A picture containing drawing

Description automatically generated

**Data Scientist – Market Place (Analysis Document)**

1. **Introduction**

This paper was prepared for Trivago Case Study. In this case study, it was requested to work with data sets taken in two different time periods. One belongs to February, th eother one belongs to May 2019.

Both datasets had very similar qualifications. When the dataset was examined in detail, it was noticed that it covers informations such as visits, clicks, bookings and revenue from different platforms daily.

1. **Datasets**

The datasets covering both the February and May had similar dimensions.

A screenshot of a cell phone

Description automatically generated

In generally, data can be classified clear. There was no structural defect in the data or a deficiency that would prevent general inferences. There were eight columns. In addition, I created some other columns like “outgoings” that is the difference between “booking\_amounts” and “revenue”. I chose to split the columns qualitatively and quantitatively.

A close up of a logo

Description automatically generated

Finally after several data preprocessing steps, the number of columns I have reached has increased considerably.

A picture containing knife

Description automatically generated

In preprocess step, I tried to seperate columns into different group. If you want to create clusters and make groups with the same inputs, you should divide the columns into smaller groups with label encoder logic, give them weights and if necessary, redesign the columns according to these results.

I prefered to determine bins clusters for “quantative\_columns”. While determining bins, different algorithms such as *([ "auto", "fd", "doane", "scott", "stone", "rice", "sturges", "sqrt"])* were used for bins methods

For “qualitative\_columns”, “***get\_dummies***” method from ***pandas*** library was used.

1. **Clustering**

After data preprocessing steps, I tried to cluster the February and May data separately. In this way, I had the opportunity to observe the breaking of features such as platforms or whether groups were selected accurattely or not.

For clstering, I created pipeline. First, one HDBSCAN algorithm works. Then, if the number of non-clustering objects is above a limit, another HDBSCAN is working again with different parameters. Finally, if the objects that cannot be clustered despite the results of the second HDBSCAN are still above another limit, all the data is clustered with the K-MEANS algorithm.

For feb dataset, my pipeline parameters are listed below:

HDBSCAN 1: 5 Clusters

A screenshot of a cell phone

Description automatically generated

The average silhouette\_score is : 0.6021047990800183

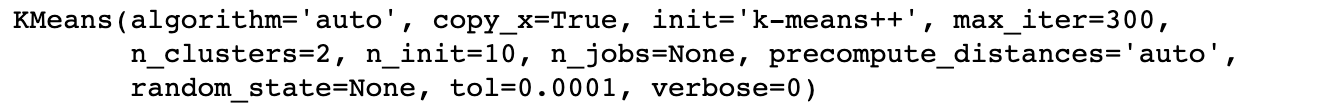
HDBSCAN 2: 2 Clusters

A screenshot of a cell phone

Description automatically generated

The average silhouette\_score is : 0.677470969697065

K-MEANS 3: 2 Clusters



For n\_clusters = 2 The average silhouette\_score is : 0.5951320111597637

Totally:

A screenshot of a social media post

Description automatically generated

For may dataset, my pipeline parameters are listed below:

HDBSCAN 1: 4 Clusters

A screenshot of a cell phone

Description automatically generated

The average silhouette\_score is : 0.6021047990800183

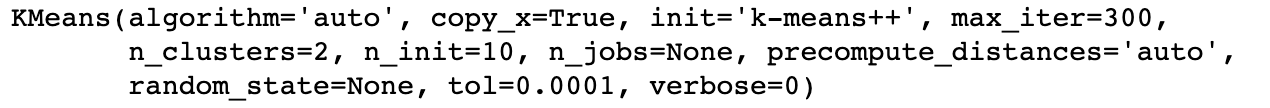
HDBSCAN 2: 2 Clusters

A screenshot of a cell phone

Description automatically generated

The average silhouette\_score is : 0.6152399045529573

K-MEANS 3: 2 Cluster



For n\_clusters = 2 The average silhouette\_score is : 0.6816187353551991

Totally:

A screenshot of a social media post

Description automatically generated

1. **Analysis**
2. Most Valuable Platform: When we analyze the data (both feb and may data), we can claim that the most valuable platform is the “US”. “Revenue” feature is not only unique parameter but it was very effective one.

