AI: Strategy + Marketing (MGT 853)

The AI $\Longleftrightarrow$ Human Interface (Session 5)

Vineet Kumar

Yale School of Management
Spring 2024

Agenda for Today's Session

- Driving as an ML Problem

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Agenda for Today's Session

- Driving as an ML Problem
- Explainability, Interpretability and Transparency
- Research on Interpretable ML models
- Domain Knowledge
- Role of AI versus Humans

Autonomous Vehicles


HYSTER
80 HYSTER
XL2
24

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s)

Expert Systems
Second-wave used computer programming (1980s to early 2000s)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

If condition X, Then do Y

- Could go to 1000s or 100K lines of code
- Need to add code for each new condition and reprogram system

Autonomous Vehicles

Three Waves

First-wave used mechanical control (1970s)

Expert Systems
Second-wave used computer programming (1980s to early 2000s)

Prediction Models
Third-wave uses (2007 – current)

Mechanical Control

- Works in very limited way
- No flexibility if environment is changed even a bit

If condition X, Then do Y

- Could go to 1000s or 100K lines of code
- Need to add code for each new condition and reprogram system

Predictive Model

- AI system learns and builds the model and delivers better (more accurate prediction) as more data is generated

Converting to Prediction Problem (In class exercise)

Autonomous Vehicles

- Consider the role of prediction in autonomous driving
- Let's walk through the AI Decision Framework

Questions to Ponder

0) What sources of data should the system use?
1) What are the possible predictive problems one might encounter?
2) How should we measure performance?
3) What are appropriate ML algorithms in our toolbox to solve them?
4) What role does judgment play in this problem?

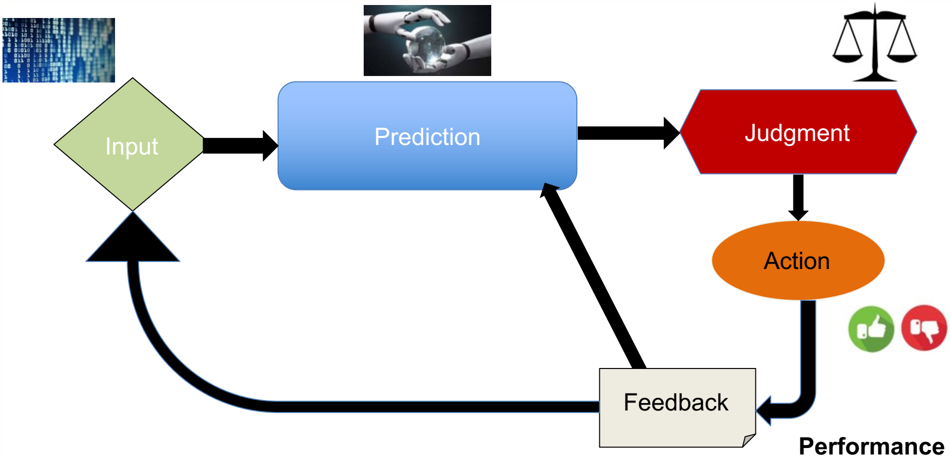
ML Pipeline

Where do humans interface?

ML Pipeline:

0. Identify Data (y, x)
1. Pre-processing
   Centering, Missing Data/Quality, Outliers, Normalization
2. Data Visualization
   Correlation, Distribution, Model
3. Data Splitting — Test
   Training, Validation
4. Feature Engineering
5. Model Selection
6. Hyperparameter Selection
   Human Analyst Decides
7. Learning (Training)
   $y = f(x)$
8. Validation
9. Interpretation
10. Testing ...

AI Decision-Making Framework



Does Feedback also inform Judgment?

Why not have a black box model?
What about Digit Classification?

Why worry about Black Box?

What if we get very high accuracy?

- 95

Why worry about Black Box?

What if we get very high accuracy?

- 95
- 99.x?

Why worry about Black Box?

