Dataiku Assignment

Morley 11/2024

Overview

Objective: Identify characteristics associated with individuals making more or less than $50,000 per year.

Dataset provided: US census data 1994/95

Requirements: Demonstrate to the business how how we can achieve this objective using data science, specifically a modelling pipeline.

Dataset overview

The US Census Dataset

- A database created from a survey that provides an understanding of the US population at a given point in time
- There is a wealth of information that describes a given individual such as age, gender, country of birth, occupation, etc.
- As it is generated from a sample of the population there is a weight that indicates the number of people in the population that each record represents.
- The target variable we are researching has been created by binning the income, and creating a split at the $50K level giving two categories, over 50,000 and under 50,000 of income a year.
- Size of dataset: 299,285 rows
- Total individuals represented
    - 1994: 253,690,657 (15,525,462 in the over 50K category)
    - 1995: 256,465,897 (17,717,307 in the over 50K category)

Business consultation

Feature exploration & Assumptions

- To simplify data exploration and modelling we have elected to look at the most recent year of the data 1995.
- Those who are under 14 have been removed from the data as that is the youngest working age in the US.
- Working with key business stakeholders we have identified a number of initial features that will best determine if someone falls into a given category, these are:
    - sex, full or part time employment stat, age, member of a labor union, country of birth self, education, major occupation code, wage per hour
- Furthermore, using our own best judgment we have also elected to include the following features:
    - capital gains, capital losses, dividends from stocks, weeks worked in year
- While duplicate records have been removed from the dataset some conflicting instances still remain and may have some impact on the results of the study, this was deemed acceptable.

Approach

Five key steps to arrive at our solution

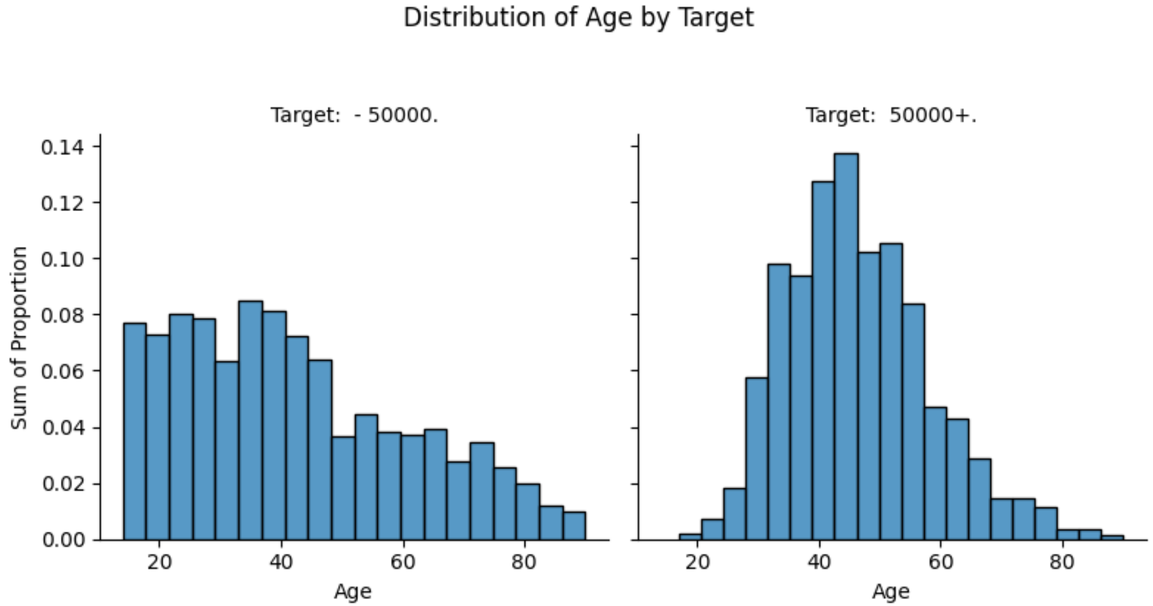
- Exploratory Data Analysis
- Data Preparation
- Data Modeling
- Model Assessment
- Results

Exploratory data analysis

What can we learn from the data?

