Project: Classification and Segmentation Models

**Objective 1:** Classification Model based on Census Data (precision:

Introduction:

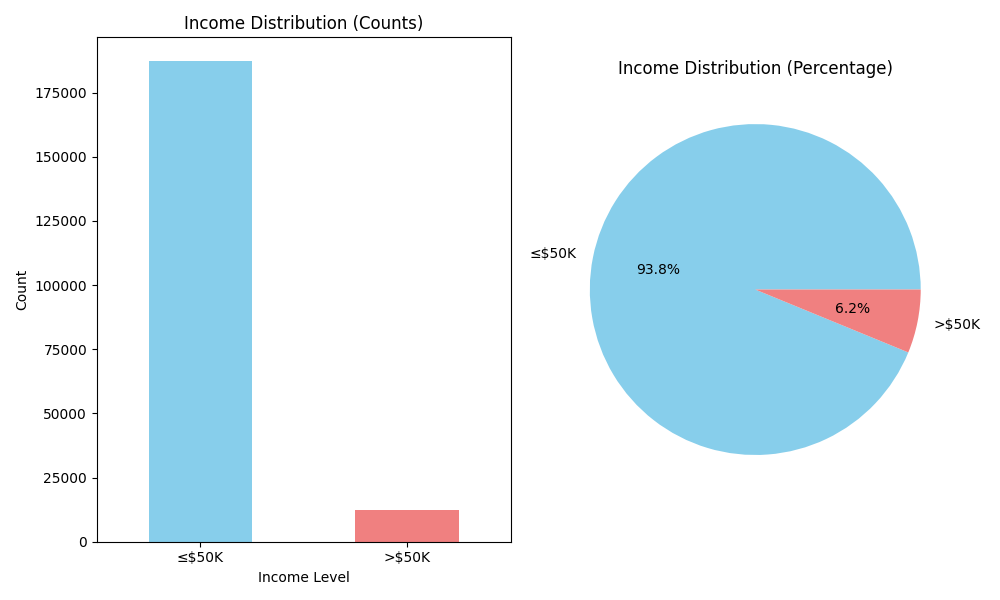
The Goal of the model is to accurately classify input data into the 2 targeted income groups (>50k or <=50k), to better suit with the marketing department in Walmart to better identify potential customers for specific advertisement.

