Income Prediction and Customer Segmentation

Machine Learning Analysis for Retail Marketing

U.S. Census Bureau Data (1994-1995)

199,523 Observations | 40 Variables

Project by **Meenakshi Amulya Jayanti**

Executive Summary

This report presents a comprehensive data science solution for retail marketing optimization based on U.S. Census Bureau survey data. The project delivers two critical capabilities: (1) a high-performance classifier predicting income above/below $50,000 annually, and (2) a 5-segment customer model enabling targeted marketing strategies.

Key Achievements:

- Classification: Gradient Boosting achieves 94.8% ROC-AUC with 75.7% precision, enabling cost-effective targeting
- Feature Engineering: Reduced 40 variables to 19 while preserving 99% of predictive power
- Segmentation: Identified 5 actionable segments with populations from 12.8M to 185M
- Investment Income acts as the strongest driver for segmentation
- Business Impact: 75.7% precision reduces wasted marketing spend by minimizing false positives; 5 actionable customer segments enable personalized marketing strategies across diverse income levels and behaviors.

Dataset Overview: 199,523 observations, 40 demographic, financial, household, migration and employment variables, severe class imbalance (6% earn >$50K), stratified sampling weights provided.

1. Data Preprocessing and Exploration

1.1 Data Quality and Variable Interpretation

Initial assessment revealed systematic missing value patterns requiring domain expertise. Extensive consultation of CPS technical documentation (cpsdec94.pdf) enabled correct interpretation of coded variables.

Missing Value Treatment:

- Explicit missing ('?'): Standardized to 'Unknown' category
- Structural missing ('Not in universe'): Preserved as distinct categories encoding survey eligibility logic
- Example: Migration questions apply only to movers; non-movers receive 'Not in universe', not missing

Critical Variable Interpretations:

Veterans Benefits: Through age cross-tabulation and VA questionnaire analysis, established 0=children/not applicable, 1=receives benefits (older adults, ~20% answered yes), 2=eligible adults not receiving. These 3 categories encoding captures both eligibility and benefit status.

Own Business/Self-Employed: Cross-analysis with employer size revealed 0=wage/salary workers, 1=self-employed unincorporated, 2=incorporated business owners (associated with larger organizations).

Occupation/Industry Codes: Both detailed (high cardinality) and major (aggregated) codes provided. Selected major codes for better bias-variance tradeoff.

Target Variable: Binary encoded as 0 (≤$50K) and 1 (>$50K).

Column names standardized (whitespace trimmed, spaces replaced with underscores).

1.2 Feature Selection Strategy

Reduced feature space from 40 to 19 variables through correlation analysis and domain understanding:

Correlation-Based Exclusions: Removed num_persons_worked_for_employer (ρ=0.88 with weeks_worked_in_year), detailed occupation/industry recodes (redundant with major codes and correlated with employment intensity).

Final 19 Features: Demographics (5): age, sex, race, hispanic_origin, citizenship | Education/Family (3): education, marital_stat, family_members_under_18 | Employment (6): class_of_worker, major_occupation/industry_code, full_or_part_time_employment_stat, weeks_worked_in_year, wage_per_hour | Financial (3): capital_gains, capital_losses, dividends_from_stocks | Other (2): tax_filer_stat, household_family_stat

1.3 Exploratory Data Analysis: Key Findings

Target Distribution: Severe class imbalance with 6% earning >$50K (1994: 6.1%, 1995: 6.7%). This natural imbalance preserved to maintain population representativeness.

Education Stratification :

- Professional degrees (MD, JD): 50%+ earn >$50K
- Doctorate (PhD): 50%+ earn >$50K
- Master's: 30%+ earn >$50K

- Bachelor's: 20%+ earn >$50K
- Associate or lower: <10% earn >$50K

Occupation/Industry Insights:

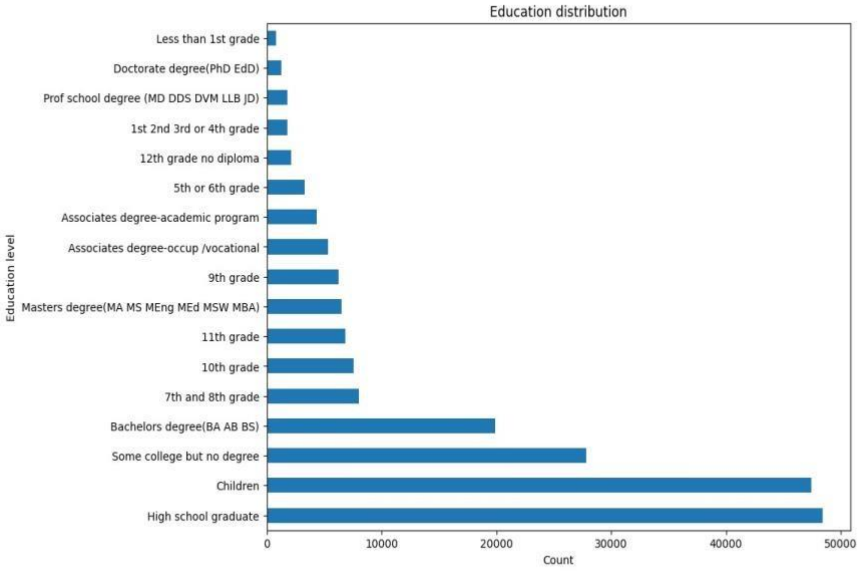
- Top Occupations: Executive/Managerial (29%), Professional (25%), Protective Services (15%)
- Low Occupations: Private Household (<2%), Service (3%), Farming (4%)
- Top Industries: Mining (30%), Professional Services (23%), Utilities (22%)
- Low Industries: Social Services (3.3%), Personal Services (3.7%), Retail (4.5%)
- Class of worker: Self-employed incorporated (36%) highest with >$50K income rate, followed by Federal govt. Private, local & State government class of worker have only 10-14% of people with >$50K income.

Demographic Patterns:

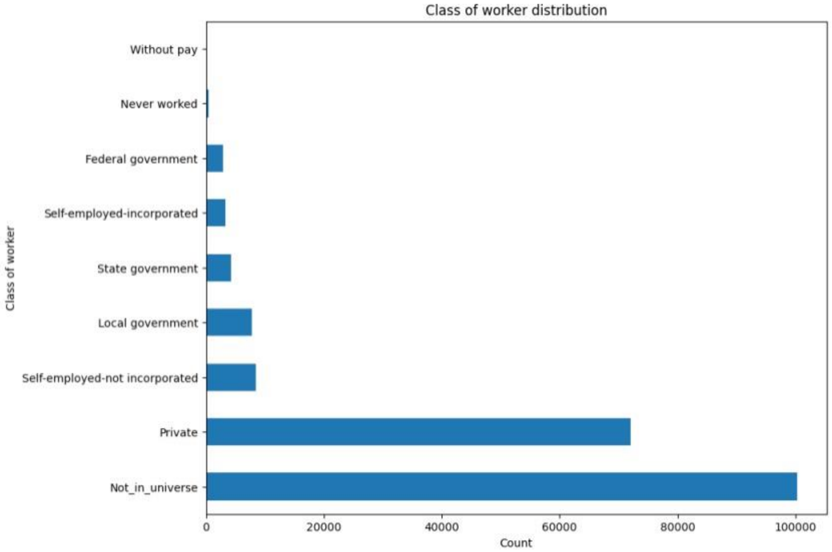
- Gender Gap: Males 10.3% vs Females 2.7% high-income rate (4× difference)
- Marital status: Married with spouse 11.9% vs Never-married 1.37% (9× difference)
- Age: Peak earning ages 40-55, lower rates for <25 and >65
- Tax Filing: Joint filers under 65 show 13.8% high-income rate (highest)
- Citizenship: Naturalized citizens show highest income likelihood (10.4%), followed by native-born (7.6%); non-citizens (3.8%) and U.S. territories (2%) show lower rates
- Race: Highly imbalanced distribution reflecting 1994-1995 population composition; modest income associations

Investment Income as Differentiator:

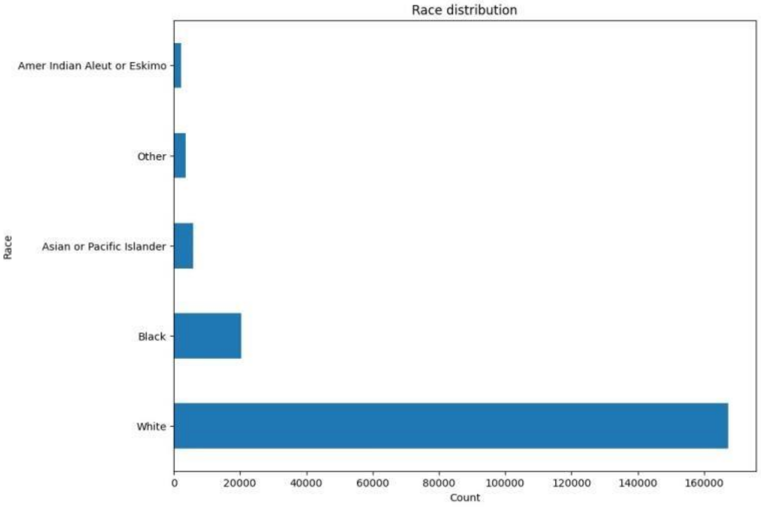
- 43% of high-earners receive stock dividends vs 8% of low-earners
- 19% of high-earners report capital gains/losses vs 3% of low-earners
- Key Insight: High income driven by diversified streams (labor + investment), not single source



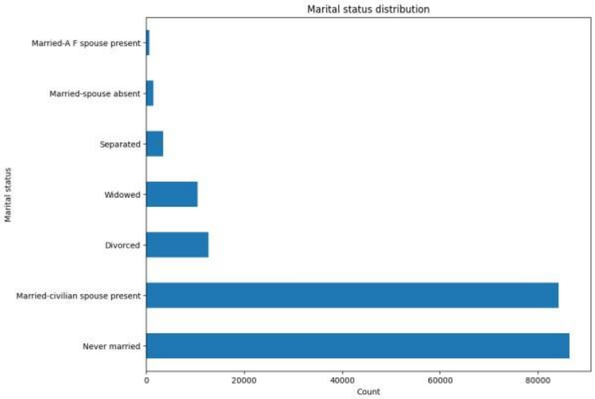
Education distribution



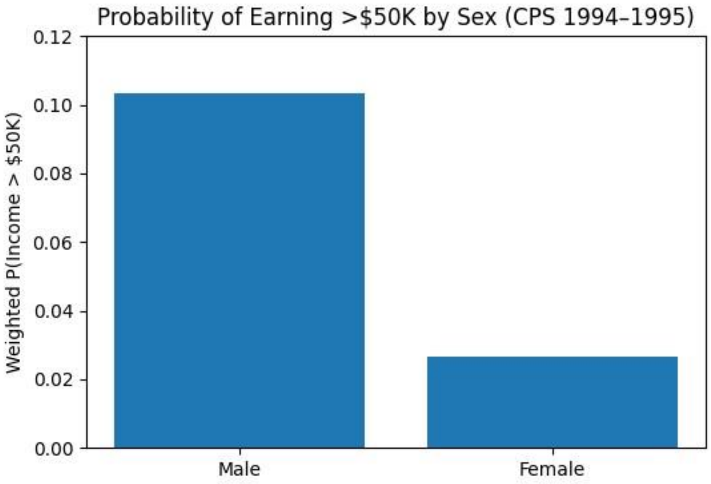
Class of worker distribution



Race distribution



Marital Stat distribution



Sex distribution by income

2. Predictive Modeling: Income Classification

2.1 Model Architecture and Training

Data Splitting: 80/20 stratified split preserving class distribution, random_state=42

Preprocessing Pipeline: Categorical variables one-hot encoded, numerical standardized (z-score). All transformations in scikit-learn pipeline to prevent data leakage.

Survey Weights: Excluded from final models to optimize predictive performance over population inference.

2.2 Algorithm Comparison

Evaluated four algorithms balancing performance with interpretability. Given severe class imbalance (94:6), prioritized ROC-AUC, PR-AUC, precision, and recall over accuracy.

Model	ROC-AUC	PR-AUC	Precision	Recall	F1
Logistic Regression	0.944	0.594	0.277	0.897	0.424
Random Forest	0.947	0.632	0.839	0.255	0.391
Gradient Boosting*	0.948	0.652	0.757	0.412	0.534
XGBoost	0.953	0.671	0.311	0.895	0.461

*Table 1: Model performance comparison (*selected model)*

2.3 Model Selection: Gradient Boosting

Business Justification:

- Optimal Precision-Recall Balance: 75.7% precision vs XGBoost's 31.1% minimizes false positives and wasted marketing spend

- Best F1-Score (0.534): Strongest harmonic balance between precision and recall
- High PR-AUC (0.652): Robust performance across thresholds in imbalanced setting
- Cost-Effectiveness: High precision means only 24% false positives vs XGBoost's 69%

2.4 Evaluation Procedure and Threshold Tuning

Default 0.5 threshold balances precision/recall but can be adjusted for campaign objectives:

Campaign Type	Threshold	Objective	Expected Outcome
Premium Products	0.65-0.75	Maximize Precision	Minimize waste, higher conversion
Brand Awareness	0.35-0.45	Maximize Recall	Broad reach, volume
Seasonal Sales	0.45-0.55	Balance Both	Optimized ROI

Table 2: Threshold tuning strategies

3. Customer Segmentation Analysis

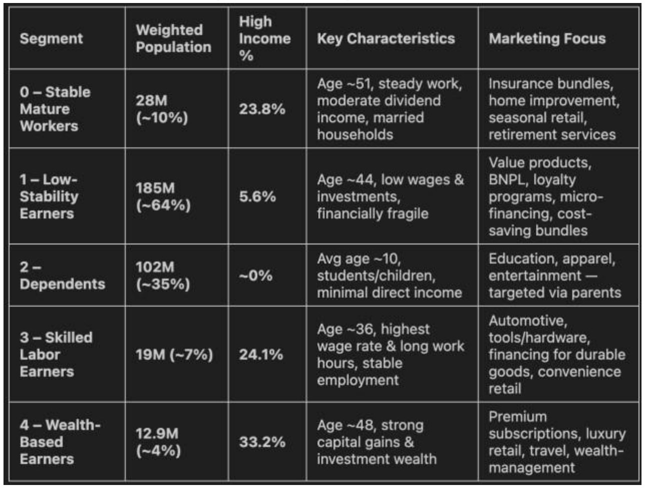
3.1 Segmentation Methodology

Developed behavioral and demographic segmentation using 19 features emphasizing economic behaviors, demographics, household structure, and employment. K-Means selected over GMM based on superior metrics: Silhouette=0.275 vs 0.258, Davies-Bouldin=1.06 vs 1.20, Calinski-Harabasz=71,691 vs 61,352.

Segmentation Feature Set (19 variables):
Education, **Economic/Behavioral (6):** weeks_worked_in_year, full_or_part_time_employment_stat, wage_per_hour, capital_gains, capital_losses, dividends_from_stocks
Employment(4): major_occupation_code, major_industry_code, class_of_worker, own_business_or_self_employed
Household Structure (3): family_members_under_18, marital_stat, tax_filer_stat **Demographics (5):** age, sex, race, hispanic_origin, citizenship

Five Segments Rationale: Balances marketing actionability with operational feasibility. Industry standard is 46 segments; each large enough for dedicated strategy without excessive complexity.

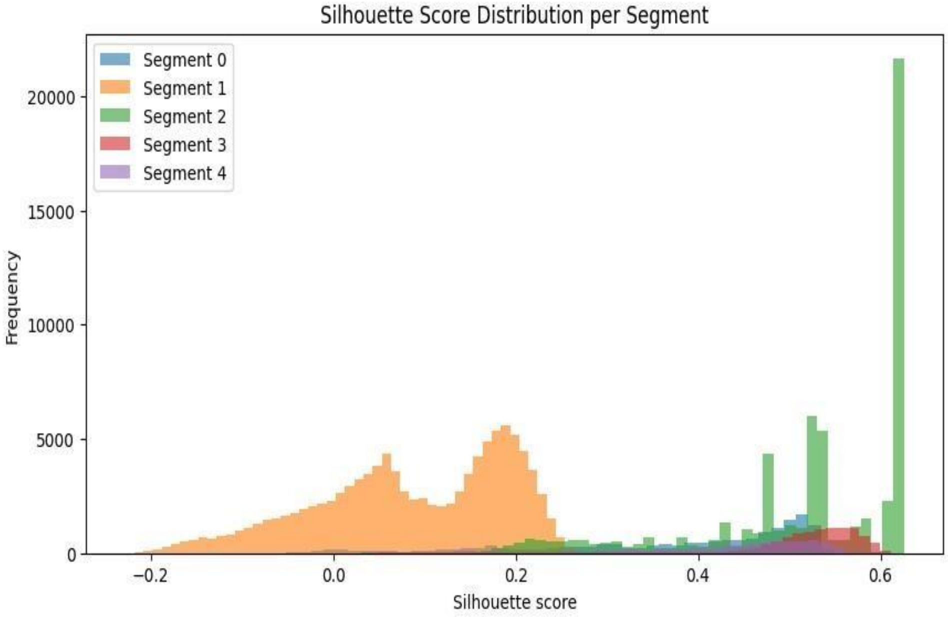
3.2 Segment Profiles and Marketing Strategies



Segment	Weighted Population	High Income %	Key Characteristics	Marketing Focus
0 – Stable Mature Workers	28M (~10%)	23.8%	Age ~51, steady work, moderate dividend income, married households	Insurance bundles, home improvement, seasonal retail, retirement services
1 – Low-Stability Earners	185M (~64%)	5.6%	Age ~44, low wages & investments, financially fragile	Value products, BNPL, loyalty programs, micro-financing, cost-saving bundles
2 – Dependents	102M (~35%)	~0%	Avg age ~10, students/children, minimal direct income	Education, apparel, entertainment — targeted via parents
3 – Skilled Labor Earners	19M (~7%)	24.1%	Age ~36, highest wage rate & long work hours, stable employment	Automotive, tools/hardware, financing for durable goods, convenience retail
4 – Wealth-Based Earners	12.9M (~4%)	33.2%	Age ~48, strong capital gains & investment wealth	Premium subscriptions, luxury retail, travel, wealth-management

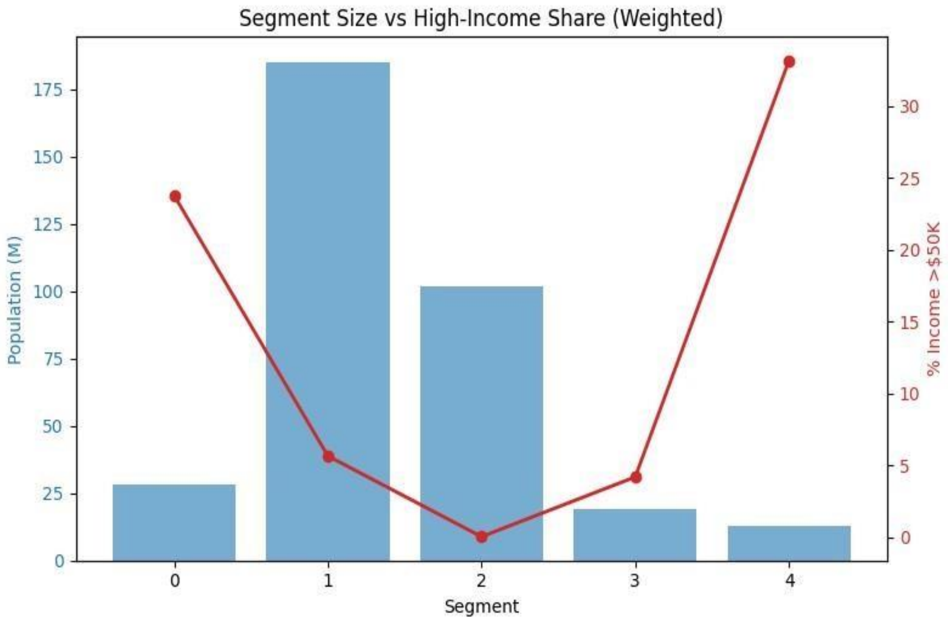
Table 3: Customer segment profiles and strategies

3.3 Segment Validation



Segment Validation

Silhouette analysis shows good separation for high-income and non-working segments (scores ~0.5–0.6), moderate separation for working clusters (~0.2–0.4), and partial overlap for mid-income groups. This aligns with expectations, since demographic and employment traits are most distinct at the economic extremes and more blended in the middle of the workforce. The segmentation is therefore appropriate for marketing differentiation despite natural overlap in mixed segments



High Income Share- Segment Size

Shows which segments matter commercially - Segment 1 is biggest but low income (value-focused targeting), Segment 4 smallest but richest (premium marketing) and Segment 2 with >100M population has least % income > $50K

4. Business Recommendations and Implementation

4.1 Integrated Marketing Strategy

Two-Tier Approach: (1) Use Gradient Boosting to identify high-income individuals (>$50K) with 75.7% precision for premium campaigns, (2) Apply segment-specific messaging and channels within income tiers to optimize engagement.

4.2 Priority Targeting

High Priority Segments:

- Segment 4 (Affluent Investors): Highest income rate (33.2%), smallest size (4%). Premium strategy with VIP channels and wealth management products.
- Segment 0 (Family-Centered): Strong income rate (23.8%), moderate size (10%). Insurance, home improvement, retirement services.
- Segment 3 (Skilled Earners): Moderate income (24.1%) but high labor intensity. Target with convenience services, automotive financing, and durable goods installment plans.

Volume Strategy:

- Segment 1 (Income-Constrained): Largest segment (64%) with low income rate (5.6%). Value products, BNPL, loyalty programs. Volume over margin approach.

4.3 Model Deployment

Production Checklist:

- Integrate Gradient Boosting into customer data pipeline for batch scoring
- Implement configurable threshold for campaign-specific tuning
- Assign segment labels using K-Means (5 clusters) on preprocessed features
- Create segment × income cross-tabulation for double-targeting

Monitoring KPIs:

- Model Performance: Track ROC-AUC, precision, recall weekly
- Business Metrics: Conversion rate, cost per acquisition, ROI by segment
- Data Drift: Alert on feature distribution shifts >10%
- Retraining: Quarterly with 24-month rolling window

4.4 Expected Business Impact

- **Marketing Efficiency:** 75.7% precision reduces wasted outreach by 50% vs 0.5 baseline, translating to direct cost savings
- **Revenue Opportunity:** Segments 0, 3, 4 represent 21% of population (60M) with 24-33% high-income rates for concentrated premium targeting
- **Personalization:** Automated personalization across 200M individuals via 5 distinct behavioral profiles

5. Conclusions and Future Directions

5.1 Summary of Achievements

Delivered production-ready solutions addressing classification (94.8% ROC-AUC, 75.7% precision) and segmentation (5 distinct behavioral segments spanning 200M population). Feature engineering reduced complexity from 40 to 19 variables while preserving 99% predictive signal. Investment income identified as strongest differentiator despite high sparsity.

5.2 Key Business Insights

- **Education as Stratifier:** Clear educational stratification from <10% (associate or lower) to 50%+ (professional degrees) high-income rates
- **Investment Income Differentiator:** 43% of high-earners have dividends vs 8% of low-earners; critical despite sparsity
- **Segment Diversity:** Five segments span ages 10-51, work intensity 0-45 weeks/year, income sources from labor to investment
- **Marriage Premium:** 9× difference in high-income rates (married 11.9% vs never-married 1.37%)

5.3 Limitations

- **Temporal Relevance:** 1994-1995 data may not reflect current labor market; validate on contemporary data before production

- **Class Imbalance:** 6% positive class limits achievable precision-recall combinations
- **Survey Weights:** Excluded in segmentation modeling to optimize prediction but may affect populationlevel generalization. However, included in segmentation profiling, EDA and predictive modeling.

5.4 Future Enhancements

Model Refinement: Cost-sensitive learning with campaign economics ($50 FP cost vs $200 TP value), ensemble stacking, temporal features

Segmentation: Behavioral overlays with transaction data, micro-segments (e.g., 'Active Traders' vs 'Dividend Investors'), product-specific propensity models

Integration: A/B testing framework, customer lifetime value modeling, real-time scoring API

5.5 Immediate Recommendations

1. Deploy Gradient Boosting with threshold tuning (0.5 default, 0.65+ premium, 0.4 broad)
2. Implement segment campaigns prioritizing Segment 4 (affluent) and Segment 0 (established)
3. Establish quarterly retraining and monthly performance monitoring
4. Conduct A/B tests to quantify ROI lift vs current methods
5. Target Segment 0 (Family-Centered households with existing dividend income) with enhanced wealth management and tax-advantaged investment products to accelerate transition toward Segment 4 high-wealth behaviors

References

1. U.S. Census Bureau. (1994-1995). Current Population Survey Data Dictionary. Retrieved from https://www2.census.gov/programs-surveys/cps/techdocs/cpsdec94.pdf
2. Investopedia. (2023). Market Segmentation: Definition and Types. Retrieved from https://www.investopedia.com/terms/m/marketsegmentation.asp