Description of your data exploration and pre-processing approaches, model architecture, training algorithm, and evaluation procedure, any interesting findings and exploration. Any business judgment and decision related to your data approaches, model selection and model usage recommendation.

Data Exploration

The dataset provided by Walmart contains ~200,000 rows and 40 demographic and employment-related variables along with a binary target variable indicating whether income was <50K or >=50K. Features include both categorical (e.g., work class, education, marital-status, occupation, relationship, race, sex, native-country) and numerical variables (e.g., age, hours-per-week, capital-gain, capital-loss).

Missing Values and Data Quality

- Several categorical features contained **"?"** and **"Not in universe"** values, representing missing or unknown categories. In a few variables, these values appeared in more than 50% of the observations. To avoid training the model on unreliable data, we dropped these features (Zheng & Casari, 2018). **It reduces noise and prevents misleading patterns in the predictions, ultimately helping the model target customers more accurately.**
- Numeric variables such as capital gains, capital losses, and dividends from stocks contained missing entries, which were imputed with zeros under the **assumption that non-reporting indicates no such income. We also assume that capital gain, loss, and dividends are for the same fiscal year.**

Class Imbalance

The income distribution is imbalanced, with most individuals belonging to the ≤50K income group. The weighted class imbalance takes into account the stratified weights of the census. There is not much of a difference between class imbalance for raw and weighted. **We will model the data on the assumption that the data is representative of the entire population.**

Label	**Raw %age**	**Weighted %age**
>=50,000	6.11	6.31
<50,000	93.88	93.69

Feature Engineering and Data Analysis

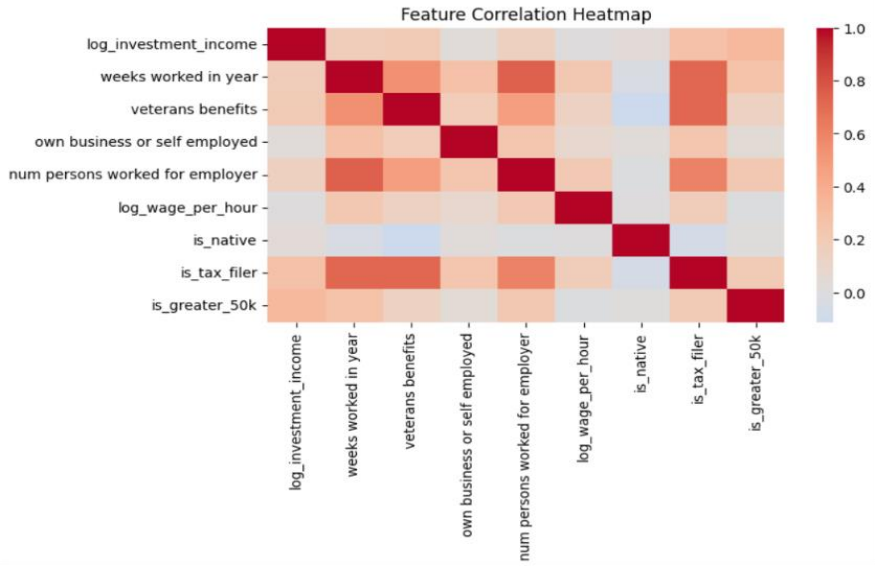
To **reduce the sparsity in categorical variables and improve interpretability**, we engineered several new features based on domain knowledge and exploratory findings:

- **Education Grouping:** Combined granular categories into broader groups to simplify interpretation and reduce sparsity.
- **Marital Status:** Collapsed into three bins: married, never married, previously married.
- **Employment:** Grouped rare job categories into "Other" to stabilize training.
- **Family Bins:** Engineered a feature indicating single, dependent, or minor household roles.
- **is_native:** Binary flag for U.S. native citizenship to capture immigration effects.
- **is_tax_filer:** Binary flag for tax filer status, a proxy for income stability.
- **log_investment_income:** Log-transformed investment income to handle skewness and outliers.
- **log_wage_per_hour:** Log-transformed hourly wage for better distribution and feature impact.

