Project: Classification and Segmentation Models

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

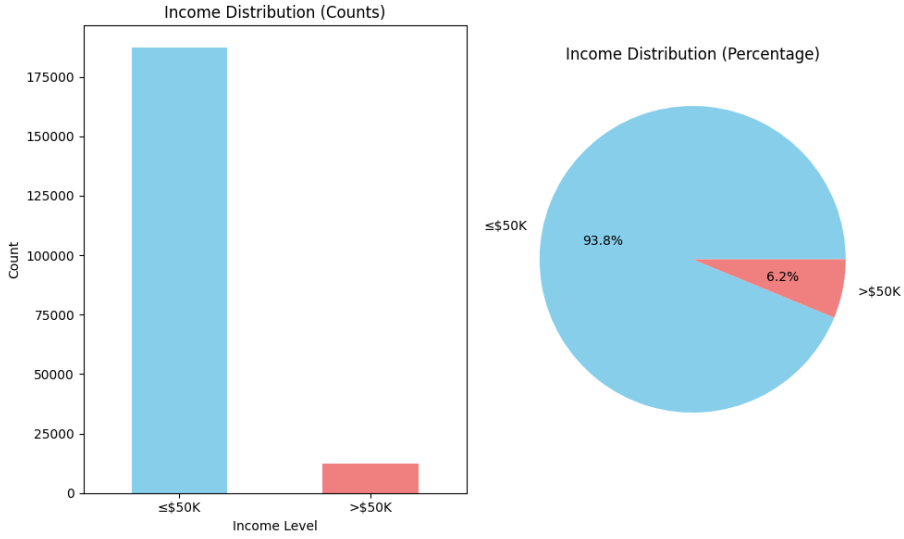
Introduction:
The goal of this model is to classify individuals into two income groups: **≤$50k** and **>$50**k. By doing so, Walmart's marketing department can better identify potential customers for **targeted advertisements and product recommendations**, both online and in physical stores.

