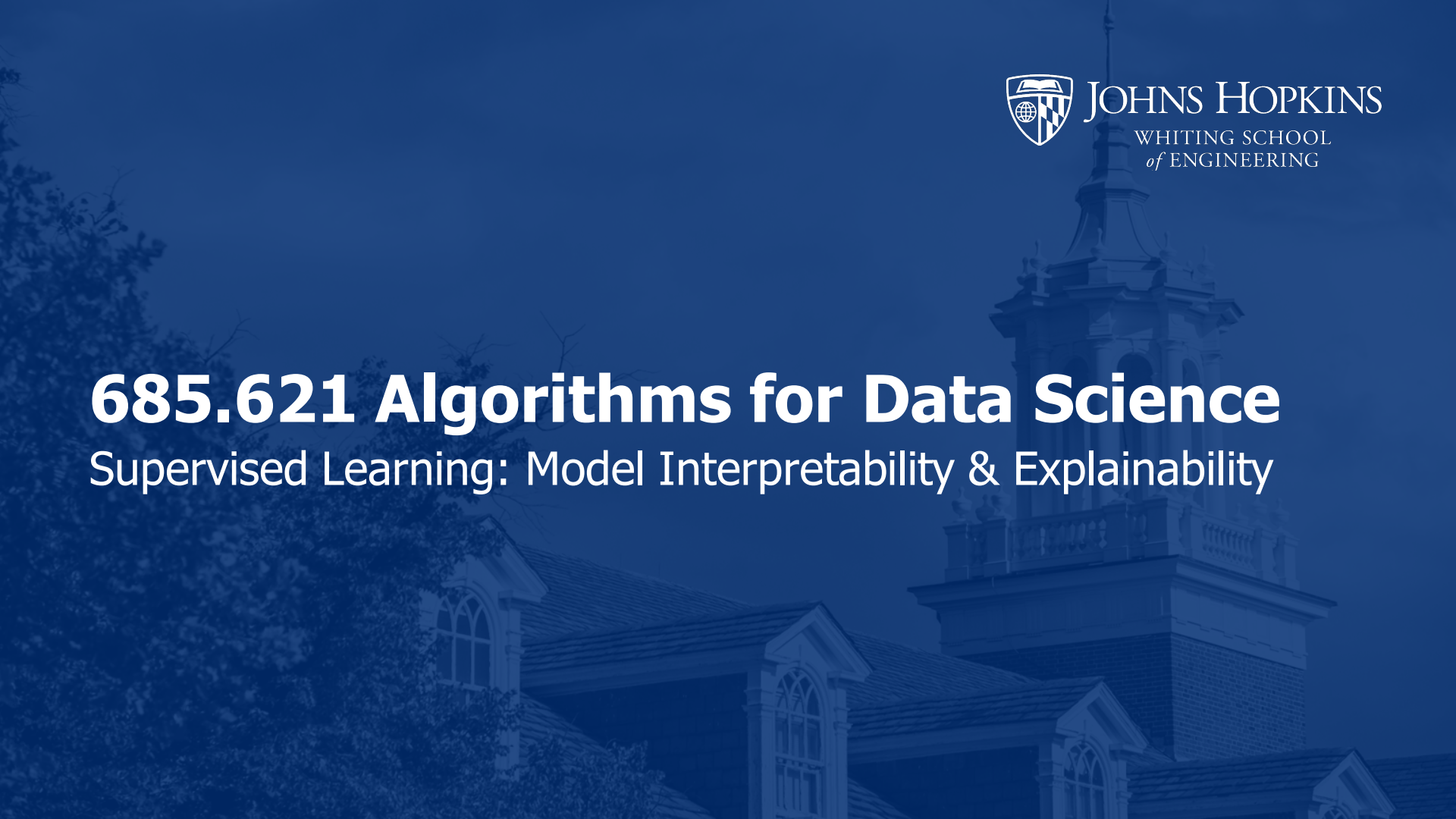


685.621 Algorithms for Data Science

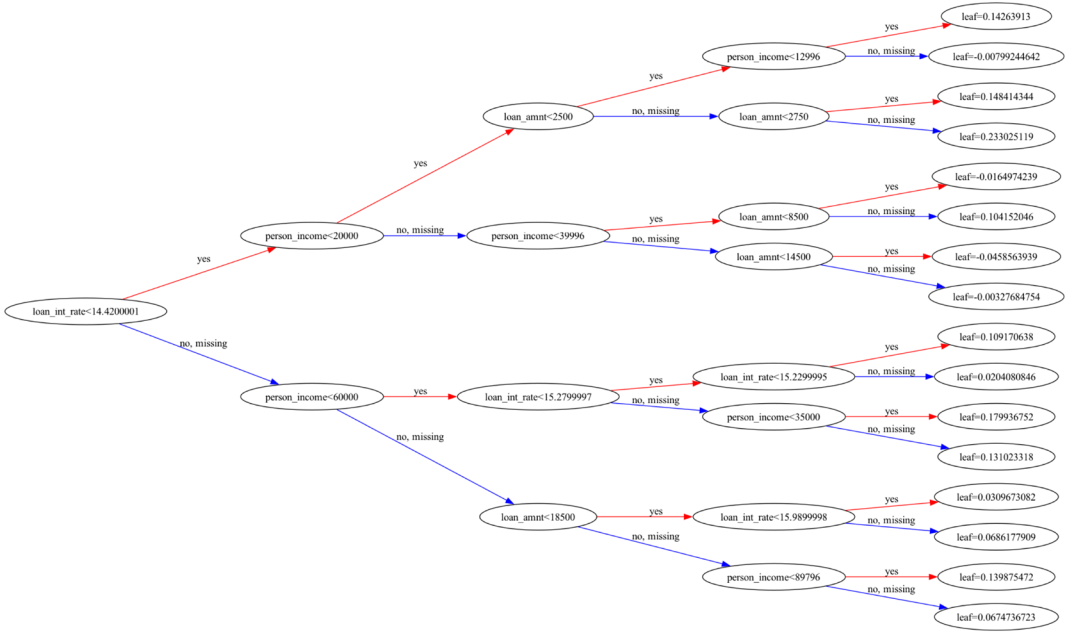
Supervised Learning: Model Interpretability & Explainability

