Report

Data exploration

1. **Missing value**

   Check whether the dataset contains missing values. The dataset does have some missing values.
2. **Nan rows**

   Check which rows contain NaN values. A total of 874 rows contain NaN values, with the exact rows listed in *data_exploration.ipynb*.
3. **Nan columns**

   Check which columns contain NaN values. Only the *'hispanic origin'* column contains NaN values.
4. **Number of columns**

   Dataset has 42 columns.
5. **All columns**

   'age', 'class of worker', 'detailed industry recode', 'detailed occupation recode', 'education', 'wage per hour', 'enroll in edu inst last wk', 'marital stat', 'major industry code', 'major occupation code', 'race', 'hispanic origin', 'sex', 'member of a labor union', 'reason for unemployment', 'full or part time employment stat', 'capital gains', 'capital losses', 'dividends from stocks', 'tax filer stat', 'region of previous residence', 'state of previous residence', 'detailed household and family stat', 'detailed household summary in household', 'weight', 'migration code-change in msa', 'migration code-change in reg', 'migration code-move within reg', 'live in this house 1 year ago', 'migration prev res in sunbelt', 'num persons worked for employer', 'family members under 18', 'country of birth father', 'country of birth mother', 'country of birth self', 'citizenship', 'own business or self employed', 'fill inc questionnaire for veteran's admin', 'veterans benefits','weeks worked in year', 'year', 'label'
6. **Number of features**

   After excluding the *weight* and *label* columns, the dataset contains 40 features.
7. **Numeric and category features**

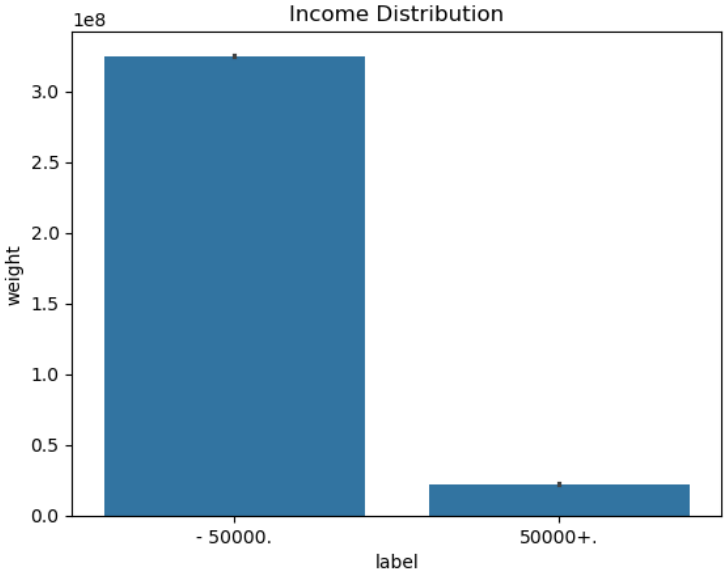
   The dataset contains 12 numerical features and 28 categorical features.

   Numerical features: 'age', 'detailed industry recode', 'detailed occupation recode', 'wage per hour', 'capital gains', 'capital losses', 'dividends from stocks', 'num persons worked for employer', 'own business or self employed', 'veterans benefits','weeks worked in year', 'year'

   Categorical features: 'class of worker', 'education', 'enroll in edu inst last wk', 'marital stat', 'major industry code', 'major occupation code', 'race', 'hispanic origin', 'sex', 'member of a labor union', 'reason for unemployment', 'full or part time employment stat', 'tax filer stat', 'region of previous residence', 'state of previous residence', 'detailed household and family stat', 'detailed household summary in household', 'migration code-change in msa', 'migration code-change in reg', 'migration code-move within reg', 'live in this house 1 year ago', 'migration prev res in sunbelt', 'family members under 18', 'country of birth father', 'country of birth mother', 'country of birth self', 'citizenship','fill inc questionnaire for veteran's admin'

8. **Class weighted samples**

   Based on the sample weights, there are 325,004,647.22 weighted negative samples and 22,241,245.25 weighted positive samples, where the negative class corresponds to *label < 50K* and the positive class corresponds to *label > 50K*. This indicates that the dataset is imbalanced.



Data cleaning

First, I checked whether the dataset contained any missing values. Only the *'hispanic origin'* column had NaN values, which were filled with 0.

The same data cleaning procedure was applied to both the classification and segmentation models.

Preprocessing

The dataset contains both numerical and categorical features. Numerical features were normalized to have a mean of 0 and a variance of 1, while categorical features were converted into one-hot encoded format.

