Training a ML model to predict a good location for a café business

Business Case

- **Introduction/Business Problem:** Train a predictive model to determine a suitable location for opening a new cafe/coffee shop business.

- **Hypothesis:** Venues in areas with a high density of cafe/coffee shops can be used to predict a suitable location for a similar business in a different area. Can training data from one location be used to predict optimal cafe locations in a different area?

Data

- *Beautiful soup* package was used to import HTML data from Wikipedia (https://en.wikipedia.org/wiki/List_of_postcode_districts_in_the_United_Kingdom)
- A table containing UK postcodes was scraped and cleaned into a pandas dataframe.

	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

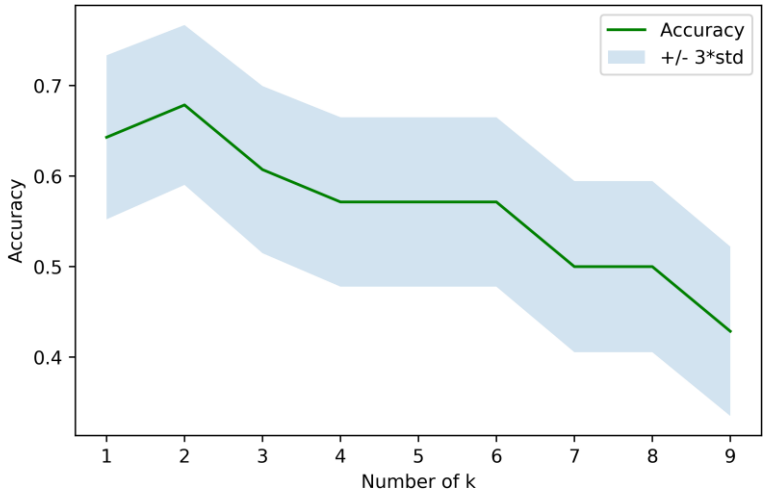
Data transformation and Train-Test Split

1. UK postcodes scraped from Wikipedia table using Beautiful Soup.
2. Latitudes and Longitudes added to table using Geocoders.
3. Foursquare API used to retrieve local venues for each postcode
4. One hot encoding applied to dataset venues.
5. Combined frequency for Café/Coffee Store calculated and normalized.
6. Postcodes for London used for training the classification models.
7. Filter data to only include London data AND postcodes in which Cafe/Coffee shop feature in top 5 most frequent AND postcodes, which include at least 5 other surrounding venues.
8. Class 1 was assigned to postcodes with normalized frequency >= normalized frequency mean. Class 0 was assigned to postcodes below the mean.
9. Coffee related venues excluded from dataset.
10. X comprised of venues data. Y comprised of Classification values 1 or 0.
11. 80% of data was used for training and 20% for testing.
12. Classification models tested: K-Nearest Neighbours, Random Forests, Decision Trees and Logistic Regression.

K-Nearest Neighbours Classifier

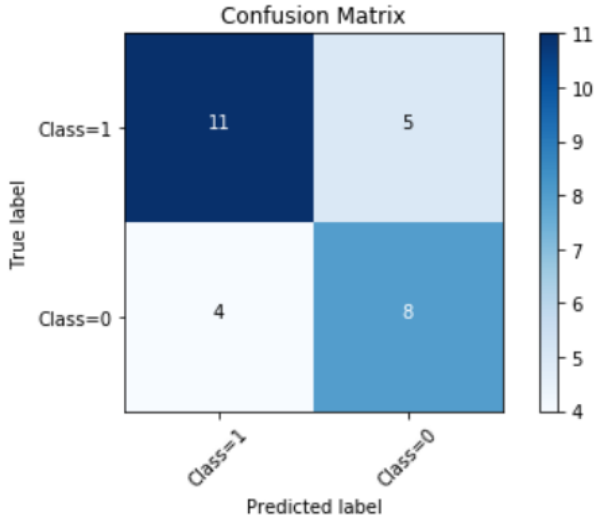
Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.71	0.71	0.59	7.62

