|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| TECHNICAL DOCUMENTATION  Authors: Kantai Melau, Adna, Calum Knight, Nnamdi Okoroji |
| |  |  |  | | --- | --- | --- | |  | 11/14/22 | Data Analytics | |

Table of Contents

[**Data Sources** 2](#_Toc119351880)

[**PostgreSQL Data-Base** 3](#_Toc119351881)

[**ETL Process** 4](#_Toc119351882)

[**Exploratory Data Analysis (EDA)** 5](#_Toc119351883)

[**Customer Segmentation** 6](#_Toc119351884)

[**Using K-Means** 6](#_Toc119351885)

[**Hierarchical Clustering** 7](#_Toc119351886)

[**Tableau** 8](#_Toc119351887)

[Analysis of visualizations from each cluster. 8](#_Toc119351888)

# **Data Sources**

There was one data set used in creating a database for this project. The data set was obtained from Maven Analytics and had 38 columns and 7043 rows. It contains information about the Churn Data for a Telecommunications company that provides phone and internet services to 7043 customers in California and includes details about customer demographics, location, services and current status.

# **PostgreSQL Data-Base**

We made a connection with PostgreSQL to act as our database. We selected to use PostgreSQL as it is a relational database. It provides an instinctive way to represent data which allows easy modification of the relationship between tables (DIMS and FACTS). The data used was structured data as it provides for faster indexing, easy updating and deleting records, complies with a data model and has a standardized format.

We created a database and made the connection on the jupyter notebook. After the connection, we imported the CSV onto the jupyter notebook and read it in pandas using “pd.read\_csv”. We then pushed the raw data to the sql database. This created a table with the required columns and data types. We then read into the data from the SQL in jupyter notebook and worked on the data that we had from the notebook.

# **ETL Process**

The “Extract, Transform and Load” process started by importing all the necessary libraries that would be used in the project. We used “Pandas”, “sqlalchemy”, “sklearn” and “matplotlib”.

After importing, the transformation process began.

We began by dropping any duplicates that existed in the data frame using “drop\_duplicates()”. We then dropped the null values in the data frame. This was in columns that were all null and not part of them using .dropna(how=”all”).

We then examined the existing columns to determine the ones that will be dropped. We dropped all the columns that we did not find necessary for the clustering process.

We then transformed the data to convert all the categorical values in the dataframe to numeric values. This was done to enable the data to be subjected to unsupervised machine learning models. After this, the data frame was ready for the EDA process to get to understand more about the data before modelling.

# **Exploratory Data Analysis (EDA)**

EDA was performed on the data to examine existing trends in the data. This was done in a different jupyter notebook.

The process began by importing the dependencies that were required for the process. These included pandas, sqalchemy, matplotlib and seaborn. We then imported the data from the database where it had been stored.

We began by examining the summary statistics in the data. From this, we were able to get the mean, standard deviation, minimum and maximum values from the different columns that are present in the data.

We plotted the age distribution of the various customers ages. This showed their various ages and how distributed they are.

The distribution of customer tenure was also plotted. This was to understand how long customers tend to stay while using the services offered.

R=The distribution of the various sexes was also plotted. From this, it was evident that there were more males than females in the data set. The people’s marital statuses were also plotted and this showed that there were more unmarried people than there were married ones.

We plotted graphs to show the most popular offers that were there and how customer uptake was. It was evident that most customers did not pick any offers. Among the offers, Offers B and E were the most popular among the customers.

We examined the types of contracts that were common among the people. Month-to-month contracts were more popular among the customers as compared with the rest. The one-year contracts were the least popular.

# **Customer Segmentation**

## **Using K-Means**

K-means was selected to find the similarity between the various items and group them into clusters. It was selected because of its simplicity to use.

Steps:

* The data was standardized using a standard scaler. This was done in order to bring all the features to a common scale without distorting the differences in the range of the values.
* We then used the Principal Component Analysis (PCA) for dimension reduction. We chose to retain 90% of the columns initially.
* We then applied the t-SNE model to further reduce the dimensions of the data. This reduced the features to just 2 features.
* We then plotted the data to see if any clusters existed visually.
* We applied K-means to the data to determine the appropriate number of clusters for the data.
* The elbow curve was plotted and 5 was found to be the representative number of clusters for the dataset.
* We then fit the models and got the predictions for the different rows of data into the various clusters.
* The clusters were appended as a column in the original dataset.
* The clusters were then visualized with the various categories used as the colour.
* For further trials, we tried the same process while using only the PCA dimensionality reduction and then the t-SNE dimensionality reduction.
* This was to also see which of the dimensionality reduction tools performs best.

## **Hierarchical Clustering**

Hierarchical clustering was done to compare the outcome with the results obtained from the K-means.

The data frame was initiated and normalized.

The method was then selected. We opted to go for “ward”. This produced better results as compared to the “complete” and the “single” method. For the 2, it was impossible to get any meaningful relationships that exist within the data and that’s why we opted for the “ward”.

We then plotted the figure to see the outcome.

# **Tableau**

A tableau dashboard was created to visualize relationships that exist between the clusters.

This was instrumental in noting some of the features in the dataset.

From this we were able to create 3 dashboards demonstrating various instances that exist in the data.

## Analysis of visualizations from each cluster.

**Demographics**

Ratio of male: female is 50:50, and this is observed between all clusters. However, the split between married/not married is also ~50:50, but this varies a lot between the clusters.

Cluster 0 & 3 - almost all married

Cluster 2 - mostly not married

Clusters 1& 3 - fewer customers that are 65+

With regards to locations, we can see a lot of customers are based in LA and the surrounding areas, but also all over California.

**Revenue**

The first graph shows the revenue generated from each cluster & second graph shows the number of customers in each cluster

Cluster 0 has the highest total revenue, but it’s the second smallest size cluster. 1158 customers producing 8m in total revenue.

Cluster 0 is closely followed by cluster 4 in terms of revenue amount brought in.

Cluster 2 has the most customers in it, but it brings in the second smallest amount of revenue.

**CONTD**

To continue looking at these clusters, here we can see how long the customers have been with the company, the contract type they are on and their churn rate.

Cluster 0 - most customers are in 2-year contracts, there is very little churn, and most of the customers had been with the company for more than 65 months).

Cluster 1 - most customers in this cluster are new to the company, and a lot are staying

Cluster 2 - highest churn rate, mostly new to the company and on month-month contracts

Cluster 3 - lowest churn rate, a lot have been with the company a long time – 2-year contract

**Usage**

Lowest revenue is brought in by cluster 1, which also has the lowest number of extra data charges & long distance charges, implying perhaps most revenue comes from extra charges.

However, cluster 0 (highest revenue), has the highest long-distance charges and extra data charges, despite most customers being on unlimited data plans (potential for offering better plans for these customers?).

The second highest cluster (4) also shows lots of extra data and long-distance charges, despite being on unlimited plans.

Shows most of the revenue for the Telecom company is coming from extra charges.

**\*\*\*\*\*\*\*\*\*\*\*\***

So the customer segments produced with the algorithms can prove very useful for the Telecom company to look at reasons for high customer churn, create better offers for customers, and know exactly which customers to target with these offers and thereby increase their revenue.