Potential Use Case

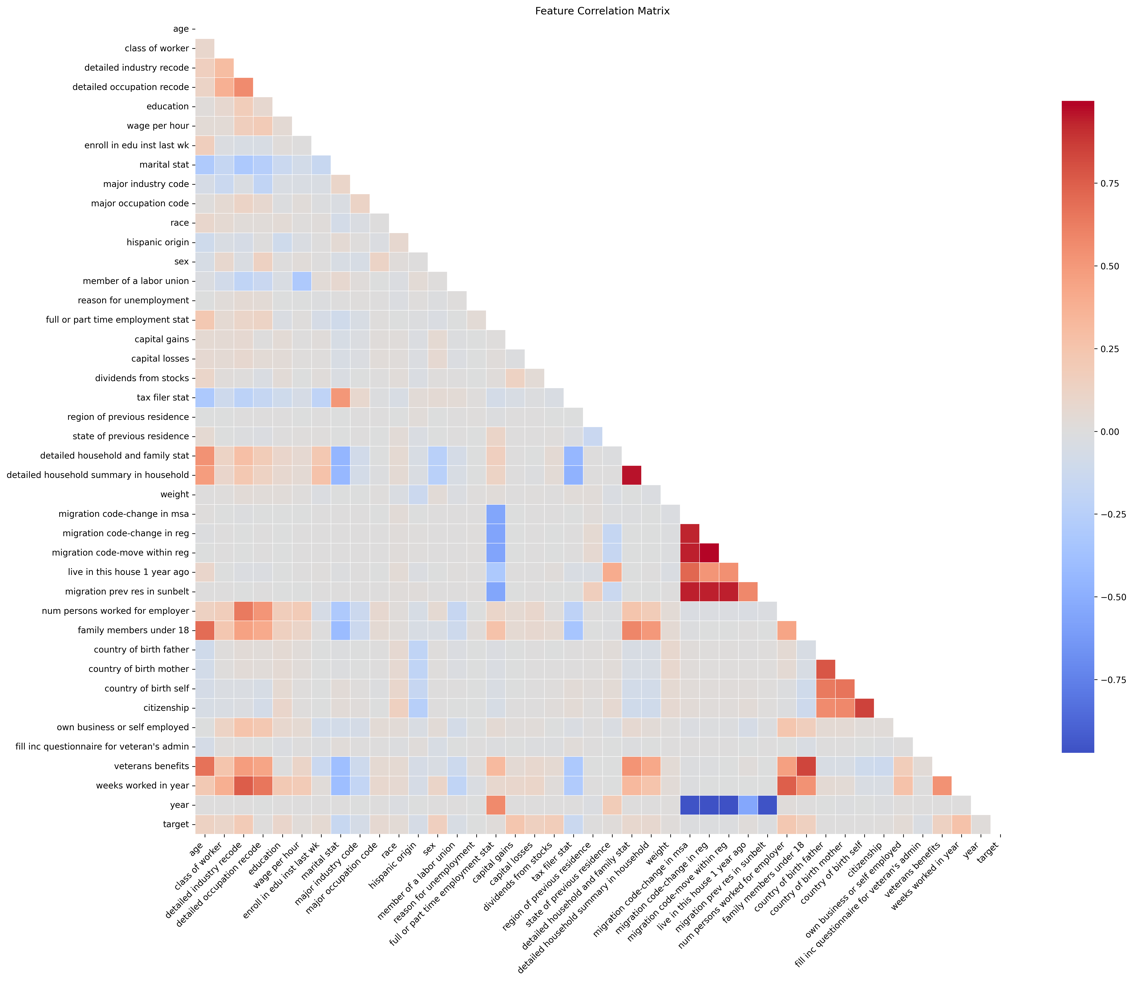
* Walmart online retail:
  + Targeted recommendation/advertisement when user fit into one of the category,
  + for instance, if based on the user information, the user can be classified into high income, the “most relevant” sorting and prioritize higher price product
* Local Product Purchase:
  + Based on geographical information + major occupancy around local store(for instance if the local business manufacture/finance/etc.) + predicted income level the local store can adjust the purchase strategy
  + For instance, if

Data Exploration

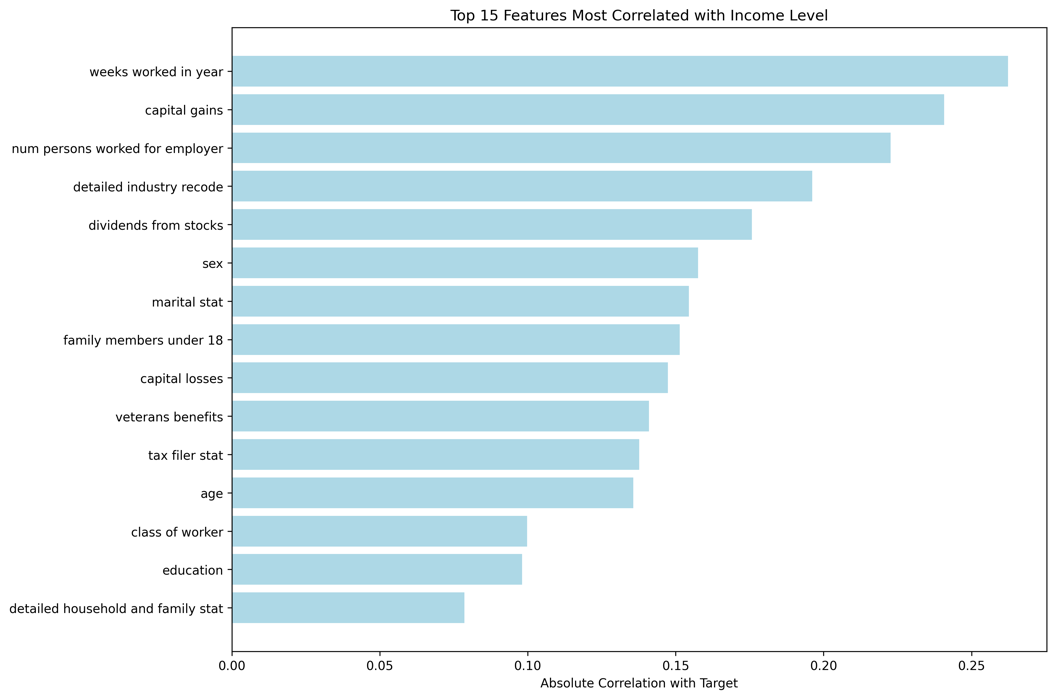
* Target Groups understanding (column\_name: “label”)
  + Skewed (15.0 : 1.0), need to resolve the imbalance of target data – indicated that for model evaluation Precision and Recall score