A larger amount of the people who earn over 50K are male than female

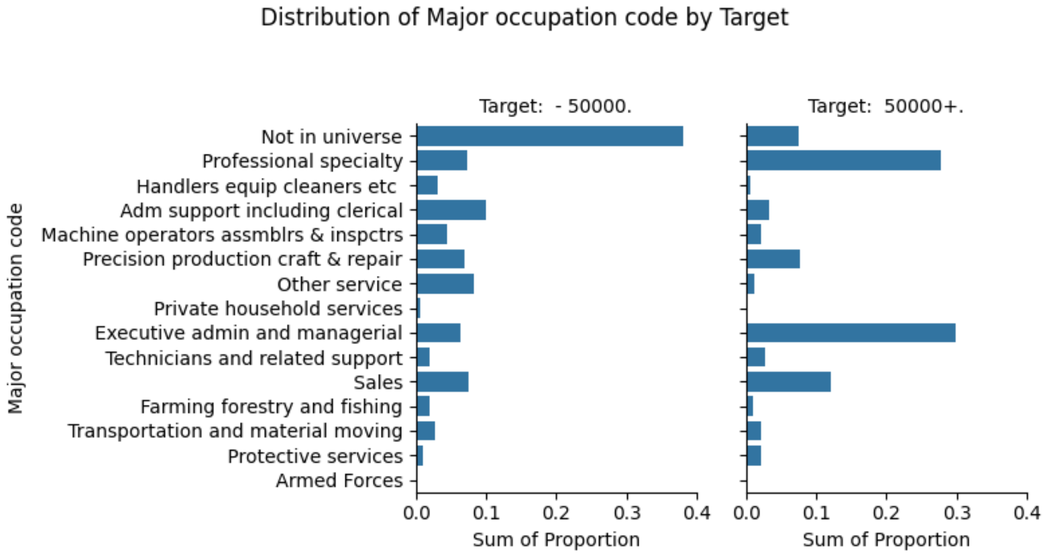
Those in the 50K+ category are largely middle aged



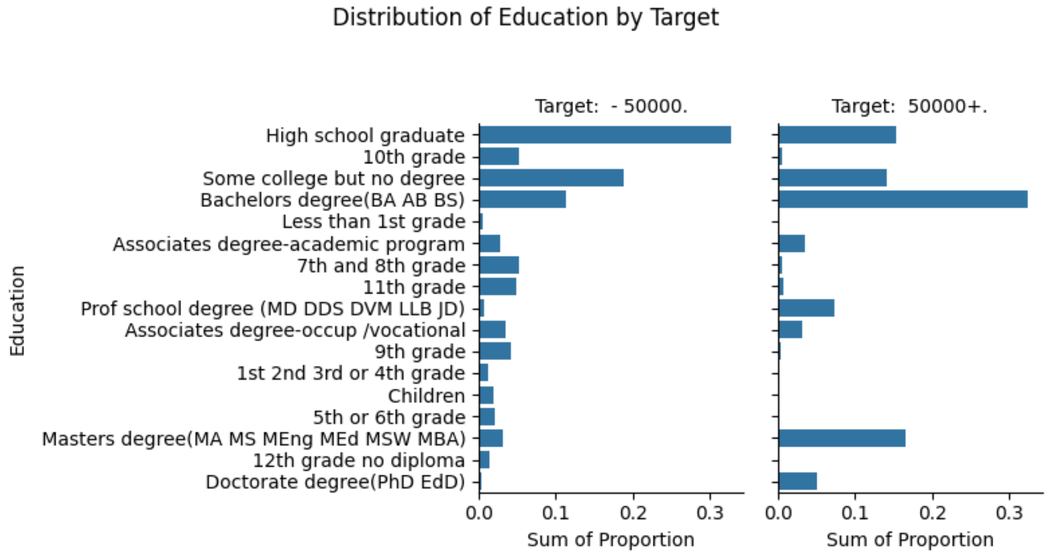
Exploratory data analysis

What can we learn from the data?

