Final Project

Blake Campbell

The best results I have gotten are:

Linear Regression:

R^2 = 3.5%

RMSE = 10,000

Decision tree:

R^2 = -91%

RMSE = 14,000

Neural net:

R^2 = -7.8%

RMSE = 11,000

K nearest neighbors:

R^2 = -15.5%

RMSE = 11,000

Random forest:

R^2 = -10.6%

RMSE = 12,000

SVM:

R^2 = -1.1%

RMSE = 10,000

The first source has an average random forest RMSE of 1000 so my error was pretty far off.

The second source used an SVD (singular value decomposition) model and got an R^2 score of 27% and 41%, so again my models accuracy was off significantly.
The last source had an R^2 score of 97%, 96%, 98%, 94%, and 98% for linear, tree, random forest, SVM, and K-NN regression models, so my scores were drastically lower than these.

If I would have to do this project over again, I would make it a classifier instead of regression where it just returns their income bracket instead of a number. I feel this would help the accuracy of the model improve significantly.

Sources:

Kukk, M., Meriküll, J., & Rõõm, T. (2022). THE GENDER WEALTH GAP IN EUROPE: APPLICATION OF MACHINE LEARNING TO PREDICT INDIVIDUAL-LEVEL WEALTH. *Review of Income and Wealth*. https://doi.org/10.1111/roiw.12596

Matz, S. C., Menges, J. I., Stillwell, D. J., & Schwartz, H. A. (2019). Predicting individual-level income from Facebook profiles. *PLOS ONE*, *14*(3), e0214369. https://doi.org/10.1371/journal.pone.0214369

Satapathy, S. K., Saravanan, S., Mishra, S., & Mohanty, S. N. (2023). A Comparative Analysis of Multidimensional COVID-19 Poverty Determinants: An Observational Machine Learning Approach. *New Generation Computing*. https://doi.org/10.1007/s00354-023-00203-8