PREDICTING SUCCESSFUL NEW CAVA LOCATIONS

Mitchell Lee

Coursera – Data Science Capstone

08 / 16 / 20

BACKGROUND

- CAVA = fast-casual Mediterranean restaurant chain
- Locations: CA, CO, CT, DC, MA, MD, NC, NJ, NY, PA, TN, TX, VA
- Successful competitor of Chipotle, Chopt, Panda Express, etc.

BUSINESS PROBLEM

- Need = identify zip codes in new states that will be successful locations
  - Assume already expanded to viable zip codes in states where located
- Commission creation of ML model
  - Proposal = start by exploring whether venues in new zip codes are good predictors of success

INTEREST & VALUE

- Identify good investment opportunities
- Minimize risk & maximize probability of return on investment

ML MODEL PROPOSAL

- Model-training dataset:
  - Zip codes of CAVAs and counts of surrounding venue types
  - 30 most-populous zip codes <u>without</u> CAVAs in states with CAVA and counts of surrounding venue types
- Prediction dataset:
  - 10 most-populous zip codes in states <u>without</u> CAVA
- Models to test:
  - Logistic regression (use for feature selection)
  - Support vector classifier
  - K-nearest neighbors classifier

DATA SOURCES

- CAVA Locations:
  - CAVA website (https://cava.com/locations)
- Zip Codes
  - Demographics websites (e.g., https://www.newjersey-demographics.com/zipcodesbypopulation)
- Geocoding
  - OpenCage Geocoding API (https://opencagedata.com/)
- Venues
  - Foursquare API "explore" endpoint (https://developer.foursquare.com/docs/api-reference/venues/explore/)

METHODOLOGY

- Exploratory Analysis
  - Visualize 10 most common venue types in CAVA and non-CAVA zip codes
  - Visualize 10 most common venue types is target zip codes
- Feature Selection
  - Drop statistically insignificant variables (t-tests, p-values > 0.05)
  - Drop collinear variables (VIF >= 10)
  - Recursive feature elimination (select best accuracy combination maintaining significance)
- Model training
  - Logistic Regression Model (using predictors selected by RFE)
  - SVC (testing subsets of predictors selected by RFE)
  - KNN (testing all possible k (max = number of zip codes in training dataset))

RESULTS

- Training dataset:
  - 101 zip codes with CAVA
  - 345 zip codes without CAVA
  - 32,505 total venues returned,
    - 502 types (e.g. Mexican restaurant, shopping mall, nail salon, etc.)