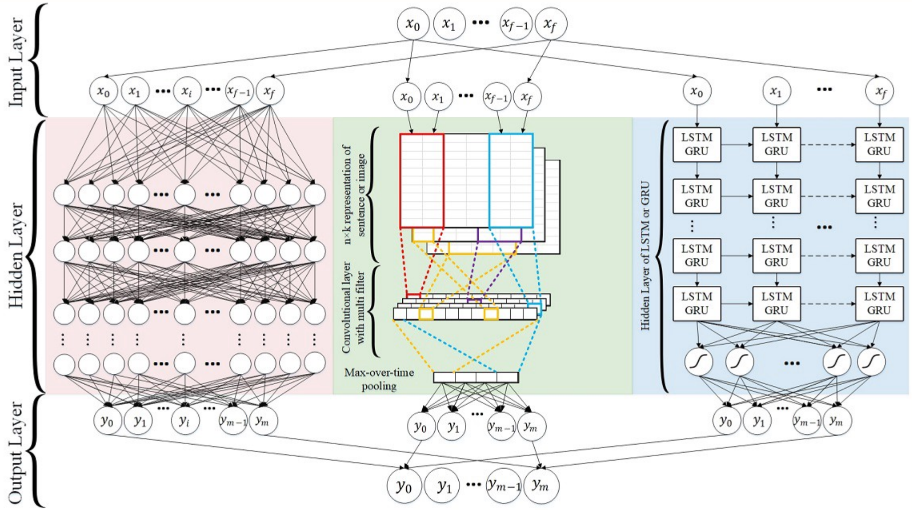
What if we get very high accuracy?

- 95
- 99.x?
- 100

Why worry about Black Box?

Wolf or Husky?

Can we understand this?



Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)
- Model produces constructs with meanings known to humans

Explainability and Interpretability

Explainability

- Capable of being understood: Plausible reasoning behind prediction
- Does not necessarily need model transparency
- Can be applied to a wide class of models (potentially all models)

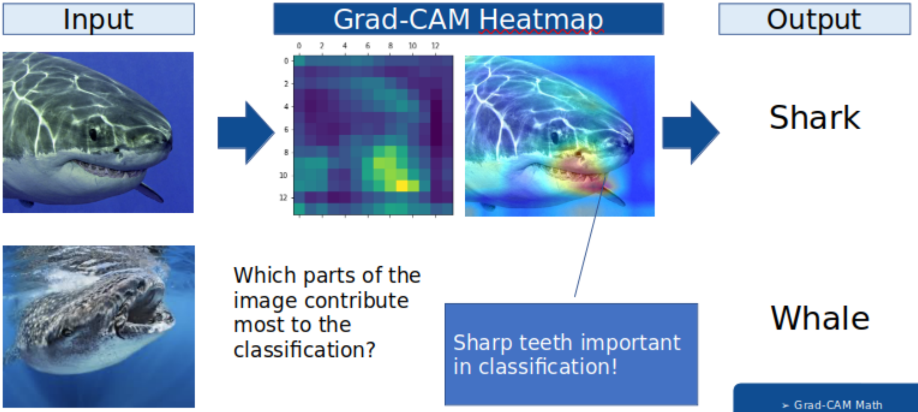
Interpretability

- Model's components are known
- Human can identify output for a specific input (with some effort)
- Model produces constructs with meanings known to humans
- May not be easy for all models

Explainability in Complex Models

Shark or Whale?

Explainability in Complex Models



Grad-CAM Math

Source: Grad-CAM Details (Technical)

Explainability $\Longrightarrow$ Interpretability

- Explainable $\neq$ Interpretable

Explainability $\Longrightarrow$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example

Explainability $\Longrightarrow$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

Explainability $\implies$ Interpretability

- Explainable $\neq$ Interpretable
- Linear Regression Example
- $\boxed{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$
- Specifying all βs fully specifies the model (with ε standard normal distribution)
- Person running the model has no hyperparameters or any other choices

nature › nature machine intelligence › perspectives › article

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

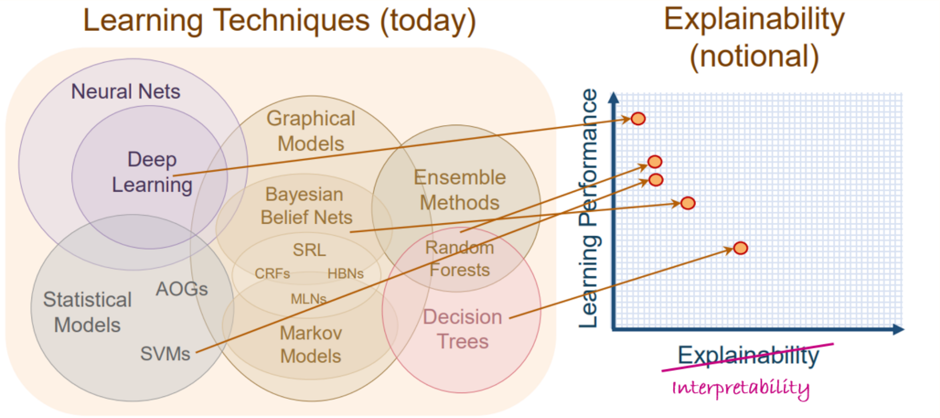
A preprint version of the article is available at arXiv.

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently

Research

Performance $\Longleftrightarrow$ Transparency Tradeoff?



Converting to Prediction Problems (In class exercise *if we have time*)

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started.
  $\geq$ 2 groups for each.

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started. $\geq$ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started.
  $\geq$ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?
- Suggest an (U)nsupervised, (S)upervised **and** (R)einforcement learning approach.

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started.
  ≥ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?
- Suggest an (U)nsupervised, (S)upervised **and** (R)einforcement learning approach.
- What variables you will use as y and X?

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started. ≥ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?
- Suggest an (U)nsupervised, (S)upervised **and** (R)einforcement learning approach.
- What variables you will use as y and X?
- Specify the actions (a), states (s), and reward (Π)

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started. ≥ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?
- Suggest an (U)nsupervised, (S)upervised **and** (R)einforcement learning approach.
- What variables you will use as y and X?
- Specify the actions (a), states (s), and reward (Π)
- What business metric would you use to measure improvement?

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Let's try this in Groups

- Choose one of (1), (2) or (3). Tell the class what you have chosen before you get started.
  $\geq$ 2 groups for each.
- What is the business problem? How do you convert it to a prediction problem?
- Suggest an (U)nsupervised, (S)upervised **and** (R)einforcement learning approach.
- What variables you will use as y and X?
- Specify the actions (a), states (s), and reward (Π)
- What business metric would you use to measure improvement?
- Role of Transparency, Interpretability and Explainability

3 Cases – Chose ONE

1) Social media (Instagram) – increase engagement
2) Content firm (Spotify) – recommend new content to its users
3) Apparel retailer – improve its product assortment

Big Picture Takeaways

- Interpretability is critically important to helping humans understand why AI makes the decisions it does

Big Picture Takeaways

- Interpretability is critically important to helping humans understand why AI makes the decisions it does
- Without that, we're guessing at how a "well performing" black box is doing its job

Big Picture Takeaways

- Interpretability is critically important to helping humans understand why AI makes the decisions it does
- Without that, we're guessing at how a "well performing" black box is doing its job
- Broadly, many applications of interpretability, e.g. with cars or watches, why are some products visually appealing?

Big Picture Takeaways

- Interpretability is critically important to helping humans understand why AI makes the decisions it does
- Without that, we're guessing at how a "well performing" black box is doing its job
- Broadly, many applications of interpretability, e.g. with cars or watches, why are some products visually appealing?
- What visual features contribute most to value?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!

Beyond Transparency

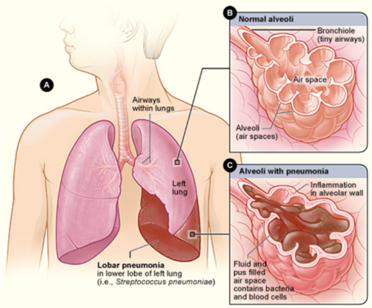
- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?

Beyond Transparency

- Researchers were comparing models for predicting likelihood of death for pneumonia patients in hospitals
- Rule based learning indicated that patients with **Asthma** were less likely to die
- Puzzling from a medical perspective!
- Why does this happen?



Research Presentation