Income Bias Analysis

"Using Data Science Methodology"

Team 3

Donghwan Kim	A0231887U
Kwang Hun Lee	A0231958W
Keyi Chen	A0218873U
Zhenzhou Tian	A0231909A

Contents

- ✓ Background
- ✓ Problem Definition
- ✓ Data
- ✓ Modeling
- ✓ Interpretation

* Analysis.(1): 1st problem
* Analysis.(2): 2nd problem

Background

[Key Question]

“What decides individual’s income?”

In other words,

- Is there any **meaningful association** between one’s income and one’s personal information, such as gender, education level, etc.?
- Can we identify the **factors** that might contribute to **income bias**?

For ”Social science researchers & Policymakers”

Problem Definition

"Find Meaningful Association" & "Predict Income Level"

>> **Identifying factors** that contribute to income bias: Explore features

>> **Prediction** on whether a person will make over $50,000 income a year
(* Income **Binary** Category: [Over USD 50,000] vs. [Under USD 50,000])

Selected Features (Potential Factors)

>> Top.3 Features based on **Feature Importance**?!

1) **fnlwgt**: Final Weight based on demographic characteristics & number of responses
2) **age**: Age of individual
3) **hours-per-work**: Working hours per work

Selected Machine Learning Methods

1) Compared total 15 different ML models
>> pycaret.classification Module
>> setup() & compare_models()

3 Selected Models (Best AUC: 0.91)

>> Logistic Regression
>> LightGBM, XGBoost

Data: Summary

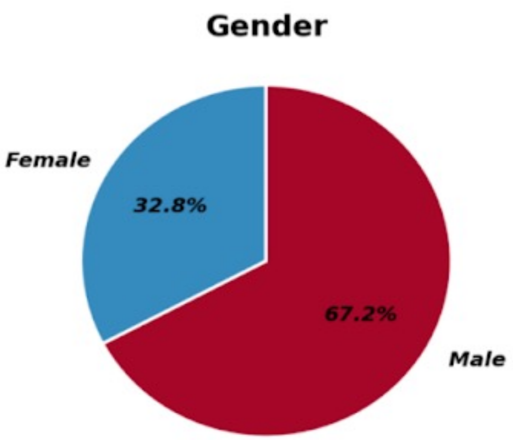
[adult_income.csv]

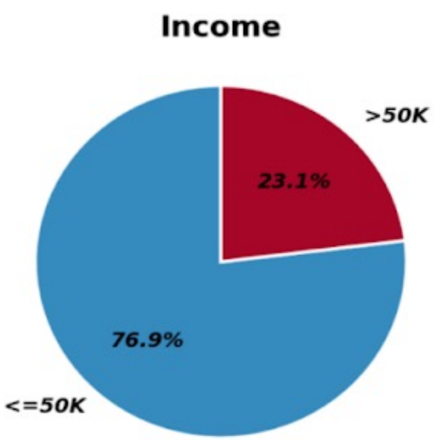
>> shape: 32,561 rows x 15 columns
>> # null values: 2,398 rows with at least 1 null value

