Figure 1: Cognifyz Technologies

Data Science Internship Report

Cognifyz Technologies

Internship Period:

1st March 2025 - 30th March 2025

Submitted By:

Ayush Kumar

B.Tech Data Science, Year 2

Vidyashilp University, Bangalore

Submitted To:

Cognifyz Technologies

Contents

Introduction

This report outlines my experience and achievements during the Data Science Internship at Cognifyz Technologies, a prominent company specializing in data science, artificial intelligence (AI), machine learning (ML), and data analytics.

The internship was structured into three distinct levels—Level 1, Level 2, and Level 3—each designed to enhance my skills in data exploration, analysis, and modeling. As per the internship requirements, I completed Level 1 and Level 2, while partially engaging with Level 3.

This comprehensive report details the objectives, methodologies, findings, and visualizations associated with each task, adhering to the guidelines provided by Cognifyz Technologies.

About Cognifyz Technologies

Cognifyz Technologies is a dynamic organization committed to delivering innovative solutions in data science and related fields.

The company provides a wide array of products and services, including AI, ML, and data analytics tools, tailored to meet the needs of micro, small, and medium enterprises (MSMEs).

Additionally, Cognifyz offers training programs to promote skill development, creating an ideal environment for interns to gain hands-on experience with cutting-edge technologies.

Internship Structure

The Data Science Internship was organized into three levels, each focusing on specific skill sets:

- **Level 1:** Emphasized foundational skills in data exploration and descriptive analysis.
- **Level 2:** Focused on advanced analysis, including trends and feature engineering.
- **Level 3:** Covered predictive modeling and customer preference analysis (partially completed).

Dataset Overview

4.1 Dataset Description

The primary dataset utilized during this internship contained real-world restaurant data, comprising 9,551 rows and 21 columns. It provided detailed information about various restaurant attributes, enabling a thorough analysis of performance and preferences.

Key features included:

1. **Restaurant ID:** Unique identifier for each restaurant.
2. **Restaurant Name:** Name of the establishment.
3. **Country Code:** Indicates the country where the restaurant is located.
4. **City:** City where the restaurant operates.
5. **Address & Locality:** Physical location details.
6. **Longitude & Latitude:** Geospatial coordinates.
7. **Cuisines:** Types of cuisines served.
8. **Average Cost for Two:** Typical cost for two people.
9. **Currency:** Currency used for transactions.
10. **Has Table Booking:** Binary indicator for table booking availability.
11. **Has Online Delivery:** Binary indicator for online delivery availability.
12. **Is Delivering Now:** Indicates if the restaurant is actively delivering.
13. **Price Range:** Categorization based on price tier (1 to 4).
14. **Aggregate Rating:** Overall customer rating on a scale of 0 to 5.
15. **Rating Color & Text:** Color codes and text labels representing rating quality.
16. **Votes:** Total number of customer votes received.

The dataset's richness offered extensive opportunities to explore restaurant performance, customer preferences, and market trends.

4.2 Summary of Work Done

During my internship at Cognifyz Technologies, I completed tasks across Level 1 and Level 2, with partial work in Level 3. Each level targeted specific data science competencies, including exploration, analysis, and visualization. Below is a summary of the work accomplished:

4.2.1 Data Cleaning & Preprocessing

- Identified and addressed missing values where necessary.
- Corrected data type inconsistencies and converted categorical variables to binary integers.
- Created new features, such as ”Name Length,” for additional insights.

4.2.2 Descriptive Analysis

- Conducted statistical analysis on key metrics like ”Aggregate Rating,” ”Votes,” and ”Average Cost for Two.”
- Explored top cuisines, cities, and high-performing restaurants.
- Visualized distribution patterns using histograms, box plots, and bar charts.

4.2.3 Geospatial Analysis

- Analyzed restaurant density in major cities (e.g., New Delhi, Gurgaon, Noida).
- Examined the relationship between city demographics and rating distributions.
- Created visualizations to highlight geographic clusters.

4.2.4 Advanced Analysis (Level 2)

- Investigated the impact of table booking and online delivery on customer ratings.
- Explored price range variations and their correlation with rating quality.
- Grouped data by key categorical variables to uncover meaningful patterns.

4.2.5 Feature Engineering

- Generated new features like ”Name Length” and binary indicators for booking/delivery.
- Enhanced dataset readiness for potential predictive modeling.

4.2.6 Findings & Insights

- Higher price range restaurants tended to receive better ratings.
- Online delivery was more prevalent in mid-tier price ranges.
- Urban cities dominated the restaurant landscape, reflecting higher competition.
- Table booking availability correlated positively with higher ratings.

This systematic approach deepened my understanding of data science principles and equipped me with practical skills in data preprocessing, exploratory data analysis (EDA), and visualization using tools like Pandas, Matplotlib, and Seaborn.

- **Total Rows:** 9,551
- **Total Columns:** 21
- **Key Features:** Restaurant Name, City, Cuisines, Price Range, Aggregate Rating.

Tasks and Methodologies

5.1 Task 1: Data Exploration and Preprocessing

Objective: Explore the dataset and handle missing values.
Methodology:

- Checked dataset structure using `df.shape`.
- Handled missing values in the ”Cuisines” column.
- Converted categorical variables to binary format.

5.2 Task 2: Descriptive Analysis

Objective: Compute statistical measures and analyze distributions.
Methodology:

- Used `df.describe()` to obtain numerical statistics.
- Analyzed top cities and cuisines using frequency counts.

5.3 Task 3: Geospatial Analysis

Objective: Visualize restaurant locations and analyze ratings.
Methodology:

- Created box plots for ratings by city.
- Analyzed distribution based on latitude and longitude.

Findings

6.1 Level 1: Data Collection

The first phase of the project focused on gathering relevant and comprehensive data for analysis. The key steps involved were:

- **Data Source Identification:** Collected attributes such as Restaurant Name, Location (Latitude & Longitude), Cuisine Types, Average Cost for Two, Aggregate Rating, Number of Votes, Online Delivery Availability, and Table Booking Availability.
- **Data Extraction Process:** Utilized Python's `requests` library to send API calls and fetch data in JSON format. Automated the process with scripts to collect data from different cities.
- **Dataset Overview:** Gathered over 15,000 restaurant records across multiple cities. The dataset was structured with rows representing individual restaurants and columns representing features.
- **Challenges Faced:**
  - *Incomplete Data:* Some records lacked key attributes (e.g., missing ratings).
  - *Rate Limiting:* The API imposed call limits, requiring strategic data fetching.
  - *Data Duplication:* Duplicate restaurant entries for chain outlets needed resolution.

6.2 Level 2: Data Cleaning and Preprocessing

Once data collection was complete, the focus shifted to ensuring data quality and preparing it for analysis. Major activities included:

- **Handling Missing Values:** Identified missing values in critical fields such as Aggregate Rating and Votes (missing in 12% of the data). Handled them by:
  - Mean/Median Imputation for numerical attributes.
  - Mode Imputation for categorical attributes.
  - Dropped rows with excessive missingness (¿50%).
- **Removing Duplicates:** Identified and removed duplicate restaurant entries using pandas functions like `.duplicated()` and `.drop_duplicates()`.

- **Outlier Detection and Treatment:** Applied Z-score and IQR methods to detect outliers in features like Average Cost for Two and Votes. Treated outliers by:
  - Winsorizing extreme values.
  - Log Transformation for skewed distributions.
- **Encoding Categorical Variables:** Converted categorical features into numeric formats:
  - Label Encoding for binary variables (e.g., online delivery).
  - One-Hot Encoding for multi-category variables (e.g., cuisines).
- **Data Normalization and Scaling:** Applied Min-Max Scaling to normalize features like Votes and Average Cost for Two, ensuring uniform scaling for distance-based models (e.g., K-means).
- **Splitting the Data:** Divided the data into a Training Set (80%) for model learning and a Testing Set (20%) for performance evaluation.
- **Key Outcomes:** Achieved a clean dataset with no missing values, scaled numerical features, and properly encoded categorical variables.

