Training a machine learning model to predict a good location for a café business based on nearby venues.

Fabio Simoes

17th March 2019

1. Introduction:

1.1. Business Problem:

Selecting the location for a new business is of high importance and directly related to its success. A café business targets mainly a young demographic, therefore a convenient location is essential. A café business should be centrally located - the ideal location would be near an office or university campus - a place where people already gather. Therefore, it is sound to assume that the venues in areas with existing cafes or coffee shops could be used to predict a good location for a new similar business in a similar city.

1.2. Who would be interested:

This project would interest an entrepreneur starting a new venture or a business chain looking to open a new store to expand its business.

2. Data:

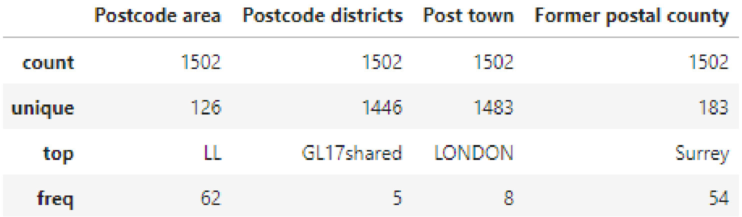
2.1. Data and sources:

The data from this project originates from 3 different sources. Initially a table of postcodes and towns for the United Kingdom was scraped from Wikipedia using Beautiful Soup library.

To obtain venue data for these locations, latitudes and longitudes were obtained from Geocoders and merged with the postcode table. Postcodes which did not return latitudes and longitudes were removed from the original table as nearby venues could not be retrieved.

Finally, Four Square API was used to retrieve nearby venues (500m radius) for each of the postcodes based on their coordinates. The following information was retrieved from Four Square: Venue Name, Venue Latitude, Venue Longitude, Venue Id and Venue Category.

One hot encoding method was used to convert the venue categories into numerical data for model training. The table in figure 1 describes the data scraped from Wikipedia showing 126 unique postcodes.



	Postcode area	Postcode districts	Post town	Former postal county
count	1502	1502	1502	1502
unique	126	1446	1483	183
top	LL	GL17shared	LONDON	Surrey
freq	62	5	8	54

Figure 1: Statistical description of postcode dataset comprising of 126 postcodes.

2.2. Preparation and cleaning of data:

To generate training data for classification, the London data was selected. This data was further filtered to ensure only postcodes with more than 5 venues are included for training data. Each selected postcode must also contain at least 1 café/coffee store venue.

The data obtained from Wikipedia is not standardized, therefore the same town name can be in upper or lowercase. As such, categorical data was modified to lower case to enable correct filtering.

West midlands data was also kept as the target for model prediction. West midlands postcodes were also selected if there are at least 5 venues (This is probably too low for proper predictions).

The frequency of Café and Coffee shops was calculated, added together and named 'Coffee' as a new column. To overcome the different number of venues in each postcode, the frequency of 'Coffee' venues was normalized.

3. Methodology:

3.1. Feature selection:

The normalised frequency of Coffee venues was used to assign a class of 1 or 0 to the training data. Values greater than the normalised frequency mean were classified as 1 while values under the mean were classified as 0. This data was then merged to the postcode table. Café related categories were removed from the feature set for training as the objective is to use the other nearby venues for training.

Number of café/coffee shops in London postcodes: 481
Number of café/coffee shops in Birmingham postcodes: 218
Number of unique venue categories: 324

The unique venue categories were selected as predictors for classification (X data) as the calculated classification was used as Y data.

3.2. Train-Test Split:

The filtered London data was randomly split into a Training Set (80% of the data) and a Test set (20% of the data).

3.3. Classification algorithms:

Four classification algorithms were selected for testing with training and test data to identify which algorithm would be the best classifier for this problem. The algorithms selected were as follow:

- K-Nearest Neighbours
- Decision Trees
- Random Forests
- Logistic Regression

3.4. Hyperparameter tuning:

For each algorithm, hyperparameter tuning was carried out. For each setting, the accuracy measurement was returned. Accuracy measurements for each parameter were then plotted against measured accuracy.

- For KNN, the value of K between 1 and 10 was tested.
- For Decision Trees, depth values between 1 and 20 were tested.
- For Random Forests, the number of trees between 1 and 20 was tested.
- For logistic regression, C values between 0.1 and 30 were tested.

The algorithm was then re-run using the optimised parameters.

