Whataburger Prediction for Houston Neighborhoods

Final Capstone Project

Introduction

- Whataburger is a popular fast food chain in Texas. There are quite a few Whataburger restaurants in Houston, TX. These Whataburger restaurants are within the 88 super neighborhoods (SNB) of the City of Houston.The management decide to open new restaurants.
- The problem is which neighborhood should be chosen for the next restaurants?
- This project is to build a model to predict which neighborhood has Whataburger restaurants in Houston.

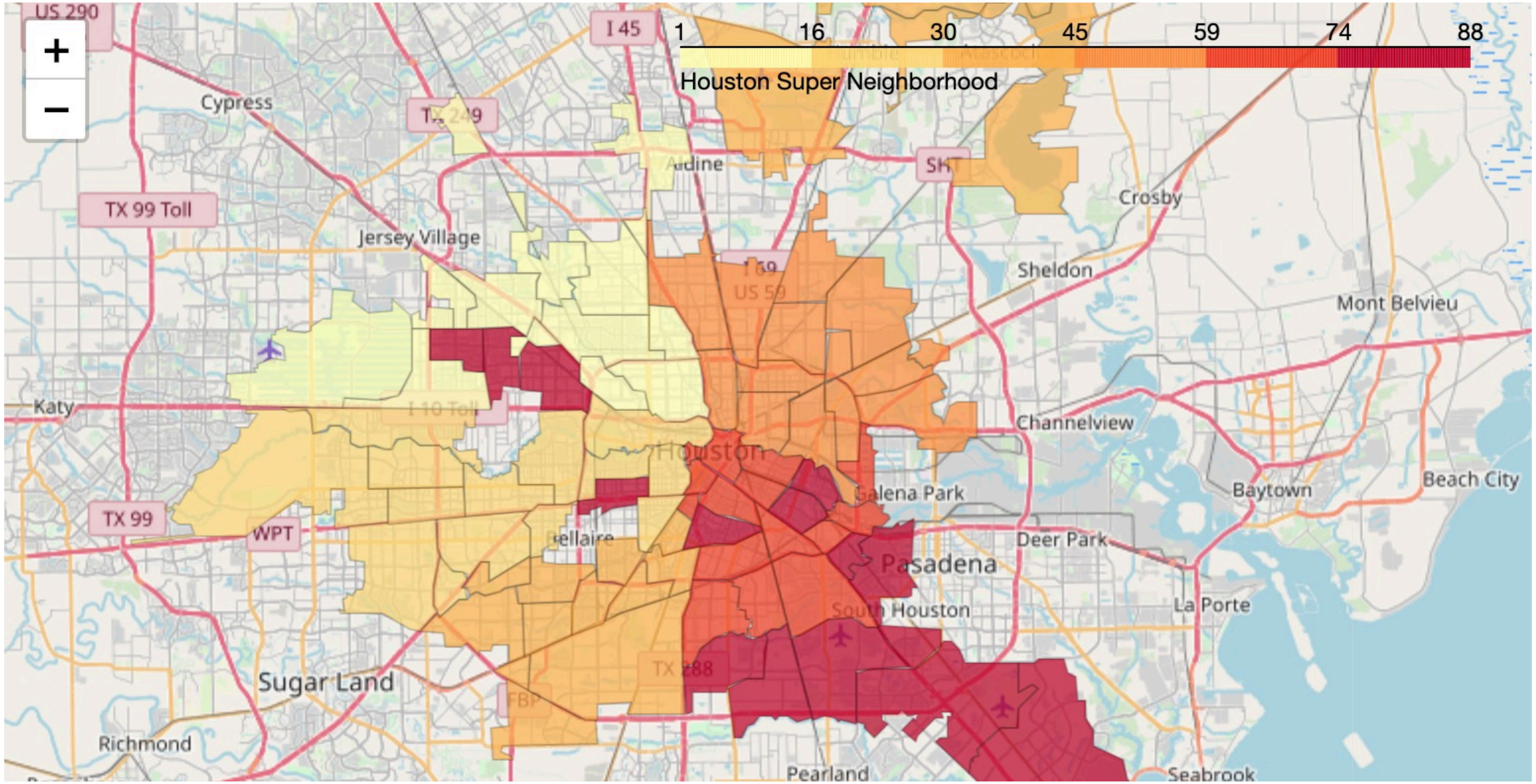
Data Collection and Wrangling

Data Source

- The Houston neighborhood geospatial information can be obtained from the city council website, which links to the ArcGis website.
- Neighborhood data is from https://cohgis-mycity.opendata.arcgis.com/datasets/super-neighborhoods?selectedAttribute=RECOGNITIO
- The shapefile data has been converted to GEO JSON data and saved in local drive for the project use.