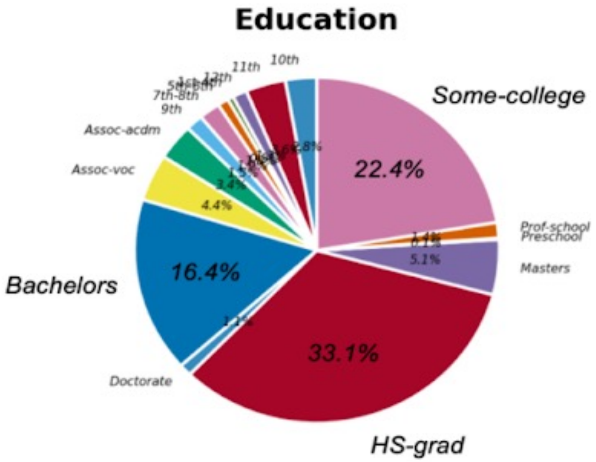
< # Null values >

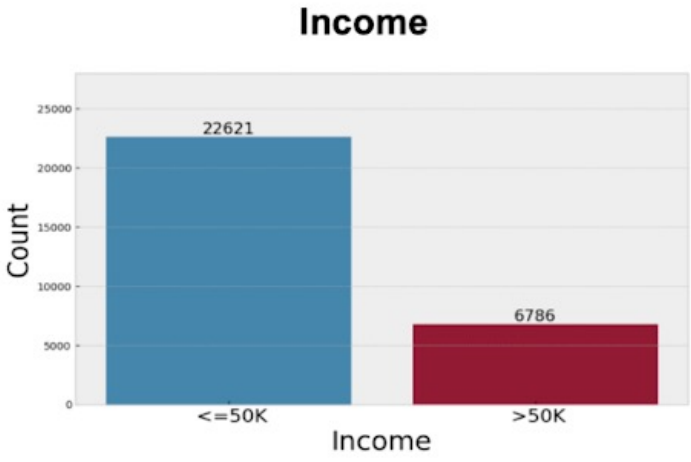
Column	Work-class	Occupation	Native-country
# Null	1,836	1,843	582

	Age	Work-class	fnlwgt	Education	Education-num	Marital-status	Occupation	Relationship	Race	Sex	Capital-gain	Capital-loss	Hours-per-week	Native-country	Income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K



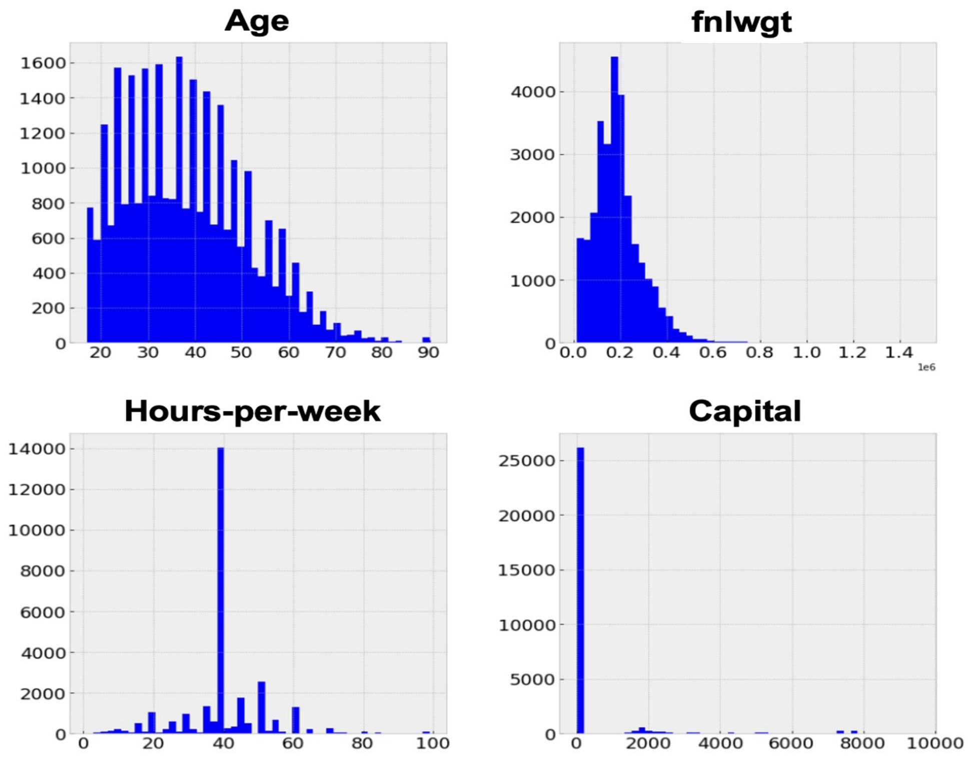




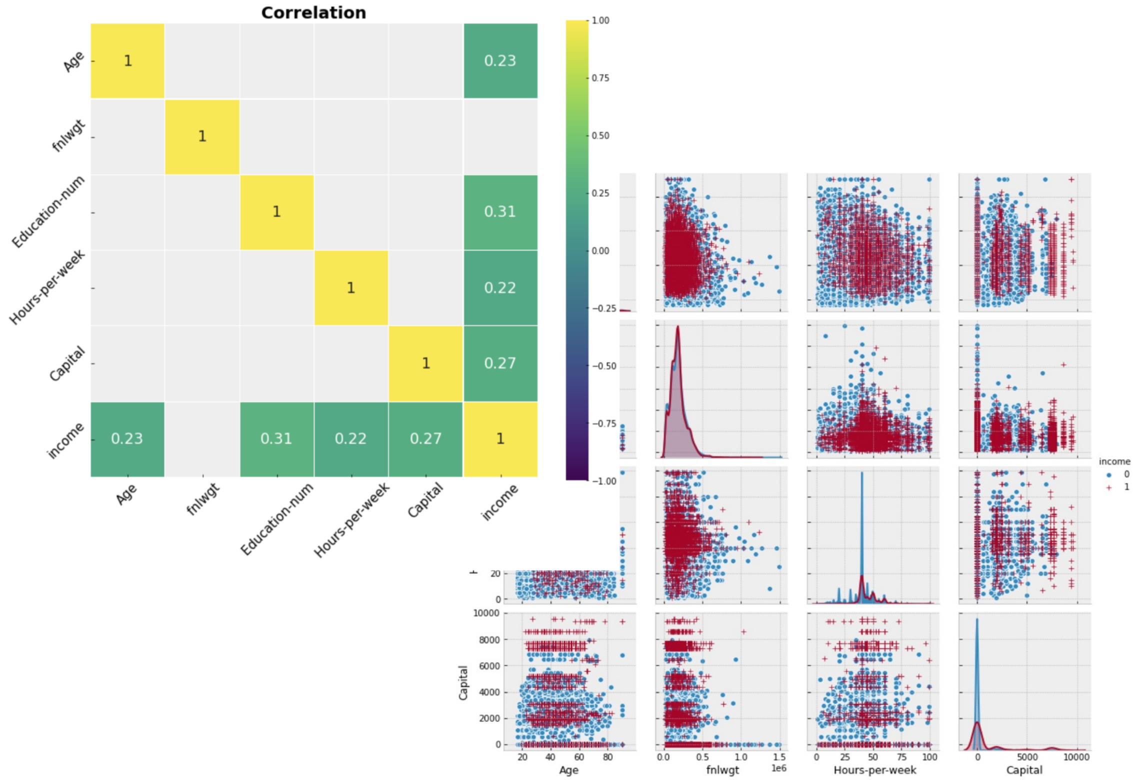


Data: Exploration

Distributions: "Data makes sense!"



Correlations with Numerical data



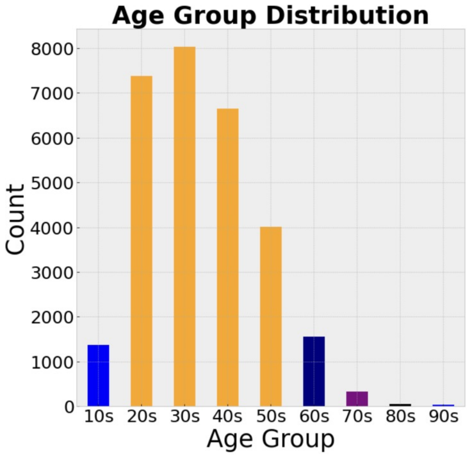
>> Identical / Similar shape in Age & fnlwgt variables

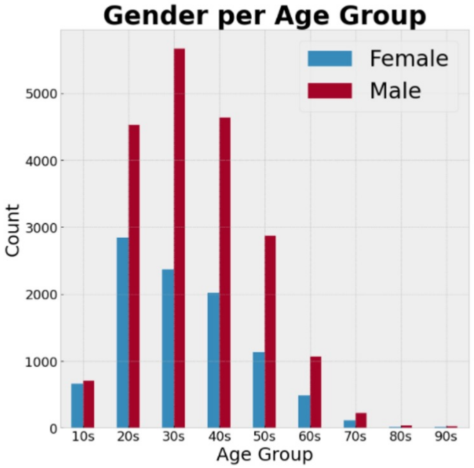
>> Usually people works 40 hours a week >> mod value of 40 for 'working hour per week' variable

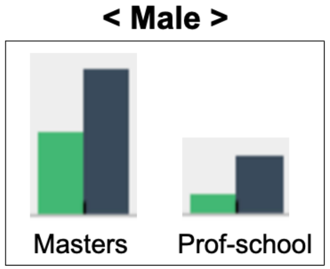
>> Majority (90.83%) of people do not invest in capital assets (e.g. stocks, real estate)

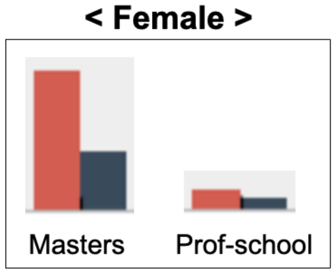
Data: Exploration

Distribution: Age Band & Sex





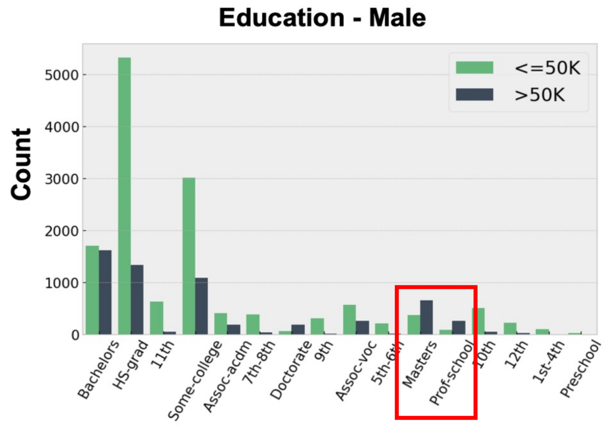


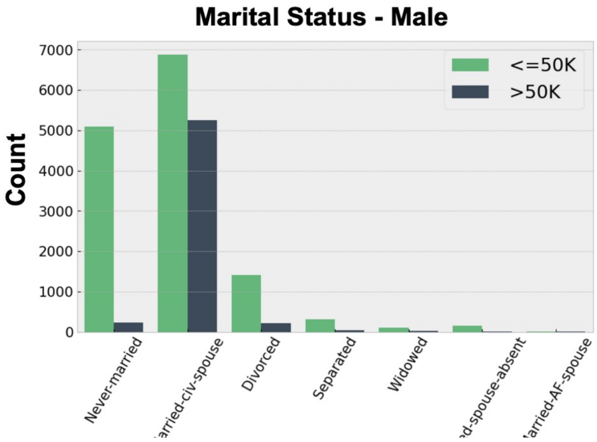


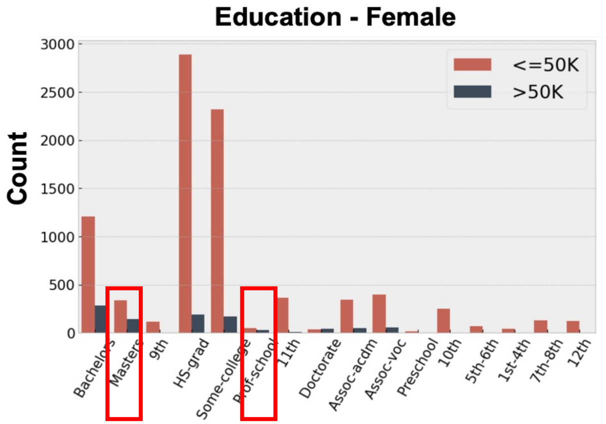
>> 2 Education Level: Masters & Professional school

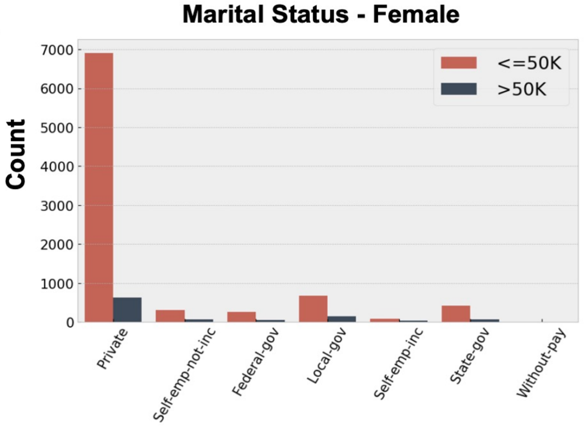
>> **<u>Opposite</u>** Trend according to **<u>Sex</u>**

Counts by Income Per Categorical Column









Data: Pre-processing

1. Duplicated Rows

>> Removed 24 duplicated rows
   * Shape: (32,561, 15) → (32,537, 15)

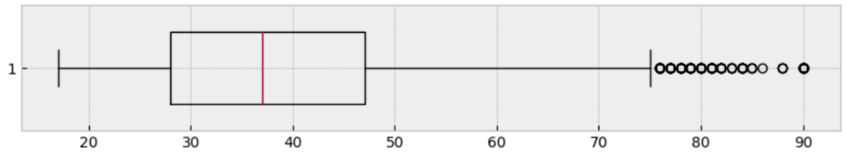
< # Null values for 3 columns >

Column	Work-class (WC)	Occupation (OC)	Native-country (NC)	Total
# Null	1,836	1,843	582	4,261
WC & OC	1,836			(1,836)
OC & NC		27		(27)
WC & NC			27	(27)
WC & OC & NC	27			27
Total				2,398

2. Missing Value Imputation

>> Listwise Deletion: 3 columns have null values & 4.36% of entire dataset are missing on average

Detecting outliers using Boxplot (Age)



3. Outlier Imputation

>> Removed outliers based on 'box and whisker plot'
   * less then lower outer fence(=Q1-3*IQ) or more than 'upper outer fence(=Q3+3*IQ)

>> Removed people with '99,999' value of capital_gain variable. (Negligible)
   * Total 159 people out of population (= 0.4%) got more than 99,999 USD from capital assets.

Modeling: Feature Engineering

1. <u>Categorical Values</u>

>> One-Hot Encoding (e.g. Marital Status, Occupation, Sex, Race)

2. <u>New Feature Generation</u>

>> Merging/Combination of features (e.g. Capital = Capital Gain - Capital Loss)

3. <u>Redundant Features</u>

>> Dropped Education feature (e.g. 1:1 mapping >> Education vs. Education number)

4. <u>Level Unbalance</u>

>> Categorical Values (e.g. ‘Native-country’ column >> US(89%) vs Others(11%))

5. <u>Normalization</u>

>> Z-Score Normalization on selected features (e.g. numerical columns: fnlwgt, capital, working hours)

Modeling: Feature Selection

Feature Importance

```
* LightGBM Feature Importance
model = lgb.LGBMClassifier()
model.feature_importances_
```

```
* Tuned LGBM Feature Importance
best_model = grid_search.best_estimator_
best_model.feature_importances_
```

```
* XGBoost
model_xgb = XGBClassifier(random_state=123)
model_xgb.feature_importances_
```