3.5. Performance evaluation of classification models:

Following calibration with the training set, the models were tested using the y test set (20% of the data, which not included in the training set).
Following hyperparameter tuning, each model was evaluated using several metrics. These include Jaccard Index, F1 score, LogLoss, Confusion Matrix, Precision and Recall. These were visualised by plotting confusion matrix results and a Receiver operating Characteristic plot of the trade-off between sensitivity and specificity.

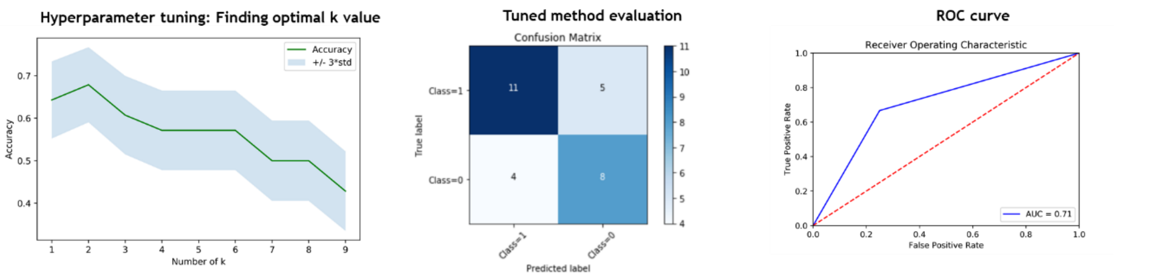
3.6. Prediction of good Cafe locations using Midlands data:

Following model evaluation, the best performing algorithm was used to predict either Class 1 or 0 for the Midlands predictor features (Nearby Venue Categories).

4. Results and Discussion:

4.1. K-Nearest Neighbours classification model:

Hyperparameter testing revelated that k = 2 resulted in optimal accuracy when compared to the test set. The highest accuracy value obtained was 0.71.
The model was then run again using k=2 and tested against the test data. The classification report showed 11/16 correctly predicted for class 1 (good location for a Café) and 8/12 for class 0 (Bad location for a Café). The ROC curve shows an AUC value of 0.71 which suggests a good balance between specificity and sensitivity. However, while this model displayed relatively high accuracy, a poor LogLoss value was obtained when compared to other classification methods.



Hyperparameter tuning: Finding optimal k value

Tuned method evaluation

ROC curve

Highest Classification Accuracy: 0.71
Best k value: 2

The model predicted 11/16 Class 1 and 8/12 Class 0 correctly

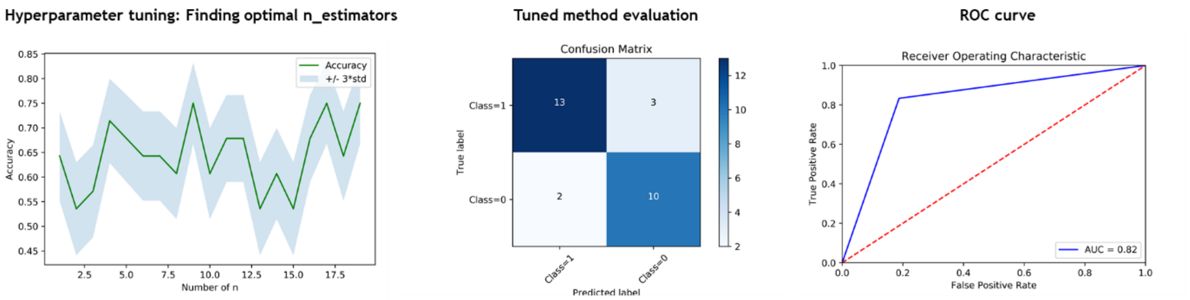
Good balance between sensitivity and specificity

Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.71	0.71	0.59	7.62

Figure 2: Hyperparameter tuning and classification evaluation for K-Nearest Neighbours algorithm.

4.2. Random Forest Classifier:

Hyperparameter testing revelated that n_estimators = 9 resulted in optimal accuracy when compared to the test set. The highest accuracy value obtained was 0.75.
The model was then run again using n estimators = 9 and tested against the test data. The classification report showed 13/16 correctly predicted for class 1 (good location for a Café) and 10/12 for class 0 (Bad location for a Café). The ROC curve shows an AUC value of 0.82 which suggests a very good balance between specificity and sensitivity. In addition to obtaining a good accuracy, the LogLoss calculated was relatively low indicating a good probability for prediction. The cross validated accuracy score was partially lower than Jaccard index or F1 score, however it still returned a relatively high value and in line with the other accuracy metrics.
This algorithm was shown to outperform other models and was therefore selected. However, the random nature of this algorithm results in variable results due to the number of features in the X data.



