Interpretability vs. Complexity

Supervised Learning

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

Contents of This Video

In this video, we will cover:

- What model interpretability means and why it matters
- The relationship between model complexity and interpretability
- Real-world examples comparing interpretable and black-box models
- When to prioritize interpretability vs. performance
- The spectrum of model interpretability options
- Practical tips for balancing interpretability and complexity

Learning Objectives

- Understand the concept of model interpretability
- Recognize the trade-off between interpretability and complexity
- Compare interpretable models vs. “black box” models
- Learn when to prioritize interpretability over performance

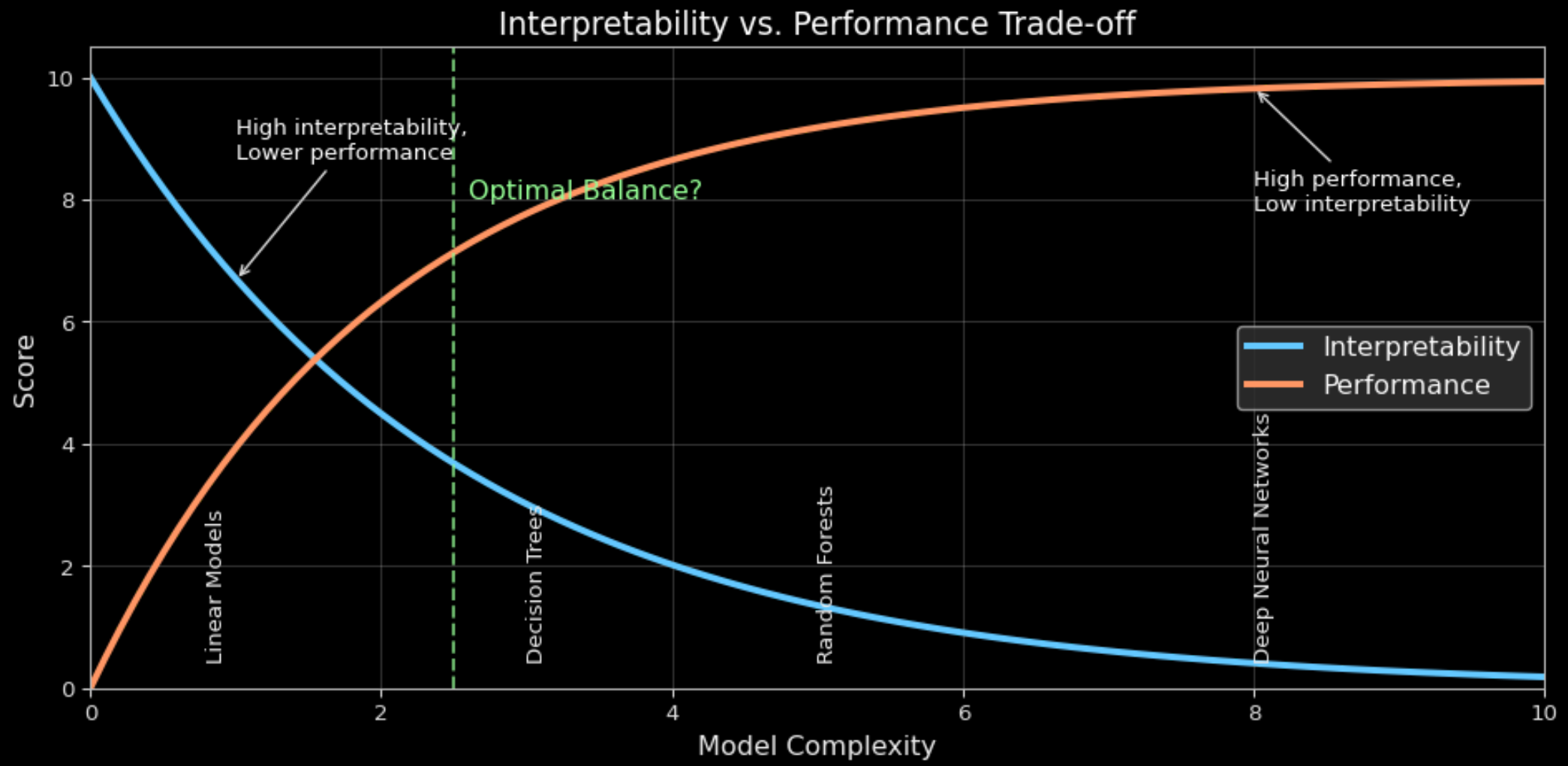
What is Interpretability?

