**Colorado Pizza: Recommendation for Location, Menu Items and Pricing**

Joshua Hiatt | 01/22/2024

1. **Introduction**
   1. **Problem Statement:** To find an underserved Coloradan community with a population of 100,000 or more that is at least 10-15 miles away from any other prominent pizzerias. Also find recommendations for menu items that Coloradans like or make like and lastly a model to predict what prices of those items would most likely do well in the market.
2. **Data Sets, Cleaning and Wrangling**
   1. **Data Set 1:** For this project I wanted to find two data sets. The first data set would contain as many restaurants from as many cities with as many menu items and prices as I could find. From Kaggle.com I was able to find the following data: <https://www.kaggle.com/datasets/cid007/pizza-restaurants-us>.

This data set contains 19 columns and 10,000 observations which come from hundreds of restaurants across the United States along with their menu items, pricing and addresses.

* 1. **Data Set 2:** I wanted city and city demographic information which I was able to get from <https://public.opendatasoft.com/api/records/1.0/search/?dataset=us-cities-demographics&rows=10000>. While this dataset was initially useful for doing some EDA and rough categorizations ultimately it was too limiting as it didn’t contain enough US cities and limited my merged dataset to 5,248 rows.
  2. **Data Sets 3 and 4:** As a result of limited observations from a merged dataframe I found two more datasets one by city, <https://simplemaps.com/static/data/us-cities/1.77/basic/simplemaps_uscities_basicv1.77.zip> which contained population, and population density numbers and another by county, <https://www.kaggle.com/datasets/asaniczka/us-cost-of-living-dataset-3171-counties> which contained cost of living and median family income statistics. These second two data sets were more robust and conserved the majority of my dataframe once merged at 9,002 observations.
  3. As far as cleansing I filtered out observations for prices higher than $50.00 which was well over 95% of my data. I also removed any special characters within the object or string columns, made all strings lowercase, dropped some unneeded columns.

1. **Exploratory Data Analysis and Initial Findings**
   1. My initial EDA was to uncover any trends or counts in the data that may indicate favored items per locality. So I bar charts and some manual groupings of menu item descriptions. For example if a menu item had the word “bianca” I would manually categorize that word into the white pizza category and if it had the word “vegetarian” or “mushroom” I would categorize it into the “veggie” pizza category. From this basic system of categorization I was able to find which Coloradan cities had higher counts of prevalent pizza menu items as displayed below:

**A graph with different colored bars

Description automatically generated**

From this rough system of categorization it becomes very clear which towns have more of a certain type of category of pizza. For example, Denver has a larger count of menu’s containing the word Margherita which would evidence a certain popularity for that item.

* 1. I also determined that if a particular city had a taste for a certain menu item like “sicilian” that there were correlations to other items like “white” meaning that restaurants would offer those two items together in many cases:

A colorful squares with white text

Description automatically generated with medium confidence

1. **Feature Engineering and Modeling**
   1. Once I saw that there were different relationships between cities and menu items and the menu items themselves I ran some rough models to explore the possibility of predicting price from menu item individual words and was able to achieve an R2 score of 0.69 on my initial smaller dataset and from only the menu item description. This prompted a deeper dive into more creative features and advanced models.
   2. Using the cleansed and merged data sets 1,3 and 4 I programed an Unsupervised categorization algorithm to create a new feature which would look at the ‘menu\_item’ column and based on prevalent words it found group all rows into a categorization ranging from 2-50. Words like ‘white’ and ‘buffalo’ were clumped into clusters that were optimized to 48 using a silhouette method:

**A colorful dots on a white background

Description automatically generated**

**A graph with a line

Description automatically generated**

* 1. Armed with this new feature that sorted menu items into like worded clusters I sought to create more categorical features and one hot encoded all of my categorical columns.
  2. I then applied a StandardScaler to my demographic information such as location (longitude, latitude), population, housing cost etc.
  3. From there the data was bootstrapped and then split into training and test sets in order to predict the menu item price or the target column ‘menus\_amountMax’.

1. **Models and Future Study**
   1. After running three manual ensemble models on my data and one automated feature engineered model I determined that the AdaBoost Model tuned to 400 decision trees and a learning rate of 0.1 was the best model to predict price at a test Mean Absolute Error of +-$1.86 or a 3.72% error rate on a price range of $0.00 to $50.00. This means that given a restaurant, a menu item description and the city for which the restaurant is located that my model can predict within $1.86 the price of a menu item between $0.00 to $50.00 price range. It also had a strong pearson correlation or R2 of 0.65 between the actual and predicted prices:

**A graph of values with blue dots

Description automatically generated**

* 1. Interestingly, while menu item words and categories were decent predictors of price the strongest predictor was healthcare cost at 20.3% of the prediction rate. This finding should prompt more study in the relationship between food descriptions, food cost and healthcare costs:

**A screenshot of a computer code

Description automatically generated**

* 1. Lastly, location had to be considered. For that I plotted the cities of Colorado with populations of 100,000 next to the pizza restaurant plots to determine an underserved Coloradan market. The most notable larger towns with sparse pizzerias were Greeley, Grand Junction and Pueblo. This is most likely due to the limitations of the pizza restaurant data but does provide some important insights to where perhaps there are fewer pizza chains as the initial EDA uncovered as well:

**A map of a restaurant

Description automatically generated**

1. **Summary and Recommendations:** 
   1. Choose either Greeley, Grand Junction and Pueblo for the site of the restaurant.
   2. Then choose 2 to 3 popular Coloradan menu items of your choice:

A bar graph with different colored bars

Description automatically generated

* 1. Choose a few more highly correlated items to the initial selections:

A colorful squares with white text

Description automatically generated with medium confidence

* 1. Model the menu item prices for your menu selection.
  2. Open a restaurant in the selected town with the selected menu items, correlated menu items and modelled prices. Here are the model’s scores:

R^2 Train: 0.9580002943937314

R^2 Test: 0.6511567335262737

MAE Train: 0.4300178092953398

MAE Test: 1.8608886277939825