1. Introduction

Background

CAVA is a privately held chain of fast-casual Mediterranean restaurants with 105 locations in California, Colorado, Connecticut, District of Columbia, Massachusetts, Maryland, North Carolina, New Jersey, New York, Pennsylvania, Tennessee, Texas, and Virginia. Since its establishment in 2006, CAVA has enjoyed considerable success as a competitor in the fast-casual, “bowl-centric” dining market (alongside competitors like Chipotle, Chopt, and Panda Express), leading to expansion to the states listed above and acquisition of the also-popular Mediterranean chain Zoës Kitchen. As such, CAVA has established itself as a strong force in the casual Mediterranean dining market.

Problem

For this project, I am pretending that representatives of CAVA have approached me to create a ML model that can predict profitable new locations for the chain. They inform me that they have ample capital to invest in new locations, but don’t want to expand to new locations unless they are confident doing so will yield a good return on investment (ROI), leading to increased profits with minimized risk for failed launches and subsequent financial losses. They also tell me that they are confident that they have already expanded to every viable zip code in states that have CAVA, so they only want to expand to new states and are open to any area of the country as long as the viabilities of proposed locations are supported by a strong model. After discussing my uncertainty as to whether ML can provide strong predictions and proposing that they should investigate what the best predictors are before fully investing in a model, they agree to fund a preliminary investigation of whether the success of a CAVA in a new location can be predicted by the type and number of nearby venues. I agree to try to develop a ML model that attempts to predict which of the most populous zip codes in states across the US that do not already have CAVA would be profitable locations for expansion based on nearby types of venues.

Interest & Value

Hopefully, this model will serve as a fruitful starting point for the development of models that enable CAVA to maximize return on investment during expansion to new areas in the US by minimizing the chance that they will expand to unsuccessful locations. The creation of such models would minimize the risk CAVA faces during expansion, leading to greater financial security during pursuit of increased profits.

2. Data Acquisition

Summary of Needed Data

Training a ML model to predict whether a zip code will be a good location for a CAVA based on nearby venues requires the following information:

- Zip codes of CAVA locations and the count of every venue type around* each of those zip codes

  (These data points serve as positive outcomes when training models)

- The 30 most populous zip codes without CAVA in states with CAVA** and the count of every venue type around* each of those zip codes

  (These data points serve as negative outcomes when training models)

Applying a trained model to predict whether zip codes in states without CAVA will be good locations for CAVA based on nearby venues requires the following information:

- The 10 most populous zip codes in states without CAVA and the count of every venue type around* each of those zip codes

* I used a 2-mile radius for my venue search calls. My logic is that success is likely only predicted by venues that are within a short travel time of a CAVA, since those are the venues that will affect traffic to the CAVA location

** I assume that CAVA has already expanded to every viable zip code in states where it is present (likely not true in reality, but ok for this exercise). This means that zip codes without CAVA in states with CAVA are not viable and thus serve as good negative outcomes for model training.

Zip Code Acquisition

Zip codes of CAVA locations were obtained by scraping the CAVA website: (https://cava.com/locations).

The 30 most populous zip codes without CAVA in states with CAVA were obtained by scraping demographic data from the internet (e.g., https://www.newjersey-demographics.com/zip_codes_by_population) and then cross referencing a list of zip codes with CAVA to eliminate overlap.

The 10 most populous zip codes in states without CAVA were obtained by scraping demographic data from the internet (e.g., https://www.newjersey-demographics.com/zip_codes_by_population).

Zip Code Geocoding

Obtaining venue data for zip codes using calls to the Foursquare API requires obtaining coordinates for those zip codes. Coordinates were obtained using the OpenCage Geocoding API (https://opencagedata.com/) with queries consisting of the zip code and the state of that zip code.

Venue Data Acquisition

Venue data around each zip code was obtained using the 'explore' endpoint of the Foursquare API. As noted earlier, a 2-mile (~3200 meter) search radius was used, with the maximum return limit allowed.

Data Cleaning

Zip codes without returned venues were eliminated from both the training and prediction datasets. Additionally, CAVA venues returned by the Foursquare "explore" endpoint were eliminated from the training dataset so that they would not impact the counts of Mediterranean restaurants within a zip code.

