2020

Battle of Neighborhoods

COURSERA CAPSTONE PROJECT

ARUN NATARAJAN

Table of Contents

1. Background 3
2. Business Problem 3
2.1. Importance of improving missing data of neighborhoods/venues 3
2.2. Stakeholders/Audience who would benefit from the solution to this problem 3
3. Data and Approach 4
3.1. Data Requirements and Data Sources 4
3.2. Analytic Approach 4
4. The Methodology 5
4.1. Data Collection, Exploration, and Clean-up 5
4.2. Exploratory Data Analysis 6
4.3. Training and Testing Dataset Split 7
4.4 Data Transformation and Feature Extraction 7
4.5. Training Classification Models 7
5. Results 7
6. Further Discussion 8
7.Conclusion 10
8. GitHub Repository 10
References 11

1. Background

Battle of Neighborhoods, this project is aimed at solving a problem involving Foursquare location data and demonstrating data science skills.

I choose to explore the city I am born and brought up; Coimbatore located in southern state of Tamil Nadu in India. Coimbatore is second largest city in the state and very well connected by road, rail, and air transport. It has mixed cultures from all over India. Well known for engineering and manufacturing, education institutions, health care facilities and services, textile businesses as well as it is a tourism hub.

Coimbatore has 64 neighborhoods (Neighbourhoods of Coimbatore, n.d.) excluding the sub-urban neighborhoods.

Initial thought was to cluster neighborhoods, which could help small business start-ups in terms of assessing opportunities and existing competition. However, when I explored neighborhoods using Foursquare I figured out a problem of missing data, which might be current limitations/challenges to location-based companies like Foursquare.

2. Business Problem

My Initial exploration of Coimbatore's places using Foursquare's explore endpoint revealed that 177 venues from 51 neighborhoods out of 64 total neighborhoods in the city, while search end point resulted in 7472 venues covering all 64 neighborhoods.

Note that there are **13 neighborhoods itself did not appear in recommendations and 376 venues appeared as uncategorized**. Second tier cities like Coimbatore may lack availability of data (categories, user reviews, etc...) for recommendations to new/prospective users (especially tourists/travellers and Foursquare's own customers).

Out of all attributes of a venue, category is plays crucial role in fetching results. In rest of project I would attempt find a solution to assign categories to uncategorized venues.

2.1. Importance of improving missing data of neighborhoods/venues

This could affect variety of people and their needs,

- As stranger/traveller one could be presented with no recommendations near him or show something far, while some venue is really close by which is unreviewed but of same category.
- As Foursquare's customer / business planer, venue details with missing categories, reviews, recommendation ratings would make it difficult to analyse and make decisions related to choose of neighborhoods/venues

2.2. Stakeholders/Audience who would benefit from the solution to this problem

- Foursquare / Similar location data providers
    - Enrich their databases for better competitive advantage and improved customer experience
- App developers/owners, who offer services based on location-data (example Food Delivery, Cab Services etc…)
    - When an un-reviewed venue is found closer than a recommended venue, enable users with optional features like "Be First to Review" and reward for doing so

- Such an approach could be taken if venues are categorized otherwise it would even be exceedingly difficult.

3. Data and Approach

3.1. Data Requirements and Data Sources

- A list of neighborhood names (Neighbourhoods of Coimbatore, n.d.) and Neighborhood location as latitude and longitude (GeoPy , n.d.)

Region	Neighborhood	Latitude	Longitude
North	Kavundampalayam	11.0452351	76.9472197
North	Chinnavedampatti	11.0629428	76.9843304
North	Press Colony	11.178903	76.9569137
North	Vadamadurai	11.084491	76.9390361
North	Kanuvai	11.0793411	76.9415814
North	KNG Pudur	11.0533111	76.9199558

- List of venues from neighborhoods and their categories (Foursquare Endpoints, n.d.)

Neighborhood	Venue	Venue Category
Kavundampalayam	Kavundampalayam	Parking
Kavundampalayam	Sabari Bakery	Bakery
Kavundampalayam	Ayyappas Restaurant	
Kavundampalayam	Ayyappa's Pearl Wedding Hall	Event Space

- List of all possible categories (Foursuare Categories, n.d.)

Top Category	Sub Category
Arts & Entertainment	Amphitheater
Arts & Entertainment	Aquarium
Arts & Entertainment	Arcade

3.2. Analytic Approach

In short problem is to categorize a venue, which could be approached **as Supervised Multi-Class Text Classification problem**.

