Predicting Personal Income: US Census Data

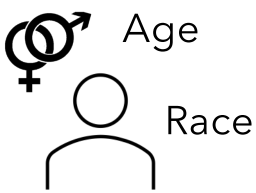
Jed Wingrove

The task and the data

"Identifying **characteristics** that are associated with a person making **more or less than $50,000** per year income"

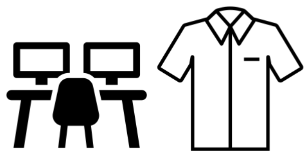
Dataset:

- US Census Data 1994/95
- Results from approx. 300,000 surveyed individuals in the US

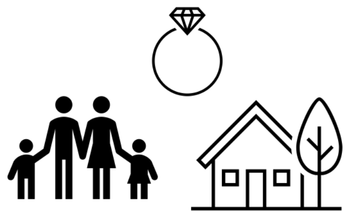


Demographics

Education



Employment



Family/Household

Define the goal

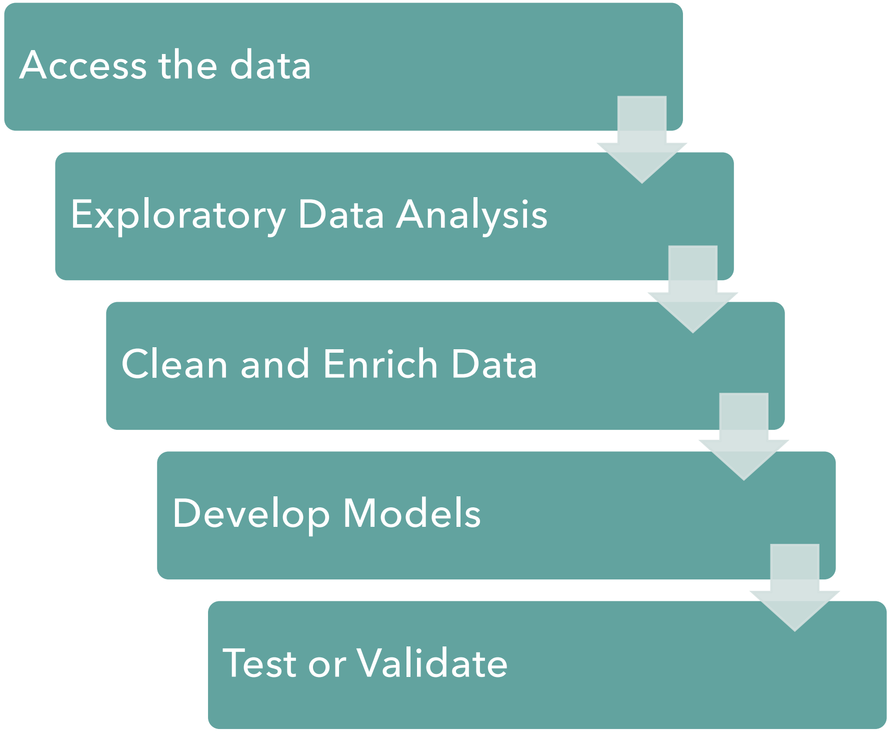
"Identifying **characteristics** that are associated with a person making **more or less than $50,000** per year income"

More or less than $50K

- Binary Classification (0= less than, 1 = More than)

Identifying Characteristics

- Explore and become familiar with the data
- Explainable models and insights



The Data and data cleaning

Dataset:

- Provided in tabular format and was split into two tables:
  - Training Set (used to train the models), n = 199523
  - Testing Set (used to test our models), n = 99762

- Contained 40 different characteristics (features)
  - Label: - $50000 or + $50000 (this is what we are looking to predict)
  - High Earners = more than $50,000
  - Low Earners = less than $50,000

Data Cleaning

Step 1:

- Remove duplicate individuals (46,627 in training set and 20,898 in test set)

