**Secrets of the Supermarket**

**RAGHUNATH MALE(811257226)**

**Abstract:**

The project on improving data quality and user experience within the OpenFoodFacts app, addressing difficulties associated with inconsistent user-contributed data. Advanced clustering and anomaly detection techniques are used in conjunction with the Gaussian Mixture Model (GMM) to identify hidden product categories and improve data integrity. The GMM, which consists of 20 components, successfully groups items based on nutritional data. Model performance is ensured by feature engineering, data cleansing, and transformation. A thorough examination shows cluster characteristics, uncertainty patterns, and anomaly detection skills. Future work will include improving the GMM, incorporating user feedback, detecting anomalies in real-time, and increasing feature sets. The nutrition score estimate model is relevance-optimized, with a focus on continual improvement and user interaction. Finally, this lays the groundwork for a better OpenFoodFacts app, stressing continual innovation and user collaboration.

**Introduction:**

In the dynamic landscape of food information accessibility, the OpenFoodFacts dataset stands as a groundbreaking initiative, aiming to harness the collective wisdom of users to compile an expansive repository of information about diverse food products. This innovative project holds immense promise in revolutionizing how consumers make informed choices about their dietary preferences. However, the venture faces a formidable challenge – the integrity of user-contributed data. Inconsistencies, errors, and the absence of robust validation mechanisms have emerged as significant hurdles, casting a shadow on the OpenFoodFacts app's capacity to deliver dependable and trustworthy product information. In response to these challenges, this project undertakes a meticulous analysis, leveraging advanced techniques, notably the Gaussian Mixture Model (GMM), to discern anomalies and unravel the intrinsic structures embedded within the OpenFoodFacts dataset. The overarching objective is to elevate the precision of product details and refine the user experience on the OpenFoodFacts app, positioning it as an even more reliable and user-friendly resource for accessing crucial information about food products.

**Project Description:**

The OpenFoodFacts project, a collaborative endeavor to crowdsource and share comprehensive information about food products, has witnessed substantial success in its mission to empower consumers with knowledge. Nevertheless, the project confronts a critical hurdle – the quality and accuracy of user-contributed data. The prevalence of inconsistent entries, errors, and the lack of effective validation mechanisms has posed a multifaceted challenge to the OpenFoodFacts app. This project, in response to these challenges, embarks on a comprehensive analysis employing sophisticated techniques, notably the Gaussian Mixture Model (GMM). The GMM serves as an advanced analytical tool, capable of identifying anomalies and unraveling the inherent structures within the OpenFoodFacts dataset. The primary objective is clear: to enhance the accuracy of product details and streamline user experiences on the OpenFoodFacts app. By doing so, this project aspires to fortify the app's position as a reliable and efficient platform for users seeking precise and trustworthy information about the foods they consume.

**Problem Statement:**

At the core of this project lies the profound challenge of compromised data quality within the OpenFoodFacts dataset. This challenge emanates from the diverse and occasionally inaccurate nature of user-contributed data, leading to inconsistencies, inaccuracies, and an overall lack of robust validation mechanisms. These issues collectively undermine the fundamental purpose of the OpenFoodFacts app – to provide users with dependable, swift, and accurate information about food products. The dataset's compromised quality poses a threat not only to the integrity of the app but also to its overarching mission of empowering users with essential details about the foods they choose. This project strategically employs the Gaussian Mixture Model to conduct a nuanced analysis, aiming to pinpoint anomalies, unravel intrinsic structures, rectify inaccuracies, and fortify data quality. The ultimate goal is to streamline and elevate user experiences on the OpenFoodFacts app, ensuring it remains a trusted resource for accessing vital information about food products.

**Background:**

The inception of OpenFoodFacts was driven by the escalating demand for transparent and easily accessible information about various food products. As consumers increasingly prioritize factors such as nutritional content, allergen information, and ethical considerations, OpenFoodFacts aimed to establish a platform where users could actively contribute and access a wealth of information about a diverse array of food items. However, the diverse nature of user-contributed data posed a significant challenge, introducing inconsistencies and inaccuracies that compromised the overall reliability of the dataset. This background underscores the pivotal importance of ensuring data quality within OpenFoodFacts. A robust and accurate dataset is not only fundamental but crucial for empowering users with reliable information about the foods they consume. Through the strategic utilization of advanced analytical methods, specifically the Gaussian Mixture Model, this initiative seeks to address data quality challenges head-on, contributing to a more trustworthy and user-friendly OpenFoodFacts app.

**1.Approach:**

A Gaussian Mixture Model (GMM) is a probabilistic model used for clustering and density estimation. In the given context of 200,000 nutrition tables, the GMM assumes latent variables corresponding to product categories, each modeled by a multivariate Gaussian distribution with features such as carbohydrates (C), sugars (S), fat (F), proteins (P), energy (E), and salt (S). The distribution of observed data (x) is expressed as a sum of these Gaussians, where each Gaussian (k) represents a product category:

A mathematical equation with numbers and symbols

Description automatically generated

Here, πk​ is the weight of the k-th Gaussian, and N(x∣μk​,Σk​) is the multivariate Gaussian distribution with mean μk​ and covariance matrix Σk​. The Expectation-Maximization (EM) algorithm is used for learning, consisting of two steps:

1. E-Step (Expectation):

A math equations with a line

Description automatically generated with medium confidence

Here, γnk​ represents the responsibility of the k-th Gaussian for the n-th data point. The responsibilities sum to 1 over all clusters for each data point.

1. M-Step (Maximization):

A group of mathematical equations

Description automatically generated

In the maximization step, the parameters of the Gaussians are updated. μk​ is the mean vector, and Σk​ is the covariance matrix. The weights πk​, means μk​ and covariances Σk​ are iteratively adjusted to better fit the data.

In the feature engineering process, several adjustments are made to the dataset to enhance its suitability for clustering analysis, particularly for Gaussian Mixture Models (GMM). Three new features are introduced:

g\_sum Feature:

Represents the rounded sum of fat, carbohydrates, proteins, and salt values in the data.

Helps identify products with potentially incorrect entries.

other\_carbs Feature:

Includes the value of carbohydrates that are not sugars, allowing the model to capture the correlation between carbohydrates and sugars.

reconstructed\_energy Feature:

Calculates the energy value of a product based on the fat, carbohydrates, and proteins content.

Aids in identifying potential discrepancies between the given energy value and the calculated one.

Furthermore, a binary meta-data frame is constructed to detect situations where any feature has a value of zero or contains at least one zero, so providing additional information for data validation.

Following feature engineering, a set of data cleaning methods are put in place to remove obvious errors. Products containing more than 100g of characteristics (save for energy-related features), negative entries, energy values greater than 3700 kJ, more sugars than carbs, and g\_sum values greater than 100g are excluded.

Following that, the dataset is feature scaled and transformed to prepare it for GMM. The Box-Cox transformation is used to deal with skewed feature distributions with a high number of outliers. This adjustment enlarges the vast majority of data points while compressing outliers, assisting in the detection of hidden substructures. Furthermore, the features are scaled to have a zero mean and unit variance to lessen the influence of greater values on cluster center and covariance updates.

A graph of different colored lines

Description automatically generated

The Box-Cox transformation's power parameter λ offers a means to control outlier compression, with lower values emphasizing more pronounced compression. When λ is fixed at 0.5 and the constant c is adjusted, minimal variation is observed in high-value compression, while lower values exhibit increased stretching, especially when c approaches zero. In summary, opting for a lower λ accentuates outlier compression, while selecting a lower c, particularly near zero, emphasizes the stretching of low values.

Hyperparameters for transforming nutrition features are individually set, influencing resulting clusters. Using specific constants and lambdas, the TransformParameter class guides the transformation process for each feature in the nutrition table. A loop applies these transformations, visually comparing the original distribution, Box-Cox transformation (for positive values), and the final transformed distribution. While improvements are evident, determining optimal hyperparameters remains a challenging task.