Supervised approach requires **a labelled dataset** to train and evaluate models, which contain minimum two columns (**text as input and category as output**)

- Input Text: Venue Names
- Output: Anyone of Foursquare Defined Categories

Neighborhood	Venue	Venue Category
Kavundampalayam	Kavundampalayam	Parking
Kavundampalayam	Sabari Bakery	Bakery
Kavundampalayam	Ayyappas Restaurant	
Kavundampalayam	Ayyappa's Pearl Wedding Hall	Event Space

Such final dataset for model training and testing could be built by analysing neighborhood, location, venue, and venue category data.

4. The Methodology

4.1. Data Collection, Exploration, and Clean-up

Since there is no pre-existing datasets available data from various sources were collected and combined to get necessary data for analysis and modelling.

1. Neighborhood Names where collected manually from (Neighbourhoods of Coimbatore, n.d.) and saved as csv file, which has 2 columns namely Region and Neighborhood
2. For each location coordinates were collected using GeoPy library. (GeoPy , n.d.)
    a. It is observed that for 8 neighborhoods GeoPy library failed to fetch the coordinates
    b. It could be due to spelling issues, for those 8 neighborhoods location coordinates were collected using Google Search. Manually updated data is saved as 'cbe_nb_loc_curated.csv'
3. For each neighborhood venue details were collected into separate data frames using Foursquare API's Search and Explore endpoints and were merged without duplicates.

Endpoint	# Neighborhoods	# Venues	# Recommended Groups	# Venue Categories	# Uncategorized Venues
Explore	51	177	1	65	0
Browse	64	7472	1	368	376

There are **13 Neighborhoods missing and 376 uncategorized venues**

4. Few venue names which were of same as neighborhood were removed. Finally, dataset has about **7459 records of venues around Coimbatore**.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Recommendation Group Type	Recommendation Group Name	Venue	Venue Id	Venue Latitude	Venue Longitude	Venue Category
Kavundampalayam	11.045235	76.94722	Not Applicable	Not Applicable	Sabari Bakery	5259e17f11d26f0160e95c89	11.042948	76.948176	Bakery
Kavundampalayam	11.045235	76.94722	Not Applicable	Not Applicable	Ayyappas Restaurant	4d5622e7cff7721e176cb2f5	11.045734	76.946890	NaN

5. Foursquare's Categories endpoint provides hierarchical structure of predefined categories, which were collected in to 2-level parent and child level. **10 Top Categories and 933 Sub/Child Categories.**

Top Category	Sub Category
Arts & Entertainment	Amphitheater
Arts & Entertainment	Aquarium

6. A base dataset was created at this stage with only two columns as we need for modelling.

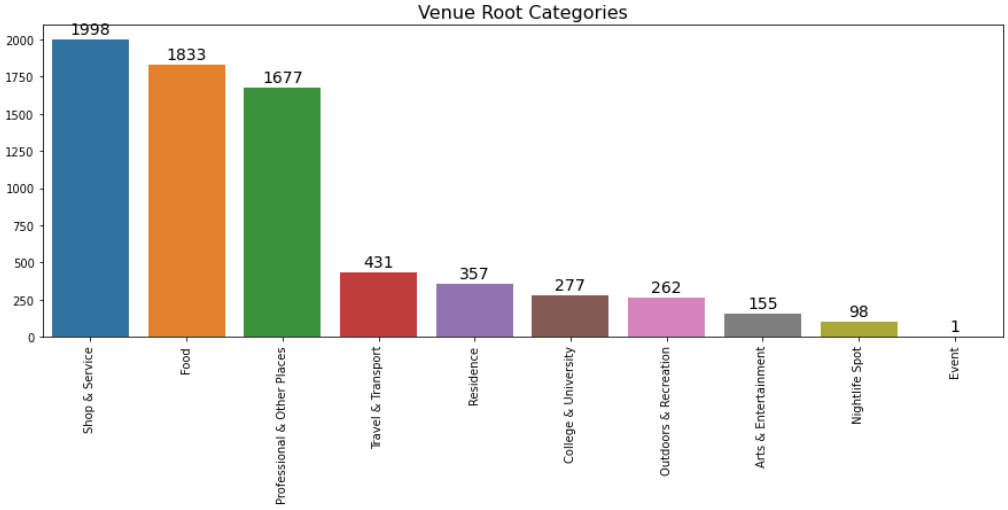
Venue	Category
Sabari Bakery	Bakery
Ayyappas Restaurant	NaN

