

FINAL PROJECT

Demo1

Leonardo Cortés Marcelo Atencio Federico López Andrés Ruiz

February 2024

Abstract

The following document presents the report of the first stage of the final project to be presented at the Data Science Boot Camp by 'Soy Henry'. This project proposes a work scheme from the ingestion of raw data provided by the client to the development of a deliverable final product.

Contents

1	Introduction	2
2	Understanding the Current Situation	2
3	Work Team	3
4	Objectives	4
5	Work Methodology	5
6	Technology Stack	5
7	Scope	7
8	Key Performance Indicators (KPI)	8
9	GitHub Repository	9
10	Project Planning	9
11	Preliminary Work	9
11.1	Global Pre ETL (Extract, Transform and Load)	9
11.2	Pre EDA (Exploratory Data Analysis)	10

1 Introduction

ARCOL Data Solutions is a company formed by a team of four passionate experts in the field of 'Data' motivated by constant curiosity and dedication to analysis. Currently, ARCOL Data Solutions takes on the challenge of preparing, analyzing, and modeling a quantity of data provided by the client in order to deliver a comprehensive analysis along with a recommendation model that allows the user to make investment decisions in the real market of the tourism sector in the United States. Their mission is to address the questions that most concern users by transforming data into intelligence. For the current project, our main source of information is the native Datasets provided by the YELP site. These datasets contain recommendations and ratings generated by users during their visits to various places throughout the United States, ranging from restaurants to hotels. This information offers us a unique and detailed insight into users' experiences at different establishments.

In addition, we complement our arsenal of data with the information provided by GOOGLE MAPS. This source includes data such as the precise location of various places registered on the site, the ratings given by users, as well as the websites associated with these places, among other relevant aspects. This combination of data allows us to conduct exhaustive analysis and provide informed answers to the questions we pose. At ARCOL, we are committed to transforming data into practical and valuable intelligence for decision making.



Figure 1: ARCOL Data Solutions.

2 Understanding the Current Situation

1. "Failure is simply the opportunity to begin again, this time more intelligently. (Henry Ford)"
2. "The worst thing that can happen to you is not to have a gold mine, but to have the mine and not extract the gold."

The quotes by Henry Ford and the analogy about the gold mine set the tone for our project proposal, highlighting the importance of learning from failures and seizing opportunities. In today's digital environment, business success is closely linked to effective presentation and communication in digital media, as well as the ability to process consumer recommendations

and feedback. In this context, we recognize that honest user feedback, facilitated by the privacy provided by the internet, has become a valuable source of information. A company's ability to understand the customer perspective and continuously improve its processes becomes crucial. Organizations that do not engage in this analysis have a high probability of facing difficulties and failures.

This is where we identify a clear Business Opportunity: focusing on projects involving businesses prone to failure, those facing challenges or having untapped growth potential. Our proposal lies in using specialized data analysis tools to identify and understand patterns, trends, and areas for improvement. By providing our clients with accurate and relevant information, we aim to equip them with the necessary tools for making informed and strategic decisions in their businesses. In summary, we seek to turn failure into an opportunity for business growth, providing intelligence from available data.

3 Work Team

The ARCOL Data Solutions team is composed of professionals with comprehensive knowledge in the field of data, allowing them to develop and interact in any role. For the specific project of "Investor Recommendation Model", although all team members are involved in each stage of the project, each member leads a specific cycle of the project, in order to have better organization and coordination.

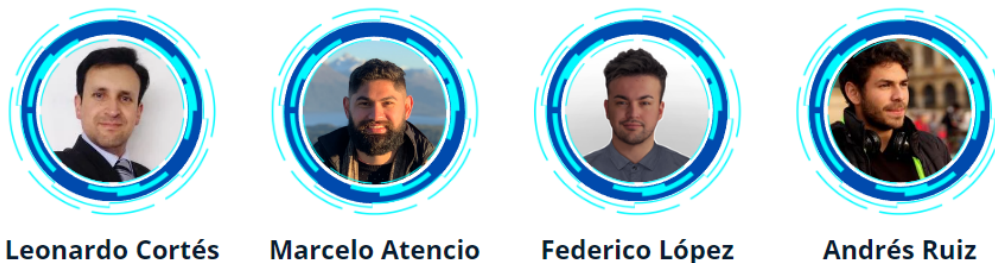


Figure 2: Team.

