JPMC Take Home Project

- Nishi Amish Modi

Executive Summary

Problem Statement - The client's goal is two-fold for marketing:

- **Classification:** Predict individuals with income less than $50K versus greater than $50K
- **Segmentation:** Develop distinct population groups for targeted marketing efforts

Data Introduction

- Weighted census data from the 1994 and 1995 Current Population Surveys (U.S. Census Bureau)
- Includes 40 variables, observation weights, and the binary income label

Quality Analysis - Data investigation revealed several issues requiring preparation:

- Missing values were present in several columns, ? character was also present
- Some numerical variables exhibited heavy left skewness
- Certain categorical variables had high cardinality (many unique values)

Data Preprocessing - series of techniques were applied to clean and transform the data:

- Missing values were handled through imputation
- Numerical features were standardized using scaling techniques
- Categorical variables were encoded using both One-Hot Encoding and Target Encoding

Classification Modeling - Six machine learning classification models were built and compared.; all models were designed within end-to-end pipelines to enable easy real-time production deployment

- **XGBoost** (eXtreme Gradient Boosting) achieved the superior performance.

Segmentation - A shallow Decision Tree with a maximum depth of 4 was trained. This choice provides high interpretability for marketing teams

- The model successfully identified **6 distinct leaf nodes**. These 6 nodes can be used as the final segments, representing groups with varying profiles suitable for specialized marketing

Data quality checks revealed some interesting insights

Data Summary -

- Observations : 199,523
- Features: 40, 1 label and 1 weight column
- Features with missing values: 16
- Features with Zero values : 19
- Features with negative values: 0

