data
iku

A Data Science Analysis of United States Census Bureau Data

ARPITHA P. BHARATHI

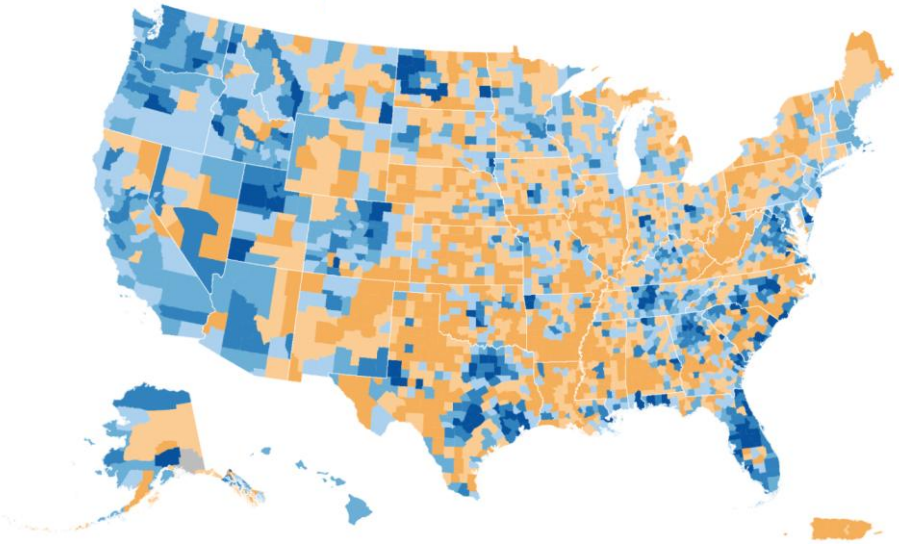
Problem definition & objectives

Problem statement:

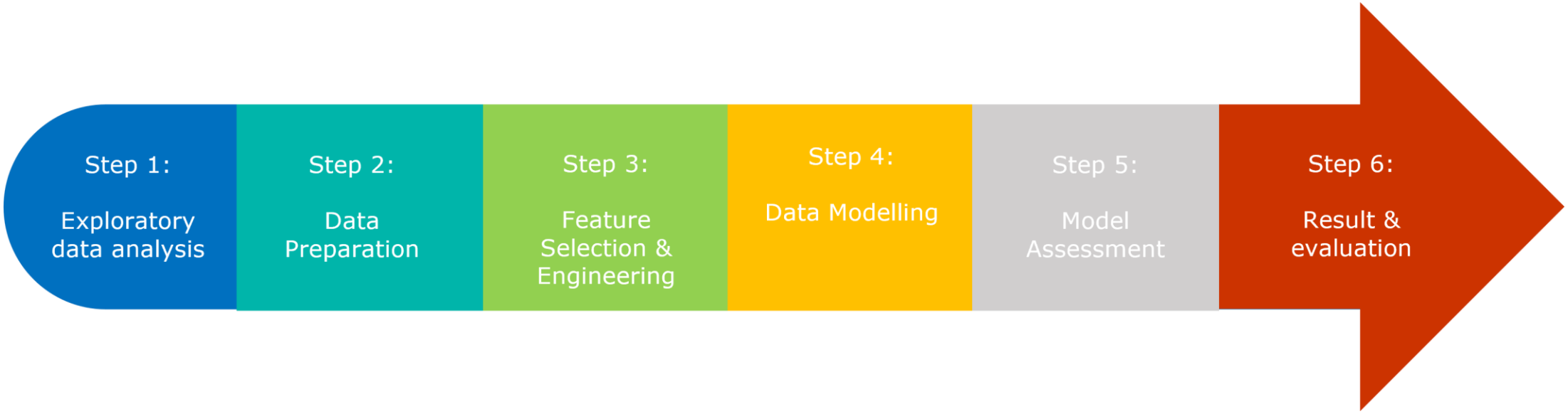
Given US census data, identify the key characteristics that are associated with a person making more or less than $50,000 a year.

Objectives:

- Understand relationships between key characteristics (41 provided) of a person within the census data and income.
- Implement machine learning models to predict income given the characteristics of an individual.
- Recommend the right machine learning model based on performance.
- Recommend further improvements / refinements to both the recommended model and data sets used if any.



Scoping of the problem



Data preparation

- Cleaned the data of blank spaces.
- Looked for outliers in the data: e.g. Null values, wage per hour, capital gains. Statistical methods were used to handle outliers.
- Converted categorical data to numerical data (label and one-hot encoding).
- Some categorical data were paid special attention and treatment:
    i. Education
    ii. Country of birth