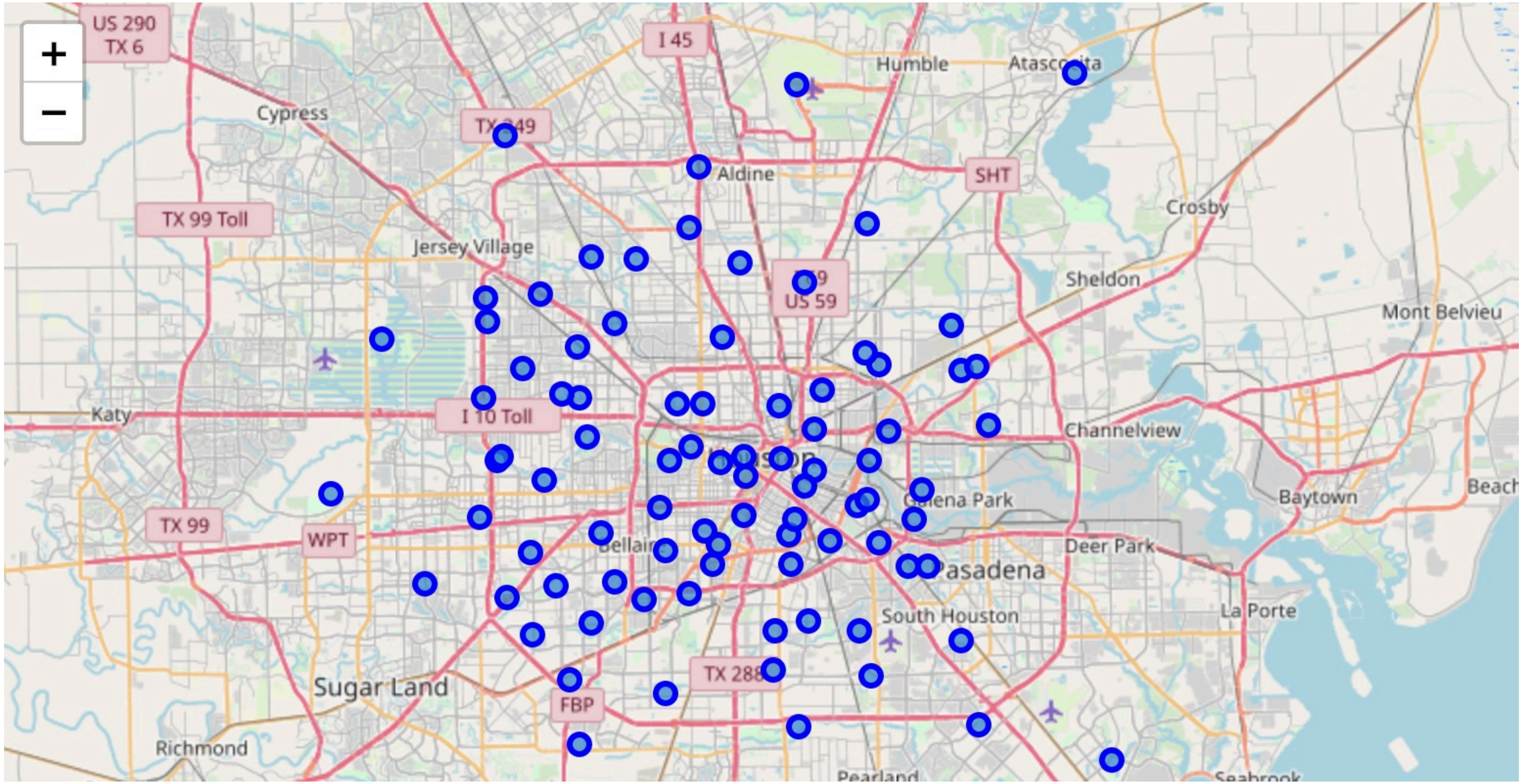
Neighborhood Boundaries

- The neighborhoods are marked on Houston map. The color scale is neighborhood ID.