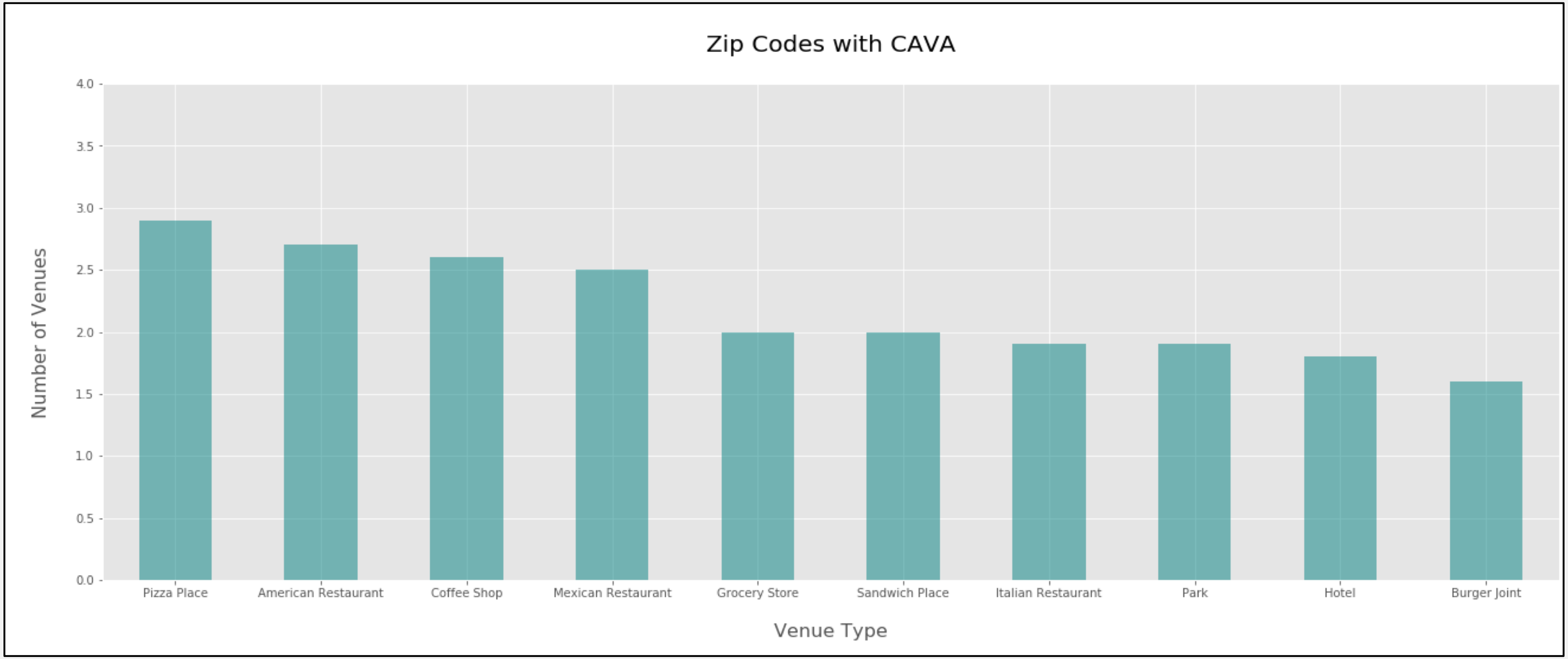
```
Dataframe shape: (32052, 7)
```

[24]:

	Has CAVA?	City	Zip Code	Latitude	Longitude	Venue	Venue Category
0	1	Anaheim, CA	92808	33.866069	-117.74321	Bodhi Leaf Coffee Traders	Coffee Shop
1	1	Anaheim, CA	92808	33.866069	-117.74321	Chipotle Mexican Grill	Mexican Restaurant
2	1	Anaheim, CA	92808	33.866069	-117.74321	Rosine's Mediterranean Grill	Mediterranean Restaurant
3	1	Anaheim, CA	92808	33.866069	-117.74321	Wood Ranch BBQ & Grill	BBQ Joint
4	1	Anaheim, CA	92808	33.866069	-117.74321	Sprouts Farmers Market	Grocery Store