Feature selection & feature engineering

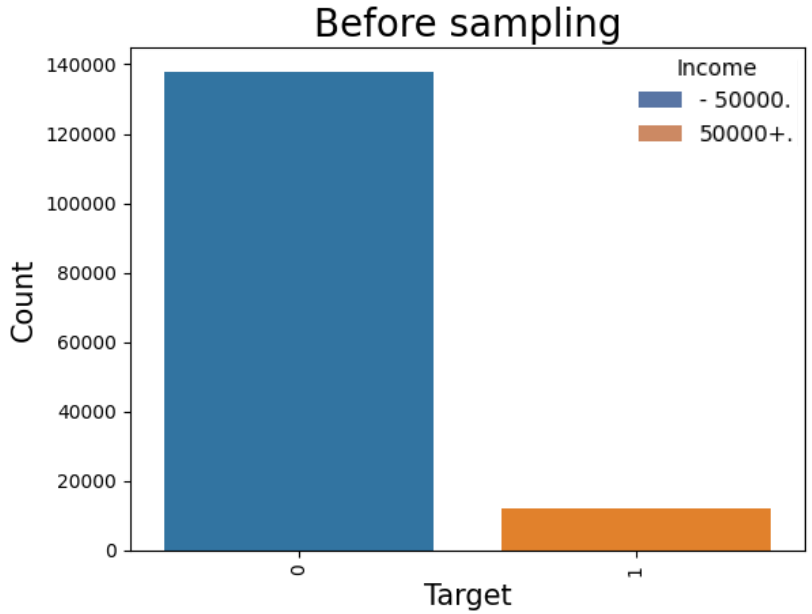
- Data scaling is applied.
- Columns that essentially give the same information are dropped using a correlation matrix (7 were dropped).
- Introduced a feature called ‘net income’.

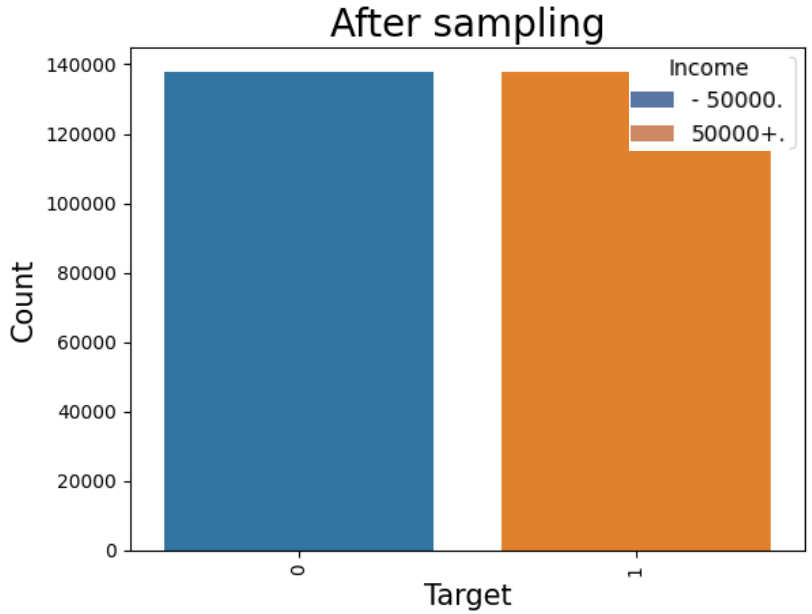
Net income = Capital gains – Capital losses

- Categories of capital gains & capital losses were dropped from the data.

Data sampling:

- Original dataset is imbalanced: over 93% are low income earners and the rest are high income earners. The data is balanced via a technique called called sampling to improve model performance.

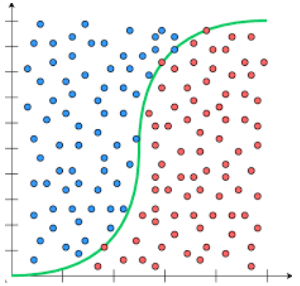


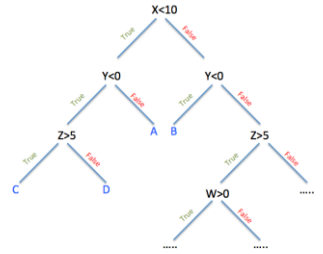


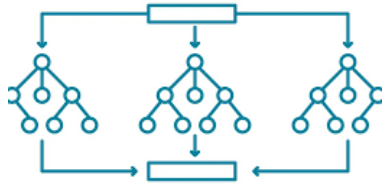
Machine learning models

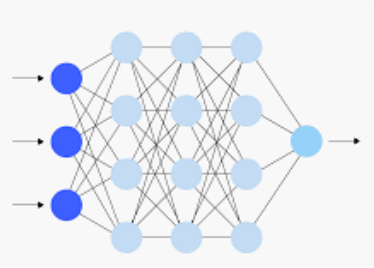
The prepared data with selected features is used to train the following machine learning models:

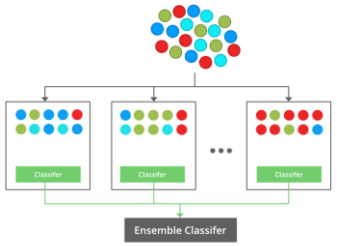
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier
- Voting Classifier
- Ada Boost Classifier
- Neuralnet







