Customer Segment Identification

Capstone Report

Ken Chen

chken4869@hotmail.com

1. Introduction

Bertelsmann, Arvato Financial Solutions client, the mail-order company, wanted to have an efficient way for targeting the customers for their mailing campaign. The objective of this project is to analyze demographics data of customers of a mail-order sales company in Germany that sells organic products and compare that to demographics data of the general population of Germany. The end goal is to be able to predict, based on demographics data, which individuals from the general population should be targeted in the mail-order campaign.

Both unsupervised and supervised learning techniques will be used. Unsupervised learning will be used to help identify segments of the general population of Germany that best matches the existing customer base of the company. A supervised learning prediction model will be developed to predict the likelihood of whether or not an individual of the general population will become a customer. The dataset was provided by a real business - Bertelsmann Arvato Analytics and represents a real-life data science task.

The two problems I tried solving here are:

1. Customer Segmentation, where using demographic information of the general population of Germany and the customer demographic information I tried representing them in cluster groups

2. Potential Customer Detection with Supervised Learning, where I created model to predict if the person will respond to mailing campaign or not.

The main goal here is to provide an opportunity for the client, the mail-order company to focus and allocate its precious resources efficiently on the potential customers rather than whole set of population.

2 Customer Segment Identification

2.1 Project Overview

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. By enabling companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention.

In this project, we will analyze demographics data for customers of a mail-order sales company in Germany, find custom segments and predict potential customs' response to market campaign. We will use unsupervised machine learning algorithms to analyze the market and segments, selecting the main characteristics that can best describe the company's consumer.

2.2 Problem Statement

- How to segment customers, and what are the features used for clustering?
- Identify the difference or similarity of population and customer data?
- What features stand out for the segmentation?
- Which clustering model is more suitable for separating customers from population?
- Which cluster is underrepresented in general population group to customer group?

2.3 Metrics

For clustering models, we do not have a solid evaluation metric to evaluate the outcome of different clustering algorithms. The predicted clusters can be used in other prediction models, where we can further evaluate the effectiveness of segmentation using downstream model's evaluation metrics. In this project, two metrics were used to select the optimal K: Elbow method by Inertia, and Silhouette Scores.

Inertia is the mean squared distance between each instance and its closet centroid. The Elbow method means plotting the inertia change against number of K and choosing K where the slope inertia started to decrease or stabilized.

Silhouette core is the mean silhouette coefficient over all the instances, ranging from [-1, 1]. A coefficient closes to 1 means that the instance is well inside its own cluster and far from other clusters, while 0 means it is close to a cluster boundary, and finally, -1 means that instance may have been assigned to the wrong cluster.

2.4 Analysis

2.4.1 Data Exploration

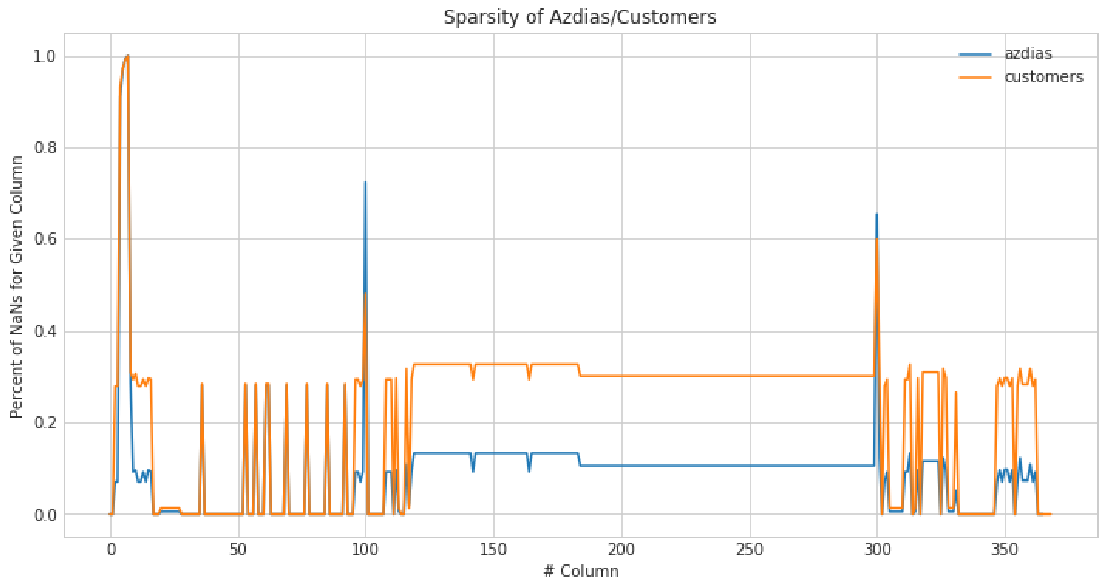
The datasets are provided by Arvato Financial Solutions. Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

Data	Usage	Description	Size
DIAS Information Levels - Attributes 2017.xlsx	Metadata	Top-level list of attributes and descriptions, organized by informational category	
DIAS Attributes - Values 2017.xlsx	Metadata	Detailed mapping of data values for each feature in alphabetical order.	
Udacity_AZDIAS_052018.csv	Custom Segmentation	Demographics data for the general population of Germany	891211 persons x 366 features
Udacity_CUSTOMERS_052018.csv	Custom Segmentation	Demographics data for customers of a mail-order company	191652 persons x 369 features
Udacity_MAILOUT_052018_TRAIN.csv	Predict Model	Demographics data for individuals who were targets of a marketing campaign	42982 persons x 367 columns
Udacity_MAILOUT_052018_TEST.csv	Predict Model	Demographics data for individuals who were targets of a marketing campaign	42833 persons x 366 columns

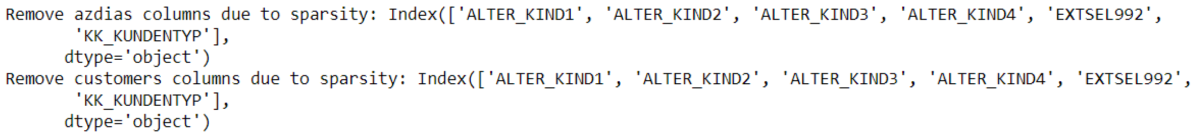
2.4.2 Data Processing

1) **Sparsity of Values by Columns**

The figure below shows the proportion of NaNs of a column feature. Similarly, Customer data also had similar fashion with a few large outliers.



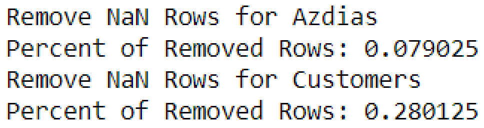
Here we can remove population and customers columns due to data sparsity:



```
Remove azdias columns due to sparsity: Index(['ALTER_KIND1', 'ALTER_KIND2', 'ALTER_KIND3', 'ALTER_KIND4', 'EXTSEL992',
       'KK_KUNDENTYP'],
      dtype='object')
Remove customers columns due to sparsity: Index(['ALTER_KIND1', 'ALTER_KIND2', 'ALTER_KIND3', 'ALTER_KIND4', 'EXTSEL992',
       'KK_KUNDENTYP'],
      dtype='object')
```

2) Sparsity of Values by Rows

Many samples(rows) have large amounts of NaN values. Therefore, we can remove these rows for our following model.



```
Remove NaN Rows for Azdias
Percent of Removed Rows: 0.079025
Remove NaN Rows for Customers
Percent of Removed Rows: 0.280125
```

3) Fill Missing Values

Here we fill missing numeric values with their median by column and fill missing categorical values with their mode by column.

3) Handle Dummy Variables

For categorical variables, we convert them to dummy variables.

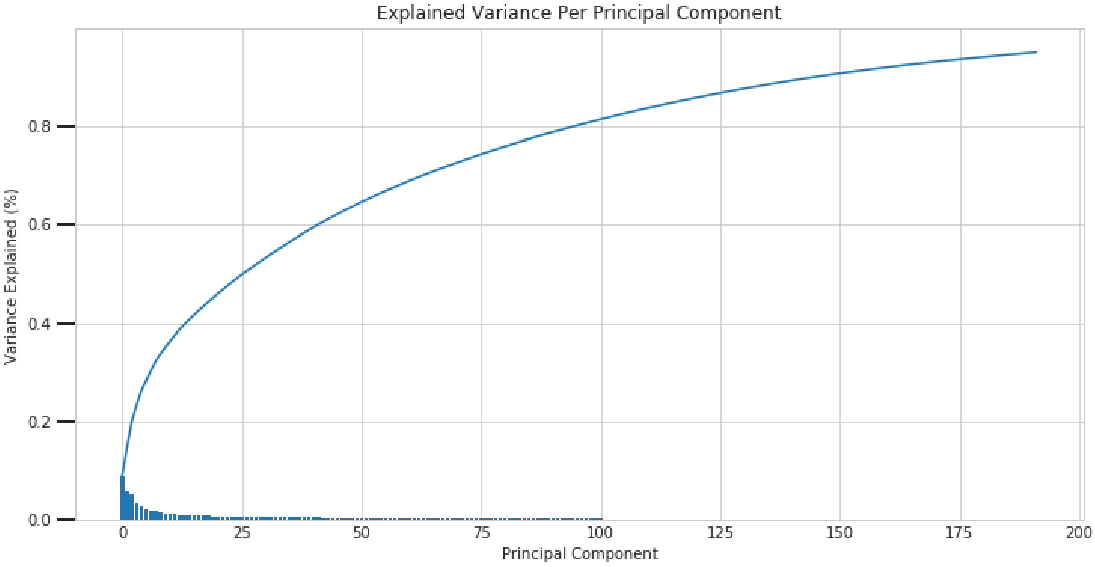
4) Feature Scaling

To standardize the data for further analysis, we apply MinMaxScaler on population data.

2.5 Dimension Reduction Using Principal Component Analysis

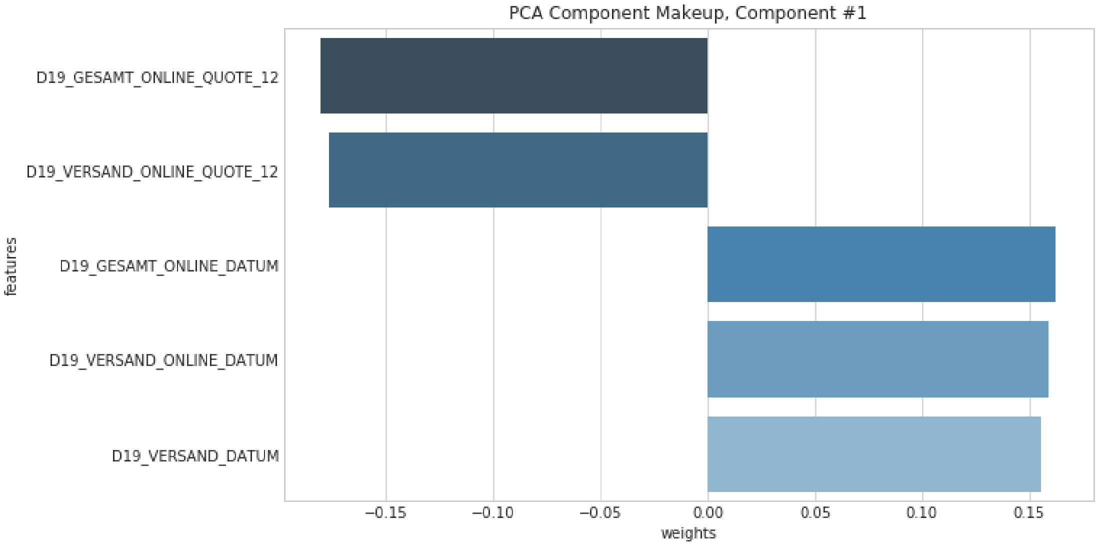
We have many components (300 features) and a lot of features are large correlations. Therefore, we need to use PCA to minimize the redundancy of the features.

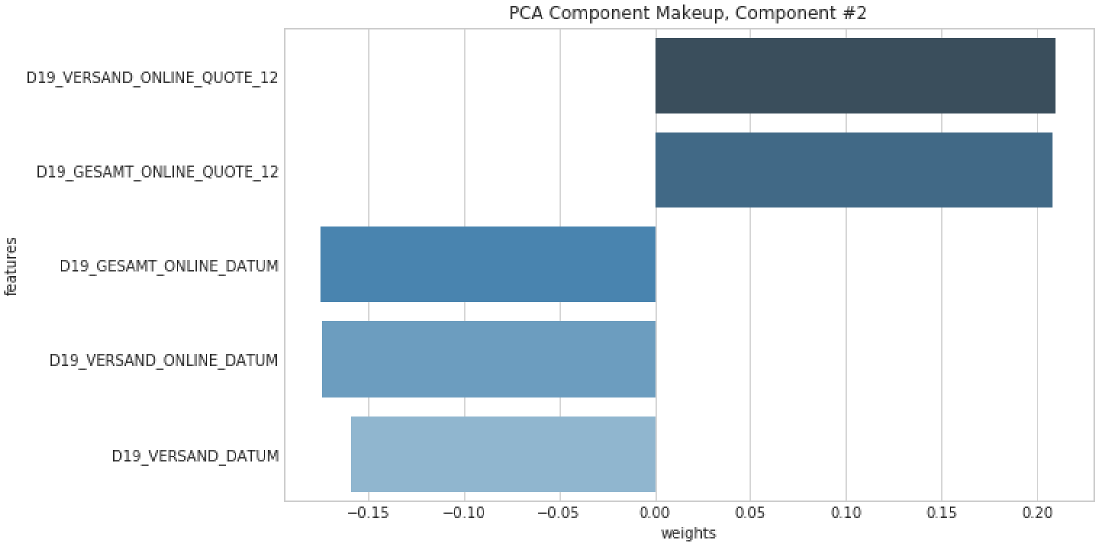
To choose the number of components, I plotted the cumulative explained variances along with the number of components, and chose 100 components corresponding to 0.81 cumulative explained variances, as showed by below figure:

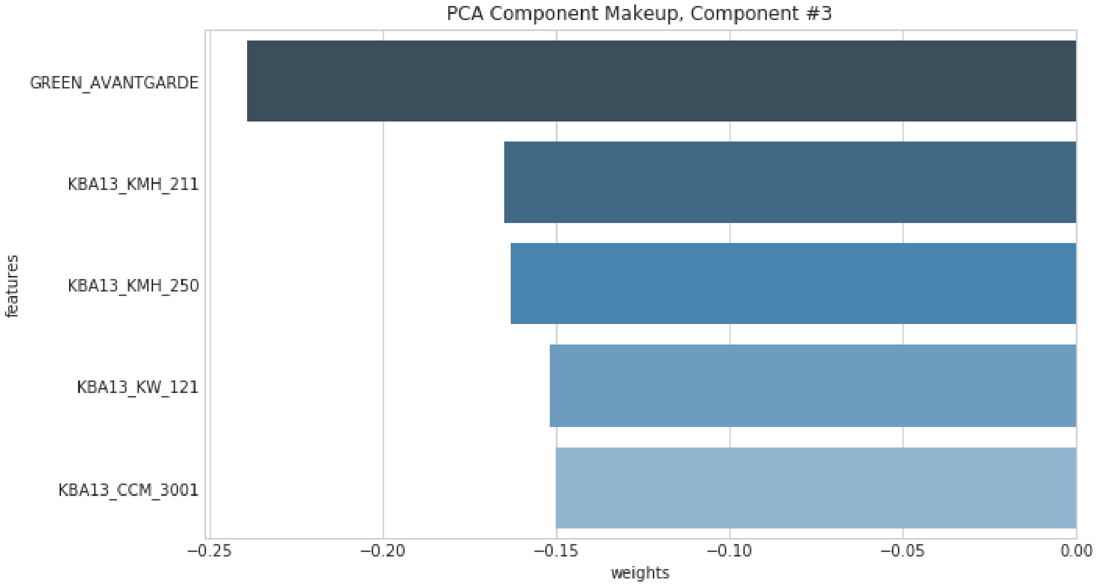


The plots above show that the explained variance plateaus a couple of times. The first and sharpest drop-off is around 15 components. At 15 components around 43 ％ of the total variance is explained. After this we can observe two others, but not as steep drop-offs around 100 and 200 components. At 100 components 81 ％ of the total variance is explained and at 150 90 ％. With each consecutive component less and less variance is explained which means that after a while the amount of additional explained variance probably is not worth the extra number of components. With that said 100 components could be considered a good middle point.

Each component had weighting scores of all original features, to represent their contribution in the component. Below figure shows the first three components and top 5 features with highest weights in this component.


PCA Component Makeup, Component #1
features
D19_GESAMT_ONLINE_QUOTE_12
D19_VERSAND_ONLINE_QUOTE_12
D19_GESAMT_ONLINE_DATUM
D19_VERSAND_ONLINE_DATUM
D19_VERSAND_DATUM
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
weights


PCA Component Makeup, Component #2
features
D19_VERSAND_ONLINE_QUOTE_12
D19_GESAMT_ONLINE_QUOTE_12
D19_GESAMT_ONLINE_DATUM
D19_VERSAND_ONLINE_DATUM
D19_VERSAND_DATUM
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
weights



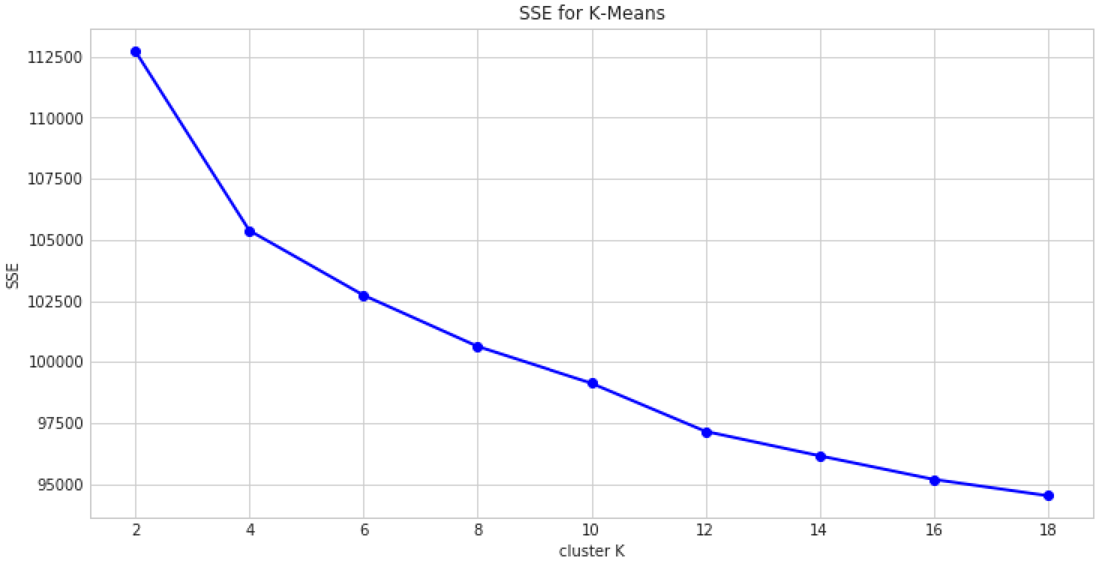
After PCA, we reduce the dimension of our data to:

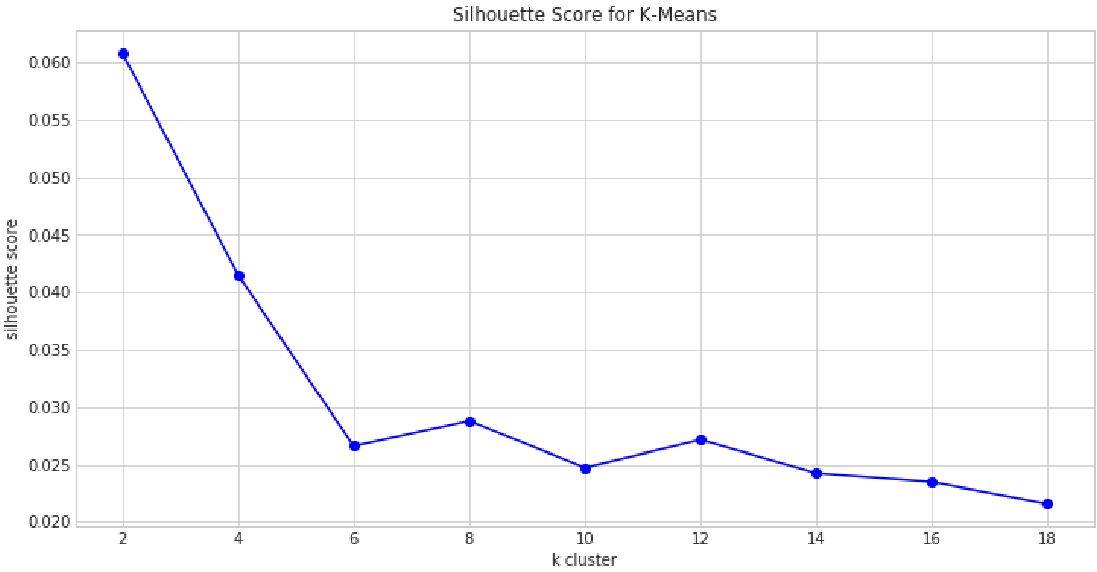
Shrinked Population Shape: 5759 * 192

Shrinked Customers Shape: 5759 * 192

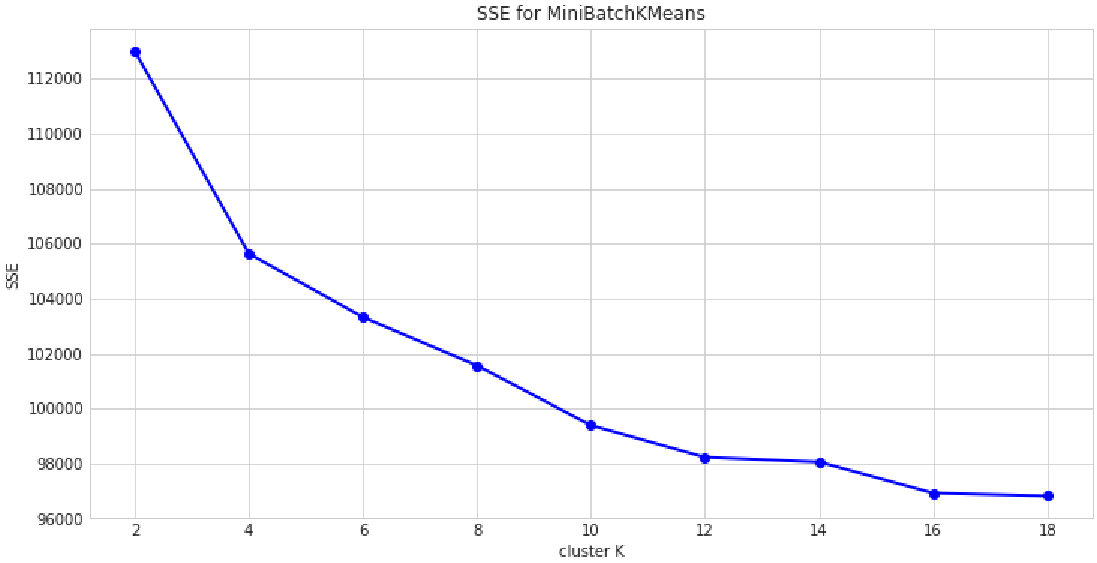
2.6 Clustering Models

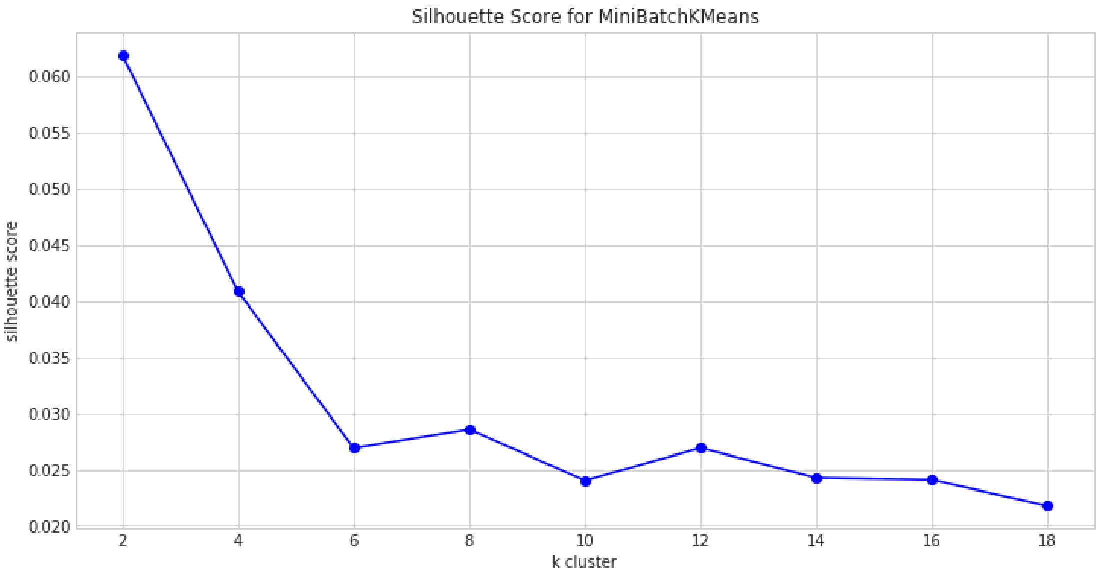
1) benchmark model: K-Means





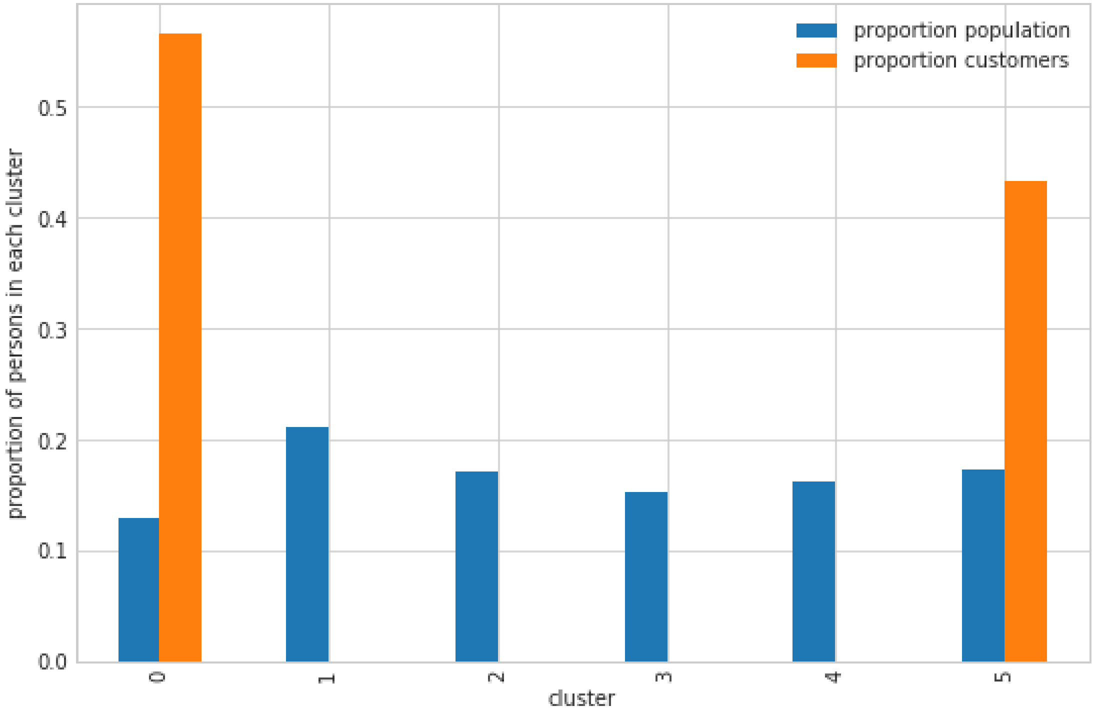
2) Mini-Batch K-Means





Although the elbow shape was not immediately apparent. At K >10, the slope of inertia started to decrease. I chose K=10 as optimal clusters. Similarly, silhouette scores became stabilized after K=10.

Assume the General Population and Customer Group were from the same population, the portion of each cluster in the data should be similar. If any segments of the population that are interested in the company's products, then we should see a mismatch in these two data. The following figure shows the cluster proportion in each general and customer group for Mini-Batch K-Means model.



Clusters 0, 5 were overrepresented in Customer compared to Population data. The proportion of people in these two clusters were significantly higher, which implies that they were the potential customers on company's products. These would be the company's interest to convert them into customers.

.

3 Detect Potential Customers

3.1 Project Introduction

After segmenting customers and identifying the potential customers, I built a classification model to make prediction if an individual would be likely to convert to customer by responding to the marketing campaign.

3.2 Problem statement

Which individual is likely to be a customer for mail-order business?

3.3 Metrics

For classification problem, we follow Kaggle's criteria AUC score, to evaluate model performance. Ideally, we would like the model to increase true positives rate but control false positive rate.

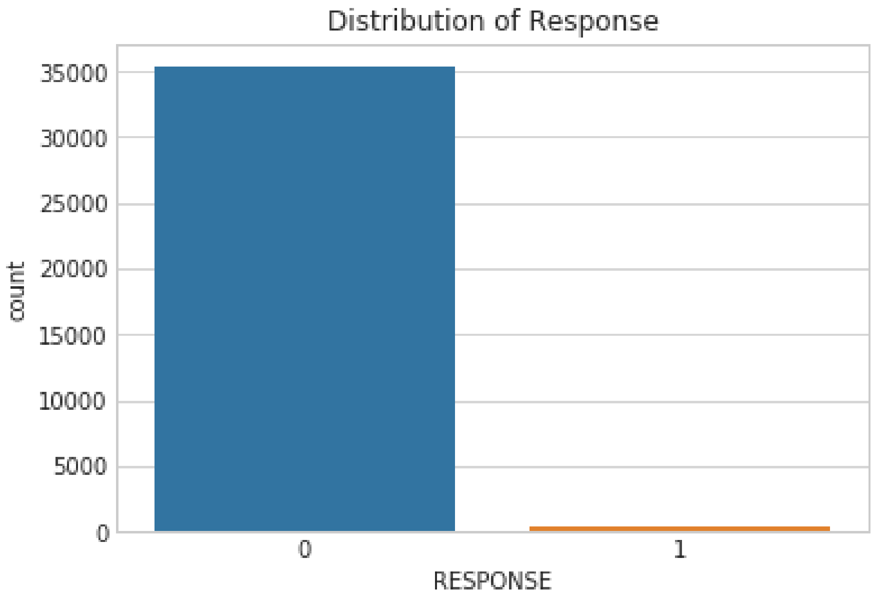
3.4 Analysis

1) Data Cleaning and Processing

The operations are the same as those for the previous project.

2) Imbalanced Data Issue

The negative class outweighs the positive class 99 percent. This problem can be solved by two resampling method: oversampling the minority class by creating synthetic data or undersampling the majority class by dropping instances.

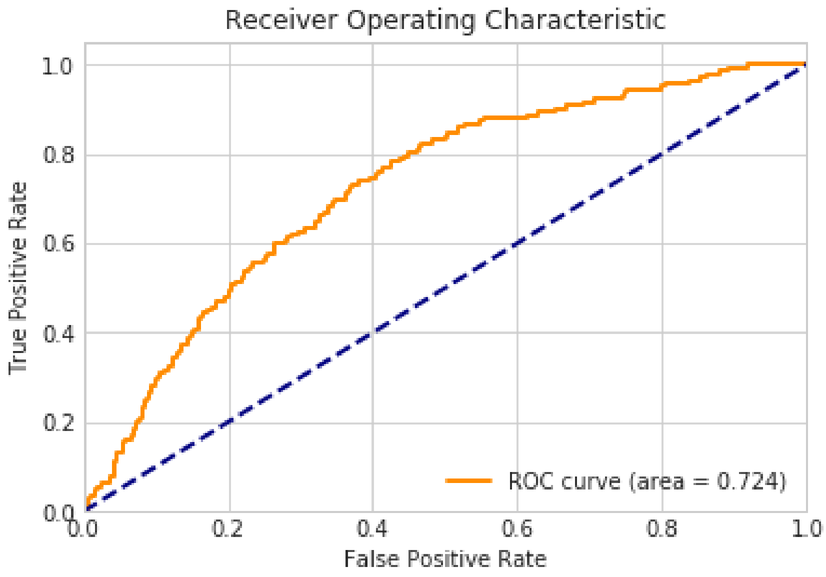


3.5 Classification Models

We build classification methods on data cleaned and processed from given datasets. We also use grid search to fine tuning the parameter for the following models to choose the model with best parameters.

1) Benchmark: Logistics Classifier

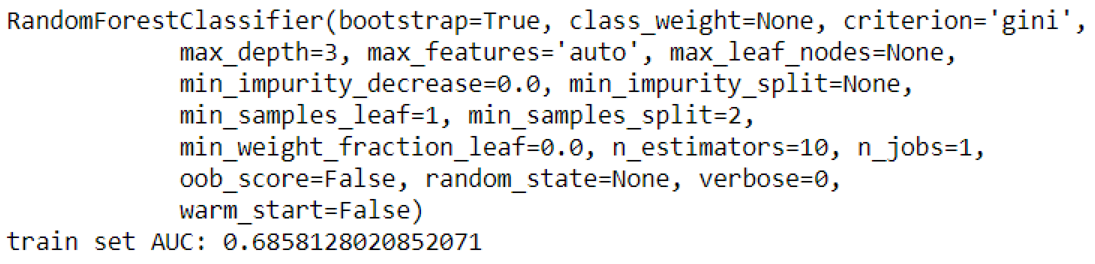
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
train set AUC: 0.6469209960897228
```



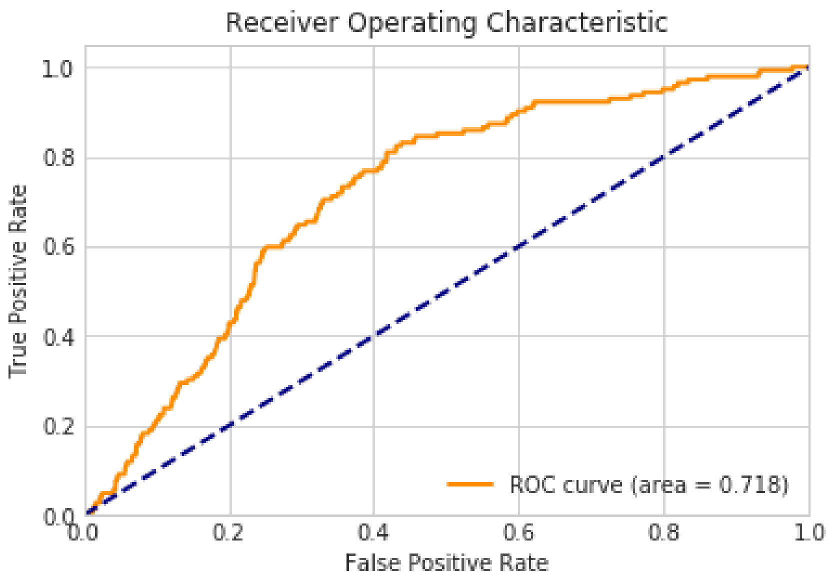
2) Random Forrest Classifier

Random forest uses a modified tree learning algorithm that inspects, at each split in the learning process, a random subset of the features. The reason for doing this is to avoid the correlation of the trees. Correlation will make bad models more likely to agree, which will hamper the majority of vote in classification problem.

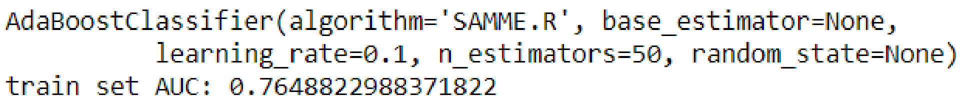
Random Forrest does not improve the performance much.



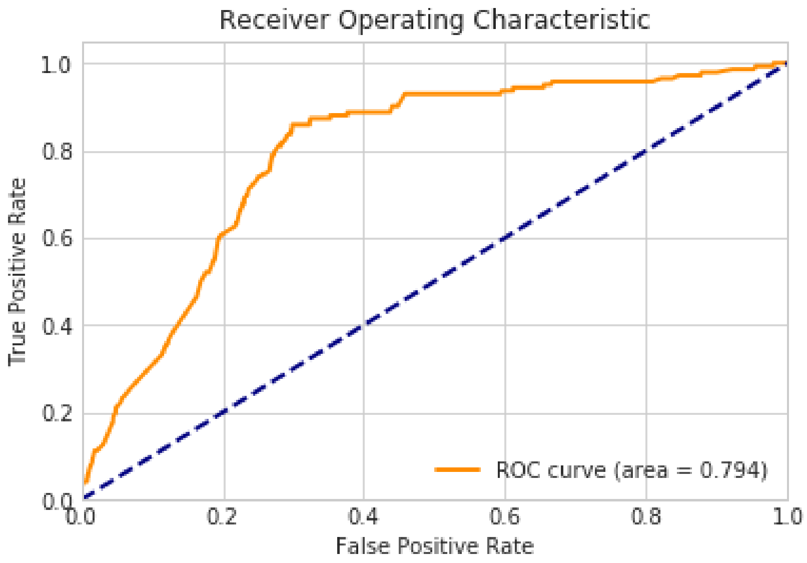
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=3, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
train set AUC: 0.6858128020852071
```



3) AdaBoost Classifier

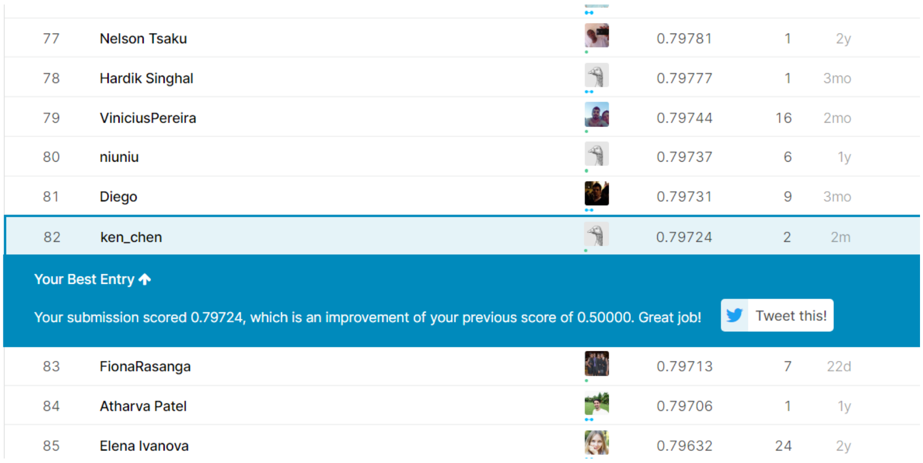


```
AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None,
          learning_rate=0.1, n_estimators=50, random_state=None)
train set AUC: 0.7648822988371822
```



3.6 Submission for Kaggle Competition

The best achiever was AdaBoost model. The score on Kaggle was 0.79, same as the performance on the testing set.



77	Nelson Tsaku	0.79781	1	2y
78	Hardik Singhal	0.79777	1	3mo
79	ViniciusPereira	0.79744	16	2mo
80	niuniu	0.79737	6	1y
81	Diego	0.79731	9	3mo
82	ken_chen	0.79724	2	2m

Your Best Entry ↑

Your submission scored 0.79724, which is an improvement of your previous score of 0.50000. Great job!

83	FionaRasanga	0.79713	7	22d
84	Atharva Patel	0.79706	1	1y
85	Elena Ivanova	0.79632	24	2y

4 Project Conclusion

In this project, I trained a K-means model and a Mini-Batch K-Means model on population data, and then used the model to cluster the customer data. 10 clusters were retrieved, among which 2 clusters were overrepresented in customer data in terms of number of people, compared against population. I also identified the prominent principle components in each cluster using heatmap. Later on, we reverted to the original features in each segment to analyze the feature distribution in both general and customer data, and I found out the divergent pattern in these features between the two data.

In the Detect Potential Customers project, I trained several classifier on training data and used it to predict test data. Even without changing models, by further cleaning data, selecting the right encoding methods, the model could achieve a better result. With hyperparameter tuning, the model was further improved by it was not significant.

References

[1]. Bertelsmann, Arvato. About.2020.

https://www.bertelsmann.com/divisions/arvato/

[2]. Custom Segmentation Definition.

https://searchcustomerexperience.techtarget.com/definition/customer-segmentation

[3]. Reasons for Customer Segmentation

https://ocw.mit.edu/courses/sloan-school-of-management/15-902-strategic-management-i-fall-2006/lecture-notes/custseg.pdf

[4]. Udacity, Bertelsmann/Arvato Project Overview. 2019.

https://classroom.udacity.com/nanodegrees/nd009t/parts/2f120d8a-e90a-4bc0-9f4e-43c71c504879/modules/2c37ba18-d9dc-4a94-abb9-066216ccace1/lessons/4f0118c0-20fc-482a-81d6-b27507355985/concepts/8400bad9-69b4-4455-826c-177d90752f00

[5]. Kaggle.com. (2020). Udacity+Arvato: Identify Customer Segments | Kaggle.

https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard .