Introduction to Machine Learning Methods for Prediction and Causal Inference

Prediction vs Causal Inference

Prediction: Identifies patterns to forecast future outcomes based on existing data.

Causal Inference: Explains how changing one variable causes changes in another.

Why We Lean Towards Prediction

- Causality is hard to prove, requiring strong evidence and more data.
- Prediction *with explainability* can identify those in need and help us understand why.
- Ex: Which features/markers correlate to homelessness?
    - Big factors include housing costs, mental and physical health issues, receipt of social safety net benefits (e.g. CalFresh)
    - Machine learning can forecast who is at the greatest risk of becoming homeless, allowing for targeted interventions

Model Selection

XGBoost (or eXtreme Gradient Boosting) is a state-of-the-art machine learning algorithm that is well-suited for prediction tasks because of its efficiency and high accuracy.

This model was utilized by **Cal Policy Lab** to predict homelessness among single adults receiving mainstream County services within Los Angeles County

Performance

Often outperforms other algorithms like Logistic Regression, Decision Trees, and Support Vector Machines (SVM) in accuracy

Scalability

Handles large datasets and complex feature interactions

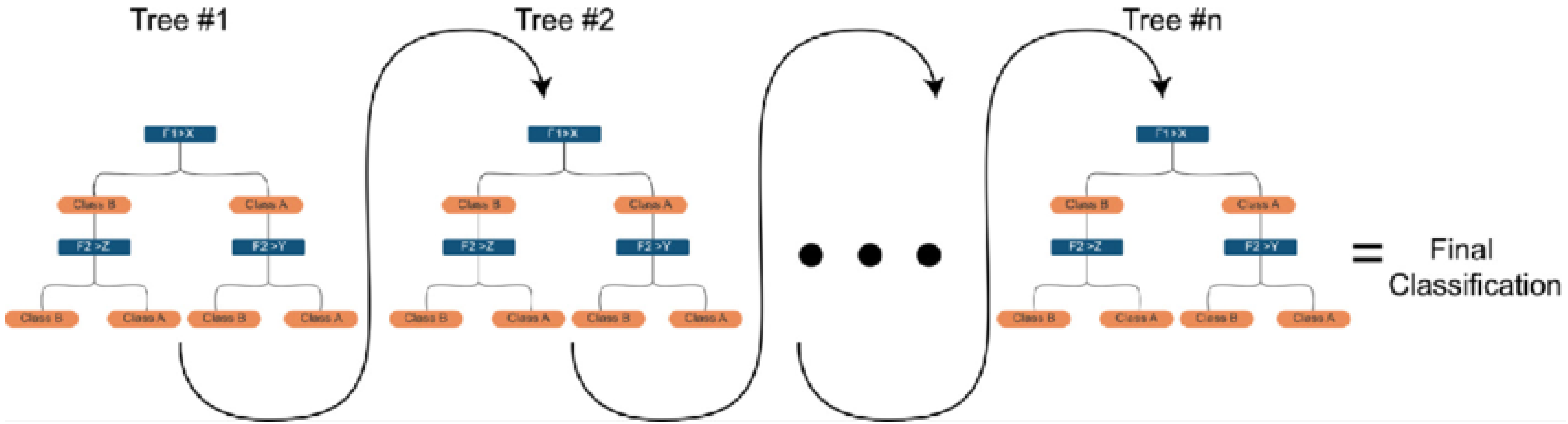
Efficiency

Parallel processing and optimization techniques significantly reduce training time

How does XGBoost work?

Gradient Boosting

- Combines the power of many simple models to create a more accurate and robust predictive model
- Each new model corrects errors made by the previous ones, allowing the model to “boost” its performance over time.



