INCOME LEVEL PREDICTION

Xin Tang
DSC 630
Professor Andrew Hua
Bellevue University

Problem Statement

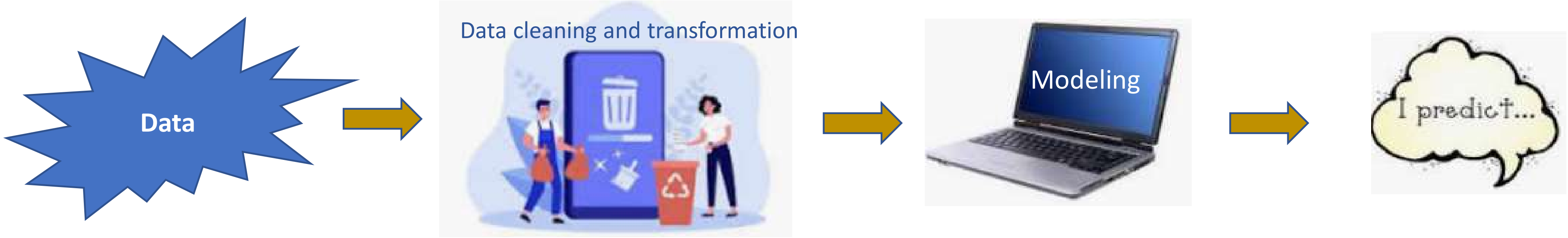
- Income information is important to make many decisions.
  - Income provides economic resources that shape many personal choices. It is also imperative for companies to know true income information to target the suitable consumer groups and mitigate the financial risk in situations like credit card or loan application, even buying a house.

- Everybody lies (Seth, 2017)
  - For various reasons, consumers often unable or unwilling to provide their true income information. which makes it difficult for enterprises to just rely on the self-reported data.

- A more objective way is needed to get the true income level information.
  - Goal is to predict the customer income level (high vs low) using objective indirect data, which are less likely to be manipulated.

Methods

- Through predictive modeling, correlated historical data can be used to build a model, which can predict the interested outcomes.
    - Predict customer's income level into 2 groups (classification)
    - Using CRISP-DM process (Cross-Industry Standard Process for Data Mining)
    - Using python as modeling language
    - Using common data modeling packages (like SCIKIT-LEARN).



Data

Data cleaning and transformation

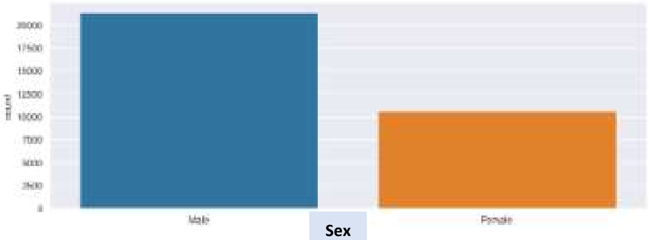
Modeling

Data set

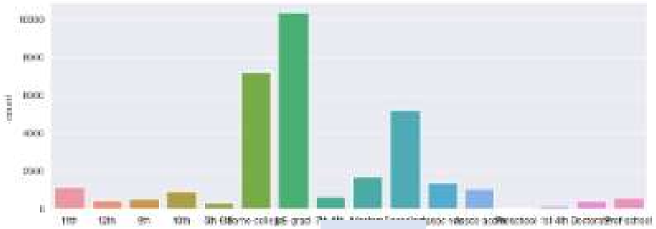
- This dataset originates from the 1994 Census Bureau database and was curated by Ronny Kohavi and Barry Becker.
- Collected information from 31934 individuals, who are old than 16 years, with AGI >$100. with work hour > 0.
- The supervised data classified individuals into 2 groups, with income exceed $50K (>) or no more than $50K (<=) in 1994. which is worth $100.6K today (per CPI inflation calculator). So, it is a still a good threshold for today.
- The dataset has mixed numerical and text information
- Information collected per individual:
  - Age: in years, numerical
  - Work class: private sector, government etc. text
  - Education level in highest grade. text
  - Education length: in years numerical
  - Marital status: text
  - Occupation: text
  - Relationship: with kid, not in family etc. text
  - Race: text
  - Sex: Male or female: text
  - Native country: text
  - Fnlwgt: final weight, an estimation of the number of people each observation represents in the population. numerical