Feature	**Categories before Binning**	**Categories After Binning**	**Top Category %ages**
Education Bin	17	6	High school - 35%
Employment Bin	8	4	Unemployed – 72%
Marital Status Bin	7	3	Never Married, Married – 43%
Age Bin	0-90	4	0-20, 20-40 – 30%
Hispanic Bin	9	4	Non-Hispanic – 86.6%
Race Bin	5	4	White – 84%
Household Bin	8	5	Head of Household – 37%
Family Bin	38	5	Householder – 37%

Exploratory analysis revealed clear patterns. Most people are salaried, as median hourly wage is zero for both groups, but high hourly rates appear more often in the >50K group. Investment income is concentrated among higher earners, while lower earners mostly have little or none. Citizenship and tax filer status also matter. High earners are usually native-born and file taxes, whereas non-filers dominate the ≤50K group. Family structure adds to this divide: single and

single-parent households are common among lower earners, while married households, often dual-income, are more prevalent among higher earners.



The heatmap shows that **investment income, weeks worked, and tax filer status have strong positive correlations with earning >50K**, while other features like being native-born or veterans benefits show weaker or negligible relationships

Feature Encoding and Modeling Preparation

Choice of Algorithm and Metric

We focused on **recall** because it focuses on identifying all positive cases; missing someone who earns over 50K is more costly for Walmart than accidentally targeting someone who earns less. In other words, we would rather include a few low-income customers than miss high-income ones.

We tested two approaches: **XGBoost** (XGBoost Developers, n.d.), a powerful algorithm for tabular data, and **IsolationForest**, treating high-income customers as "anomalies." After running baseline models, XGBoost gave much better recall than IsolationForest (see appendix), so we chose XGBoost as the main model**. XGBoost works well here because it handles many variables, deals with class imbalance, and shows which features matter most. Its gradient boosting approach also helps capture complex patterns between demographics, jobs, and income.**

Categorical Feature Encoding

To prepare categorical variables for modeling, we transformed them into numerical representations. **Nominal variables** with no inherent ordering (such as family_bin and race_bin)

were **one-hot encoded**, producing binary indicator variables for each category. This ensured that the model did not mistakenly assign ordinal meaning to unordered groups. For **ordinal features** with natural rankings, **label encoding** was applied to preserve their intrinsic hierarchy.

Handling Class Imbalance

To address class imbalance, we applied two different approaches. First, we trained the **baseline** XGBoost model using the **scale_pos_weight parameter**, set to **15.34** (the ratio of class imbalance in the training data)**.** This weighting mechanism increased the penalty for misclassifying minority-class instances, thus forcing the model to pay closer attention to individuals earning above $50,000.

Next, we performed **stratified undersampling** (keeping every other parameter constant) of the majority class in the training set. This ensured that both income classes were more evenly represented during training, while still preserving the original class proportions in the validation and test sets for realistic evaluation. Stratified undersampling gave us better results so we went with the undersampling technique.

Modeling

Training and Hyperparameter Tuning

Three modeling pipelines were evaluated:

1. **Baseline XGBoost** with **scale_pos_weight** - This model served as the initial benchmark and directly addressed the class imbalance through weight adjustment.
2. **Baseline XGBoost** with **undersampling -** The **undersampled** dataset was used to train an unmodified XGBoost model, offering insight into whether balancing the dataset itself could outperform weighting. **(scale_pos_weight = 1)**
3. **Fine-tuned XGBoost with RandomizedSearchCV, Regularization (L1, L2), and Cross Validation -** The **undersampled** training set was further optimized using randomized hyperparameter search. This allowed for exploration of learning rates, maximum tree depth, number of estimators, and subsampling parameters to maximize predictive performance while avoiding overfitting. **(scale_pos_weight = 1)**

Evaluation and Results

Model performance was assessed using a **confusion matrix**, **classification report**, and **ROC-AUC score**.