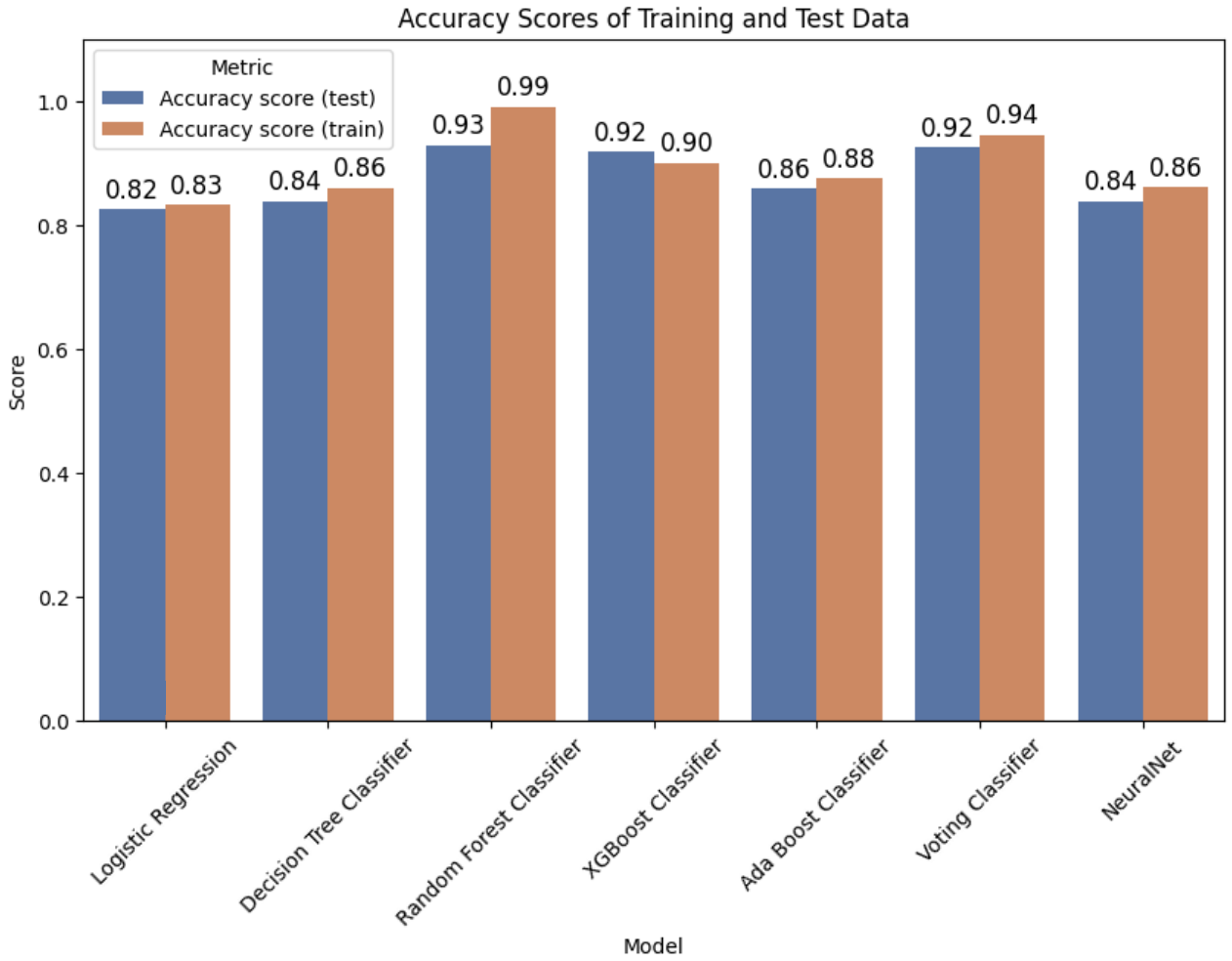




Machine learning model performance: Accuracy

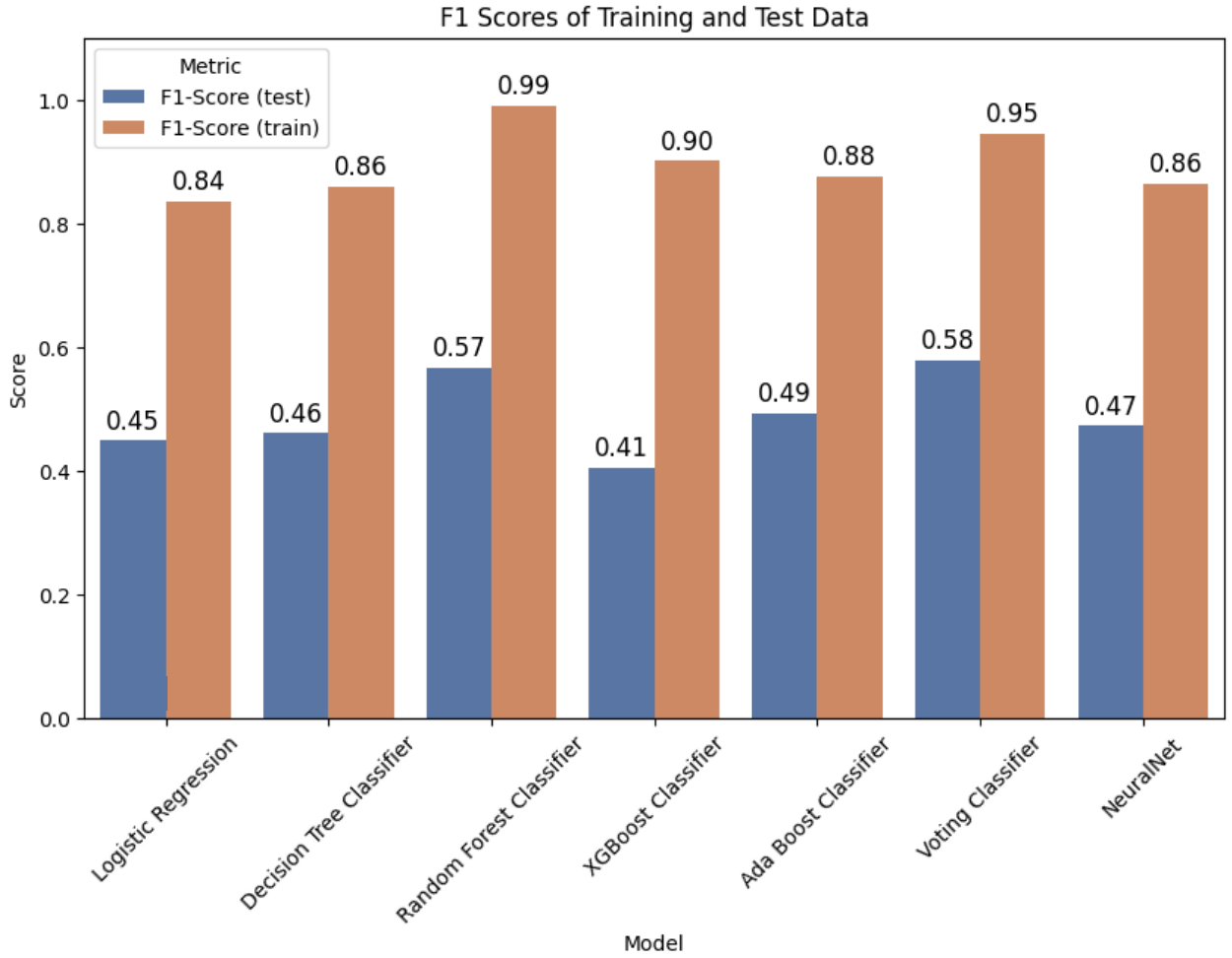
Machine learning models are evaluated based on their accuracy and F1 scores.

- Accuracy scores measure the percentage of data correctly predicted.
- Accuracy scores for training data and test data are measured.
- Random forest classifier and voting classifier perform the best among all the models.



Machine learning model performance: F1 Score

- F1-scores measure how correctly the data has been classified.
- Random forest classifier and voting classifier perform the best among all the models on the F1 score metric.
- On further inspection, Random forest classifier performs too well on the training data indicating the model might be overfit.
- Thus, Voting classifier is chosen as the recommended model.



data
iku

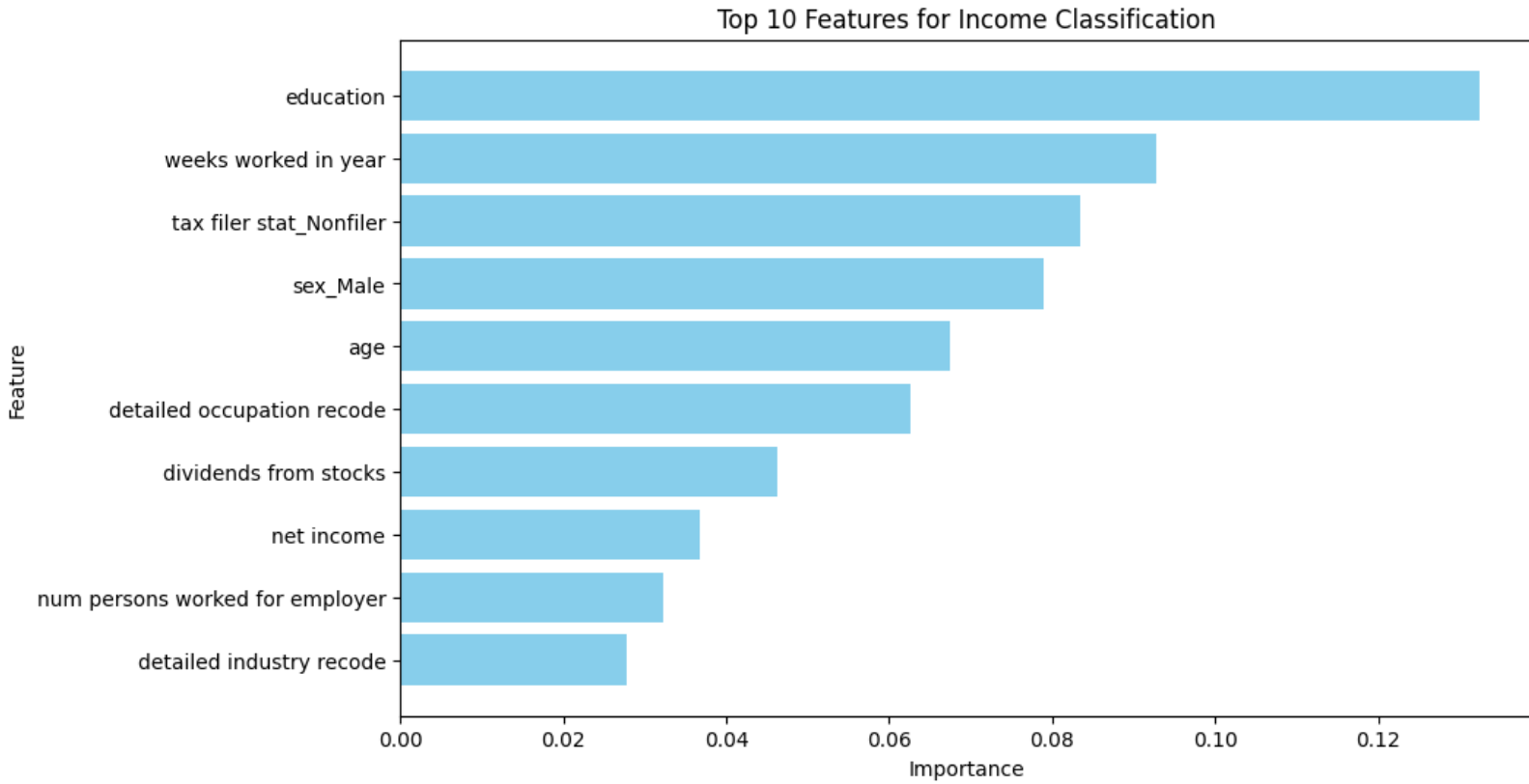
Evaluation of Results & Next Steps

Characteristics of individuals less/more than $50k

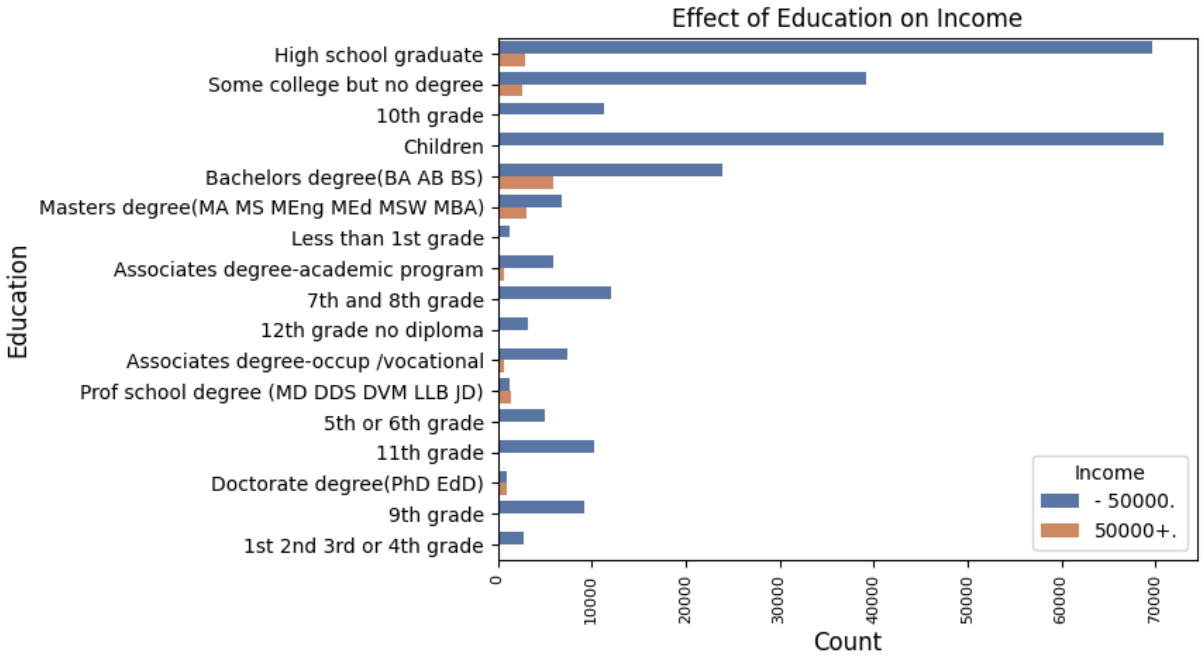
- The voting classifier is used to derive the top features that play a pivotal role in determining income.

