Customer Segmentation

Technical Report

18 Sep 2019

Problem statement

We want to help our marketing team to spot the target customers easily and efficiently, so the marketing campaign reaches its prospective customers.

Today our supermarket mall has thousands of customers. Only 60% of our campaigns reach target customers. In other words, 40% of our campaigns reach the wrong customers. That means we are wasting our resources, additionally, there is a notable decrease in revenue.

We will use Machine Learning and Deep Learning techniques to segment and categorize our customers. That is going to improve identifying the target customers for each marketing campaign.

Exploratory Data Analysis

We need to Explore our data to answer some important questions

- What is the form of the data?
  - Structured
    - CSV file
    - SQL DB
  - Unstructured
    - Text
    - Images
- Does the data have labels?
- How many instances does the data have?
  - Small Dataset – less than 10000 Instances
  - Large Dataset – more than 10000 Instances
- How many features does the data have?
  - Multidimensional Dataset
    - Small Dataset – less than 10 dimensions
    - Large Dataset – more than 10 dimensions
- What is the type of each feature?
  - Quantitative (Interval)
  - Qualitative (Nominal)
- How to describe the data?
  - Sample/Population
  - Central Tendency (Mean, Median or Mode)

- Are there Outliers?
- Are there any Null/Missing values?
- What does The Visualization tell us?
- Is the Data normally distributed?

Findings

The data we will work on is

- Structured data in a CSV file
- The data is Unlabeled
- A Small Dataset – it has 200 instances
- A Small Multidimensional Dataset – it has 5 features
- There are 2 types of Features
  - 3 Quantitative Features (Interval)
    - Age
    - Annual Income
    - Spending Score
  - Qualitative
    - Customer ID (numerical categories 1:199)
    - Gender (labels: Female-Male)
- The Dataset is a Sample of 200 instances, and we can describe it as follows:
  - Age
    - customers' age ranges from 18 to 70 y/o.
    - mean and median are slightly close that means there is No Outliers.
  - Annual Income
    - Customers` annual income ranges from 15K$ to 137K$
    - Customers who are 49 y/o and more earns more than 75% of other customers and their spending score is 73 or higher.
  - Spending Score
    - the average spending score is 50
  - In general, about 50% of Customers are
    - more than 36 y/o
    - earn more than 61.5K$ per year
    - and have Spending Score more than 50

- There are Outliers in the Annual Salary column (137)
  - After investigation these Outliers do not affect our data for many reasons
    - Mean and median are slightly close
    - They had no effect on the model performance
    - It is reliable, in this city some employees earn more than 137K$ per year

- There are no Null/ missing values
- Data Visualization
  - Figure 1 {3 Box plots}
    - Age
      - IQR from 28 to 48
      - 50% of our Customers ranges from 28 to 48 y/o

    - Annual Income (k$)
      - IQR from 41 to 77
      - 50% of our customers` annual income ranges from 41 K$ to 77 K$
      - Annual Income (k$) has an outlier (137)

    - Spending Score (1-100)
      - IQR from 34 to 72
      - 50% of our Customers have Spending Score from 34 to 72 points
  - Figure 2 {Box plot}
    - Annual Income (k$)
      - 75% of our customers earns less than 78 K$ per year
  - Figure 3,4,5 {3 Histograms}
    - Age
      - Most of our customers are less than 50 y/o
    - Annual Income (k$)
      - Most of our customers earns less than 80 K$ per year
    - Spending Score (1-100)
      - Spending score looks normally distributed
  - Figure 6 (Scatter plot)
    - Annual Income and Spending Score
      - There are 5 clusters
  - Figure 7,8 (Histogram)
    - Female Spending Score
      - Spending Score for female customers is normally distributed
    - Male Spending Score
      - Spending Score for female customers looks normally distributed

- After applying Shapiro Normality test the results was
  - Female Spending Score is normally distributed
    - p-value is greater than alpha .05 so it is normally distributed
  - Male Spending Score is not normally distributed
    - p-value is less than alpha .05 so it is not normally distributed
  - Spending Score is not normally distributed
    - p-value is less than alpha .05 so it is not normally distributed

Note:

- Further investigations and Hypothesis tests need to be done to make our inferences.

- The Gender Column needs to be encoded.

Hypothesis Testing

There are many questions need to be answered to make our inference.

Hypothesis test 1:

Question 1: Does the gender of customers affect Spending Score?

H0: the gender does not affect Spending Score
H1: the gender affects Spending Score

Alpha = .05

Used Tests:

- one-way ANOVA
  - Rules
    - The data should be normally distributed
      - Our data is not normally distributed so we can not trust the result of the test, but we will try it.
- Kruskal-Wallis
  - Has no rules
    - It is a nonparametric test so we can use it.

Method:

1. Divide the dataset into two samples, So
    - The are two independent categories: Male – Female
    - And one dependent variable: Spending score

2. Apply one-way ANOVA test
3. Apply Kruskal-Wallis test

Results Analysis:

one-way ANOVA and Kruskal-Wallis tests have p-value are greater than .05

p-value > alpha

0.57 > .05 & 0.41 > .05

Decision:

we will **ACCEPT** the Null Hypothesis

Conclusion:

Being a female customer or a male customer does not affect spending score.

Hypothesis test 2:

Question 2: Is the average spending score for male customers differ from female customers?

H0: the mean spending score for male customers equals the mean of female customer.

H1: the mean spending score for male customers does not equal the mean of female customers.

alpha= .05

Used Tests:

- Welch's t-test (Independent Samples t-test)
  - Rules
    - Two Independent samples
    - The data must be continuous
    - the samples do not have to hold the same number of instances.

Method:

1. Choose 88 random instances of Female Spending Score (not mandatory)
2. The two samples have the same number of instances
3. Apply Welch's t-test

Results Analysis:

Welch's t-test has p-value greater than .05

p-value > alpha

0.69 > .05

Decision:

we will **ACCEPT** the Null Hypothesis

Conclusion:

The average spending score for male and female customers is the same.

Hypothesis test 3:

Question 3: Is the Annual Income a good variable to predict the Spending Score ?

H0: The coefficients are zero (no relation between the two variables)
H1: The coefficients are not zero (there is a relation between the two variables)

alpha= .05

Used Tests:

- ► Regression Analysis
  - ▫ Rules
    - ▪ One Independent variable
    - ▪ One Dependent variable
    - ▪ The data must be continuous

Method:

1. Choose Annual Income as the Independent Variable
2. Choose Spending Score as the Dependent Variable
3. Apply Regression Analysis

Results Analysis:

Regression Analysis has p-value less than .05

p-value < alpha

0.00 < .05

Decision:

we will **REJECT** the Null Hypothesis

Conclusion:

The Annual Income is an important variable to predict the Spending Score, so it will be fed to the model.

Hypothesis test 4:

Question 3: Is the Age a good variable to predict the Spending Score ?

H0: The coefficients are zero (no relation between the two variables)
H1: The coefficients are not zero (there is a relation between the two variables)

alpha= .05

Used Tests:

- Regression Analysis
  - Rules
    - One Independent variable
    - One Dependent variable
    - The data must be continuous

Method:

4. Choose Age as the Independent Variable
5. Choose Spending Score as the Dependent Variable
6. Apply Regression Analysis

Results Analysis:

Regression Analysis has p-value less than .05

p-value < alpha

0.00 < .05

Decision:

we will **REJECT** the Null Hypothesis

Conclusion:

The Age is an important variable to predict the Spending Score, so it will be fed to the model.

Note:

- Now we are confident about the inference we made about our data.
- The next step will be preparing our data then we will build the model.

Data preparation

There is a categorical variable in the data (Gender) and its datatype is *object* so it needs to be encoded to numerical representation as the model accept only numerical values.

Techniques used:

- ► Label Encoder
- ► One Hot Encoder

We used both to be able to retrieve the actual values Male-Female.

Note:

- ► We used both Encoders to be able to retrieve the actual values Male-Female.
- ► The data is ready to be feed to the model
- ► The next step will be choosing and building the Machine Learning model.

Machine Learning Modelling

The data is ready and now we can start building the model but there are some steps involved to choose the best algorithm.

1-Choosing the algorithm

From EDA and Visualization, we conclude that

- It is an Unsupervised Learning task (Unlabeled data)
- It is a batch learning task (small dataset)
- It is instance based (each instance will be compared with other instances)
- The data is grouped in 5 regions

So,
The best choice for this problem will be a K-Means Clustering algorithm.

2-Validation:

After choosing the K-Means Clustering, we need to choose the best number of clusters. By sight it is 5 but we need to make sure it is the best choice.
So, the question now is, how many clusters does K-Means need to perform well?

