Bertelsmann Arvato Capstone project

Umar Ul-Hassan

July 2021

In this project I will be analysing data about the general population of Germany along with data about existing customers in Bertelsmann Arvato in order to determine which members of the general population are the most likely to respond to mail-outs. To do this I will use unsupervised learning techniques to find clusters of the customer data and compare these clusters with the general population. Then I will create a machine learning model which will be used to determine whom are most likely to respond to the mail-out.

1 Datasets and Inputs

All the data presented below was used for creating the machine learning models and gaining insights on the data. The data listed below was provided by Bertelsmann Arvato.

- **Udacity_AZDIAS_052018.csv**: Provides the demographics for the general population. Contains 891211 entries $\times$ 366 features.
- **Udacity_CUSTOMERS_052018.csv**: Provides the demographics for customers of a mail-order company. Contains 891211 entries $\times$ 366 features.
- **Udacity_MAILOUT_052018_TRAIN.csv**: Provides the demographics for individuals whom were targeted by a marketing campaign. Contains 42932 entries $\times$ 366 features.
- **Udacity_MAILOUT_052018_TEST.csv**: Provides the demographics for individuals whom were targeted by a marketing campaign. Contains 42833 entries $\times$ 366 features.

There are also two additional sources of metadata which describe the features and the meanings behind their values.

- **DIAS Information Levels - Attributes 2017.xlsx**: Explains what the feature names represent.
- **DIAS Attributes - Values 2017.xlsx**: Explains what the value encodings represent.

2 Solution statement

In order to accomplish the objective stated at the beginning the following steps will be taken:

1. Data exploration and preprocessing The first step will be to explore the data and understand the data better. The next step will be to deal with unusual entries and to rectify them. After this columns and rows with too many missing values should be dropped. Check this
2. Feature selection Not all the features will be required for training the unsupervised model. Therefore PCA will be used as an unsupervised method to estimate the optimal number of features to use for clustering. The yellowbrick library will be used to determine the optimal number of clusters and the Calinski Harabasz metric will be used for scoring each cluster number.
3. Supervised machine learning The final step will be to produce a machine learning model which accurately predicts which individuals are more likely to respond to a mailout. The proposed algorithms which will be used are the following:
    - Logistic regression
    - K-nearest neighbours
    - XGBoost

3 Benchmark model

The benchmark model all other models will be compared to will be logistic regression due to it's ease of training and interpretability.

4 Evaluation metric

There are 3 evaluation metrics used for this report. The distortion score, chi squared and the area under the receiver operator characteristic curve (AUROC) which where chosen for unsupervised clustering, supervised feature extraction and supervised learning respectively. The reasoning for selecting each metric will be given in this section.

4.1 Distortion score

The Calinski Harabasz score is the sum of squared distances from each point to its assigned center. This makes sense for measuring how good a set of clusters are because the clusters should be well spaced from one another. [1].

4.2 Chi squared

Prior to selecting and evaluating machine learning models it is best to first select what the best features are. For a categorisation problem such as determining whom will respond to a mailout. The higher the chi squared score is, the more likely it is that the variables are dependent on each other. Therefore by creating a chi squared contingency table against the response variable we can determine which variables are the most highly correlated with the response variable and therefore which variables are the most important [2].

4.3 AUROC

The Receiver Operator Characteristic curve is an evaluation metric for binary classification problems which is precisely the problem which this report aims to solve. The higher the area under curve (AUC) is of the ROC curve, the better performing the model is at distinguishing between positive and negative classes [3].

5 Project Design

The following steps performed in the report wil be outlined in this report.

1. **Data Cleaning and visualisation**: This section explores the data for improper values and missing values along with the nature of all the data sets outlined in section 1. Based on the insights attained the datasets will be cleaned accordingly.
2. **Feature selection**: Not all of the dataset is necessary to obtain insights into the data and in some cases can lead to overfitting. Therefore for the unsupervised clustering stage of the project pca willbe performed to obtain the most relevant features and then K-Means clustering will be performed to determine the optimal number of clusters and glean insights on the nature of the typical Arvato Bertelsmann customer.
3. **Model tuning**: A variety of different algorithms will be evaluated and the best model will be selected.
4. **Evaluation and testing**: The final model will be used to perform predictions on the test set and a submission will be made.

References

[1] cg, "Easily understand k-means clustering," Nov 2016. [Online]. Available: https://avidml.wordpress.com/2016/10/29/easily-understand-k-means-clustering/

[2] S. K. Gajawada, "Chi-square test for feature selection in machine learning," Oct 2019. [Online]. Available: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

[3] A. B. am on a journey to becoming a data scientist. I love to unravel trends in data, "Auc-roc curve in machine learning clearly explained," Jul 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/