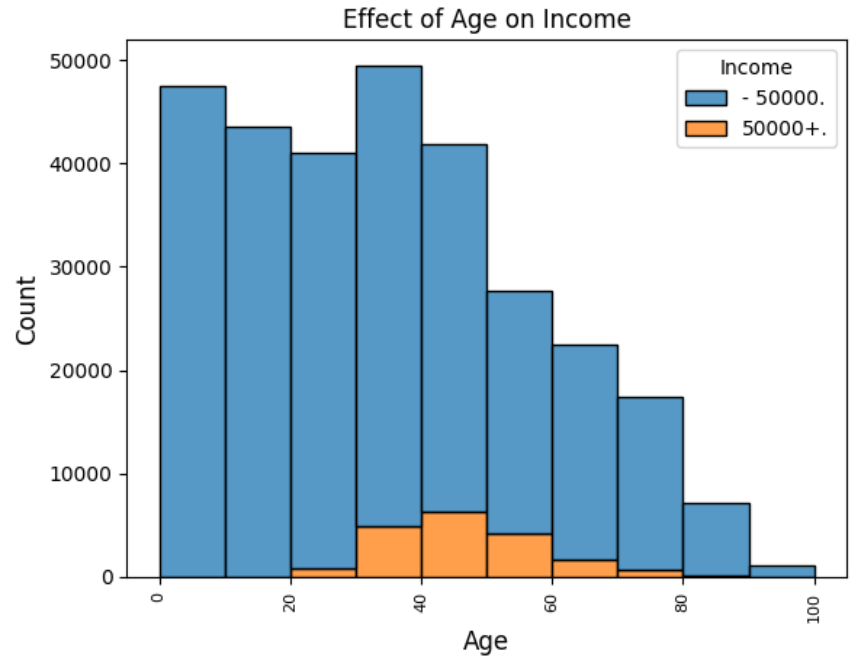
Key Observations

- *Net Income* which is an added feature is one among the top 10 features.
- *Education* is shown to be the most important among the top 10 features. This was a category which was given special attention while converting to numeric data.



Characteristics of individuals less/more than $50k





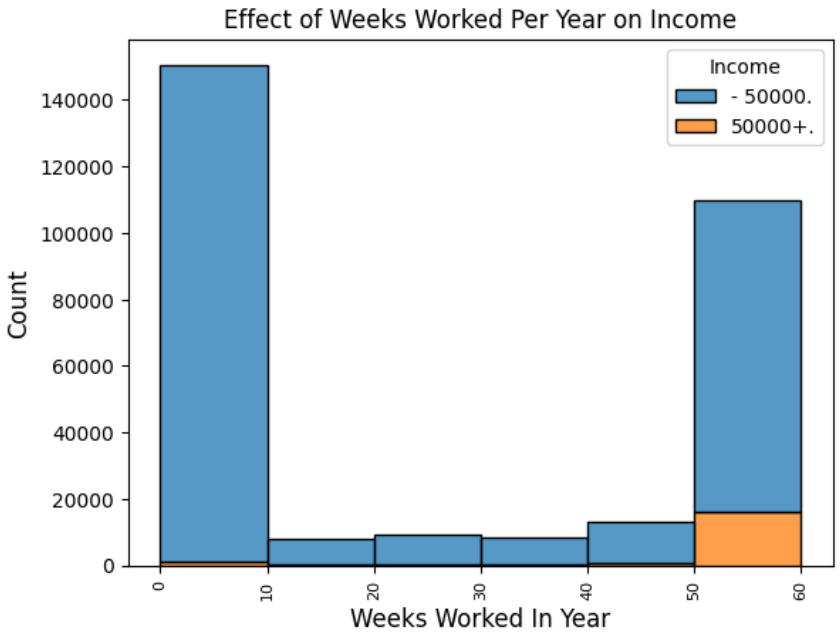
Effect of education on income

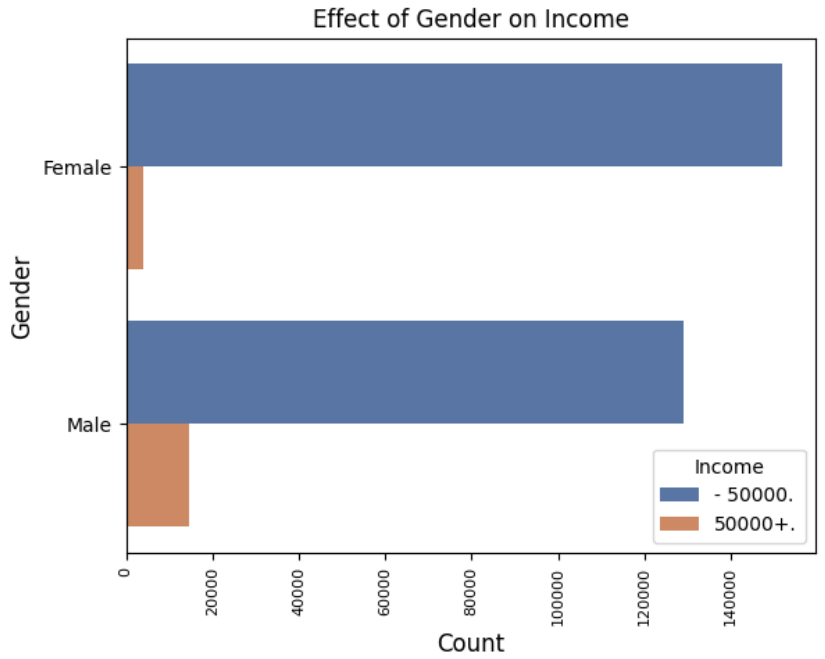
- Most of the people earning > $50k tend to have graduated high school, attended college and have a bachelor's or master's degree

Effect of age on income

- Most of the people earning > $50k tend to fall into bands between 20-80 years of age.

Characteristics of individuals less/more than $50k





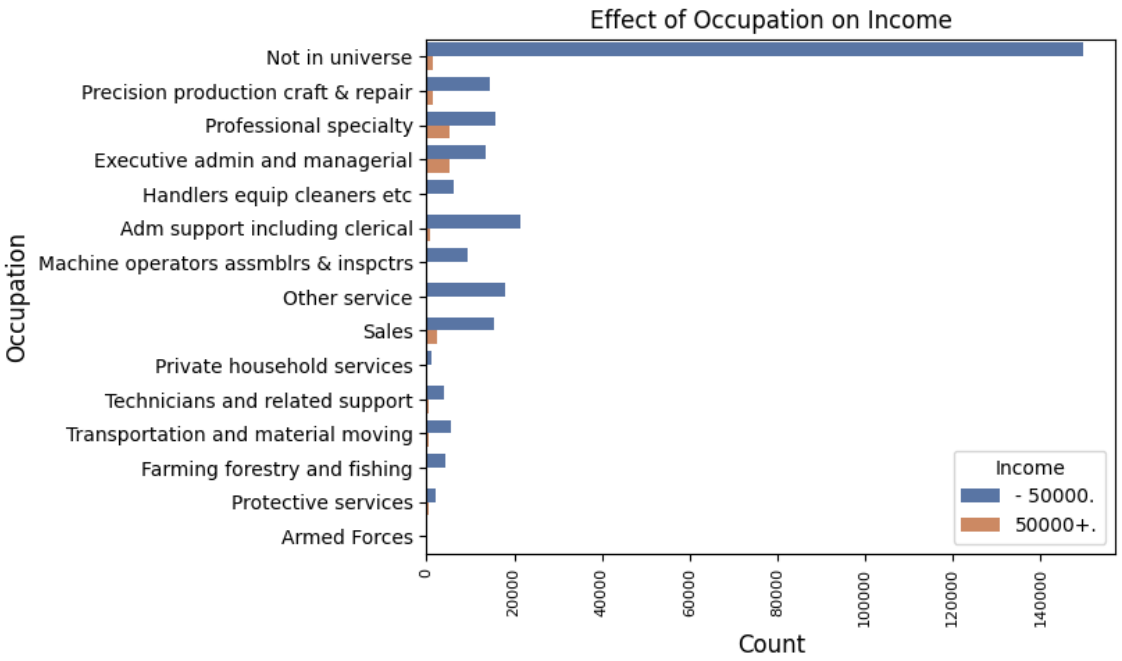
Effect of weeks worked in a year on income

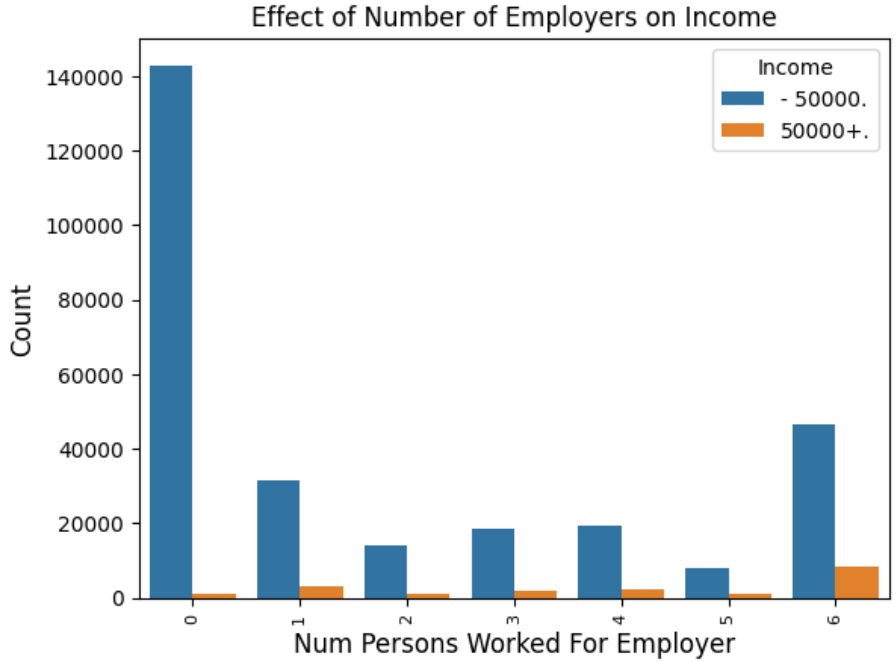
- People earning > $50k tend to work all 52 weeks.

Effect of gender on income

- People earning > $50k tend to be male.

