A Data Science Analysis of United States Census Bureau Data

ARPITHA P. BHARATHI

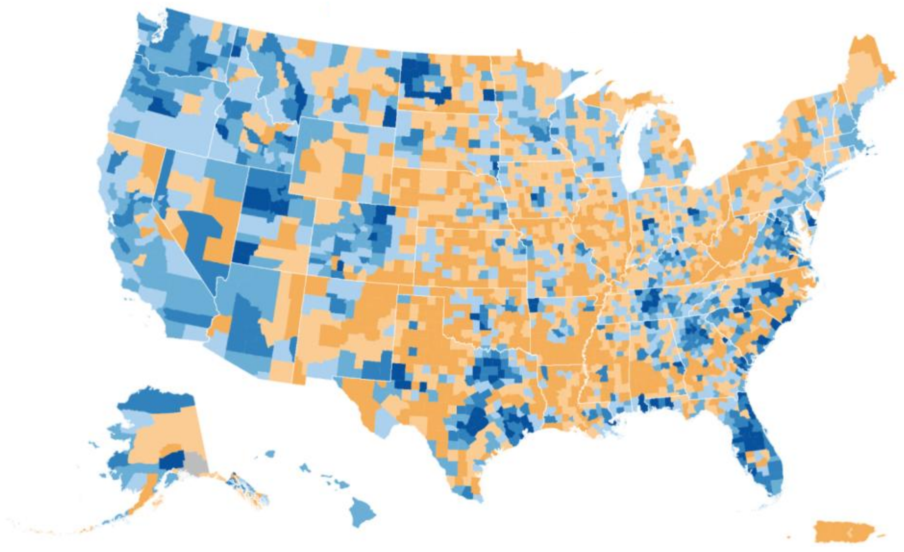
Problem Definition & Objectives

Problem statement:

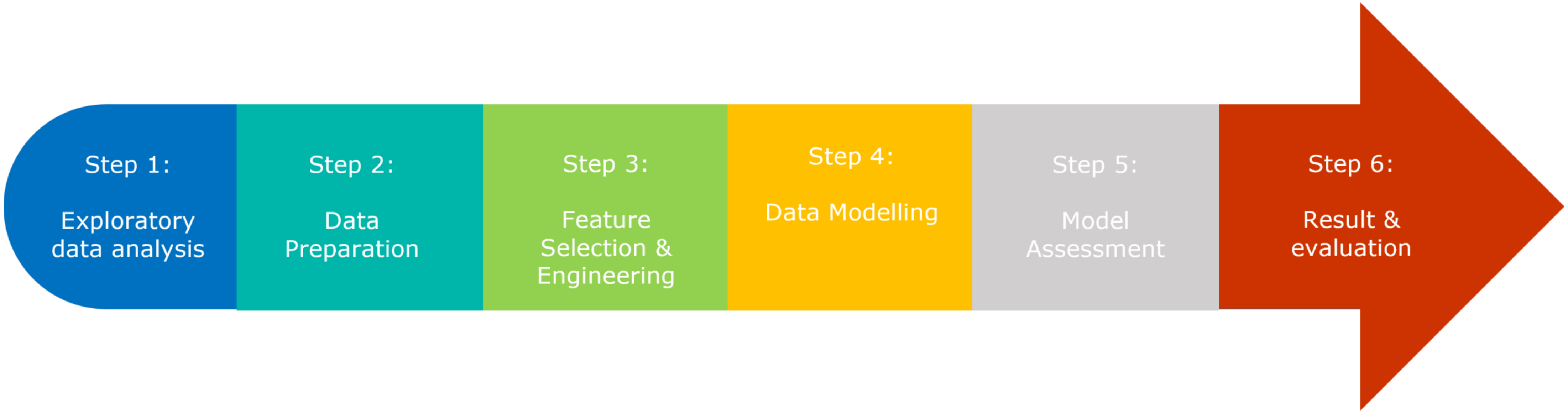
Given US census data, identify the key characteristics that are associated with a person making more or less than $50,000 a year.

