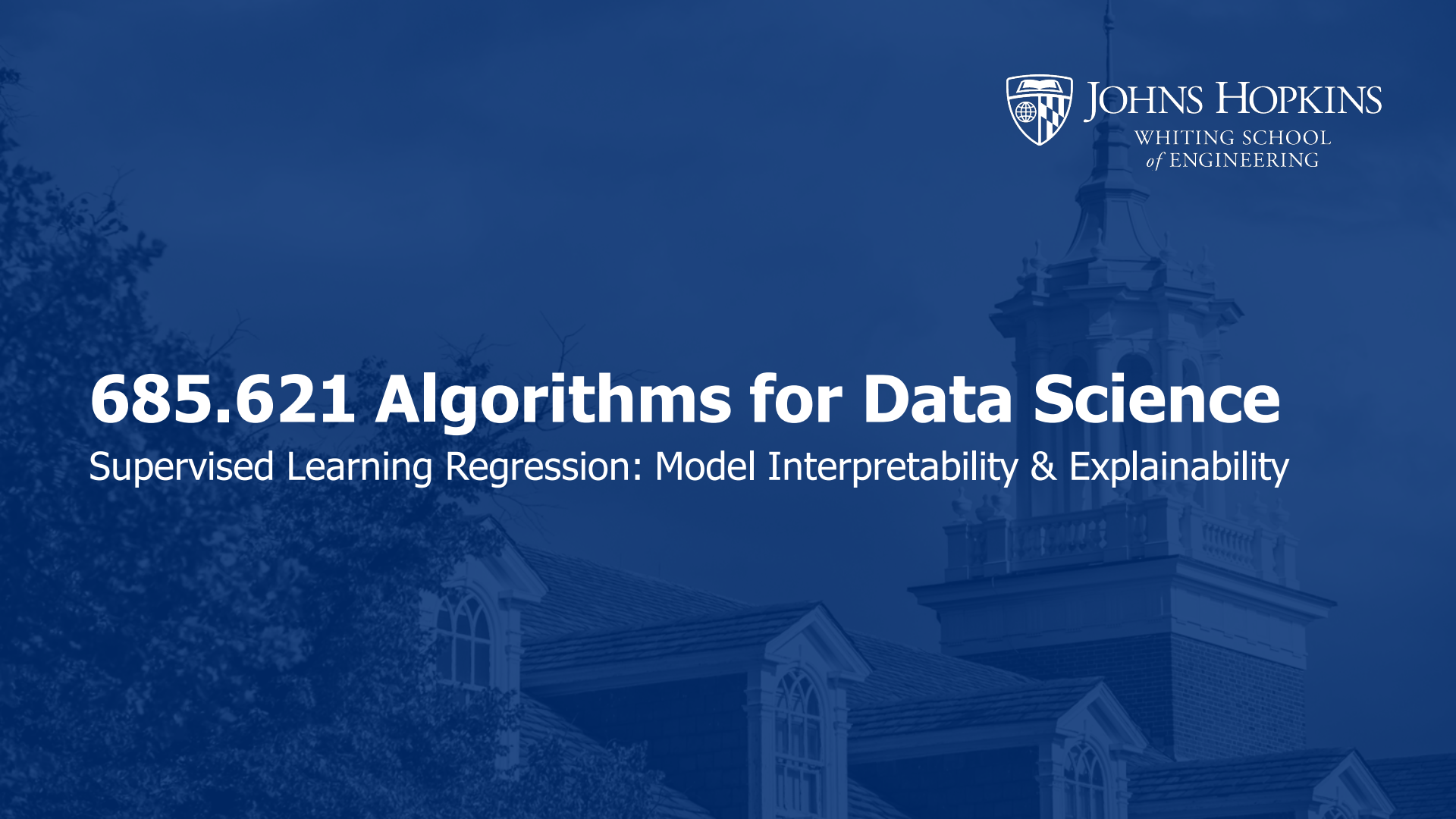
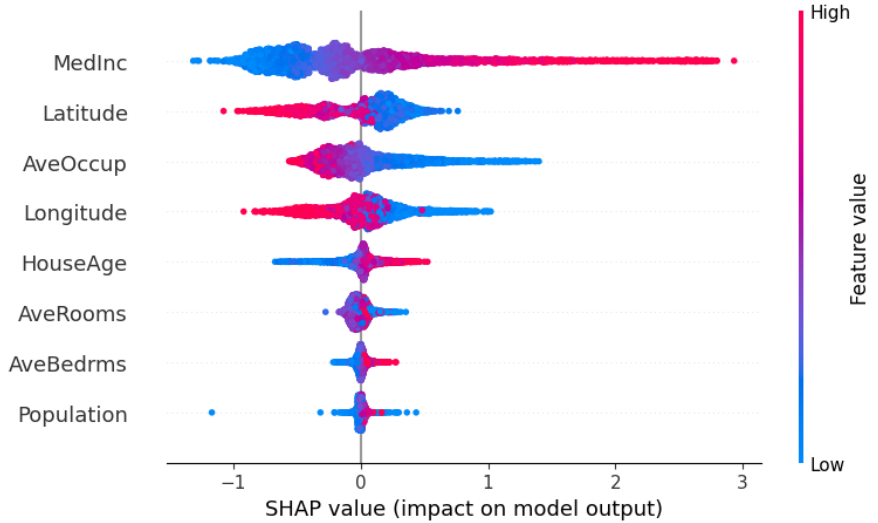


685.621 Algorithms for Data Science

Supervised Learning Regression: Model Interpretability & Explainability

Why Interpret Regression Models?

- **Understand** how input features influence predictions
- **Build trust** and transparency in model decisions
- **Diagnose model behavior** and uncover biases
- **Communicate results** to non-technical stakeholders.



Coefficients vs Feature Importance

Coefficients

- Each coefficient βj represents the **change in prediction per unit change in feature** xj, holding others constant
- Sign and magnitude indicate **direction and strength of influence**
- **Watch out for multicollinearity**: interpretation breaks down when features are correlated

Feature Importance

- Decision Trees and Random Forests assign **importance scores** based on **split contributions** to reducing error
- Useful for identifying **most influential features**
- **Less interpretable** in structure, but **intuitive at feature level**

Model-Agnostic Methods (SHAP & LIME)

- **LIME**: Locally approximates model with an interpretable surrogate model
- **SHAP**: Uses game theory to assign contributions of each feature to individual predictions
- Works for any **black-box model** (SVR, Random Forest, Ensembles)