Income Projections in America

A Data-Driven Approach to Predicting Income Trends in the U.S

Author: Constance Gontier

Tuesday, April 2nd 2024

CONTEXT

Goal

- Examine demographic characteristics of subpopulations across the US
- Identify the key influential factors on income prediction to allocate funding
- Predict whether an individual earns more or less than 50'000 $ a year

Data

- A dataset provided by the US Census Bureau
- ~ 300'000 individuals

PROJECT STEPS

Data Analysis

Numerical insights
Graphical representations

Data Preparation

Data preprocessing
Feature engineering

Data Modeling

Data Selection
Model implementation

Model Assessment

Performance comparison
Choice of the best model

Results

Key factors of influence
Future work

Data Analysis

- A first glance at the data
- The influence of age
- The influence of sex
- The influence of race
- The influence of industry

Tuesday, April 2nd 2024

Data Analysis - A first glance at the data

The training dataset

- Population: **199 523 individuals**
- Attributes: **9 continuous** columns and **32 categorical**
- Income Distribution:
  - **6.2%** earn more than 50'000$
  - **93.8%** earn less than 50'000$

The testing dataset

- Population: **99 762 individuals**
- Same Income Distribution

Data Analysis - The influence of age

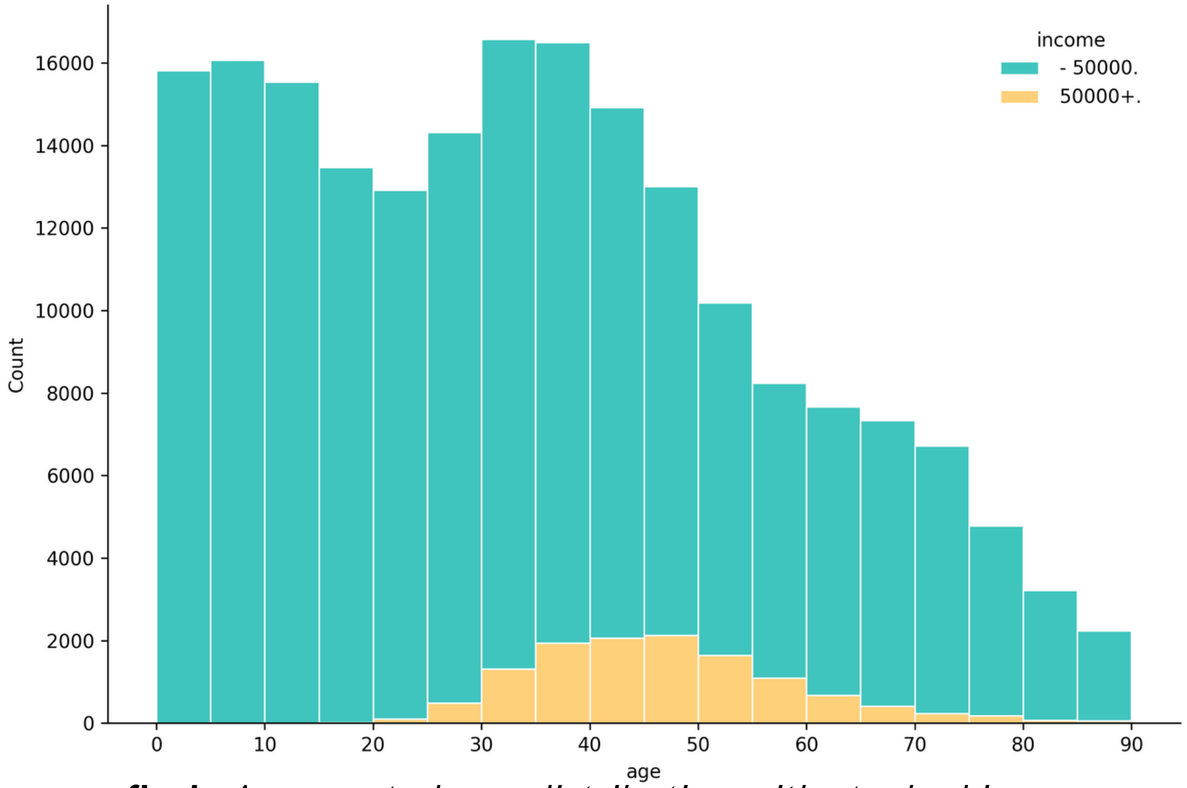


fig 1: *Aggregated age distribution with stacked income*

Remarks

- Distribution of higher earners centered around **age 45**
- Dip in population around **age 20**

