Census Income Classification

Marwan Mousa

The Problem

- The US Census Bureau collects demographic and economic data about Americans to help inform strategic initiatives.
- It covers all segments of the population to give a better understanding of its characteristics.
- Using this dataset we want understand what characteristics drive **high income**.
- To achieve this we
  - Build a classifier to **predict** an individual’s income status from their demographic data.
  - Use the model to understand which characteristics **affect** income the most.

The Dataset

- The dataset contains **42** demographic and economic data variables describing individuals.
- The variable of interest is **total person income**, which represents a person's total annual income.
- In this case, the total income is a **binary variable** representing whether an individual is a high income earner (> $50,000) or not.
- The dataset is already split into a training set and test set for evaluation.
    - The training dataset includes **199,523** instances
    - The test dataset includes **99,762** instances
- The dataset is imbalance however with only ~6% of instances high income.

Feature Engineering - Reducing Features

- With 33 categorical variables with **over a hundred distinct values**, the number of features used needed to be reduced to avoid the curse of dimensionality.
- Some variable are also very similar and essentially provide the same information while others applied to a minority of the population and were invalid for the rest limiting their usefulness.
- Variables were excluded if they **didn’t affect income**, or would result in many invalid instances.
- The effect of variables on income was decided **heuristically** based on the feature meaning or due to l**ack of correlation with the target**.
- The final list of variables used were:
    - Age, Race, Sex, Citizenship, Education and Marital Status.
    - Class of Worker, Employment Status and Employer Size.
    - Capital Gains/Losses, Dividends and Weeks Worked per Year.

The Algorithms

- Several machine learning algorithms can be used to create a classifier.
- Each algorithm has its advantages and disadvantages with respect to different tasks.
- We will be concerned mostly with their **explainability** and **predictive accuracy**.
- Generally as models become **more complex** their **predictive performance increases**.
- However this makes the models **less explainable**.
- The algorithms used were:

Simple

1. Logistic Regression
2. K Nearest Neighbours
3. Decision Tree

Complex

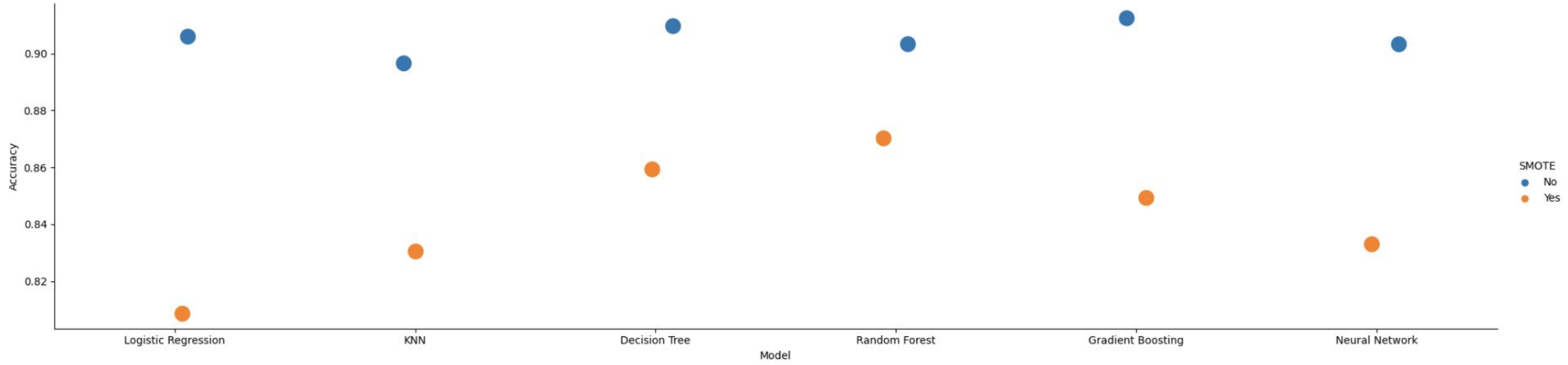
1. Random Forest
2. Gradient Boosting
3. Artificial Neural Network

