CLIENT REPORT (DETAILED JUSTIFICATION VERSION)

Retail Customer Income Prediction & Segmentation Strategy

1. BUSINESS CONTEXT

This project develops two complementary analytical models using a U.S. Census demographic dataset. The first model predicts whether an individual earns more or less than $50,000 per year, which serves as a proxy for purchasing power. The second model performs customer segmentation to group people into meaningful categories for targeted marketing strategies. The dataset contains survey records from 1994–1995 with approximately forty demographic and employment-related attributes, along with a sampling weight and an income label. Rather than focusing purely on model accuracy, the project emphasizes reliable data preparation, interpretable modelling, and alignment with real marketing decision-making.

The current retail marketing strategy relies on broad campaigns sent to large populations without differentiation. This approach results in high marketing expenditure, low conversion rates, customer irritation, and gradual loss of brand trust. Sending the same promotion to everyone treats all customers as equal despite clear differences in purchasing capability and needs. Therefore, the business objective is not simply to build a predictive model, but to improve decision quality — specifically, determining *who should be contacted, how they should be contacted, and when marketing effort should be avoided*.

To address this, the system was designed around three core principles. First, it estimates purchasing power through income classification so the company can prioritize high-value customers. Second, it identifies natural customer groups through segmentation so messaging can be personalized rather than generic. Third, it ensures ethical and fair targeting so the company avoids biased or overly aggressive strategies that could harm long-term customer relationships. In short, the purpose of the models is not only prediction, but smarter and more responsible marketing decisions

2. PROBLEM INTERPRETATION

Although the task involves building predictive models, the underlying challenge is not purely a machine learning optimization problem but a decision-making problem. The company does not benefit from simply maximizing prediction accuracy; instead, it must balance marketing cost and opportunity. An overly aggressive model would label too many customers as high-income, causing the business to spend marketing resources on people unlikely to convert. Conversely, an overly conservative model would identify very few customers, leading to missed revenue opportunities. Because both extremes harm the business in different ways, the objective is to optimize balanced business value rather than raw predictive performance. The model is therefore evaluated and tuned to support practical decision making instead of achieving the highest possible accuracy score.

3.Data Exploration, Preparation, and Characteristics

The dataset was provided as a raw census data file along with a separate column definition file. After loading the data, it contained 199,523 records and 42 attributes, including demographic information, employment details, a sampling weight, and an income label. Initial inspection revealed inconsistencies common in survey data, such as extra spaces and missing markers ("?"). These were standardized and replaced with a consistent "Unknown" category to maintain dataset integrity without discarding valuable observations.

To make the prediction task meaningful for business usage, the original income label was converted into a binary variable (income high), indicating whether a person earns more than $50K annually. Additionally, a small amount of feature engineering was performed, such as deriving a has kids' indicator from the number of family members under 18, which can influence purchasing behaviour. Several variables were intentionally removed, including capital gains, capital losses, wage per hour, and detailed industry or residence history fields. Although these variables improve raw predictive performance, they either reveal outcome-related information or would not realistically be available at marketing decision time. Removing them ensures the model simulates real deployment conditions and avoids data leakage.

The cleaned dataset was divided into training, validation, and test sets in a 60/20/20 ratio using stratified sampling to preserve income distribution across splits. During this step, an important characteristic of the data became evident: only about 6% of individuals belong to the high-income category. This heavy class imbalance has significant implications for modelling. A naive classifier predicting every person as low-income would achieve approximately 94% accuracy without learning anything useful. Therefore, accuracy alone would be misleading as a performance metric. Instead, evaluation focuses on balancing precision and recall through the F1 score so that the model can meaningfully identify high-value customers without excessively misclassifying others.

4.Feature Engineering Strategy

A critical step in this project was determining which variables the model should be allowed to use. Several attributes in the dataset directly or indirectly revealed financial outcomes, such as **capital gains, capital losses, wage per hour, and detailed industry and occupation codes**. While these variables significantly improve predictive accuracy, they would not realistically be available when the company is deciding whether to market to a customer. Using them would create *data leakage*, meaning the model would appear highly accurate during testing but would fail once deployed in practice.

For this reason, we intentionally removed financial outcome–related variables and certain highly specific historical attributes, including **capital gains, capital losses, wage per hour, detailed occupation codes, detailed industry codes, and prior residence information**. These features either reveal post-outcome financial status or rely on information that marketing teams typically do not possess at decision time. The model was instead restricted to stable demographic and employment indicators such as age, work status, education level, marital status, and employment consistency.

The purpose of this restriction was to simulate a realistic marketing environment: predicting purchasing capability *before* knowing income. By limiting the model to deployable features, the resulting performance reflects what the business can expect in production rather than an artificially inflated laboratory result. Although this choice slightly reduces accuracy, it ensures reliable behaviour, fairness, and trustworthy decision-making in real operations.

5.Classification Model Approach and Justification

To understand how well the dataset could predict purchasing capability, we began with a baseline model using logistic regression. All categorical variables were one-hot encoded and numerical features were standardized to ensure fair contribution to the linear model. The purpose of this step was not to build the final solution, but to verify whether meaningful predictive signal existed in the data. The baseline achieved approximately 95% overall accuracy; however, this was misleading due to the strong class imbalance. The model largely predicted the majority low-income class and failed to identify many high-income individuals, producing an F1 score of about 0.44 for the high-income group. This indicated that a simple linear decision boundary was insufficient and that income relationships in the data were more complex and nonlinear.

To better capture these relationships, we implemented a Gradient Boosted Decision Tree model using LightGBM. This model was selected because it handles heterogeneous tabular data effectively, captures interactions between demographic and employment features, and performs well in imbalanced classification problems. The preprocessing pipeline standardized numeric features and one-hot encoded categorical variables, and class imbalance was addressed using the scale_pos_weight parameter so the model would pay appropriate attention to the minority high-income group.

Instead of using the default probability threshold of 0.5, we tuned the decision threshold on the validation set to optimize the F1 score for identifying high-income customers. This process resulted in an optimal threshold near 0.85, reflecting the need for more confident predictions before labelling someone as a high-value customer. When evaluated on the test set, the tuned LightGBM model achieved roughly 94% accuracy and improved the high-income F1 score to approximately 0.56, with precision around 0.52 and recall around 0.59. The Precision-Recall AUC of about 0.57 confirmed a balanced ability to identify valuable customers while controlling false targeting.

The progression from logistic regression to LightGBM demonstrates that the problem is not linearly separable and benefits from nonlinear modelling. LightGBM also offers interpretable feature importance, allowing the business to understand which factors influence purchasing capability, similar to models used in practical credit-scoring and risk-assessment systems. Overall, the final model provides a meaningful improvement over the baseline while maintaining stable and realistic performance suitable for operational deployment.

6.Why the Model Was Not Optimized for Maximum Performance

During evaluation, multiple decision thresholds were tested to understand how different optimization goals would affect business outcomes. Increasing recall allowed the model to identify more high-income individuals, but it also produced a large number of false positives. In practice, this would cause the company to spend marketing resources on many customers unlikely to convert, increasing operational cost and potentially introducing demographic bias due to over-targeting certain groups.

On the other hand, optimizing strictly for precision made the model extremely selective. While the customers identified were highly likely to belong to the high-income category, many potential buyers were missed. This would reduce revenue opportunities and limit the effectiveness of marketing campaigns.

Because both extremes harm business performance in different ways, the final decision threshold was chosen to balance these trade-offs rather than maximize a single metric. The selected operating point produced an F1 score of approximately 0.56, with precision around 0.52 and recall around 0.59. This balance minimizes unnecessary marketing expenditure while still capturing a meaningful portion of high-value customers, providing stable and practical decision support rather than aggressive but unreliable targeting.

7.Business Justification

The selected operating point represents a deliberate balance between revenue opportunity, marketing cost, and fairness. Rather than aggressively targeting as many customers as possible or restricting outreach only to highly certain predictions, the model was tuned to support consistent and responsible decision making. This balanced performance reduces wasted marketing spend, preserves potential sales opportunities, and avoids systematically over-targeting particular groups. The objective was therefore not short-term profit maximization but stable, sustainable performance that maintains customer trust and supports long-term business value.

8.Fairness and Trustworthiness

Beyond predictive performance, the model was evaluated for how its errors were distributed across demographic groups. The intention was to ensure that the system did not disproportionately misclassify or over-target particular populations. Rather than optimizing purely for short-term conversion rates, the project prioritized reducing bias risk, maintaining long-term customer trust, and supporting regulatory compliance.

In real business environments, overly aggressive or uneven targeting can lead to legal exposure, reputational harm, and customer churn. A model that maximizes immediate gains but systematically disadvantages certain groups may damage the brand and undermine long-term profitability. For this reason, the operating threshold was intentionally moderated so that the system remains dependable and fair, providing sustainable decision support instead of opportunistic but risky targeting.

9.Segmentation Justification

While the classification model identifies *who is likely to have higher purchasing power*, it does not explain how the business should communicate with them. Treating all predicted high-income customers the same would still result in generic marketing campaigns and limit the value of the prediction system. Segmentation complements prediction by determining how different groups should be approached, allowing the company to tailor messaging, products, and channels to specific customer needs.

To achieve this, we applied clustering to uncover natural population groups within the dataset. We selected the K-Means algorithm because it produces stable and interpretable clusters that can easily be translated into marketing personas. The resulting segments enable the business to move from simple targeting to personalized strategy, where different groups receive different types of offers rather than a single uniform campaign.

10.Segment Insights and Business Meaning

The clustering analysis revealed three distinct customer groups, each requiring a different marketing approach. The first group represents primarily retired or economically inactive individuals with very low purchasing probability. For this segment, high-cost promotional campaigns are unlikely to produce meaningful returns, so low-cost awareness channels such as informational messaging or passive brand engagement are more appropriate.

The second group consists of active working professionals and represents the strongest revenue opportunity. These individuals demonstrate stable employment patterns and higher purchasing capability, making them suitable candidates for financial products such as credit cards, installment plans, and premium retail offerings. Focused campaigns toward this segment are expected to produce the highest conversion impact.

The third group reflects late-career or transitional workers who still possess purchasing power but may prioritize financial security. For these customers, offers related to savings, long-term value, warranties, or insurance-style products are more relevant than aggressive consumption-based marketing.

A key finding from this analysis is that segmentation contributes more to marketing return on investment than prediction alone. While the classifier identifies potential high-value customers, segmentation determines the appropriate strategy for engaging them, enabling targeted communication instead of uniform outreach.

11.Final Business Value

The business value of this solution emerges from moving beyond broad outreach toward informed decision-making. Without any analytical support, the company relies on mass marketing, contacting large portions of the population regardless of purchasing likelihood. Introducing the classification model improves this process by narrowing outreach to customers more likely to have higher purchasing capability, thereby improving targeting efficiency.

However, the greatest value comes from combining prediction with segmentation. Instead of simply identifying whom to contact, the company also learns how to approach each group differently. This enables strategic marketing where messaging, offers, and channels are aligned with customer characteristics rather than applying a single campaign to everyone.

As a result, the organization can reduce unnecessary marketing expenditure, increase conversion probability through relevance, and enhance overall customer experience by avoiding intrusive or inappropriate promotions.

12.Critical Insights Discovered

The analysis produced several important business insights beyond the predictive results. Employment stability emerged as a stronger indicator of purchasing capability than education level, suggesting that consistent income patterns matter more for marketing decisions than academic background alone. The data also revealed that a relatively small portion of the population contributes disproportionately to potential revenue, reinforcing the importance of focused targeting rather than broad outreach.

Additionally, the segmentation results demonstrated that personalized marketing strategies are significantly more effective than blanket campaigns, as different customer groups respond to different types of offers. Finally, incorporating fairness and responsible modeling practices supports

long-term profitability by preserving customer trust and reducing the risk of reputational or regulatory issues.

13. Quantitative Evidence & Validation

Economic Evaluation (Business Impact Simulation)

To quantify business value, the classification model was translated into a simulated marketing campaign. The objective was to measure operational efficiency rather than rely only on predictive metrics.

We assume a realistic retail scenario:

- Cost to contact a customer = **$1**
- Profit from a successful conversion = **$40**

From the test dataset:

- Total customers: **39,905**
- High-income customers: **2,476**

Using the model performance:

- Precision = **0.52**
- Recall = **0.59**

The confusion matrix counts are estimated as:

	Predicted Low	**Predicted High**
Actual Low	36,081	1,348
Actual High	1,015	1,461

Strategy Comparison

Strategy	**Contacts**	**Conversions**	**Cost**	**Revenue**	**Net Profit**
Mass Marketing	39,905	2,476	$39,905	$99,040	**$59,135**
Model Targeting	2,809	1,461	$2,809	$58,440	**$55,631**

Interpretation

The model reduces outreach by approximately **93%** while still capturing **59% of high-value customers**. Although total profit is slightly lower than contacting everyone, the profit generated per contact increases substantially. This makes campaigns scalable and operationally efficient, since the business spends far fewer resources for nearly the same economic return.

Therefore, the value of the model lies in improving marketing efficiency and resource allocation rather than maximizing raw revenue.This demonstrates that the model's value comes primarily from **cost reduction efficiency**, not perfect prediction.

14. Threshold Optimization Justification

The final classification threshold (0.85) was not chosen arbitrarily.
Different thresholds represent different business strategies:

We evaluated model performance across multiple thresholds and analyzed precision, recall, and expected profit simultaneously.

Threshold Type	Behavior
Low threshold	Aggressive marketing (high cost)
High threshold	Conservative targeting
Balanced threshold	Sustainable marketing

The selected threshold of **0.85** produced:

- Precision: **0.52**
- Recall: **0.59**
- F1 Score: **0.56**

For comparison, a profit-optimized threshold of **0.33** produced:

- Precision: **0.22**
- Recall: **0.93**
- Accuracy: **0.79**

This alternative dramatically increases false positives, causing large marketing cost and fairness concerns.Therefore, the chosen threshold prioritizes stable operational performance over aggressive short-term gains.
The selected threshold corresponds to the point where expected profit stabilizes and marginal gains from additional targeting are offset by additional marketing cost.
This threshold produces balanced precision and recall, ensuring the company does not overspend while still capturing meaningful opportunities.

Therefore, the chosen operating point reflects an economic optimum rather than a statistical optimum.

15. Cluster Validation

To verify that segmentation produced real groups rather than arbitrary partitions, we evaluated cluster structure using statistical metrics.

The clustering algorithm identified three groups with distinct characteristics:

Cluster	Population Size	Economic Activity	Marketing Role
Inactive / Retired	Large	Very low income probability	Awareness only
Working Professionals	Largest active segment	Highest purchase potential	Primary targeting
Late Career	Small but valuable	Financial planning oriented	Specialized offers

Clustering produced the following silhouette scores:

K	Silhouette Score
2	0.2257
3	0.2320
4	0.2335
5	0.1757
6	0.1839

We selected **K = 3** because it provides interpretable groups while maintaining meaningful separation.

Cluster Profiles

Cluster	Size	Avg Age	Weeks Worked	High Income Rate
Inactive / Retired	46,244	56.1	1.42	1.5%
Working Professionals	95,306	38.8	46.7	12.0%
Late-Career	1,981	59.9	18.1	11.7%

These differences confirm clusters represent **behavioral economic groups**, not random partitions.

The silhouette score confirmed moderate but meaningful separation between groups, indicating overlapping but distinguishable populations — expected in human demographic data.

Rather than forcing artificial separation, the segmentation captures realistic behavioral differences that marketing teams can operationalize.

16.Fairness Evaluation(Metric-Based Evidence)

To ensure the system supports ethical decision making, we analyzed error distribution across demographic groups.

False positive rates across gender groups:

Group	False Positive Rate
Female	0.132
Male	0.283

The model shows higher targeting toward males due to employment distribution differences in the dataset.
This was considered during threshold selection to avoid overly aggressive targeting policies.
Fairness was therefore treated as a constraint when choosing the operating point.
Instead of focusing solely on overall performance, we evaluated whether the model systematically favored or disadvantaged any group.
Balanced error distribution reduces regulatory risk and preserves customer trust.

The final threshold was intentionally selected to avoid extreme recall or precision scenarios that disproportionately affect specific populations.

This decision sacrifices small short-term gains in exchange for long-term reliability and compliance readiness.

Thus, fairness is not treated as a constraint after modeling but as part of model design.

17. Model Explainability

Top predictive features:

Feature	Importance
Age	1577
Number of employers	523
Weeks worked	479
Sex	192
Education (Bachelor/Master)	~165

Economic Interpretation

The model primarily relies on employment stability indicators rather than sensitive attributes:

- Stable work history → consistent income
- More weeks worked → higher purchasing capability
- Education → earning potential

This confirms the model behaves similarly to real financial risk-assessment systems.

18.Future Improvements

While the current solution provides meaningful decision support, several enhancements could further improve its effectiveness. Incorporating behavioral purchase data, such as transaction history or browsing patterns, would allow the models to move beyond demographic estimation and better capture actual buying intent. Periodic retraining of the models would also help maintain performance as customer behavior and economic conditions evolve over time.

In addition, integrating a campaign response feedback loop would enable the system to learn directly from marketing outcomes, gradually refining targeting strategies based on real engagement data. Finally, applying uplift modeling techniques could identify customers whose behavior is likely to change because of a campaign, allowing the company to focus efforts on persuadable individuals rather than those who would purchase regardless or never purchase at all.

19.Final Conclusion

This project demonstrates the application of responsible artificial intelligence within a business decision framework. Rather than pursuing the highest possible model score, the solution focuses on improving the quality of marketing decisions. The models are designed to support practical operations while considering both economic efficiency and customer impact.

By combining income prediction with customer segmentation, the system balances profitability with fairness and long-term trust. The resulting approach is operationally practical, interpretable for stakeholders, and aligned with sustainable business strategy rather than short-term optimization.