For feb data,

A screenshot of a cell phone

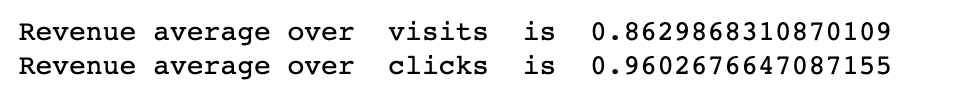
Description automatically generated

For may data,

A picture containing bird

Description automatically generated

For total data, revenue values had highly correlationship with visits and clisks



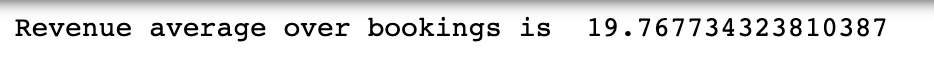
A picture containing object, comb

Description automatically generated

A picture containing object, comb

Description automatically generated

In addition to this, averagelly, company earned 20 unit Money per every booking process.



A close up of a mans face

Description automatically generated

1. Global and platformaverage revenue per clicks: In both datasets, “US” is the most important platform in revenue point of view.

For whole data set, average revenue per click in global is nearly 0.96. If we are based on platform data, the first three platform are listed as follows; US, AU and UK.

A screenshot of a social media post

Description automatically generated

In Feb;

A screenshot of a cell phone

Description automatically generated

In May;

A screenshot of a cell phone

Description automatically generated

1. Trends in data: Acording to me, it makes sense to identify trends on revenue column. In this point of view, calculating correlation between columns provides realy valid solutions.

First of all, we can easily see relation between columns in below graph but there are lots of relational plots in notebook files.

A picture containing object, comb, sitting

Description automatically generated

Secondly the correlation between columns;In both month values, correlation is higher than 0.95. This shows that features in dataset were chosen in right logic.

In Feb,

A screenshot of a cell phone

Description automatically generatedA picture containing table, bird

Description automatically generated

In May,

A screenshot of a cell phone

Description automatically generated A picture containing table

Description automatically generated

As a second point of view, it may be useful to examine the data in time diffraction. In may and february dataset, same trends based on time diffraction cans be seen. In both of them, weekends are high peaks and middle part pf week days are also low peaks. Sunday is the high revenue based day in both ones. Starurday is following it with a low diff. Thursay and Wednesday are low peak generally. Moreover in May, more revebue and booking was executed.

A picture containing bird

Description automatically generated

1. Relation between visits-bookings and clicks-bookings: If it is necessary to be honest, the ratio between clicks and bookings or visits and bookings are not high. Bu the help of accurate campaigns and higher discounts, these ratios can be higher. But until now, the ratios are very similar, around 4 percent.

A screenshot of a cell phone

Description automatically generated

A close up of a white wall

Description automatically generated

A picture containing object, large

Description automatically generated

I think clicks and visits have highly correlationship and they effects to each other direcly.

CLICKS average over VISITS is 0.8986940441744298

A picture containing object, measure, pencil, comb

Description automatically generated

1. **A/B Test Analysis**

In general, the data showed a stable character. As a result, my studies to examine the distribution of the data, gave accurate and rapid results. Density-based cluster models could aggregate a very high percentage of data.

We couldn’t say anything about selection of the months but we cab say that the differences between February and May could be negligible.

The platform distribution of the data was fairly even. This allowed the data to be examined as objectively.

However, we can not say the same for test and control groups. In the clustering studies, they belonged to very different clusters. Just as this has reduced the success of the model, it may cause other problems. For example, it is very difficult to make prediction about the test group by looking at the control group we have.

A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generated

In addtion to these, it was observed that totally only 2 clusters with test group data were formed. A total of 85 transactions were seen in these two clusters.

The number of clusters with only control group data was 8. There were a total of 226 line data in these eight clusters. This number is particularly high. It corresponded to 63 percent of almost all control data.

A close up of text on a white background

Description automatically generated

Again, in generally, evenry numeric columns have the similar ratio between control and test groups. The lower the rate, the more consistent the results.

A close up of text on a white background

Description automatically generated

I think that the data should be chosen with a different logic when it is divided into control and test groups.

One more suggestion is that if we want to analyse and predict more accurate, we need more row of data and more columns that explain transactions.

Prepared by

Eser İnan Arslan