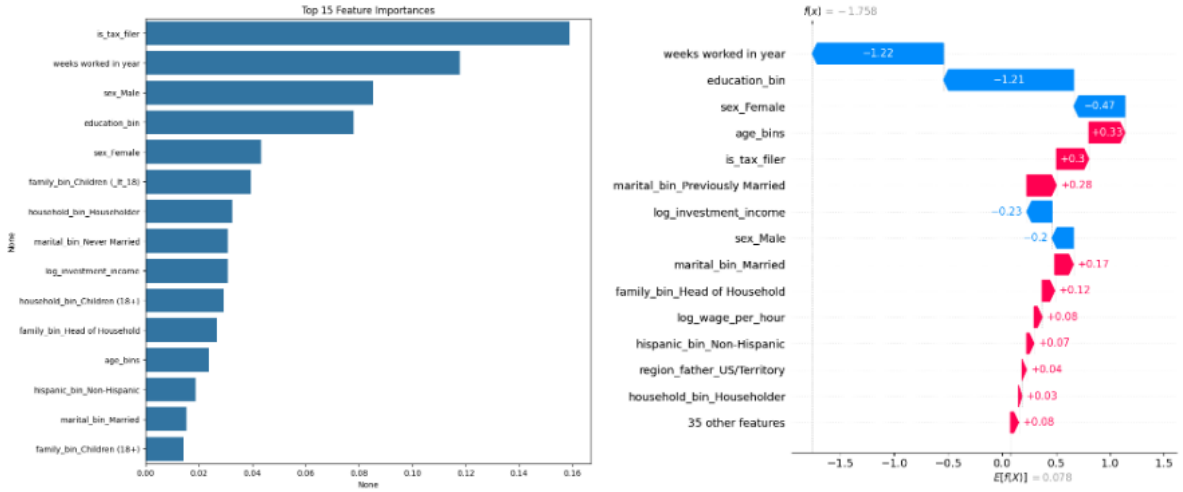
Model	Class	Precision	Recall	F1 Score	ROC-AUC

Xgboost (baseline)	<50K	0.989	0.855	0.917	0.935
	>=50K	0.277	**0.853**	0.418	
XGBoost (undersampled)	<50K	0.99	0.835	0.905	0.932
	>=50K	0.257	**0.877**	0.397	
Fine-tuned Xgboost	<50K	0.992	0.826	0.901	0.935
	>=50K	0.251	**0.894**	0.392	

- The **undersampled baseline XGBoost** performed comparably, demonstrating that sampling was an effective alternative but did not drastically outperform weighting.
- The **fine-tuned XGBoost via RandomizedSearchCV** achieved the best overall results, improving recall for the minority (>$50K) class without sacrificing much precision, and yielding a strong ROC-AUC score.

Why These Features Matter For Walmart?

The most important factors in predicting income, through **feature importance and SHAP** (Lundberg & Lee, n.d.), include **tax filer status, weeks worked in a year, gender, education, family structure, and investment income.** Customers who file taxes and work more weeks tend to have higher and more stable incomes, making them good candidates for **loyalty programs and premium offers**. Education level also matters, as higher education often means greater earning potential, which can guide **marketing for higher-end products and financial services.** Family structure plays a big role too. The households with young children need baby products, groceries, and school supplies, while single customers often prefer convenience meals and entertainment. Investment income is another strong signal of financial security, pointing to opportunities for promoting high-margin products or credit services. Finally, age and gender can help **shape product recommendations**, such as health and wellness for older shoppers and lifestyle or tech products for younger ones.



The plots show the features significant to model training and an example of how these variables are contributing to predict an outcome for any given datapoint.

