**NEBOLISA KOSISOCHUKWU**

**COGNIFYZ TECHNOLOGIES INTERNSHIP DOCUMENTATION**

**07/02/2025**

**COGNIFYZ TECHNOLOGIES**

**MACHINE LEARNING INTERNSHIP TASKS**

**1. Introduction**

Food is one of the three basic human needs, and the restaurant industry has become a competitive global market, with an estimated 15 million restaurants worldwide. Understanding competition and key success factors is crucial for anyone entering this business.

This internship focused on using machine learning techniques to analyze restaurant data and build predictive models to assist in decision-making.

**TASK 1: Predicting Restaurant Aggregate Ratings**

The task was to build a machine-learning model that predicts restaurant ratings based on essential features.

**Dataset Overview**

* **Scope:** Restaurants across 15 countries
* **Size:** 9,551 rows, 22 columns

**Data Preprocessing**

* **Handling Missing Values:**
  + The Cuisines column had a few missing values, which were dropped due to their low frequency.
* **Encoding Categorical Variables:**
  + **One-Hot Encoding:** Applied to "Country Code," "Currency," "Has Table Booking," "Has Online Delivery," and "Switch to Order Menu."
  + **Label Encoding:** Applied to "City" due to its high cardinality.
  + **MultiLabelBinarizer:** Used for "Cuisines" to handle multiple cuisines per restaurant.
* **Feature Selection:**
  + Some columns were dropped due to redundancy or irrelevance:
    - **Dropped:** Restaurant ID, Name, Address, Locality, Longitude, Latitude, Rating Color, Rating Text.
    - **Kept:** Country Code, City, Cuisines, Average Cost for Two, Currency, Has Table Booking, Price Range, Has Online Delivery, Votes**.**

**Exploratory Data Analysis (EDA)**

* **Correlations:**
  + **Votes and Price Range** had the **highest correlation** with **Aggregate Rating.**
  + **Restaurants with Table Booking** had higher ratings compared to the other Yes/No categorical features
* **Country-Level Insights:**
  + **90% of restaurants** in the dataset were from **India,** causing an imbalance.
  + **The highest mean rating was from the Philippines (Code 162), although only 22 restaurants were represented.**
* **Rating vs Votes Analysis:**
  + **Many restaurants had high ratings but very few votes, leading to potential bias in model training.**
  + **India was a country that had restaurants with the highest votes and restaurants with the lowest votes as well.**

**Model Training & Evaluation**

* **Algorithm Used:** XGBoost Regressor
* **Data Split:** 70% training, 30% testing
* **Performance Metrics:**
  + **Mean Squared Error (MSE)**
  + **R-squared (R²) Score:** 94%

**Challenges & Improvements**

**Successes:** The model performed well with a high R² score.

**Challenges:** The dataset was imbalanced, with a dominance of Indian restaurants affecting generalizability.

**Future Improvement:**

**-** Try different regression models (Random Forest, Decision Tree) and handle class imbalance using resampling techniques.

- Put more energy and effort into the analysis section to be able to give a detailed explanation of the data and how each feature affected the target variable

**TASK 2: Restaurant Recommendation System**

The task was to develop a content-based filtering recommendation system based on user preferences.

**Data Preprocessing**

* **Used the same techniques as Task 1**
* **Scaling:**
  + Standard Scaler for Price Range
  + Robust Scaler for Average Cost for Two (to handle outliers)

**Feature Selection for Similarity Computation**

* Used City, Cuisines, Online Delivery, Table Booking, Price Range, and Average Cost for Two for filtering.

**Recommendation Approach**

* **Step 1:** Filter by Country Code → Ensured accessibility.
* **Step 2:** Filter by Cuisine → Primary factor influencing restaurant choice.
* **Step 3:** Filter by Online Delivery & Table Booking → Features correlated with higher ratings.
* **Step 4:** Filter by Price Range & Cost for Two → Ensured price-based relevance.
* **Step 5:** Compute Cosine Similarity → Top 5 most similar restaurants were recommended.

**Challenges & Future Work**

**Successes:** Achieved relevant recommendations with a structured approach.

**Challenges:**

-       Some cuisines were rare in certain cities, limiting recommendation diversity and number

-       Also experienced difficulty in ordering the features for filtering to obtain optimal results

**Improvements:**

* **Use K-Means Clustering** for better segmentation.
* **Expanding Recommendations Beyond the country** of the selected restaurant to include neighbouring countries of the selected restaurant using longitude & latitude data.
* **Create a column named Reviews** that encompasses everything about the restaurant, use NLP Techniques to interpret and encode the reviews which would be used for the recommendations.

**TASK 3: Cuisine Classification Model**

The task was to build a model to classify restaurants based on their cuisines**.**

**Data Preprocessing**

* **Encoding:**
  + **Label Encoding for "City."**
  + **MultiLabelBinarizer for "Cuisines."**

**Feature Selection**

* **Kept**: Country Code, City, Cuisines (Target), Average Cost for Two, Price Range, Aggregate Rating.
* **Dropped:** Redundant columns like Restaurant Name, Address, Locality, Currency, Has Online Delivery.

**Model Training & Evaluation**

* **Algorithm:** Random Forest
* **Data Split**: 80% training, 20% testing
* **Performance Metric:** Hamming Loss = 0.0145
* **There was an imbalance in the cuisine distribution in the data.** Some cuisines only appeared once in the entire dataset and had no instances in the testing data, so obtaining the performance metrics of those cuisines failed.
* **Many cuisines had a lower fraction of presence in the dataset, affecting readings.** The model could easily identify when a cuisine was absent but struggled when it was present.
* **The accuracy score of each cuisine was biased to the majority class.**

**Successes & Improvements**

**Successes:** Altogether, the model handled multi-label classification efficiently.

**Improvements:**

* **Use oversampling/undersampling techniques** like SMOTE for class balancing.
* **Try out other techniques/algorithms for classification**

**TASK 4: Geographical Distribution Analysis**

The task was to classify and analyse the distribution of restaurants globally**.**

**Process:**

* **Converted dataset to GeoDataFrame** and created a geometry column for mapping on the world map.
* **Observations:**
  + Most restaurants were **in Asia (congested in India), North America (a bit dispersed, and Australia**
  + **Few restaurants in Africa, highlighting a lack of representation.**
  + **New Delhi** had the most restaurants in the dataset.
  + **Price Range 2** was the most common across cities.
  + **The Philippines** had the highest mean rating but had very few restaurants.

**Challenges & Improvements**

**Successes:** Converted data to geospatial format successfully**.**

**Challenges:** More work needed to visualize and analyze trends effectively.

**Next Steps:**

* **Finish implementing interactive maps (Folium, Plotly) for better visualization.**
* **Learn more about Geopandas** and its vast features to tackle geospatial problems easily

**Overall Conclusion & Learnings**

**This internship provided valuable experience in the following:**

**1**. **Data preprocessing and feature selection for** machine learning: I obtained an extensive understanding of different techniques in processing data and feature engineering processes.

 2. Implementing **XGBoost, Random Forest, and Recommender Systems:** I had the opportunity to delve into ensemble learning techniques and the key aspects of selecting algorithms for your machine learning model.

3. Addressing **data imbalances, missing values, and model biases**: I understood how an imbalance in a dataset affects the model's accuracy.

4.  Using **geospatial analysis** to understand business distribution.

**Future Work**

**1.** Go in-depth in statistical data analysis, especially geospatial

**2**. Improve model generalization by balancing class distributions.

**3.**Experiment with **Neural Networks** and **Deep Learning** for better performance**.**

**4.** Accomplish an End-to-end ML and Deep Learning Project with AWS, AZURE deployment