Mridul S Gangwar
Udacity Machine Learning Nanodegree
Capstone Proposal

Domain Background: Finding new customers for a product is always a challenge. In an ideal world where you have a lot of funds available to acquire new customers, you can run promotions, discounts, online and offline campaigns to get new customers. However, when you have limited funds and staff hours available, you should narrow down your efforts to a target segment, which is more probable of buying your products. "If I'm trying to expand sales, I have to find out who my existing customers are. What are their demographics? What do they look like?" Jerry Osteryoung, director of outreach for the Jim Moran Institute for Global Entrepreneurship at Florida State University[1].

Problem Statement: In the capstone project, I will be working on data provided by Arvato Bertelsmann. I will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. The learnings from the previous comparison will be used to create a model to predict which individuals are most likely to convert in the marketing campaign.

Datasets and Inputs: We are provided with four datasets in this project:
AZDIAS: demographics data for the general population of Germany
CUSTOMERS: demographics data for customers of a mail-order company
DIAS Attributes: detailed mapping of data values for each feature in alphabetical order
DIAS Information Level: top-level list of attributes and descriptions, organized by informational category

Solution Statement: This problem can be divided into two different subproblems -

1) Use Unsupervised Learning to perform customer segmentation and identify clusters/segments that best match the company's customer base.
2) Use Supervised Learning to identify targets for the marketing campaign who could become their customers.

The probability predicted from the model is used to identify the likelihood of responding to a mail campaign.

Benchmark Model: A predictive model of AUC 50% is not capable of identifying positive class vs. negative class. We need to create at least one model which beats our benchmark model. I generally start with a simple model like logistic regression, in this case, to begin as a benchmark model and then build on creating a more complex model like the random forest, xg boost, etc.

Evaluation Metrics: It's a classification problem, whether the person will respond to the mail or not. I am expecting the data to be unbalanced as very few people react to the mail campaigns; the best metrics to be used are AUC/ROC and f1 score metrics.

Project Design: Like other data science projects, I am expecting to go through several steps-

- Data exploration – this includes steps like distribution of null values across the columns, highly correlated columns, etc.
- Feature engineering – dropping columns with a high percentage of missing values, imputing missing values, creating new columns based on the domain experience and the data, dimensionality reduction, etc.
- Unsupervised analysis - creating customer segmentation
- Baseline modeling - creating simple logistic regression model to see how it is performing
- Secondary modeling - more advanced models like the random forest, xgboost to improve the evaluating metrics
- Conclusion and prediction for the test data set

This project workflow is not fixed, and I might add some additional steps and remove some based on the need of the project. I am expecting myself to be jumping around and switching between different stages.

[1] – How to find new customers and increase sales, by Elizabeth Wasserman
https://www.inc.com/guides/find-new-customers.html