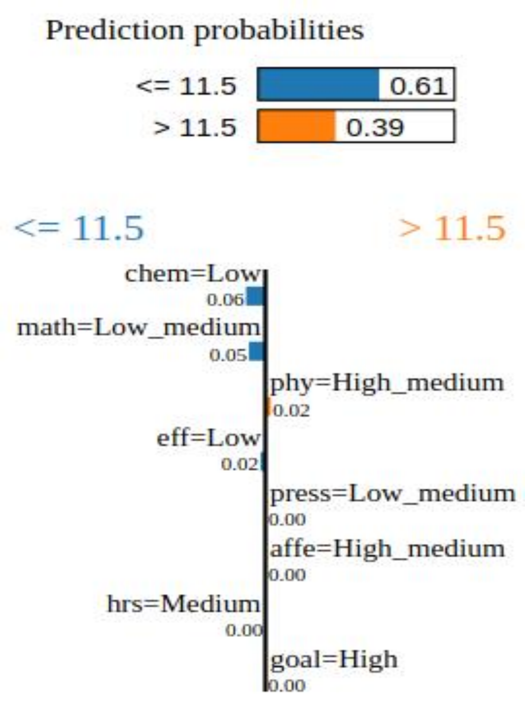


Explanatory Modeling	Explanatory Modeling with Predictive Validity	**Predictive Modeling and Comprehensible Approximation**

Trusting correct identification of “at-risk” students - LIME Explanations

Example 3





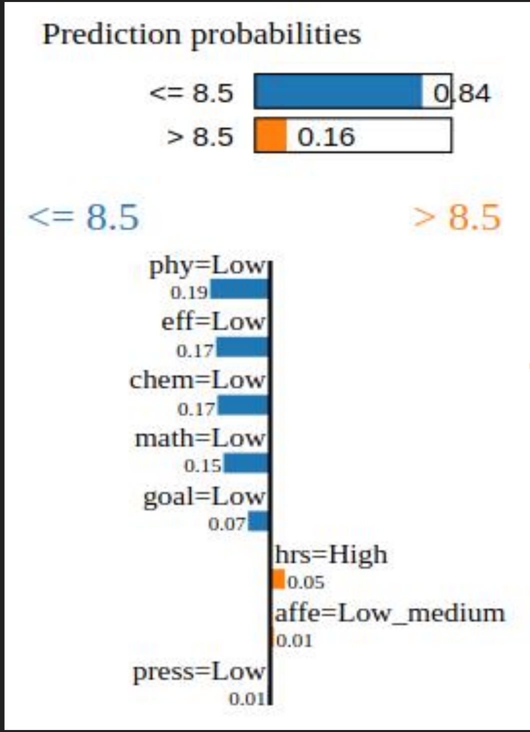
Observation:

Moving towards the higher levels of an IV increases the chances of being not classified as “at-risk” (i.e, “>8.5”)

Practical Implication:

Individual counselling

Further exploitation of LIME Explanations...



Type I

The frequencies of different values taken by an input variable

Type II

The frequencies of an input variable displayed at different positions in the lists of contribution

Subtle average-case patterns

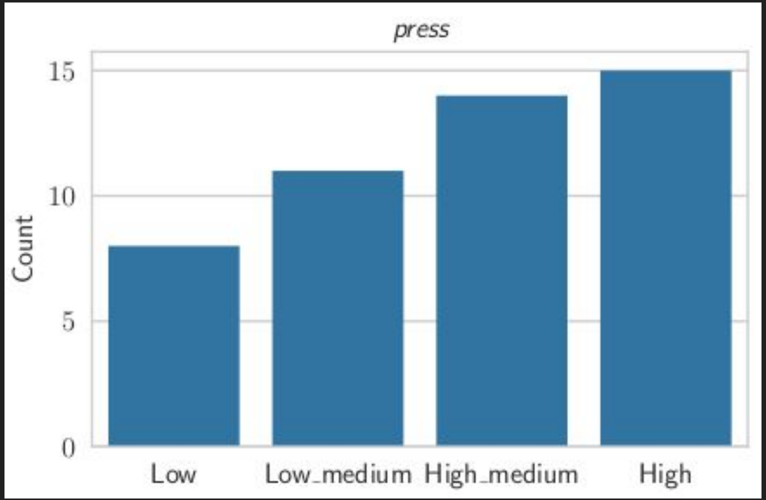
Identify **complex patterns** that are hard to hypothesize

Explanatory Modeling | Explanatory Modeling with Predictive Validity | **Predictive Modeling and Comprehensible Approximation**