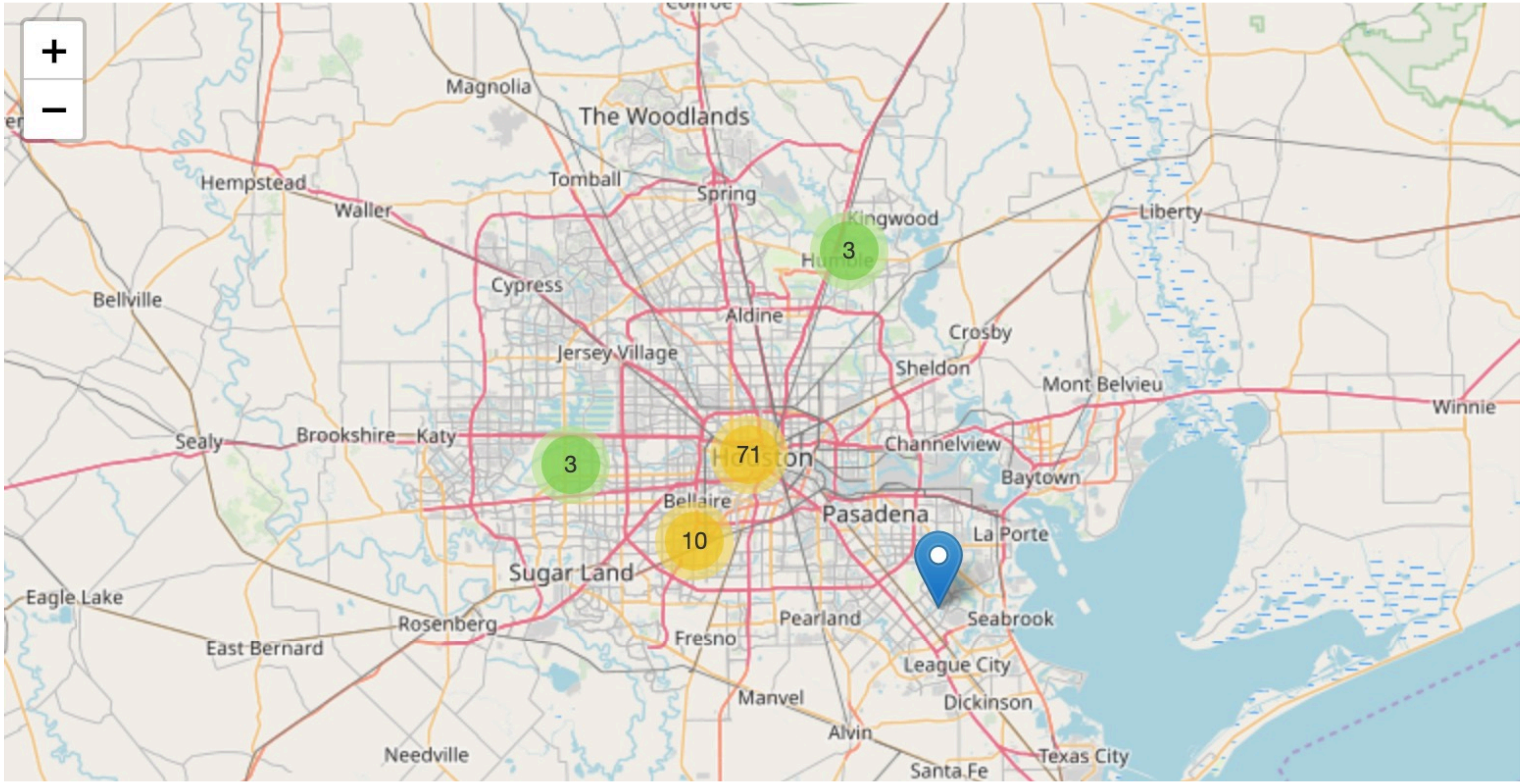
Neighborhood Centroids

- Simple calculations are used to get neighborhood center coordinates from the JSON file.