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

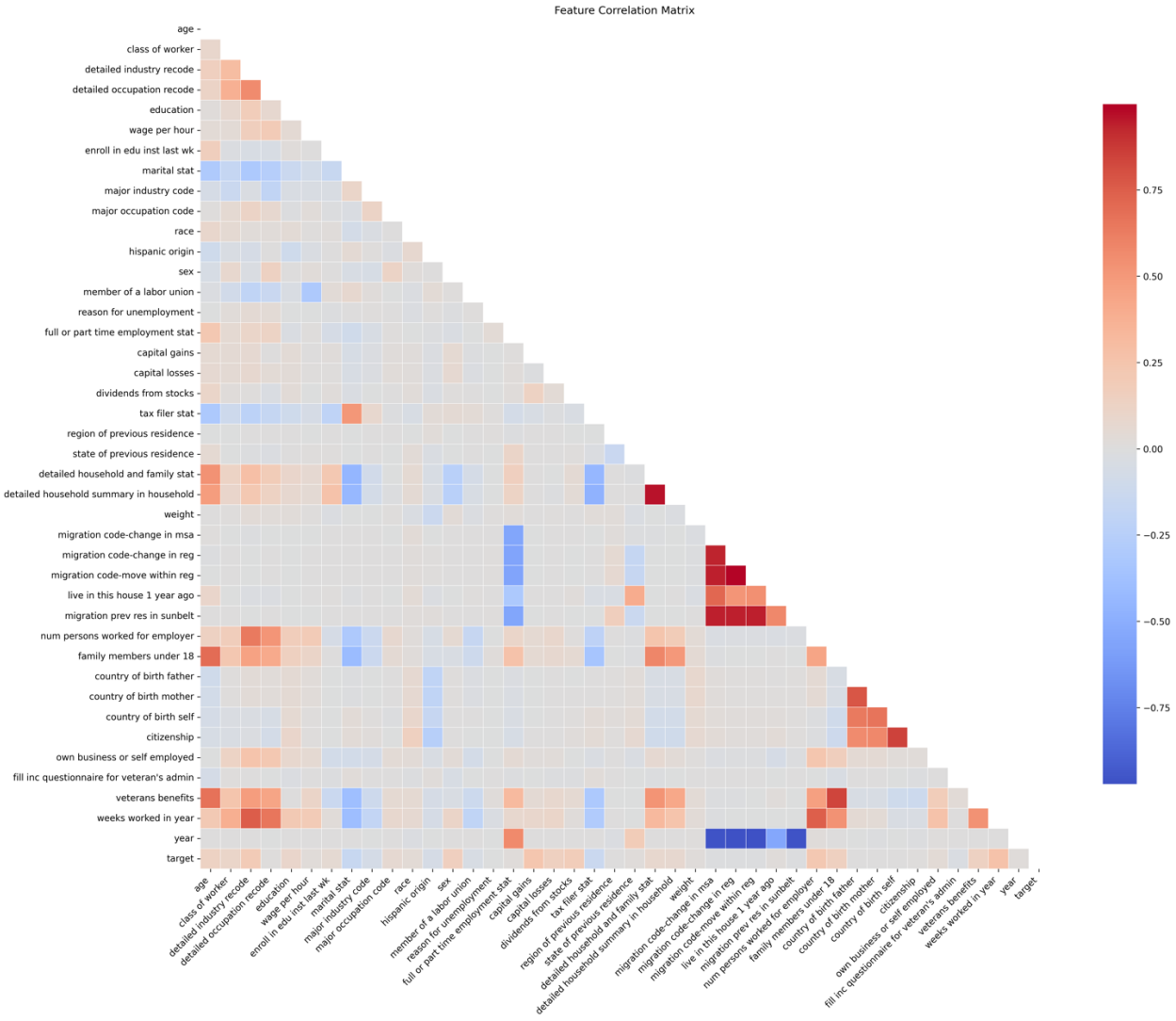
Data Exploration

- Target Groups understanding (column_name: "label")
  - Highly imbalanced dataset (15.0 : 1.0), need to resolve the imbalance of target data – indicated that for model evaluation accuracy alone is misleading, so precision, recall, and pr-auc are emphasized

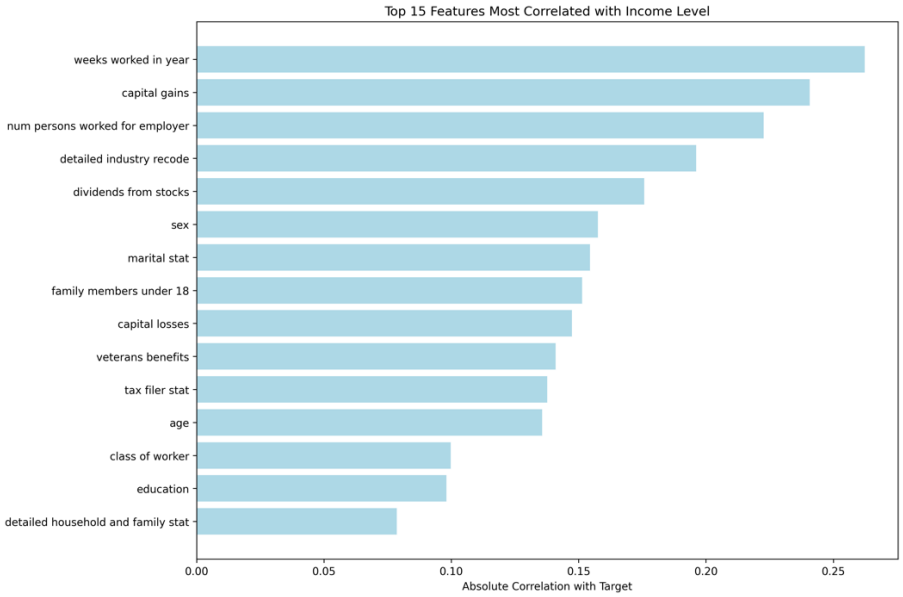


- Categorical Feature understanding:
  - Data skewed and missing:
    - Many features contain "Not in universe" values
  - Collinearity: Filtered out below columns that have high collinearity with other columns
    - detailed industry recode <-> weeks worked in year: 0.754
    - detailed household and 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
- Outlier Handling:
  - Replaced invalid/extreme values (age=0, capital gains/losses outliers) with mode values
- 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 -- Quantile transformation used to normalize skewed distributions (tested with Maxmin, standard, Robust , Quantile-- quantile has the best performance)
  - 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	Accuracy	ROC-AUC	PR-AUC	Notes
Logistic Regression	0.8483	0.9263	0.9160	Balanced precision/recall
Random Forest	0.8307	0.9137	0.9013	Slightly lower recall
XGBoost (Best)	**0.8635**	**0.9365**	**0.9314**	Best overall performance

- Evaluation:
  - Logistic Regression, Random Forest, and XGBoost. Evaluation used Accuracy, ROC-AUC, and PR-AUC (most relevant given imbalance).

