Machine Learning Engineer Nanodegree - Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Rajesh Mittal – Submitted: 4/20/20

Domain Background

Arvato is a software services company that helps its business customers to analyze data and gather insights on a global scale.
In this project, Arvato is helping a mail-order company which sells organic products in Germany. The goal is to understand the customers base to identify which customers will be interested in company's products.
The project is relevant because customer targeting is a universal problem that is applicable to multiple industries where human cost to screen potential customers in cost prohibitive. ML methods are quite useful for this analysis. I got a chance to learn about the techniques and methods in this area when I was working on census data analysis during ML cases studies as part of my nanodegree project. I also found some good articles and research papers in this area.

https://towardsdatascience.com/market-segmentation-with-r-pca-k-means-clustering-part-1-d2c338b1dd0b
https://www.ritchieng.com/machine-learning-project-customer-segments/

Problem statement

"Given the demographic profile of a person in Germany, how can a mail order company identify new customers that are likely to use its products".
The problem as its core is a two fold. First, clustering user segments to understand correlation between two groups using unsupervised techniques. Clustering can help identify latent features that describe general population which can then be used to identify which features are good qualifiers for company's target customer base. Applying insights from this clustering, the final job is to build a supervised model to predict if a given customer is potential target for mail order campaign.

Database and inputs

There are four data files associated with this project:

- 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).

Additionally, there are two meta-data files

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

All files are provided by ML nanodegree program with partnership from Arvato. As part of working on this problem, I will have to abide by terms of services put forward as part of data sharing agreement.

Solution statement

There are two parts to the solution approach
Part1 – Do customer segmentation.

- The first step would involve data inspection and cleaning like removing missing values, encoding non-numeric data in to numerical features and normalizing the data to ensure all features have balanced ranges.
- The second step will involve dimensionality reduction to identify critical latent features from 366 high level features. I plan to use PCA for this segmentation.
- In final step, population data will be clustered using K-means to identify high level clusters in target population and identify how two customers segments are related.

Part 2 – Build a predictor

- Data cleaning and normalization of training data (very similar to first step above).
- Feature engineering to decide if features should be dropped or mapped to reduced set using what we did as part of PCA analysis.
- Run different supervised algorithms to improve prediction(XGBoost, Linear model, Random forest etc).

Benchmark model

I plan to use Logistic regression model as it is simple, map to the prediction problem well(binary classification) and also easy to train.

Evaluation metrics

For first part of the project(segmentation), I plan to use evaluate

- PCA using the number of feature needed based on variance ratio of individual features.
- K-means by distance between identified clusters, this will likely based on number of clusters.

For second part(Prediction), I plan to split training data to train and validation set. The model will be trained on the training set and will be tuned based on the its prediction performance on the validation set. I have following choices in terms of evaluation criteria

- Accuracy
- Confusion Matrix – F1 score, Recall, Precision
- Area Under the Receiver Operating Curve

The final selection will depend on data distribution of training samples. If the target labels are highly imbalanced then accuracy would be a bad choice to evaluate the model. In that case Recall or AUROC will be a better metric to use.

Project design

Here us project flow I am proposing

1. Data Cleaning and Visualization: The data will be checked for any missing values. Visualize features to

understand any noticeable patterns in the data.

2. Feature Engineering: Identify explained variance of features and determining the required number of features for maximum variance using a PCA dimensionality reduction. Use PCA to isolate any redundant/highly correlated features.
3. Clustering using unsupervised learning: Use K-Means on PCA features to segment the data into clusters.
4. Prediction on training data using supervised algorithms: Employ supervised algorithms to predict if a customer will be interested in the offer. Logistic Regression, Decision Tree, Random Forests and XGBoost will be used to make predictions and will be evaluated. The proposed evaluation metrics will be used to determine the best model.
   Finally, the best model will be used to make predictions on the test data and the predictions will be submitted on the Kaggle competition page.