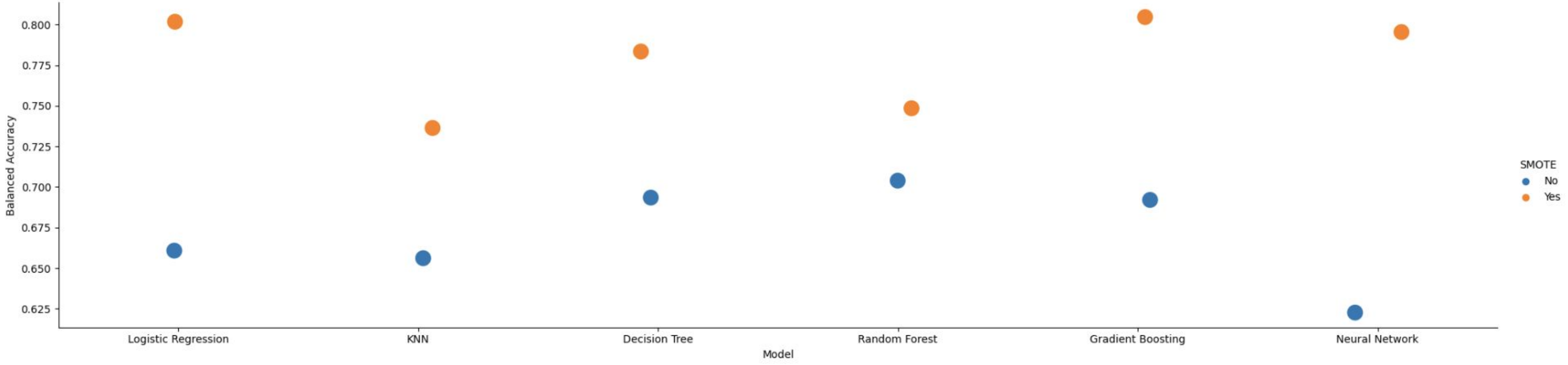
Model Evaluation Comparison

- We train two versions of each model from the algorithms previously described,
    - One is trained on the regular training dataset.
    - One is trained on a version where the **high income class is upsampled**.
- The **SMOTE**[1] method was used for the upsampling.
- After training the different models we compare them using the standard classification metrics.
- Given the heavy imbalance, we avoid using accuracy as a benchmark and focus on **balanced accuracy, f1 and recall**.