Model	Top10 Features	Importance
Xgboost	fnlwgt	817
	Age	680
	Hours-per-week	484
	Education-num	310
	Capital	271
	Work-class_Private	79
	Marital-status_Married-civ-spouse	70
	Occupation_Exec-managerial	51
	Occupation_Adm-clerical	48
	Relationship_Not-in-family	47

Model	Top10 Features	Importance
LightGBM	Hours-per-week	788
	Age	666
	fnlwgt	609
	Capital	521
	Education-num	309
	Occupation_Adm-clerical	62
	Occupation_Machine-op-inspct	58
	Occupation_Craft-repair	58
	Occupation_Other-service	54
	Occupation_Transport-moving	52

Modeling: Best Models

< Preliminary Model Comparison >

Model	Notation	Accuracy	AUC	F1
Light Gradient Boosting Machine	lightgbm	0.8698	0.9248	0.7164
Extreme Gradient Boosting	xgboost	0.8662	0.9217	0.7103
Gradient Boosting Classifier	gbc	0.8612	0.9160	0.6813
Ada Boost Classifier	ada	0.8542	0.9101	0.6781
Random Forest Classifier	rf	0.8501	0.8998	0.6736
Linear Discriminant Analysis	lda	0.8360	0.8891	0.6325
Ridge Classifier	ridge	0.8350	0.0000	0.6076
Extra Trees Classifier	et	0.8305	0.8774	0.6404
Decision Tree Classifier	dt	0.8020	0.7392	0.6067
Logistic Regression	lr	0.7872	0.5785	0.3298
Naive Bayes	nb	0.7862	0.8039	0.3372
K Neighbors Classifier	knn	0.7640	0.6495	0.3858
Dummy Classifier	dummy	0.7513	0.5000	0.0000
Quadratic Discriminant Analysis	qda	0.6380	0.5822	0.3839
SVM - Linear Kernel	svm	0.4615	0.0000	0.3473

>> pycaret package automatically run & compare different models using compare_models function

< Actual Model AUC Comparison >

1) Split & Balance

>> Split: Train 70%, Test 30%
>> random_state = 5151
>> Balance: SMOTE

2) Tried 5 different models

>> KNN / Random Forest / Logistic Regression / LightGBM / XGBoost

3) Model Selection

>> Selected 3 best models
>> Highest ROC-AUC score from LightGBM

Interpretation

[Analysis.1] What factors influence the income bias?

>> From EDA based on Correlation, ML Feature Importance and Visualizations
  * Main factors: **Age**, **working hours**, Education-level, capital margin
  * Minor factors: **Occupation**, Marital status of individual, **sex**

[Analysis.2] How accurately can we predict new person's income bias?

>> If we know one's particulars, we can predict if he/she makes over 50,000 USD or not by up to **90% of AUC**, on average.

[Deployment] Trend Reporting & Policy Proposal

>> We can capture trends as we get more census data over time.
>> ***Policymakers*** can develop
  incentives *to boost* ***women's career*** *considering their* ***education level***.

References

* ‘Bias’ in Social Science
Hammersley, M. and Gomm, R. (1997). Bias in Social Research. Sociological Research Online, [online] 2(1), pp.7–19. Available at: https://journals.sagepub.com/doi/full/10.5153/sro.55.

* Light Gradient Boosting Machine
neptune.ai. (2020). Understanding LightGBM Parameters (and How to Tune Them). [online] Available at: https://neptune.ai/blog/lightgbm-parameters-guide.

* Light Gradient Boosting Machine
lightgbm.readthedocs.io. (n.d.). Parameters Tuning — LightGBM 3.3.2.99 documentation. [online] Available at: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html.

* Outlier Definition in Box and Whisker Plot
Nist.gov. (2019). 7.1.6. What are outliers in the data? [online] Available at: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm.

THANK YOU