Limitations of the Study

Data Bias: The data is old and might not capture current trends. According to a study done by NCRC on racial wealth, Asians were one of the highest earning individuals nowadays in 2021 with a median income of $101k (National Community Reinvestment Coalition). However, the 94' census data shows that 0% of the Asians earned over $50k in 1994. **Modeling on current data would help capture the most recent trends.**

Precision-Recall Trade-offs: Improving recall reduced precision significantly which could be a marketing inefficiency in the longer run. **To improve the precision, I tried different thresholds (0.6, 0.7, 0.8)**. It dropped the recall value but there was no significant improvement in the precision. Therefore, finding a better model with a balanced precision and recall score would help mitigate this issue. **Oversampling for minority class using SMOTE or some other technique might help.**

Appendix

Incorporating Population Weights

To align the model with population representation, we trained XGBoost using census-provided sample weights and evaluated the results using weighted metrics. While ROC-AUC and overall accuracy remained similar to the imbalance-handled model (~0.936), recall for the >$50K income group was substantially lower (~40% vs >85% in imbalance-handled models). Given

Walmart's objective to maximize recall for high-income individuals, we prioritized imbalance handling strategies for the final model.

Anomaly detection using IsolationForest

We tested an unsupervised approach by treating the >$50K income group (~6% of the data) as anomalies using Isolation Forest. At the default threshold, the model achieved **ROC-AUC = 0.426** and recall for the positive class was only **6%**, far below the baseline XGBoost model. This confirmed that supervised methods with imbalance handling were more effective for this task.

2. Your retail client is interested in developing a segmentation model of the people represented in this dataset for marketing purposes as well. Using your favorite machine learning or data science techniques, create the segmentation model and demonstrate how the resulting groups differ from one another and how your retail client can use this model for marketing.

Approach

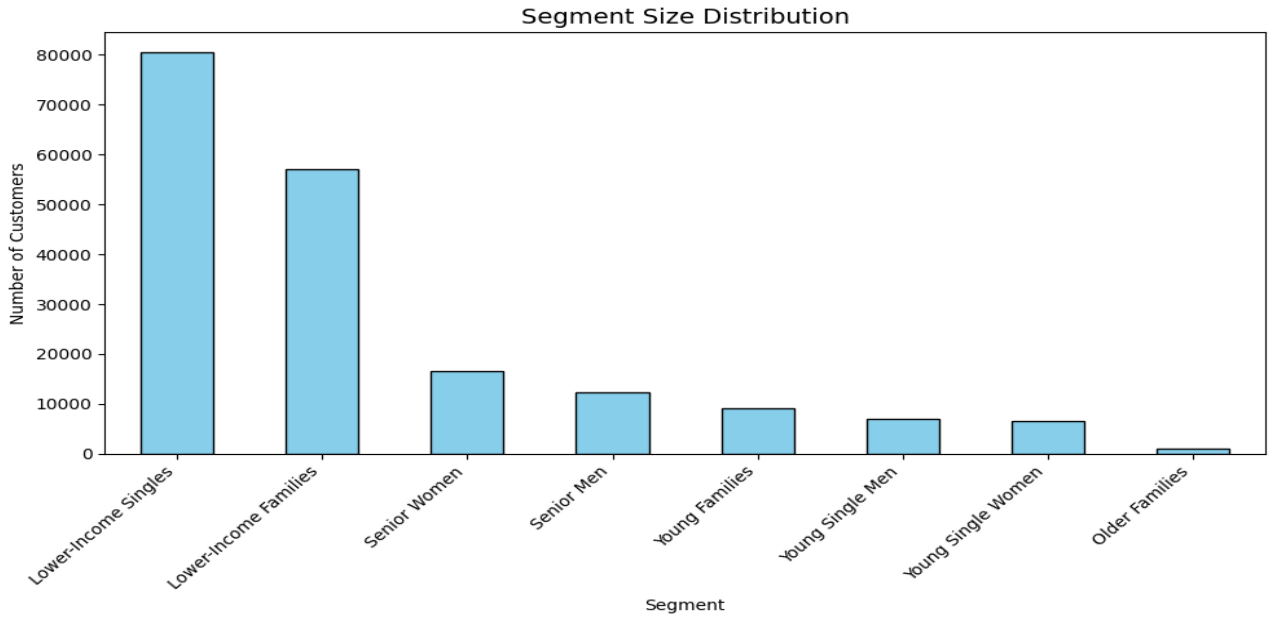
Walmart serves a wide range of customers, from those shopping for basic groceries to those buying higher-end products. This means our segmentation must reflect the diversity across income levels, education, and employment status, covering almost the entire social spectrum (except perhaps the top 1%). Creating very distinct and niche segments can be challenging with a purely unsupervised machine learning approach, as it may not fully capture the business logic we need. For this reason, a **rule-based segmentation approach** (Wedel & Kamakura, 2000) makes the most sense here. It gives us the flexibility to define meaningful groups based on our understanding of Walmart's customer base and their unique needs. It also gives us the transparency such that business teams can easily understand why a customer belongs to a specific segment.