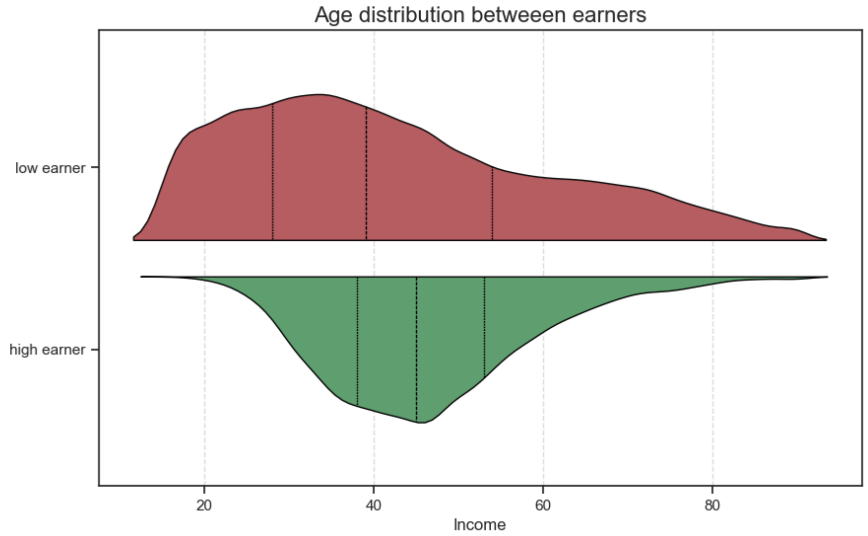
Step 2:

- Identify any conflicting instances (identical characteristics but different label).
- Can't tell which label is correct, so have to remove all conflicting instances.

Data Filtering and EDA

Person's Age

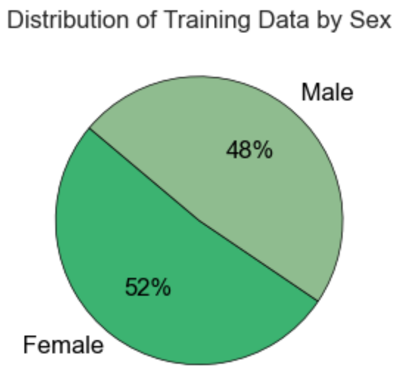
- Continuous Data (actual values)
- Range 0 – 90 years old
- Survey notes described 15 years and older as an adult
- 15 years feasible age to be in employment
- Filtered the data to have a range between 15 – 90 years

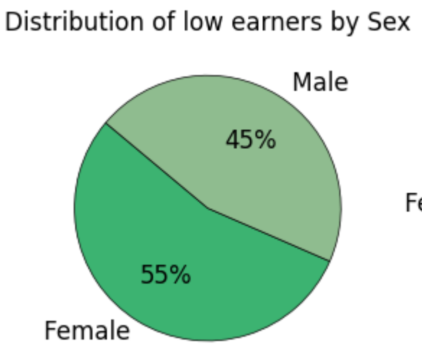


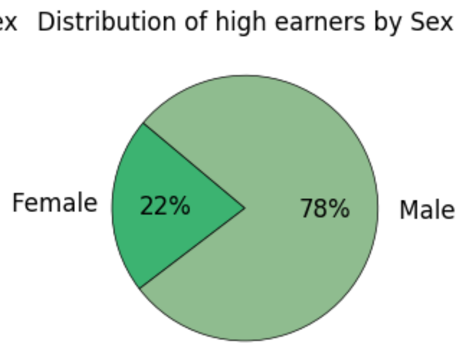
High earners – on average, slightly older
Low earners – bigger spread and variation