- **Leonardo Cortes.** *Project Manager (PM) and Data Scientist:*

Leonardo manages and coordinates the overall project, setting clear objectives, deadlines, and deliverables to ensure timely and successful project delivery to the client. He assigns resources, tools, and technologies to use to ensure that the project is completed efficiently. He is also responsible for the analysis, modeling, and interpretation of the data obtained, using advanced statistical and machine learning techniques to extract valuable insights and patterns.

- **Marcelo Atencio.** *Back-End Developer and Cloud Engineer:*

Marcelo is responsible for designing and implementing the data ingestion pipeline, ensuring that data is collected, stored, and processed efficiently and securely. He develops the necessary scripts, APIs, and database architectures to handle large volumes of data, guaranteeing data integrity and reliability. He also manages the deployment and scaling of cloud infrastructure to support the project's computational requirements.

- **Federico López.** *Front-End Developer and UI/UX Designer:*

Federico designs and develops the user interface and experience for the final product, ensuring that it is intuitive, responsive, and visually appealing. He collaborates with the client to understand their requirements and preferences, incorporating feedback to iteratively improve the design. He also implements interactive features and visualizations to enhance user engagement and understanding of the data.

- **Andrés Ruiz.** *Data Engineer and DevOps Specialist:*

Andrés is responsible for acquiring, cleaning, and preprocessing the raw data obtained from various sources, ensuring that it is consistent, complete, and formatted correctly. He develops automated workflows and pipelines to streamline the data processing and transformation process, reducing manual intervention and minimizing errors. He also implements monitoring and logging mechanisms to track the performance and quality of the data pipeline.

4 Objectives

1. **3-6 Month Business Objective:** Guide our business towards success by increasing it proportionally and adapting to the individual needs of each client. This achievement is realized through the implementation of a robust recommendation system based on constantly updated indicators.
2. **Project Description:** Create a recommendation system designed especially for new investors. This system is based on real-time updated financial indicators, allowing our clients to access precise and relevant information about the most attractive projects in the market.
3. **Objective of the Recommendation System:** Provide our clients with informed and accurate suggestions about the most attractive projects. We aim to be the reliable guide in identifying promising business opportunities, enabling investors to make decisions supported by updated and relevant data.
4. **Accuracy and Reliability:** Provide recommendations based on constantly updated data, ensuring the accuracy of the information.
5. **Informed Guidance:** Offer a system that acts as a compass for new investors, providing informed guidance tailored to the individual needs of each client.
6. **Opportunity Identification:** Facilitate the identification of promising business opportunities, optimizing the decision-making process for our clients.
7. **Success Measurement:** Measure through the percentage growth of client satisfaction, the increase in investment profitability, and the consolidation of our business as a benchmark in personalized financial advice.

5 Work Methodology





Figure 3: Work Methodology.

For the project execution, we propose the implementation of a Group Monitoring of Individual Progress methodology, with a visible responsible person for each stage. Under this methodology, we carry out both collaborative tasks and individual assignments. In both instances, comments are made on the progress achieved, and after unanimous team approval, the new progress is integrated into the corresponding repository or file. This approach promotes transparency, collaboration, and efficiency in project development.







This methodology is complemented by evenly assigned roles among the four project members. This enables knowledge exchange at each stage of the project, thus promoting versatility acquisition in the field of Data Science.

The assignment of responsibilities for each task and phase involves assigning specific roles, such as overseeing task completion, reviewing content, and adequately documenting each step. These roles are clearly specified in the Gantt chart, providing an organizational structure that facilitates efficient and collaborative project execution.






