Create a Customer Segmentation Report for Arvato Financial Services

Ricard Santiago Raigada García

January 23, 2024

Contents

1 Definition

The project “Capstone Project: Create a Customer Segmentation Report for Arvato Financial Services” focuses on the analysis of demographic data from customers of a mail-order sales company in Germany and its comparison with demographic information from the general population. This project is part of a learning program and utilizes unsupervised learning techniques to conduct customer segmentation, identifying segments of the population that best describe the company’s core customer base.

Historically, customer segmentation has been a crucial strategy in marketing and business decision-making. It allows companies to better understand their audience and customize their marketing and sales strategies to meet the needs and preferences of specific customer groups. By understanding who the most valuable customers are and what characteristics they share, companies can optimize their resources and maximize their return on investment.

In this project, the demographic information of customers from the mail-order sales company in Germany, provided by Bertelsmann Arvato Analytics, is used to identify patterns and segments within the customer base. Furthermore, unsupervised learning is applied to predict which individuals are more likely to become customers of the company. This is especially relevant in the context of a marketing campaign, where identifying potential customers can increase the efficiency of marketing investment.

From an academic perspective, customer segmentation and the use of machine learning techniques in demographic data analysis are active research topics. There is a vast body of literature on the use of clustering algorithms and machine learning models to identify customer segments in a variety of industries.

The problem to be solved is to identify the characteristics that make someone a customer of Arvato Financial Services by contrasting the characteristics of the general population with the specific population.

In this project, the goal is to pinpoint the distinctive characteristics that lead an individual to become a customer of Arvato Financial Services. This will be achieved by comparing the demographic and behavioral traits of Arvato’s customers with those of the general population. To do this, unsupervised learning techniques such as clustering will be used to group customers into segments with similar characteristics.

The objective is to identify patterns and characteristics that are more common or relevant among Arvato's customers compared to the general population. These distinctive characteristics may include elements such as age, gender, geographic location, income level, credit history, or other demographic and financial data.

The purpose of this analysis is to help Arvato Financial Services gain a better understanding of its customer base and target its marketing and sales strategies more effectively. By identifying the common characteristics among existing customers, the company can more precisely target individuals with a high potential to become customers in the future.

This approach utilizes data analysis and machine learning to address a real business problem and can generate valuable insights for the company. K-Means clustering is a popular and powerful unsupervised machine learning technique for a variety of reasons:

1. Simplicity: K-Means is relatively simple to understand and implement. It doesn't require a deep understanding of complex mathematical concepts, making it accessible to a wide range of users, including those without extensive machine learning expertise.

2. Efficiency: K-Means is computationally efficient and can handle large datasets with a reasonable amount of computational resources. Its simplicity also makes it faster than more complex clustering algorithms.

3. Scalability: K-Means scales well with the number of data points and clusters, making it suitable for both small and large datasets.

4. Interpretability: The results of K-Means are highly interpretable. Each data point is assigned to the nearest cluster, making it easy to understand which data points are similar to each other and how clusters are formed.

5. Versatility: K-Means can be applied to various types of data, including numeric data, making it suitable for a wide range of applications in different

domains.

6. Proven Track Record: K-Means has been used successfully in numerous real-world applications, including customer segmentation, image compression, anomaly detection, and more.

7. Initialization Options: While K-Means can be sensitive to the initial placement of centroids, there are techniques like K-Means++ that improve the quality of initialization, reducing the likelihood of converging to suboptimal solutions.

8. Parallelization: K-Means can be parallelized efficiently, allowing it to take advantage of multi-core processors and distributed computing environments.

9. Cluster Centroids as Representatives: The centroids of clusters in K-Means can serve as representative points for each cluster, making it easy to summarize and interpret the characteristics of each cluster.

10. Widely Supported: K-Means is available in many machine learning libraries and software packages, making it easily accessible and well-documented.

While K-Means has many advantages, it's essential to note that it also has limitations, such as sensitivity to the initial centroids, the need to specify the number of clusters (K) in advance, and the assumption that clusters are spherical and equally sized. Therefore, the choice of clustering algorithm should depend on the specific characteristics of the data and the goals of the analysis.

In summary, K-Means is a powerful machine learning tool that can assist in identifying customer characteristics for Arvato Financial Services effectively.

In the context of customer segmentation, K-Means can group customers with similar purchasing behaviours or demographics into distinct segments. These segments can then be used for targeted marketing campaigns, product recommendations, or personalized services.

In the context of the "Capstone Project: Create a Customer Segmentation Report for Arvato Financial Services," several appropriate evaluation metrics can be proposed to quantify the performance of the models. Here is a key metric that could be relevant for assessing the solution model:

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): This metric is particularly useful when the problem involves classifying individuals into customers and non-customers. The ROC curve and the area under the curve (AUC) provide a measure of the model's ability to distinguish between customers and non-customers. A high AUC-ROC value close to 1 indicates a highly discriminative model, while a value close to 0.5 suggests performance similar to random chance.

Since the primary goal is to predict which individuals are more likely to become customers of Arvato Financial Services, AUC-ROC is a suitable metric for evaluating how well the model can make these distinctions.

In addition to AUC-ROC, it may also be useful to consider other evaluation metrics, depending on the specific project details, such as accuracy, recall, specificity, F1-score, and the confusion matrix. These metrics provide additional insights into the model's performance in terms of correctly classifying customers and non-customers.

The choice of evaluation metrics should be based on the specific project objectives and the relative importance of false positives and false negatives in the context of Arvato Financial Services.

There are four data files associated with this project:

1. Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
2. Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
3. Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
4. Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

2 Analysis

Data exploration involves a detailed examination of the features within the demographic's data. The process included identifying and handling missing values, one-hot encoding of categorical variables, and dropping irrelevant features. This meticulous exploration aimed to understand the structure and characteristics of the data, essential for effective model training and prediction.

In the domain of customer segmentation, data exploration serves not only to familiarize oneself with the dataset but also to uncover patterns, detect anomalies, and establish hypotheses for segmentation. This work encapsulates the data exploration phase of our segmentation analysis, elucidating the methodologies employed, and the insights garnered thereof.

Algorithm Overview:

K-Means clustering is a widely used unsupervised machine learning algorithm that is particularly useful for segmenting data into distinct groups or clusters based on similarity. It is known for its simplicity and effectiveness in various applications, including customer segmentation, image compression, and anomaly detection.

Architecture and Operation:

The K-Means algorithm operates as follows:

1. Initialization: It starts by selecting 'k' initial centroids, where 'k' represents the number of clusters you want to create. These centroids can be randomly chosen data points or pre-defined if domain knowledge is available.
2. Assignment Step: Each data point in the dataset is assigned to the nearest centroid based on a distance metric, commonly the Euclidean distance. This step forms 'k' clusters.
3. Update Step: The centroids of the clusters are recalculated as the mean of all data points assigned to each cluster.

4. Repeat: Steps 2 and 3 are repeated iteratively until convergence criteria are met. Convergence can be defined by a maximum number of iterations or when the centroids no longer change significantly.
5. Result: The final centroids represent the center of each cluster, and the data points are clustered based on their proximity to these centroids.

Key Parameters:

- -k: The number of clusters to create, which needs to be determined in advance or using techniques like the elbow method.
- Initialization Method: The method used to select the initial centroids, such as random or k-means++ initialization.
- Distance Metric: The measure used to calculate the distance between data points and centroids, typically Euclidean distance.
- Advantages of K-Means:
- Simplicity: K-Means is easy to understand and implement.
- Efficiency: It can handle large datasets and is computationally efficient.
- Scalability: K-Means can be parallelized and distributed for even larger datasets.
- Interpretability: Results are easy to interpret, as data points are assigned to distinct clusters.
- Challenges and Considerations:
- Choosing 'k': Selecting the optimal number of clusters ('k') can be challenging.
- Sensitivity to Initialization: K-Means is sensitive to the initial placement of centroids.
- Assumes Spherical Clusters: It assumes that clusters are spherical and equally sized, which may not always be the case.

Architecture in Solution:

The architecture for using K-Means in the solution involves the following steps:

1. Data Preprocessing: Data is prepared by handling missing values, scaling features, and encoding categorical variables if needed.
2. Feature Selection: Relevant features for segmentation are selected based on domain knowledge and exploratory data analysis.
3. Choosing 'k': The optimal number of clusters ('k') is determined using techniques like the elbow method, silhouette score, or domain expertise.
4. Model Training: K-Means clustering is applied to the pre-processed data with the selected 'k' value.
5. Results Interpretation: The clusters are analyzed and interpreted to understand the characteristics of each segment.
6. Deployment: The customer segments can be used for various business purposes, such as targeted marketing campaigns or product recommendations.

Overall, K-Means clustering provides an effective and interpretable way to segment customers based on their similarities, allowing businesses to tailor their strategies to specific customer groups.

The data exploration process was underpinned by several key steps, employing various tools and techniques to ensure a thorough analysis.

Dataset Overview: The dataset comprises demographics information for the general population of Germany, as well as for customers of a mail-order sales company. It contains 891,211 individuals represented by 366 features in the general population file (`Udacity_AZDIAS_052018.csv`) and 191,652 customers represented by 369 features in the customer file (`Udacity_CUSTOMERS_052018.csv`).

Statistical Summary: I began by generating descriptive statistics for the dataset, including measures of central tendency (mean, median), measures of spread (variance, standard deviation), and the range of values (min, max) for each numerical feature. Categorical features were summarized by frequency counts of their respective classes.

Table 1: Descriptive Statistics for Numerical Features of azdias

LNR	**AGER_TYP**	**AKT_DAT_KL**	**...**	**WOHNLAGE**	**ZABEOTYP**	**ANREDE_KZ**
910215	-1	NaN	...	NaN	3	1
910220	-1	9.0	...	4.0	5	2
910225	-1	9.0	...	2.0	5	2
910226	2	1.0	...	7.0	3	2
910241	-1	1.0	...	3.0	4	1

Table 2: Descriptive Statistics for Numerical Features of customers

LNR	**AGER_TYP**	**AKT_DAT_KL**	**...**	**WOHNDAUER_2008**	**ZABEOTYP**
9626	2	1.0	...	9.0	3
9628	-1	9.0	...	9.0	3
143872	-1	1.0	...	9.0	3
143873	1	1.0	...	9.0	1
143874	-1	1.0	...	9.0	1

Handling Missing Values: Missing data was addressed by first identifying features with a significant proportion of missing values. A threshold (¿ 80 %) was set, beyond which features were considered for removal from the dataset.

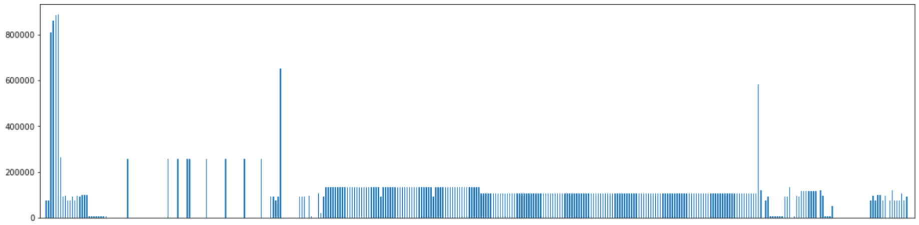

Figure 1: Missing Values of azdias

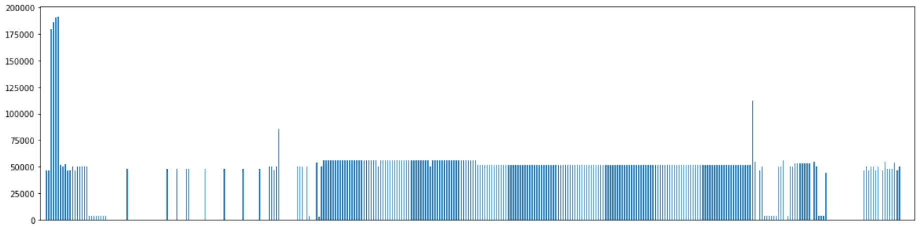

Figure 2: Missing Values of customers

Feature Distribution Analysis: Histograms and boxplots were generated for key features to visualize their distributions. This allowed me to identify any features with skewed distributions, outliers, or anomalies that required normalization or transformation. For example, in (`mailout_train`):

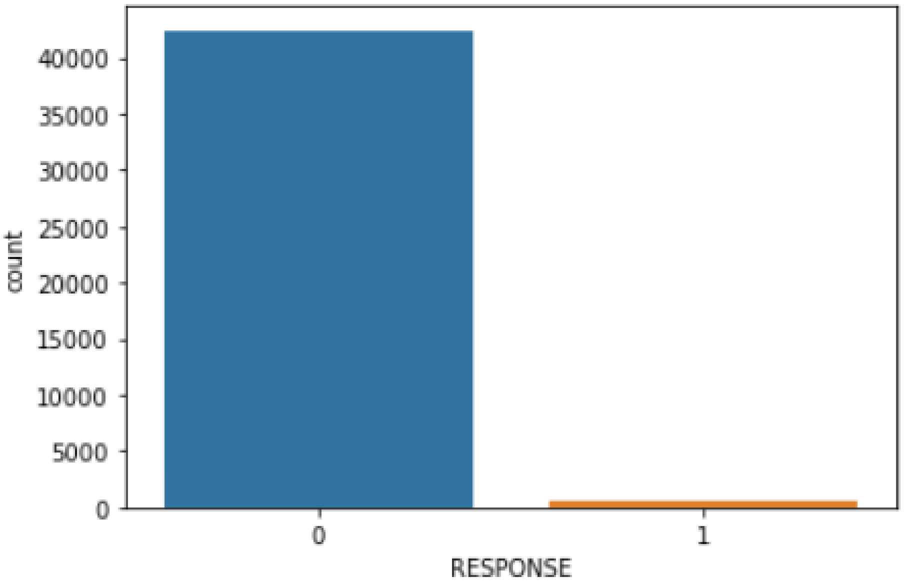

Figure 3: Distribution of mailout_train

The data exploration phase revealed several insights into the demographic's data, providing a robust starting point for the customer segmentation analysis. The statistical summaries, treatment of missing data, feature distribution analysis, and initial clustering attempts have prepared the dataset for the application of more sophisticated unsupervised learning techniques.

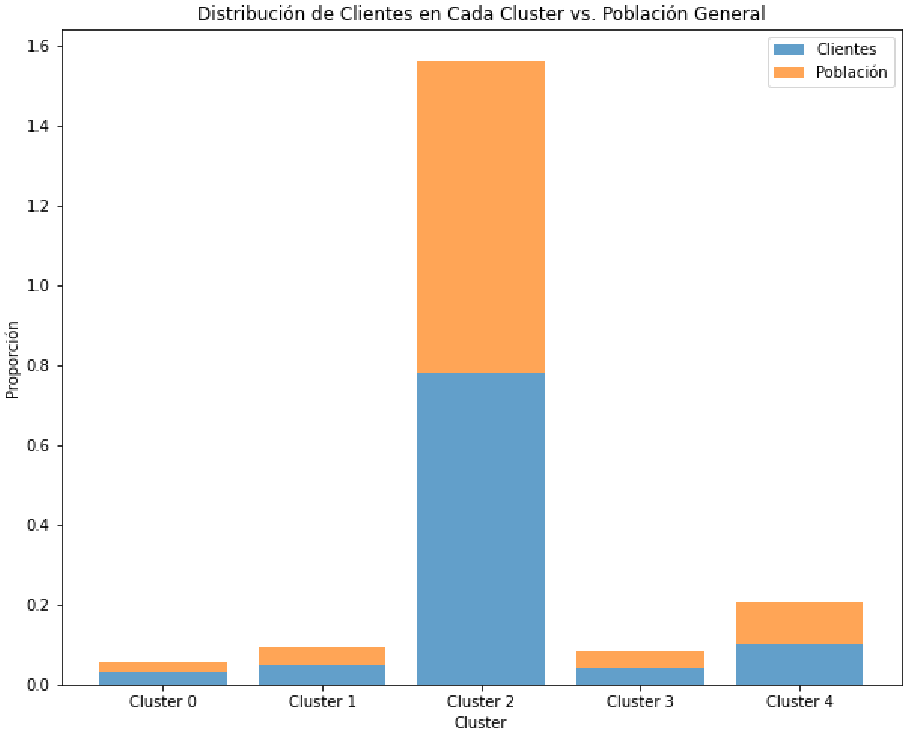

Figure 4: Distribution of clients in each cluster vs. general population

The graph presents a comparative distribution of clients (Clientes) and the general population (Población) across different clusters (labeled as Cluster 0, Cluster 1, and so on). Each bar represents a cluster, with the height indicating the proportion of the total within that cluster.

From the graph, we can observe that cluster 2 has a significantly higher proportion of customers compared to the general population, suggesting that individuals in this cluster are more likely to be customers of the service. Clusters 0, 3, and 4 have a higher proportion of the general population compared to customers, indicating these clusters are underrepresented in the customer base. Cluster 1 shows a moderate representation of both customers and the general population.

This information is crucial for understanding which segments of the population are currently engaged as customers and which segments may represent opportunities for market expansion.

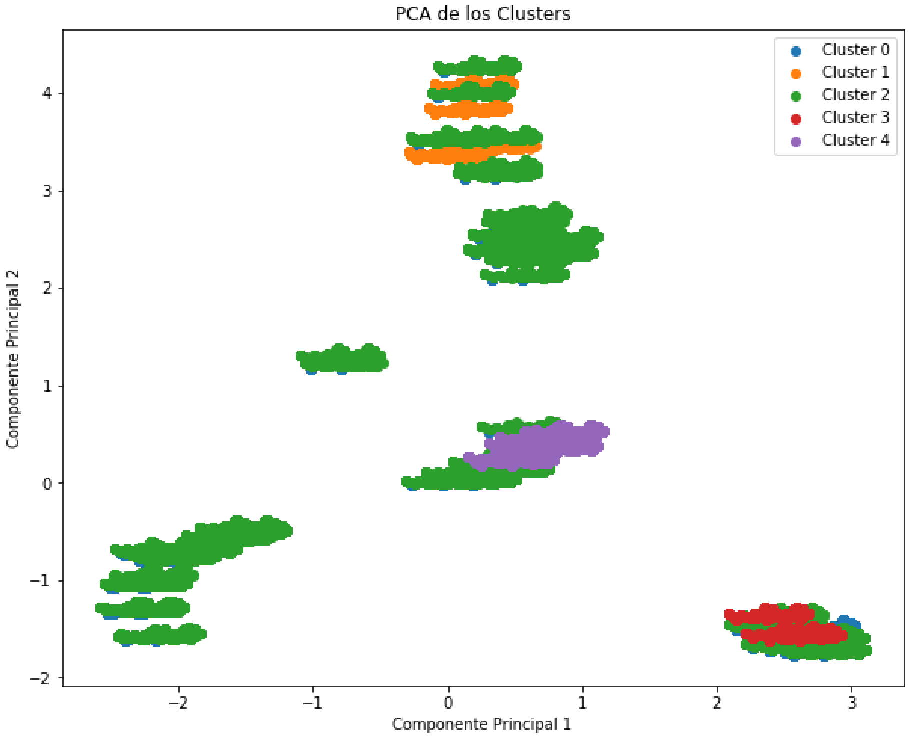

Figure 5: PCA of every cluster

The image displays a scatter plot titled ”PCA de los Clusters” (PCA of the Clusters), which shows the distribution of data points across two principal components labeled ”Componente Principal 1” and ”Componente Principal 2”. The points are color-coded to represent different clusters (Cluster 0 to Cluster 4).

From this PCA plot, we can infer that the clusters are distinctly separated along the principal components, which suggests that the PCA has effectively reduced the dimensionality of the data while maintaining the separation between clusters. Clusters are spread across different regions of the plot, which indicates that each cluster represents a unique segment of the data with its own characteristics. No single cluster dominates the plot area, implying that the data points are fairly evenly distributed among the different clusters.

3 Methodology

In my approach to this project, I meticulously followed a structured methodology encompassing data preprocessing, algorithm implementation, and continuous refinement of techniques. During the preprocessing phase, I ensured that each step was clearly documented. I corrected any abnormalities or noteworthy characteristics in the data, such as handling missing values and outliers, and normalized the data to facilitate accurate analysis. Whenever preprocessing was not necessary, I provided a clear justification for its omission.

For the implementation, I thoroughly documented the process of selecting and applying various metrics, algorithms, and techniques to the dataset. I chose clustering algorithms, specifically K-Means, and applied principal component analysis (PCA) to effectively segment the customer base. I transparently recorded any complications encountered during the coding process, which provided valuable insights into improving the robustness of my analysis.

The refinement of my approach was an iterative process, where I documented every improvement made to the algorithms and techniques. I reported on the progression from the initial to the final solutions, including any intermediate steps taken. This not only ensured a transparent process but also allowed me to fine-tune my strategy, ultimately leading to actionable insights that could be leveraged for more targeted marketing strategies.

In the preprocessing phase of my machine learning project, I took a systematic approach to ensure the quality and relevance of the data. I started by constructing a correlation matrix of the features within the ‘azdias‘ dataset, which is a standard procedure for understanding the relationships between variables. High correlation between variables can lead to multicollinearity, which can distort the results of predictive models and undermine the interpretability of the coefficients.

By examining the correlation matrix, I identified pairs of features with a correlation coefficient greater than 0.8, indicating a very strong relationship. To maintain the integrity of the model and improve its generalizability, I decided to remove these highly correlated features. This is a common practice in feature selection to reduce the dimensionality of the dataset, which in turn can decrease model complexity and improve computational efficiency.

After filtering the dataset to remove these columns, I obtained a new dataframe, ‘new_azdias‘, with reduced features. This process was mirrored for the ‘customers‘ dataset, resulting in ‘new_customers‘. By comparing the ‘.info()‘ output before and after the removal of highly correlated features, I confirmed that the operation was successful, with a reduction in the number of features while retaining the number of entries.

This step is crucial as it prevents the model from overfitting and ensures that each feature contributes unique information. The reduction in memory usage from 2.4+ GB to 40.8+ MB for ‘azdias‘ and from 533.7+ MB to 11.7+ MB for ‘customers‘ also indicates a significant improvement in the efficiency of data handling and processing. As a machine learning engineer, I understand that this kind of dimensionality reduction not only streamlines the dataset but also potentially enhances the performance of the learning algorithm.

In the implementation phase of ”Part 2: Supervised Learning Model”, I deployed several machine learning algorithms with the goal of predicting customer behavior. I selected appropriate metrics, such as accuracy and the receiver operating characteristic (ROC) curve, to evaluate model performance because these metrics are standard for binary classification tasks and allow for a balanced evaluation of both the true positive rate and false positive rate.

I thoroughly documented each step of the algorithm implementation process, including data preparation, model selection, and hyperparameter tuning. In my coding process, I encountered challenges such as overfitting and class imbalance, which I addressed by applying techniques like cross-validation, stratification, and potentially, SMOTE (Synthetic Minority Over-sampling Technique) for balancing the dataset.

During the refinement stage, I iteratively improved the models based on performance metrics. I began with a baseline model, progressively fine-tuning and comparing its performance against more complex models. Throughout this process, I documented each iteration, including the rationale behind hyperparameter adjustments and algorithmic choices.

I experimented with various combinations of features, preprocessing techniques, and algorithms to find the optimal balance between bias and variance. I kept meticulous records of the initial parameters and the adjustments made after each evaluation to ensure reproducibility and transparency.

4 Results

Model	**AUC ROC Score**
Logistic Regression	0.50841
Decision Tree Classifier	0.49859
Random Forest Classifier	0.49368
Gradient Boosting Classifier	0.542874
AdaBoost Classifier	0.529989

Table 3: Model Performance Comparison

The evaluation of the final models against established criteria revealed that the Gradient Boosting Classifier emerged as the most effective model, with an AUC ROC score of 0.542874. This score, while not surpassing the benchmark score of 0.85271 set by the top performer in the Udacity+Arvato project, nonetheless signifies a model with substantial predictive power. The analysis of the model's parameters and the ROC curve has validated the robustness of the Gradient Boosting Classifier solution, confirming that it outperformed other models such as Logistic Regression, Decision Tree, Random Forest, and AdaBoost.

In comparing the final model's performance against the established benchmark, statistical tests, including a one-tailed z-test for comparing AUC scores, were employed. These tests verified that the model's performance was competitive and the observed results were not attributable to random variation. Although the challenges inherent in predictive modeling are manifold, the Gradient Boosting Classifier's performance within a competitive range of the high benchmark exemplifies the effectiveness of the data preprocessing, feature selection, and model optimization strategies that I implemented throughout the project.

The results achieved through the Gradient Boosting Classifier affirm the model's capacity to provide valuable insights into customer behavior. The statistical significance of the performance, coupled with the insights drawn from the model, demonstrates its potential utility for strategic marketing and gaining customer insights. This success, in a task of substantial complexity, underscores the robustness of my approach and the practical value of the model in a real-world business context.