High earners – Middle aged, White, Men

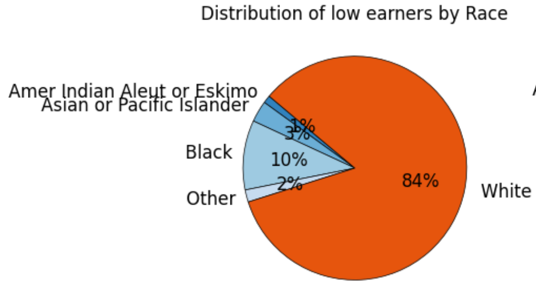
Sex

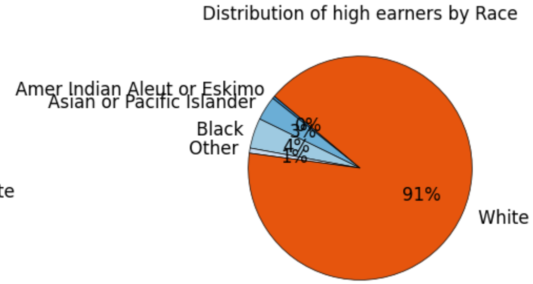


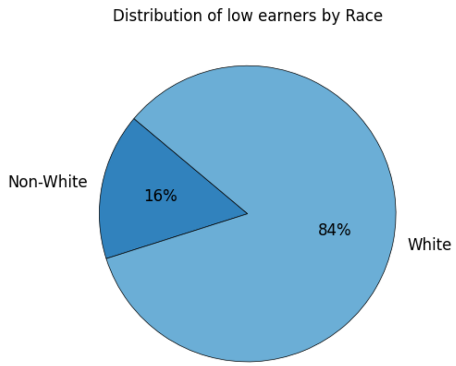


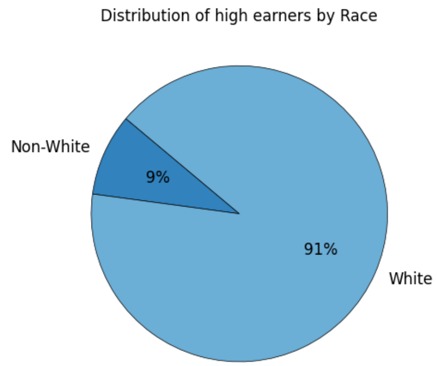


Race



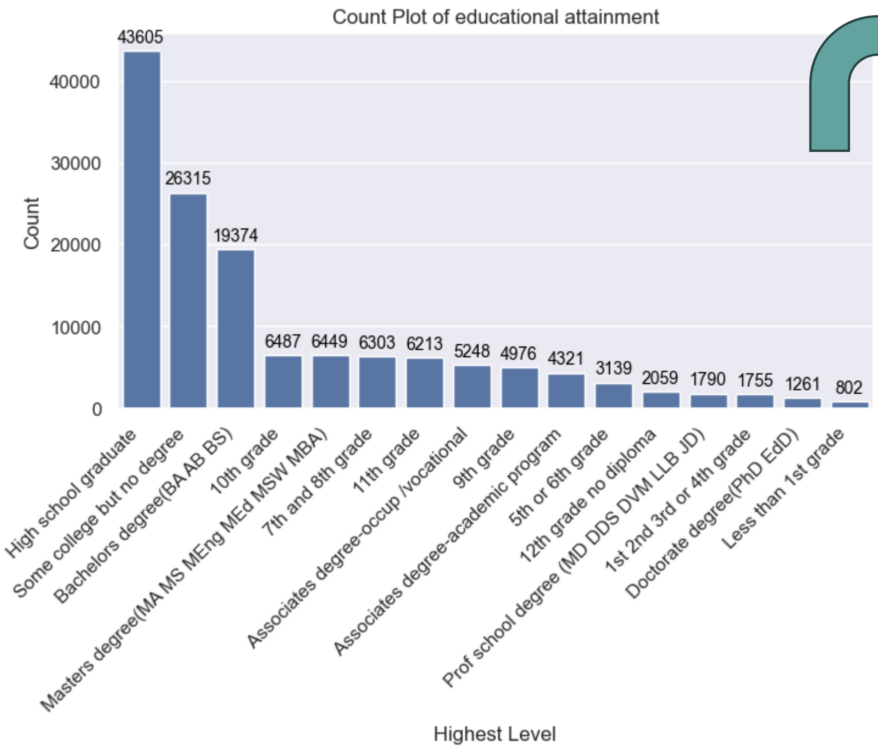






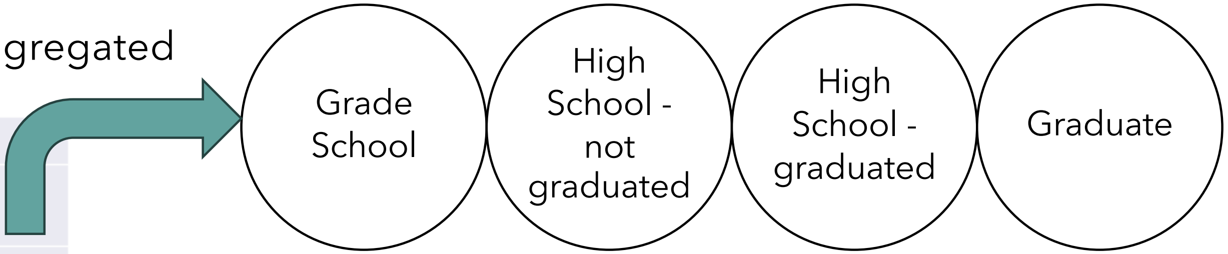
High earners – more qualifications

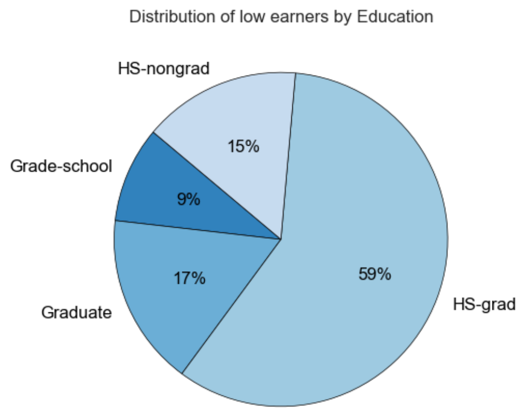
Highest Education Level

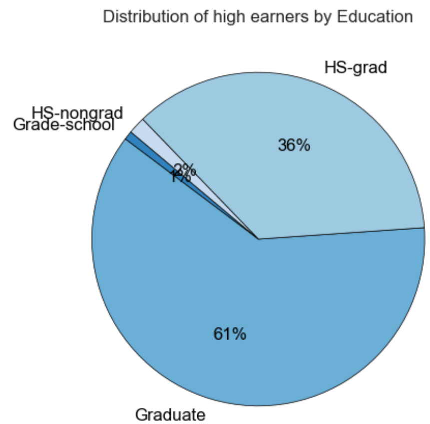


Aggregated

4 Major Categories

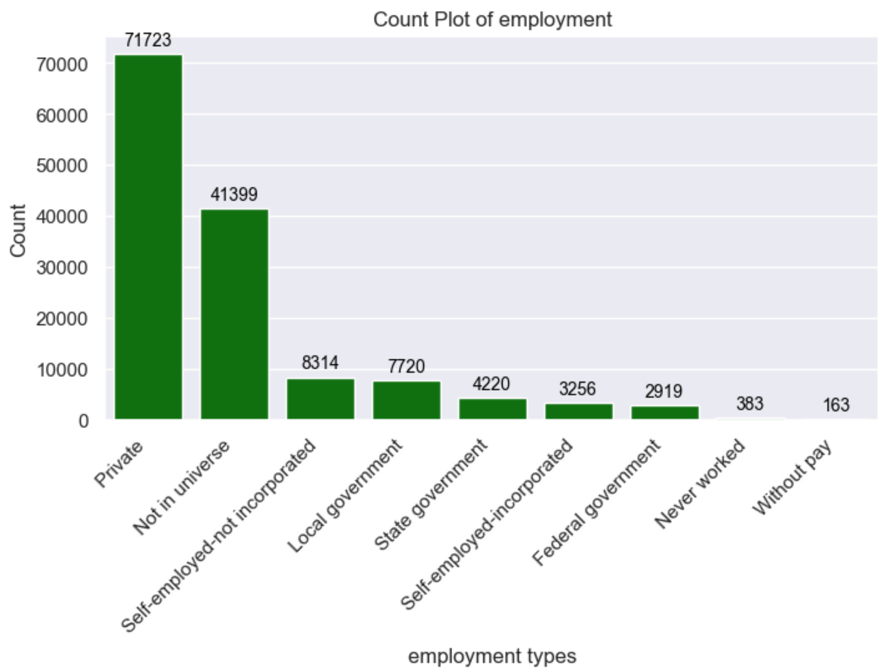


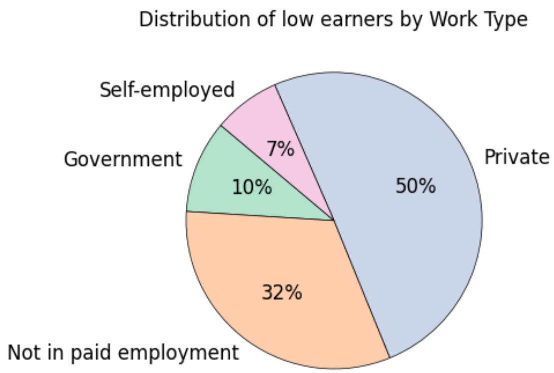


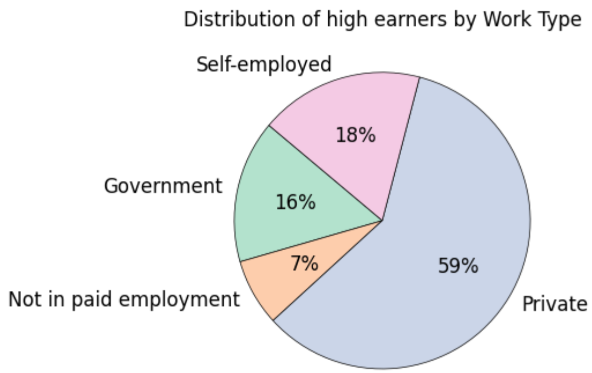


High earners – self-employed, in work?

Employment



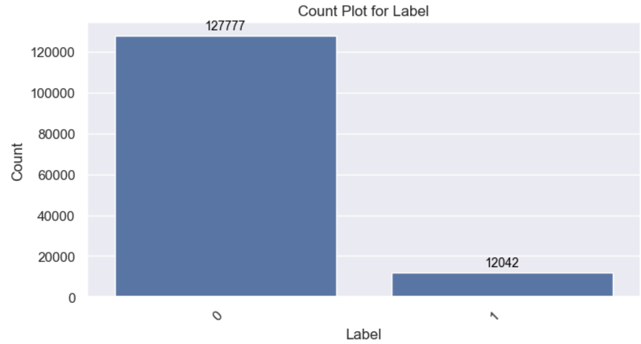




Ready for Modelling – Features, Labels

Feature	Type
Age	Continuous (15-90)
Male	Binary (0,1)
Married	Binary (0,1)
Race White	Binary (0,1)
Education – Grade School	Binary (0,1)
Education – High School (not Graduated)	Binary (0,1)
Education – High School (Graduated)	Binary (0,1)
Education – University Graduate	Binary (0,1)
Employment - Government	Binary (0,1)
Employment – not employed	Binary (0,1)
Employment - Private	Binary (0,1)
Employment – self employed	Binary (0,1)
Householder - Yes	Binary (0,1)
Householder - Live with householder	Binary (0,1)
Householder - Child	Binary (0,1)
Parents US born	Binary (0,1)