Hyperparameter tuning: Finding optimal k value



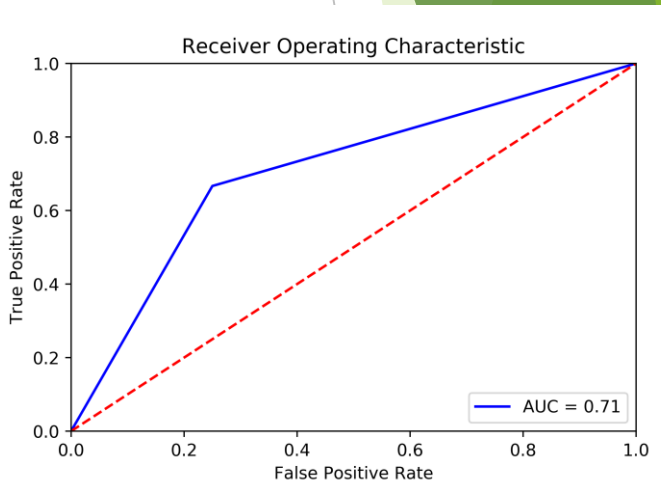
Highest Classification Accuracy: 0.71
Best k value: 2

Tuned method evaluation



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

ROC curve



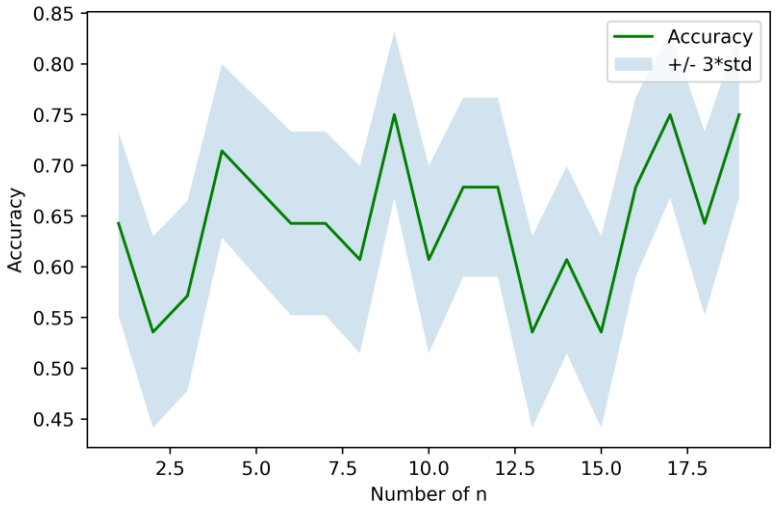
Good balance between sensitivity and specificity

- KNN model was able to provide a good accuracy. However relatively high logloss

Random Forest Classifier

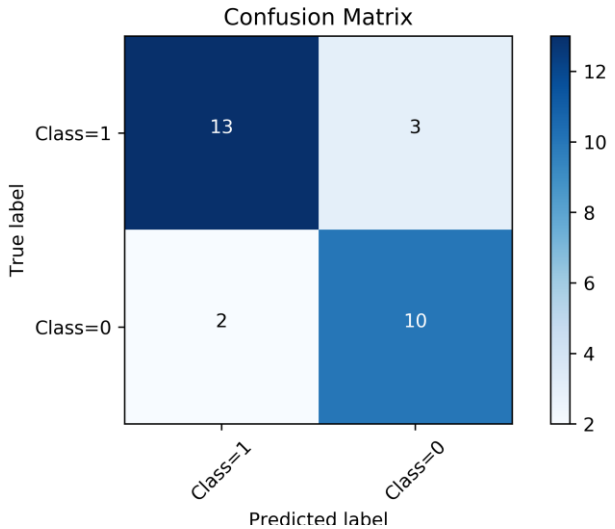
Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.82	0.82	0.70	0.58