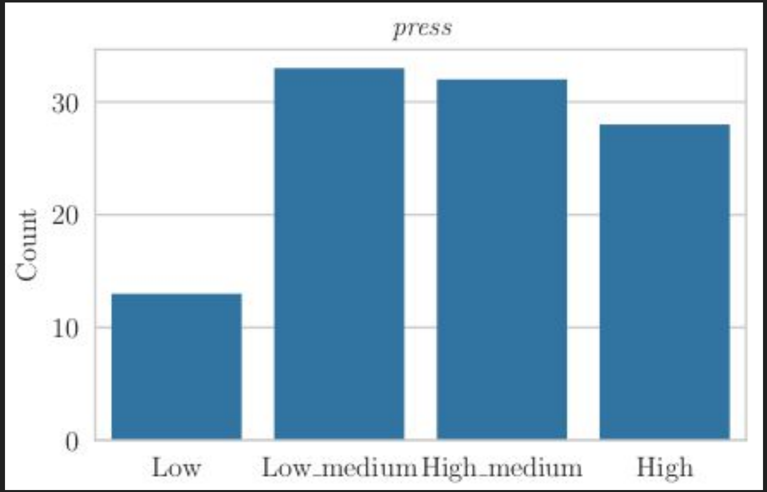
Type I plot using LIME Explanations

High levels of "preference for time pressure" are associated with students in the "at-risk" and "moderate-risk" zones

at-risk students ("<= 8.5")



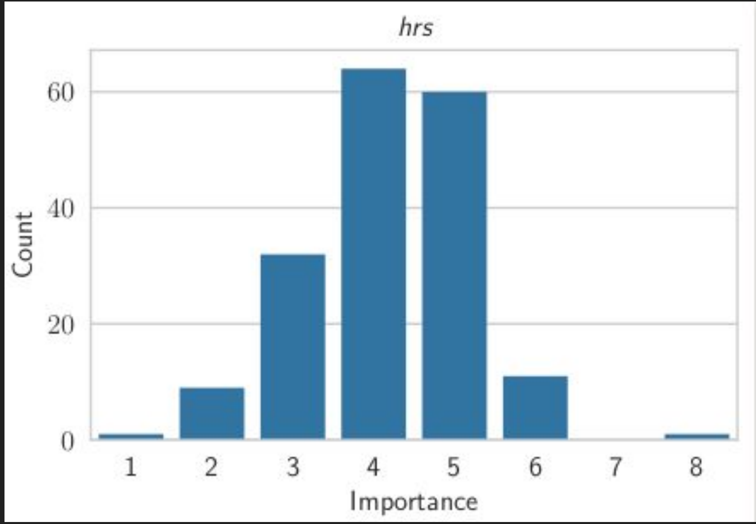
at-risk + moderate-risk students ("<= 11.5")



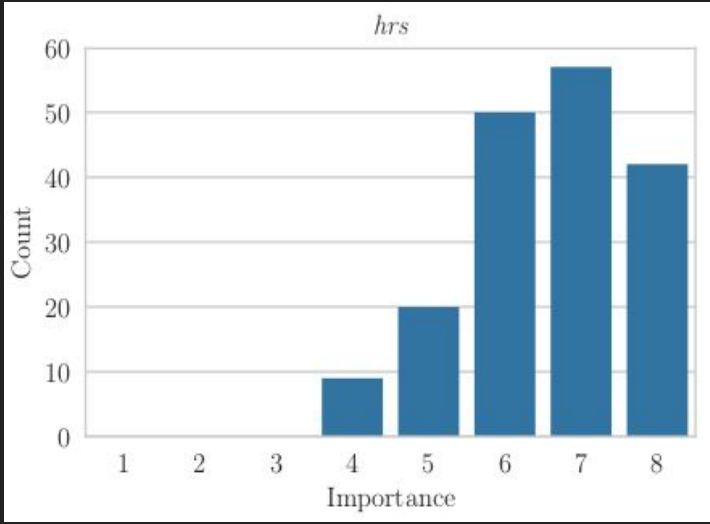
Type II plot using LIME Explanations

Different effects of "Hours spent for Math per week" for different classification

"<= 8.5" vs "> 8.5"
at-risk vs (moderate-risk + no-risk)



"<= 11.5" vs "> 11.5"
(at-risk + moderate-risk) vs no-risk



TAKEAWAYS

- Indiscrimination between different statistical modeling approaches in related literature
- Three important approaches :

Explanatory

Explanatory
+
Predictive

Predictive
+
Some method
to make predictions
interpretable

- For the current problem, all three approach gave many similar results showing that our understanding of the phenomenon is almost correct
- However, some subtle effects were uncovered only in predictive modeling
- Finally, we should not get settled at explanatory modeling,
  But at the same time, we should not get fascinated by predictions


ANY
QUESTIONS