Objectives:

- Understand relationships between key characteristics 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

- Imported the census data into a dataframe and combined both the test and training data.
- 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.
- Some categorical data were paid special attention and treatment:

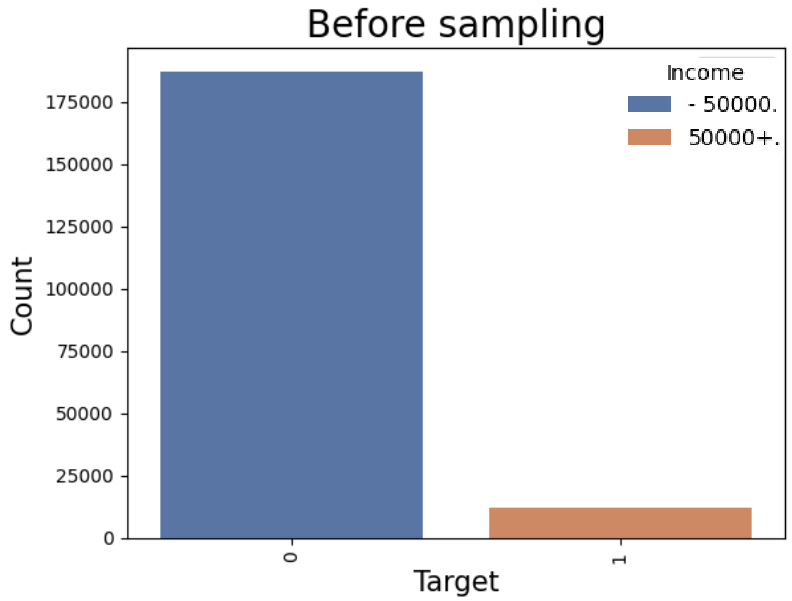
    i. Education
    ii. Country of birth

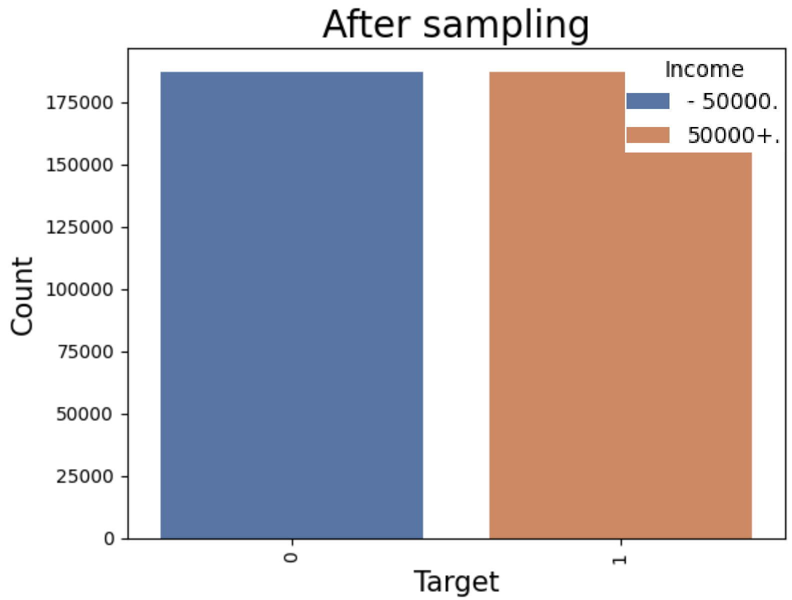
Feature selection & feature engineering

- Introduced a category of data called 'net income'.

Net income = Capital gains – Capital losses + Dividends from stocks.

- Data scaling is applied.
- The data is split back into test and training data sets.
- Columns that essentially give the same information e.g. Citizenship are dropped using a correlation matrix.
- Original dataset is imbalanced: over 93% are low income earners and the rest are high income earners. The data is balanced via sampling to allow better model performance.

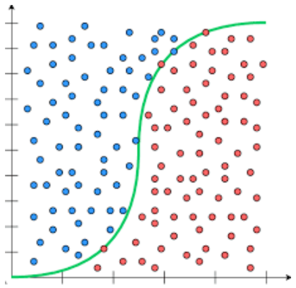


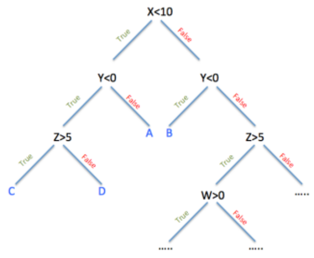


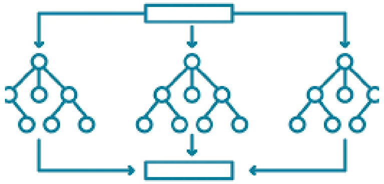
Machine learning models

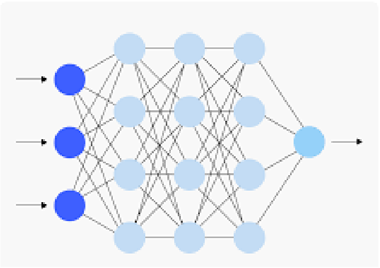
The prepared data with selected features is used to train the following machine learning models:

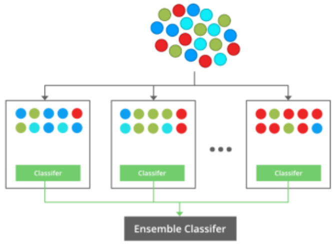
- Logistic Regression
- Decision Tree Classifer
- Random Forest Classifier
- XGBoost Classifier
- Voting Classifier
- Ada Boost Classifier
- Neuralnet







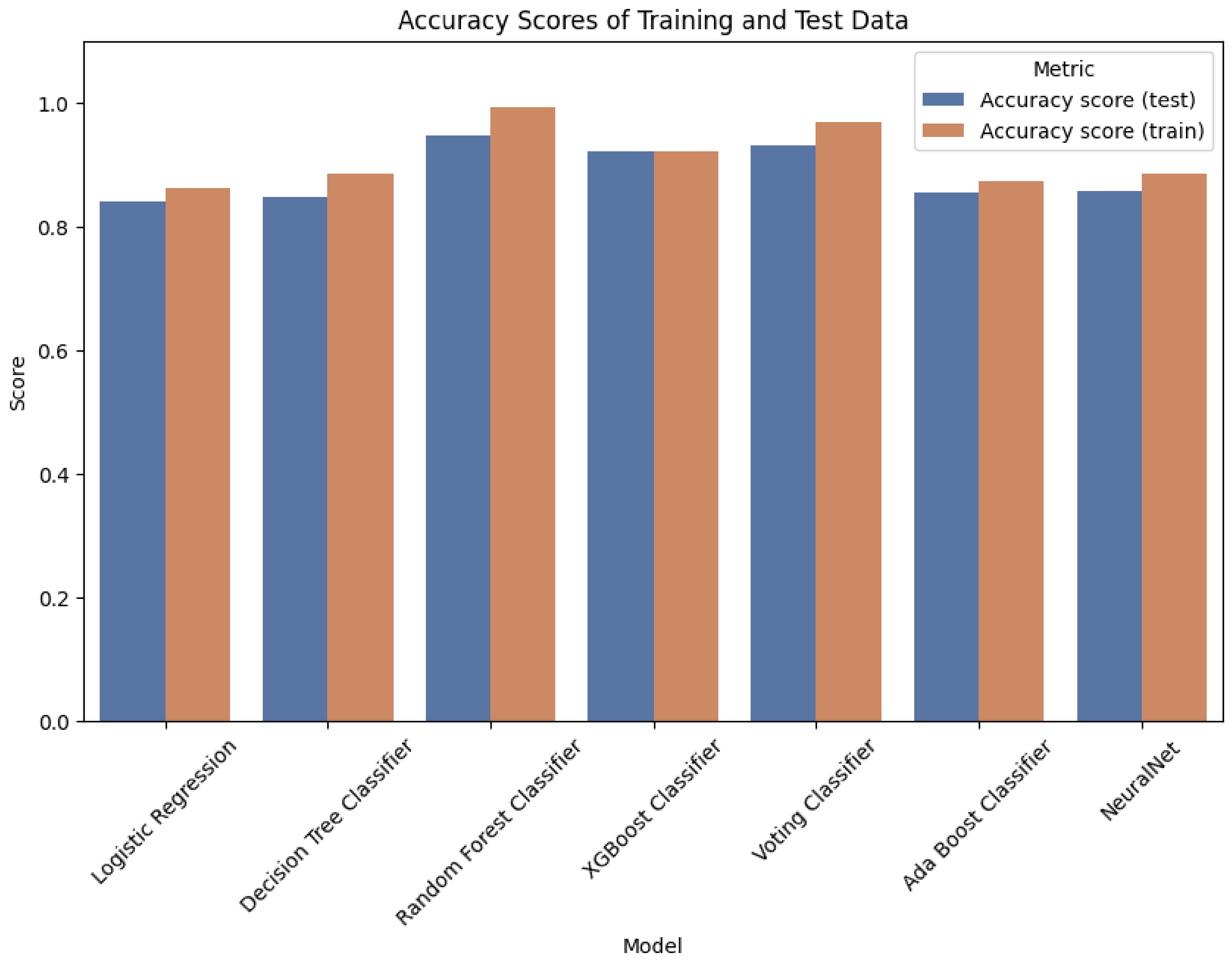