Neighborhood Clustering

- When the map scale is changed, neighborhood points will be automatically clustered.

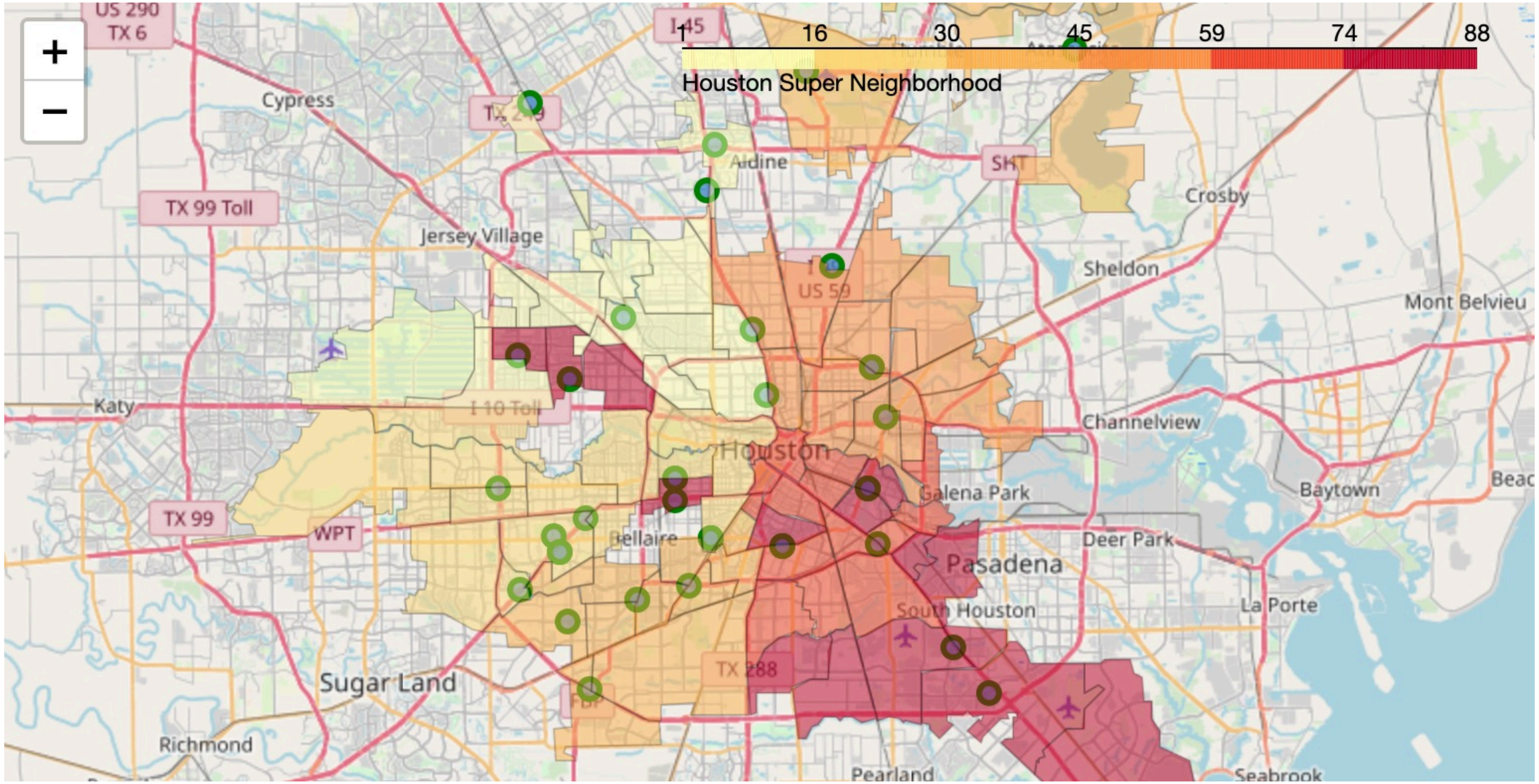


Data Acquisition

- Foursquare is used to get nearby venues for the neighborhoods. Besides venues type and numbers, the data also shows if there's an existing Whataburger in the neighborhood. The features of different venue category characterizes each neighborhood. The features can be used to classify each neighborhood.