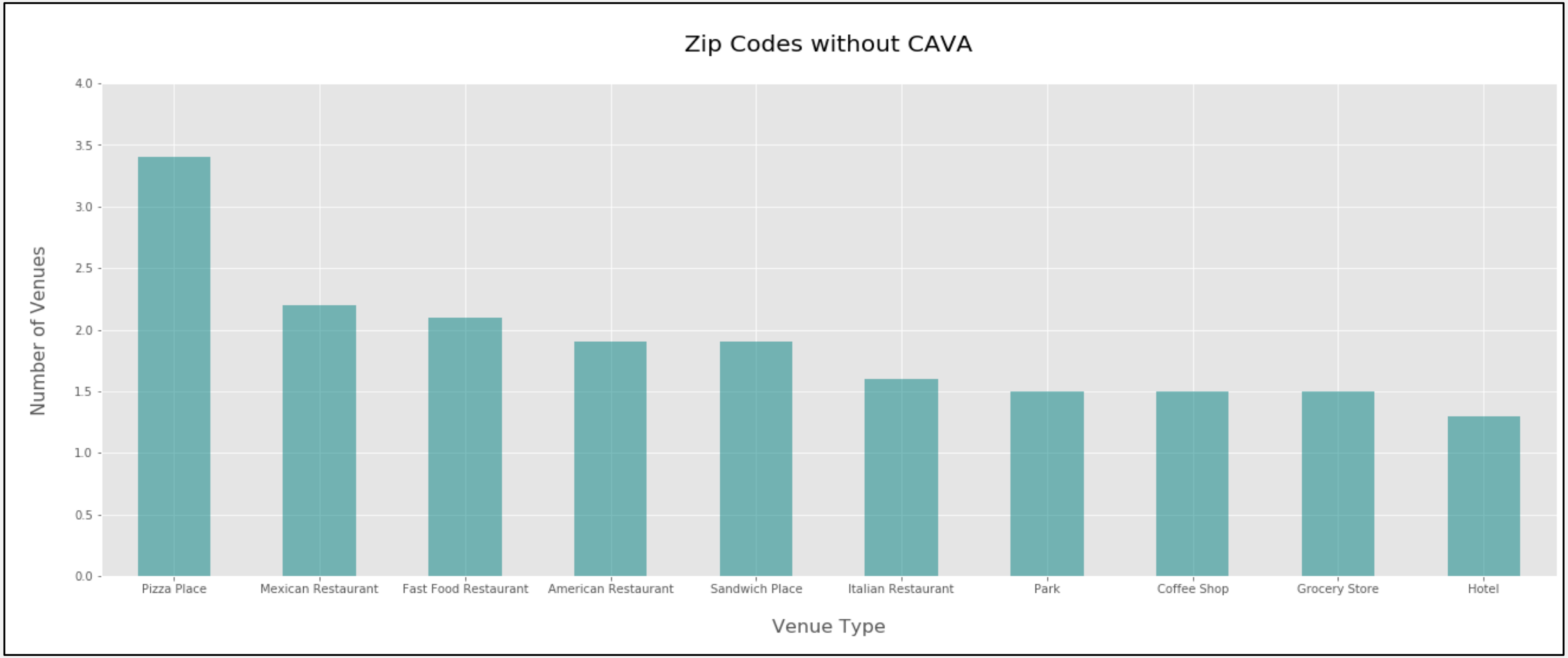
RESULTS

- Exploratory Data Analysis:



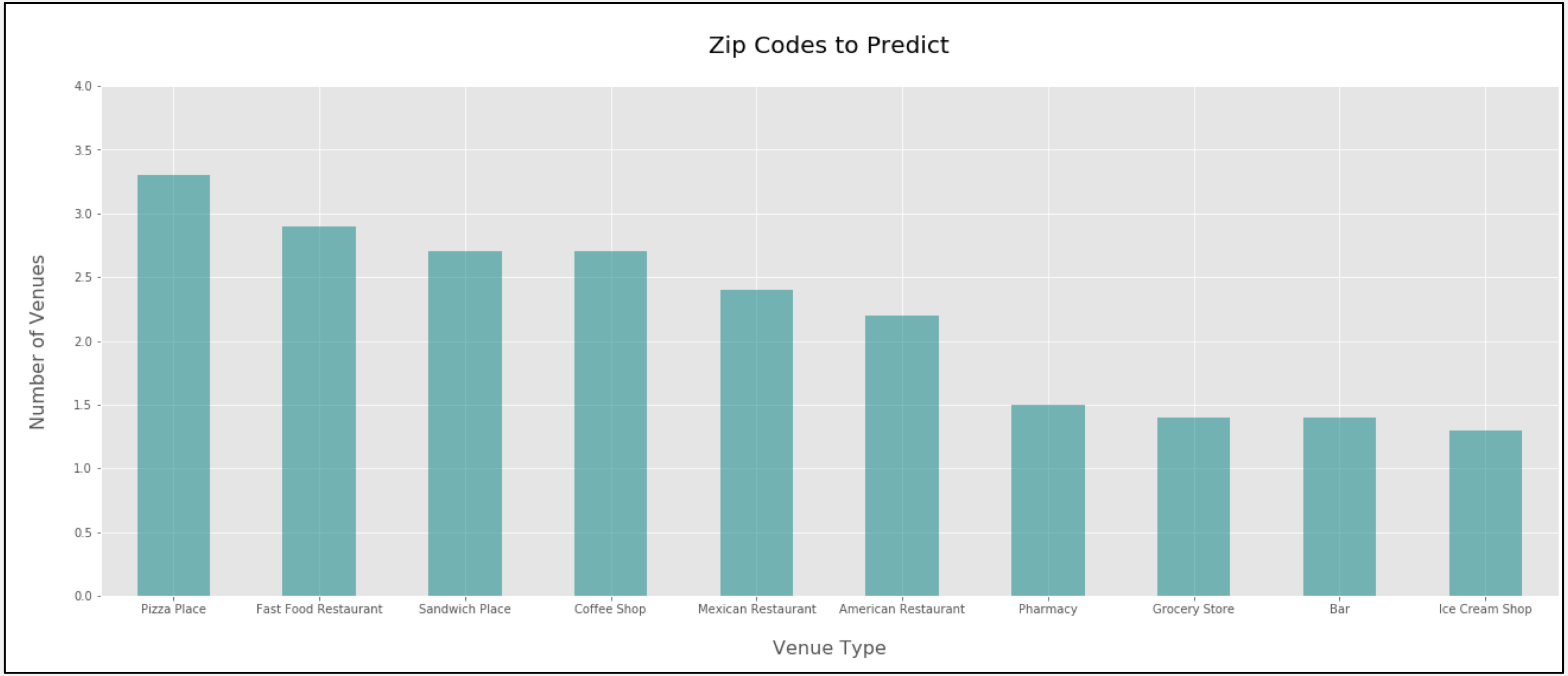
RESULTS

- Exploratory Data Analysis:



RESULTS

- Exploratory Data Analysis:



RESULTS

- Features Selection
  - 502 starting venue types
  - 376 venue types not significantly different between CAVA and non-CAVA zip codes (126 venue types remaining)
  - 2 collinear venue types (VIF > 10) (124 venue types remaining)
  - RFE selects 9 venue types as optimal set of predictors based on logistic regression
    - Largest number of venue types where all venues within are statistically significant

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

RESULTS

- Model training

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
SCV*	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

* Best model (by accuracy) used 7 predictors (subset of predictors identified by initial RFE)

** Best model (by accuracy) used 5 neighbors

RESULTS

- Prediction Summary (1 = Good CAVA location)

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

RESULTS

- Prediction Summary (1 = Good CAVA location)

Prediction	Zip Code	City
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

RESULTS

- Prediction Summary (1 = Good CAVA location)

Prediction	Zip Code	City
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	03820	Dover, NH
1	05401	Burlington, VT

RESULTS

- Prediction Summary (1 = Good CAVA location)

Prediction	Zip Code	City
1	98012	Bothell, WA
1	04330	Augusta, ME
1	98115	Seattle, WA
1	04103	Portland, ME
1	98052	Redmond, WA
1	05403	South Burlington, VT
1	04210	Auburn, ME
1	04240	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	02908	Providence, RI
1	96819	Honolulu, HI

RESULTS

- Prediction Summary (1 = Good CAVA location)

Prediction	Zip Code	City
1	02907	Providence, RI
1	97124	Hillsboro, OR
1	97045	Oregon City, OR
1	99709	Fairbanks, AK
1	97006	Beaverton, OR
1	02909	Providence, RI

DISCUSSION

- Model Performance
  - Good positive-predictive value (PPV)
    - Confident that positive predictions are correct (good return on investment)
  - Mediocre sensitivity
    - Model likely misses many good locations (missed investment opportunities)
- Model Limitations
  - Positive-predictive value and sensitivity could be improved
  - Model ignores other important variables
    - Cost of business
    - Tax laws and business incentives
    - Brand awareness (and more)

CONCLUSION

- Nearby venues = decent predictor
  - Potential for further optimization
- Could be combined with other predictors to improve performance
  - Should be assessed individually before incorporation
- Overall exercise was successful
  - Showed the nearby venues are decent predictor
- CAVA should continue investment in model creation