Machine learning model performance: Accuracy

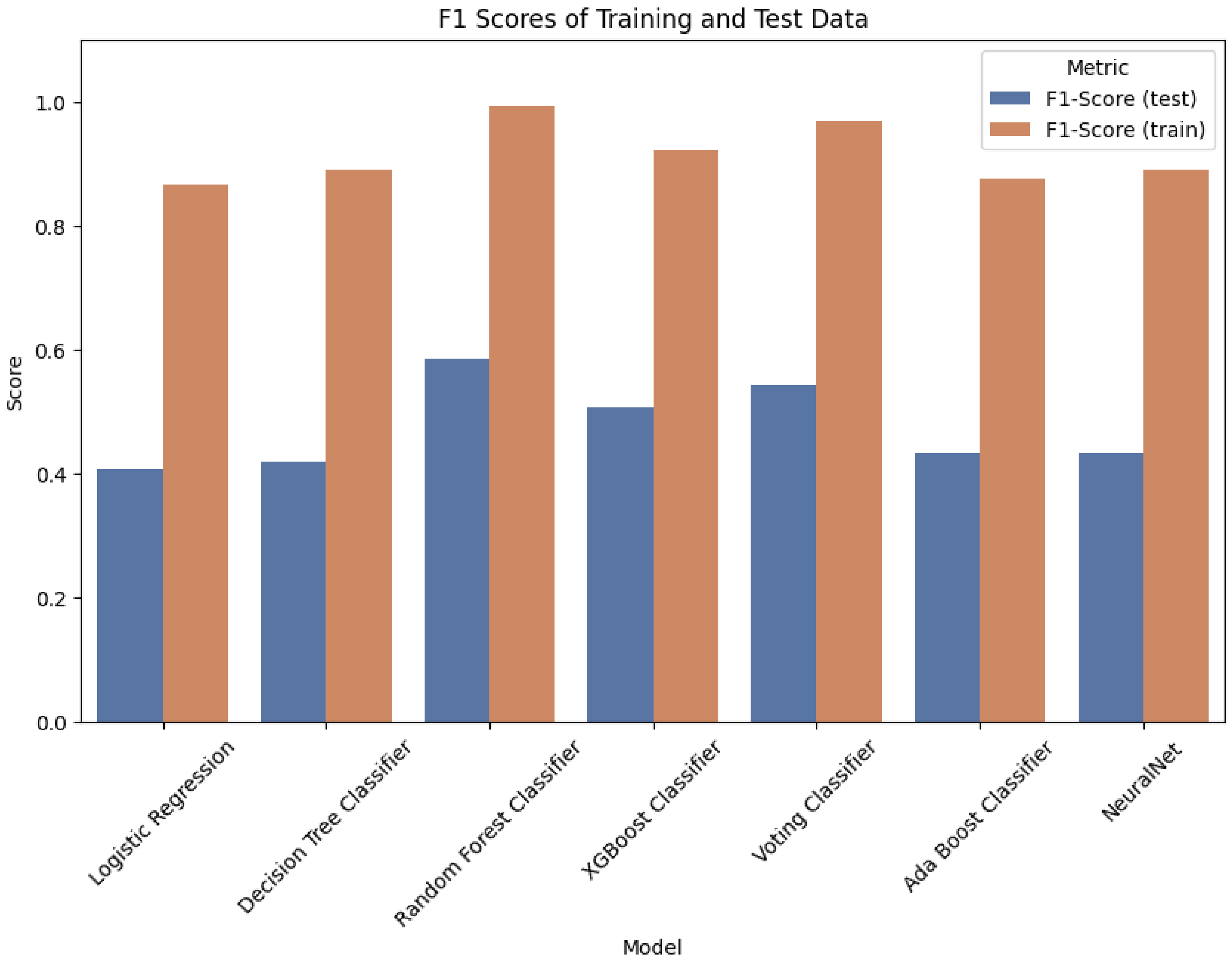
Machine learning models are evaluated based on accuracy and F1 scores.

- 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 seem to perform the best among all the models.



Machine learning model performance: F1 Score

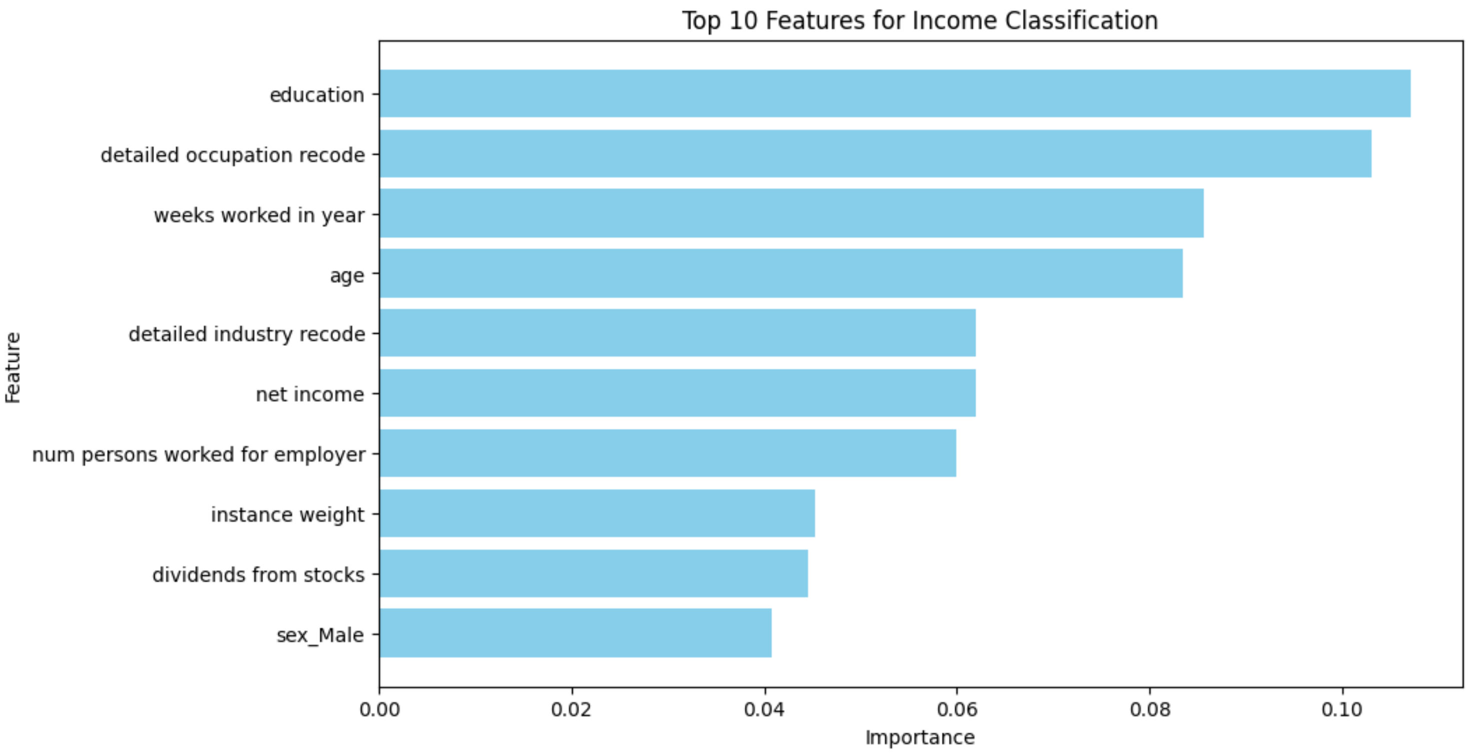
- F1-scores measure how correctly the data have been classified.
- F1-scores for training data and test data are measured.
- Random forest classifier and voting classifier seem to perform the best among all the models on the F1 score metric.
