- Developed a clustering model for Zomato restaurants based on reviews and features, allowing the identification of different restaurant segments and assisting in strategic marketing decisions.
- Faced challenges in text preprocessing for sentiment analysis, conducted feature engineering for clustering, and resolved high dimensionality using PCA for better visualization.
- Achieved a silhouette score of 0.67 for clustering, indicating well-defined clusters, and obtained 84% accuracy for sentiment classification using the Naive Bayes model.
- Deployed using Streamlit, creating a user interface to explore clusters and review sentiments. This solution has the potential to enhance targeted marketing and improve customer satisfaction by 30%.

Keywords in CSV format:
Clustering, PCA, Silhouette Score, Sentiment Analysis, Naive Bayes, Feature Engineering, Text Preprocessing, High Dimensionality, Data Visualization, Customer Segmentation, Targeted Marketing, Python, Scikit-learn, NLTK, SpaCy, Streamlit, Model Deployment, Customer Satisfaction, Predictive Analytics, Restaurant Analysis

Zomato Restaurant Clustering Workshop: 10 Project-Specific Interview Questions

1. **What was the main objective of the Zomato Restaurant Clustering Workshop?**
    - To segment restaurants into different clusters based on attributes like cost, ratings, and cuisine types and to analyze customer sentiments from reviews to derive insights for restaurant recommendations.
2. **What clustering algorithm did you use, and why was it suitable for this project?**
    - The K-Means clustering algorithm was used because it effectively partitions restaurants into distinct clusters based on numerical features like cost and ratings, helping identify homogeneous groups.
3. **What were the challenges faced during clustering, and how did you address them?**
    - Challenges included selecting the optimal number of clusters and dealing with high-dimensional data. Solutions involved using the Elbow Method and applying Principal Component Analysis (PCA) to reduce dimensions.
4. **Explain the significance of the Silhouette Score in evaluating clustering quality.**
    - The Silhouette Score measures how similar a data point is to its own cluster compared to other clusters. A higher score indicates well-separated and meaningful clusters.
5. **How did you preprocess the text data for sentiment analysis, and which libraries did you use?**
    - Preprocessing involved tokenization, stop-word removal, lemmatization, and vectorization using NLTK and SpaCy. The text was then converted into numerical features using TF-IDF.

6. **What approach did you take to perform multi-dimensional clustering on restaurant data?**
    - Multi-dimensional clustering was performed using PCA to reduce feature space and then applying K-Means clustering on the reduced dimensions. This approach helped capture complex patterns without overfitting.
7. **What features did you include in the clustering, and how did you decide on these features?**
    - Features like cost, ratings, cuisine types, and collection tags were included. Feature selection was guided by exploratory data analysis and domain knowledge to ensure they reflected distinct customer segments.
8. **How did you validate your clusters to ensure they were meaningful and actionable?**
    - Clusters were validated using the Silhouette Score and visual inspection (scatter plots, PCA projections). Additionally, cluster characteristics (e.g., average cost, predominant cuisine) were analyzed to ensure each cluster had business significance.
9. **What insights did you gain from the sentiment analysis of reviews, and how did it complement the clustering?**
    - Sentiment analysis revealed areas of high and low customer satisfaction, helping to understand why certain clusters were preferred. For example, clusters with positive sentiment often had better ratings and higher customer retention.
10. **How would you improve the clustering model if new restaurant data were to be included?**
    - I would update the clustering model by recalibrating it with new data points, use a dynamic clustering approach, and incorporate new features such as location or demographic data to make the clusters more robust.

---

20 Technical Interview Questions for Zomato Clustering and Sentiment Analysis

21. What is the curse of dimensionality, and how did it affect your clustering model?

- The curse of dimensionality refers to the difficulty in analyzing high-dimensional data due to the exponential growth of feature space. It affects clustering because distance metrics become less meaningful. This was mitigated using PCA to reduce dimensions.

22. Why did you choose K-Means over Hierarchical Clustering for this project?

- K-Means is computationally efficient for large datasets and can handle updates to clusters when new data is added, unlike Hierarchical Clustering, which is better suited for smaller datasets and does not allow incremental updates.

