# Shapley Additive Explanations

Research Seminar

Mentor: dr. Branko Kavšek
Student: Nemanja Cvetić

# Why this matters?

Machine learning models can be very accurate but completely unreadable.

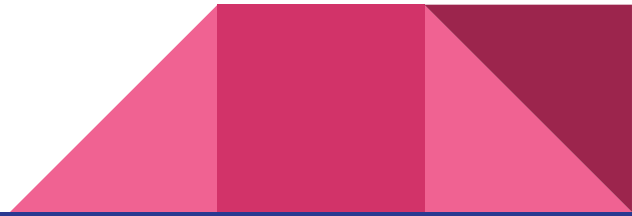
A Random Forest with 200 trees makes thousands of internal decisions, but nobody can read it.

if a bank model rejects a loan, regulators in Europe (GDPR) legally require an explanation.

SHAP and LIME were invented to solve this.

# How the research is structured

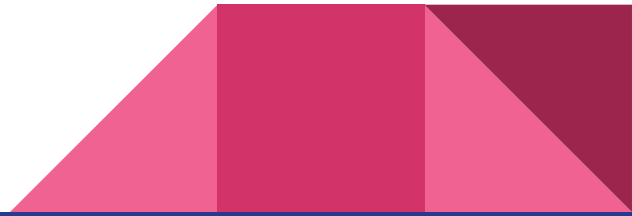
- Two models of different complexity: **Linear Regression** (simple) and **Random Forest** (complex)
- Three explanation methods: **LinearSHAP, TreeSHAP, and LIME**
- One dataset: California Housing - 20,640 rows, 8 features, predict price
- Three hypotheses to test, each with a clear confirmation condition



# What the dataset actually is

The 8 features:

- MedInc
- HouseAge
- AveRooms/AveBdrms
- Population/AveOccup
- Latitude/Longitude



# The Models

- **Linear Regression**
  - Linear Regression draws the best straight-line relationship between each feature and the target.
- **Random forest**
  - A Random Forest is an ensemble of 200 decision trees. Each tree asks a sequence of yes/no questions about the features and ends in a prediction.

# SHAP

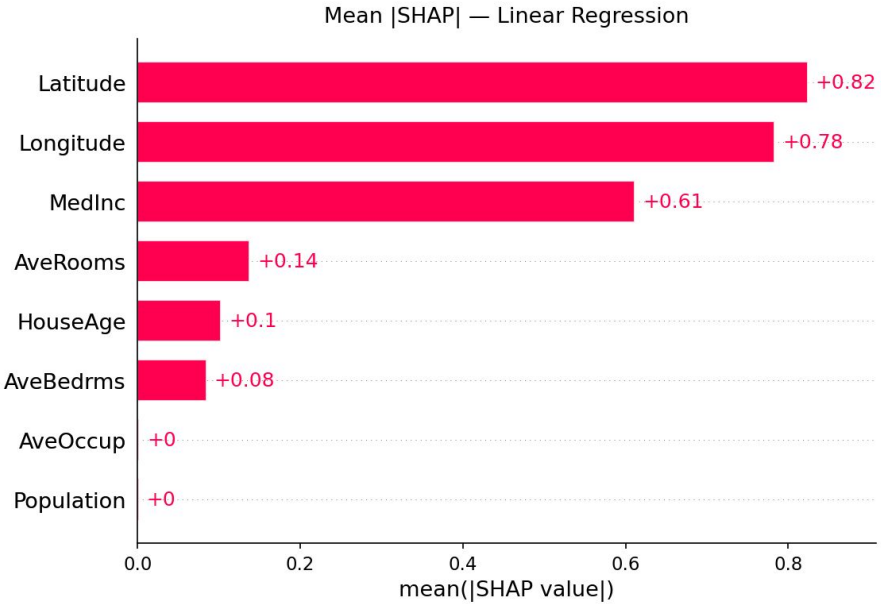
- SHAP is based on Shapley values, invented by economist Lloyd Shapley in 1953.
- He was solving the "team credit problem".
- His answer: consider every possible order in which workers could join the team.
- SHAP (Lundberg & Lee, 2017) translates this directly to machine learning: features are the workers, the model prediction is the team's output.

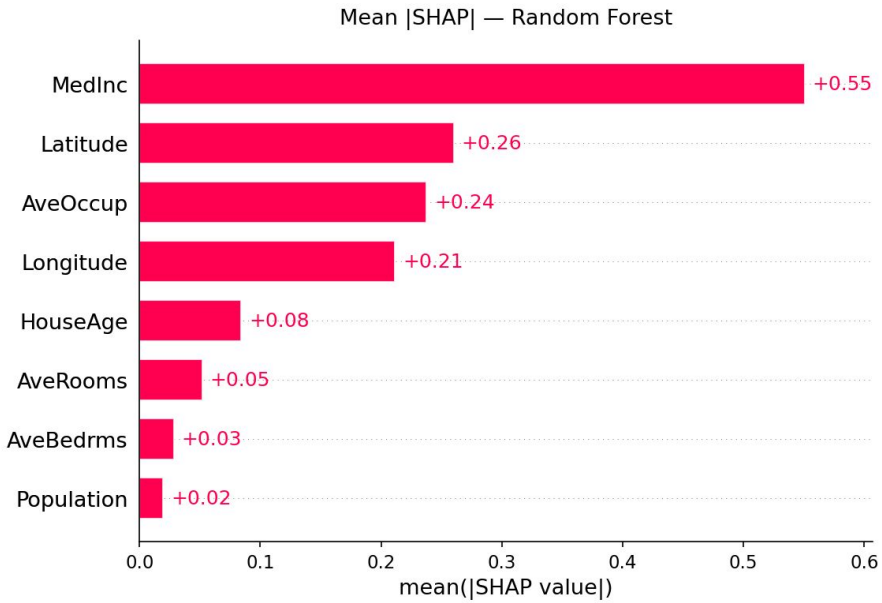
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS 2017), 30, pp. 4765–4774.

# Results

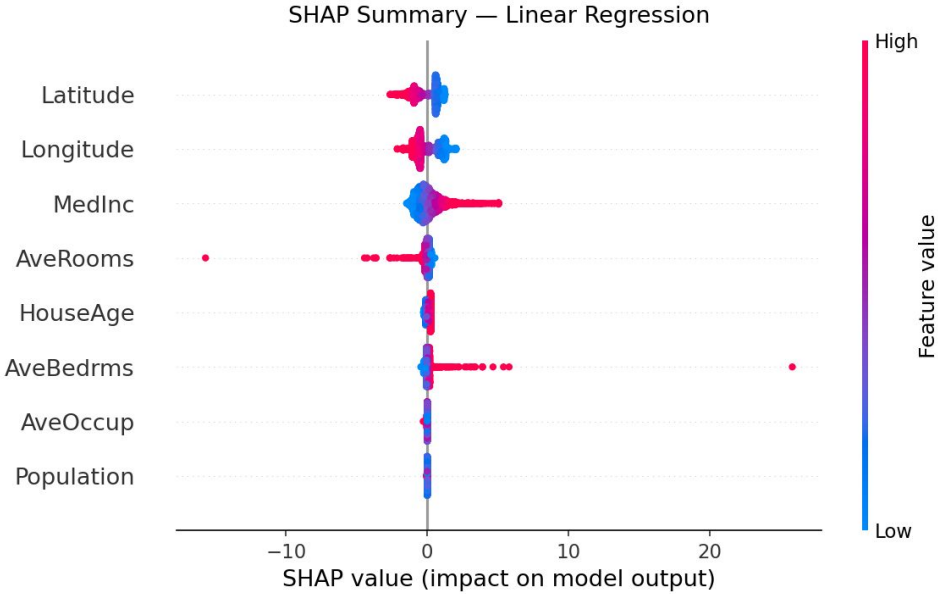
- **H1** - Income dominance: **CONFIRMED**
  - MedInc (median household income) ranked #1 in both the TreeSHAP and LIME attributions for the Random Forest
- **H2** - Geographic clustering: **CONFIRMED**
  - The SHAP dependence plots for Latitude and Longitude show clear non-uniform spatial structure.
- **H3** - Non-linearity matters: **CONFIRMED**
  - The Random Forest improved $R^2$ by 40% over Linear Regression (0.8062 vs 0.5758), proving non-linear patterns exist in the data

