From Coffee Machines to Machine Learning, Accenture

Steven Shi

Accenture 2B

AI Studio Project Write-Up

Fall 2023

Table of Contents

Business Focus

This project centers on applying machine learning methodologies to assist a client in the coffee shop industry in optimizing various facets of their business operations. The primary objectives include:

1. Location Optimization:
   Utilizing machine learning algorithms and cleaned datasets to pinpoint the most advantageous location for the client's coffee shop in New York City. Considerations involve variables such as population density, demographic insights, competition analysis, foot traffic patterns, and proximity to key transportation hubs.
2. Specialty Items Selection:
   Leveraging customer and business data to discern unique consumer preferences and current trends within the coffee shop market. Identifying three distinct specialty items that align with customer demands, as well as discovering any findings for items to avoid for the appeal of the coffee shop.
3. Business and Marketing Strategies:
   Analyzing data and findings in graphs and other mediums to craft tailored marketing strategies for both the business strategy and individual success of the coffee shop. Analyzing customer reviews to determine attributes correlated with successful and underperforming coffee shops, enabling data-driven decisions regarding operational and marketing priorities.

This project aims to use the power of machine learning to optimize decision-making for the coffee shop stakeholder, fostering customer satisfaction, refining business strategies, and driving growth and sustainability in the NYC market for a morning's cup of coffee.

Data Preparation and Validation

DATASET DESCRIPTION

For the location model, our dataset encompasses diverse variables:

- Business Information: `ID, alias, name, reviews_count, categories, rating, transactions, location.zip_code, location.display_address.`
- Demographics Data: extracted from `'demographics.xlsx'`, including `Best Population Estimate`.

The reviews model dataset comprises of the following columns:

- Business Details: `business_id, name, address, city, state, postal_code, latitude, longitude, stars_x, review_count, is_open, attributes, categories, hours, business_name.`

DATASET PREPROCESSING

Location Model:

- **Initial Data Assessment**:
  - Dropped irrelevant columns: `'is_closed', 'url', 'image_url'`, and others.
- **Merging Demographics Data**:
  - Loaded and merged 'demographics.xlsx' based on the `'location.zip_code'` and `'geography'` columns.
  - Replaced zero values with NaN for further processing.

Menu Model:

- **Yelp Dataset Integration:**
  - Combined `'business.json'` and `'reviews.json'` from the Yelp Dataset and joined them into one .csv file.
- **Data Inspection**:
  - Examined columns such as `business_id, name, address, stars_x, stars_y, useful, funny, cool, text, date,` among others.
- **Filtering Relevant Data:**
  - Focused on filtering reviews specifically relevant to coffee shops, employing a text corpus and for menu item detection.

Marketing:

- **Unified Dataset**:
  - Merged the location-based information with the filtered menu data to create a consolidated dataset for marketing analysis.

EDA (EXPLORATORY DATA ANALYSIS)