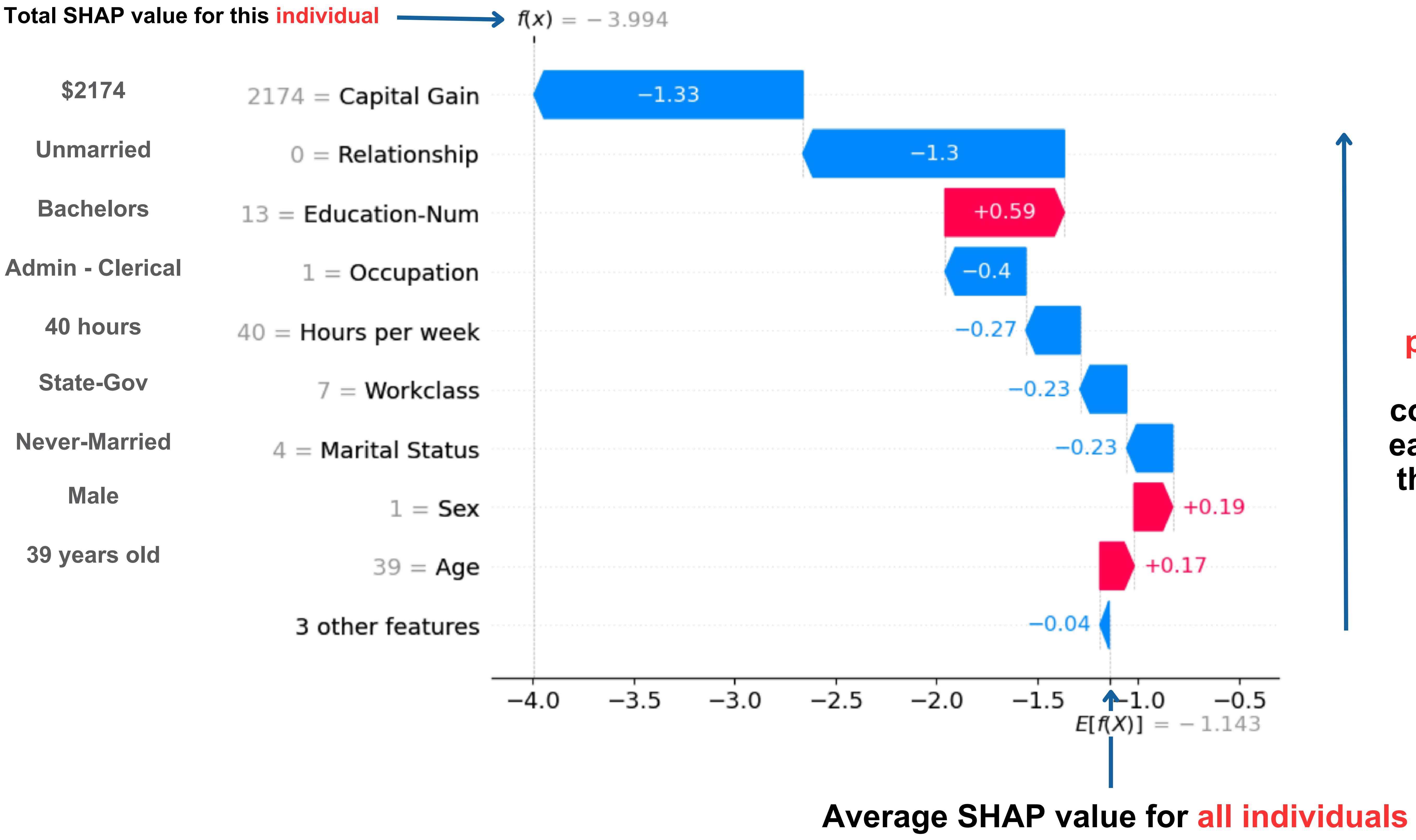
Demo

Census Income Dataset

Goal: Predict whether annual income of an individual exceeds $50K/yr

	Age	Workclass	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Over_50K
0	39.0	State-gov	13.0	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	False
1	50.0	Self-emp-not-inc	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	False
2	38.0	Private	9.0	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	False
3	53.0	Private	7.0	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States	False
4	28.0	Private	13.0	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba	False
5	37.0	Private	14.0	Married-civ-spouse	Exec-managerial	Wife	White	Female	0.0	0.0	40.0	United-States	False
6	49.0	Private	5.0	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0.0	0.0	16.0	Jamaica	False
7	52.0	Self-emp-not-inc	9.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	45.0	United-States	True
8	31.0	Private	14.0	Never-married	Prof-specialty	Not-in-family	White	Female	14084.0	0.0	50.0	United-States	True
9	42.0	Private	13.0	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178.0	0.0	40.0	United-States	True

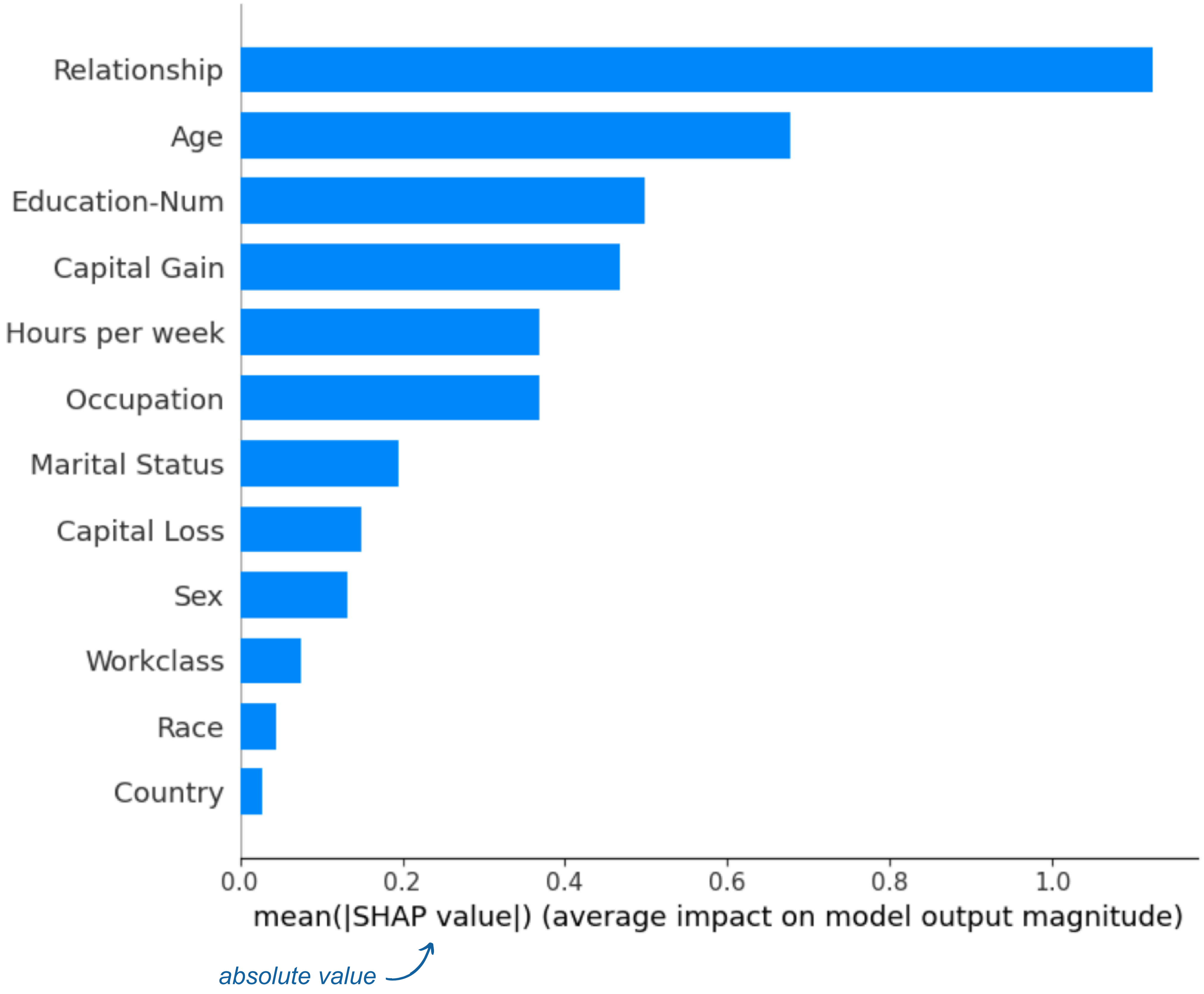
SHAP Waterfall Plot



Each row shows the positive and negative contribution of each feature to the prediction

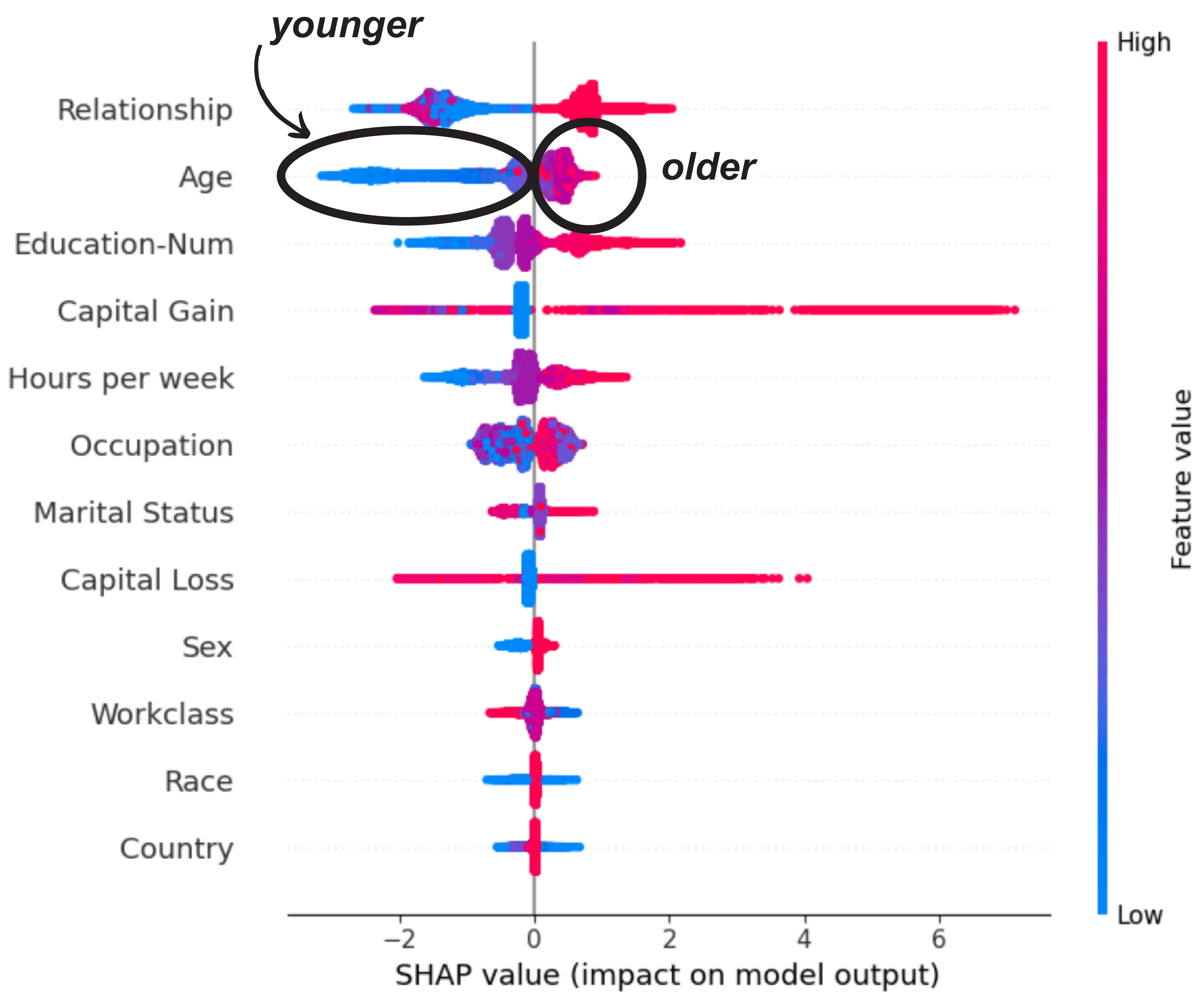
Mean Importance Bar Chart

Global feature importance plot

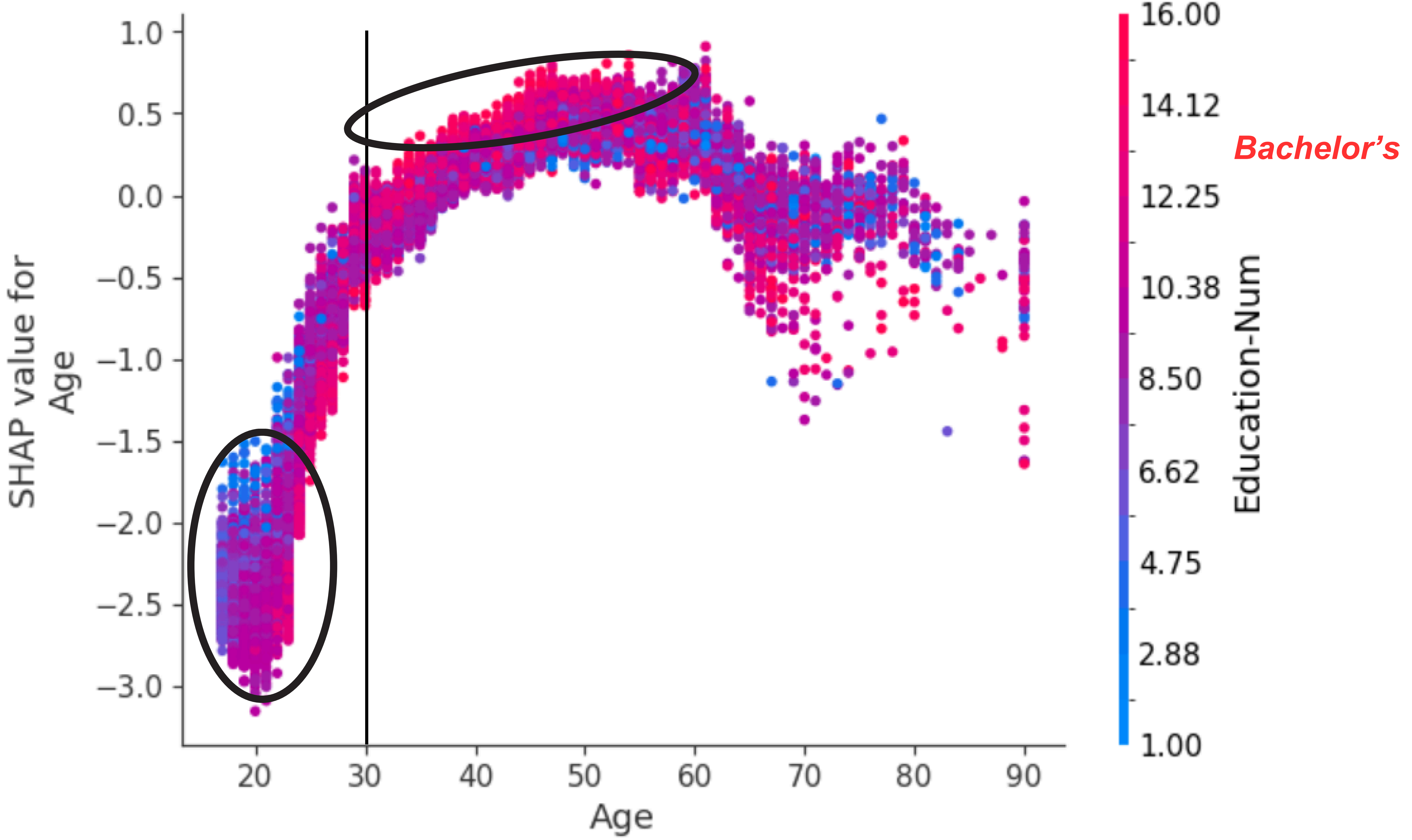


SHAP Summary Plot

Distribution + Variability



SHAP Dependence Plots



Conclusion and Next Steps

Conclusion

- The journey with XGBoost and its application to the homelessness predictive model is just beginning. While we've made significant progress, there is still much to discover when we interpret the data and when we understand the broader implications of our findings.

Next Steps

- Data discovery
- Results coming soon (to theaters near you!)

Sources

Abualdenien, J. (n.d.). Ensemble-learning approach for the classification of Levels Of Geometry (LOG) of building elements. Retrieved from https://www.researchgate.net/figure/eXtreme-Gradient-Boosting-XGBoost-Schematic-Representation-it-builds-decision-trees_fig2_357741497

Brownlee, J. (2016, August 30). Feature Importance and Feature Selection With XGBoost in Python. Retrieved from Machine Learning Mastery website: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

Lundberg, S. (2018). Be careful when interpreting predictive models in search of causal insights — SHAP latest documentation. Retrieved from Readthedocs.io website: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html

Lundberg, S. (2020, October 6). Interpretable Machine Learning with XGBoost. Retrieved from Medium website: https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Płoński, P. (2020, August 17). Xgboost Feature Importance Computed in 3 Ways with Python. Retrieved from MLJAR Automated Machine Learning website: https://mljar.com/blog/feature-importance-xgboost/