The Importance of Model Interpretability

- **Trust & Transparency**: Users need to understand why a model makes certain predictions.
- **Regulatory Compliance**: Industries like healthcare and finance require explainable AI.
- **Debugging & Improvement**: Interpretability helps identify model biases and errors.
- **Human-in-the-Loop Decision Making**: Decision-makers rely on interpretable outputs.

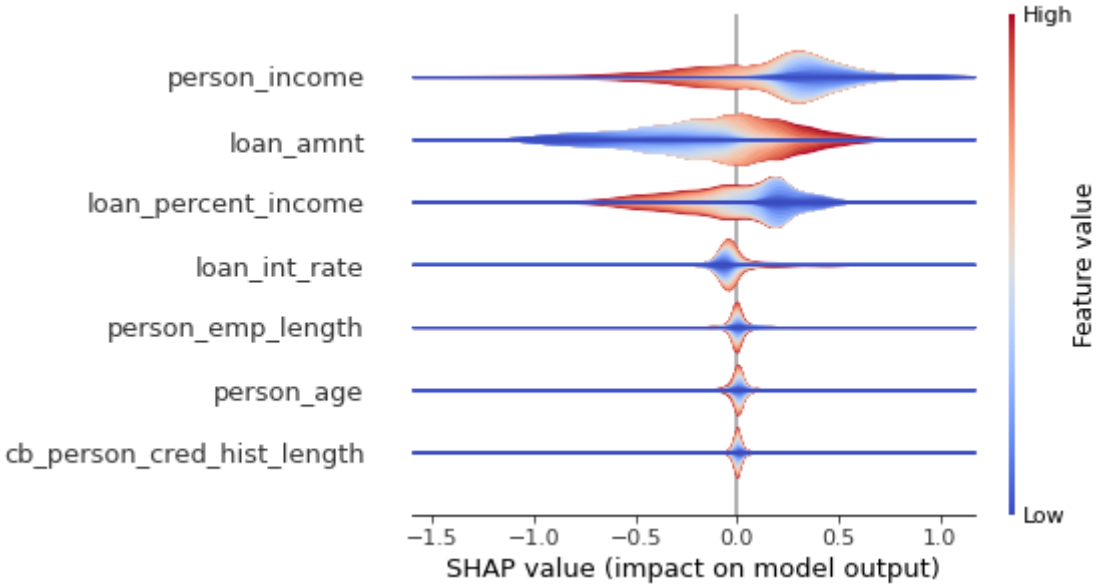
How Do We Interpret Models?

- **Intrinsic Interpretability:**
  - Simple models like Decision Trees & Logistic Regression.
  - Built-in transparency—directly explainable.
- **Post-Hoc Interpretability:**
  - Needed for complex models like Neural Networks & Random Forests.
  - Uses techniques like Feature Importance, SHAP, and LIME.

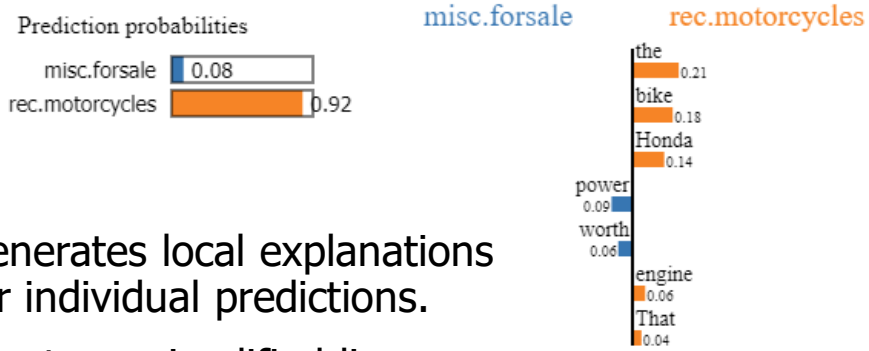


Explaining Predictions Using SHAP Values

- Provides global and local feature importance scores.
- Helps interpret black-box models like Random Forest and Neural Networks.
- The summary plot shows which features impact predictions the most.



LIME: Explaining Individual Predictions



- Generates local explanations for individual predictions.
- Creates a simplified linear approximation of the complex model.
- The LIME explanation shows how each feature contributes to a specific prediction.