Hyperparameter tuning: Finding optimal n_estimators

Tuned method evaluation

ROC curve

Highest Classification Accuracy: 0.75
Best n_estimator value: 9

The model predicted 13/16 Class 1 and 10/12 Class 0 correctly

Good balance between sensitivity and specificity

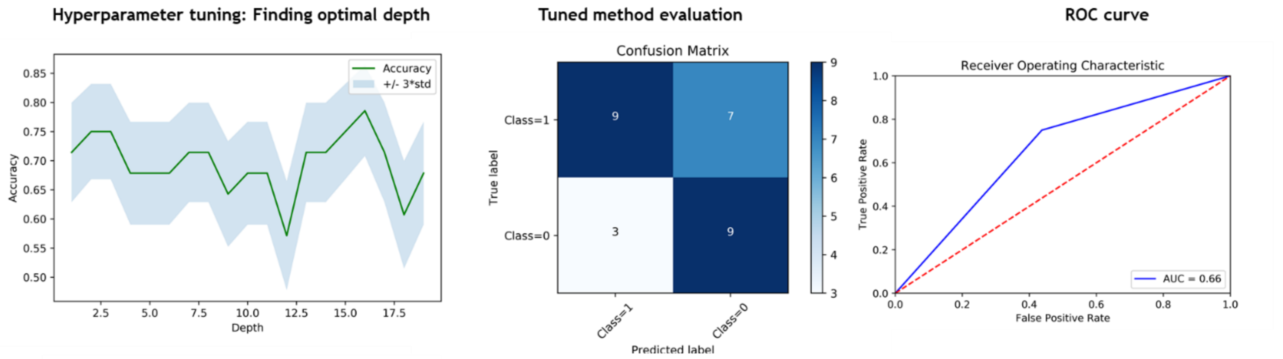
Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.82	0.82	0.70	0.58

Figure 3: Hyperparameter tuning and classification evaluation for Random Forest algorithm.

4.3. Decision Tree Classifier:

Hyperparameter testing revelated that optimal depth was 13. The algorithm was run criterion=”entropy”. The highest accuracy value obtained was 0.64.

The model was then run again using depth set as 13 and tested against the test data. The classification report showed 9/16 correctly predicted for class 1 (good location for a Café) and 9/12 for class 0 (Bad location for a Café). The ROC curve shows an AUC value of 0.68 which suggests relatively poor balance between specificity and sensitivity. In addition to the relatively poor accuracy obtained from this model, a very high LogLoss value of 12.3 was obtained suggesting a low probability of correct prediction when compared with other models. This algorithm did not perform as well as Random Forest Classification. Possibly because of its tendency to overfit data.
The cross validated score was very similar to the Jaccard index and F1 score showing consistent accuracy values.



Highest Classification Accuracy: 0.78
Best depth value: 16

The model predicted 13/16 Class 1 and 7/12 Class 0 correctly

Acceptable balance between sensitivity and specificity

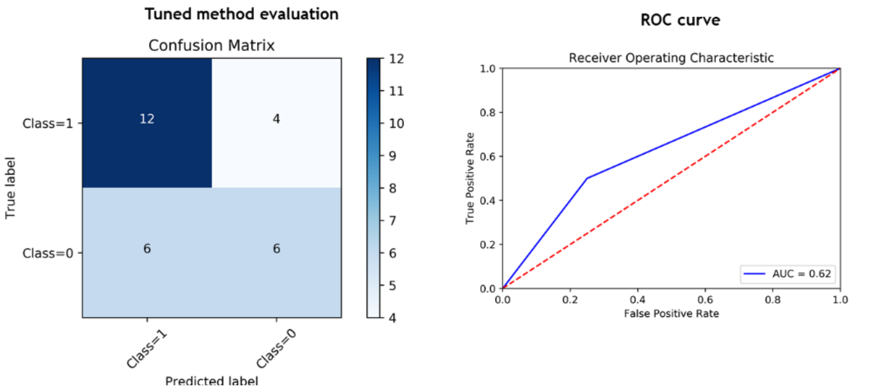
Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.64	0.63	0.65	12.3

Figure 4: Hyperparameter tuning and classification evaluation for Decision Tree algorithm.

4.4. Logistic Regression Classification:

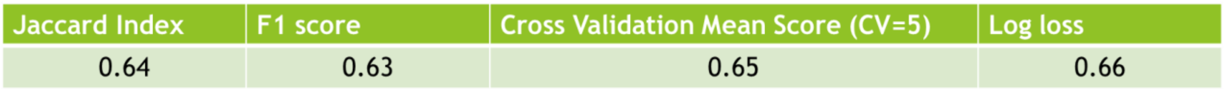
Hyperparameter testing revelated that optimal C value was 10. The highest accuracy value obtained was 0.68.
The model was then re-run using C=10 and tested against the test data. The classification report showed 12/16 correctly predicted for class 1 (good location for a Café) and 6/12 for class 0 (Bad location for a Café). The ROC curve shows an AUC value of 0.62 which suggests relatively poor balance between specificity and sensitivity. A low LogLoss value of 0.66 was obtained.
The cross validated score was very similar to the Jaccard index and F1 score showing consistent accuracy values.

Hyperparameter tuning: Best C value = 10
Best Classification accuracy for C = 10 → 0.68



The model predicted 6/12 Class 1 and 12/4 Class 0

Poor balance between sensitivity and specificity



Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.64	0.63	0.65	0.66

Figure 5: Hyperparameter tuning and classification evaluation for Logistic Regression algorithm.

4.5. Summary of Classification Model Performance:

Algorithm	Jaccard Index	F1 Score	Logloss	Cross Val Score	Number of True Positives	Number of true Negatives	Number of False Positives	Number of False Negatives	ROC AUC
KNN	0.714286	0.714286	7.623964	0.592430	11	8	4	5	0.71
Random Forests	0.821429	0.822120	0.575979	0.695136	13	10	2	3	0.82
Decision Trees	0.642857	0.642857	12.335277	0.649552	13	7	5	3	0.68
Logistic Regression	0.642857	0.637128	0.664771	0.649552	12	6	6	4	0.62

Figure 6: Classification performance summary

4.6. Using Random Forest Model training with London data to predict good location in the midlands to open a new Café business.

The optimised random forest model in section 4.2 was selected as the best performing algorithm and used with the data for midlands postcodes to predict either class 0 (bad location for café) or 1 (good location or a café). This data consisted of 22 unique postcodes. The predicted classification results were then merged into the midlands data table to plot this data for each postcode in the table. Folium maps library was used to render the map and the circle maker feature was applied to differentiate between Class 1 (Blue circles) and class 0 (red circles).

As expected most locations in Birmingham city are classed as suitable to open a café as these are locations that have several surrounding venues. Other smaller locations are classed as bad location for a café. This is probably due to the low number of surrounding venues or infrequent venues for areas with Café businesses.

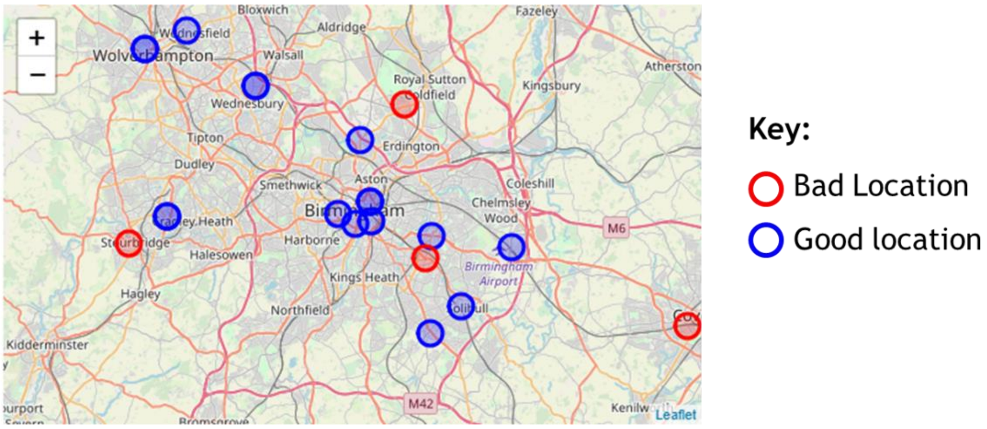


Figure 7: Rendered folium map displaying class 1 or 0 for postcode locations in the midlands.

5. Summary and conclusions:

Wikipedia data combined with Geocoders coordinates and Foursquare API data were used to generate classification models to predict ideal locations to open a Café/Coffee store business.
K-nearest neighbours, Decision Tree, Random Forest and Logistic regression models were tuned to establish best parameters for highest accuracy.
Random Forest classifier was chosen as the best performing model due to high accuracy and low logLoss value. The model was applied to the midlands dataset and the predicted results were rendered as a folium map.
The venues in areas which have high frequency of Cafes may be used as predictors for estimating whether a location in a similar city is suitable to open a Café business or not.
While random forest classifier was the best performing algorithm, it might not be the best suited for this task as its random nature, which can result in variable results.