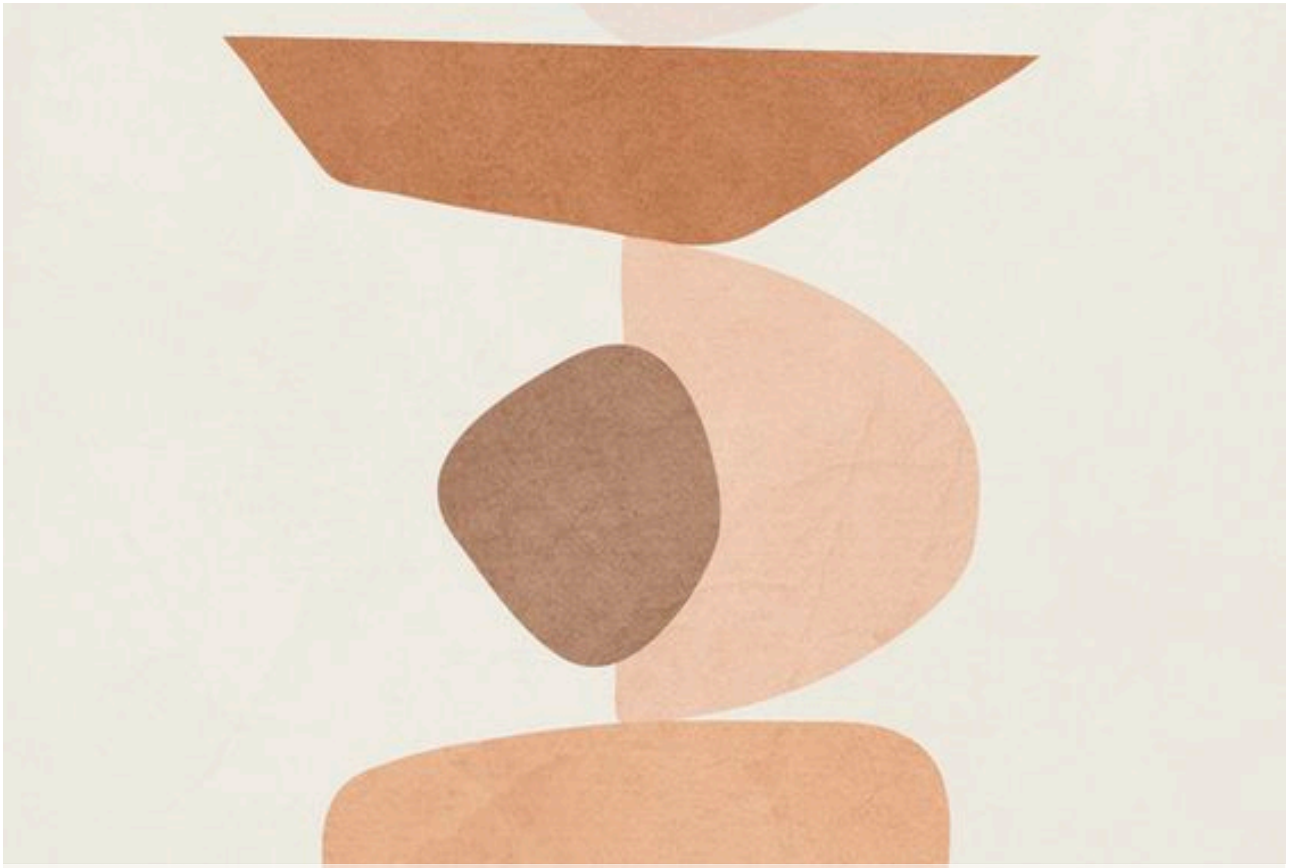
Capstone Project

Explaining Customer Ratings of Tesco and Sainsbury's



Zixiong Long

August 2020

Table Of Contents

1. Introduction

1.1 Background

Grocery stores are unquestionably an essential part of urban life. From sandwiches to ibuprofen, local grocery stores are perhaps the third mostly visited places in one's day, apart from his or her flat and office. With the popularity of mobile apps and experience-sharing websites, ratings from vast amount of users are shared and facilitate the exploring of other users to the venue.

Under this backdrop, it would be of great interest to the management of the grocery stores or a curious user to find out the driving forces behind the rating of the store. The management team of the stores can then utilise these findings to improve the customer experiences of existing stores or find a better location for a new store, and potentially increase the revenue of each store.

Which factors could provide explaining power to the different ratings across those seemingly identical stores? Although the level of services and the completeness of variety of goods could be valid answers to address this question, those data are generally difficult to for an outsider to obtain. Rather, the relevant exogenous factors, such as the location of the store, the house prices of the neighbourhood, the competitive environment, are available through location data providers and other free online resources, and these factors are the focus of this research paper.

1.2 Define the problem

The key question of this research is to uncover the exogenous factors that drive users' ratings of a grocery store. To narrow down the problem, London is chosen as the city, and Tesco and Sainsbury's are chosen as the grocery store companies of this analysis. It is constructed as a classification problem and the user ratings are the labels to be predicted. The final fine-tuned machine learning model will provide the best explanatory inputs that contribute to the ratings.