The same preprocessing procedure was applied to both the classification and segmentation models. The classifier used separate training, validation, and test sets during preprocessing, whereas the segmentation model used the entire dataset.

1.Classifier

Dataset splitting

The dataset was divided into training, validation, and test sets in proportions of 70%, 10%, and 20%, respectively.

Models

I trained three models: Logistic Regression, Random Forest, and XGBoost. The Logistic Regression and Random Forest models were trained using their default settings. XGBoost served as the primary model of focus, with the following hyperparameters: learning rate = 0.05, maximum depth = 6, subsample = 0.8, colsample_bytree = 0.8, and a maximum of 2000 estimators.

Training

The Logistic Regression and Random Forest models used only the training set during training. In contrast, XGBoost utilized the validation set for parameter tuning and early stopping.

Sample weight

The sample weights were applied to all three models during both training and evaluation.

Model evaluation

1. Logistic regression

Test set evaluation

	Precision	Recall	F1	Accuracy	Balanced accuracy	AUC
<50k	0.96	0.99	0.97			
>50k	0.69	0.39	0.5			
				0.95	0.69	0.94

The Logistic Regression model performed well on the <50K class, achieving a precision of 0.96 and a recall of 0.99. Since <50K is the majority class, this result is expected.
However, for the minority class (>50K), the model achieved only 0.69 precision and 0.39 recall, indicating that it failed to identify most of the >50K samples.

2. Random forest

Test set evaluation

	Precision	Recall	F1	Accuracy	Balanced accuracy	AUC
<50k	0.96	0.99	0.97			
>50k	0.74	0.38	0.5			
				0.95	0.68	0.94

The Random Forest model showed similar performance to Logistic Regression on the <50K class. However, its performance on the minority class (>50K) improved, achieving a precision of 0.74.

3. Xgboost

Test set evaluation

	Precision	Recall	F1	Accuracy	Balanced accuracy	AUC
<50k	0.97	0.99	0.98			
>50k	0.75	0.48	0.59			
				0.96	0.74	0.95

XGBoost is the proposed model for this classification task. It outperforms both Logistic Regression and Random Forest across all evaluation metrics.

XGBoost predicts the <50K class almost perfectly. It effectively recognizes low-income individuals, achieving a recall of 0.99—meaning it correctly identifies 99% of them—and a precision of 0.97, indicating that 97% of predicted low-income cases are correct.

For the minority >50K class, the model achieves a precision of 0.75, meaning that 75% of predicted high-income cases are accurate. Its recall of 0.48 indicates that it captures nearly half of all high-income individuals, which is acceptable for marketing applications.

The XGBoost model achieved an overall accuracy of 96%, which is high, largely due to the dominance of the <50K majority class.

Considering the class imbalance, the balanced accuracy was also calculated and found to be 74%, which is reasonably good. The model achieved an AUC of 0.95, indicating strong overall performance.

Therefore, XGBoost was selected as the final classifier for this task.

Feature importance

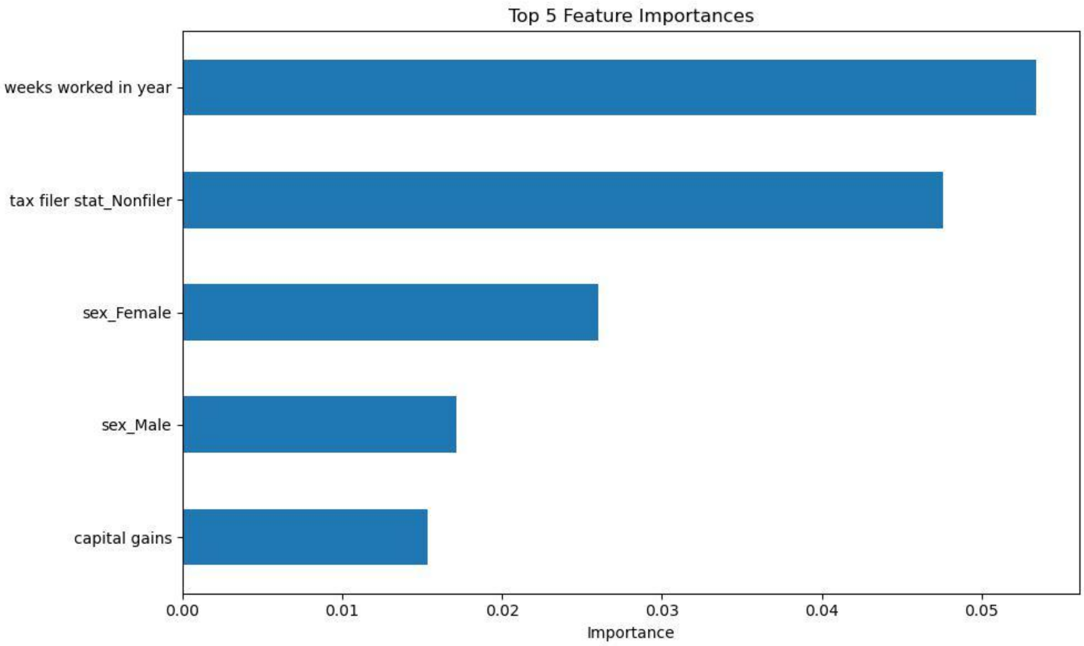
I also computed the feature importance from the XGBoost model and selected the top five features for potential marketing use.