A group of graphs showing different colored lines

Description automatically generated

**2. Implementation and Results:**

The Gaussian Mixture Model (GMM) refines the M-step update during the training phase by leveraging transformed and scaled data. The issue of determining the ideal number of components is complex, with evaluations based on log-likelihood, Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC). Striking a balance is critical, because too few components might result in broad, mixed clusters, while too many can result in overly particular or duplicate clusters. Another strategy is to examine feature correlations between clusters when determining a reasonable number. A GMM with 20 components is trained and successfully converges in a practical demonstration. The selection of a suitable number of clusters necessitates careful evaluation of statistical indicators as well as the intricacies of real-world data.

**2.1. Clustering of product types:**

The decision to fit the Gaussian Mixture Model (GMM) with 20 components proves effective when examining cluster center correlations in subsequent analyses. To gain insights into interesting clusters, a selection of three features among carbohydrates, fat, proteins, salt, other\_carbs, sugars, energy, reconstructed\_energy, and g\_sum are considered. It's noteworthy that transformed features, denoted as "transformed\_" + your\_feature, can also be included in the analysis. For stability reasons in utilizing Plotly, the plot focuses on 40,000 data points. This approach allows for a detailed exploration of cluster behaviors and relationships among the specified features.

A diagram of a protein source

Description automatically generated with medium confidence

The Gaussian Mixture Model (GMM) allows for a wide range of cluster sizes and densities, each with its own covariance matrix. While some clusters have sparse boundary regions, motivating more inquiry into outlier identification and model confidence, feature transformation during preprocessing broadens the data space, revealing unanticipated substructures. Notably, the discovery of a cluster that acts as a protective barrier for outliers adds an unexpected and fascinating layer to the investigation.

Data Covered per cluster:

A graph and diagram of a graph

Description automatically generated with medium confidence

Most clusters encompass approximately 5% of the data, yet variations exist with some clusters displaying notably low or high coverage. Further exploration is needed to understand the product types and feature ranges associated with these clusters, determining their homogeneity and assessing the potential need for additional components in the mixture model.

We discovered hidden product categories only based on nutrition table information by successfully applying Gaussian Mixture. Chocolate, pasta, ice cream, cheese, yoghurts, juice, grains, sauces, meat, water, and nuts were discovered in the clusters. While the coarse-grained approach to clustering was extremely successful, closer examination reveals concerns about clusters including comparable or mixed-type products, as seen in cluster 6. Other questions include determining the model's confidence in cluster assignments and analyzing data veracity in the context of probable app user faults. The cluster\_names dictionary contains named categories for each cluster, and a new "category" column in the nutrition\_table aids comprehension of the recognized product types.

A close up of words

Description automatically generated

Similarity between different Clusters

A screenshot of a graph

Description automatically generated

While some clusters are dissimilar and cover different products, others are highly correlated despite their nearby centers, suggesting a subtle difference between products like pasta and whole grain.

A screenshot of a graph

Description automatically generated

Clusters like pasta and grain, though closely positioned in the transformed feature space, exhibit nuanced differences. Pasta has slightly lower fat and higher protein than the whole grain cluster, which contains more variation in sugars and salt. The mixed-type cluster 6 encompasses outliers in various features, explaining broad distributions on low values. Clusters 6 and 12 are remarkably similar, differing mainly in fat, proteins, and carbohydrates, revealing challenges posed by numerous discrete 0-entries in these nutrients, a topic explored further in the outlook.

Density of the clusters:

A screenshot of a computer screen

Description automatically generated

Cluster densities differ significantly, with beverages in Cluster 5 showing the highest density (minimal fat, protein, salt, varying sugar) and outliers in Cluster 6 leading to the lowest (mixed types and nutritional values). Additionally, specific clusters exhibit notable high or low densities in certain features (e.g., Cluster 12: candy, fruits, sauces with low density in carbohydrates).

A red and orange striped chart

Description automatically generated

**2.2. Certainty Analysis:**

The Gaussian Mixture Model offers the advantage of probabilistic cluster assignments, providing insights into products that lie between two or more clusters. By examining the normalized responsibilities (πkN(μk, Σk) ) per cluster component, obtained during the E-Step, we can gauge the model's certainty in assigning data points to clusters. This probabilistic approach distinguishes Gaussian Mixture from non-probabilistic models like K-Means, allowing for a more nuanced understanding of uncertain cluster assignments.

A close-up of a graph

Description automatically generated

With high certainty for most data points (>95%), our model suggests only a few outliers with ambiguous cluster assignments (around 0.5).

A graph of arrows with text

Description automatically generated with medium confidence

Despite investigating the possibility of clusters exhibiting higher uncertainty or similar clusters sharing increased uncertainty, no discernible pattern was identified.

A group of words on a white background

Description automatically generated

High certainty data points are closely associated with product names or categories, but lower certainty data points frequently reflect different product types. For example, the pasta and grain clusters have overlaps in unclear data points, implying that some cereals in the pasta cluster may potentially have a second peak of responsibility in the grain cluster, showing items "in-between."

A screenshot of a computer

Description automatically generated

While Pasta cluster 9 exhibits high uncertainty (~78%) towards Grain cluster 16, the converse is not true (Grain cluster 16 has low uncertainty towards Pasta with only 11% responsibility). This suggests non-mutuality in cluster competition. Furthermore, the heatmap reveals intriguing interactions: Creamy, Cheesy cluster 4 attracts several others, including Meat cluster 10, Cheese cluster 14, and Virgin Oil & Sauces cluster 17. Analyzing features through violin plots reveals the reason: cluster 4 encompasses a broad range of proteins, fat, and salt, leading to its high second-highest responsibilities.

A screen shot of a diagram

Description automatically generated

Above is the result just to get an idea of how cluster assignment uncertainty looks in the feature space.

**2.3. Anomaly Detection:**

The Gaussian Mixture Model (GMM) is a key component in anomaly detection, relying on the probability density function formulation:

p(x)=∑k=1Kπk⋅N(x∣μk,Σk)

Here, πk represents the weight of each cluster, andN(x∣μk,Σk) is the Gaussian distribution function. The logarithm of this function is utilized for scoring each data point:

lnpn(x)=ln∑k=1Kπk⋅N(xn∣μk,Σk)

The resulting value, obtained through score samples, reflects the density of the region where data spot xn is situated. Lower values indicate low density regions, making it a potential measure for anomaly detection. Anomalies could include errors or rare products located in sparse regions.

Anomalies are defined based on a threshold derived from the p-quantile, with a chosen value (e.g., 7%) determining the proportion of anomalies in the data.

Visualizing Anomalies:

A graph with red and blue dots

Description automatically generated

**2.4. Estimating Nutrition Scores:**

The nutrition scoring process integrates clustering outcomes, binary anomaly scores, log-likelihood (logL), and cluster responsibilities to predict nutrition scores for products in the database. Categorized into groups for both beverages and food, these scores are associated with specific ranges reflecting different healthiness levels. The "nutri\_score" column in the nutrition table captures the original nutrition scores, with approximately 8.4% of entries marked as missing. The prediction model utilizes available clustering and nutrition information to estimate scores, offering insights into the health attributes of products within the database.

A close up of numbers

Description automatically generated

**Future Work:**

The Gaussian Mixture Model (GMM) will be continuously refined in the future by experimenting with different hyperparameters, cluster counts, and feature engineering methodologies to improve clustering accuracy. The incorporation of user feedback methods within the OpenFoodFacts app could collectively validate and rectify data entries, contributing to ongoing data quality improvement. Real-time anomaly detection skills should be created to allow for the rapid discovery and correction of data mistakes, resulting in a dynamic and responsive user experience. Including more nutritional and product-related features in the feature set for clustering and anomaly detection could provide a more thorough understanding of food goods. Furthermore, enhancing the nutrition score estimate model, taking into account aspects such as dietary standards or user-specific preferences, has the potential to improve the relevance and utility of the model.

**Conclusion**

The initiative has made substantial progress in addressing the issues related to poor data quality in the OpenFoodFacts dataset. The study not only found anomalies and underlying structures but also uncovered hidden product categories based only on nutrition table information by employing advanced approaches such as the Gaussian Mixture Model. The findings highlight the potential for improving the OpenFoodFacts app's user experience, making it a more credible and informative resource for people seeking accurate information about the food they consume. While the current analysis gives useful insights, it is critical to acknowledge the ever-changing nature of user-contributed data, which necessitates ongoing efforts for continual development.