- On further inspection, Random forest classifier seems to perform too well on the training data indicating the model might be overfit.
- Voting classifier seems to be the best model.



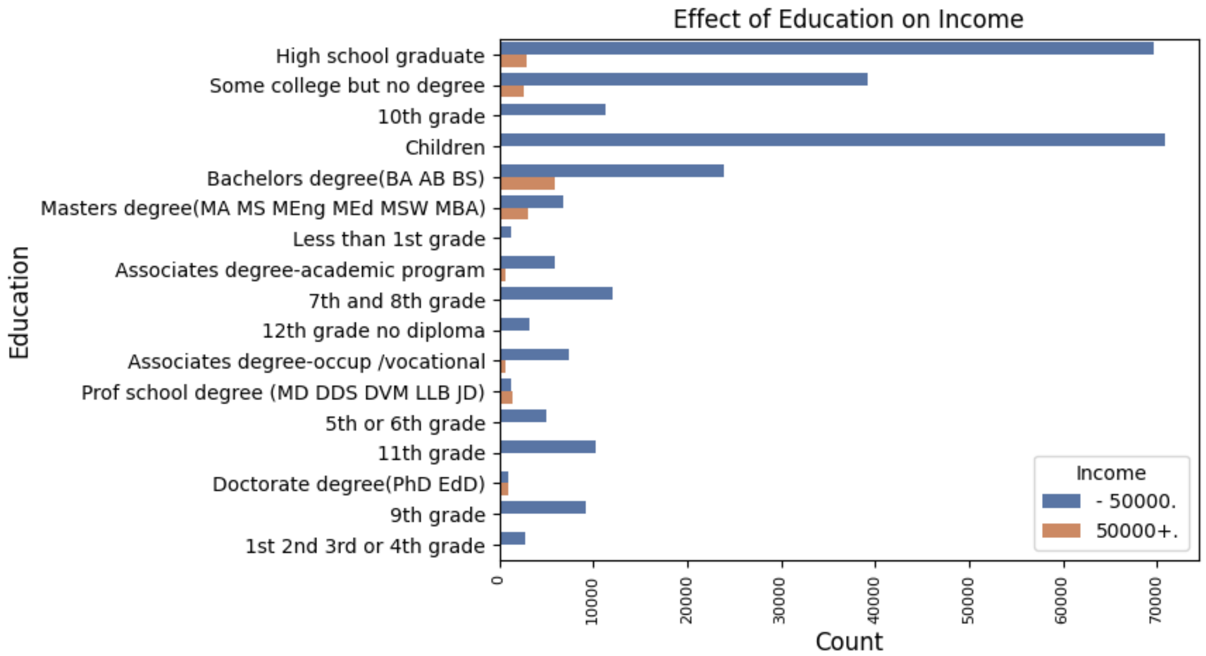
Characteristics of individuals less/more than $50k

