Statistics 208 Project Proposal

Project Team

Team Leader: Nathanial Zhu

Team Members: Xujuan Liang, Xiaoya Wei, Youyi Fu

Topics

Topic Title: Customer Segmentation Analysis

Topic Motivation: Customer segmentation is a powerful analytical tool that enables businesses to enhance marketing efficiency, improve customer experiences, and drive product innovation. By dividing customers into distinct groups based on their behaviors and preferences, companies can create targeted marketing campaigns that yield higher conversion rates and provide a better return on investment. This segmentation also facilitates tailored product offerings and personalized service, which significantly boosts customer satisfaction and loyalty. Additionally, understanding different customer segments assists businesses in making strategic decisions, from resource allocation to product development, and enables predictive analytics for forecasting purchasing behaviors and potential churn. Ultimately, customer segmentation not only helps in achieving personalization at scale but also provides a competitive edge in understanding geographic and demographic influences on customer behavior. This comprehensive approach not only maximizes business efficiency but also enhances overall customer engagement and retention.

Research

In a research context, leveraging Kaggle's customer segmentation datasets provides opportunities to advance methodologies and practical applications in business analytics and marketing. Research could focus on developing and comparing innovative statistical and machine learning techniques to enable more effective customer segmentation. Researchers can also drill down into consumer behavioral insights to assess how different market segments respond to marketing strategies and product features. Additionally, assessing the impact of segmentation on key business outcomes such as sales, retention, and profitability provides empirical evidence of its effectiveness. It will also be critical to explore ethical considerations of data use, the impact of cultural and socioeconomic factors on segmentation effectiveness, and the potential of emerging technologies such as artificial intelligence and machine learning to refine data collection and personalization strategies. Such research not only enriches the academic literature but also provides actionable insights to businesses and policymakers, promoting a deeper understanding of effective customer engagement strategies in different market contexts.

Objective Proposal: Give 3, can help figure new ones if needed.

1. **Predictive Customer Profiling**: This involves creating predictive models to understand customer behaviors and predict outcomes such as spending scores. It directly applies machine learning, a core skill in business analytics, and has a direct impact on marketing and strategic business decisions.
2. **Modeling Customer Lifetime Value (CLV)**: Building models to estimate the lifetime value of customers is crucial for making informed business decisions and optimizing marketing strategies. It's highly relevant for demonstrating the application of predictive analytics in financial forecasting and customer relationship management.
3. **Feature Importance Analysis**: Identifying key features that influence customer segmentation and spending can help businesses focus their efforts on the most impactful data points. This objective not only hones skills in data analysis but also contributes to more effective and targeted business strategies.
4. **Imputation Model Development**: While crucial, developing models for imputing missing data is slightly more technical and less directly tied to business outcomes compared to the objectives listed above. However, it's essential for ensuring the quality and completeness of data used in further analyses.
5. **Impact Analysis of Imputation on Segmentation**: Analyzing how different imputation techniques affect segmentation accuracy is important for data integrity and analytical accuracy. It is ranked lower mainly because it is somewhat more specific and serves as a supportive analysis to broader objectives like predictive profiling and CLV modeling.

Data

We propose utilizing the "Customer Segmentation Dataset" available on Kaggle (https://www.kaggle.com/datasets/abisheksudarshan/customer-segmentation) for our analysis. This dataset contains information on customer demographics, spending scores, and annual income, providing valuable insights into consumer behavior. It consists of 200 records with five features, including CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1-100).

Data Gaps

Here are some potential data gaps in our dataset, and we will aim to solve the data gaps one by one once we start the project and in our introduction video.

1. **Missing Values**: As evident in the dataset, several columns like 'Ever_Married', 'Graduated', 'Profession', 'Work_Experience', and 'Family_Size' have missing data. The absence of this information can decrease accurate analysis and predictive modeling efforts. We will determine a way to solve the missing values.
2. **Categorical Data Encoding**: Many of the variables (e.g., 'Gender', 'Ever_Married', 'Graduated', 'Profession', 'Spending_Score', 'Var_1') are categorical. Depending on the modeling techniques

used, you might need to convert these into numerical formats through encoding. This conversion must be done carefully to preserve the meaning and relationships within the data.

3. **Data Balance**: If the dataset or certain important categories within it (like customer segments in 'Segmentation') are imbalanced, this could bias the model outcomes. It's important to check if all categories are adequately represented.
4. **Outliers and Inconsistencies**: We might observe outliers and inconsistencies in the dataset. Any outliers or inconsistent entries in the data can distort predictive modeling and analysis. For instance, unusually high or low values in 'Age' or 'Work_Experience' might need investigation and potential treatment.
5. **Depth of Information**: The dataset might lack certain variables that could be critical for more comprehensive analysis. For example, detailed income levels, educational background specifics, or more granular geographic information could provide more insights but may not be included.