1. N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 321-357, 2002.

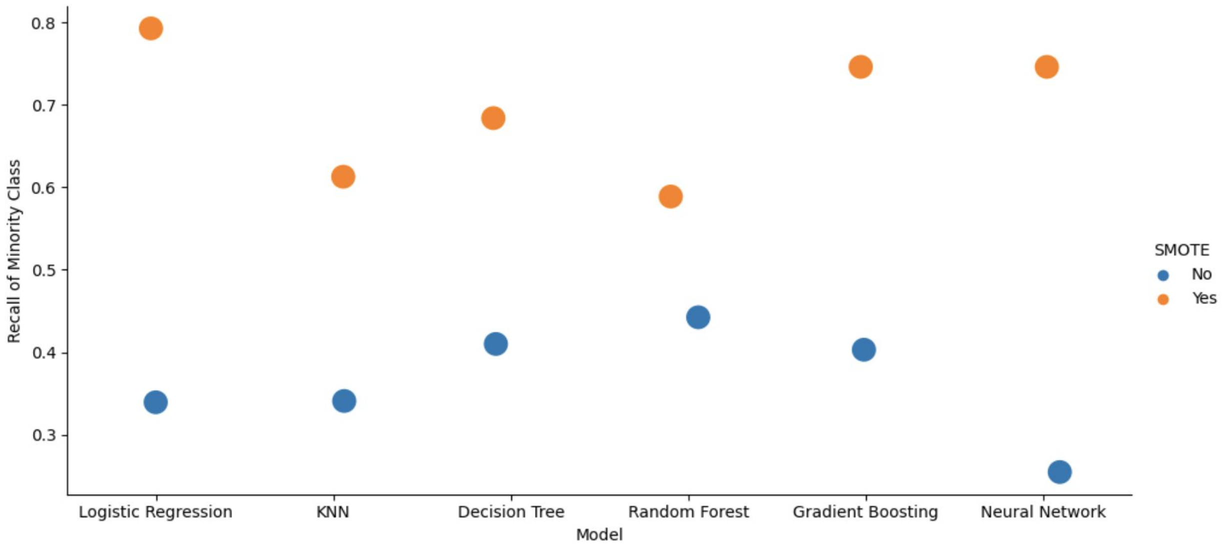
Model Evaluation Comparison - Accuracy



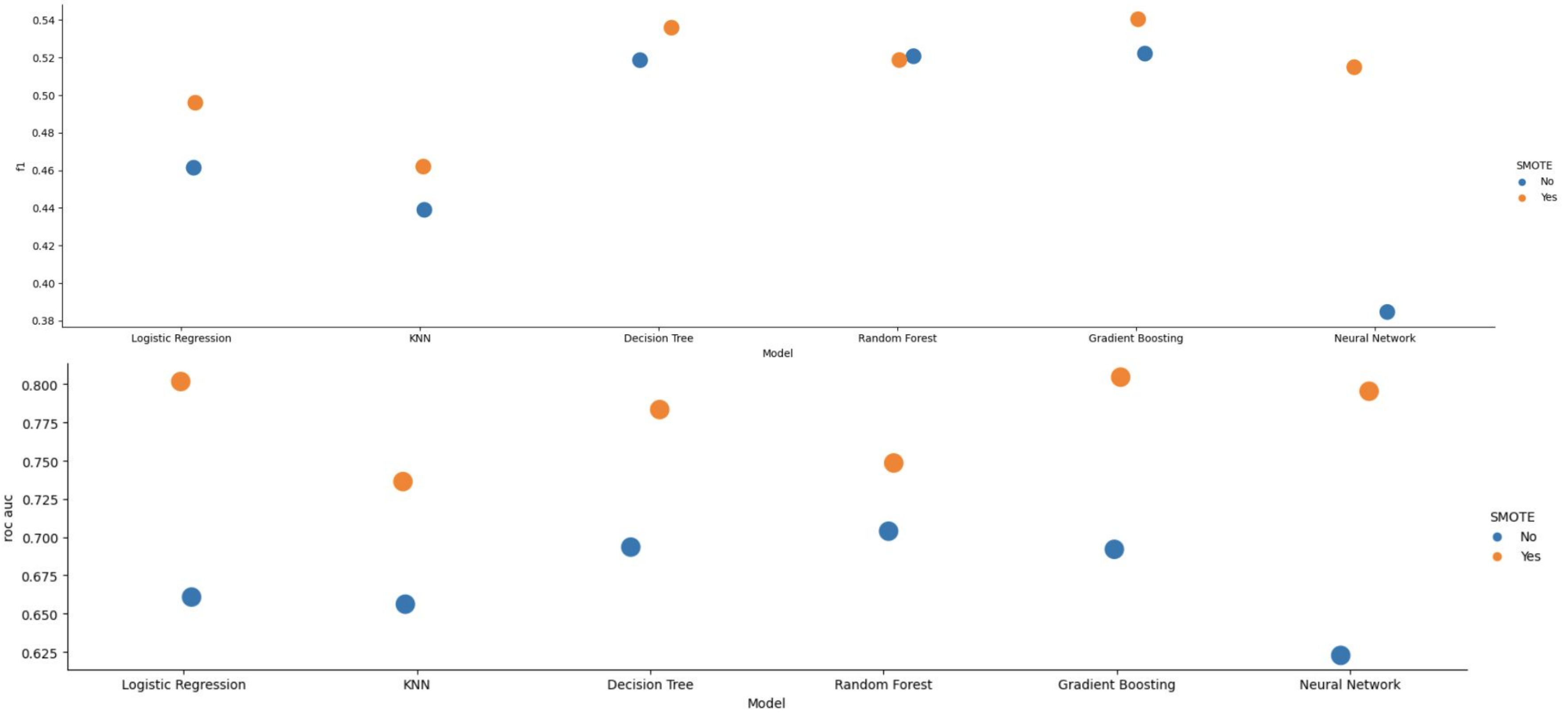


Model Evaluation Comparison - Recall

Using SMOTE greatly improved all models ability to identify the minority class



Model Evaluation Comparison - f1



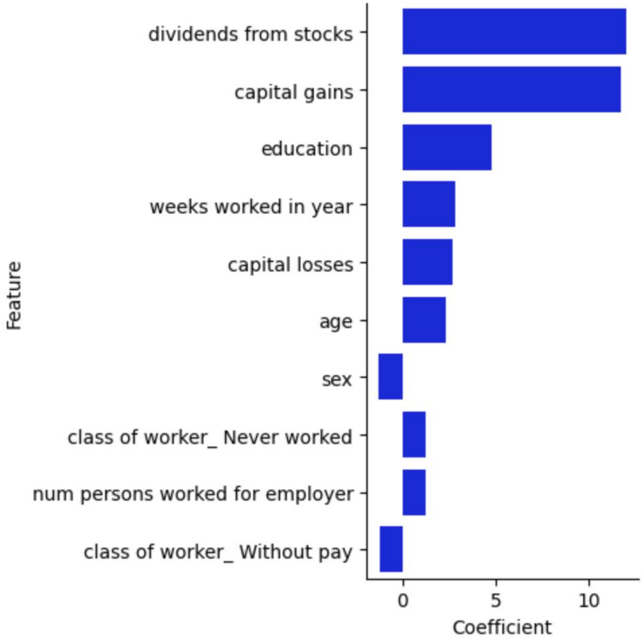
Beyond Prediction

- Having good predictive models doesn't necessarily tell us how the different features affect an individual's income.
- For policy makers to make use of these models they need to **understand** how the different variables **drive** the outcome, and which have the greatest effect.
- We focus on two models to attempt to understand the effect of the different features:
    - Logistic Regression
    - Gradient Boosting
- Logistic Regression is **inherently explainable** i.e. the coefficients represent the effect*.
- Gradient Boosting isn't explainable, but we can get a posteriori "explainability" by assessing feature importance using methods like SHAP.