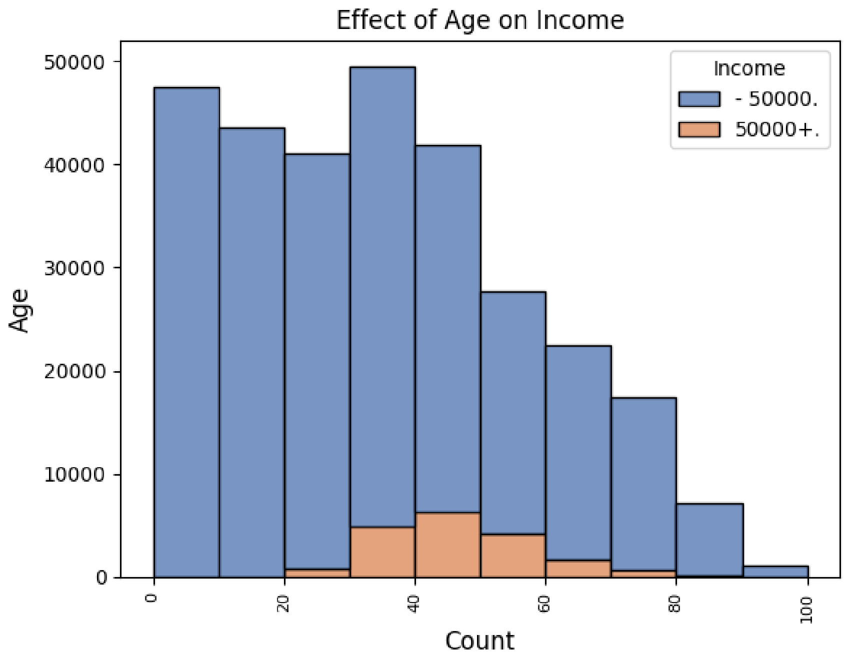
Key Observations

- *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





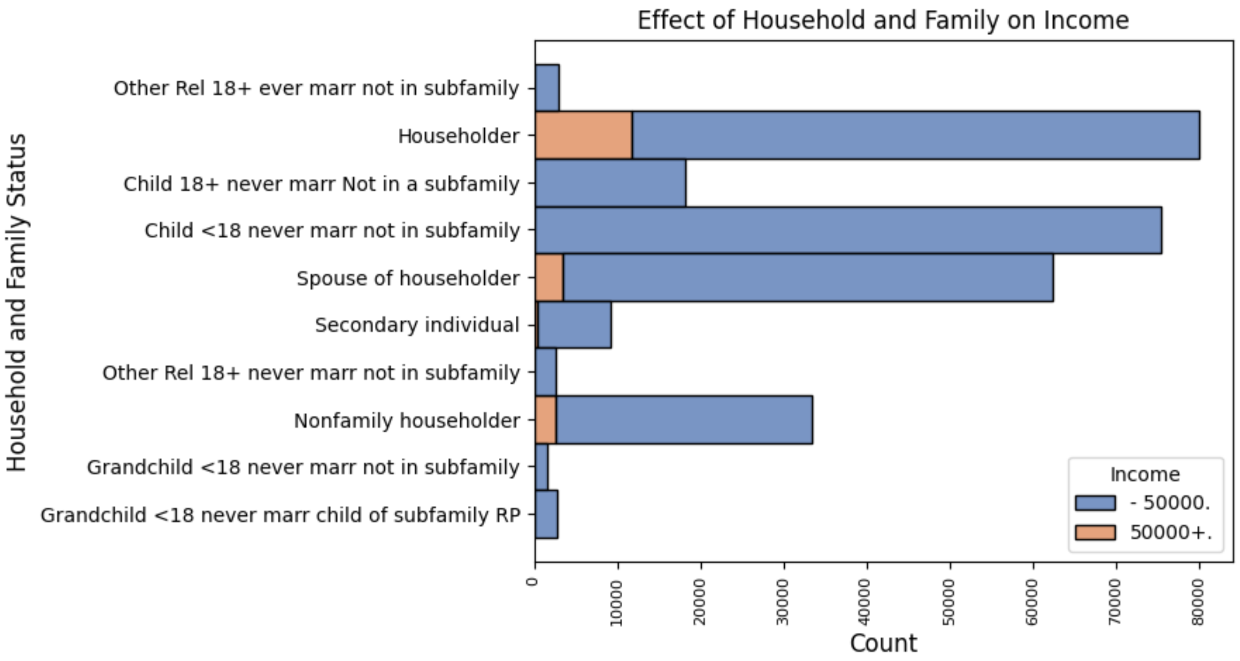
Effect of education on income

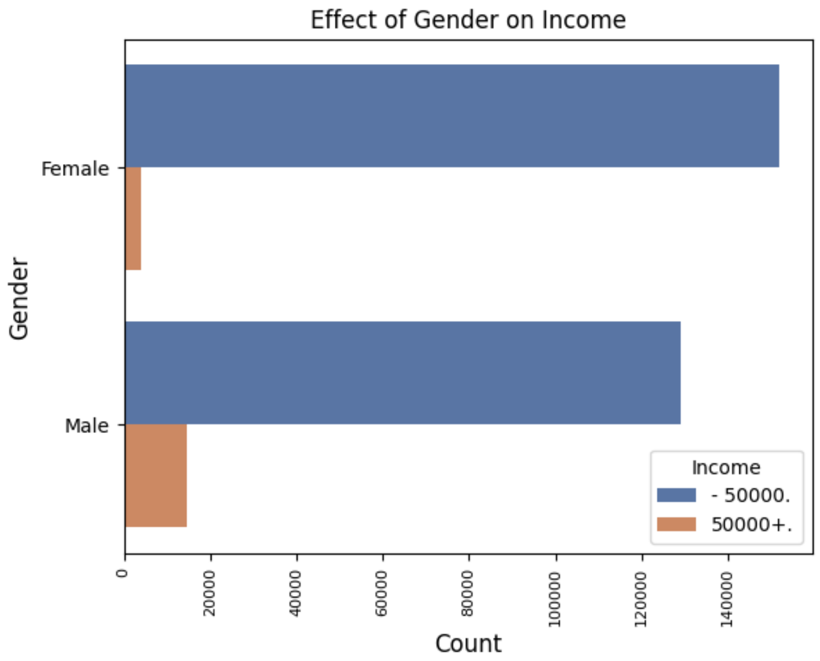
- Most of the people earning > $50k tend to have graduated high school, attended college and have a bachelor’s or master’s degree