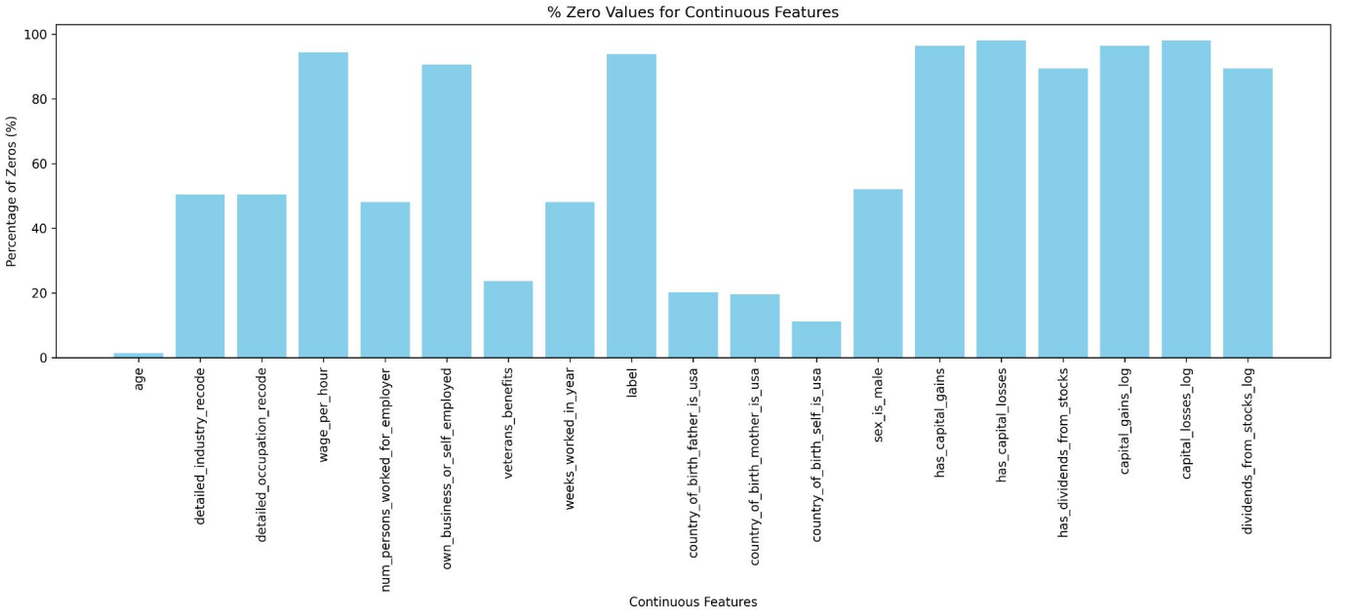
Missing Value characters identified -

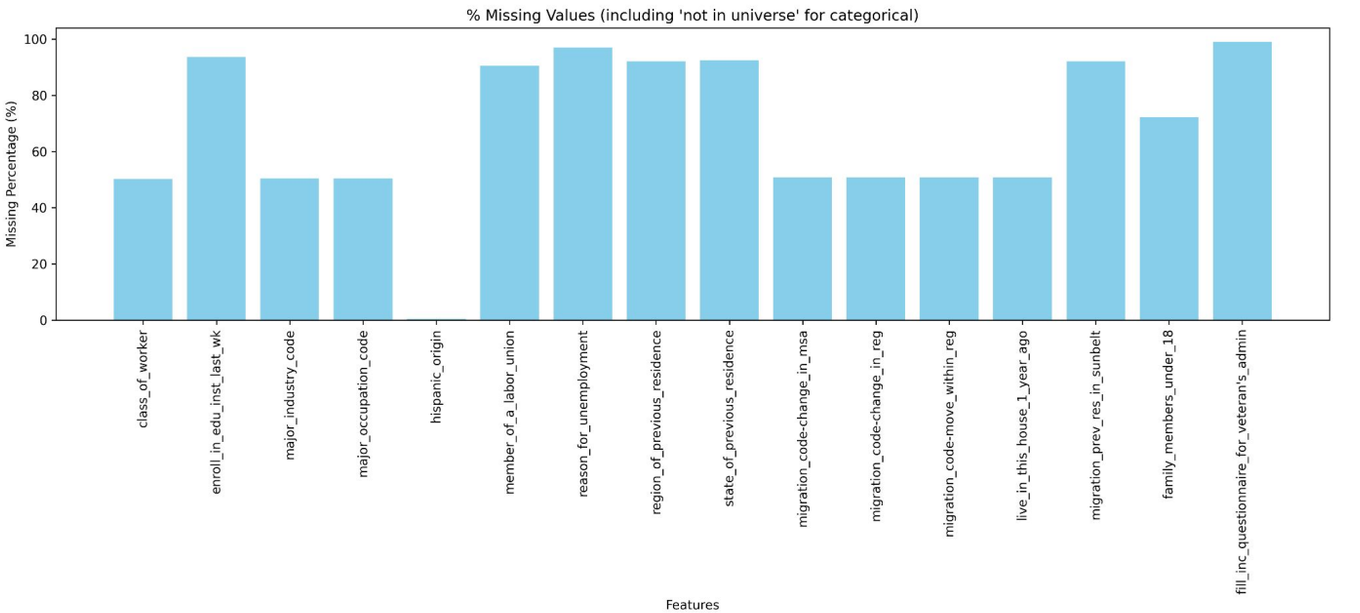
- ?
- nan
- Not In Universe

Insights

- ~50% missing values in four migration variables
- Over 60% zero values in financial/work-related continuous features (e.g., industry/occupation codes)
- Numerous categorical variables are dominated by "not in universe" defaults, requiring to treat them as missing value
- 24.7% of data in the veterans benefits column are outliers

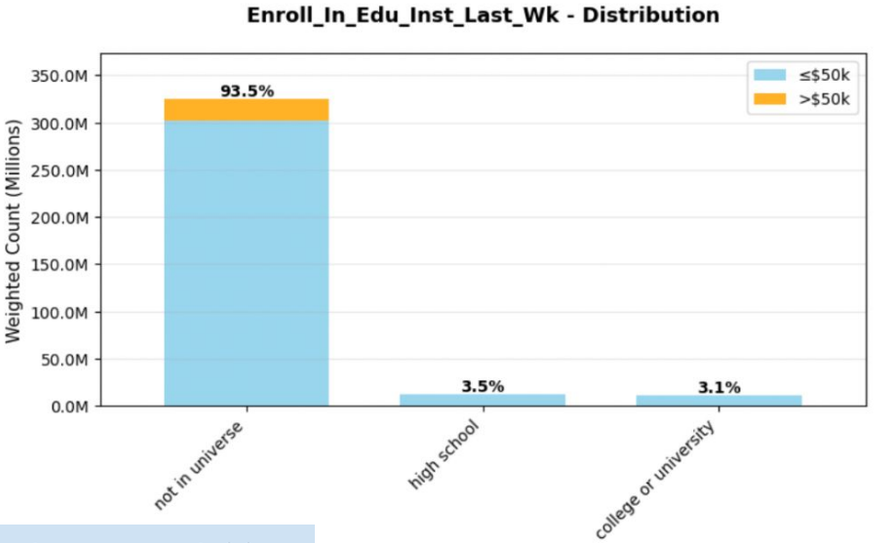
Imputation: Missing values in categorical features were created with category "not in universe" before any encoding

