Executive Summary for Data Science Practicum Project

Introduction/Motivation

Our project delves into the analysis of Yelp data, specifically focusing on coffee and tea shops in Pennsylvania (PA). The primary objective is to uncover key factors influencing ratings for these establishments, offering valuable insights for business owners and consumers.

1. **Domain and Questions:**

Domain: Coffee and tea shops in Pennsylvania.
Key Questions:

- What are the critical factors that impact ratings for coffee and tea shops in PA?
- Which aspects can business owner take action to improve their ratings?

2. **Thesis Statement:**

Our analysis unveils crucial insights into the determinants of PA coffee and tea shop ratings.

Background Information

Our project utilizes the data source from Yelp datasets, a subset of "Yelp Dataset Challenge", comprising JSON files with 6,990,280 reviews and 150,346 businesses, focusing on recommended reviews and filtered for quality. We specifically focus on PA coffee and tea shops, the reason why we selected Pennsylvania lies in its abundance of businesses, which makes data of the region most representative in the whole datasets.

Data Cleaning, Merging and Pre-Processing

Meticulous data cleaning and preprocessing, were essential for maintaining the integrity of our analysis. The datasets merging was applied for further analysis.

Data Cleaning:

- Addressed missing values, duplicates, and ensured data quality through Pandas functions.
- Addressing Outliers: Utilized z-score analysis and IQR filtering to identify and handle outliers appropriately.

Data Merging and Preprocessing:

- Merging Techniques: Applied Pandas merge() function for consolidating relevant data, ensuring a comprehensive analysis.
- Review Data Integration: Integrated review data by the unique business_id identifier, to correlate all reviews for each business, which laid the foundation for comprehensive analysis, providing a holistic view of customer sentiments.
- Text Preprocessing: Implemented rigorous text preprocessing methods before modeling, involving removal of common stop words, extraneous spaces elimination and punctuation handling, which in order to ensure the meaningfulness and cleanness of text data.

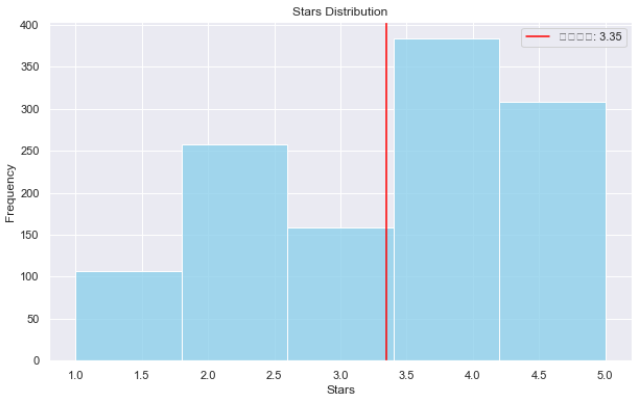
Exploratory Data Analysis (EDA)

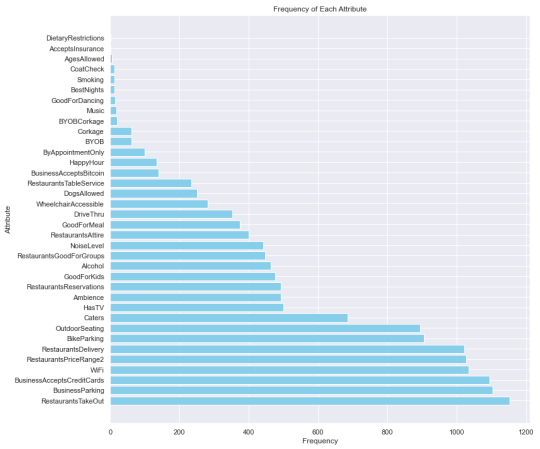
In this section, we further analyzed the selected data, including missing values handling, data transformation, feature extraction and visualization. The goal is to facilitate subsequent modeling, identify key factors influencing ratings, and uncover patterns behind business operations.

1. **Visualizing the stars ratings**

Calculate the average coffee shop rating and present the distribution of star ratings using a histogram, which highlights the prevalence of 4-star and 5-star ratings, with 2-star ratings more common than 1-star and 3-star. This pattern suggests that customers are likely to leave reviews only when particularly satisfied or dissatisfied with the service.

*





2. **Visualizing Attribute Frequencies:**

Utilized horizontal bar plots as shown above to visualize the frequencies of each attribute and excluded less frequent attributes.

3. **Data Transformation**

(1) Dummy Variable Transformation:

Dummy Variable Transformation: Converted categorical parking features into dummy variables by creating a subset (BP) to filter out missing values and replacing boolean values with string representations ("True" and "False"). This step improves compatibility with diverse modeling techniques.

(2) One-Hot Coding and Missing Values Handling

Apply one-hot encoding to categorical columns, like WiFi, RestaurantsPriceRange2, Alcohol, NoiseLevel, RestaurantsAttire, using pd.get_dummies, transforming them into suitable format for modeling. Employ the 'fillna' method to replace missing values with 0.