The most important feature is *weeks worked in year*, indicating that working time is a key factor influencing income. The second most important feature is *tax filer stat_Nonfiler*, suggesting that individuals who do not file taxes are unlikely to have high income. The third and fourth features relate to *sex*, which are less relevant for marketing purposes. The fifth most important feature is *capital gains*, which aligns with expectations, as individuals with higher capital gains typically have higher income.

Overall, *weeks worked in year*, *tax filer stat*, and *capital gains* are the most actionable features for marketing. Individuals who work more weeks, file taxes, and have higher capital gains can be considered primary marketing targets.

Feature importance

Weeks worked in year	Tax filer stat_Nonfiler	Sex_Female	Sex_Male	Capital gains
0.053	0.048	0.026	0.017	0.015



Future work

AutoML can be explored in future work, as it can automatically search across various models, including neural networks. Due to time constraints, AutoML was not implemented in this study.

2.Segmentation model

Model

K-Means was selected as the model for this task, as it is an unsupervised learning problem.

Decide number of segments

First, I determined the optimal number of segments (*k*) for the K-Means algorithm. Two methods were used: the Elbow Method and the Silhouette Score. Based on the results from both methods, *k = 4* was selected.

K-Means training

The preprocessed dataset was fed into the K-Means algorithm for training, with sample weights applied during the training process.

Sample weight

Sample weights were applied when calculating the mean and standard deviation.

Select features

The dataset contains 40 features, from which key features need to be selected to characterize the four clusters. Z-scores were used to identify features that differ the most across segments.

4 segments of people

Segment 0

Basic Summary:

Age	Education	Sex	Marital stat	High income ratio
56	High school graduate	Female	Married-civilian spouse present	0.017

Segment 0 represents retired older females who are not in the labor force. Their *weeks worked in year* is 0.88 standard deviations below the mean, averaging only 2.22 weeks.

This group generally has lower income but considerable free time. They can be targeted with marketing campaigns for digital entertainment (e.g., Netflix), health-related, and family-oriented products.

Segment 1

Basic Summary:

Age	Education	Sex	Marital stat	High income ratio
8	Children	Male	Never married	0

Segment 1 represents children and can therefore be excluded from the marketing target group.

Segment 2

Basic Summary:

Age	Education	Sex	Marital stat	High income ratio
38	High school graduate	Male	Married-civilian spouse present	0.11

Segment 2 consists mainly of married, middle-aged, high school–educated males. Their *weeks worked in year* is 0.85 standard deviations above the mean, averaging 44.43 weeks compared to the overall average of 23.58, indicating long working hours. Additionally, this group has relatively high income, with 11% classified as high-income earners.

As a high-income group with limited free time, they can be considered a key marketing target for high-value products such as automobiles and real estate.

Segment 3

Basic Summary:

Age	Education	Sex	Marital stat	High income ratio
38	High school graduate	Male	Married-civilian spouse present	0.123

Segment 3 is similar to Segment 2 but has a slightly higher proportion of high-income individuals (12.3%). Based on the top selected features, this group's *full-time schedule* is 1.47 standard deviations above the mean, with 80.74% working full time. The *children or armed forces* variable is 1.25 standard deviations below the mean, accounting for only 0.64% compared to the overall 61.31%.

Their *weeks worked in year* is 0.87 standard deviations above the mean, averaging 44.89 weeks versus the overall average of 23.58—similar to Segment 2.

This group likely contains more individuals in full-time employment than Segment 2. Given their comparable characteristics, the same marketing strategies can be applied to both segments, or Segments 2 and 3 can be treated as a single target group.