Rule-Based Segmentation

We created 8 segments that covers Walmart's customer base:

- **Lower-Income Singles (42.3%)** – Price-sensitive, often unemployed or working few weeks; focus on low-cost essentials.
- **Lower-Income Families (30.0%)** – Budget-conscious households with kids; prefer bulk groceries, family packs, and savings programs.
- **Senior Women (8.7%)** – Women 60+, not employed; prioritize groceries, health products, and shopping convenience.
- **Senior Men (6.5%)** – Men 60+, retired; buy groceries, home maintenance, and affordable goods.

- **Young Families (4.8%)** – Adults 20–40 with kids; high demand for childcare items, groceries, and bundled offers.
- **Young Single Men (3.7%)** – Men under 40, no kids; spend on electronics, gaming, and convenience foods.
- **Young Single Women (3.5%)** – Women under 40, no kids; interested in fashion, beauty, and wellness products.
- **Older Families (0.5%)** – Adults 40–60 with kids (teens or college-aged); buy seasonal goods and home upgrades. (can be ignored)



Most Walmart customers belong to **value-driven segments**. Lower-Income Singles and Families make up over 70% of the customer base.

Key Characteristics of Segments

If we look deeper into employment and education, we uncover some key characteristics:

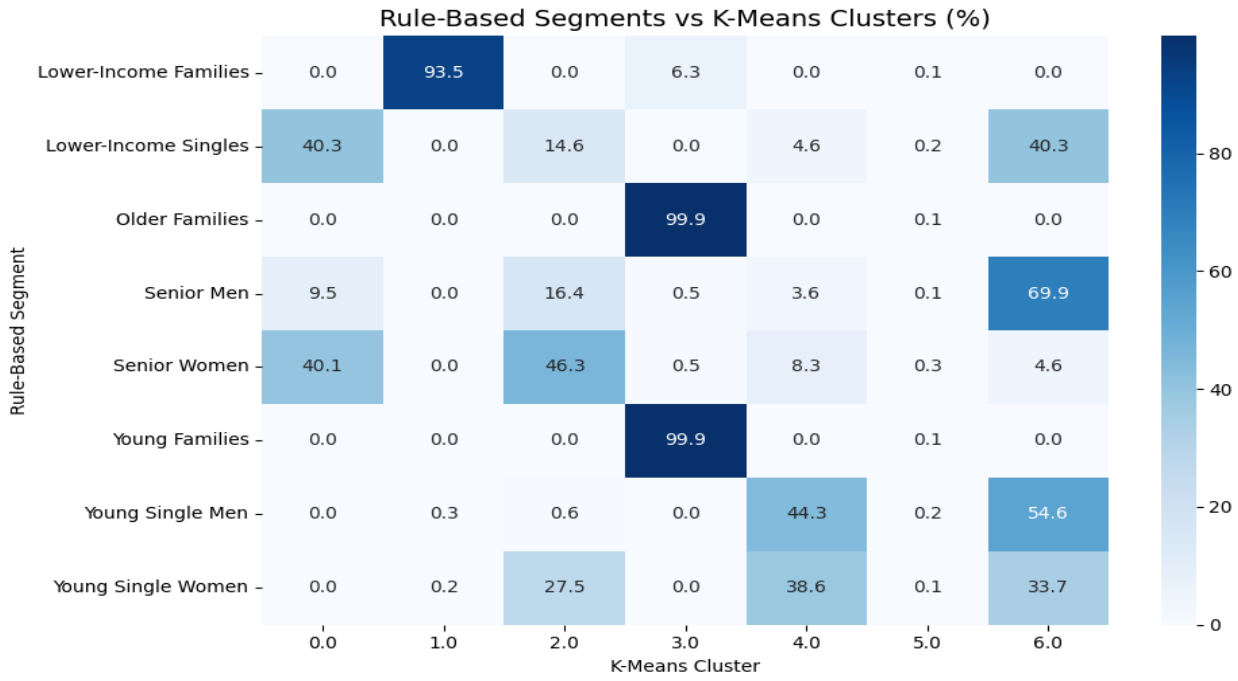
- **Low-Income Families** hardly work (avg 2.5 weeks/year), with 96% unemployed and have the lowest education levels. They need low-cost bundles and savings programs.
- **Young Singles** work the most weeks and have higher education, making them ideal for digital campaigns and product promotions.
- **Seniors** work very little, rely on fixed income, and prioritize health and groceries.

Validation with K-Means

To provide an empirical foundation for our rule-based segmentation, we ran a K-Means clustering algorithm (Arthur & Vassilvitskii, 2007) with 7 clusters to check if the clusters formed by K-Means aligned with those from our rule-based approach. Family segments showed strong

alignment, confirming that household structure is a key driver. Singles and seniors, however, were split across multiple clusters, indicating that factors like education and work stability also play an important role.