23. How would you use PCA for this project, and what is the impact of choosing the number of components?

- PCA was used to reduce the dimensionality of features like cost, ratings, and collection tags. Choosing the number of components impacts how much variance is retained in the dataset, which directly influences clustering quality.

24. What are the limitations of using K-Means for clustering text data?

- K-Means relies on Euclidean distance, which may not capture the semantic relationships in text data. Text features are high-dimensional and sparse, making it harder to find meaningful centroids.

25. Explain how you would handle imbalanced classes in sentiment analysis.

- Techniques like oversampling the minority class, undersampling the majority class, or using algorithms like SMOTE (Synthetic Minority Over-sampling Technique) can help balance classes for more accurate sentiment classification.

26. How would you evaluate the performance of a sentiment analysis model?

- Metrics like Accuracy, Precision, Recall, F1-Score, and Confusion Matrix would be used. Additionally, ROC-AUC scores and qualitative analysis (e.g., examining misclassified samples) would provide deeper insights.

27. What is the role of feature scaling in clustering algorithms like K-Means?

- Feature scaling ensures that all features contribute equally to the distance metric used in K-Means. Without scaling, features with larger ranges dominate the clustering process.

28. You used NLTK and SpaCy for text preprocessing. What are the main differences between them?

- NLTK is more suitable for educational purposes and basic NLP tasks, while SpaCy is optimized for production and provides faster tokenization, parsing, and named entity recognition.

29. How would you modify your clustering algorithm to include text-based features?

- Use TF-IDF to convert text data into vectors and include these vectors as additional dimensions in the feature space. Alternatively, use Word2Vec or BERT embeddings for more semantic-rich representations.

30. What are stop words, and why are they removed during text preprocessing?

- Stop words are common words like "and", "the", and "is" that do not add significant meaning to the text. They are removed to reduce noise and improve the performance of text-based models.

31. Explain the use of the Elbow Method in choosing the optimal number of clusters.

- The Elbow Method plots WCSS (Within-Cluster Sum of Squares) against the number of clusters. The "elbow point" represents the optimal number of clusters, where adding more clusters does not significantly reduce WCSS.

32. How does the silhouette score work, and what does a low score indicate?

- The silhouette score measures how well each data point is assigned to its cluster, with a score close to 1 indicating good clustering. A low score (< 0.25) suggests overlapping clusters or poorly defined boundaries.

33. How would you incorporate a cost-benefit analysis into your clustering framework?

- Assign cost and benefit scores to each restaurant cluster based on average spending and customer reviews. Use these scores to identify clusters with high customer satisfaction relative to cost, guiding business strategies.

34. What is the impact of choosing an incorrect number of clusters in K-Means?

- Choosing too few clusters results in underfitting, while too many clusters lead to overfitting, capturing noise rather than meaningful patterns. It can lead to misinterpretation of the data.

35. Describe a scenario where K-Means clustering would fail.

- K-Means would fail in datasets with non-convex shapes or varying cluster sizes. For example, clustering two concentric circles would not work as K-Means tries to create spherical clusters.

36. What is the difference between soft and hard clustering?

- **Hard Clustering:** Assigns each data point to exactly one cluster.
- **Soft Clustering:** Assigns data points to multiple clusters with a probability score, used in models like Gaussian Mixture Models (GMM).

37. How would you determine if text preprocessing has improved model performance?

- Compare metrics (e.g., accuracy, F1-score) of models with and without preprocessing. Observe qualitative improvements in misclassified examples and check for more distinct cluster boundaries.

38. If you notice that some restaurant clusters are overlapping significantly, what would be your approach?

- Consider using advanced clustering techniques like DBSCAN that can handle arbitrary cluster shapes. Use domain knowledge to add new features or try kernel methods to transform the feature space.

39. How would you use sentiment scores from reviews in clustering analysis?

- Calculate sentiment scores using NLP models and use them as an additional feature in the clustering model. This can highlight clusters with high or low customer satisfaction, making the segmentation more meaningful.

40. Explain the concept of Term Frequency-Inverse Document Frequency (TF-IDF).

- TF-IDF measures the importance of a word in a document relative to its frequency in a collection. It is calculated as TF * IDF, where TF is the frequency of a word in a document and IDF is the inverse of its frequency across all documents. It helps filter out common words and highlights rare but important terms.