2. Data Sources

Datasets used in this analysis are from three sources: Foursquare location data, Wikipedia and The London Datastore. The data extracted from these sources will become the inputs of the machine learning model to classify the ratings of grocery stores.

2.1 Foursquare API

Foursquare API is the main source of data in this report. It is a popular location data vendor that provides venues locations, user ratings and tips, etc. It is utilised to find the coordinates of London's Tesco and Sainsbury's stores, explore the neighbourhood of these stores and obtain the ratings of these stores. Here is an example of Foursquare data:

id	name	categories	referralId	hasPerk	location.address
4b8d4de5f964a52038f332e3	Tesco Metro	[{'id': '4bf58dd…	v-1595789668	False	17-25 Regent St
4bbe6094006dc9b64512fb3f	Tesco Express	[{'id': '4bf58dd…	v-1595789668	False	1-4 Charing Cross
4acf3e3ef964a520dcd220e3	Tesco Metro	[{'id': '4bf58dd…	v-1595789668	False	22-25 Bedford St, Covent Garden

Figure 1. Example of Foursquare data

2.2 Wikipedia

The "List of areas of London" (https://en.wikipedia.org/wiki/List_of_areas_of_London) from Wikipedia is used. It provides the areas and boroughs of London and the corresponding post codes, which will be used for the locations of the grocery stores in the model. Here is an example of the Wikipedia data:

Location	London borough	Post town	Postcode district
Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2
Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4
Addington	Croydon[8]	CROYDON	CR0
Addiscombe	Croydon[8]	CROYDON	CR0

Figure 2. Example of Wikipedia data

2.3 The London Datastore

The London Datastore is a free and open data-sharing portal where anyone can access data relating to the capital. The median house prices of each areas in London is downloaded from this website (https://data.london.gov.uk/dataset/average-house-prices). Two types of house prices are provided by the website: organised by Borough and by Wards. Here is an example of The London Datastore data:

New code	Ward name	Borough name	Year ending Dec 1995	Year ending Mar 1996	Year ending Jun 1996	Year ending Sep 1996	Year ending Dec 1996	Year ending Mar 1997	Year ending Jun 1997
E09000001	City of London	City of London	-	-	-	-	-	-	-
E05000026	Abbey	Barking and Dagenham	53,000	50,000	50,000	53,500	49,998	46,995	45,000
E05000027	Alibon	Barking and Dagenham	45,000	45,000	45,500	45,998	45,968	46,148	48,000
E05000028	Becontree	Barking and Dagenham	49,000	50,000	50,000	50,000	51,000	53,995	56,250
E05000029	Chadwell Heath	Barking and Dagenham	59,000	57,000	56,000	58,000	58,000	58,250	59,738

Figure 3. Example of The London Datastore data

3. Data Collecting and Wrangling

3.1 London location data

The raw data from the website consists of 534 rows and 5 columns, with “London”, “London borough", “Post town”, “Postcode district”, “Dial code” and “OS grid ref” being the columns. Considering the “Post town” contains locations which is not London, and column “Dial code” and “OS grid red” are not needed, those rows and columns are dropped. Subsequently the coordinates of each location is searched through geopy package in python and added to the DataFrame. The cleaned dataset is a 297x5 matrix and the first five rows of cleaned data is showed as Figure 4:

	Location	London borough	Postcode district	Latitude	Longitude
0	Abbey Wood	Bexley, Greenwich	SE2	51.4876	0.11405
1	Acton	Ealing, Hammersmith and Fulham	W3, W4	51.5081	-0.273261
2	Aldgate	City	EC3	51.5142	-0.0757186
3	Aldwych	Westminster	WC2	51.5131	-0.117593
4	Anerley	Bromley	SE20	51.4076	-0.0619394

Figure 4. Cleaned London location data

3.2 Grocery stores data

The Foursquare API is utilised to search the grocery stores of Tesco and Sainsbury’s. The coordinates of locations from previous section is used as centroids and 1km (1000 meters) is the radius for the search, with the keywords of “Tesco” and “Sainsbury” respectively. The search is designed to extract id, names and categories of the searched grocery stores. The corresponding location and borough information from London location data is also appended to the dataset for future uses. The returned raw data from the search is presented as Figure 5:

	id	name	categories	latitude	longitude	location	borough
0	4bdaced33904a593ec64479e	Tesco Express	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	51.508483	-0.281283	Acton	Ealing, Hammersmith and Fulham
1	4f575dd7e4b05abedfc29890	Tesco Express	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	51.514446	-0.076851	Aldgate	City
2	4b1abe09f964a520c4f023e3	Tesco Metro	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	51.517222	-0.080311	Aldgate	City
3	5332d2c2498e49c84573186a	Tesco Express	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	51.507199	-0.073972	Aldgate	City
4	4fdfa939e4b0c441c2653cbe	Tesco Metro	[{'id': '4bf58dd8d48988d118951735', 'name': 'G...	51.516805	-0.066368	Aldgate	City