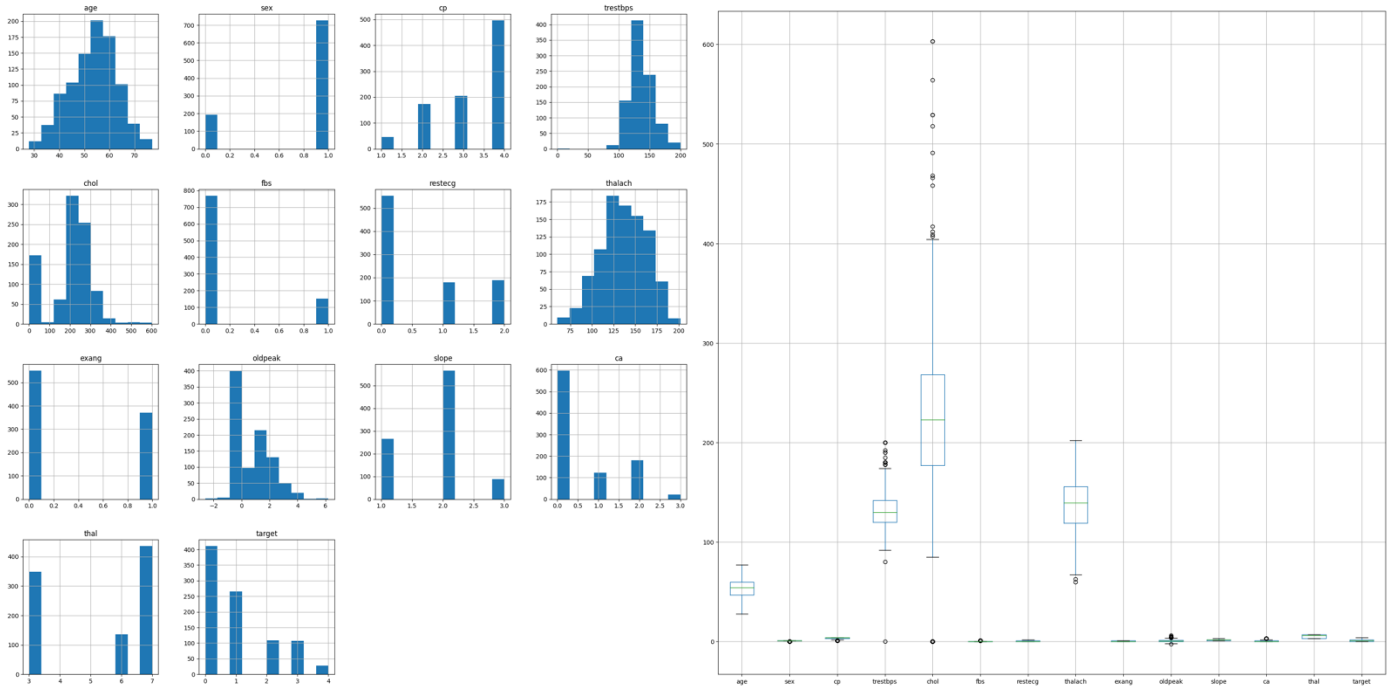
During the Data Understanding and Preparation phases, our team had many insights while exploring the datasets using our Google Colab notebooks. See our specific EDA in Google Colab, here: location_model and menu_model.

Insights from Census Data:

- **Demographic Analysis**:
  - Leveraging EDA on the census dataset, we identified areas characterized by the highest income levels and racial demographics. Visualization techniques aided in understanding geographical nuances through insightful graphical representations.
- **Visualization Impact**:
  - Visualization via correlation plots and graphs facilitated a deeper comprehension of our dataset, transcending the limitations of traditional CSV file scrutiny.

Challenges Encountered:

- **Yelp API Limitations**:
  - Obtaining comprehensive data from the Yelp API was a challenge. The API's limitations, providing only 50 data points for individual business data per call, led us to devise a workaround. Employing a script to request multiple API calls, we eventually created our CSV file.
- **Review Data Limitation**:
  - Restricted reviews per cafe (only 3 per API call) called for an alternative approach. We used a Yelp dataset with over 30,000 reviews, circumventing limitations inherent in the Yelp API.



FEATURE SELECTION

To enhance our location model's predictive capability, we employed various techniques to select relevant features and generate new columns for evaluation:

- **Feature Engineering based on Zip Code**:

    - We calculated an `'average rating'` column for each coffee shop using the zip code information. This new feature helps determine the average rating based on coffee shops in the same area.
- **Aggregating Café Names by Zip Code:**
    - We created an `'all_names'` column listing all café names in a particular zip code area.
- **Optimal Location Labeling**:
    - Introduced an `'optimal_location'` binary classification based on predefined criteria like high average ratings and reviews to population ratio.
- **Data Refinement**:
    - To enhance clarity and manageability, we rounded the `'Reviews to Population Ratio'` to 3 decimal places and sorted the DataFrame for better analysis.

Approach

Location Model: Random Forest Binary Classifier

Objective: The location model aims to categorize coffee shop locations as 'optimal' or 'non-optimal' based on specific criteria like 'Reviews to Population Ratio' and 'average ratings' within zip codes.

Methodology:

1. **Data Preparation:**
    - **Zip Code Grouping:** Organized coffee shop data by zip codes.
    - **Criterion Definition:** Derived 'Reviews to Population Ratio' and 'average ratings' for each zip code.
2. **Model Selection and Training:**
    - **Model Chosen:** Employed a Random Forest Binary Classifier.
    - **Feature Selection:** Utilized 'review_count', 'Best Population Estimate', 'Reviews to Population Ratio', 'average rating'.
    - **Training and Testing:** Split the data into training and testing sets.
3. **Hyperparameter Tuning:**
    - **Parameter Optimization:** Explored hyperparameters for the Random Forest Classifier to enhance model performance.
4. **Model Evaluation:**
    - **Performance Metrics:** Assessed model accuracy, precision, recall, and F1-score.
    - **Prediction Validation:** Tested predictions for new location data.
5. **Final Model Selection and Deployment:**
    - **Model Validation:** Evaluated the best-performing model based on validation results.
    - **Deployment Strategy:** Prepared the model for deployment considering scalability and robustness.