Effect of age on income

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

Characteristics of individuals less/more than $50k





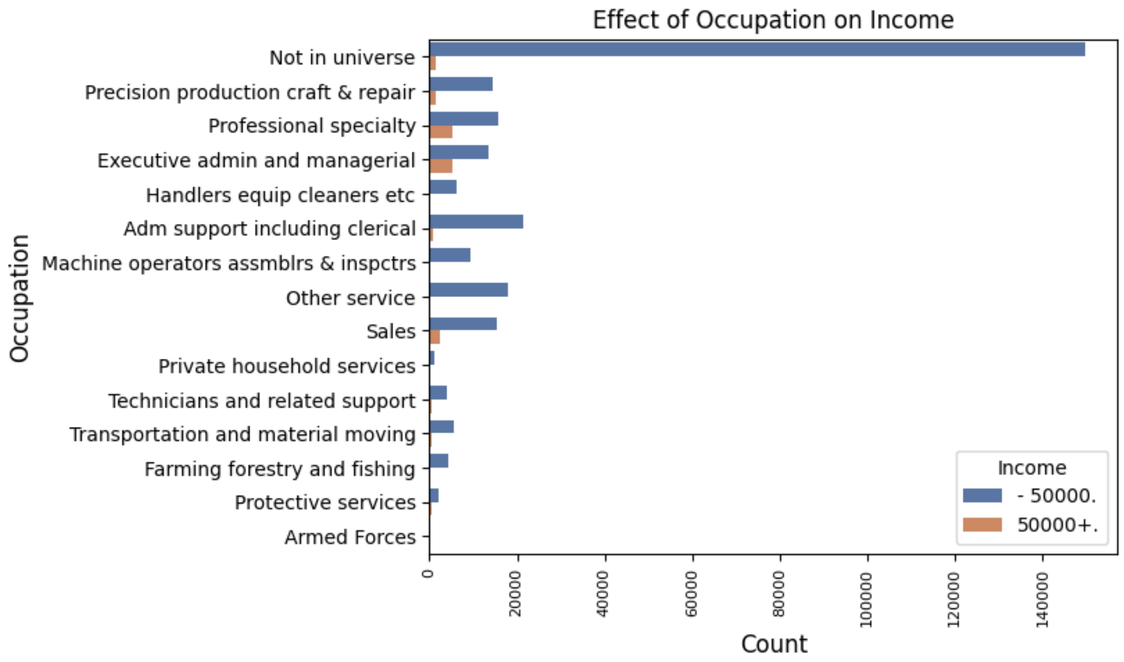
Effect of household and family income

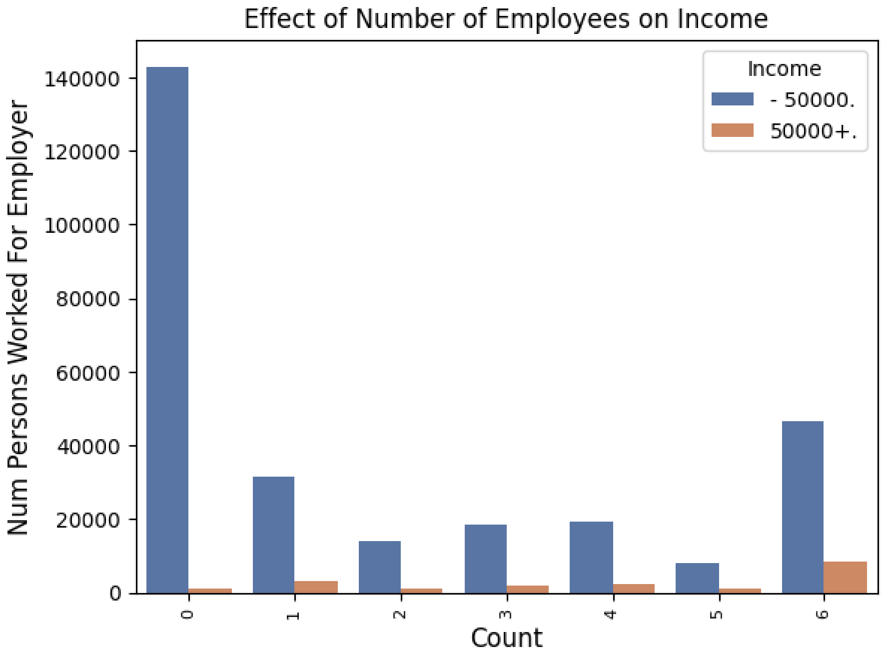
- People earning > $50k tend to be householders or married to one.

Effect of household and family 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 employees

- People earning > $50k tend to also be company owners hiring a lot of workers.

Recommendations & suggested improvements

Data improvements:

- Effect of children seems to sway highly towards low income earners. Perhaps one can make an assumption to consider only 18+ year olds in order to gain more information regarding income.
- Consider selecting fewer children in the dataset. Around 22% of the data set is under 18 years of age. Or collect additional data.
- More attention can be given to feature selection to aggressively group together some categorical data.

Model improvements:

- Data improvements might automatically increase model performance.
- Introduce regularization, or train the model on most relevant features.
- Apply techniques such as RFE to select more relavant features.
- Perhaps spend more time on ensemble learning methods.

Thank you

Back-up

Model performance

Model	Accuracy score (test)	Accuracy score (train)	F1-Score (test)	F1-Score (train)
Logistic Regression	0.8397	0.8612	0.4074	0.8648
Decision Tree Classifier	0.8506	0.8855	0.4188	0.8884
Random Forest Classifier	0.9472	0.9928	0.5834	0.9929
XGBoost Classifier	0.9238	0.9180	0.5063	0.9200
Voting Classifier	0.9305	0.9670	0.5349	0.9675
Ada Boost Classifier	0.8546	0.8739	0.4338	0.8766
NeuralNet	0.8569	0.8866	0.4336	0.8899

Characteristics of individuals less/more than $50k