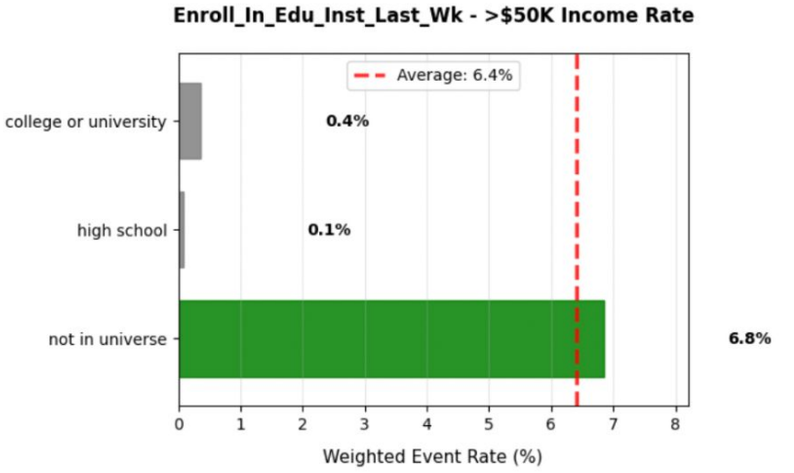




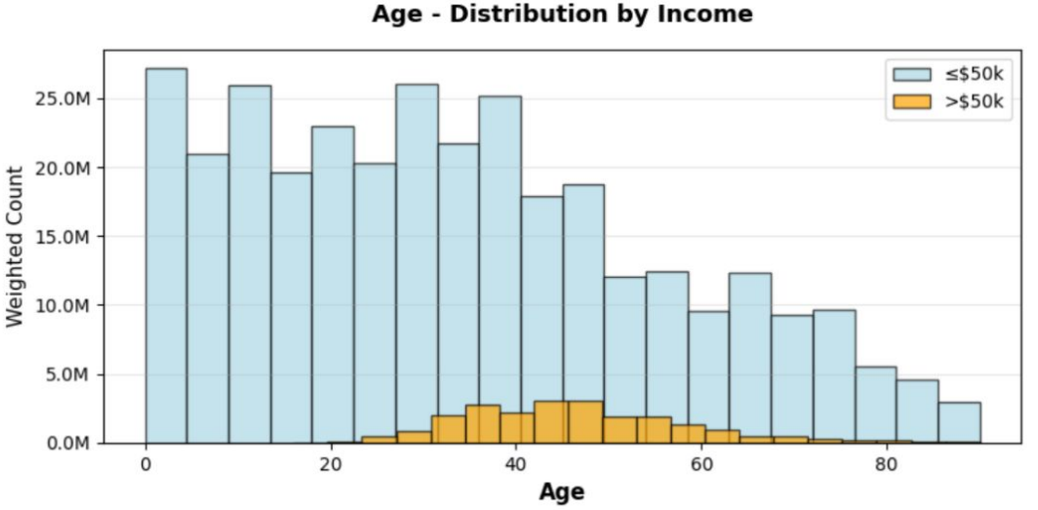
Analyzing variable distributions and event rates guides feature engineering, ensuring optimal model performance

Sample plots for categorical variables





Sample plots for continuous variables



Age - Statistics

Income	Mean	Std	Median	Q1	Q3	Min	Max
≤$50k	33.68	22.35	31.00	14.00	49.00	0.00	90.00
>$50k	46.13	11.77	45.00	38.00	53.00	16.00	90.00

Several data transformations were performed to continuous and categorical features

Variable(s)	Initial Variable Type	Transformation	Reasoning
Label (Target)	Categorical (2 unique values)	Converted to 1/0 binary Mapping - 50000+. to 1 and - 50000. to 0	Converted to binary target for modelling
Education	Categorical (17 unique values)	education level was converted using ordinal encoding to an integer scale ranging from 1 (lowest education level) to 17 (highest education level)	Implied order ascertain from education level of an individual
Country of Birth of Father, Country of Birth of Mother, Country of Birth of Self,	Categorical (each with 42 unique values)	Converted to 1/0 binary; USA vs. Non-USA	High cardinality with infrequent categories, binary categories with ~20% Non-USA values
State of Previous Residence	Categorical (42 unique values)	Combined into three categories; USA, Non-USA, and Missing	High cardinality with infrequent categories, three categories with ~80% missing values
Sex	Categorical (2 unique values)	Converted to 1/0 binary Mapping - Male to 1 and Female to 0	No need to create dummies as there are only two unique categories
Capital Gains, Capital Losses, Dividends from Stocks	Continuous	Non-zero indicator columns (0/1) for each; Log(1+x) transformation for each	To tackle skewness and maintain information about zeros

Weight feature was used along with undersampling and encoding for model training

Weight Feature Usage:

- It represents the relative distribution of people in the general population due to stratified sampling
- **<u>Not used directly</u>** as a model feature
- **Used in two ways:**
  - Calculated **weighted event rates** and **counts** for QA
  - Passed as **sample weights during model training**

Event Rates (Original Data):

- Unweighted event rate: 6.21%
- Weighted event rate: 6.41%

Sampling Strategy:

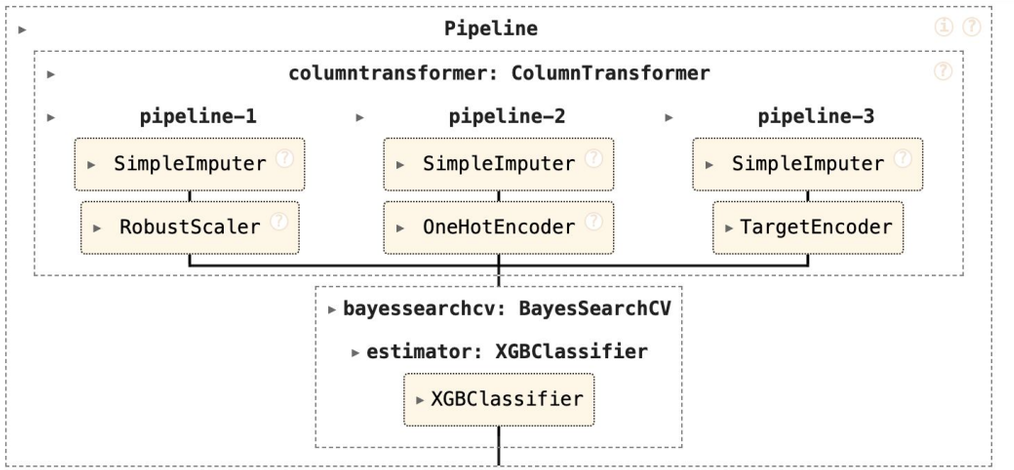
- Random undersampling applied to majority class (0s)
- Post-sampling event rates:
  - Unweighted: 10.30%
  - Weighted: 10.00%
- It address class imbalance and improve model learning

Encoding Type/Scaling	Variable(s)
Scaling using Robust Scaler	All continuous features
Target Encoding (Features will more than 9 categories)	Major Industry Code Major Occupation Code
Ordinal Encoding	Education
One Hot Encoding (Features will less than 10 categories)	Remaining categorical features
Dropped Features	*Major Industry Code* and *Major Occupation Code* (both have >45 unique categories and are covered by major industry/occupation features) *Year* (only 1994 and 1995 as values)

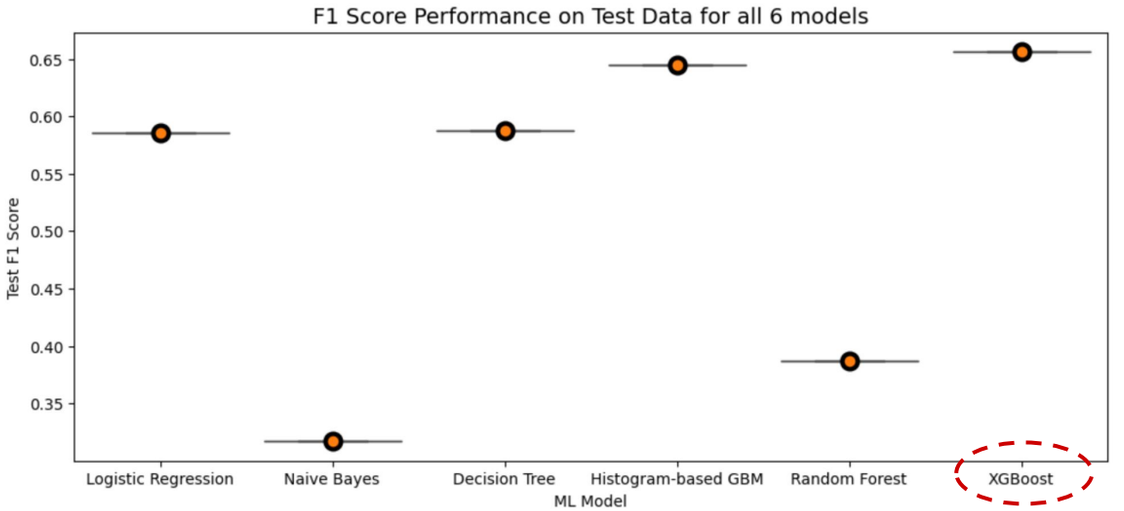
Six machine learning models were trained using scikit-learns pipelines that can leveraged for end-to-end deployment

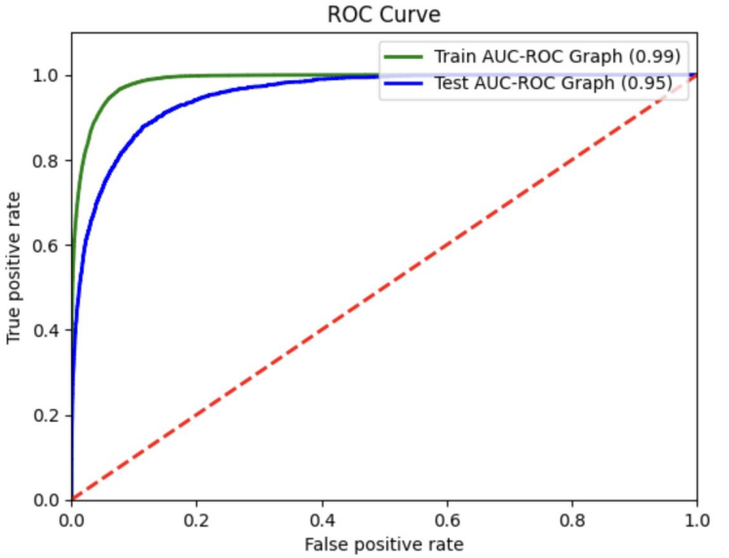
Model	HP Tuning Method	Tuned Hyperparameters
Logistic Regression	L2 Regularization K-fold Cross Validation	Used regularization
Naive bayes	-	-
Decision Tree	Bayesian Search with K-fold Cross Validation	max_depth ccp_alpha
Histogram based GBM	Bayesian Search with K-fold Cross Validation	l2_regularization learning_rate max_depth
RandomForest	Random Search with K-fold Cross Validation	max_ features max_depth n_estimators
XGBoost	Bayesian Search with K-fold Cross Validation	max_depth, reg_alpha, colsample_bytree, learning_rate gamma, reg_lambda, subsample, n_estimators, min_child_weight