Data Analysis - The influence of sex

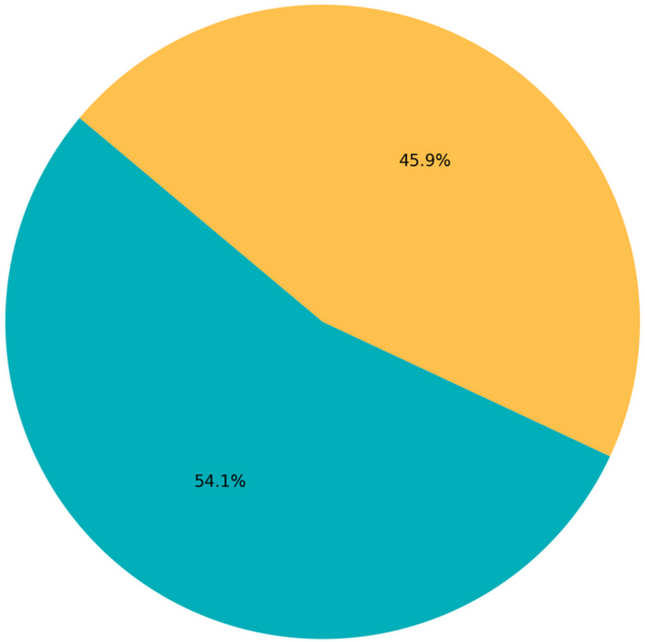


fig 2.a: Pie chart of sex distribution for lower income

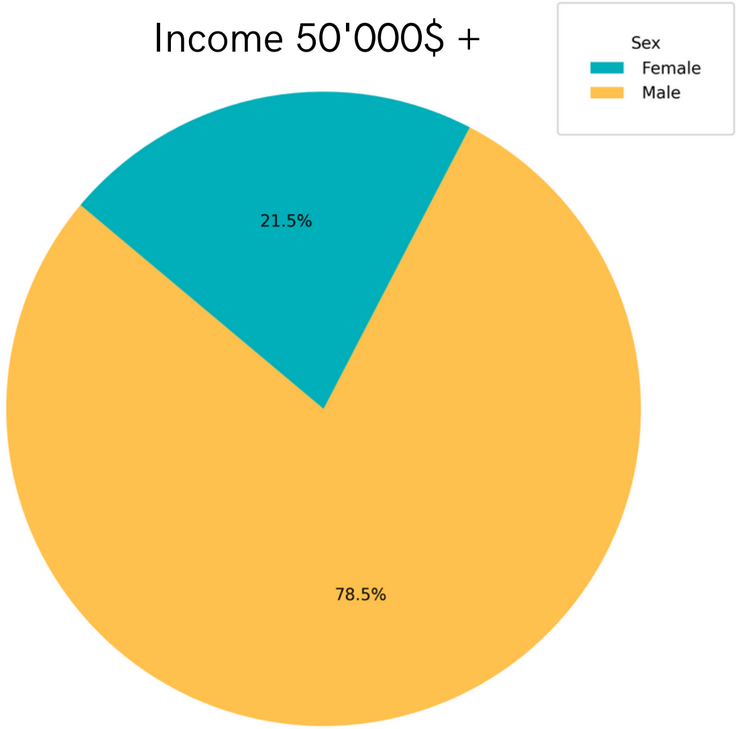


fig 2.b: Pie chart of sex distribution for higher income

Remarks

- **Balanced** sex distribution for an income **lower** than 50'000$
- Significantly **higher male population** for an income **higher** than 50'000$