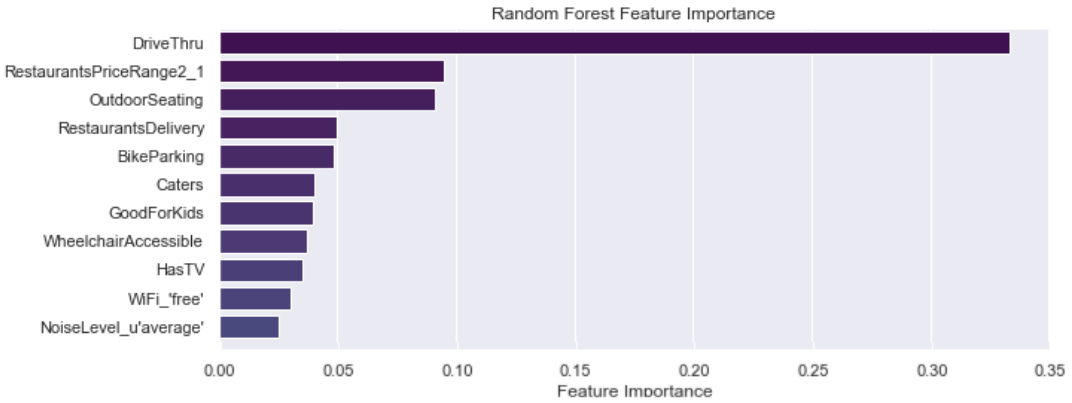
Key Findings
About Coffee & Tea Shops in PA Market on Yelp

This section covers building a rating prediction model, analyzing feature importance, conducting model diagnostics and explanations. We also applied Latent Dirichlet Allocation (LDA) to reviews data and visualize the results. And summarized the strengths and weaknesses of both models.

Predictive Model: Random Forest (RF)

1. Random Forest Model Construction

A RF Regressor model was built using the RandomForestRegressor, which was configured with parameters: n_estimators=100, max_depth=10, min_samples_split=2, min_samples_leaf=2, and random_state=42. The dataset was split into training and testing sets respectively 80% and 20%.



2. Feature Selection and Importance Visualization
We visualized importance scores and ranked feature importance through a bar chart, and we noticed that "Drive through", "Price Range", "Outdoor Seating", "Restaurants Delivery" and "Bike Parking" are the most important features to influence stars of shops.

3. Model Diagnostics (Model Performance Evaluation)
After predicting both training and test sets, performance metrics like MAE, MSE, and R-squared were computed for model evaluation.The R-squared value, approaching 1, indicates a better fit of the model for predicting coffee shop ratings. Model diagnostic results are as follows: Training MSE: 0.3988, Training R-squared: 0.6326
Test MSE: 0.6508,Test R-squared: 0.4151

4. Model Interpretation
Our Random Forest model uncovers patterns and correlations, predicting coffee shop ratings. Feature selection and metrics like MAE, MSE, and R-squared quantify accuracy, guiding business decisions. Overall, the model contributes significantly, offering insights, interpreting data, and making predictions.

LDA Model

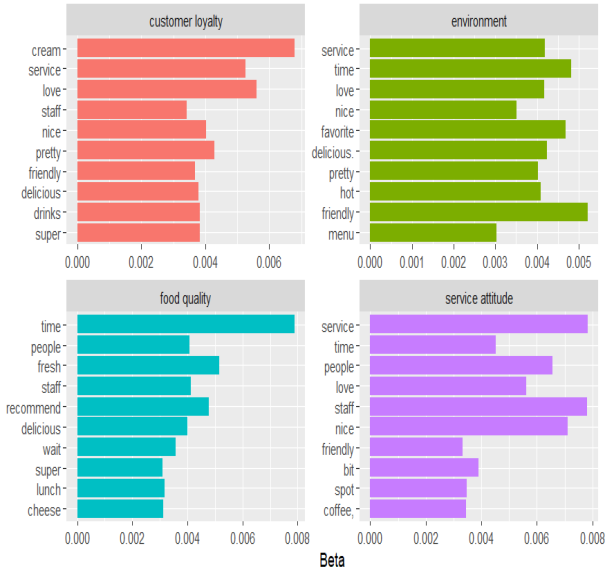
LDA is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

1. Generate Topic Word Cloud:
Visually represent high-frequency words for each topic, providing a quick understanding of key features in each theme.

2. Topic Model Construction:
We opted for 4 topics, creating a Document-Term Matrix (DTM) pre-modeling. Stopwords were removed, and a stemming function applied. Leveraging unsupervised learning with LDA, the model assumes each document is a mixture of topics represented by a distribution of words. This approach helps unveil underlying patterns in the text.

3. Model Results and Explanation:
We conducted topic modeling to analyze the review of the coffee shop. The topic modeling categorized reviews into 4 groups: customer loyalty, environment, food quality, service attitude. We found that in the topic of service attitude, customers mentioned most about "friendly, time, nice", which indicates that they care about serving speed and good attitude when consuming. And in the topic of food quality, people will have higher expectations on

freshness, taste, and wait time. There are several limitations of LDA, 1. There is a limit to the amount of topics we can generate 2. LDA is unable to depict correlations which led to occurrence of uncorrelated topics



Random Forest Model Strengths and Weaknesses
strength:
Handling Overfitting/Accuracy: Mitigates the risk of overfitting by averaging multiple trees
Feature Importance: Capable of handling a large number of input variables and identifying the most significant variables.
Weakness:
Model Interpretability: poor interpretability for decision making.
Bias towards Certain Attributes: RF might be biased towards attributes which dominate in cases, leading to skewed results.

Conclusion/Suggestions

In conclusion, our comprehensive analysis offers actionable insights for stakeholders in Pennsylvania's coffee and tea shop market. Transparent acknowledgment of assumptions and limitations provides a nuanced interpretation of results. We provide suggestions based on topic modeling: 1. Focus on staff training to make the service more friendly. 2. Improve food quality by freshness and taste. Additionally, we provide suggestions based on business attribute analysis:1. Offer reasonable prices. 2. Invest more in outdoor seating.

Contribution:

Zongliang Han: Clean Data and Filter Data, EDA**,** LDA, Random Forest, Presentation ppt making

Chuyi Lin: Clean Data and Filter Data, EDA, LDA, summary report writing

Ziyi Yang: Clean Data and Filter Data, Shiny app, Random Forest