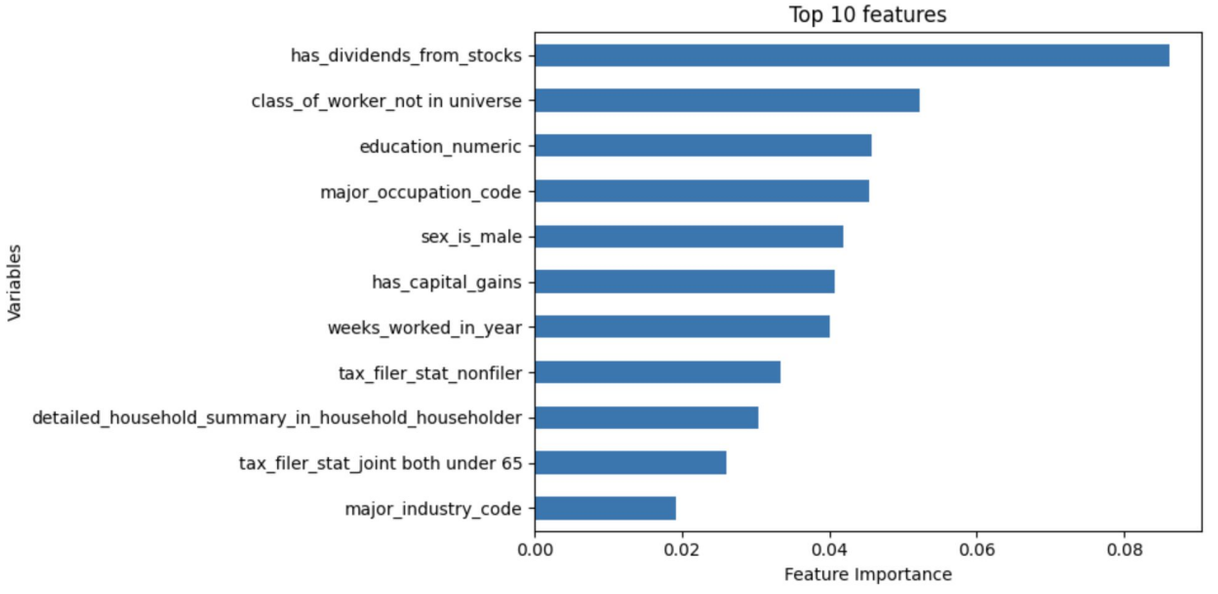
Sample Pipeline Architecture

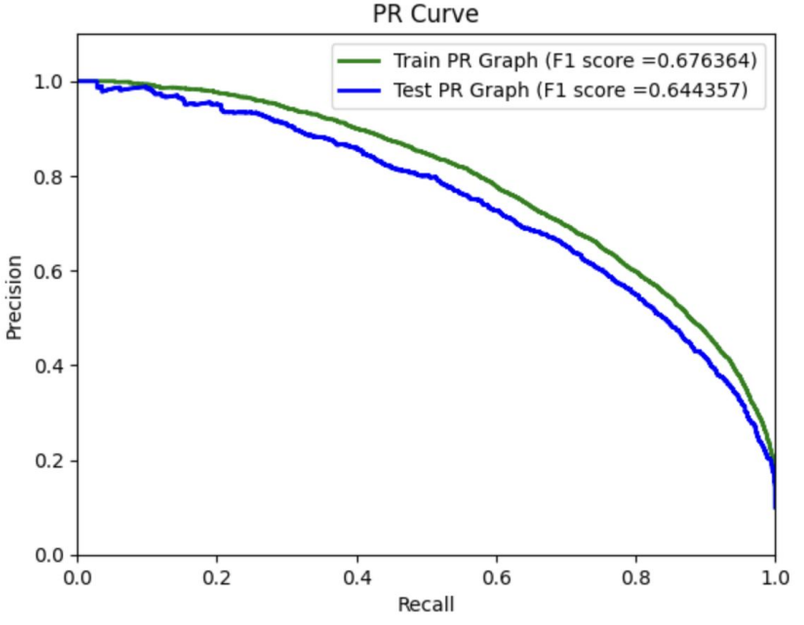


Education Status, Dividends, Occupation Code, and Sex are the four top features driving targeting people with Income greater than $50K