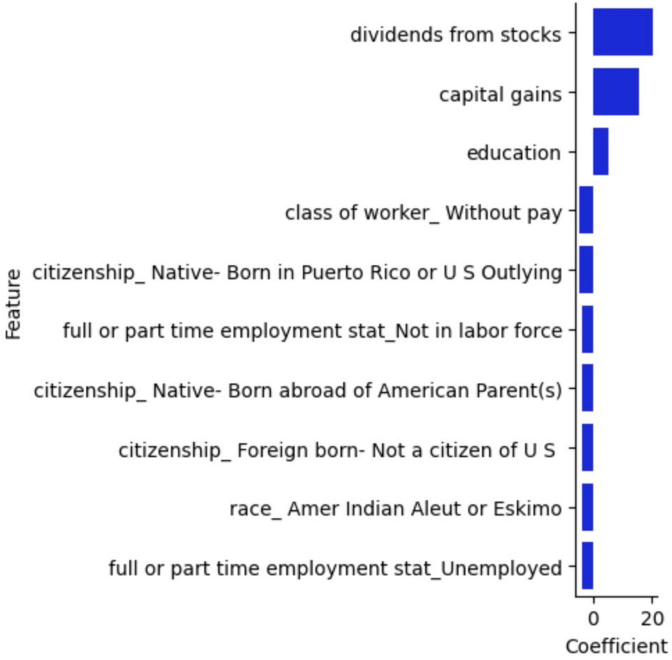
* assuming no confounding

Logistic Regression Explainability

Without Resampling

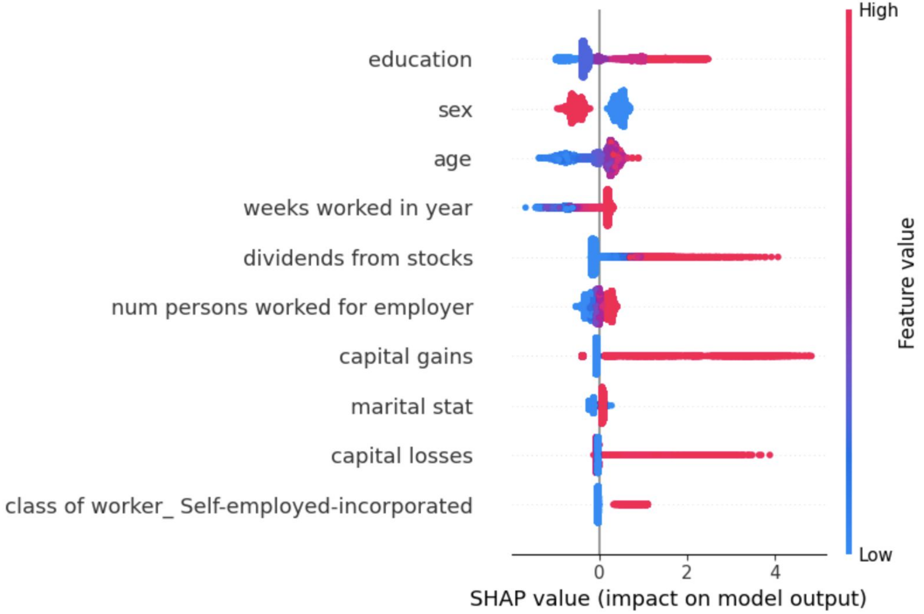


With SMOTE upsampling

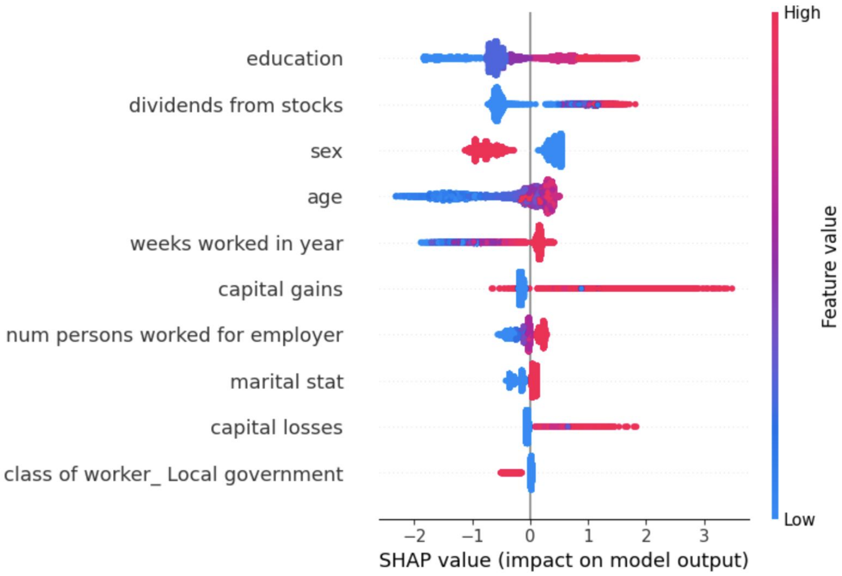


SHAP values

Without Resampling



With SMOTE upsampling



Explainability Summary

- Logistic regression coefficients can be viewed as effect of features while SHAP values shows the impact of the features on the model's prediction
- With SMOTE upsampling, the coefficients don't make intuitive sense, possibly due to artificial data points not being as meaningful
- With any oversampling, the logistic regression model seems to **agree** with the gradient boosting SHAP values on the most important features
- **Education** and **income for stocks** seem to be the best predictors of high income.

Future Improvements

- **Feature Engineering**
  - Variable were removed heuristically from an understanding of their definitions or if they weren’t correlated with the target.
  - Feature selection could be done more rigorously by assessing the correlation of each feature with the target **given** all other possible features.
  - This would include only features with a **causal or significant effect** on the target.

- **Training**
  - Hyperparameters weren’t tuned when training the models.
  - A fair comparison between algorithms would require hyperparameter tuning.
  - It is much more likely to result in **better performing** models as well.

Beyond Explainability

- In standard logistic regression, the model does not take into account how the features are **related**.
- It assumes the features are correlated with the **target only**.
- We can perform conditional independence tests to understand how the variables are related.
- This can allow us to build a causal graph, which would inform on which variables to include or exclude in the regression to avoid confounding.
- This will give a more accurate representation of how features **affect** the outcome.