Hyperparameter tuning: Finding optimal n_estimators



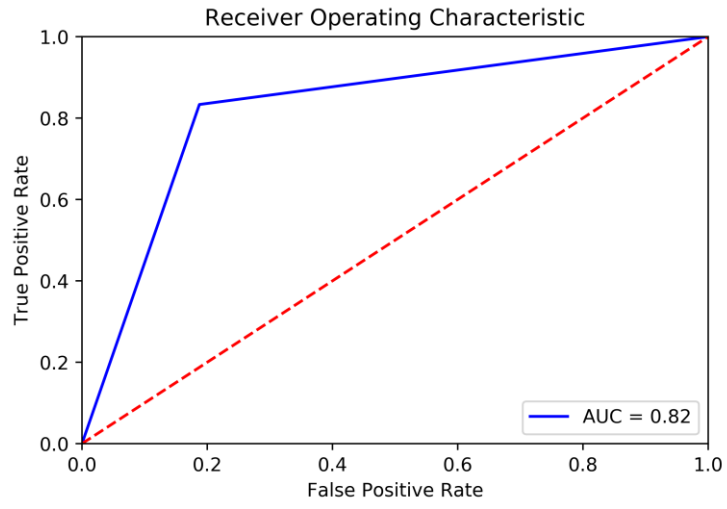
Highest Classification Accuracy: 0.75
Best n_estimator value: 9

Tuned method evaluation



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

ROC curve



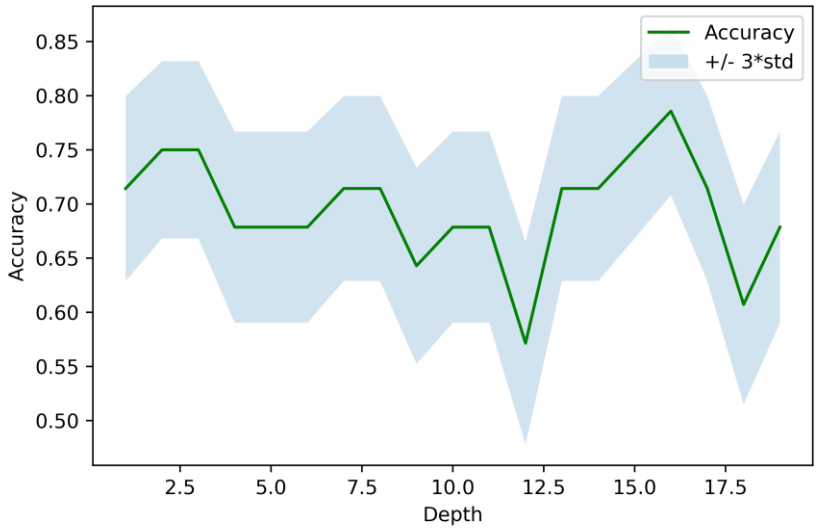
Good balance between sensitivity and specificity

- Random Forest model was able to provide a good accuracy with a low logloss. However results vary with each run due to its random nature

Decision Trees

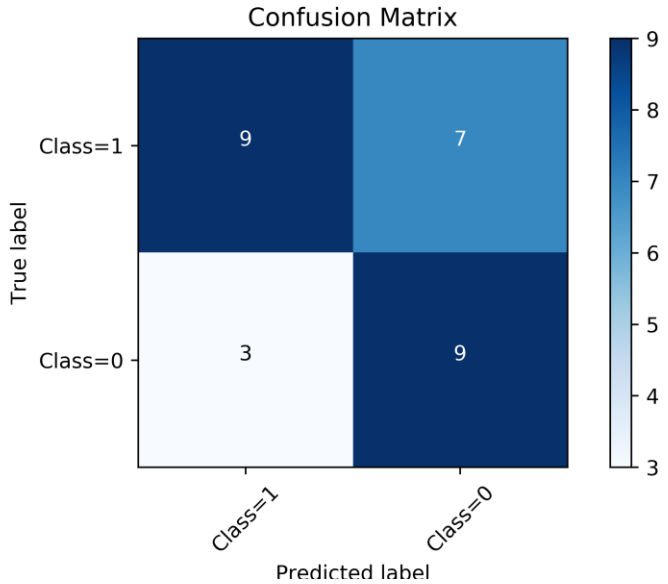
Jaccard Index	F1 score	Cross Validation Mean Score (CV=5)	Log loss
0.64	0.63	0.65	12.3