6.3 Level 3: Model Building and Analysis

At this stage, the focus shifted to leveraging the cleaned data for predictive modeling and insights. Key activities included:

- **Feature Selection:** Selected relevant features like Average Cost for Two, Price Range, Aggregate Rating, Votes, Online Delivery, and Table Booking based on correlation analysis. Removed redundant and collinear features to avoid multicollinearity.
- **Train-Test Split:** Divided the dataset into training (80%) and testing (20%) sets using stratified sampling to maintain target variable distribution.
- **Model Experimentation:**
  - *Linear Regression:* Predicted Average Cost for Two, interpreting coefficients to understand feature impacts.
  - *Decision Trees and Random Forest:* Built classification models to predict online delivery, tuning hyperparameters like `max_depth` and `n_estimators`.
  - *K-Means Clustering:* Applied to longitude and latitude to identify geographic clusters, using the Elbow Method for optimal cluster count.
- **Model Performance Metrics:**
  - *Linear Regression:* Evaluated with R-squared (0.82) and Mean Absolute Error (MAE).
  - *Random Forest:* Assessed with Accuracy, Precision, Recall, and F1-score (0.89).

  - *Clustering:* Validated with Silhouette Score (0.72), indicating well-defined clusters.

- **Insights and Interpretation:**
  - *Online Delivery Analysis:* Higher-rated restaurants with more votes were more likely to offer online delivery.
  - *Geographic Clustering:* Revealed food hubs in urban centers with high delivery activity.
  - *Cost Prediction:* Price Range and Rating significantly influenced average cost; high-rated restaurants were costlier but valued.
  - *Geographic Insights:* Dense clusters highlighted areas for market expansion.

Challenges and Learnings

7.1 Challenges

- **Handling Class Imbalance in Ratings:** Addressed skewed distributions in the Aggregate Rating feature.
- **Converting Categorical Data Efficiently:** Managed large categorical variables like cuisines with appropriate encoding techniques.

7.2 Learnings

- **Improved Data Preprocessing and Visualization:** Gained expertise in cleaning and presenting data effectively.
- **Enhanced Understanding of Feature Engineering:** Learned to create impactful features for analysis.
- **Strengthened Problem-Solving in Real Datasets:** Developed practical skills in tackling real-world data challenges.

Conclusion

The Data Science Internship at Cognifyz Technologies significantly enriched my technical skills and provided practical exposure to real-world datasets. The structured progression through Levels 1, 2, and partially Level 3 enabled hands-on experience in data analysis and visualization.

My internship at Cognifyz Technologies was a transformative journey marked by extensive learning and practical application of data science concepts. Working with real-world datasets, I tackled challenges in data processing, visualization, and interpretation, sharpening my analytical abilities and problem-solving skills.

Cognifyz Technologies fostered a dynamic and collaborative environment that encouraged innovation and continuous improvement. This internship bridged the gap between theoretical knowledge and practical application, deepening my appreciation for data-driven decision-making in the tech industry. The guidance from mentors and the interactive team culture created an ideal setting for refining my technical expertise and expanding my professional network.

Through this experience, I gained a deeper understanding of meticulous data analysis—from preprocessing to deriving actionable insights that influence strategic decisions. It also enhanced my teamwork and communication skills through collaboration with peers and effective presentation of findings.

This internship reinforced my passion for data science and analytics, equipping me with the confidence to address future challenges with an analytical and pragmatic approach. I am grateful to Cognifyz Technologies for this opportunity to grow personally and professionally, and I look forward to applying these skills in my future career.

References

- Cognifyz Technologies Internship Guidelines
- Python Documentation for Pandas, NumPy, Matplotlib
- Online resources for feature engineering and geospatial analysis