ARVATO FINANCIAL SOLUTIONS

Project Proposal

Prepared for: Arvato Financial Solutions
Prepared by: Juanita Smith
13 November 2021

ARVATO FINANCIAL SOLUTIONS

DOMAIN BACKGROUND

Arvato Financial Solutions provides a wide range of financial services to businesses.

Their mission statement is to be backbone for seamless and secure financial transactions between businesses and consumers around the world.

Their portfolio of uses cases offered to their clients are mainly credit management, e-commerce, mobility, telecommunications and banking.

Industries covered are:

- Banking and Finance
- Energy and Utilities
- E-commerce
- Internet and Telecoms
- Insurance
- Public Sector
- Media and entertainment

Arvato partnered with Udacity, to make available a sample of the data of one of of their clients which are a mail order sales company in Germany, to help students to practise data science using real data, and a real use case

The dataset contain a lot of fields from the services offered above, like insurance, telecom, banking and therefore worth a mention to help put context behind the data.

ARVATO FINANCIAL SOLUTIONS

PROBLEM STATEMENT

We need to help Arvato answer 3 questions for their mail order sales client

1. Help the mail order company to understand who their customers are
2. How does their client base compare to the rest of Germany ?
3. How can the mail order company acquire new clients more efficiently ?

Answering these questions will help the mail-order company to:

- Increase efficiency in customer acquisition process by targeting the right people
- Target advertising to reduce costs
- Target marketing for only relevant new clients with high probably of responding
- Data-driven decision making instead of gut feel

