Arvato Machine Learning AWS Project Proposal

1. Domain Background

Effective marketing campaigns require a deep understanding of the target audience, especially in the competitive mail-order retail industry. In Germany, mail-order companies face the challenge of distinguishing their core customer base from the general population to optimize campaign targeting and resource allocation. Leveraging machine learning, businesses can analyze large-scale demographic datasets to gain actionable insights and build predictive models to improve marketing outcomes.
This project is based on real-life data provided by **Bertelsmann Arvato Analytics**, a global leader in data-driven marketing solutions. It involves analyzing customer and general population demographics to uncover key traits of the company's customer base and applying these insights to predict responses to a mail-order campaign.

2. Problem Statement

The mail-order company struggles to identify which individuals within the general population are most likely to convert into customers. This inefficiency leads to wasted marketing resources and suboptimal campaign performance.

The specific challenges are:

1. Understanding how existing customers differ from the general population.
2. Predicting which campaign recipients are most likely to become customers based on demographic data.

This problem is measurable and replicable, with a solution directly impacting marketing efficiency and ROI.

3. Solution Statement

The solution involves two major steps:

1. **Customer Segmentation Report**:
    - Perform unsupervised learning on demographic data from existing customers and the general population to identify patterns and segments.
    - Identify demographic groups most aligned with the company's core customer base and those less likely to engage.
2. **Supervised Learning Model**:
    - Develop a predictive model using labeled campaign data (TRAIN subset) to determine which individuals are likely to respond to future campaigns.
    - Use the resulting model to generate predictions for the TEST dataset, where response labels are withheld for Kaggle evaluation.

This structured approach combines exploratory insights and predictive power to solve the problem effectively.

4. Datasets and Inputs

The project utilizes four datasets:

1. **General Population Data** (`AZDIAS`): 891,211 rows x 366 columns; demographic data for the general German population.
2. **Customer Data** (`CUSTOMERS`): 191,652 rows x 369 columns; demographic data for the company's existing customers, including three additional columns: `CUSTOMER_GROUP`, `ONLINE_PURCHASE`, and `PRODUCT_GROUP`.
3. **Campaign Training Data** (`MAILOUT_TRAIN`): 42,982 rows x 367 columns; demographic data for individuals targeted in a campaign, with a `RESPONSE` column indicating whether they became customers.
4. **Campaign Test Data** (`MAILOUT_TEST`): 42,833 rows x 366 columns; demographic data for campaign targets, with the `RESPONSE` column withheld for evaluation.

Each dataset provides detailed demographic attributes, including individual, household, building, and neighborhood-level information. The datasets require extensive preprocessing, including handling missing values, standardizing formats, and encoding categorical data.

5. Benchmark Model

For the predictive phase, a logistic regression model will serve as the benchmark. Logistic regression is widely used in binary classification due to its simplicity and interpretability. It will provide a baseline for evaluating the performance of more advanced models, such as gradient-boosted trees or neural networks.

6. Evaluation Metrics

- **Customer Segmentation**:
  - **Silhouette Score**: Measures the quality of clustering by evaluating cohesion and separation.
  - **Davies-Bouldin Index**: Quantifies the average similarity between clusters.
  - **Cluster Visualization**: Use PCA or t-SNE to visualize high-dimensional data in a 2D space.
- **Predictive Model**:
  - **Accuracy**: Overall correctness of predictions.
  - **Precision**: Correctly predicted positive responses.
  - **Recall**: Ability to identify all true positive responses.
  - **F1-Score**: Harmonic mean of precision and recall.
  - **ROC-AUC**: Measures the model's ability to distinguish between classes.

These metrics ensure both segmentation quality and predictive model performance are appropriately evaluated.

7. Project Design

Step 1: Customer Segmentation Report

1. **Data Cleaning and Preprocessing**:
   - Handle missing data using imputation strategies.
   - Normalize numerical features and encode categorical variables.
2. **Exploratory Data Analysis**:
   - Investigate distributions and correlations in the `AZDIAS` and `CUSTOMERS` datasets.
3. **Clustering Analysis**:
   - Apply k-means clustering to identify segments within the customer base and compare them to the general population.
   - Validate clusters using silhouette scores and Davies-Bouldin Index.

4 **Insights and Reporting**:
    - Identify demographic groups most likely to engage with the company and those less likely to do so.

Step 2: Supervised Learning Model

1 **Preprocessing and Feature Engineering**:
    - Clean and preprocess the `MAILOUT_TRAIN` and `MAILOUT_TEST` datasets.
    - Incorporate insights from segmentation to create new features.
2 **Model Development**:
    - Train a baseline logistic regression model on `MAILOUT_TRAIN`.
    - Experiment with advanced models like random forests, XGBoost, and neural networks.
3 **Evaluation and Tuning**:
    - Use cross-validation and hyperparameter tuning to optimize model performance.
    - Compare models using evaluation metrics (precision, recall, F1-score, etc.).
4 **Prediction**:
    - Apply the best model to the `MAILOUT_TEST` dataset to generate predictions for Kaggle competition evaluation.

8. Presentation

The final deliverables will include:

- **Customer Segmentation Report**: Detailed insights into the customer base and actionable demographic profiles.
- **Predictive Model Results**: Predictions for the TEST dataset, presented alongside evaluation metrics and feature importance analysis.

This project combines exploratory analysis and predictive modeling, demonstrating the value of machine learning in solving real-world marketing challenges.

Conclusion

This project will provide actionable insights through customer segmentation and predictive modeling, enabling the financial institution to optimize its marketing strategies. By leveraging unsupervised and supervised learning, the solution will improve campaign response rates while reducing costs, ensuring a substantial return on investment.