Feature	Type
Label	**0=-$50,000, 1= + $50,000**



Class Imbalance in our labels

- Less data to learn the association between features and higher income.
- Choose appropriate models
- Choose appropriate metrics for scoring

Models and Scoring

Classification Models

Logistic Regression
- Simple, interpretable, computationally cheap
- Good for binary classification

Random Forest
- Interpretable, handles all data types
- Non-linear, ensemble method, Voting

Decision Trees
- Interpretable, handles all data types
- Non-linear

Scoring

Precision (0-1)
- A measure of quality

Recall (0-1)
- A measure of quantity

F1 Score (0-1)
- Average between Precision and Recall

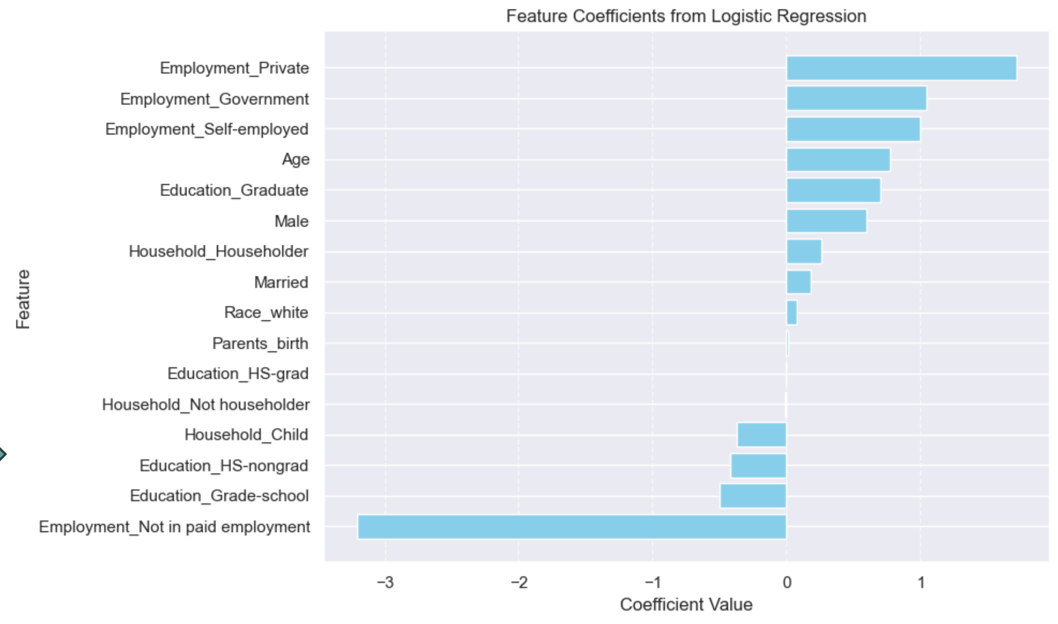
AUC ROC (0-1)
- Single value that represents classifier performance across many thresholds.

Results

	Precision	Recall	F1 Score	AUC ROC
Logistic Regression	0.57	0.27	0.36	0.62
Decision Trees	0.54	0.30	0.38	0.64
Random Forest	0.54	0.31	0.39	0.64

With Class Imbalance Parameter Added

	Precision	Recall	F1 Score	AUC ROC
Logistic Regression	0.24	0.89	0.38	0.82
Decision Trees	0.26	0.80	0.40	0.80
Random Forest	0.28	0.76	0.41	0.79



Insights and closing remarks

"Identifying **characteristics** that are associated with a person making **more or less than $50,000** per year income"

- Data preparation, cleaning, feature filtering and engineering. Typical data science pipeline

- EDA highlighted some interesting characteristics and features likely to be informative of higher income.
  - Slightly older, well educated, employed (self-employed), Male
  - Room for more feature engineering to uncover interesting and insightful characteristics

- Limitations:
  - Class imbalance discovered - modelling trickier, (imbalance-learn, SMOTE)
  - Understanding of the data and some of the features and how the US census survey system works.