Income Classification in the United States

by Nick Horton

Background

IPUMS USA

- Originated in 1790
- Constitution requires its collection every 10 years
- In 2000, collection became annual

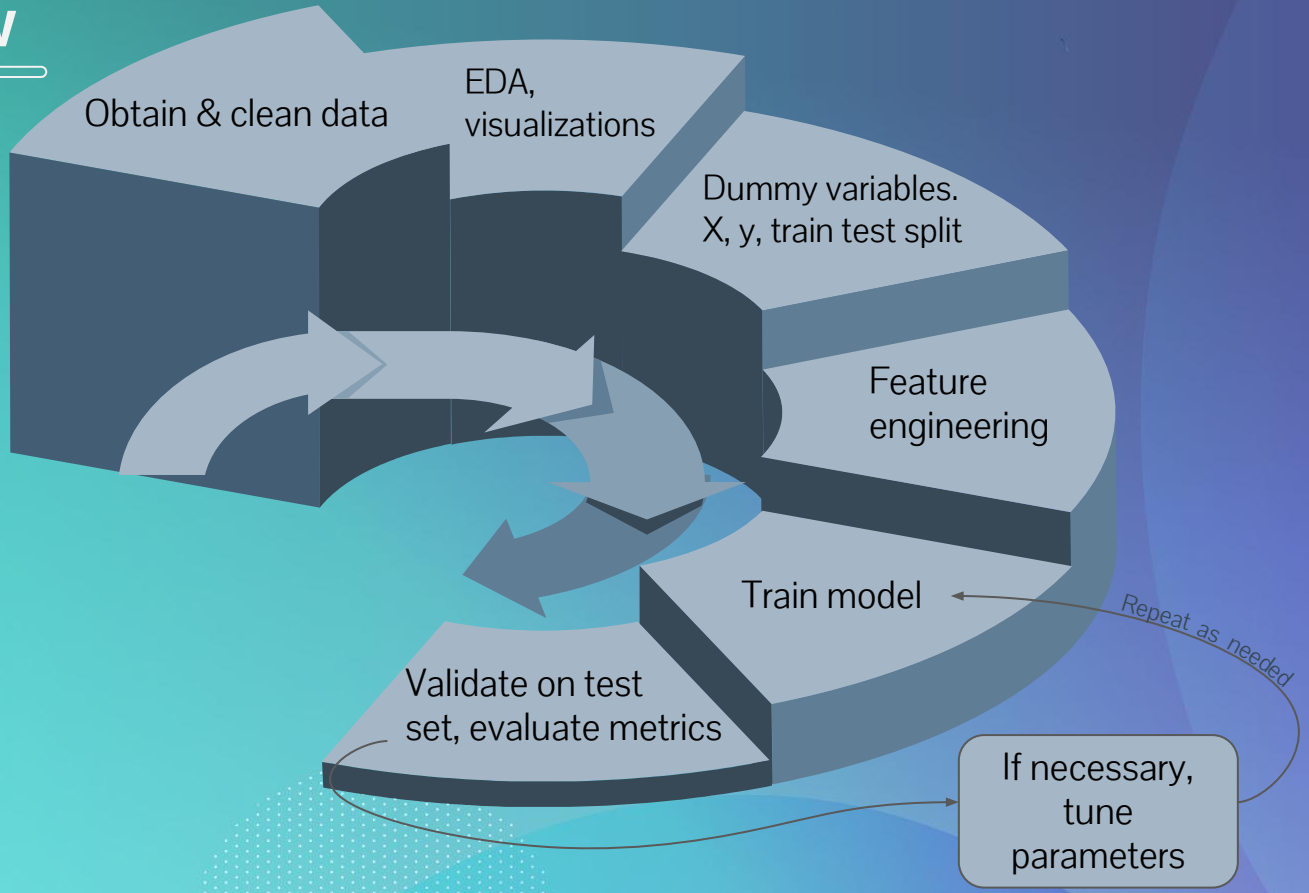
Collects a wealth of demographic & economic data

Can we use machine learning to predict whether someone's income is above or below $50,000/yr based on their answers to census questions?