Those earning over 50K look to trend to more professional services



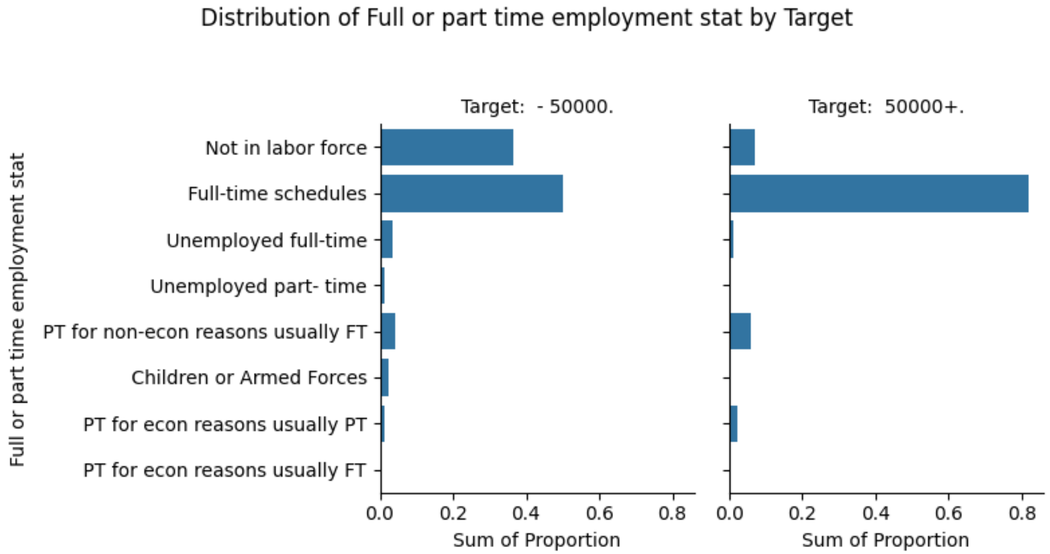
A larger proportion of those in the 50K plus bracket look to have education greater than a high school graduate



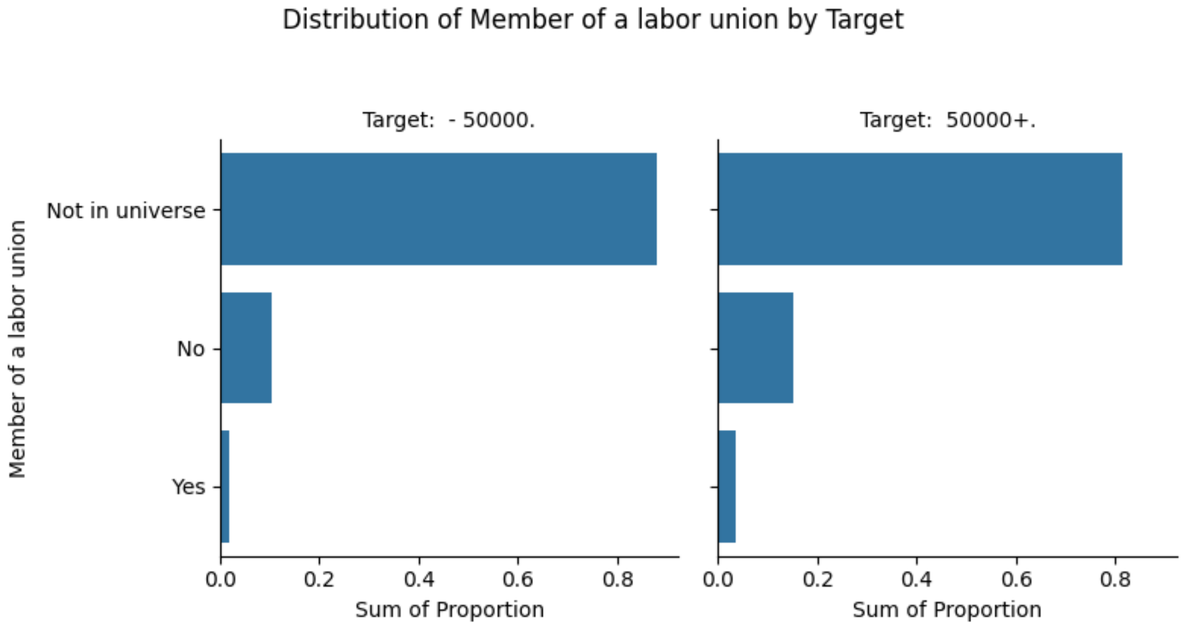
Exploratory data analysis

What can we learn from the data?