Menu and Marketing Model: NLP Analysis of Reviews Data

Objective: The menu and marketing model focuses on using NLP techniques to derive insights from customer reviews, categorize them based on sentiment, and extract common words/phrases related to coffee shops.

Methodology:

1. **Data Filtering and Preprocessing:**
    - **Keyword Filtering:** Identified relevant keywords to filter coffee-related reviews.
    - **Category Identification:** Categorized reviews based on specific coffee-related categories.
2. **NLP Analysis:**
    - **Keyword Extraction:** Extracted food items and specific terms related to coffee from reviews.
    - **Sentiment Analysis:** Categorized reviews into positive, neutral, and negative sentiments.
3. **Feature Extraction and Insights:**
    - **N-gram Analysis:** Explored common word sequences (bigrams, trigrams) in different types of reviews.
    - **Customer Sentiment Insights:** Derived insights into customer preferences and sentiments based on review analysis.
4. **Final Model Evaluation:**
    - **Review-Based Insights:** Analyzed the most common words/phrases across different review types.
    - **Final Insights for Marketing Strategy:** Extracted actionable insights for menu and marketing strategy based on sentiment and customer preferences.

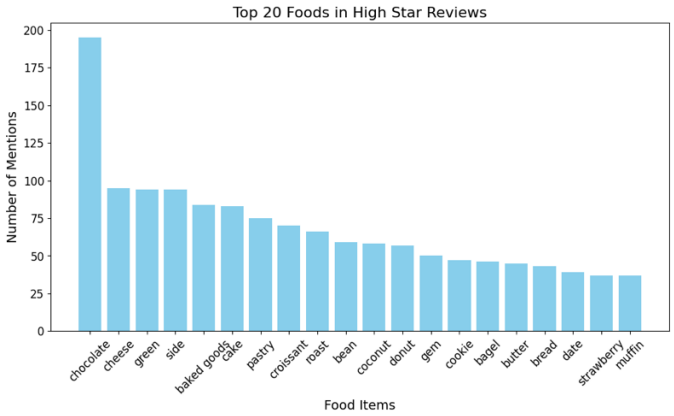
Key Findings And Insights

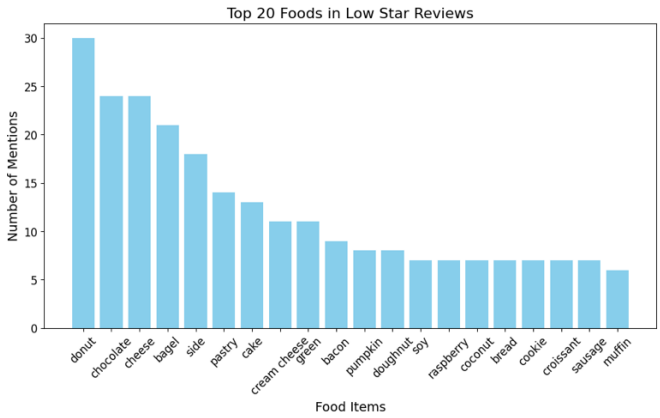
KEY RESULTS

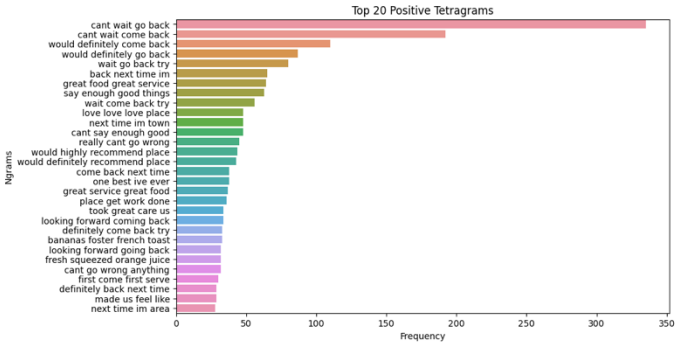
Location Model Insights:
Our conclusive analysis derived from the location model identified three key areas—Soho, Williamsburg, and Midtown—as optimal coffee shop locations. These locations consistently demonstrated strong potential for successful coffee shop establishments, providing a robust foundation for strategic business placement. Soho, Williamsburg, and Midtown emerged as prime coffee shop locations following a comprehensive analysis. Soho's appeal lies in its high foot traffic, affluent demographics, and proximity to upscale stores, offering an ideal post-shopping relaxation spot. Williamsburg's vibrant arts scene and diverse community make it an attractive destination, while Midtown's central position and bustling commercial activity cater to a steady stream of commuters and professionals. However, while promising, these areas may face increased competition, requiring a thorough market evaluation to ensure sustained success in the competitive coffee shop landscape.