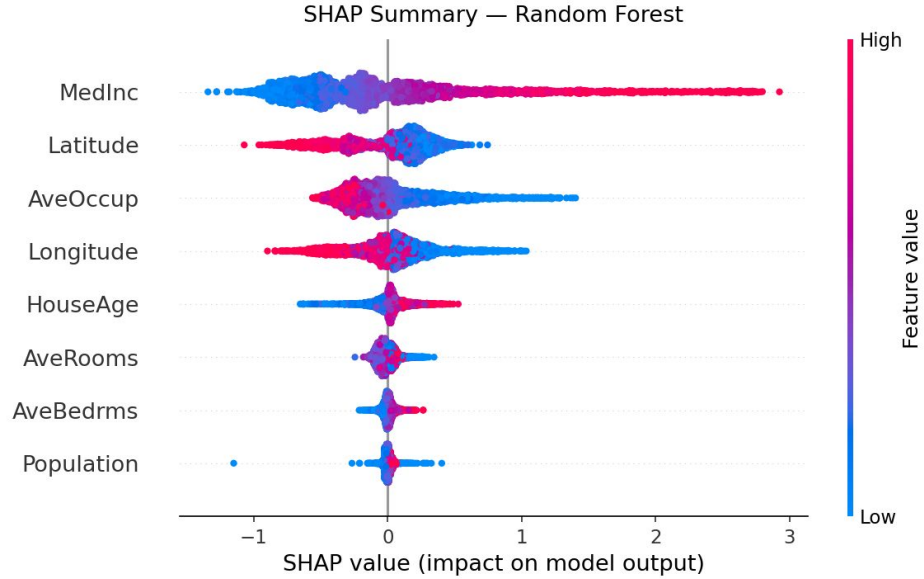
# SHAP Linear regression vs random forest



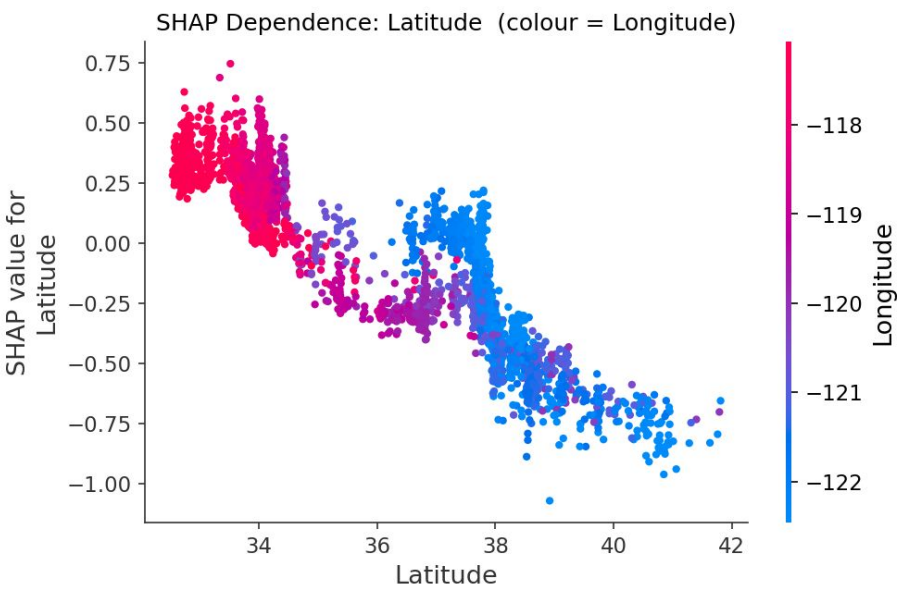



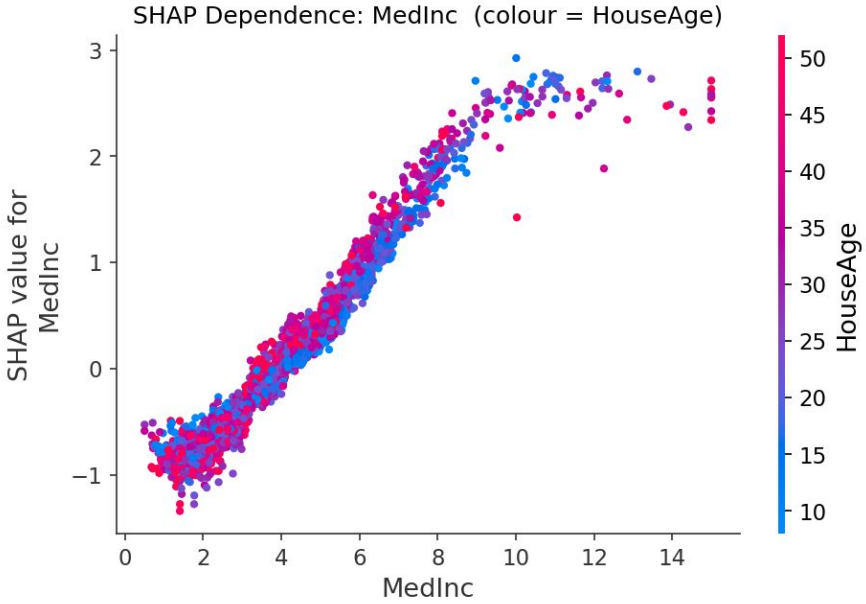
# SHAP Linear regression vs random forest (beeswarm)



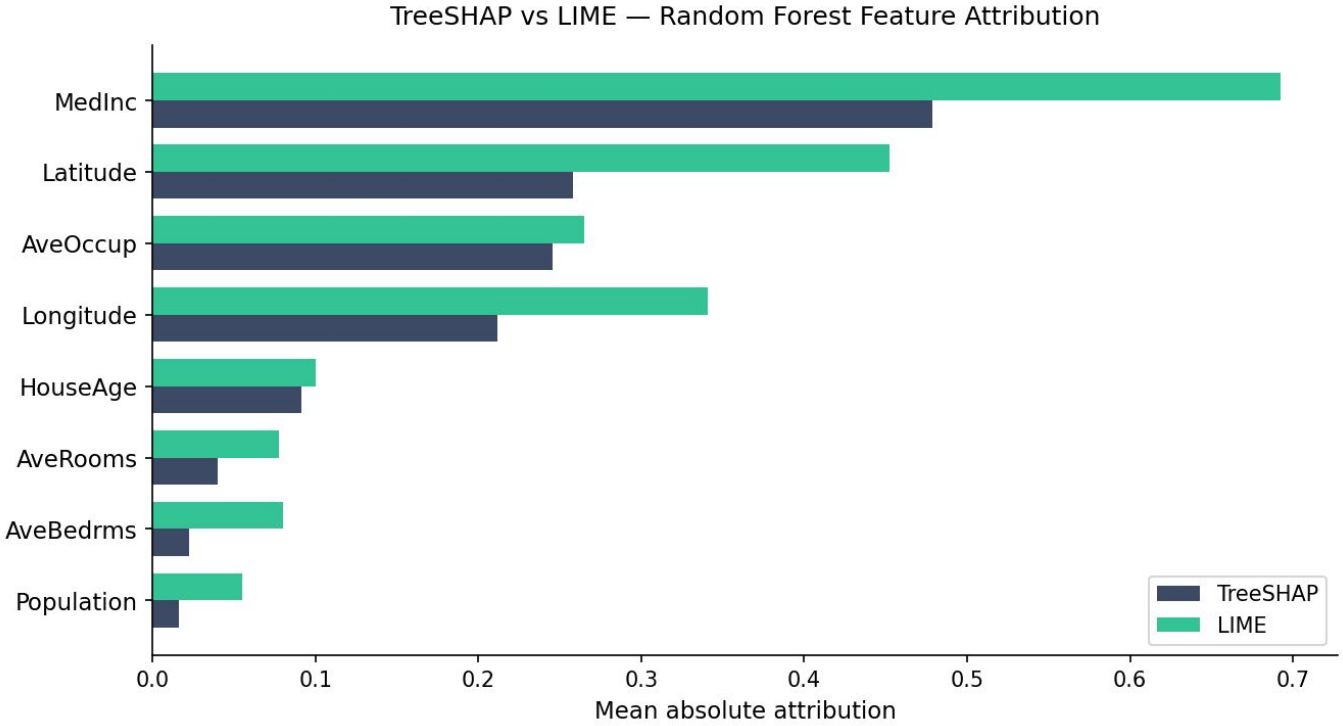



# SHAP Latitude and MedInc dependence



# SHAP vs LIME random forest



# Thank you for the attention!

# References


Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the Theory of Games II (pp. 307–317). Princeton University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS 2017), 30, pp. 4765–4774. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. Journal of the Operational Research Society, 73(1), 91–101. https://doi.org/10.1080/01605682.2020.1865846