FINAL PROJECT

Demo 3

Leonardo Cortés Marcelo Atencio Federico López Andrés Ruiz

February 2024

Abstract

The following document presents the report of the third and final stage of the final project to be presented at the Data Science Boot Camp of 'Soy Henry'. This stage includes the selection and development of the Machine Learning model, the final development of a graphical analysis in a dashboard, as well as the implementation of the proposed KPIs and finally, the implementation of both the model and the dashboard into a single deliverable product to the final client.

Contents

1 Feature Engineering

In this stage, we proceed to develop a business recommendation model that, when entering a category and/or business state, can suggest to the investor the highest-rated business with the

most relevant words among users, in addition to 5 other similar business recommendations.

1.1 Sentiment Analysis

To carry out sentiment analysis, we use the function "sentiment_analysis" located in our resources. This function evaluates each of the user comments and assigns a rating as follows:

- Value 0 for negative comments.
- Value 1 for neutral comments.
- Value 2 for positive comments.

It is of our special interest, and especially for the investor, to reduce the information to those businesses whose score is greater than 4 and with strictly positive comments (2). This narrowed information is processed in the next stage and stored in a dataframe named "df_filtered".

1.2 Minimum Distance

To the reduced information, we proceed to divide it by states and store it in individual dataframes. Then, we execute the function "distances", which calculates the distance for each non-tourist business with respect to the nearest tourist business. The resulting information is stored in a new dataframe. Finally, we unify all the processed information into a single table.

Since the number of records is excessive to perform the Machine Learning model, we reduce the data to a sample that allows having a minimum number of businesses for each state, category, non-null minimum distance, and number of nearby businesses less than 3. This sample is saved as "model.parquet".

2 Recommendation Model (ML)

In this stage, we will develop a business recommendation model. Upon entering a category and/or business state, we will suggest to the investor the highest-rated business along with the most relevant words among users. Additionally, we will provide 5 other similar business recommendations. To perform this recommendation, we have chosen cosine similarity as a comparative model between business features.

We will define 3 functions:

2.1 Function `find_businesses`

This function takes a state and optionally a business category as arguments and returns the three highest-ranked businesses with the best location that are within the specified characteristics.

For example, if we enter:

- State: California
- Category: Pub

We will get the following information:

Business Name	State	Category
Black Diamond Tavern	California	Pub
Woodhouse Blending & Brewing	California	Pub
MCG Service LLC	California	Pub

If we enter:

- State: New York
- Category: None

We will get the following information:

Business Name	State	Category
Blue Mountain Reservation Trail Lodge	New York	Hostel
Rosie Dunn's Victorian Pub	New York	Pub
Sushi Nonaka	New York	Restaurant

2.2 Function `get_recommendations`

This function utilizes cosine similarity and takes the name of a business (generated from the `get_business_info` function) as an argument and returns 5 similar businesses. It considers the variables of state, category, and the nearest non-tourist business category.

2.3 Function `get_business_info`

This function takes the name of a business as an argument and generates a word cloud based on user comments and basic business information (Name, Address, City, State, category, Nearest tourist spot, and the distance to it).

For example, if we enter:

- Black Diamond Tavern"

We will get the following information:

- Business Name: *Black Diamond Tavern*
- Address: *Black Diamond Tavern, 42172 Moonridge Way, Big Bear Lake, CA 92315*
- City: *Big Bear Lake*
- Distance: *0.839082*
- Nearest Tourist Business: *Trails End*
- Tourist Category: *Park*

- Ranking: *4.100000*

Additionally, a corresponding word cloud will be generated.

Figure 1: Wordcloud.

Moreover, it provides recommendations given by the `get_recommendations` function.

Business Name	State	Category
Tiburon Tavern	California	Restaurant
Huntsman Tavern	Nevada	Pub
HOB Tavern	New Jersey	Pub

3 Entity-Relationship Model

Once the tables for the relational model have been defined, we proceed to visualize both the final Entity-Relationship diagram and the Primary Keys (PK) and Foreign Keys (FK) for linkage.