6 Technology Stack

-  **Microsoft Fabric:** It is the heart of the project, as the integrated tools allow Microsoft Fabric to be present at all stages of the project development cycle, covering from the initial data ingestion, through processing, transformation, and analysis to the visualization of the resulting information and automation of the entire project. It also allows integrating Machine Learning processes to perform the necessary models to deliver the final product. The choice of this tool is based on its capacity to provide a professional and authentic experience in an environment similar to that in which companies specialized in engineering, analysis, and data science operate in the sector.
-  **Apache Spark:** Apache Spark is a powerful distributed processing engine designed to handle large volumes of data quickly and efficiently. One of Apache Spark's distinctive features is its ability to perform data processing operations in memory, making

it significantly faster than traditional data processing systems. Due to its large set of libraries that versatily facilitate data processing in Python (Pyspark) and querying structured data with SQL language (Spark SQL), Spark is also a comprehensive tool that will be used during the cycles of each stage of the project.

-  **Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows creating and sharing interactive documents containing live code, visualizations, explanatory text, and mathematical equations. The distinctive feature of Jupyter Notebook is its ability to combine rich text, such as Markdown or LaTeX, with executable code in various programming languages, including Python. It is a tool that is integrated into the Microsoft Fabric environment for processing code under the Pyspark and SparkSQL language, and therefore, it is a tool that will be used throughout the project.
-  **Visual Studio Code:** (VS Code) is a popular source code editor that, due to its intuitive interface, wide compatibility, and built-in tools, will be used for the development of Python code for the functions integrated into the project, especially during the data engineering stage and Machine Learning development.
-  **Python:** Python is a high-level, interpreted, and multipurpose programming language. Python will be used in all stages of the project due to its simplicity, versatility, and the powerful libraries it offers. In this case, Python will be used as a fundamental tool thanks to libraries such as NumPy, pandas, Matplotlib, Seaborn, Scikit-Learn, TensorFlow, tools that we will use in the first stage with data ingestion and primary data analysis. During stage 2 with the development of EDA and the respective data selection for storage in the warehouse. And in the final stage for the development of the different Machine Learning models that need to be performed.
-  **SQL:** SQL, (Structured Query Language), is a programming language designed to manage and manipulate relational databases. Being the de facto standard language for working with databases, it will be the instrument used particularly at the end of the data engineering process and in the data analysis process, when performing the respective queries and modifications to create a dimensional and related model and then load the necessary data into the tool that will be used for visualization analysis.
-  **Power BI:** Power BI is an analytics platform that allows visualizing and sharing data effectively to make informed decisions. It offers a wide range of tools for data preparation, data visualization, analysis, and collaboration in one place. For the particular project, this tool (which is implemented within the Microsoft Fabric environment) will be used during the data analysis stage, taking data already worked and structured previously during the data engineering stage, to present a visualization that allows understanding the data and thus allowing the client to make the best decisions based on them.
-  **LaTeX:** LaTeX is a technical text composition system for producing professional documents with high typographic quality, using a simple markup language that describes the logical structure of the document, such as titles, sections, lists, tables, and

equations. In this case, the present document, as well as all reports and documentary processes required in the project, will be developed with this tool and presented as a high-quality and professional PDF document.

-  **Git:** Git is a distributed version control system that will be used during all project cycles to track progress and collaborative work among the team.
-  **GitHub:** GitHub is a web-based collaborative development platform that uses the Git version control system. There we will store step by step the instances, along with the documentation and code made to develop the project, so it is a tool that will be present during each of the project stages.
-  **Scikit Learn:** Scikit Learn is a machine learning library for the Python programming language. It is used to develop and apply machine learning algorithms efficiently and easily. It is the library that we will use to work during the data processing and machine learning stage, to perform most of the proposed models in our projects, due to the wide range of supervised and unsupervised learning algorithms, as well as tools for data preprocessing.
-  **Streamlit:** Streamlit is a Python library that allows creating interactive web applications for data analysis and visualization quickly and easily. With Streamlit, you can turn Python scripts into interactive web applications. It offers a wide range of widgets for data input, interactive graphics, and visualization capabilities, making it a powerful tool for rapid prototyping and implementation of data analysis and science applications.
-  **Jira:** Jira is a software application mainly used for project management and issue tracking. In our case, Jira will be used to plan, track, and manage the tasks proposed by and for teamwork.