Business Insight:

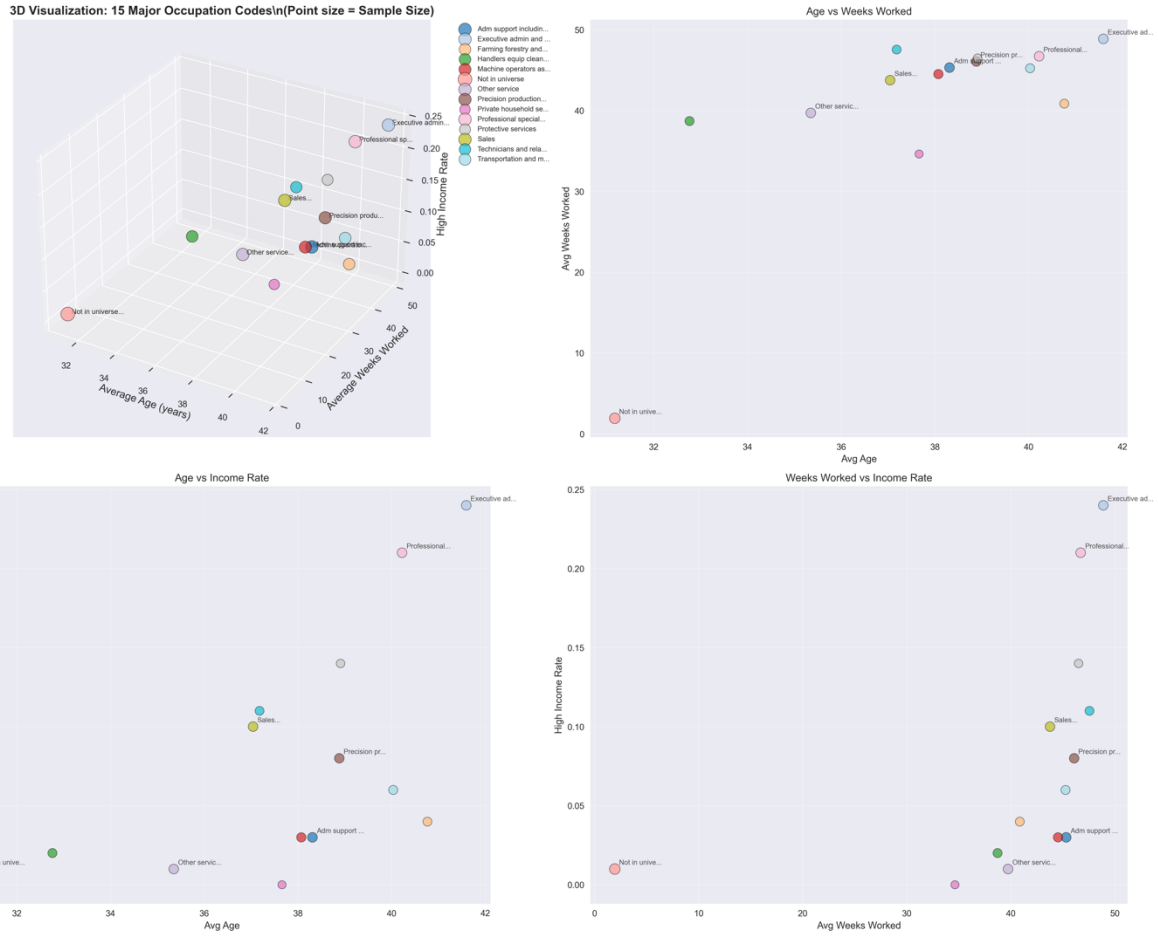
- Targeted Marketing: Walmart can deliver income-aware recommendations, improving ad relevance and conversion
- Optimized Inventory: Local stores can adapt purchase and stocking strategies to match income distribution + occupation mix in the area
- Customer Segmentation: The classifier complements occupation-based clustering, allowing multi-dimensional segmentation (occupation * income prediction)

Objective 2: Segmentation Model based on Census Data

Introduction:

Using 3D clustering for better understanding on the census data, the clustering analysis can highlights clear differences in age, work stability, and income potential across occupation groups. By tailoring marketing strategies to these segments, the retail client can optimize product mix, pricing, and messaging—maximizing engagement across all income levels:

- (Reason using a different approach from the classification model) **take "weight" into consideration** – for better scalability of the model in real life
- analyzed 199,487 customers across 14 major occupation groups using three key factors:
    - Average Age (professional maturity)
    - Average Weeks Worked (employment stability)
    - High-Income Rate (earning potential)
- Clustering revealed three distinct customer segments that vary significantly in income potential, work stability, and lifestyle. These groups provide a strong foundation for occupation-based retail marketing.
- Logic: based on the major Occupation recommend specific items