Unsurprisingly a higher proportion of people earning over 50K are employed full time

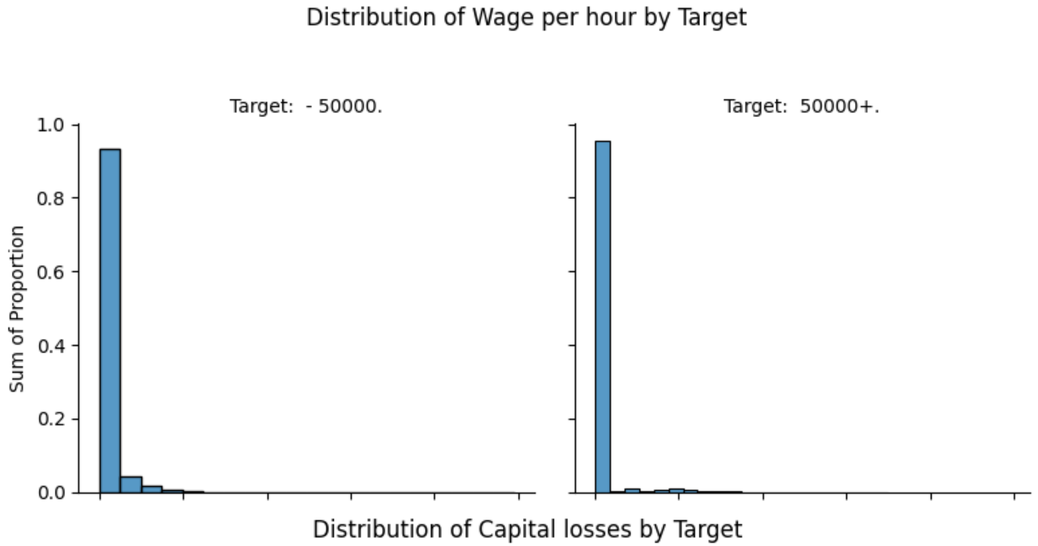


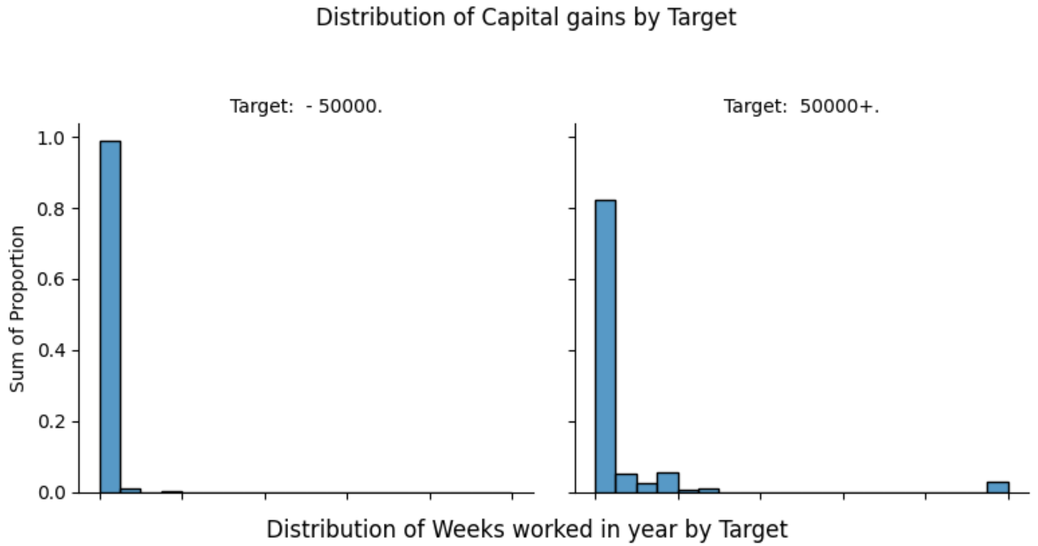
Union membership doesn't appear to be meaningfully different across the target variables

