Capstone Project Proposal

Domain Background:

Arvato is a global services company headquartered in Gütersloh, Germany. Its services include customer support, information technology, logistics, and finance. The history of Arvato goes back to the printing and industry services division of Bertelsmann; the current name was introduced in 1999. Today, Arvato is one of eight divisions of Bertelsmann, the media, services and education group. In 2016, Arvato had about 68,463 employees and an overall turnover of 3.84 billion euros.

Problem Statement:

In this capstone project, I will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. I will then identify the parts of the population that best describe the core customer base of the company. Then, I will apply what I've learned on a third dataset with demographics information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company

Datasets and Inputs:

There are 3 data files that will be used in this project:

- **`Udacity_AZDIAS_052018.csv`**: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0	11.0
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0	10.0
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	3.0

5 rows × 366 columns

- **`Udacity_CUSTOMERS_052018.csv`**: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	9626	2	1.0	10.0	NaN	NaN	NaN	NaN	10.0	1.0
1	9628	-1	9.0	11.0	NaN	NaN	NaN	NaN	NaN	NaN
2	143872	-1	1.0	6.0	NaN	NaN	NaN	NaN	0.0	1.0
3	143873	1	1.0	8.0	NaN	NaN	NaN	NaN	8.0	0.0
4	143874	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	7.0

5 rows × 369 columns

- **`Udacity_MAILOUT_052018_TRAIN.csv`**: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- **`Udacity_MAILOUT_052018_TEST.csv`:** Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Solution Statement:

This project will be achieved in 2 parts:

- **Customer Segmentation Report :** Use unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments.
- **Supervised Learning Model:** Use the previous analysis to build a machine learning model that predicts whether or not each individual will respond to the campaign from a third dataset with attributes from targets of a mail-order campaign

Benchmark Model:

Since the problem can be modeled as a binary classification problem where 1 = converted to customer and 0 = not converted to customer, I propose using a logistic regression model as a benchmark model as this is a common, simple and time-efficient way to benchmark classification problems.

Evaluation Metrics:

I propose using F1-score as an evaluation metric as we are dealing with a binary classification problem and since we are looking at an imbalance dataset, a metric like the F1 score would paint a better picture of the model's efficacy than a simple metric like accuracy or precision and recall.

$$F1\ Score\ =\ 2\ *\ \frac{Precision * Recall}{Precision + Recall}$$

Where,

$$Precision\ =\ \frac{\#\ of\ True\ positives}{\#\ of\ True\ positives + \#\ of\ False\ positives}$$

$$Recall\ =\ \frac{\#\ of\ True\ positives}{\#\ of\ True\ positives + \#\ of\ False\ negatives}$$

Project Design:

1. **Prepare the data:**
    a. Clean the data to remove missing values or replace them with some default values
    b. Understand the data (like mean, median etc.) and perform exploratory data analysis to get a sense of the nature of the dataset that I am working with
2. **Perform feature engineering:**
    a. Scale the data so the ranges are similar (b/w 0 and 1) and not varying greatly from column to column
    b. Encode categorical data so that it is easier for the model to ingest and understand
    c. Handle outliers either by removing them or by replacing them with default values (this depends on the exploratory data analysis)
3. **Create the customer segmentation report:**
    a. Segment the customers into group using clustering methods like K-means
    b. Visualize the clusters and keep tuning the hyperparameters like no. of clusters to arrive at an optimum number
    c. Identify clusters with the highest customer concentration after settling on the right hyperparameters

4. **Build prediction model:**
    a. Build and train a binary classification model to determine if a customer would convert because of a mail-order campaign
    b. Run through different models and compare it to our benchmark logistic regression model
    c. Make predictions on test data using the model that gives us the most optimal F1 score

Platform used:

I will be using **AWS Sagemaker studio** to work on this capstone project since I am already very familiar with the platform and it provides out-of-the-box solutions for storage, endpoint for accessibility and training/tuning hyperparameters

References:
https://en.wikipedia.org/wiki/Arvato

https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226

https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10