7 Scope

- In addition to the data provided by our client, we incorporate relevant information about real-market economic data in the sector.
- Use of technological tools included within the service provided by Microsoft Fabric (Data Lake, Data Warehouse, Data Factory, Power BI).
- Development of a dashboard for visualizing information analysis.
- Offer a recommendation system through an API.
- The analysis will include representative commercial businesses in the tourism sector within the USA, with a subsequent expansion of the segment according to the needs of our clients.

8 Key Performance Indicators (KPI)

We formulate a set of indicators designed to provide a comprehensive assessment of the current business situation. These indicators are primarily aimed at facilitating informed decision-making for our investors, aligning with the achievement of our main objective. The proposed indicators are:

1. Customer Satisfaction Index (CSI):

Formula: Let:

x = Average user ratings.

y = Maximum possible rating

$$CSI = \frac{x}{y} \times 100$$

Objective: Evaluate overall customer satisfaction compared to the previous year. A higher CSI indicates higher satisfaction, which can be attractive to investors.

2. Bad Experience Index (BEI):

Formula: Let:

x = Number of negative reviews.

y = Total number of reviews

$$BEI = \frac{x}{y} \times 100$$

Objective: Annually evaluate service quality at specific locations. Provides detailed information on performance in different geographical areas.

3. Online Reputation Index (ORI):

Formula: Let:

x = Total number of positive reviews.

y = Total number of reviews

$$ORI = \frac{x}{y} \times 100$$

Objective: Measure the online reputation of the business each year. A high proportion of positive reviews can be attractive to investors.

4. Check-in Measurement (CM):

Formula: Let:

x = Total number of Check-ins.

$y = 1.1$

$$CM = x \times y$$

Objective: Improve the number of Check-ins by 10% each year.

9 GitHub Repository

The following link provides access to the project repository on GitHub:

[Investment Recommendation System - GitHub Repository](#)

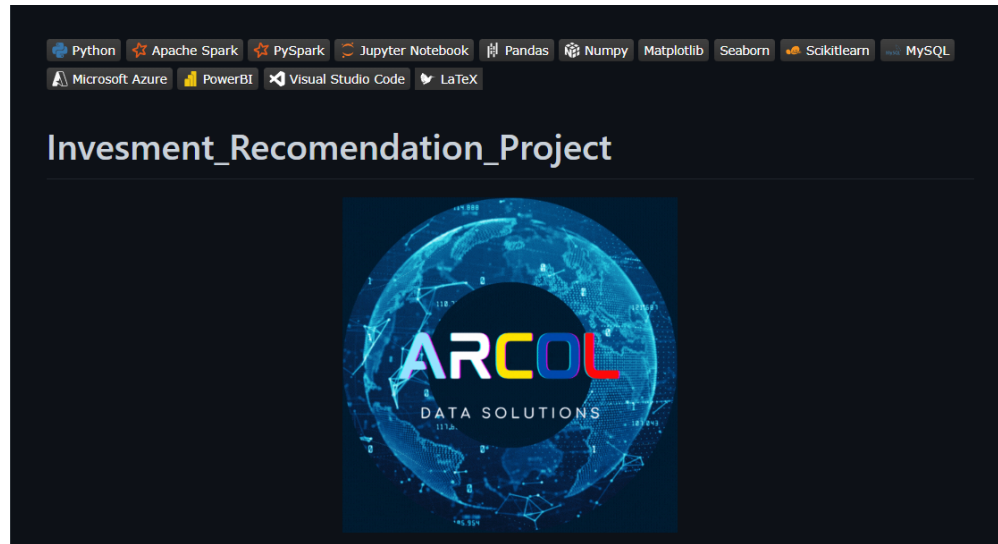


Figure 4: GitHub Repository.

10 Project Planning

Using a Gantt chart, utilizing the Task Management software 'Jira', we detail all tasks to be executed to achieve the goal, their duration, and dependencies.

11 Preliminary Work

11.1 Global Pre ETL (Extract, Transform and Load)

In the initial phase of the project, we established a Data Lake in Microsoft Fabric and directly loaded the datasets from Google Maps and Yelp from a Google Drive link. We developed automated code to read files in .json and .pickle format, converting them to .parquet for easier manipulation. We highlight the creation of the 'Utils.py' file containing essential functions, accessed from the main code.

We chose to store files in parquet format from the beginning to reduce memory consumption and streamline data manipulation in later stages. This is achieved thanks to the efficiency and performance advantages offered by the parquet format compared to others. The modularity of the code and accessibility to key functions from "utils.py" contribute to effective project organization and maintenance.

11.2 Pre EDA (Exploratory Data Analysis)

1. **Memory Usage** During this process, we conducted a comparison of memory consumption between the original files and their parquet format counterparts within the Data Lake. Initially, the native data represents an approximate consumption of 36 GB. However, after applying further processing, the parquet files significantly reduce this consumption to 10 GB, implying a nearly 70% reduction in resource usage. This efficient memory management not only optimizes storage capacity but also enhances the overall system efficiency.
2. **Preliminary Data Quality Analysis** Through an initial data review, our aim is to visualize the information provided in each file, understanding the data type, the number of records, and the presence of null values. Additionally, we seek to intuitively verify if we have the necessary information for the proposed indicators. In this process, we identified five key aspects that require special attention regarding data quality:
 - (a) *Velocity*: The constant data generation speed poses a significant challenge for timely insights and indicators, as they may become obsolete upon creation. In our case, the database encompasses an extensive set of customer reviews collected from two websites that continuously incorporate new information, minute by minute. The main challenge lies in achieving comprehensive automation of the entire process to effectively adapt to the constant expansion of the database. This automation will not only ensure agility in analyzing existing data but also facilitate efficient management of incoming information, ensuring the relevance and constant updating of the indicators and conclusions obtained. We will address this issue specifically in the ETL pipeline development.
 - (b) *Variety*: The original information is stored in files in .json and .pickle formats. Transforming these files to parquet format will provide us with the ability to optimize resource usage and improve data accessibility.
 - (c) *Volume*: The database has a sufficient number of records to carry out precise analysis. In each folder we have:
 - Metadata-sites:
 - 275,000 records
 - 0 completely null records
 - 7 columns containing null data; of which 5 have more than 20% null data
 - Reviews-states:
 - 51 files with between 300 thousand and 2.8 million records
 - 0 null records
 - 3 columns have above 35% null data; the same for all files
 - Yelp:
 - 4 files with a total of 9.1 million records.
 - 0 null records
 - 1 file with 3 columns containing a low percentage of null data

- (d) *Value:* In this context, we recognize the immense value of all the provided information to achieve the established objectives. This valuable range of data includes ratings, comments, locations, prices, business descriptions, customer information, and other key attributes. Since our top priority is to maximize the utility of this data, we anticipate the need to carry out a thorough filtering and structuring process during the ETL stage. This approach will allow us to accurately assign relevance to the information, aligning with our specific purposes.
- (e) *Veracity:* Despite the data coming from two widely recognized review websites, it is imperative to conduct a comprehensive Exploratory Data Analysis (EDA). Through this analysis, we are positioned to identify possible biases, assess consistencies, handle null values, measure the reliability of the information, and address any noise present in the data. This process enables us to gain a deep understanding of the quality and intrinsic characteristics of the data, a crucial aspect for conducting informed analysis and ensuring the reliability of the obtained results.