Characteristics of individuals less/more than $50k





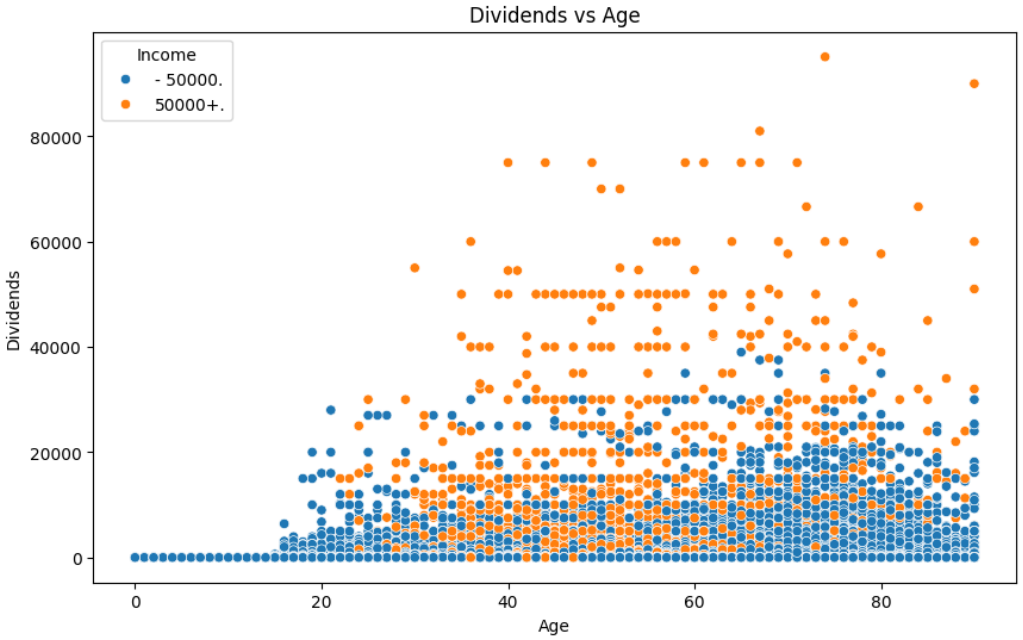
Effect of occupation on income

- People earning > $50k tend to be employed in management, sales or in specialist fields.

Effect of number of employers

- People earning > $50k tend to switch jobs/companies.

Characteristics of individuals less/more than $50k



Effect of dividends on income

- People earning > $50k tend to be older.
- We observe an increase in dividend income with age.

Effect of net income

- People earning > $50k tend to be older.
- One expects a trend of increasing net income with age.
- We cannot draw too many conclusions due to inconsistent trends in data.

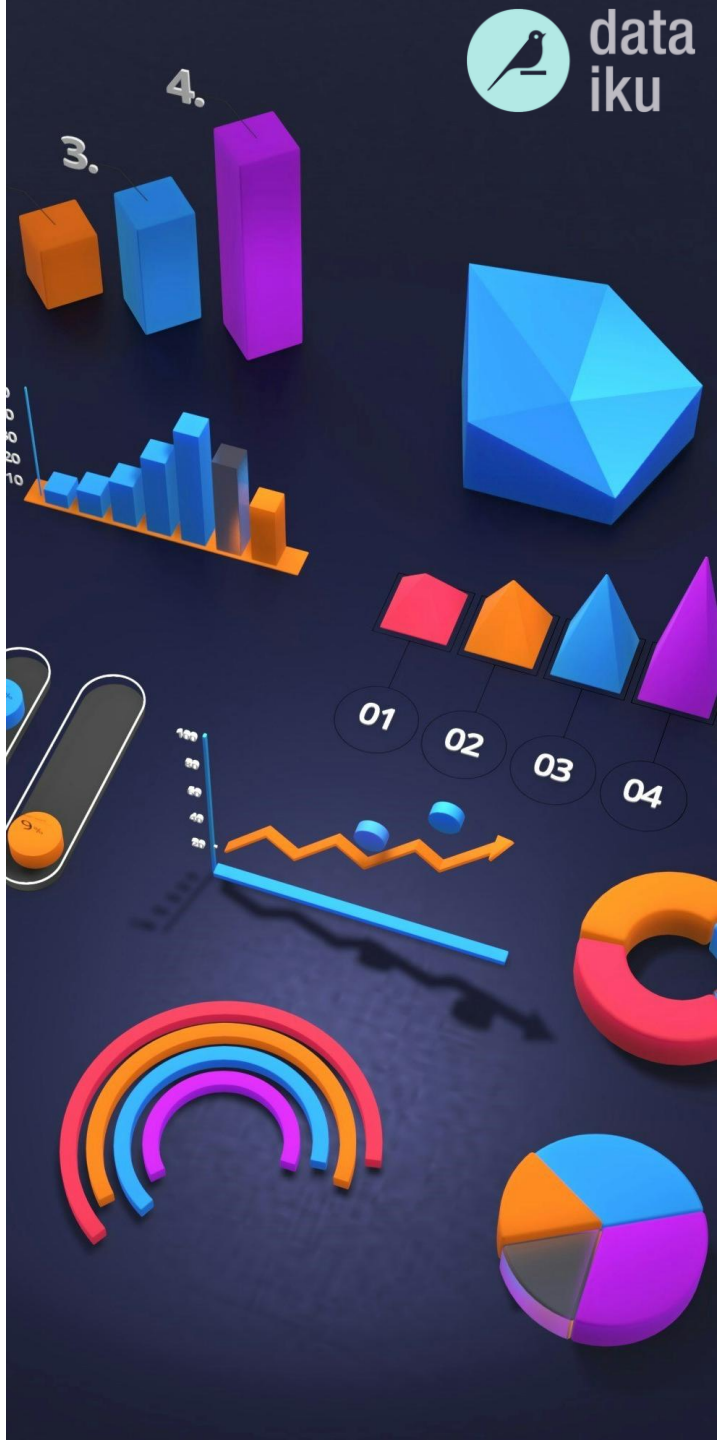
Future improvements

- Effect of children seems to sway highly towards low income earners (Approx 22%). Perhaps only 18+ year olds can be considered in order to gain more information regarding income.

- More attention can be given to feature selection to aggressively group together some categorical data.

- More data on the category *'Not in universe'* as it comprises of the largest portion of the occupation.

- Employ hyperparameter tuning techniques such as *GridsearchCV.*



data
iku

Thank you

Back-up

Characteristics of individuals less/more than $50k