3. Methodology

Exploratory Data Analysis

The only exploratory analysis performed was to find and compare the ten most common venue types, on average, in zip codes with CAVA and in zip codes without CAVA to get a sense of what venue types were truly most common and/or most frequently returned by Foursquare's "explore" endpoint. The same analysis was performed for the prediction dataset (zip codes in states without CAVA).

Feature Selection

After the training dataset was constructed as described above, feature selection was performed on the training dataset to optimize prediction.

First, t-tests were performed to examine whether each venue type in the training dataset differed significantly in number between zip codes with CAVA and zip codes without CAVA, and venue types that did not differ significantly were dropped from the training dataset.

Next, collinear venue types were eliminated from the training dataset by iteratively calculating the VIF score for each remaining venue type in the training dataset, eliminating the venue type with the highest VIF score above or equal to 10, and then repeating that process until all remaining venue types had VIF scores below 10.

Finally, the best combination of remaining venue types for prediction was then selected using recursive feature elimination (RFE) with a logistic regression model to find the combination with the best accuracy where all venue types in that combination are statistically significant (some combinations with higher accuracy were discarded because not all venue types in those combinations were statistically significant as assessed by logistic regression, meaning the accuracy scores reported for those combinations are especially untrustworthy).

*Notably, further feature selection was performed while training a support-vector classifier (SVC) model, since another round of RFE with the SVC model found a subset of the features selected by the initial round of RFE that performed better for that model.

Model Fitting and Optimization

With features optimally selected as assessed by logistic regression complete, the feature-selected training dataset was used to train a logistic regression model, a SVC, and a k-nearest neighbor classifier using sklearn packages. Every possible k was assessed while optimizing the k-nearest neighbors classifier.