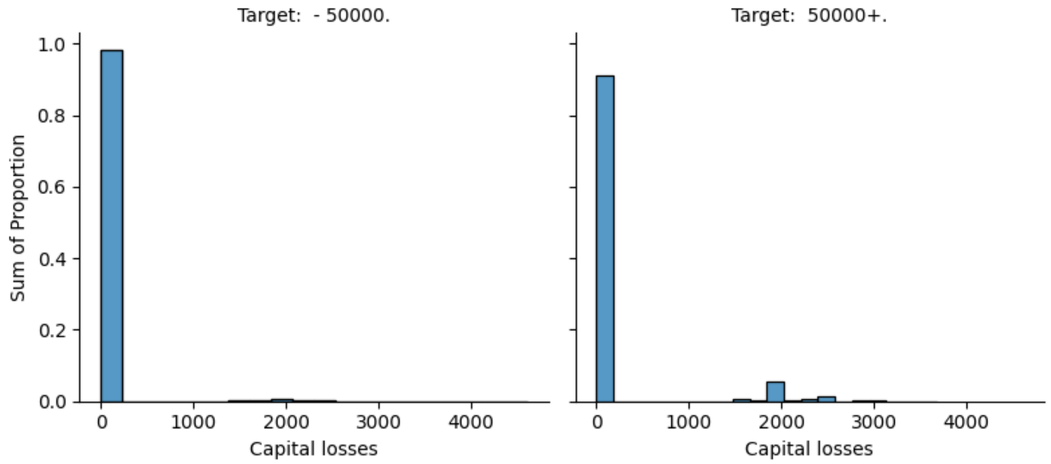


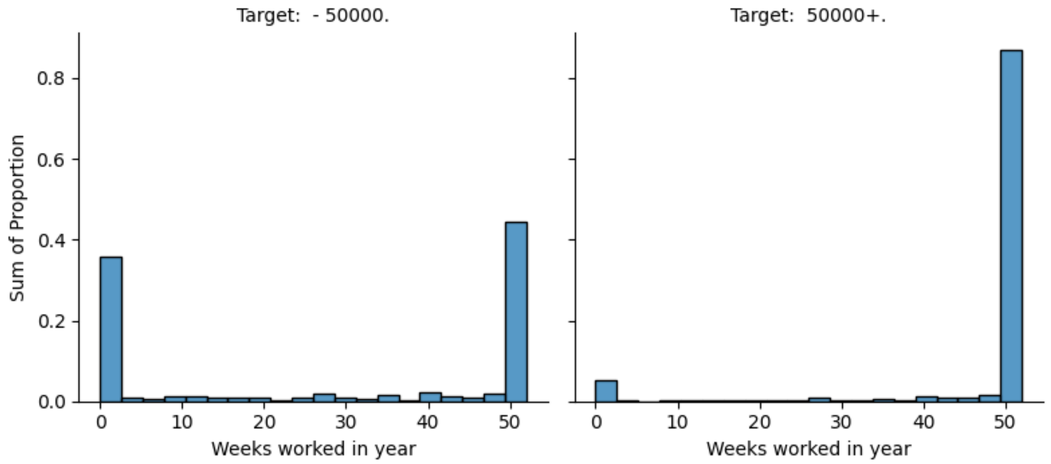
Exploratory data analysis

For almost all of the income related continuous variables over 80% are 0 across both target variables.









Data Preparation

Transform the data so that the model is able to understand it

- Reduced dimensionality of some variables
  - Age to age_group
  - education to education_categories
- Categorical variables encoded using one hot encoding
- Transform continuous variables with large numbers of 0 into flags
- Weeks worked per year min max scaled

Data Modeling

Model Selection, training, and tuning

- Classification task with an unbalanced dataset
- Use of three differing models to evaluate performance
  - Logistic regression, a benchmark model
  - Random forrest, a model with high explainability
  - Boosted tree, a model that should perform well with an unbalanced dataset
- Test train split of the data
- Class weight consideration
- Hyper paramater tuning
  - C, Depth of tree, n_estimators, learning rate, subsample
- Model evaluation metric, f1 score (macro)

Data Modeling

Model evaluation

	Logistic Regression	Random Forrest	Gradient Boosted
Accuracy	0.82	0.83	0.93
F1 Macro	0.67	0.67	0.70
Performance Under 50K	Precision: 0.98 Recall: 0.81	Precision: 0.98 Recall: 0.83	Precision: 0.94 Recall: 0.98
Performance Over 50K	Precision: 0.30 Recall: 0.85	Precision: 0.31 Recall: 0.83	Precision: 0.63 Recall: 0.35