Hyperparameter tuning: Finding optimal depth



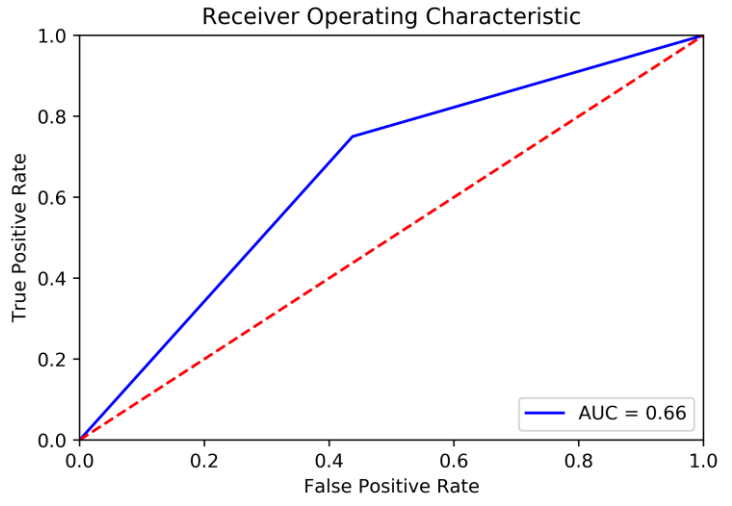
Highest Classification Accuracy: 0.78
Best depth value: 16

Tuned method evaluation



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

ROC curve



Acceptable balance between sensitivity and specificity

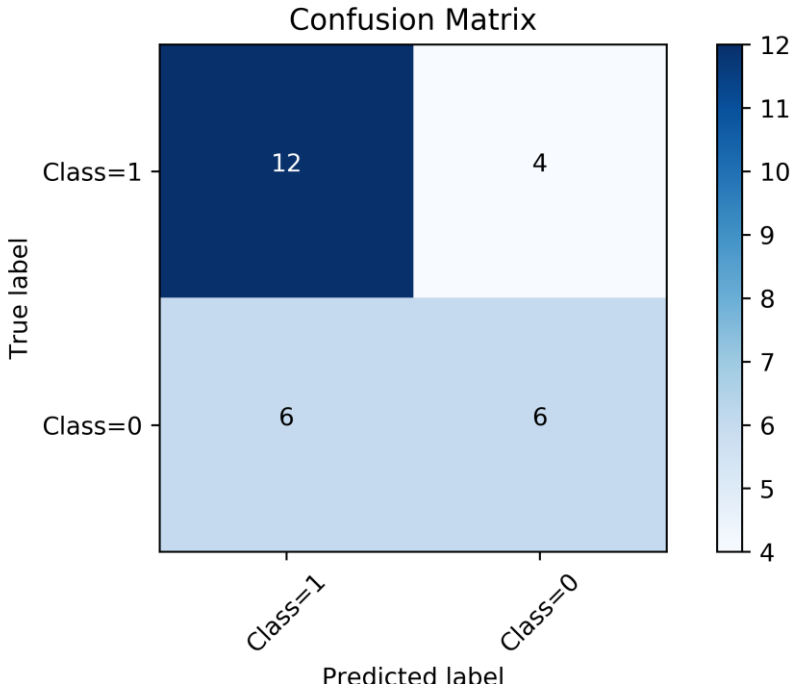
- Decision Trees model was able to provide a resonabe recall for class 0 but poor recall for class 1. High LogLoss value.

