Capstone Project

Customer Segmentation Report for Arvato Financial Solutions:

1. Definition

Project Overview

Arvato Financial Services is one of eight divisions of Bertelsmann and it is headquartered in Gütersloh, Germany.

It uses cutting edge Artificial intelligence to provide financial solutions to its client in the field of

1) Identity and fraud management
2) Credit risk management
3) payment and financial services etc.

In this project both supervised and unsupervised learning techniques are used to analyze demographics data of customers of a mail-order sales company in Germany against demographics information for the general population.

Data for this project is provided by Arvato itself.

The aim of this project was to apply unsupervised learning techniques to identify segments of the population that form the core customer base for a mail-order sales company in Germany. These identified segments can then be used for direct marketing campaigns towards audiences that will have the highest expected rate of returns.

Problem Statement

Arvato is helping one of its client– a mail order sales company.

- Challenge is to analyze demographics data of existing customers of company and compare it against demographics information for the general population of Germany using clustering algorithms.
- Once a sample of population is selected, it will use supervised algorithms to predict its potential customers.
- The goal is to predict whether a person became a customer after mailout campaign.

There are 4 steps:

1. **Exploratory data analysis**: clean and transform data, remove missing values etc.

2. **Segmentation**: use unsupervised learning techniques to create clustering of customer and map it to general population.

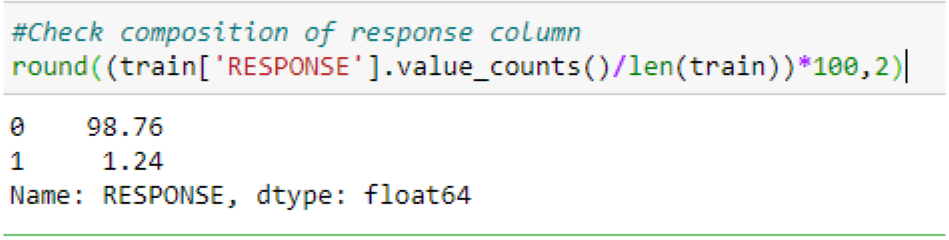
3. **Prediction**: use the demographic features to predict whether a person became a customer after a mailout campaign.

4. **Kaggle**: Use the same algorithm to predict and submit to Kaggle competition to get evaluation.

Metrics

Generally, for classification kind of problems ‘Accuracy’ is used for model evaluation purpose.

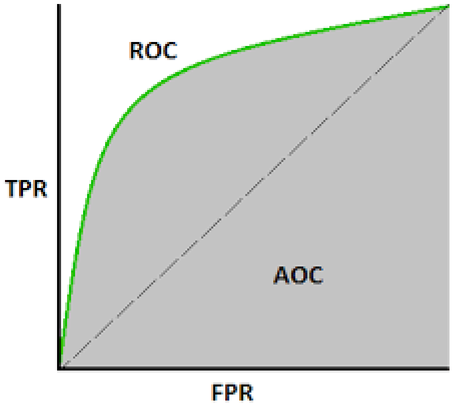
After analyzing the Train dataset we have found data is highly imbalanced. Only 1.24% data of Train dataset has positive response.



```
#Check composition of response column
round((train['RESPONSE'].value_counts()/len(train))*100,2)
```

```
0    98.76
1     1.24
Name: RESPONSE, dtype: float64
```

Hence, we have scrapped the idea of using ‘Accuracy’ as evaluation metric and choose AUC score instead.



Area under the receiver operating characteristic curve (ROC_AUC) from predicted probabilities has been used to evaluate performances of the models.

AUC (Area under the ROC Curve) provides an aggregate measure of performance across all possible classification thresholds.

2. Analysis

Exploratory data analysis

Data exploration:

Data was provided by Arvato and has 4 major components along with features dataset:

- Azdias: 891 211 persons (rows) x 366 features (columns)
- Customers: 191 652 persons (rows) x 369 features (columns).
- Train: 42 982 persons (rows) x 367 (columns).
- Test: 42 833 persons (rows) x 366 (columns).

Tasks performed on data

Missing Value Treatment.

1) Inspected features dataset and found out value [0,-1] corresponds to unknown or missing. Converted all [0,-1] values to 'NaN'
2) Found out percentage of total missing values in each columns.
3) Dropped columns which has 30% or more missing values.
4) Segregated columns into Numeric and categorical type.
5) High cardinality categorical variables have been dropped.
6) 'NaN' values have been imputed by below logic.
   - Categorical -> Mode value
   - Numeric – Median value

Scaling

7) Numeric values have been scaled using MinMaxScaler.

Encoding

8) Categorical values have been labelEncoded.

Above steps were performed on all 4 datasets in EDA phase.

Exploratory Visualization

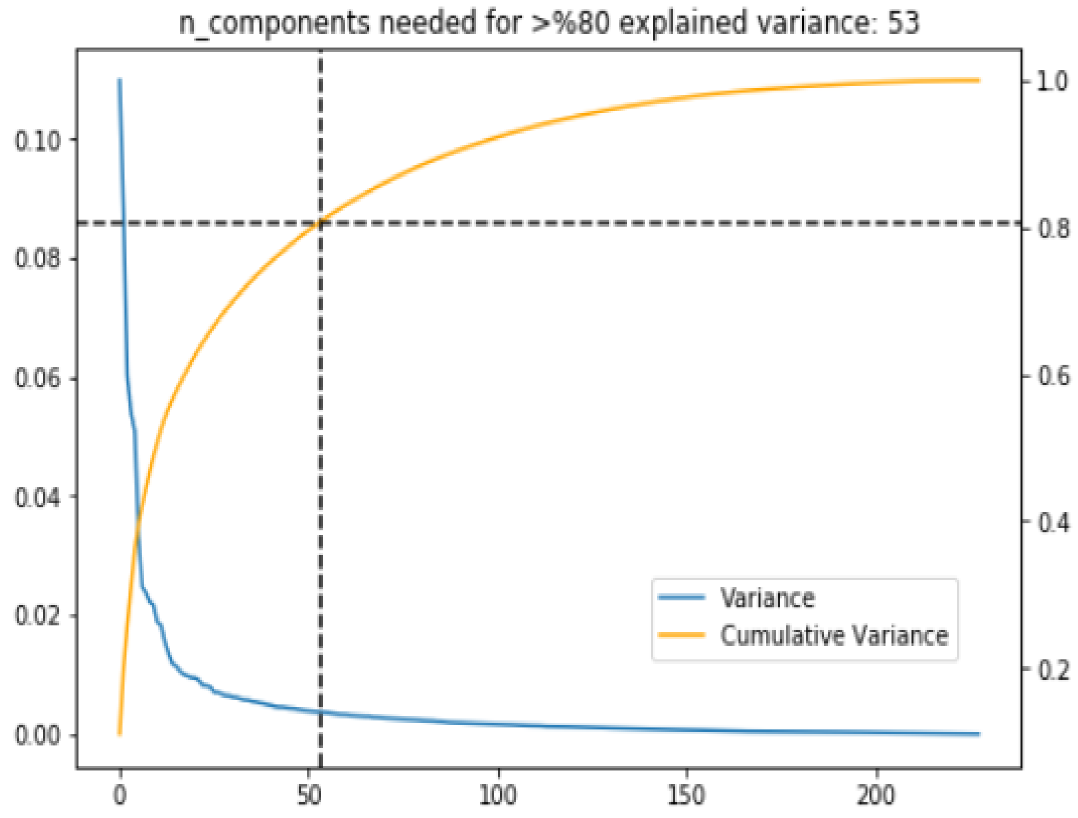
Segmentation

Applied unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general population of Germany.

Algorithms and Techniques

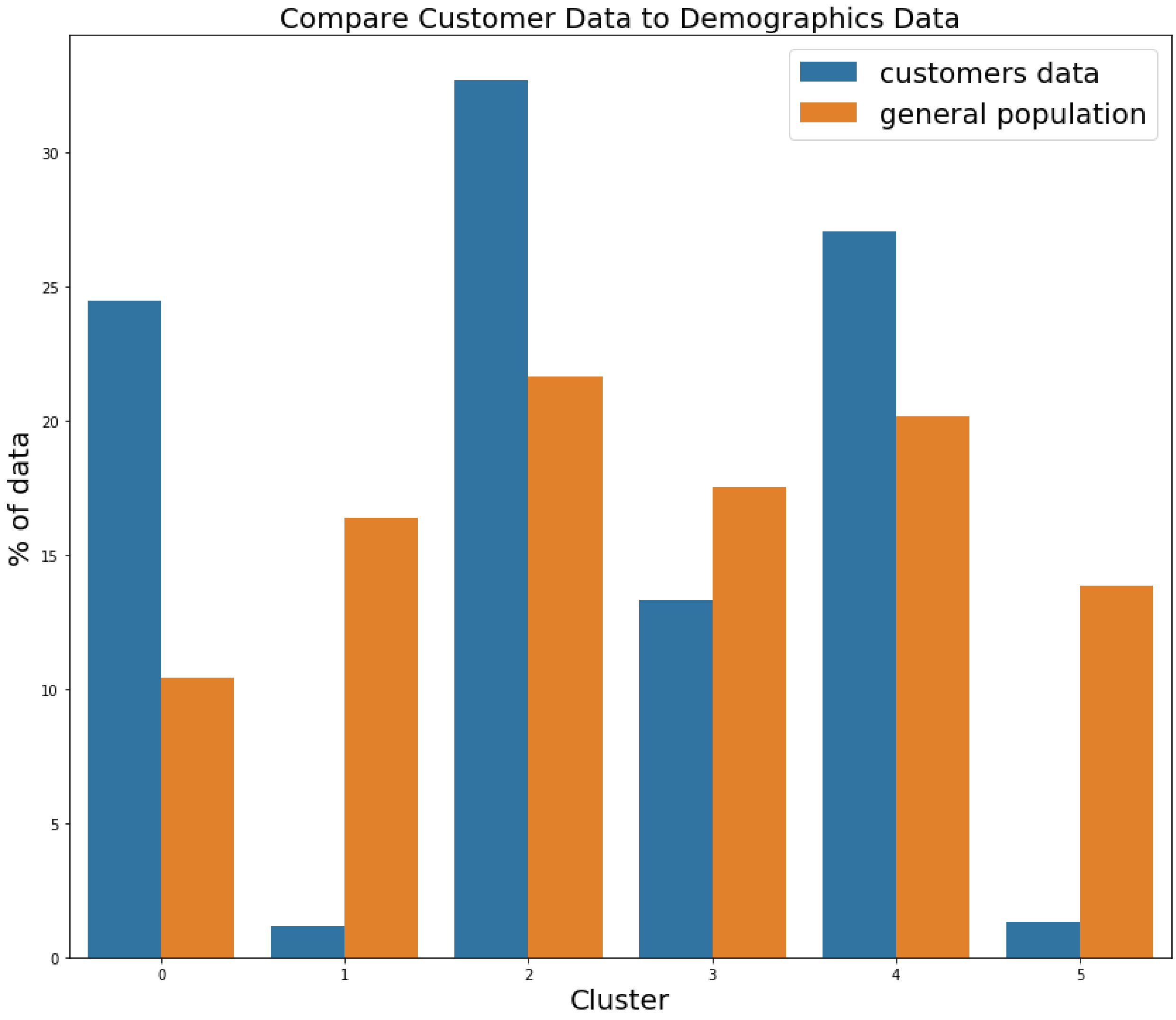
PCA

1) Applied PCA on Azdias dataset obtained after data preprocessing.
2) Goal is to find out most important components and sort them in descending order of there imporatance.
3) We took collection of top components who could explain 80% Variance.
4) 53 components were able to explain 80% variance in data.



KMeans

5) Re-applied PCA with 53 components .
6) Used K-Means clustering technique.
7) Applied elbow method to obtain optimal number of clusters.
8) Found 6 clusters to be suitable.
9) Applied K-means technique on Azdias data and customer data.



Benchmark : I used Logistic regression model as benchmark in supervised section.

3. Methodology

Data Preprocessing

All the 4 datasets were preprocessed in the EDA phase.

- Azdias
- Customers
- Train
- Test

Steps :

- Filling NAN values in place of [0,-1]
- Treating all missing values.
- All the columns with more than 30% missing values were dropped.
- Scaling data by applying MinMaxScaler
- Applying Label Encoder on Categorical data.

Implementation: Supervised Learning Techniques

Prediction

Now that we have found which parts of the population are more likely to be customers of the mail-order company, it's time to build a prediction model.

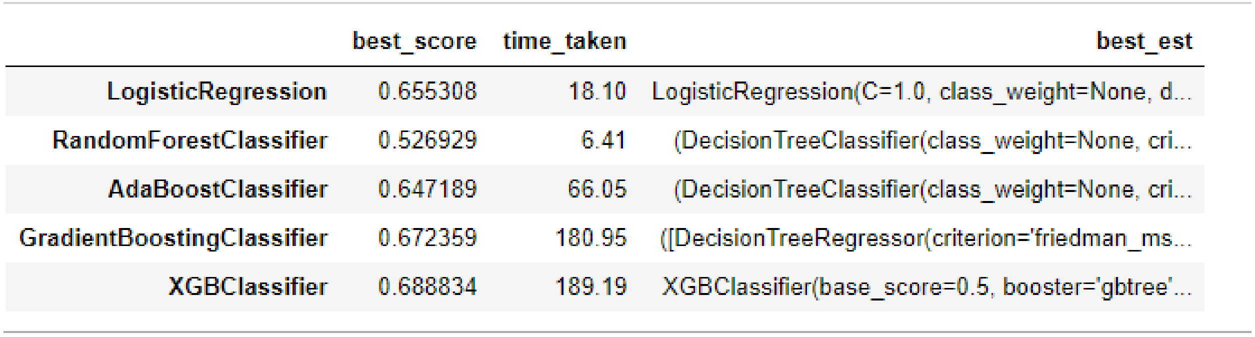
Goal is to use the demographic information from each individual to decide whether it will be worth it to include that person in the campaign.

1) We have used classification models to learn from ‘Response’ column of Train dataset whether a person became a customer of the company.

2) Classification models used are as below :

   a) Logistic Regression (Benchmark Model)
   b) Random Forest Classifier
   c) AdaBoost Classifier
   d) Gradient Boosting Classifier
   e) XGBoost library.

3) We have used Grid SearchCV along with K fold cross validation technique to obtain best results



	best_score	time_taken	best_est
LogisticRegression	0.655308	18.10	LogisticRegression(C=1.0, class_weight=None, d...
RandomForestClassifier	0.526929	6.41	(DecisionTreeClassifier(class_weight=None, cri...
AdaBoostClassifier	0.647189	66.05	(DecisionTreeClassifier(class_weight=None, cri...
GradientBoostingClassifier	0.672359	180.95	([DecisionTreeRegressor(criterion='friedman_ms...
XGBClassifier	0.688834	189.19	XGBClassifier(base_score=0.5, booster='gbtree'...

4) We are getting best results with XGBoost library, which is generally a goto model for Kaggle competitions as well.

Model Evaluation

5) We have used AUC-ROC as evaluation metric because of highly imbalanced data. It is generally one of the best metrics for classification problems.

Refinement

Model Tuning

6) After a bit of model tuning in XGBoost results were increased from

```
0.6888 to .7018
```

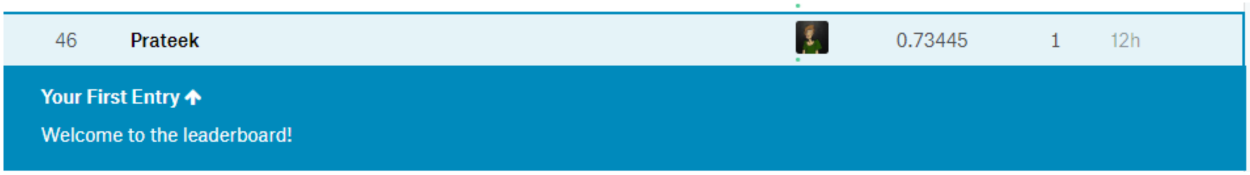
Results

Model Evaluation and Validation

- We have got AUC score of .6888 using XGBoost.
- After a bit of performance tuning results were improved to .70
- We are getting best results with XGBoost library,
- which is generally a goto model for Kaggle competitions as well.

Kaggle

I have scored rank 46th on the leaderboard.



46	Prateek	0.73445	1	12h

Your First Entry ↑

Welcome to the leaderboard!

Justification

1) AUC score of .70 is quite good indicator of a successfully classification model
2) Kaggle rank of 46 is very good indicator of a robust model

Conclusion

Reflection

- I have performed Exploratory data analysis on all datasets at once to focus on more important tasks like Segmentation and classification.

- In the Segmentation part, I have performed data pre-processing and used PCA method combined with kmeans to get the clustering of different population.

- I used 6 clusters to continue with clustering process.

- The difference was discussed, and I have known that which population might be potential customers and which population might not. With understanding the difference, a company can be much more focus on their target, and then increase conversion rate or lower down their marketing cost. The impact is large.

- In the supervised learning part. I applied various classification model and choosed the best out of it.

- XGboost model gave me good results in Kaggle competition as well.

Improvement

In order to increase the performance of the supervised learning model, the following parts might be conducted.

- Use better missing value treatment approach.
- Use better imputation strategy.
- SMOTE techniques for under-sampling or over- sampling data.
- Increase PCA components.