- The ability to understand and explain **how** a model makes decisions
- Knowing the “why” behind a prediction
- Being able to trace the reasoning from input to output
- Understanding feature importance and relationships

What is Complexity?

- **Model complexity**: Number of parameters, non-linear relationships, etc.
- **Flexibility**: Ability to fit intricate patterns in the data
- More complex models can capture nuanced relationships
- But they often become harder to interpret

The Trade-off Visualized



Examples of Models Along the Spectrum

Model Type	Interpretability	Complexity/Flexibility
Linear Regression	High	Low
Logistic Regression	High	Low
Decision Trees	Medium-High	Medium
Random Forests	Medium	Medium-High
Support Vector Machines	Medium-Low	High
Neural Networks	Low	Very High

Example: Linear Regression vs. KNN

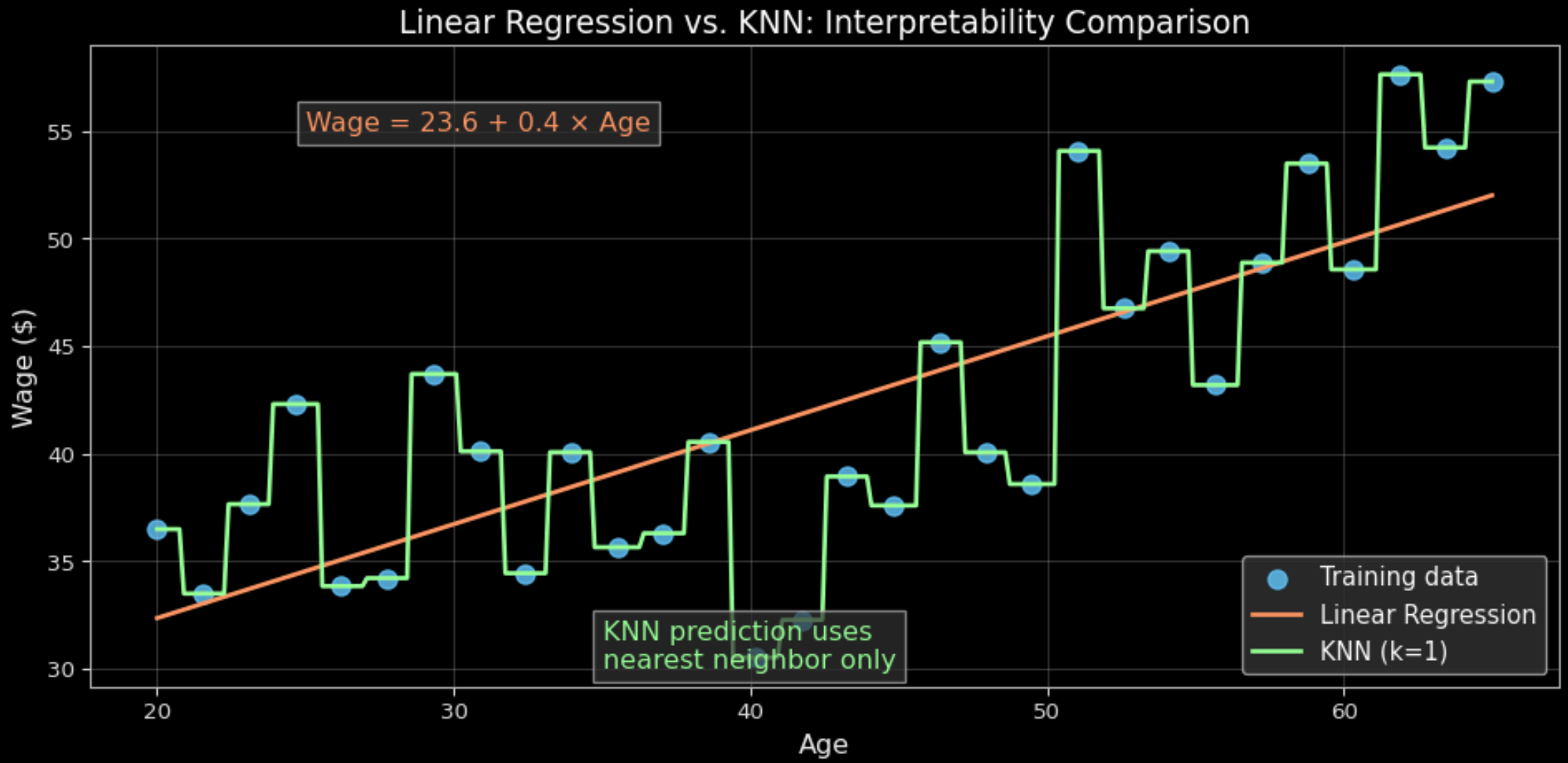
Linear Regression

- Equation: `Wage = 50 + 0.7*Age`
- Interpretation: “Each additional year of age is associated with a $0.7 increase in wage”
- Clear, global relationship

K-Nearest Neighbors (K=1)

- No equation, just memorized data points
- Interpretation: “Your prediction is based on the most similar case in our database”
- Local, instance-based reasoning

Visualizing the Linear vs. KNN Contrast



Why Does Interpretability Matter?

- **Trust and Transparency**: Stakeholders need to understand why decisions are made
- **Debugging**: Easier to identify and fix problems
- **Regulatory Compliance**: Some fields require explainable decisions
- **Knowledge Discovery**: Learning about the domain from the model itself
- **Ethical Considerations**: Ensuring fairness and avoiding bias

When to Prioritize Interpretability

- **High-stakes decisions**: Medical diagnosis, criminal justice
- **Regulatory requirements**: Financial services, healthcare
- **Need for insights**: Understanding factors driving outcomes
- **Building trust**: New systems being introduced to stakeholders
- **Detecting bias**: Ensuring fairness and equity

When to Prioritize Performance/Complexity

- **Low-stakes predictions**: Product recommendations, weather forecasting
- **Accuracy is paramount**: Competitive scenarios where small improvements matter
- **Underlying relationship is highly complex**: Natural language, images
- **Large amounts of data**: Taking advantage of patterns human can't see