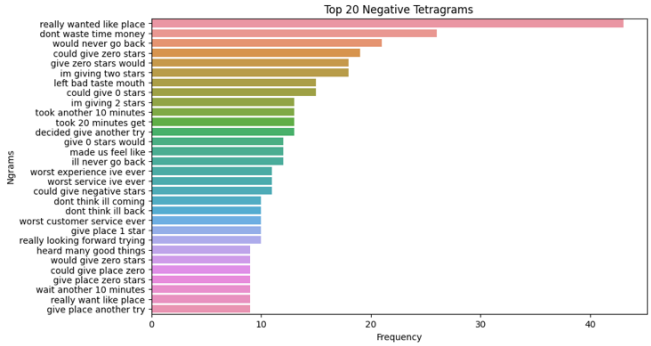
Menu Model Insights:
Furthermore, our investigation into preferred menu items underscored a significant correlation between specific choices and heightened customer satisfaction. Gooey butter cake, Cafe du Monde, and ice cream emerged as favored menu options, emphasizing a distinct inclination towards dessert choices over savory offerings. Complementing these findings, our research on key amenities highlighted crucial preferences among coffee shop-goers: an inclination towards intimate settings, efficient service with minimal wait times, and a diverse range of seating options. These insights collectively construct a comprehensive understanding of the elements that contribute to an optimal coffee shop experience, guiding strategic decisions in location selection and menu design.

Our menu model insights, generated through detailed NLP analysis involving bi, tri, and tetra-grams, provided a nuanced understanding of customer sentiments. Positive trends, including phrases like "get work done," "staff super friendly," and mentions of desirable seating options, were indicative of favorable customer experiences. Contrarily, neutral mentions of "drive-thru" and "almond milk," along with certain references to "Dunkin Donuts," conveyed less emotive or neutral sentiments. Conversely, negative sentiments encapsulated phrases highlighting issues such as "bad customer service" and "long wait times" that significantly impacted customer experiences. By amalgamating insights from our location and menu analyses, we meticulously crafted a blueprint defining the quintessential coffee shop—encompassing the top 20 foods correlated with high and low-star reviews, alongside positive and negative tetra-grams. These findings serve as a definitive guide for curating an appealing and customer-centric coffee shop experience, delineating the essential elements pivotal for success in this domain.









INSIGHTS

- **Teamwork:**
  - This project marked my initial venture into a substantial technical endeavor, and it has been pivotal in my growth in data science. The collaborative environment exposed me to collective problem-solving, fostering innovative solutions that surpassed individual contributions. This collective approach not only elevated the project's quality but also cultivated a strong sense of camaraderie among team members.
- **Time Management:**
  - Balancing school commitments alongside this project demanded rigorous time management. Each team member consistently dedicated over three hours per week, often stretching into late hours to accommodate the project workload. Coordinating schedules for weekly meetings and finding additional work hours aligned with our varying schedules was a significant challenge that we collectively managed to navigate.
- **Asking for Guidance:**
  - Acknowledging the significance of seeking guidance, our team actively engaged with the challenge advisor, teaching assistants, and utilized platforms like Slack for discussions and queries. These interactions provided invaluable insights, contributing significantly to our problem-solving process and overall project success. Regular team meetings, status updates, and openly sharing when we got stuck helped to ensure that everyone was on the same page. I am eager to bring these lessons learned to an internship experience this summer!

Acknowledgements

As our project culminates on this launch day, I'm astounded by the immense learning journey these past 10 weeks have offered. My heartfelt gratitude goes out to everyone at Accenture for their invaluable insights and guidance. I extend my sincerest appreciation to our Challenge Advisors, Timo and Celine, whose support has been instrumental. To my amazing teammates—Caroline, Farhin, Hafsa, Jing, and Felice—your collaborative spirit and dedication have been the bedrock of our achievements. Each one of you has enriched this experience profoundly.

I'd like to express my deepest thanks to our TA, Amber, for her unwavering assistance and course support. Additionally, immense gratitude goes to Break Through Tech, the Cornell Tech AI Program team, and specifically, Erika and Abby, for orchestrating an exceptional program experience. Your efforts have been invaluable in shaping this enriching journey. This experience has been transformative, and I'm excited to see the knowledge we've gained propel us forward. Here's to the future and the countless opportunities it holds!