- The existing Whataburgers are marked on the map.

Existing Whataburgers



Data Pre-Processing

- The data from Foursquare is input into a dataframe. There is a total of 5022 records from the Foursquare. There is a total of 327 distinct categories.

- The data is grouped by neighborhoods with 327 columns for different categories. The frequencies of each categories are calculated and stored in the dataframe. The frequencies are the features of neighborhoods and will be used in the following modeling. There's an additional column to show if there's Whataburger. Integer number '1' indicates there is an existing Whataburger. '0' means not existing.

Methodology

Data Split and Classification Methods

- 80% of the data will be used for training and 20% will be used for testing.
- The following Classification algorithms will be used in the current project:
  - K Nearest Neighbor(KNN)
  - Decision Tree
  - Support Vector Machine (SVM)
  - Logistic Regression
- The train data was first used. Different "Ks" were tested for the KNN method. Different depths of Decision Tree were tested. Different kernels were tested for Support Vector Machine (SVM). Different solvers and inverses of regularization strength (C parameter) were tested for Logistic Regression.

Evaluation Methods

- Accuracy was used to check the train data modeling. By tuning the above mentioned parameters, the best accuracies were achieved for the train data.

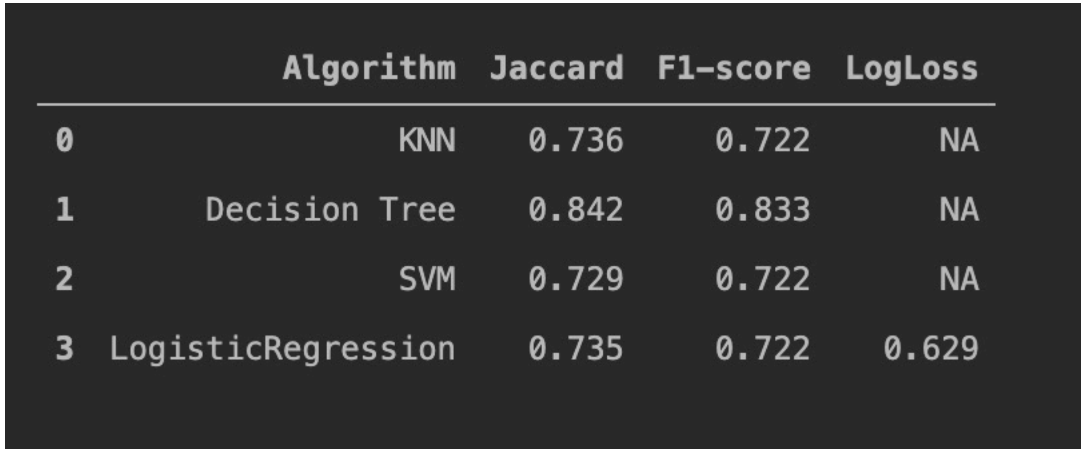
- After this, test data was used. The test data was, 'out of sample', exclusive from the train data. F1-score, Jaccard, and LogLoss were used to evaluate the test data modeling results.

- This is an iterative process. If the evaluation metrics are not good, the parameters in training section will tuned until the best results are achieved.

Results

Comparison of Different Methods

After the itineration, the best results are obtained and presented below:

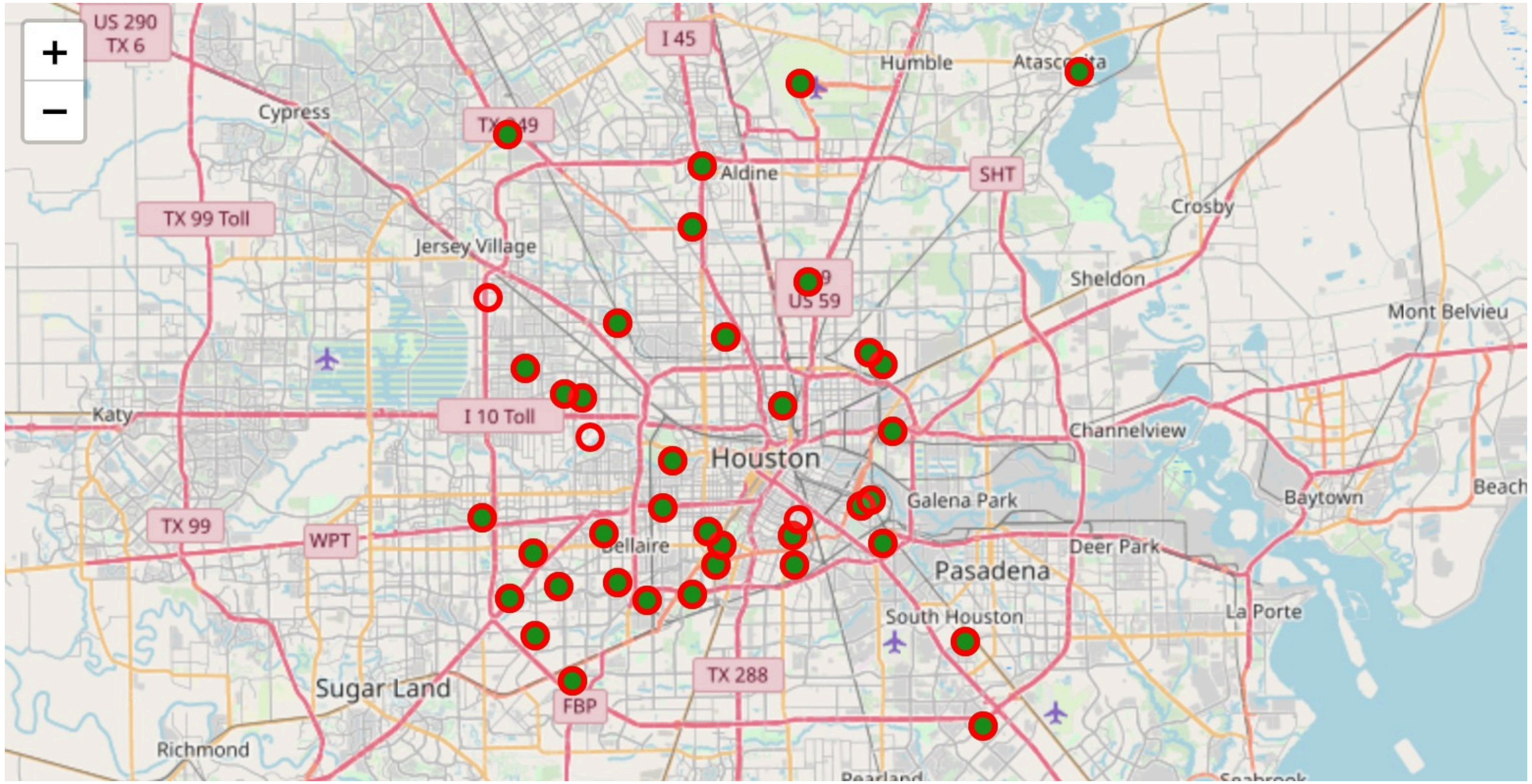


	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.736	0.722	NA
1	Decision Tree	0.842	0.833	NA
2	SVM	0.729	0.722	NA
3	LogisticRegression	0.735	0.722	0.629

Decision Tree gives the best results; therefore, it is selected.

Comparison of Actual and Predicted

The decision tree was used to predict the whole dataset. The comparison of the results is presented below. Green dots are actual neighborhoods that have Whataburger. Red circles are predicted results. It shows a great agreement.



Discussion

Discussion

- The data from Foursquare is mainly about the individual venue locations and categories. Other information, such as population, income, traffic, etc., can provide more features of neighborhood. These features can help classify neighborhood.

- Only limited parameters have been tuned for each algorithm. Other parameters such as probability threshold can be further tuned to get even better results.

Conclusion

Conclusion

- A classification project was conducted to find the best method to predict the existence of Whataburger restaurant in Houston. A total of 5022 records were collected from Foursquare for the 88 super neighborhood. 327 distinct categories were analyzed.

- Four classification methods have been tested.

- The results show Decision Tree is the best method for the current project.