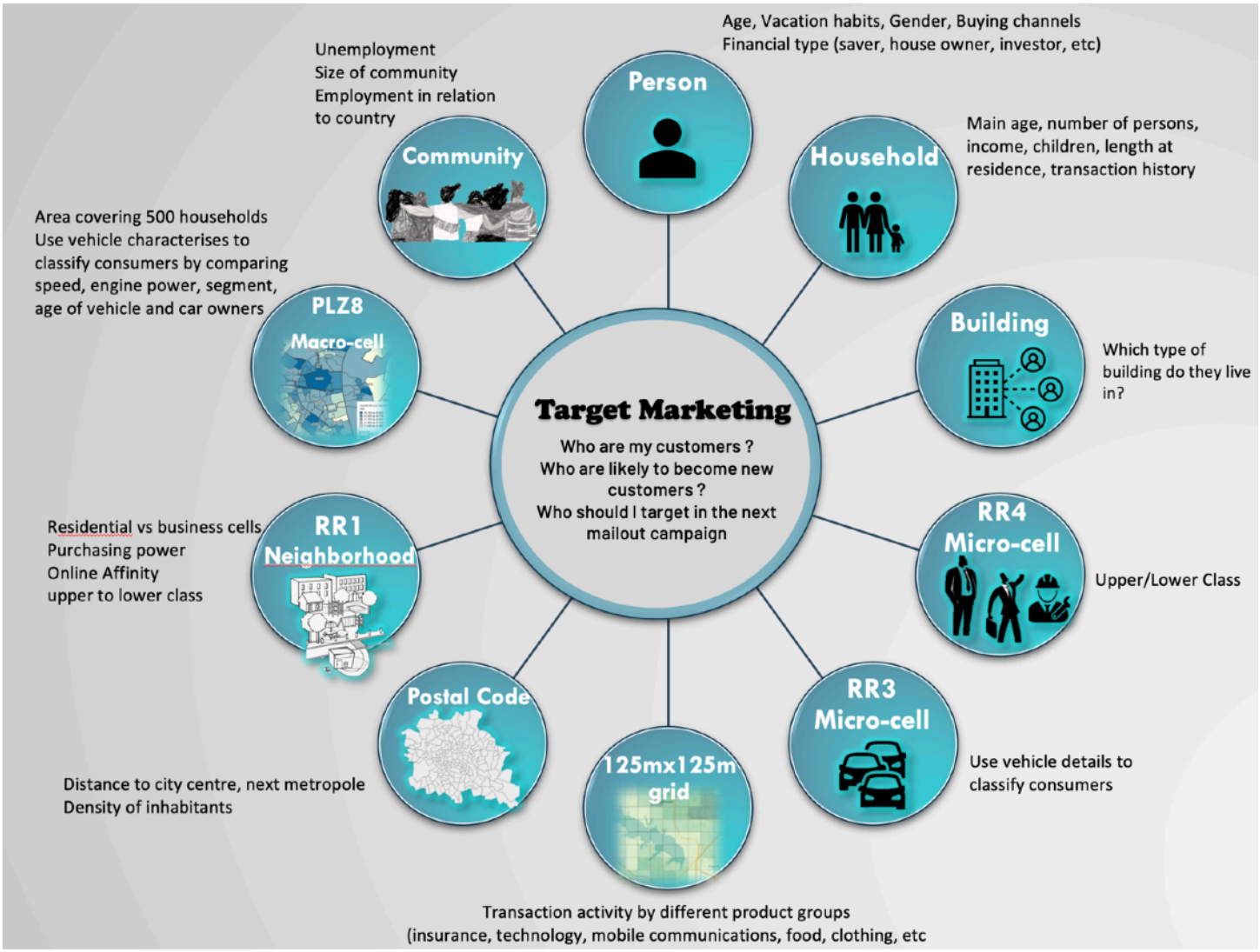
DATASETS AND INPUT

3 datasets were provided by Arvato:

1. Demographics and own attributes of the mail-order company's **existing customers**
2. A bigger dataset containing the same demographics and attributes for the wider **general population of Germany**
3. Demographic data for individuals who were **targets of a marketing campaign**, which are a subset of the same persons as mentioned in the customer dataset 1

Demographic data were obtained using open public data sources in Germany and were provided to students to downloaded from Udacity

Summary of the features provided, which can be found in all 3 datasets



Technical information:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns). This dataset is a subsample of dataset 1. A column 'RESPONSE' were provided to indicate if the individual has responded to the campaign or that. This will be used to trained the model

- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns). In this dataset, the columns 'RESPONSE' are omitted, means it's a great test to see how will the model generalise to unseen new data.

Further information given by Udacity and Arvato:

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighbourhood. Use the information from the first two files to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS"), then use your analysis to make predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed; it is against that withheld column that your final predictions will be assessed in the Kaggle competition.

Otherwise, all of the remaining columns are the same between the three data files. For more information about the columns depicted in the files, you can refer to two Excel spreadsheets provided in the workspace. [One of them](./DIAS Information Levels - Attributes 2017.xlsx) is a top-level list of attributes and descriptions, organised by informational category. [The other](./DIAS Attributes - Values 2017.xlsx) is a detailed mapping of data values for each feature in alphabetical order.

SOLUTION STATEMENT

The project will be dividend into 3 parts

1. Customer Segmentation

Use unsupervised learning methods to analyse attributes of established customers and the general population in order to create customer segments.

2. Supervised Learning Model

Use the third dataset with attributes from targets of a mail order campaign.

I'll use the previous clustering analysis to build a machine learning model that predicts whether or not each individual will respond to the campaign.

3. Kaggle Competition

Once I've chosen a model, I'll use it to make predictions on the campaign testing data as part of a Kaggle Competition.

I'll rank the individuals by how likely they are to convert to being a customer, and see how my modelling skills measure up against my fellow students.

BENCHMARK MODEL

For supervised learning, I will benchmark the model against a few scikit-learn algorithms, to get a first feel for how will the models perform without any hyper parameter tuning.

Algorithms to try at minimum:

DecisionTreeClassifier

AdaBoostClassifier - Due to previous experience in another project this algorithm performed well

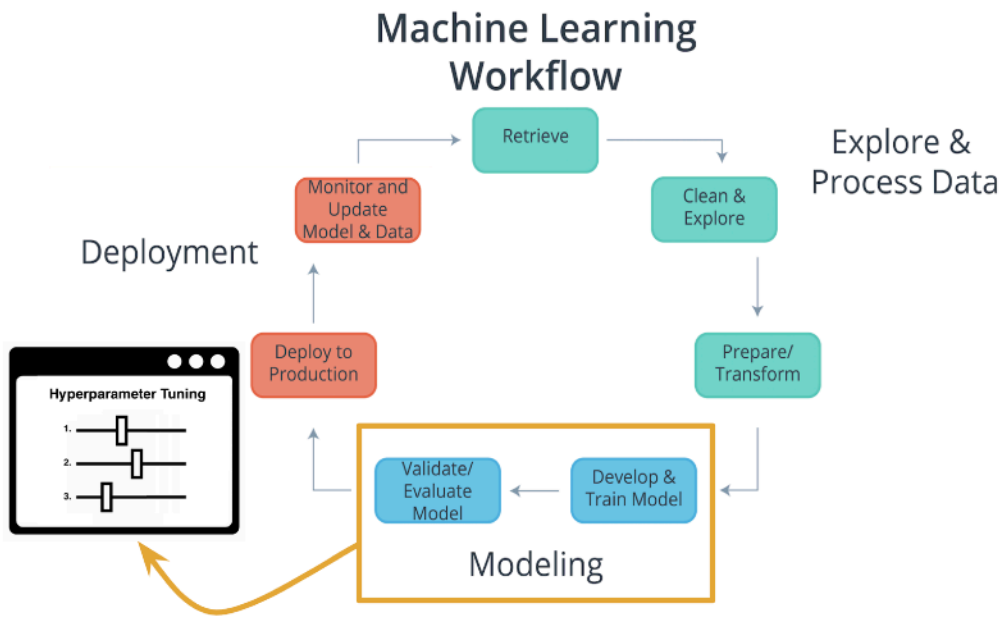
XGBClassifier - As Udacity trained a lot using the algorithm and it wins a lot of kaggle completions his has me curious

EVALUATION METRICS

There is a large class imbalance, where most individuals did not respond to the mail-out campaign. Thus, predicting individual classes and using accuracy does not seem to be an appropriate performance evaluation method. Instead, the competition will be using AUC to evaluate performance. The exact values of the "RESPONSE" column do not matter as much: only that the higher values try to capture as many of the actual customers as possible, early in the ROC curve sweep.

PROJECT DESIGN

I will follow the ML Workflow cycle as taught in the Udacity course



Project steps

1. **Retrieve and Understanding the data** provided, features, their meaning, grouping/levels and relationships.
2. **Exploratory data analysis** to help understand the data by using notebooks and a lot of plotting
3. **Data cleaning**

   As the data are obtained from public sources, the data are expected to be very messy, we were warned ! A substantial amount of effort needs so be invested in cleaning the data following steps below

   - Identifying missing values - not all data have missing data imputed as np.nan. Identify all these cases and choose how each case should be handled
   - Removing columns and rows with too many missing values
   - Removing columns with too high correlations
   - Feature engineering
     - Handling of categorical data and converting them into numerical numbers through one-hot encoding
     - Cleaning of columns where too much data are captured into one column, rather split the content

- Imputation
  - IterativeImputing will be used to impute missing values, as this use machine learning to use existing features to predict missing ones. This should produce a more accurate result than just taking mean or frequently used.
- Scaling all datasets

4. **Feature reduction through PCA**

   There are 366 features in the datasets, summarising the features first should give better clustering and predictions results

5. **Clustering** - Use K-means to cluster the data

Summarise the dimensions of the data and cluster unlabelled data into groups with similar properties

Look at relationships between demographics features, organise the population into clusters, and see how prevalent customers are in each of the segments obtained.

6. Supervised modelling and evaluation to **predict individuals who will respond to marketing campaigns**

As this course was entirely focused on deployment using AWS and Sagemaker, I will use Sagemaker's python SDK and hosted environment to further train XGBOOST with objective binary:logistic, focusing on hyper parameter tuning

7. **Validate/Evaluate model** results by focussing on feature importance and making sure the trees make sense and are actionable for marketing

8. **Final Recommendation**