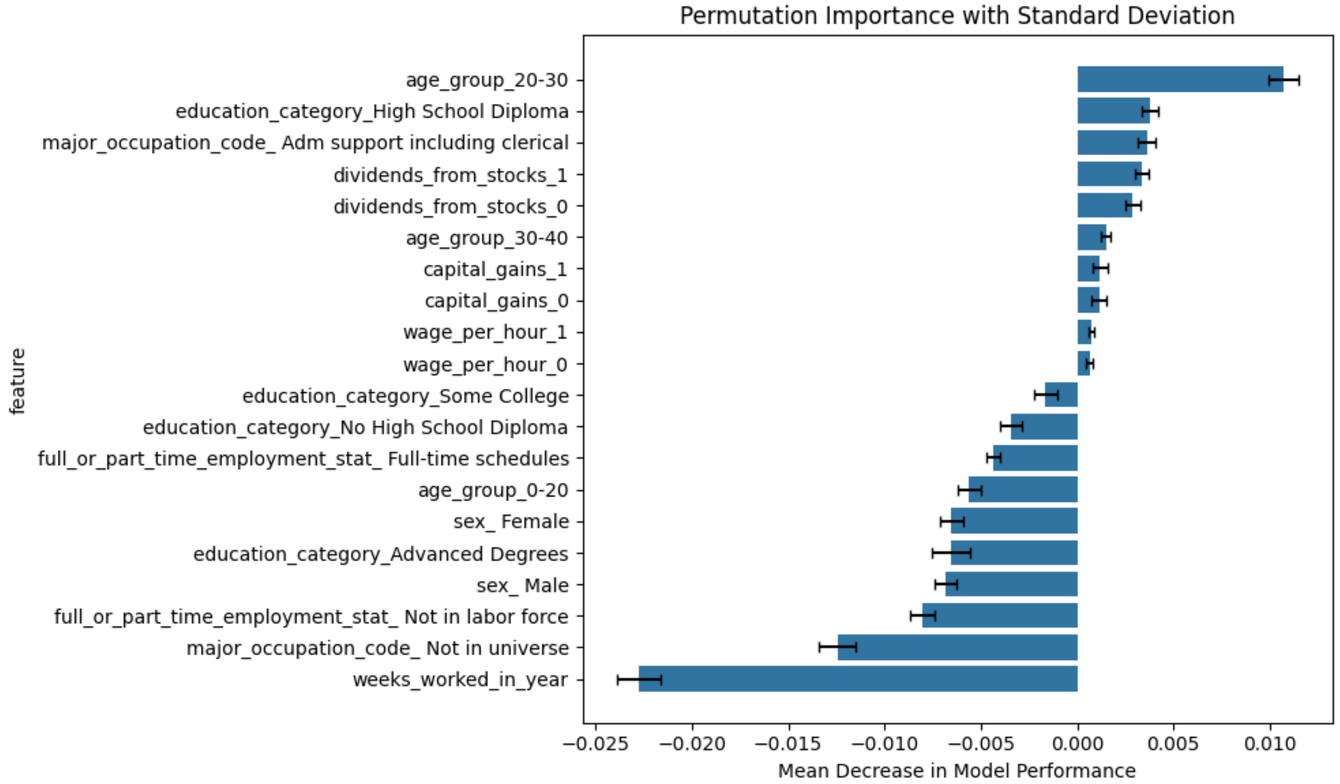
Results

Understanding the biggest drivers of differing incomes

Using our random forest model we are able to individually asses the importance of each feature on predicting the outcome.

The most impactful characteristics of an individual when understanding their income

- Age
- Education
- Income sources



Next steps

- Using the information we've learnt we could opt to add more features to the model
  - Use feature importance information to refine feature engineering and selection
  - Remove some low performing features
- Explore alternative measures for dealing with the unbalanced dataset such as oversampling using SMOTE
- Explore more complex models that may be able to better handle the complex data
- Use a more complex feature exploration tool like shap to explain what characteristics drive a person to be in one or the other categories

Q & A