> $50K/yr

< $50K/yr

Modeling Workflow



Best Results:

Random Forest

Accuracy

Train: 97.55%

Test: 81.40%

Precision: 72.72%

Recall: 67.17%

Honorable Mentions:

Extra Trees Classifier

Accuracy Train: **97.55**% Test: **80.64**%

Precision: 71.25%

Recall: 66.41%

K Nearest Neighbor

Accuracy Train: **96.70**% Test: **74.61**%

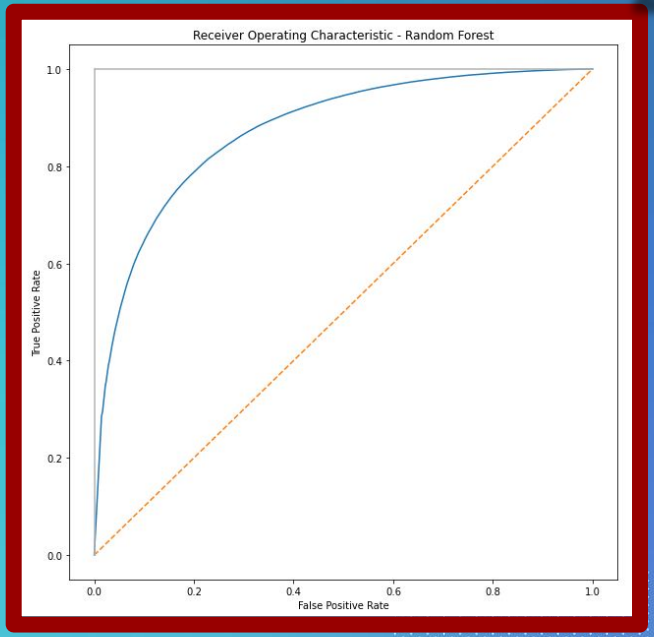
Precision: 84.93%

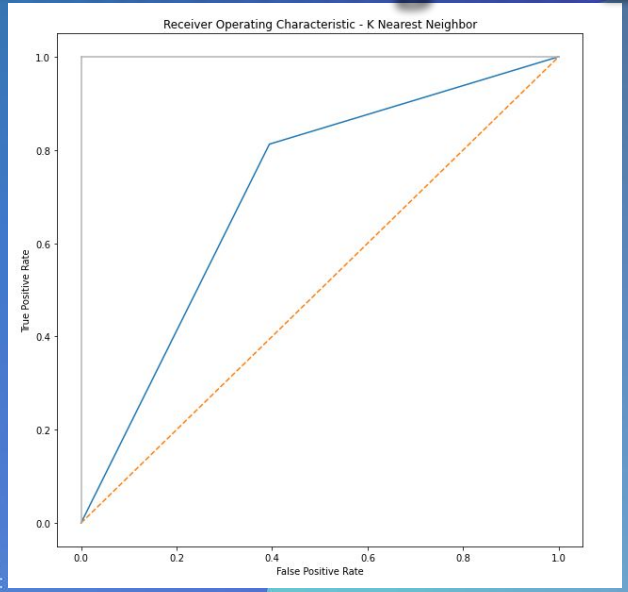
Recall: 74.95%

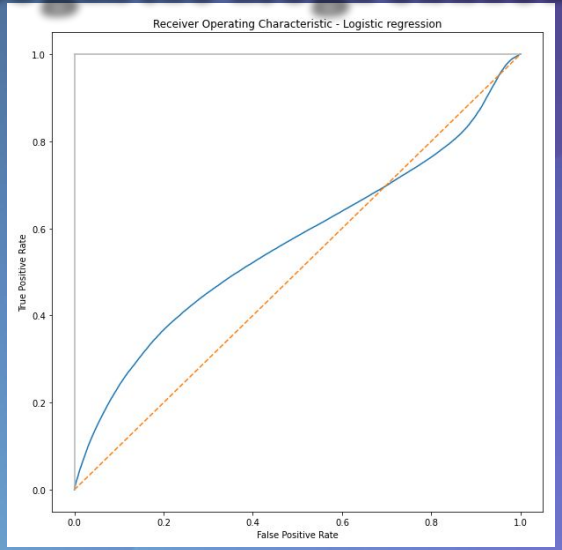
Best Results:

Random Forest

K Nearest Neighbor Logistic Regression







Next Steps

XGBoost

Fit an XGBoost model with fully optimized hyperparameters

Multiclassification

Predict for upper, middle, and lower classes instead of a binary split at 50K

Flask App

Create an interactive flask prediction app

IPUMS USA

- Originated in 1790
- Collected every 10 years as mandated by the Constitution
- In 2000, collection became annual

Collects a wealth of demographic & economic data

SUPPORTED BY

NIH National Institutes of Health
NSF
ancestry
Stat/Transfer
University of Minnesota


slidesgo