Data Analysis - The influence of race

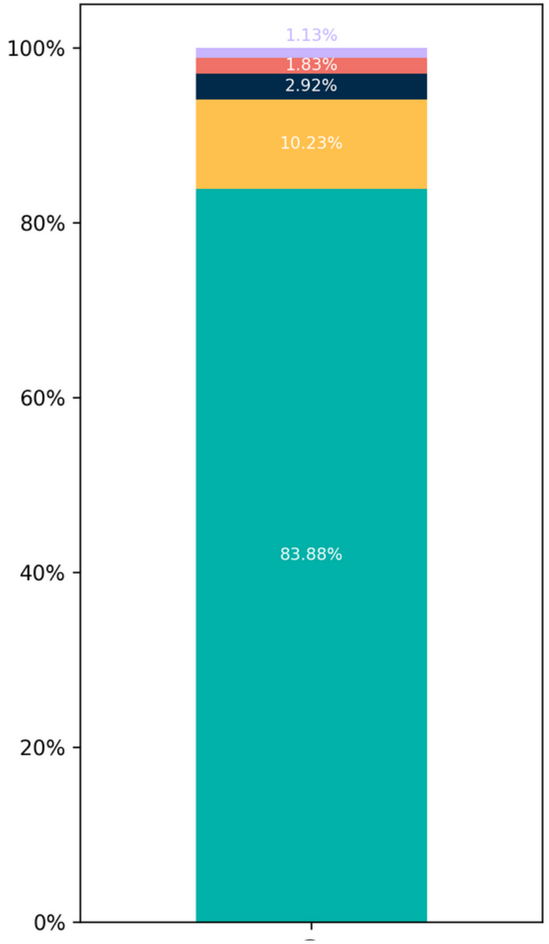


fig 3.a: *Grouped race distribution*

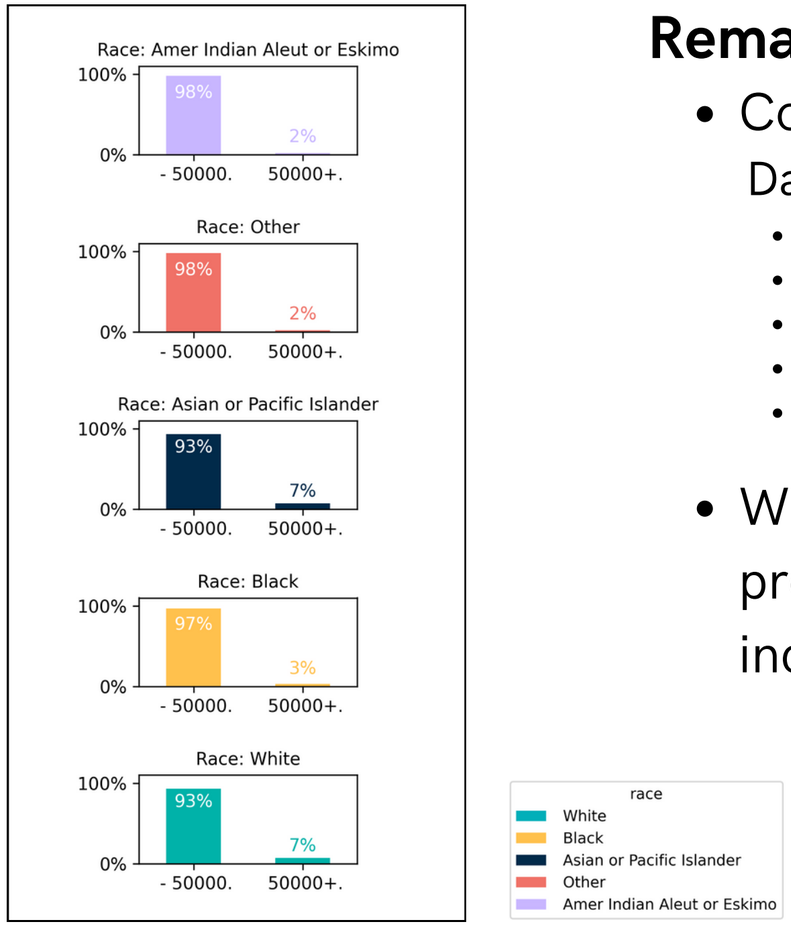


fig 3.a: *Proportion of population for each income category by race*

Remarks

- Comparison with actual U.S population

Dataset population:	US population*:
• White: 83.9%	• White: 60.1%
• Black: 10.2%	• Black: 12.2%
• Asian/Pacific: 2.9%	• Asian/Pacific: 5.6%
• Other: 1.8%	• Other: 21.4%
• Native: 1.1%	• Native: 0.7%

- White and Asian/Pacific have a higher proportion of population with high income (**7%** vs. 2-3%)

*__Source:__ US Census Bureau

Data Analysis - The influence of industry

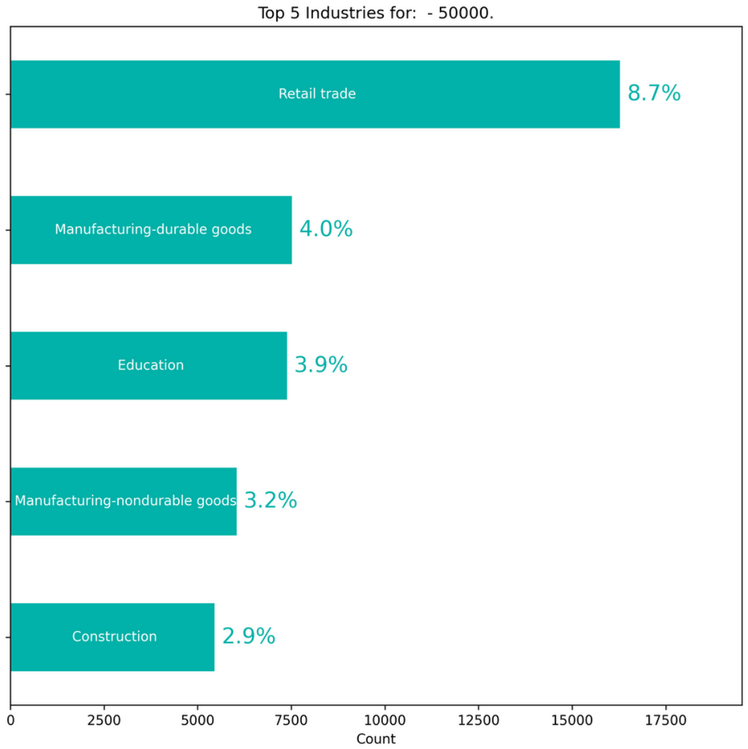


fig 4.a: *Top 5 industries with proportion (%) in lower income values (- 50'000$)*

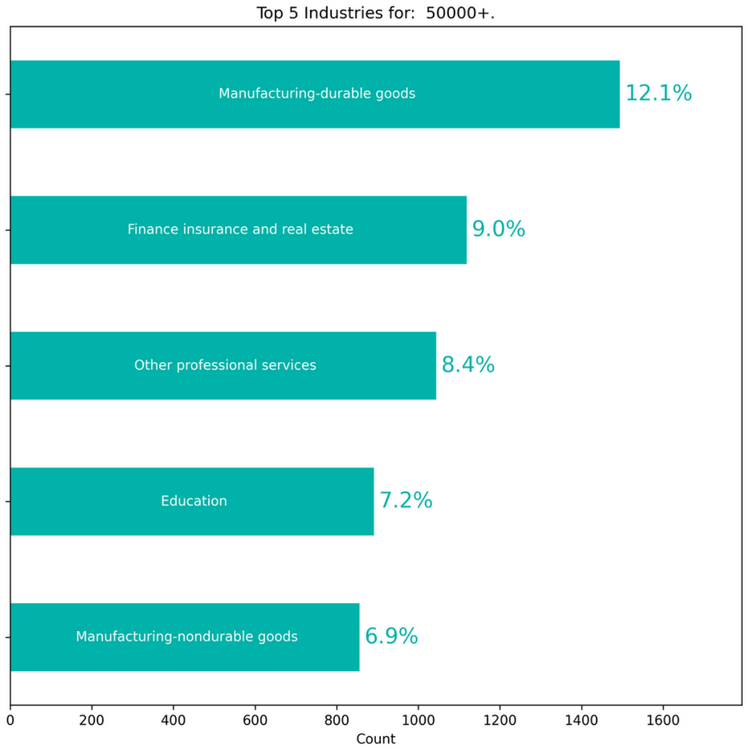


fig 4.a: *Top 5 industries with proportion (%) in higher income values (50'000$ +)*

Remarks

- Industry with highest proportion in each income category:
  - - 50'000$: **Retail trade** (8.7%)
  - 50'000$ +: **Manufacturing-durable goods** (12.1%)
- Proportion percentage for lower income is evenly distributed
- Proportion percentage for higher income has a focus on certain industries (fig 2b)

Data Preparation

- Cleaning the data
- Engineering the features
- Creating data sub-sets

Tuesday, April 2nd 2024

Data Preparation - Cleaning the data

Missing Data

- *Missing* or *Not in Universe (or Children)* values represent 2 076 058 cells -> **32.93%** of the dataset
- Columns with
  - More than 40% missing -> **Remove 14 columns**
  - Less than 40% missing -> **Imputation on 3 columns**

Duplicate Data

- 4316 duplicate rows -> **2.16%** of the dataset
- **Removed**

Data Preparation - Engineering the features

Feature Creation

- After cleaning:
  - **8 continuous** columns (age, wage…)
  - **19 categorical** columns (education, sector…)
- From 3 columns of numerical data -> **Binning**
  - *Age*, *Wage per hour*, *Weeks worked in year*
- From 2 columns of numerical data -> **Combining** for new feature
  - *Total wage for year* = *Wage per hour* x *Weeks worked in year*

Feature Engineering

- Continuous columns:
  - **Scaling** (Standardization)
- Categorical columns:
  - **One hot encoding**

Data Preparation - Creating data sub-sets

Created 5 New Datasets

- **All features** (369 features)
- **Best features** (108 features)
  - Based on correlation for continuous features -> 8 continuous features
  - Based on Chi2 for categorical features -> Isolate 100 best features
- **PCA**
  - Retain 90% of variance
  - 43 dimensions kept
- **Downsampling**
  - Original dataset size: 195'207
  - Balanced dataset size: 24'764
- **Oversampling**
  - Original dataset size: 195'207
  - Balanced dataset size: 365'650

Data Modeling

- Random Forest
- Logistic Regression
- XG-Boost
- Neural Network

Tuesday, April 2nd 2024

Data Modeling - Random Forest

Principle

- Utilises a multitude of decision trees to improve accuracy and control over-fitting through bagging and feature randomness
- Each tree votes for the most popular class -> the majority vote determine the final prediction

	All features	Top Features	PCA	Downsampled	Oversampled
f1 on evaluation set	0.52	0.53	0.48	0.84	**0.98**
f1 on testing set	0.52	0.54	0.48	0.54	**0.56**

Data Modeling - Logistic Regression

Principle

- Statistical method for predicting binary outcomes
- Estimate probabilities using a logistic function, assuming a linear relationship between input features

	All features	Top Features	PCA	Downsampled	Oversampled
f1 on evaluation set	0.53	0.52	0.44	**0.85**	**0.85**
f1 on testing set	**0.51**	**0.51**	0.44	0.47	0.47

Data Modeling - XGBoost

Principle

- A novel tree learning algorithm
- A gradient boosting library that uses regularization to enhance performance and prevent overfitting

	All features	Top Features	PCA	Downsampled	Oversampled
f1 on evaluation set	0.58	0.58	0.52	0.86	**0.90**
f1 on testing set	**0.58**	0.57	0.52	0.48	0.53

Data Modeling - Neural Network

Principle

- Layers of neurons with activation functions
- Capable of modeling complex non-linear relationships by adjusting weights during training
- Utilizes a forward pass through layers with ReLU and sigmoid activations to predict probabilities

	All features	Top Features	PCA	Downsampled	Oversampled
f1 on evaluation set	0.55	0.56	0.47	0.85	**0.87**
f1 on testing set	**0.56**	0.53	0.50	0.42	0.48

Model Assessment

- Comparing our models
- Selecting the best model

Tuesday, April 2nd 2024

Model Assessment - Comparing our models

Metrics evaluation: Unbalanced dataset -> f1 score

	Random Forsest	Logistic Regression	XGBoost	Neural Network
Best perform on	Oversampled dataset	All and Top features dataset	All features datasets	All features dataset
f1-score	0.56	0.51	**0.58**	0.56
Training time	11.6 seconds	4.7 seconds / 3.2 seconds	**1.3 seconds**	36 seconds
Testing time	1.1 seconds	0.3 seconds / 0.3 seconds	**0.2 seconds**	0.1 seconds

Model selection

Based on f1 sore and run times:

- **XGBoost** is the best option

Results

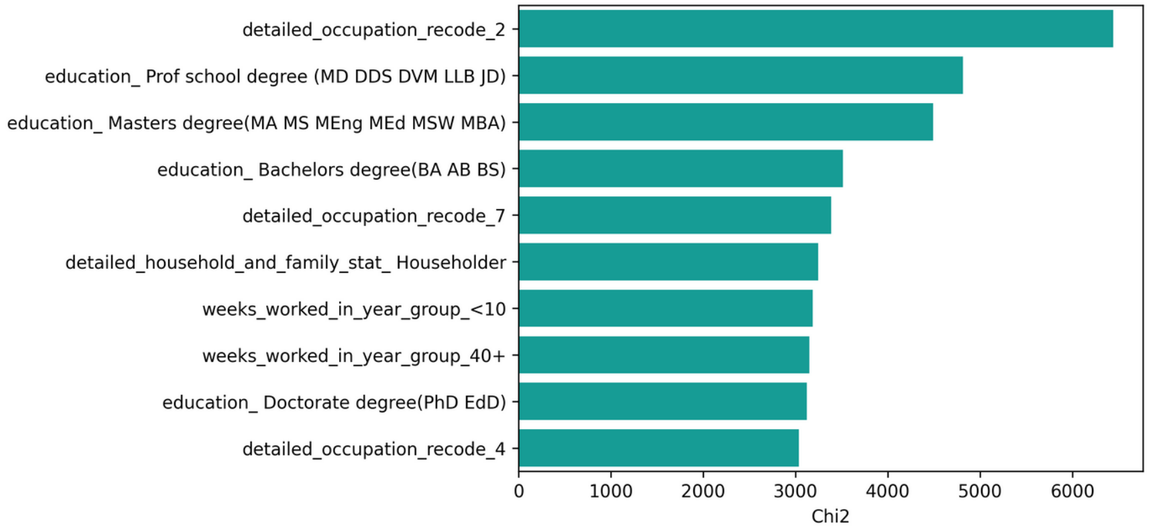
- Key features
- What next?

Tuesday, April 2nd 2024

Results - Key features

Statistical Method

- Chi2 gives us a list of Chi scores



Using a model

- Random Forest gives us an order of feature importance

Results - What next?

Potential Ideas

- **Dive into *Not in Universe* Values**: Analyze patterns behind "Not in Universe" entries to identify potential biases or systemic issues, potentially leading to targeted models for adults and children
- **Segmented Modeling:** Implement separate models for different demographic groups to address dataset imbalance and tailor predictions to specific populations (adults vs children for example)
- **Incorporate External Data**: Enrich the analysis with external datasets to add context, or to collect more data for the underrepresented class
- **Explore Other Models:** Test other more complex Neural Networks

Thank You!

- Questions?

Tuesday, April 2nd 2024