In conclusion, while our rule-based approach works well for family segments, singles and seniors may need further refinement. For example, separating young professionals from unemployed singles or distinguishing active retirees from those fully dependent on fixed income could improve targeting.



Marketing Recommendations

Here is how Walmart can use these segments for marketing:

- **Lower-Income Families and Singles:** Promote discounts, bulk offers, and SNAP-friendly programs (Arizona Department of Economic Security, n.d.).
- **Young Families:** Push bundled childcare products and loyalty programs.
- **Young Singles:** Target electronics, fashion, and convenience items through app-based offers.
- **Seniors:** Highlight pharmacy services, senior discounts, and easy ordering options.
- **Older Families:** Seasonal promotions and large household items.

Limitations

Over-simplification: Although it is important from a business perspective to improve interpretability of the model, a rule-based system might still be over-simplified and it might not reflect nuanced customer behaviors. **Trying a rule-based and clustering hybrid model would help**

No Behavioral Data: Behavioral data, i.e. purchase frequency, spending habits, etc. is completely missing from the census data which might affect the way clusters are formed. **Incorporating that data into segmentation would make better clusters.**

References

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 1027-1035.

Arizona Department of Economic Security. (n.d.). *Nutrition Assistance (formerly the Food Stamp Program)*. Retrieved August 24, 2025, from Arizona Department of Economic Security website.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, 4765-4774.National Community Reinvestment Coalition. (2023, August 23). *Racial wealth snapshot: Asian Americans and the racial wealth divide*. Retrieved from NCRC website.

Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations (2nd ed.). Kluwer Academic Publishers.

XGBoost Developers. (n.d.). *XGBoost documentation*. Retrieved August 24, 2025, from xgboost.readthedocs.io — the XGBoost documentation pages

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media.