Logistic Regression Classifier

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

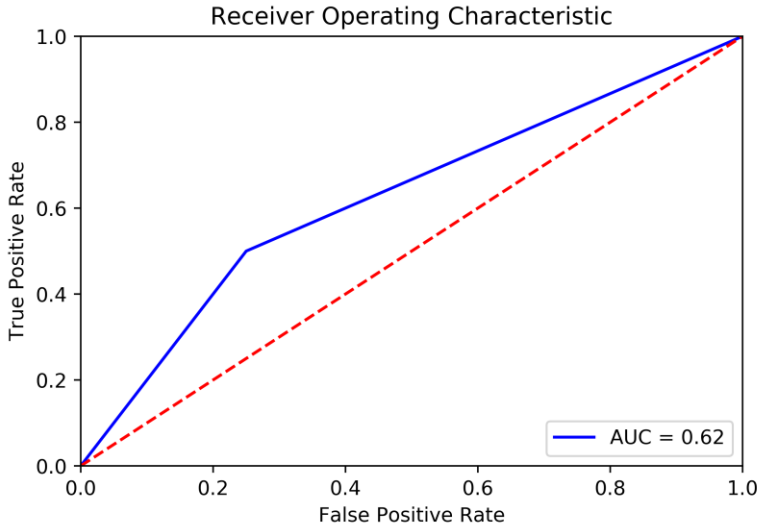
Hyperparameter tuning: Best C value = 10
Best Classification accuracy for C = 10 → 0.68

Tuned method evaluation



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

ROC curve



Poor balance between sensitivity and specificity

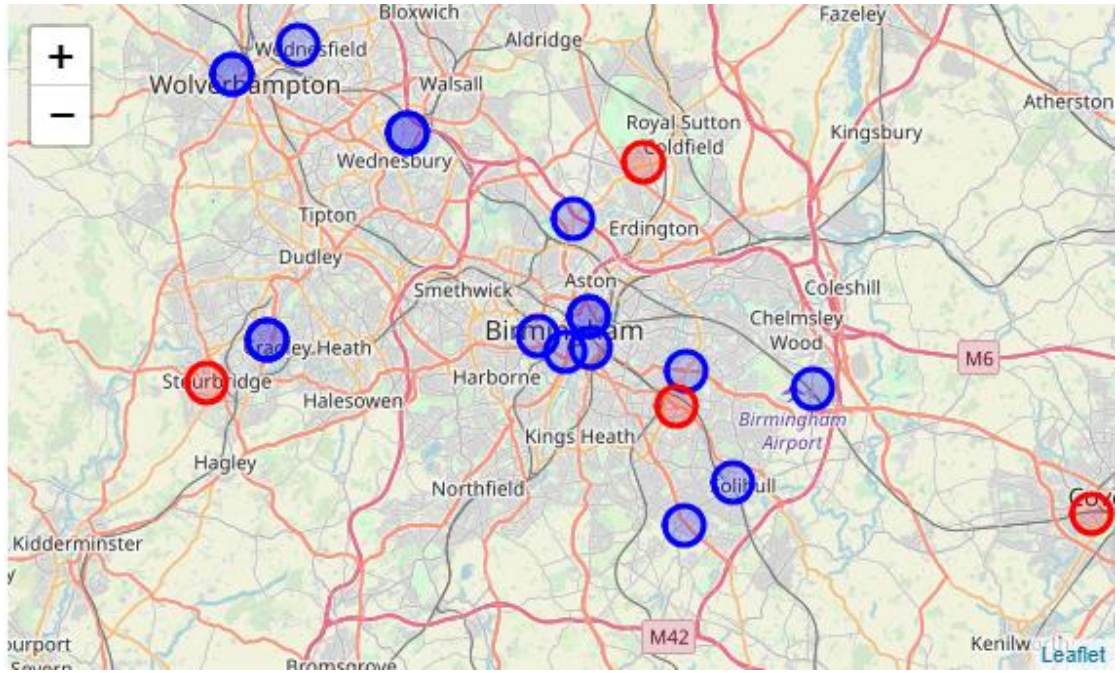
- Logistic Regression model was able to provide acceptable accuracy with a low logloss.

Summary of 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

Predicting good Café locations in the Midlands, UK

- Selected Model: **Random Forests.**
- This model was trained with all data from the London venue dataset and used to predict good Café locations in the midlands.
- Folium maps used to render good and bad locations.



Key:

Bad Location

Good location

Summary

- 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. However the random nature of this model makes its performance variable.
- 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 can be used as predictors for estimating whether a location in a similar city is suitable to open a Café business or not.