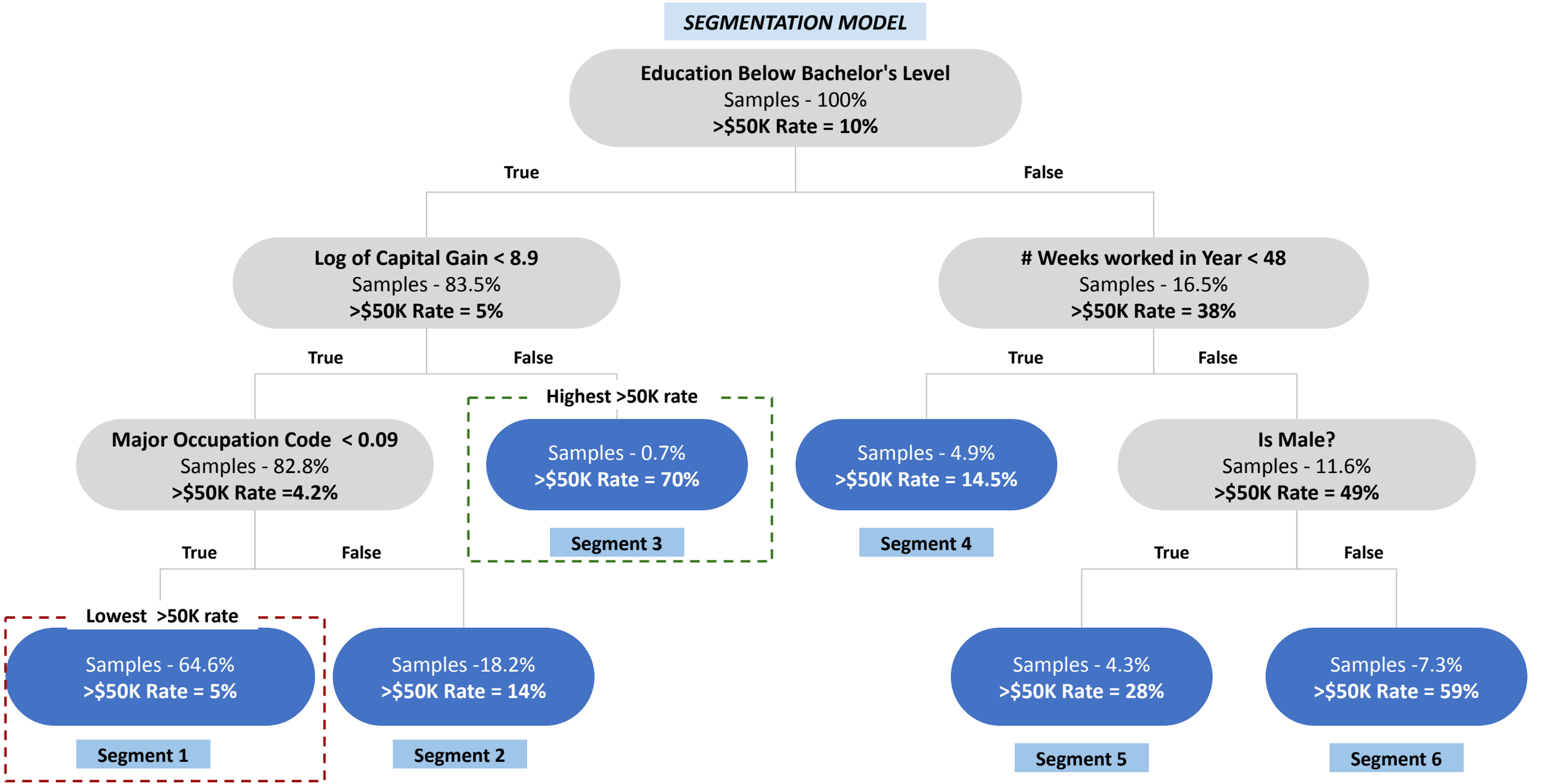






PR Curve, F1 Score, Precision and Recall are better suited than Accuracy

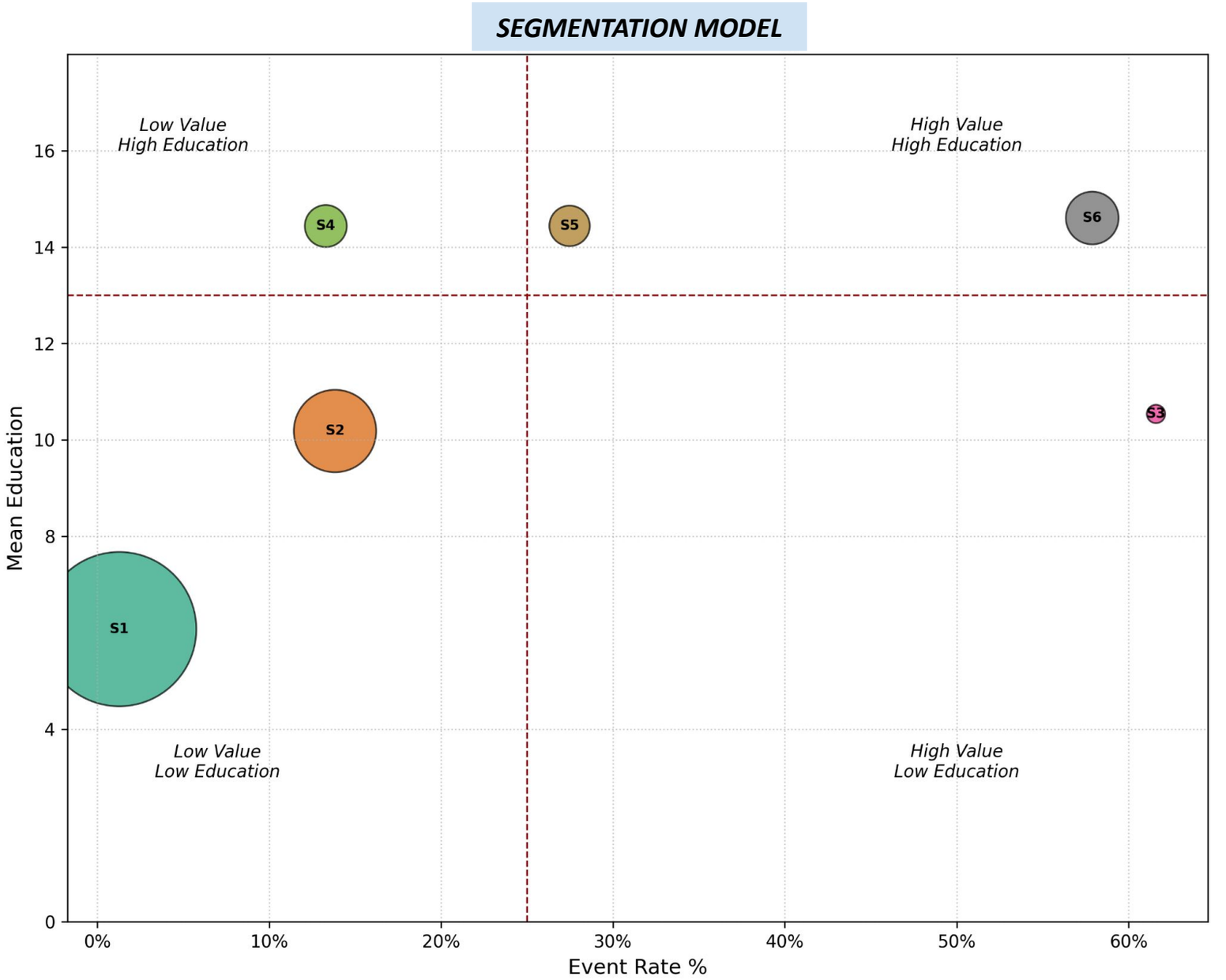
Shallow decision trees can be interpretatively used to define segments



Six distinct personas enable tailored marketing across varying Income and Education profiles

Segment	Persona	If-Then Logic	% Population	Income >$50K Rate	Mean Age	Mean of Education
1 Mass Market	Youngest, Low Education, Minimal Financial Activity; Require low-cost, automated campaigns	Education below Bachelor's Level AND Capital Gains ≤ 8.8613 AND Occupation Code ≤ 0.0941	64.1%	1.3%	31	6.1
2	Mid-Age (39), Moderate Education & Financial Activity	Education below Bachelor's Level AND Capital Gains ≤ 8.8613 AND Occupation Code > 0.0941	18.3%	13.8%	39	10.2
3 Elite	Elite, Highest Priority, Low Volume, 100% has Capital Gains; Slightly older generation	Education below Bachelor's Level AND Capital Gains > 8.8613	0.9%	61.6%	47	10.5
4	Older and highly educated low earners. Oldest Segment (49.7), Highly Educated, High Dividends	Education Bachelor's Level or above AND Weeks Worked ≤ 47.5	4.8%	13.3%	50	14.4
5 Female medium earner	100% Female, Highly Educated, Dividend Active	Education Bachelor's Level or above AND Weeks Worked > 47.5 AND Is Female	4.4%	27.5%	40	14.4
6 Premium Male Segment	100% Male, Highly Educated, High Financial Activity	Education Bachelor's Level or above AND Weeks Worked > 47.5 AND Is Male	7.5%	57.9%	43	14.6

High-value and highly educated groups (segment 5 and 6) represent key opportunity despite small volumes



Bubble Size - Population %

Appendix

References

- https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
- https://archive.ics.uci.edu/dataset/2/adult
- https://matplotlib.org/
- Google for syntax and function APIs