T-DAT-901
RECOMMENDER

STUDENTS: Chase Lawrence, Clément LeCroart, Sylvain Raya
30 January 2022

Overview

The aim of this project is to analyze a dataset of client purchase history and provide insights into the company's client population. Our end goal would be to leverage Big Data techniques and Artificial Intelligence to provide the company with prototypes for customer segmentation model and a product recommender model.

Data Analysis

The data provided came in a single batch which appeared to have been a database export. The file was in a csv format and was relatively heavy at approximately 730 MB with over 7,000,000 records.

Our first challenge came from the sheer size of the file. Data analysis initially was drastically slowed due the the time it took just to load the data into memory. To resolve this we came up with two possible solutions:

- Insert the provided data into a managed database
- Store the file in a more efficient file format

The first solution seemed promising but quickly proved to be even more inefficient than reading the original file. This was due the fact that before being able to perform any statistical analysis the program would need to query the database to recover this information and then load it into a dataframe which increased the program overhead.

The next and retained solution was to save the data into a more file efficient format (HDF5) after having converted the different columns into more space efficient types. This allowed us to reduce our file size from 730MB to 180MB!

With the data in a more accessible format now we were now poised to perform an initial data analysis. We were able to identify 1484 products, 900,000 clients and over 2,000, 000 tickets.

The only anomaly that we were able to identify was that the product prices were not uniform. For example two products that are exactly the same would have different not only across the same month but the same ticket as well.

Segmentation

Data preprocessing

To get the data ready for segmentation we would need to take the data from being a ticket export to a client summary dataset. To kick this step off we grouped the ticket information by the client ID and for each group performed aggregate calculations. This provided us with an abundance of statistical information per client. For each client we were able to calculate their mean number of monthly visits, the standard variation in their ticket totals, the variance in their monthly spending and much more.

Once this was done we had on our hands a very high-dimensional data set. To reduce the dimensions we used Pricipal Compenent Analysis (PCA) to identify the number of components to retain at least 95% of the variance in our dataset. From our experiment we determined that the two most influential features of the dataset were the variance in monthly spending and the variance in ticket totals.

Model

For this model we chose to use the K-Means clustering algorithm. To choose the optimal number of clusters we ran the algorithm with cluster values from 2 to 10. We noticed that the inertia of the model drastically slowed down after the k was set to 7, therefore we made the assumption that the optimal number of clusters would have been 7.

Results

After training and fitting the model with the number of clusters we identified the centroids of each cluster. The dimension reduction was also reversed on their components to get back an approximation of the original feature values. We then used this to give us a clearer vision of what each cluster represented.

Recommender

Data preprocessing

To build our recommender we would first need to create datasets containing the unique values of the data we intend to embed for both the clients and products. This was a relatively simple task and required little effort.

Model

For our model we chose an item-based collaborative filter style recommender. The model was built using the two tower model architecture. This neural network uses two sub-models to learn candidates and queries separately. The recommendation score would be the dot product of the output of these two models.

Results

When training was completed for the model we attained a top-5 categorical accuracy metric of 0.33 which means that 33% of the time the true positive was in the top 5 products recommended.