Used technique

- ► Elbow method (Figure 9)
    - □ choose the number with a low inertia

Results:

5 is the number with the low inertia (located on the elbow)

So,
The number of clusters = 5 (K=5) will perform well.

Building K-Means

Used Libraries:

- Sciekit-Learn
- Pipeline
- Matplotlib

Method:

1. Dividing the dataset into two sets
   - Training set
   - Test set
2. Creating a pipeline with K-Means
   - n_clusters=5
3. Train the K-Means with the training set
   - 160 instances
4. Testing the K-Means with the test set
   - 40 instances
5. Visualizing the result

Results:

Figure 10 (scatter plot)

The algorithm performed well and there are 5 well separated clusters.

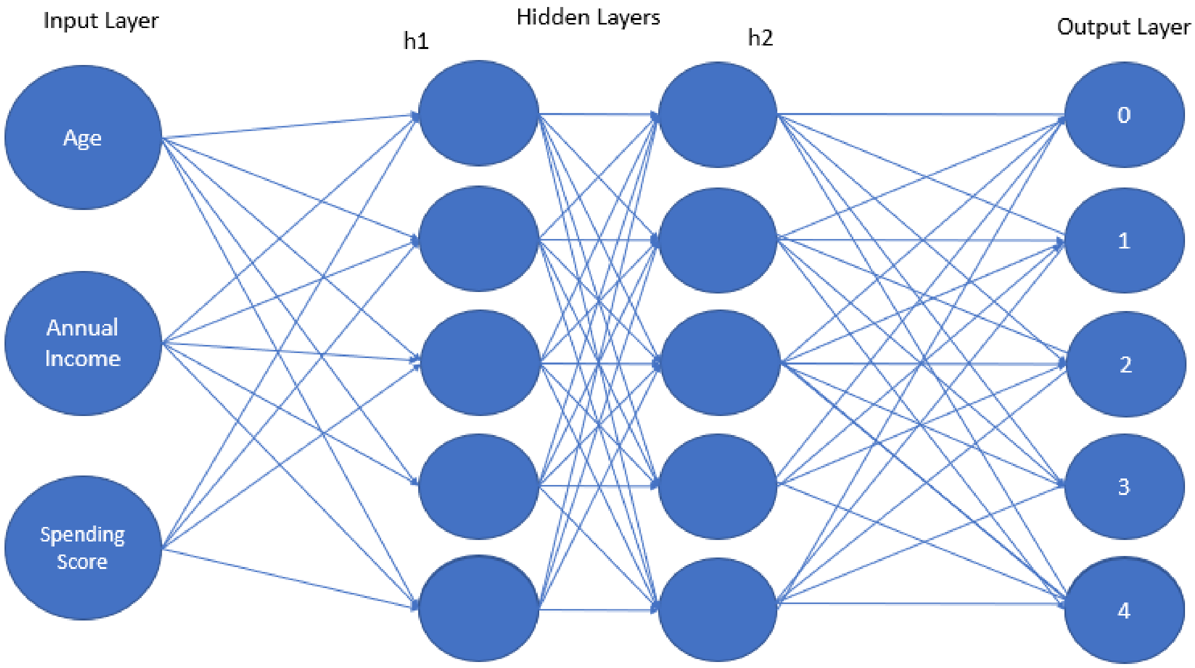
Data Labeling

The K-means will give each instance a label from 0 to 4, so we can label the data easily and use it later as a labeled dataset to be fed to a Deep Learning model.

Deep Learning Architecture

The data is labeled and can be used in a Supervised Learning model such as Artificial Neural Networks. So, we will start building the model.

Artificial Neural Networks Architecture



Used Libraries:

- Keras

Method:

1. Dividing the dataset into two sets
   - Training set
   - Test set
2. Creating an instance of the model
   - Sequential()
3. Fine tuning the model
   - Activation functions
     - relu
     - sigmoid
   - Loss Function
     - categorical_crossentropy
   - Metrics
     - Accuracy
   - Optimizer
     - RMSProp
       - Learning rate = .02

Validation

- Using a Test Set of 40 instances the model was tested, and the result was satisfying the model achieved high Accuracy and the Loss was closer to 0.
- After predicting the labels, it was the same as the real labels.

Results:

```
Accuracy = 94.99
Loss = 0.48
```

Saving the model

The model was saved as **pickle** file, to be used in deployment.