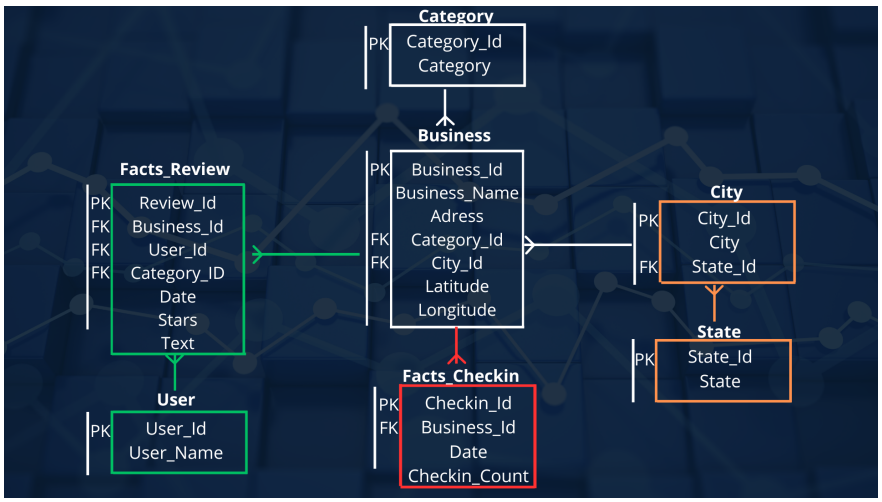

Figure 2: ERD.

Where we can see the details:

- **Category Table**
  - `Category_Id`: Integer (Primary Key)
  - `Category`: String
- **City Table**
  - `City_Id`: Integer (Primary Key)
  - `City`: String
  - `State_Id`: Integer (Foreign Key)
- **State Table**
  - `State_Id`: Integer (Primary Key)
  - `State`: String
- **User Table**
  - `User_Id`: Object (Primary Key)
  - `User_Name`: Object
- **Business Table**
  - `Business_Id`: Integer (Primary Key)
  - `Business_Name`: Object
  - `Address`: Object
  - `Category_Id`: Integer (Foreign Key)
  - `City_Id`: Integer (Foreign Key)
  - `Latitude`: Integer
  - `Longitude`: Integer
- **Review Table**
  - `Review_Id`: Integer (Primary Key)
  - `Business_Id`: Integer (Foreign Key)
  - `Date`: Date
  - `User_Id`: Object (Foreign Key)
  - `Ranking`: Integer
  - `Stars`: Integer
  - `Text`: Object
  - `Sentiment_Analysis`: Integer

- **Checkin Table**
  - `Checkin_Id`: Integer (Primary Key)
  - `Date`: Date
  - `Business_Id`: Integer (Foreign Key)
  - `Checkin_Count`: Integer

4 Dashboard and KPIs

4.1 Global Analysis

For data visualization, we will only provide a dashboard where a global analysis of the information can be seen: business locations on the map, number of businesses by category, TOP 10 cities with the most businesses, and states with the most businesses.

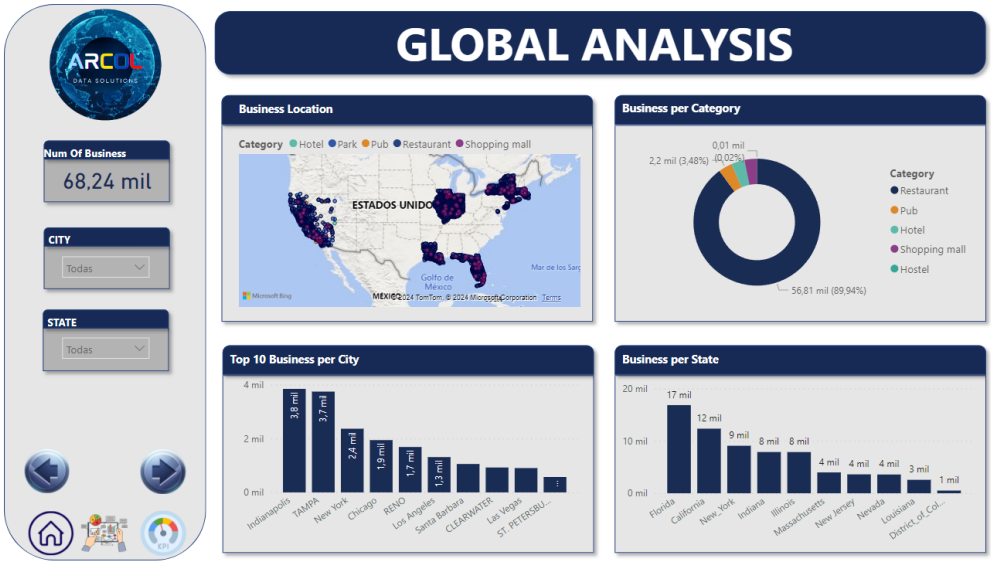

Figure 3: Global Analysis.