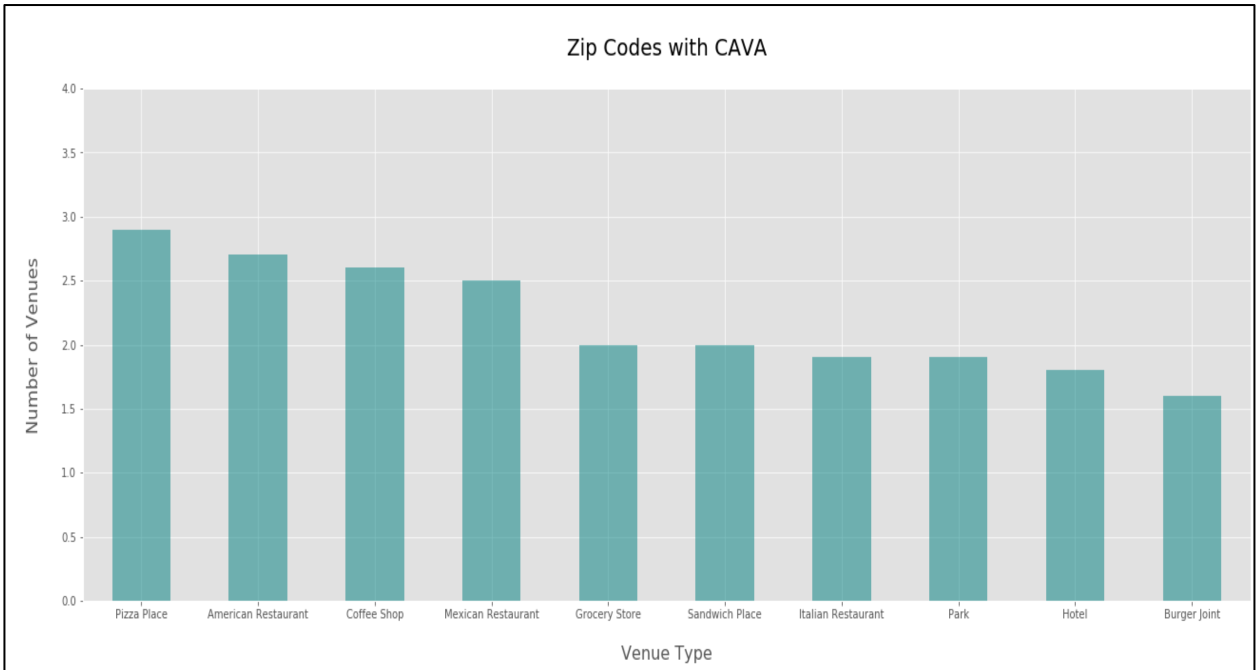
4. Results

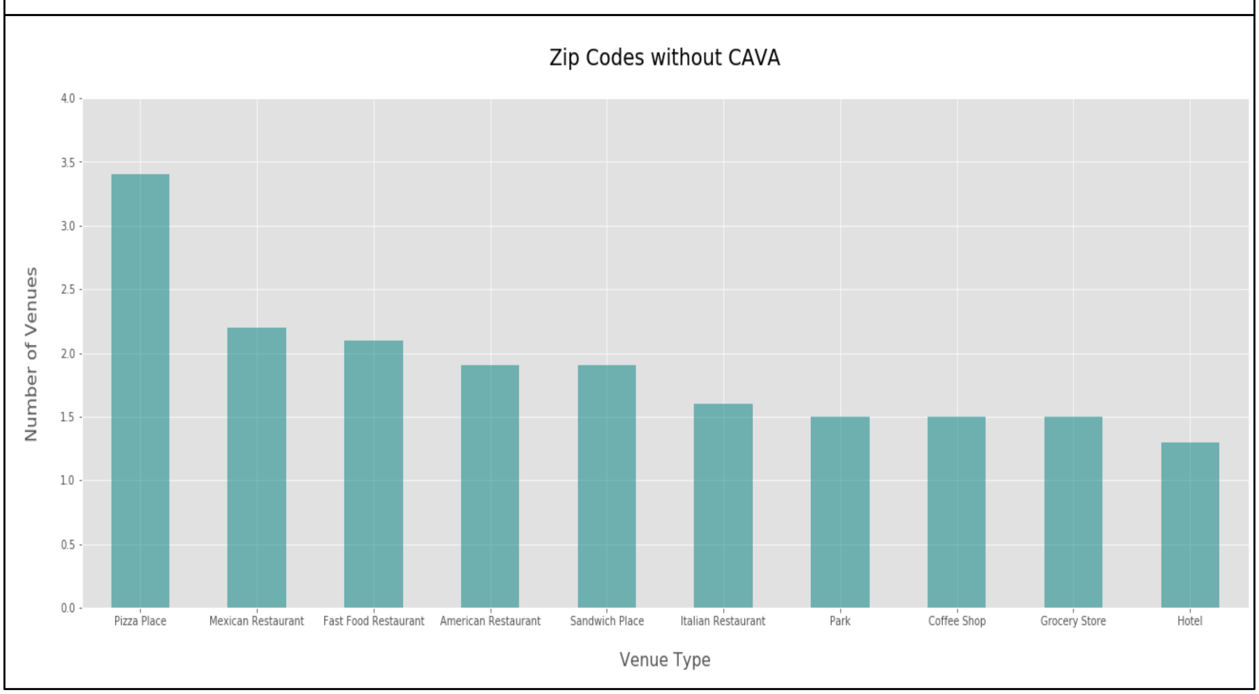
Initial Data Acquisition

For the training dataset, 101 zip codes containing CAVA and 345 zip codes without CAVA in states with CAVA were obtained. When queried with those zip codes, the Foursquare API returned a total of 32,052 venues covering all but 1 zip code (07104). After coding of those venues by venue type and grouping by zip code, the training dataset contained 502 venue types across 445 zip codes.

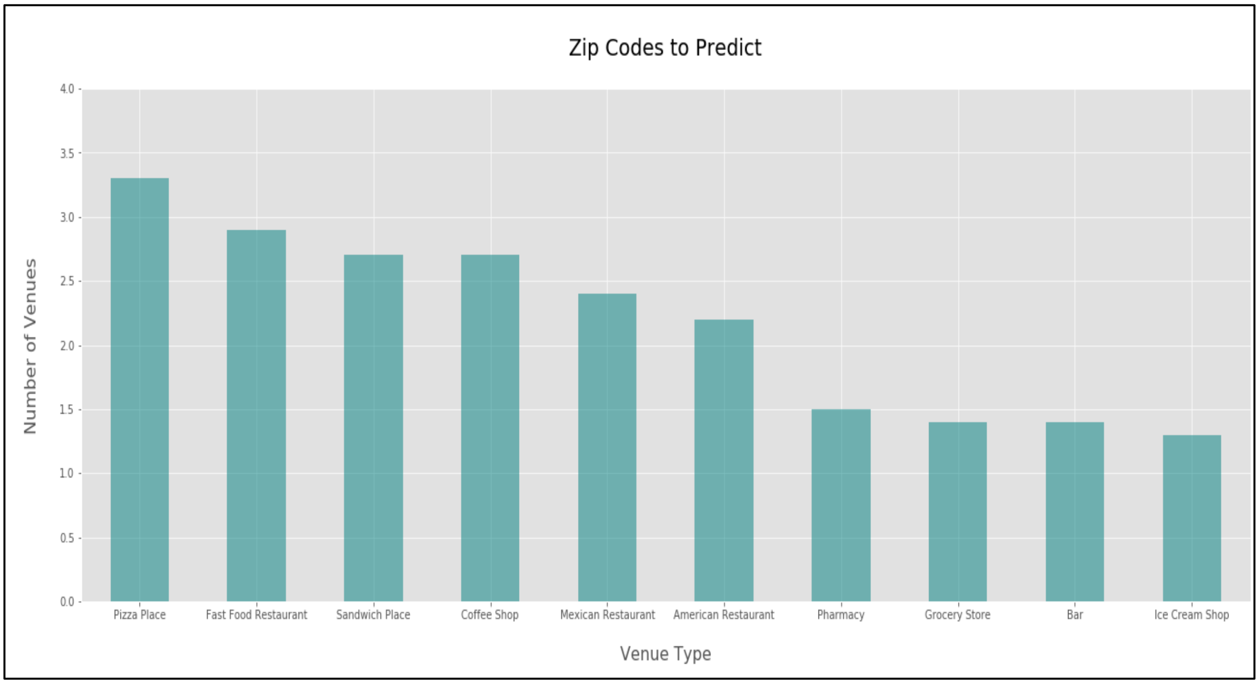
Exploratory Data Analysis

The bar graphs below show the ten most common venue types, on average, among zip codes with CAVA and among zip codes without CAVA:





The bar graph below shows the ten most common venue types, on average, among the most populous zip codes in states without CAVA:



<u>Feature Selection</u>

T-testing performed as the first step of feature selection revealed that 376 venue types do not differ significantly in frequency, on average, among zip codes with CAVA and zip codes without CAVA. Elimination of those venue types from the training dataset left 126 venue types remaining in the dataset.

Collinearity testing performed as the second step of feature selection revealed that 2 of the 126 venue types remaining after t-testing had VIF values >= 10. Elimination of those venue types from the training dataset left 124 venue types remaining in the dataset.

RFE based on logistic regression accuracy performed as the third step of feature selection determined that the following nine venue types provide the most accurate classification while maintaining statistical significance:

Venue type	Coefficient	Lower 95% CI	Upper 95% CI
Smoothie shop	0.2536	0.0295	0.4777
Garden center	0.4377	0.0843	0.7910
Sushi restaurant	0.2596	0.0101	0.5091
Shopping mal	0.4169	0.1825	0.6513
Salad place	0.5737	0.2529	0.8945
Pharmacy	-0.2721	-0.5173	-0.0270
Discount store	-2616	-0.5114	-0.0117
Gourmet shop	0.5115	0.1634	0.8596
Coffee shop	0.2568	0.0143	0.4992

Training and optimization of a logistic regression model, SVC model, and knn model follwing the feature selection completed as described above (with additional selection for the SVC model) resulted in the following performances:

Model	Avg. Accuracy	Avg. PPV*	PPV Lower 95% CI	PPV Upper 95% CI	Avg. Sensitivity**	Sensitivity Lower 95% CI	Sensitivity Upper 95% CI
Log Reg	0.80	0.74	0.61	0.87	0.20	0.14	0.26
SVC	0.84	0.74	0.64	0.84	0.49	0.39	0.59
KNN	0.75	0.52	0.39	0.65	0.45	0.32	0.58

*PPV = positive predictive value (the number of positive predictions that are correct)
** Sensitivity = the number of good locations that are predicted

The logistic regression model used the following venue types:
- Smoothie shop
- Garden center
- Sushi restaurant
- Shopping mall
- Salad place
- Pharmacy
- Discount store
- Gourmet shop
- Coffee shop

The SVC model used the following venue types:
- Smoothie shop
- Garden center
- Shopping mall
- Salad place
- Discount store
- Gourmet shop
- Coffee shop

The k-nearest neighbors classifier model used 5 neighbors.

Prediction Summary

Applying the optimized SVC model (which appears to provide the best predictions given a comparable PPV and a better sensitivity than the logistic regression model) to a prediction dataset containing venue data for the 10 most populous zip codes in every state without CAVA (380 zip codes total) resulted in classifying the following 70 zip codes as good CAVA locations:

Prediction	Zip Code	City
1	99801	Juneau, AK

1	84043	Lehi, UT
1	68801	Grand Island, NE
1	70003	Metairie, LA
1	48103	Ann Arbor, MI
1	36830	Auburn, AL
1	83642	Meridian, ID
1	83646	Meridian, ID
1	83704	Boise, ID
1	83709	Boise, ID
1	85281	Tempe, AZ
1	68516	Lincoln, NE
1	30041	Cumming, GA
1	89052	Henderson, NV
1	89123	Las Vegas, NV
1	96706	Ewa Beach, HI
1	96734	Kailua, HI
1	96744	Kaneohe, HI
1	96789	Mililani, HI
1	96797	Waipahu, HI
1	48823	East Lansing, MI
1	68116	Omaha, NE
1	96817	Honolulu, HI
1	58102	Fargo, ND
1	59601	Helena, MT
1	59901	Kalispell, MT
1	59102	Billings, MT
1	59101	Billings, MT
1	60618	Chicago, IL
1	58201	Grand Forks, ND
1	58104	Fargo, ND
1	58103	Fargo, ND
1	57701	Rapid City, SD
1	53704	Madison, WI
1	60639	Chicago, IL
1	60647	Chicago, IL
1	57105	Sioux Falls, SD
1	65807	Springfield, MO
1	55106	Saint Paul, MN
1	55104	Saint Paul, MN
1	55044	Lakeville, MN
1	53711	Madison, WI

1	96816	Honolulu, HI
1	59715	Bozeman, MT
1	96818	Honolulu, HI
1	99577	Eagle River, AK
1	3820	Dover, NH
1	5401	Burlington, VT
1	98012	Bothell, WA
1	4330	Augusta, ME
1	98115	Seattle, WA
1	4103	Portland, ME
1	98052	Redmond, WA
1	5403	South Burlington, VT
1	4210	Auburn, ME
1	4240	Lewiston, ME
1	97301	Salem, OR
1	99504	Anchorage, AK
1	99507	Anchorage, AK
1	99515	Anchorage, AK
1	98208	Everett, WA
1	97223	Portland, OR
1	2908	Providence, RI
1	96819	Honolulu, HI
1	2907	Providence, RI
1	97124	Hillsboro, OR
1	97045	Oregon City, OR
1	99709	Fairbanks, AK
1	97006	Beaverton, OR
1	2909	Providence, RI

5. Discussion

Model Performance

Overall, the SVC model (which performed best among the models tested) provided good positive predictive value with mediocre sensitivity. As such, the CAVA representatives can be cautiously confident that any recommended zip code will indeed be a good CAVA location, but should be aware that the model will likely miss many locations that are also good options. Of course, there is certainly room for improvement of the model, especially regarding its sensitivity.

Model Limitations

In addition to sub-optimal PPV and sensitivity, the model is limited by its exclusion of many other factors that likely determine the success of a restaurant in a new location, including perhaps: brand awareness in new locations, local and state tax laws and corporate incentives, zoning laws, population density and more.

6. Conclusion

Future Improvements to Model

Keeping with the original purpose of this exercise—to begin to build an effective ML model for identifying successful new locations for CAVA—the model developed here seems to suggest that nearby venue types would be a good predictor to include in a final model. I suspect that combining nearby venue types with other predictors (like those mentioned in the "Model Limitations" section), could lead to a very effective model.

Thus, this exercise was a success.