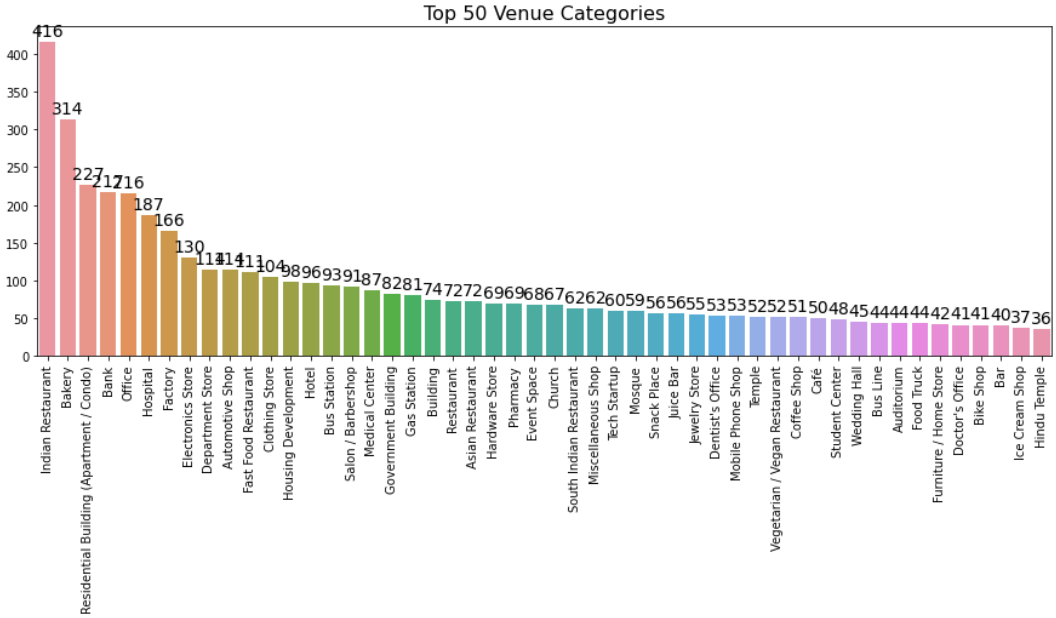
7. Further the rows without categorization was removed as we need labelled dataset for model building.
8. Finally, Root Category column was added using the data fetched from Foursquare Category hierarchy.

Venue	Category	RootCategory
Sabari Bakery	Bakery	Food
Ayyappa's Pearl Wedding Hall	Event Space	Professional & Other Places
Thangam Arisi Mandi	Grocery Store	Shop & Service
Chandra Hyundai Service Centre	Automotive Shop	Shop & Service
Kalpana Theatre	Theater	Arts & Entertainment

4.2. Exploratory Data Analysis

Exploration of categories revealed that the remarkably high imbalance among Venue Categories and the Top Category named ‘Event’ has only one venue.





Hence Event Root Category was removed from dataset and **It would be simpler to assign Root/Top Category for uncategorized venues, which will still meet our project goal of categorizing venues**.

Final resulting dataset summary is as given below

```
# Venues: 7088
# Root Categories: 9
```

	Venue	RootCategory
1	Sabari Bakery	Food
3	Ayyappa's Pearl Wedding Hall	Professional & Other Places
4	Thangam Arisi Mandi	Shop & Service
5	Chandra Hyundai Service Centre	Shop & Service
6	Kalpana Theatre	Arts & Entertainment

4.3. Training and Testing Dataset Split

Base dataset split in to 80% for training and 20% for testing. sklearn's **StratifiedShuffleSplit** was used to split the dataset so that from each class 20% records are split for testing remaining is reserved for training.

This method was used to ensure every class is represented in training and test set, otherwise popular random selection may miss, over/under sample classes from imbalanced dataset.

4.4 Data Transformation and Feature Extraction

Bag of Words is a simple and popular technique to convert text data into numeric data. However, this method does not capture any information other than plain word count.

sklearn's TfidfVectorizer is used to transform venue names to numeric vector, which does of following two stage process:

- First use sklearn's CountVectorizer to transform venue names to bag of words vector by which each word in venue name is counted full vocabulary of words from all Venue Names
- Second use sklearn TfidfTransformer to transform the word count vector into TF-IDF feature vector (For explanation of TF-IDF https://en.wikipedia.org/wiki/Tf%E2%80%93idf)

4.5. Training Classification Models

There are several algorithms that can be training and evaluated, with given time constraints to complete the project following algorithms are considered.