4.2 Key Performance Indicators (KPIs)

Regarding KPIs, we have determined 4 representative indicators for each business. These indicators are calculated based on a specific period and objectives are set in comparison with the previous period.

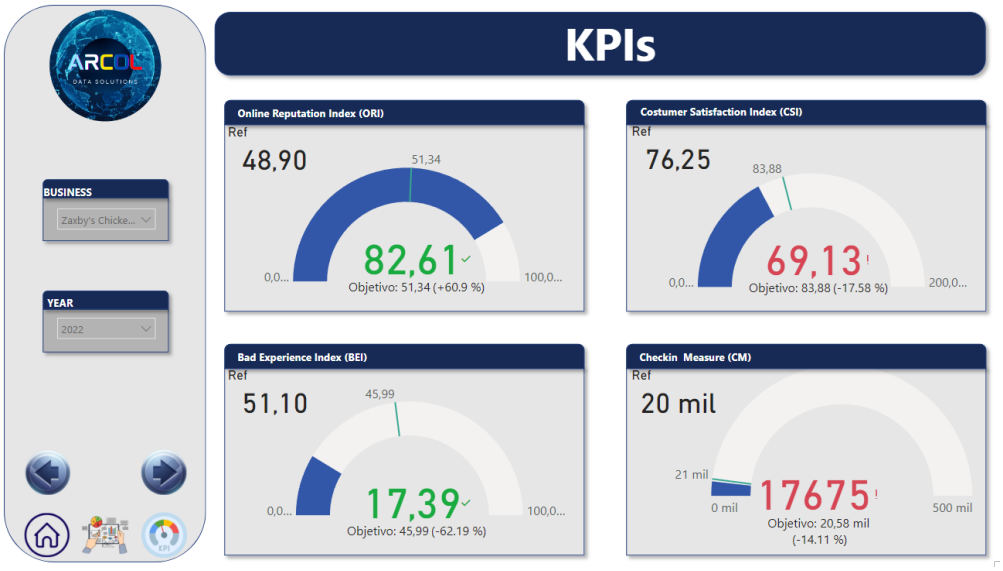

Figure 4: KPIs.

Online Reputation Index (ORI)

$$ORI = \left(\frac{\text{Number of positive reviews (SA} = 2)}{\text{Total reviews}} \right) \times 100$$

Objective: Calculated annually, increase online reputation by 5% compared to the previous period.

Customer Satisfaction Index (CSI)

$$CSI = \left(\frac{\text{Average user rating}}{\text{Maximum possible rating}} \right) \times 100$$

Objective: Calculated annually, increase customer satisfaction by 10% compared to the previous period.

Bad User Experience Index (BEI)

$$BEI = \left(\frac{\text{Number of neutral and negative reviews (SA ¡ 2)}}{\text{Total reviews}} \right) \times 100$$

Objective: Calculated annually, decrease bad user experience by 10% compared to the previous period.

Check-In Records (CM)

$$CM = \text{Number of Check-Ins recorded}$$

Objective: Calculated annually, increase the number of Check-Ins recorded in the business by 5% compared to the previous period.

The investor can select a specific business and view the indicators for the registered years, allowing them to evaluate the performance of the business over time.

5 Final Product

Finally, after conducting various tests with the Machine Learning model and the data used to feed the dashboard, a "main.py" document is created containing the necessary functions to execute both the ML model and the visualization of the dashboard with the respective analysis and KPIs, all in a single deliverable final product, deploying the final result of our project through a "web app" coded with the "streamlit" tool. Additionally, a presentation page is created within the same app where the user can view information about the ARCOL Data Solutions team.

Figure 5: ML Model.

Figure 6: Dashboard.

6 Conclusion

In conclusion, ARCOL Data Solutions presents a comprehensive and strategic approach to the development of an investor recommendation model for the tourism sector in the United States. By leveraging the power of data analysis and machine learning techniques, we aim to provide our clients with valuable insights and actionable recommendations for making informed investment decisions. Our team of experienced professionals is committed to delivering high-quality results and exceeding client expectations. We look forward to the opportunity to collaborate with you and contribute to your success in the dynamic and competitive market landscape.