Data Visualization

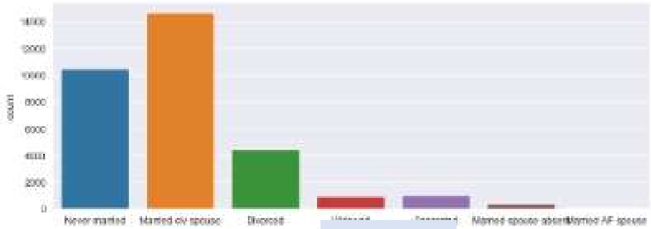
White is the dominating race
High income to low-income ratio is 0.32: 1



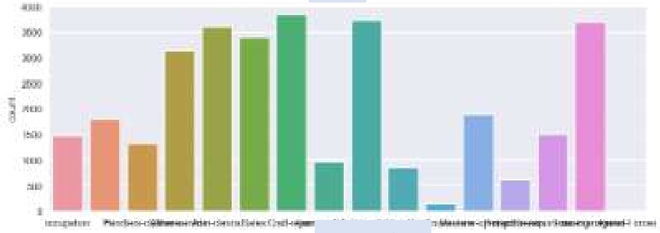
Sex



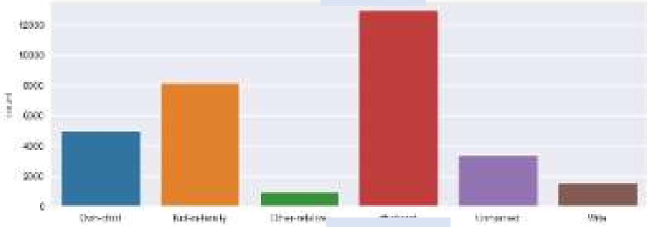
Education



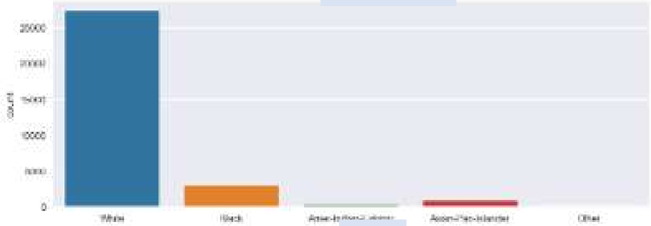
Marital status



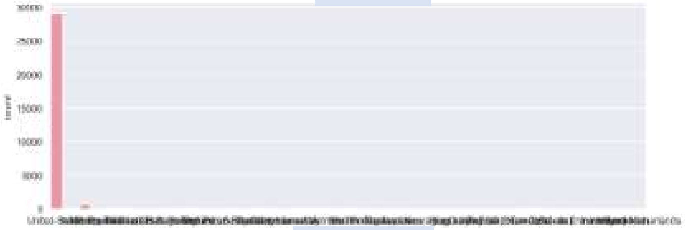
Occupation



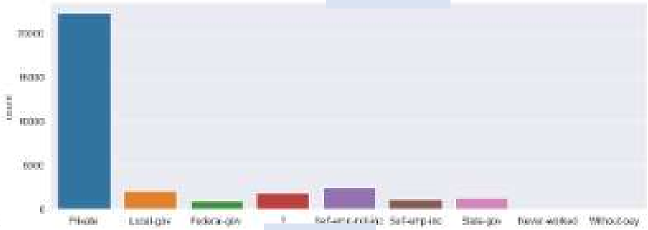
Relationship



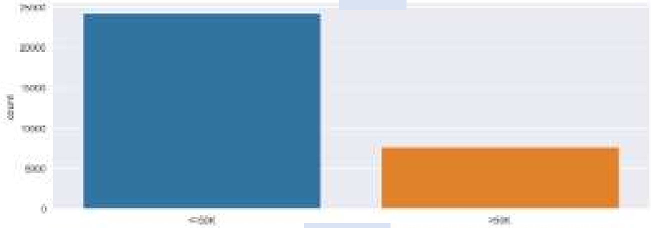
Race



Native country

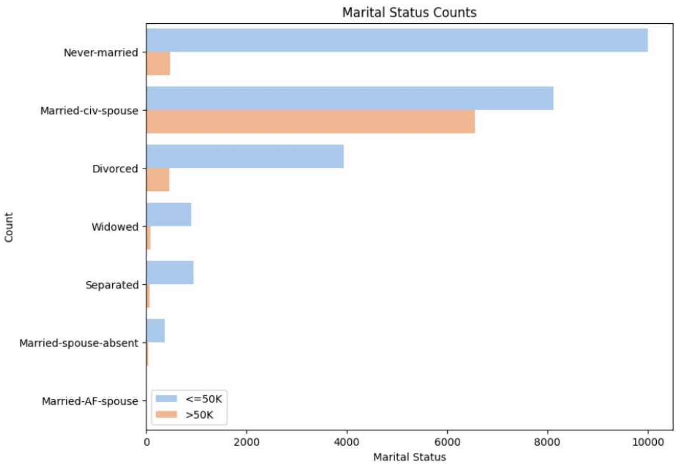


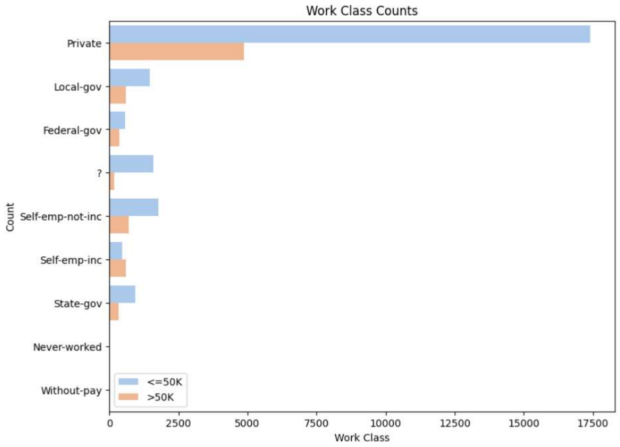
Work class

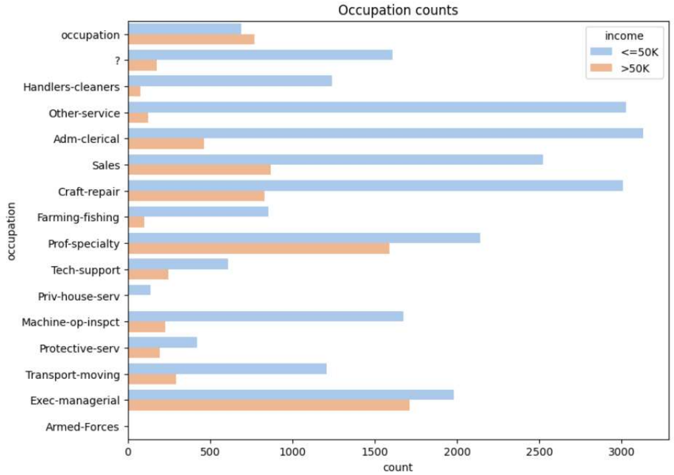


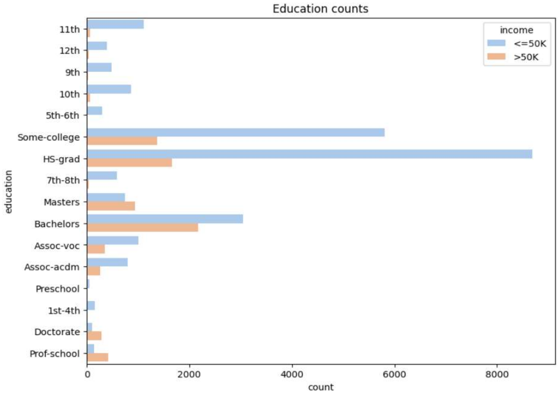
income

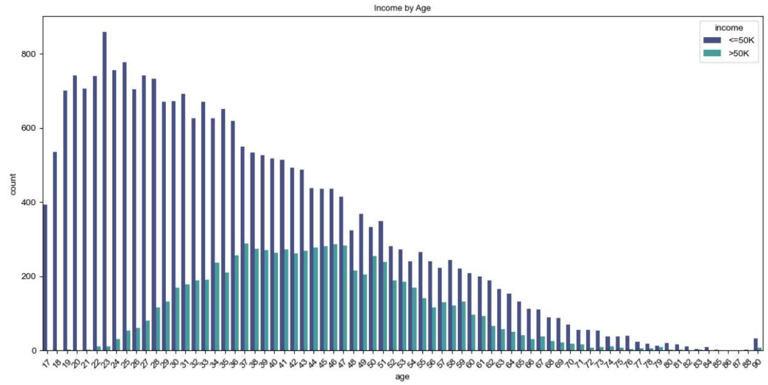
Data Visualization

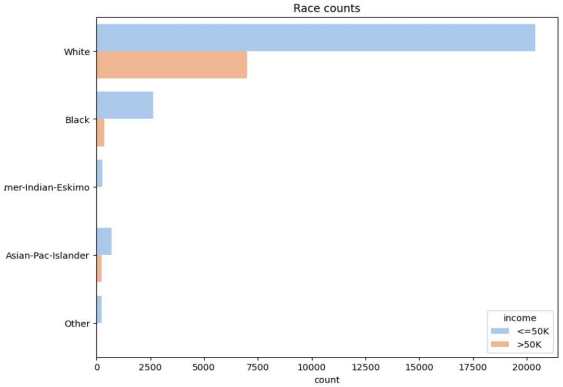












Marital status, work class and Age are driving income difference

Data Cleaning and feature engineering

Data cleaning:

- ”?” and meaningless answer (answer o occupation is occupation) are converted to NA.
- NA is less than 1% of all answers, so NA removed
- Education and education length number are similar thing so keep only education length.
- fnlwgt is also removed since it is not relevant to any other variables or income.

Feature engineering:

- Convert income to 0 and 1 (0 is <=$5k and 1 is >$5K).
- Convert sex to 0 and 1 (0 is male and 1 is female)
- Convert work class to numerical based on its mean income of this class.
- Convert all categorical columns (marital status, occupation, relationship, race, native country) to numerical values.
- Since income count is not balanced, so use SMOTE to get it balanced for model training
- Data is being standardized using standard scaler before train/test split.

Model selections

Models:

This is a classification question, so 3 classification models are selected.

- Random Forest classifier
  - A flexible, easy to use ML algorithm with build in benefit of overfitting avoidance.

- Random Forest classifier with hyperparameter tuning.
  - Buy tuning and select optimal model parameters, it normally yield improved performance.

-
- XGB Classifier
  - Often has better accuracy and scalable to fit the data size.

Model evaluations

Models need to be evaluated to ensure it is robust and effective.

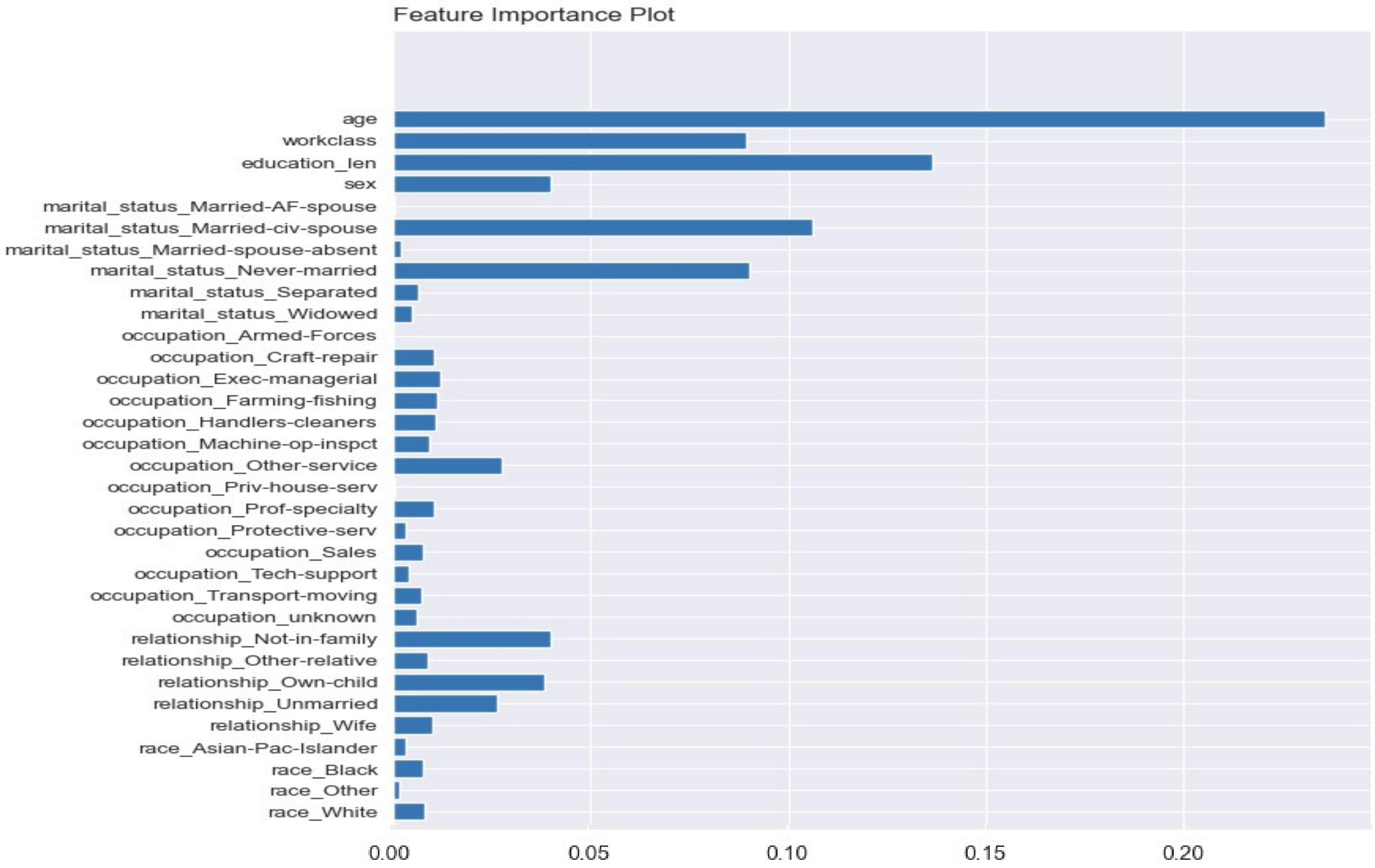
This is a classification question, so 3 classification models are selected.

