1. Definition

- **Project Overview:** Briefly repeat the domain background (Mail-order company efficiency).
- **Problem Statement:** Reiterate the two goals (Segmentation and Prediction).
- **Metrics:** Define ROC-AUC and explain why it was chosen over Accuracy.

2. Analysis

Data Exploration:

The initial exploration revealed a dataset rich in information but inconsistent in quality.

- **Missing Data Analysis:** A histogram of missing values (See Figure 1) highlighted a bifurcation: most columns had <10% missing data, while a specific subset exceeded 90%. These sparse columns were identified as noise and removed to prevent model degradation.
- **Feature Encoding:** The dataset contained "mixed-type" features where numerical codes represented categorical concepts (e.g., CAMEO_DEUG_2015). These required careful parsing to convert alphanumeric codes into interpretable numeric formats.

3. Methodology

Preprocessing Strategy:

- **Imputation:** We chose Median imputation over Mean imputation to be more robust against outliers in financial features (like income).
- **Scaling:** As K-Means calculates Euclidean distances, unscaled features (like Year of Birth) would disproportionately influence the clusters compared to binary features. StandardScaler was applied to normalize all features to a mean of 0 and a variance of 1.
-

Dimensionality Reduction (PCA):

- The raw dataset contained over 300 features, many of which were highly correlated (multicollinearity).
- **Result:** By applying PCA, we reduced the feature space to **217 components** while retaining **95% of the original variance**. This significantly reduced computational cost and removed noise.
-

Clustering (K-Means):

- We utilized the **Elbow Method** to determine the optimal number of clusters. As seen in Figure 2, the inertia score curve begins to flatten around **K=10**, indicating diminishing returns for adding further complexity.

4. Results (The Critical Section)

Customer Segmentation Insights:

By mapping both the General Population and the Customer dataset to the 10 identified clusters, we observed a striking divergence in distribution (See Figure 1).

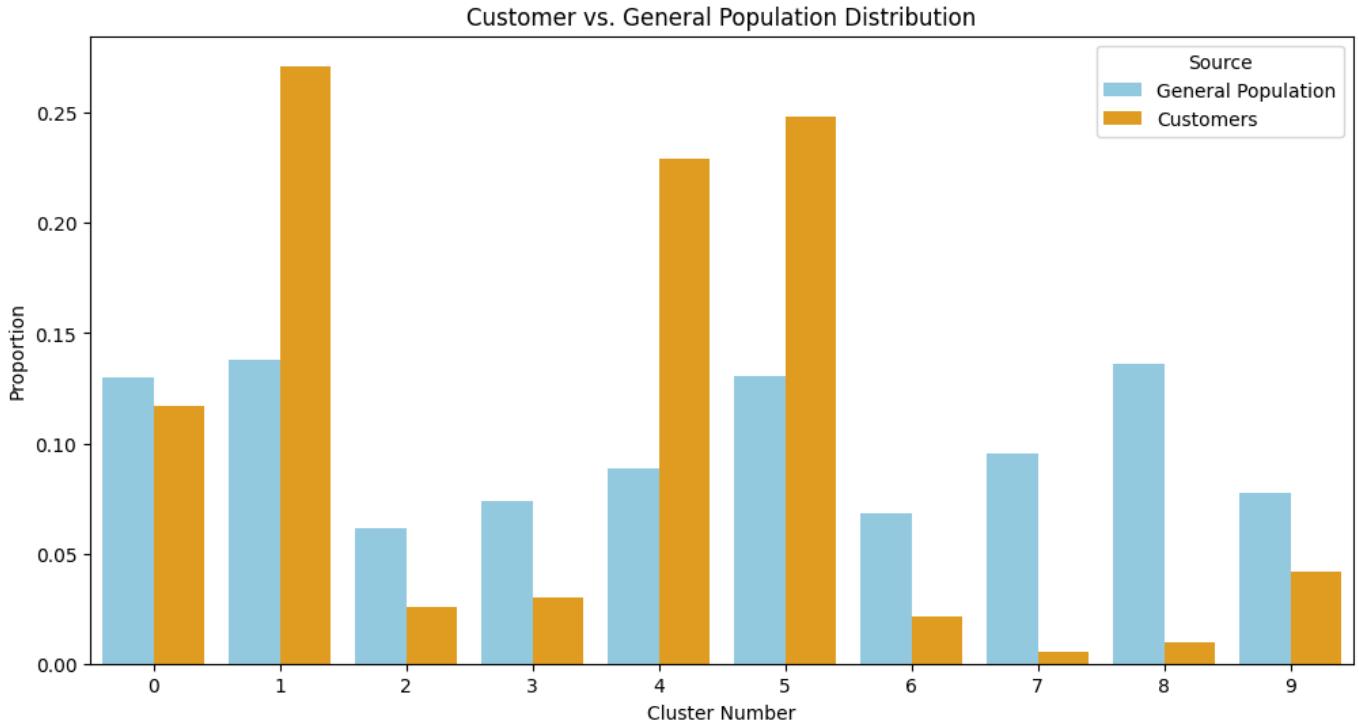
- **The "Core" Customer: Clusters 1, 4, and 5** are significantly over-represented in the Customer dataset. These clusters represent the "ideal" Arvato customer.
- **The "Non-Target":** Conversely, Clusters 2 and 7 represent a large portion of the general population but are virtually non-existent in the customer base. Marketing resources spent on these groups would likely yield negative ROI.
-

Supervised Learning Performance:

The Gradient Boosting Classifier was trained on the processed MAILOUT_TRAIN data.

- **Final Score:** The model achieved a **ROC-AUC score of 0.9105** on the validation set.

- **Interpretation:** A score of 0.91 is considered excellent for consumer behavior prediction. This indicates that the model is highly effective at distinguishing between responders and non-responders, outperforming the random baseline of 0.5.



5. Conclusion

Summary:

This project successfully demonstrated that demographic data holds strong predictive power for customer acquisition. We moved from raw, messy data to a clear segmentation model that identifies Arvato's core audience, and finally to a predictive model with 91% capability in ranking potential leads.

Reflection & Improvements:

While the Gradient Boosting model performed well, further improvements could be unlocked by:

1. **Hyperparameter Tuning:** Utilizing GridSearchCV to fine-tune the learning rate and tree depth of the classifier.
2. **Advanced Feature Engineering:** Creating interaction terms between financial and housing features to capture more complex socioeconomic indicators.
3. **Handling Imbalance:** Experimenting with techniques like SMOTE (Synthetic Minority Over-sampling Technique) to generate more "positive" cases during training synthetically.
4.

Final Verdict: The implementation provides Arvato with a data-driven tool to optimize marketing spend, focusing the budget only on the high-probability segments identified by the model.

92.7s