All algorithms fed with same training set.

- K Nearest Neighbour (KNN)
  - Best k/model is evaluated using 20% training set itself.
- Decision Tree
- Support Vector Machine
  - 3 different models using various kernel ['linear', 'rbf', 'sigmoid'] are trained
- Logistic Regression
  - 5 different models using various solvers ['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga'] are trained

In total 10 different models were trained to evaluate and select best performing algorithm.

5. Results

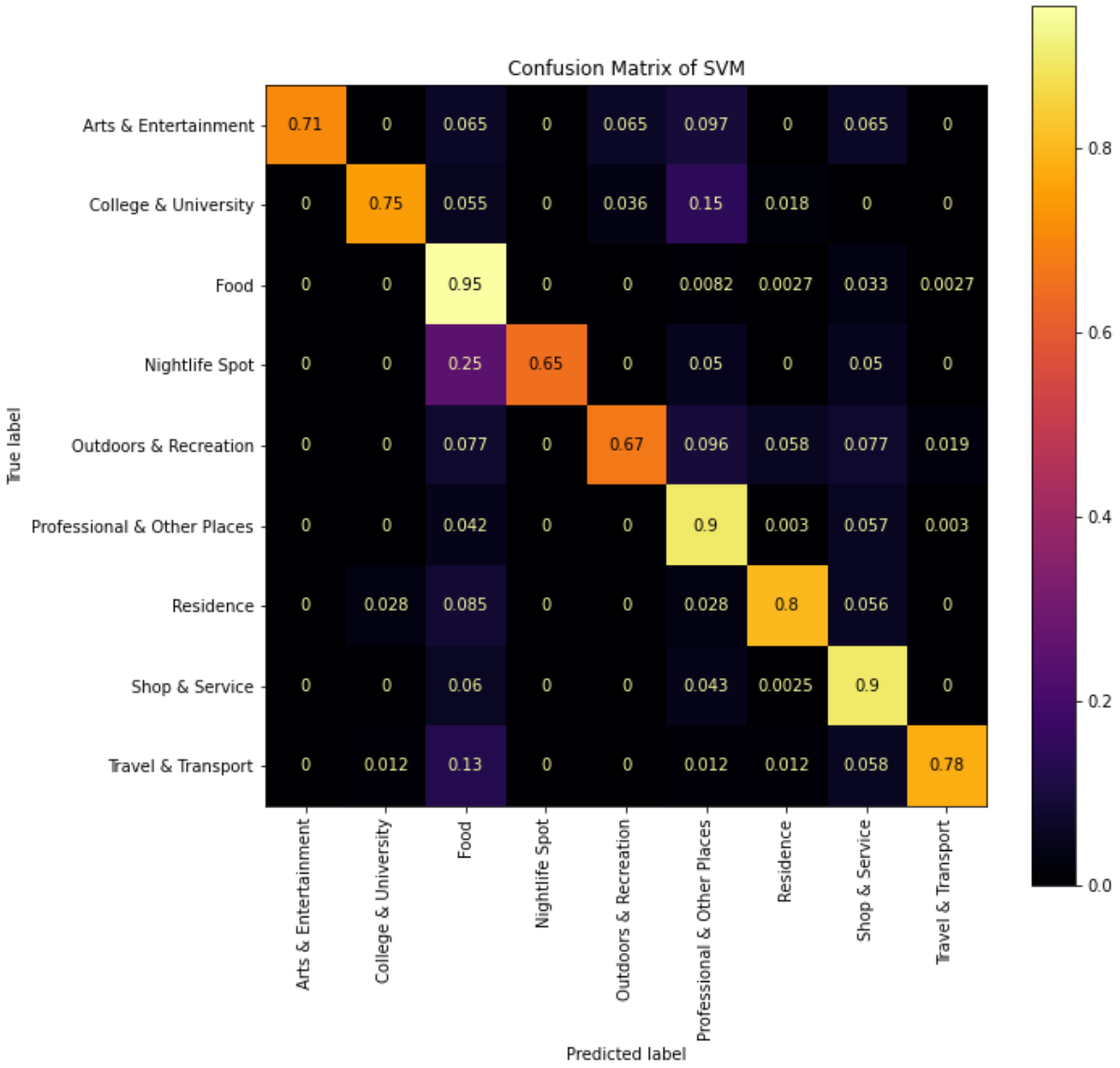
All trained models were run on same testset and predicted values were used to compare against expected values and calculated Accuracy/Jaccard, F1-Score and Logloss(Logistic Regression only)