* Categorical Feature understanding:
  + Data skewed:
  + Data Missing:
  + Collinearity: Filtered out below columns that have high collinearity with other columns
    - detailed industry recode <-> weeks worked in year: 0.754
    - detailed household and family stat <-> detailed household summary in household: 0.962
    - migration code-change in msa <-> migration code-change in reg: 0.934
    - migration code-change in msa <-> migration code-move within reg: 0.939
    - migration code-change in msa <-> live in this house 1 year ago: 0.721
    - migration code-change in msa <-> migration prev res in sunbelt: 0.938
    - migration code-change in msa <-> year: -0.959
    - migration code-change in reg <-> migration code-move within reg: 0.999
    - migration code-change in reg <-> migration prev res in sunbelt: 0.939
    - migration code-change in reg <-> year: -0.971
    - migration code-move within reg <-> migration prev res in sunbelt: 0.942
    - migration code-move within reg <-> year: -0.971
    - migration prev res in sunbelt <-> year: -0.962
    - num persons worked for employer <-> weeks worked in year: 0.747
    - family members under 18 <-> veterans benefits: 0.843
    - country of birth father <-> country of birth mother: 0.781
    - country of birth self <-> citizenship: 0.846



* + Feature Importance:
    - Detailed output for feature selection place sees the output of
      * Categorical data:
      * Numerical data:
      * Mutual Information:
    - **Top predictive features**: weeks worked in year, capital gains, num persons worked for employer, detailed industry recode, dividends from stocks



* Conclusions:
  + 40 features are really detailed, but does include a lot of Null(“not in universe”) data points

Pre-processing approach:

* Feature selection:
  + 11 features
* Down sampling data:
  + Balance the distribution of data, resample the target majority to match with the target minority
  + The imbalanced data is raising an issue of low precision score, but high accuracy, meaning that the model will place 99% of the input to <=50% and will score a 93% accuracy
  + Tested with up-sampling, but due to the extreme skewed data, the performance is not significantly different from original data
* Scaler selection:
  + Quantile (tested with Maxmin, standard, )
  + Potential issue – if the distribution of the data changed, this approach will reduce accuracy of the model
* Since the data is down sampled, it would be better to just use a simpler model

Training:

* Train test split

Model Selection and Evaluation:

* Selected 3 different models Logistic Regression, XGBoost, and RandomForest
* Model Evaluation and Performance:
  + Logistic Regression
  + RandomForest
  + XGBoost

**Objective 2:** Segmentation Model based on Census Data

Introduction:

Using 3D clustering for better understanding on the census data

* (Reason using a different approach from the classification model) take “weight” into consideration – for better scalability of the model
* Use case:
  + When there is a
* Logic, based on the major Occupation recommend specific items

A screenshot of a graph

Description automatically generated

A group of graphs and charts

Description automatically generated

Potential Next Step:

* Better Data augmentation
* Data Pipeline:
  + Data gathering + Cleaning + scale
* Model CI/CD:
  + The current approach is hard coded
* Internal Tool:

Things to Improve: