CUSTOMER SEGMENTATION

CL311-Technical Writing

Written by: Tanu Agarwal **Roll no.: 180107070**

Abstract: In this report, we designed and implemented a customer segmentation model for an automobile company. In this process, our model learnt using the past groupings of customers on the basis of age, gender, interests etc., then analyzed the current data and applying the given algorithms predicted the groups of given customers. Our algorithms are shown to perform better than majority classifiers.

Key words: Customer Segmentation, Algorithm, Random Forest, Decision trees, dataset, features and labels.

Introduction: An automobile company has plans to enter new business sectors with their current items (P1, P2, P3, P4 and P5). After escalated statistical surveying, they've concluded that the conduct of new market is similar to their current market.

Companies utilizing customer segmentation work under the way that each client is extraordinary and that their showcasing endeavors would be better off in the event that they target explicit, littler gatherings with messages that those purchasers would discover significant and lead them to purchase something. Companies likewise want to increase a more profound comprehension of their customers' inclinations and requirements with finding what each fragment finds generally important to all the more precisely tailor advertising materials toward that portion.

In their current market, the business group has ordered all customers into 4 segments (A, B, C, D). At that point, they performed divided effort and correspondence for various segments of customers. This system has work exceptionally well for them. They intend to utilize similar procedure on new business sectors and have recognized 2627 new possible clients.

We were needed to assist the manager with anticipating the correct groups of the new clients.

Methodology: Customer segmentation is the practice of partitioning a customer base into groups of people that are comparative in specific ways relevant to marketing, such as age, gender, interests and spending habits.

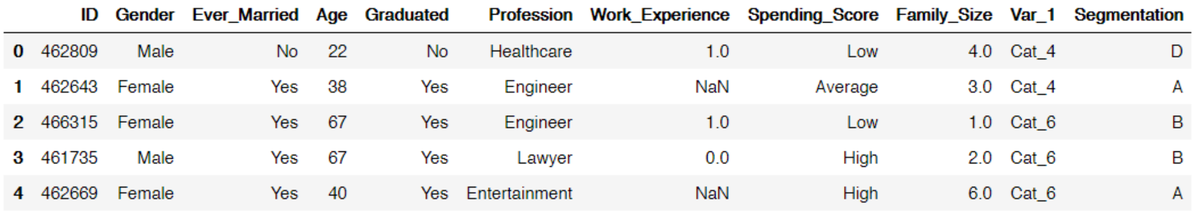
Our model used the previous data given by the company with all the features of the customers.

Data description:

- Variable - Definition
- ID - Unique ID
- Gender - Gender of the customer
- Ever_Married - Marital status of the customer
- Age - Age of the customer

- Graduated - Is the customer a graduate?
- Profession - Profession of the customer
- Work_Experience - Work Experience in years
- Spending_Score - Spending score of the customer
- Family_Size - Number of family members for the customer (including the customer)
- Var_1 - Anonymized Category for the customer
- Segmentation - (target) Customer Segment of the customer

The following is a sample of the given data:



	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

Using the above data and **RandomForestClassifier** as the classifying algorithm, we achieved a very high accuracy on the given test data set.

Algorithm of RandomForestClassifier involves the following steps:

- Bootstrap
- Bagging/Aggregation

Random Forest uses **Decision trees** as its base learning models. We haphazardly do row sampling and feature sampling from the dataset framing test datasets for each model. This part is called **Bootstrap.**

Bagging, then, is combining the results of different Decision trees. Using the following formula:
Given training set X= x_1, …, x_n and labels Y= y_1, …, y_n, bagging B times,

Then for b = 1, …, B;
We take n examples from X, Y and train a tree f_b on X_b and Y_b, so, for unseen samples x',

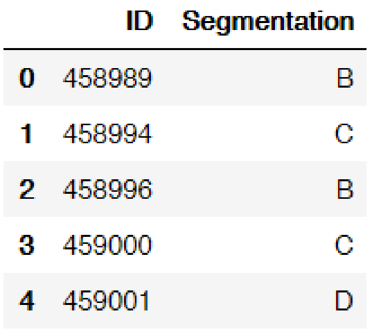
$$\hat{f} = (1/B) * (\textstyle\sum_{b=1}^{B} f_b(x'))$$

Or we can also take majority votes in trees.

Observations: The following observations came after examining the data given:

- The maximum number of people belonged to group D.
- The accuracy reached on validation data was 92.3%.

Results: The accuracy achieved on test data was 90.8%. A sample of the final result:



	ID	Segmentation
0	458989	B
1	458994	C
2	458996	B
3	459000	C
4	459001	D

Discussion: As shown by the above high accuracy of the test data, the growth in company's sales is expected. It actually increased by around 15% because of better customer understanding and thus better customer relations and a better sale.

References:

- Report Content: Self Written
- Problem Statement: https://datahack.analyticsvidhya.com/contest/janatahack-customer-segmentation/#ProblemStatement