Project: Customer Segmentation and Identifying Potential Customers for Arvato Financial Services

Problem statements: A mail-order sales company in Germany would like to identify the parts of the Germany population that best describe the core customer base of the company. The company would then like to use the insights obtained from the customer segmentation to help developing models that can predict which individuals are more likely to convert to their customers so that they can perform more targeted and successful market campaign in the future.

Domain Background: This project is provided by Bertelsmann Arvato Analytics, which is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include Supply Chain Management solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation and data/analytics.

This project lies in the growing field of customer analytics. Consumer product focused companies start to use internal data (such sales/transaction data, customer information) together with external data (such as population demographics, industry information) to better understand customers and their consumption behaviors. The insights from customer analytics can help companies improve business performance by conducting more targeted marketing to acquire potential customers as well as developing products and services that can better serve their existing customers and thus increasing sales.

I choose this project because I'm interested in the consumer and marketing analytics space and would like to improve my skills in this area.

Datasets are provided by Bertelsmann Arvato Analytics. There are four datasets in total.

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

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. 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 "MAILOUT TRAIN" file includes one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. All the insights learned from the first three data sets will be used to predict whether individuals in "MAILOUT TEST" set will respond or not.

Two Excel spreadsheets are provided.

- "DIAS Information Levels - Attributes 2017.xlsx" is a top-level list of attributes and descriptions, organized by informational category.
- "DIAS Attributes - Values 2017.xlsx" is a detailed mapping of data values for each feature in alphabetical order.

I do find that not all variables in the data sets have corresponding descriptions in the excel spreadsheets. However, we should not exclude a variable simply because there is no description for it. Instead, if the variables appear to be important in the analysis, we should ask more information about those variables.

According to the excel files, there are only seven numeric variables. I also checked the number of unique values for each variable to identify potential more numeric variables.

var_summary.csv summarizes the unique value for each variable and whether description about the variable can be found in the excel files. **Proportional_Missing.csv** summarizes proportional missing values for each variable, which will be used to determine whether the variable will be included in later analysis.

Solution statements:

I will first preprocess the data for modeling. The data set includes both numeric data and categorical data. Any numeric variables that have more than 30% missing values will not be used for modeling. Missing values will be filled by using median of the variable. I will also try using KNN imputer, but this may take a very long time. All categorical data will be encoded to dummy (indicator) variables. For each encoded dummy variable, I will only include it in the candidate variables for modeling if the difference between the group means (population vs customers) is statistically and economically significant. Furthermore, a minimum threshold (5%) of mean of the dummy variable is imposed. This ensures that at least a meaningful amount of people are in that encoded group and thus the dummy variable can provide valuable information.

For customer segmentation, K-means clustering will be trained on Germany population. The number of clusters can be determined based on silhouette score. The trained model will be used to predict the clusters of the customers of the mail-order company. I will then compare the distribution of clusters of the customers with the distribution of the clusters of the Germany population to understand where the customers reside. I will then examine the clusters where the proportion of the customers are significantly higher than that of the population. I will then use model (such as Random Forest Classifier) to predict the clusters and find the most important features that help define those clusters.

For supervised modeling part, I will use all the candidate variables (obtained using criteria mentioned previously) and augment them with the predicted clusters for each individual.

Evaluation metric used is area under ROC curve. I will use Naïve Bayesian Classifier as the **Benchmark model**. Other models/algorithms that will be tested include Logistic regression, Random Forest Classifier, Gradient Boosting Classifier and XGBoost Classifier. Cross validation will be used to select hyperparameters of each model. Final selection of the model will be based on the performance on the test set. Ensemble of the models results will be tested as well to further improve performance.