- Key Findings:
    - Premium Segment (High-Income Occupations)
        - Market Size: ~39,900 customers (20% of market)
        - Income Rate: 20.8% (well above average of 9.0%)
        - Profile:
            - Average Age: ~ 40 years
            - Weeks Worked: ~ 47 / yr
        - Target Occupations:
            - Executive & Managerial
            - Professional Specialty
            - Sales
            - Protective Services
        - **Interpretation:** Stable, established professionals with strong earning potential
        - **Potential Strategy**:
            - Products: Luxury goods, high-end electronics, premium brands
            - Pricing: Premium pricing (emphasize quality and status
            - Channels: Upscale stores, online premium platforms
            - Messaging: Status, convenience, time-saving benefits
    - Mainstream Segment (Mid-Income Occupations)
        - Market Size: ~41,900 customers (21% of market)
        - Income Rate: 6.5% (near average)
        - Profile:
            - Average Age: ~ 39 years
            - Weeks Worked: ~ 45/ yr
        - Target Occupations:
            - Clerical/Admin Support
            - Precision Craft & Repair
            - Machine Operators
            - Transportation
            - Farming/Fishing
            - Technicians
        - **Interpretation:** Reliable workforce, moderate income levels, family-oriented lifestyles
        - **Potential Strategy:**
            - Products: Mid-range products with practical utility
            - Pricing: Competitive pricing with seasonal promotions
            - Channel : Department stores, online marketplaces
            - Messaging: Value, reliability, family-focused
    - Value Segment (Budget-Conscious Occupations)
        - Market Size: ~117,700 customers (59% of market)
        - Income Rate: 1.0% (well below average)
        - Profile:
            - Average Age: ~ 34 years
            - Weeks Worked: ~29/ yr
        - Target Occupations:
            - Not in Universe (unemployed/unstable work)
            - Other Service Workers
            - Handlers & Laborers
            - Private Household Services
        - **Interpretation:** Younger, less stable employment, highly price-sensitive
        - **Potential strategy:**
            - Products: Essentials, generic brands, budget-friendly items
            - Pricing: Low prices, deep discounts, bulk deals
            - Channels: Discount stores, dollar stores, online deal sites
            - Messaging: Affordability, necessity, saving

Potential Next Step:

- Better Data augmentation, for instance:
    - Enrich the dataset with additional demographic, geographic, or behavioral features to improve the model performance / making insightful decisions
- Add SHAP or LIME explanations to help non-technical teams understand why customers are classified into each group
- Data gathering + Cleaning + scale/ Model CI/CD
    - The current approach of data conversion and process has largely been hard coded (feature columns etc.) if the model need to be deployed for CI/CD, there is a need to make the feature selection to be dynamic
- Internal Tool Development(Two-stage decision flow):
    - Combination of the segmentation model and classification model:
        - Segment assignment (who are they?)
            - Input: occupation, age, weeks worked, etc.
            - Output: f(Input) = segment = Premium | Mainstream | Value} + a segment-affinity score (0 ~ 1)
        - Candidate generation (what do segments want?)
            - Map each segment to 8–15 product families with high historical lift for that segment (e.g., Premium → high-end electronics; Value → essentials/discount).
    - Income classification (what can they likely pay?)
        - Run XGBoost income model.
        - Output: p_high_income (0–1) → translate to price band (e.g.
            - $P \geq 0.7$ → premium
            - 0.4 – 0.7 → mid
            - < 0.4 → budget
    - Scoring & ranking (match need + price)
        - For each candidate SKU, compute a blended score:
        - Score = w1·segment_affinity + w2·price_match + w3·personal_signals + w4·freshness + w5·margin_weight
            - price_match: 1.0 if SKU price falls in the user's band, decays if above/below
            - personal_signals: recent views, add-to-carts, brand affinit
            - freshness: seasonal/new arrival boost
            - margin_weight: optional business control
    - Guardrails
        - Always include 1–2 cross-band options (e.g., stretch premium or value alternative) to capture upsell/downsell.
        - Cap the % of any single brand/category for diversity.

References:

Handling imbalance data using upsampling and dowsampling
https://medium.com/codex/handling-imbalanced-data-upsampling-and-downsampling-in-machine-learning-10f33ff0620b

QuantileTransformer Scaler
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

Maxmin Scaler
https://www.analyticsvidhya.com/blog/2020/12/feature-engineering-feature-improvements-scaling/

Upsampling SMOTE reference (ends up using down sampling due to limited features and performance)
https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Walmart Marketing Insights – shifting to online shopping
https://businessmodelanalyst.com/walmart-marketing-strategy/#Walmart_Marketing_Goals_and_Objectives

Walmart Market Shift
https://www.thestreet.com/retail/walmart-makes-drastic-move-to-keep-customers-from-fleeing-stores