Algorithm	Jaccard	F1-score	LogLoss	Description
SVM	0.877292	0.875764	NaN	kernel=rbf
Decision Tree	0.861777	0.861950	NaN	criterion=Entropy
KNN	0.813822	0.815391	NaN	Best K=1 at Validation Accuracy=0.77
LogisticRegression	0.428773	0.354079	1.691702	solver=newton-cg, C=0.01

SVM with Radial Basis Function (rbf) kernel and Decision Tree performing better than others. **SVM with 88% stands top performer.**

6. Further Discussion

Further from below confusion matrix plot form results of SVM model suggests that majority of the results are as expected (on the diagonal). Confusion matrix is represented as normalized values(i.e, percentage of accuracy)



However, there are misclassifications to explored further, especially the ones with accuracies less than 90%.

Especially, Major chunk of accuracy loss happens due to misclassification of "Nightlife Spot"(Expected) as "Food"(Predicted). Extracted view of such venues (given below) shows that there are **wrong categories assigned to certain venues in existing database of Foursquare Place API and mix of words present in different categories.**

Such examples are highlighted on below table.

Venue	Expected	Predicted
G3 water Doctor	Food	Professional & Other Places
Kalanikethan	Shop & Service	Food
Spartan	Shop & Service	Food
saraswati hospital	Shop & Service	Professional & Other Places
Paatiamma Pottikadai	Shop & Service	Food
K.M.M Fish Stall	Shop & Service	Food
Compu Soft Systems	Shop & Service	Professional & Other Places
kovai pazhamudir	Shop & Service	Food
Black Box	Nightlife Spot	Food
Arafa JAS Briyani Hall	Food	Professional & Other Places
TNEB Office	Shop & Service	Professional & Other Places
bullet miami	Shop & Service	Food
Hotel Clarion	Nightlife Spot	Food
Spectra	Shop & Service	Food
Bullmen Royal Enfield	Shop & Service	Professional & Other Places
Nair Kadai	Shop & Service	Food
Kongu Masala	Shop & Service	Food
Lakshmi Ceramics	Shop & Service	Food
Knotty Spirit	Nightlife Spot	Food
Bullmen Royal Enfield	Shop & Service	Professional & Other Places
R v store	Food	Shop & Service
Mercedes Benz	Shop & Service	Food
G3 water Doctor	Food	Professional & Other Places
Bread Boys	Food	Shop & Service
Karpagam Complex Fish Stall	Food	Shop & Service
All Seasons Restaurant	Food	Residence
Nongu Kadai	Shop & Service	Food
Thomas Cook India Private Limited	Shop & Service	Professional & Other Places
Anandhas Rs Puram	Food	Shop & Service
Kathi express	Food	Shop & Service
The Circus - Creative & Development Solutions ...	Shop & Service	Professional & Other Places

7.Conclusion

Overall approach to solving the problem of enriching venue details using text classification is appropriate, which is evident from 88% accuracy shown by Support Vector Algorithm.

However, misclassification of the model also revealed that the existing categories used for training itself is susceptible and could be wrong. Thus, leading machine learning models also to be learning and predicting wrong.

The project can be further continued in following aspects.

- For model training and testing recommended venues only can be used, as they would be verified records
- Sophisticated text clean-up can be considered to avoid any confusing words.
- Further specialized algorithms such as Word2Vec, Embeddings can be considered for feature extraction.
- In addition to the scope uncategorized venues, we could validate venue categories of un-reviewed venues using same model.

8. GitHub Repository

Entire Source code and the report itself are available at
https://github.com/arunnatarajan80/Coursera_Capstone

s

References

Foursquare Endpoints. (n.d.). Retrieved from Foursqare Places API: https://developer.foursquare.com/docs/places-api/endpoints/

Foursuare Categories. (n.d.). Retrieved from Foursquare Places API: https://developer.foursquare.com/docs/api-reference/venues/categories/

GeoPy . (n.d.). Retrieved from GeoPy Documentation: https://geopy.readthedocs.io/en/stable/

Neighbourhoods of Coimbatore. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Neighbourhoods_of_Coimbatore