- **Post-hoc explanations available**: When global interpretability isn't required

The Spectrum of Interpretability

- **Glass Box Models**: Linear/logistic regression, small decision trees
- **Grey Box Models**: Random forests, gradient boosting, shallow neural networks
- **Black Box Models**: Deep neural networks, complex ensembles
- **Post-hoc Explanation Tools**: LIME, SHAP values, partial dependence plots

Real-World Scenario: Loan Approval

Simple Decision Tree

```
if income > $50k AND
   credit_score > 700:
   approve
else if credit_score < 600:
   deny
else:
   review manually
```

Complex Ensemble Model

- Uses dozens of factors
- Non-linear interactions
- 5% more accurate in predicting defaults
- Cannot easily explain individual decisions

Case Study: Medical Diagnosis vs. Image Recognition

Medical Diagnosis System

- Needs to explain why a patient is high-risk
- Doctors must understand the reasoning
- Interpretability often valued over small accuracy gains

Image Recognition System

- Tagging photos in a social media app
- Low-stakes decisions
- Performance and user experience prioritized
- “Black box” acceptable

Practical Tips for Balancing the Trade-off

- Start simple, add complexity incrementally
- Use domain knowledge to create interpretable features
- Consider using interpretable models for exploration
- Use complex models for prediction when stakes are low
- Hybrid approach: interpretable model + black-box “advisor”
- Employ post-hoc explanation techniques for complex models

Quiz

Which statement about interpretability and complexity is MOST accurate?

A. Simpler models are always better because they’re more interpretable
B. Complex models are always better because they’re more accurate
C. The appropriate balance depends on the specific application context
D. Interpretability and complexity can always be maximized simultaneously

What We've Covered

In this video, we've discussed:

- The concept of model interpretability and its importance
- The fundamental trade-off between interpretability and complexity
- Examples comparing linear models to more complex alternatives
- Scenarios where interpretability should be prioritized
- Situations where performance may be more important than interpretability
- Practical strategies for balancing both considerations