- Accuracy: Ratio of correctly predicted observation to total observation. Not work well since data is imbalanced.
- Precision: Proportion of positive prediction which are correct.
- Recall: Proportion of actual positives that are classified correctly.
- F-1 Score: a harmonic mean of precision and recall.

	Random Forest Classifier	RF with tuning	XGB classifier
Accuracy	86%	86%	86%
F-1 score for income >$5K	0.86	0.87	0.86
Precision for income >$5K	0.84	0.84	0.83
Recall for income >$5K	0.89	0.89	0.89

I would recommend random forest classifier with tuned parameters

Factors impact income level most



Age, Education and Marriage status are the top 3 factors determining income level

Conclusion

- This project successfully built a machine learning model, which can classify individual’s income level into high/low group, using indirect objective information.

- The model is with 86% accuracy, 89% of classified high income are actual high income individuals.

- This method can be used to predict individual’s income level, or cross check if the self-reported income is trustworthy.

Recommendations/Next Step

- Getting more recent customer data to validate the model.
- Gather subject expert input if additional variables need to be considered
- Gather expert input on feasibility of the model
- ROI (return on investment) study
- Develop an implementation plan for executive approval.

Ethical implication

- The outcome of the model will help to make decisions which will shape people’s life, so regular monitor and review are recommended.

- The success of model is heavily depended on the data selected. Avoid discrimination on data selection.

- As society evolves, so are the factors impacting income, continuous improvement of model is suggested.

- This model will use a lot of personal information, please be aware of the data leak risk.

References and Acknowledges

- Income dataset:
  - ANAGHA K, P (Oct. 2023). Adult Income Census, Predict whether income exceeds $50K/yr based on census data, Retrieved June 10, 2024 from https://www.kaggle.com/datasets/anaghakp/adult-income-census/data

- Evaluating classification methods:
  - Brandon Wohlwend (Ju16,2023): Evaluating classification models
  - https://medium.com/@brandon93.w/machine-learning-evaluating-classification-models-18713af3d764

- What is XGBoost and why does it matter?
  - From Nvidia.com, author unknown
  - https://www.nvidia.com/en-us/glossary/xgboost/

- Thanks for the invaluable input/review from the instructor and classmates:
  - Professor Andrew Hua (instructor)
  - Ander Eguluz (classmate)
  - Krista-fer Knuckey (classmate)

Any Questions?

Break up