Figure 5. Raw grocery stores data

The raw data of Tesco has 1181 entries and the raw data of Sainsbury's has 1001 entries. After removing duplicates and extract category names from "categories" column, the final cleaned datasets for Tesco and Sainsbury's have 389 and 327 rows respectively. The first five lines of Tesco data is exhibited in Figure 6:

	id	name	categories	latitude	longitude	location	borough
0	4bdaced33904a593ec64479e	Tesco Express	Grocery Store	51.508483	-0.281283	Acton	Ealing, Hammersmith and Fulham
1	4f575dd7e4b05abedfc29890	Tesco Express	Grocery Store	51.514446	-0.076851	Aldgate	City
2	4b1abe09f964a520c4f023e3	Tesco Metro	Grocery Store	51.517222	-0.080311	Aldgate	City
3	5332d2c2498e49c84573186a	Tesco Express	Grocery Store	51.507199	-0.073972	Aldgate	City
4	4fdfa939e4b0c441c2653cbe	Tesco Metro	Grocery Store	51.516805	-0.066368	Aldgate	City

Figure 6. Cleaned grocery stores data

It is worth noting that Tesco has four kinds of stores: Tesco is supermarket, Tesco Express is high street grocery stores, Tesco Extra is larger hypermarkets, and Tesco Metro is grocery store in underground station. Sainsbury's has two kinds of stores: Sainsbury's is supermarket, and Sainsbury's Local is high street grocery stores. Their corresponding names and categories are reflected in "name" and "categories" columns.

The locations of each store are plotted in Figure 7 with red points being Tesco and orange points being Sainsbury's:

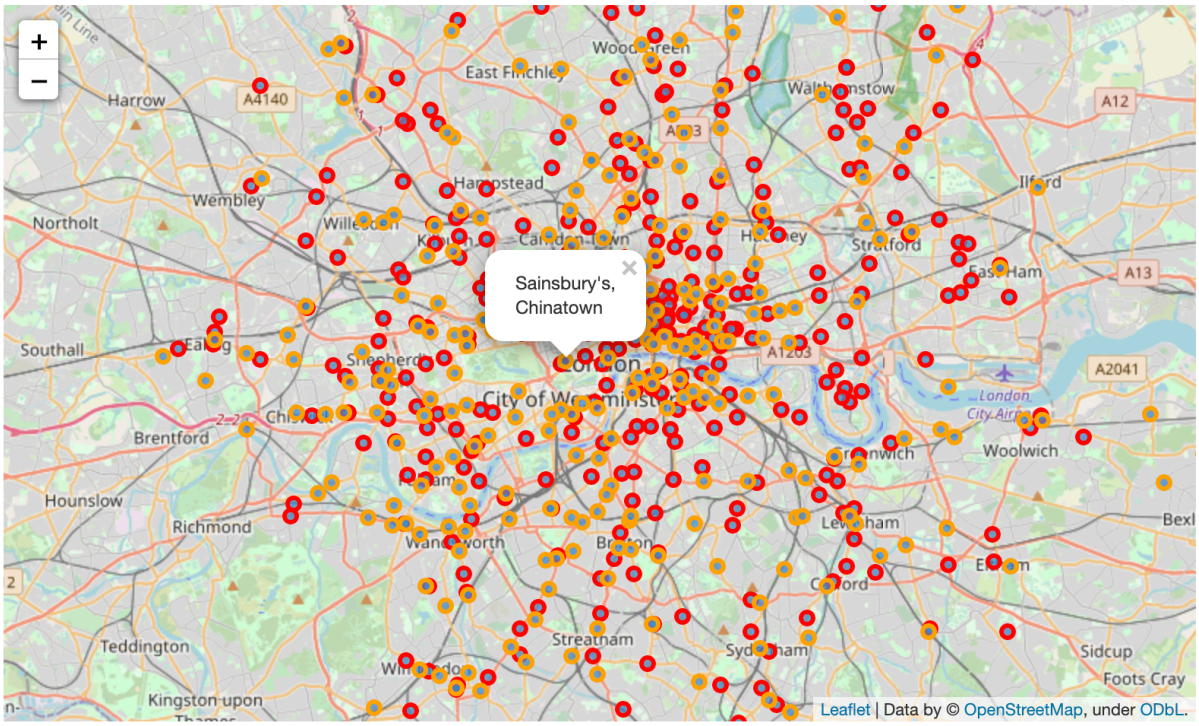


Figure 7. Grocery stores on map

3.3 London house price data

Two files, organised by borough and ward, from the website are downloaded and read into python. The borough file contains 45 rows and ward file contains 630 rows. Median house prices are provided from Dec 1995 to Dec 2017, and only the most recent data is considered. An example of ward data is showed in Figure 8:

	Ward name	Borough name	Year ending Dec 2017
1	City of London	City of London	-
2	Abbey	Barking and Dagenham	231000
3	Alibon	Barking and Dagenham	295000
4	Becontree	Barking and Dagenham	300000
5	Chadwell Heath	Barking and Dagenham	310000

Figure 8. London median house price by ward

The house price is added to each store of the previous Tesco and Sainsbury datasets by matching its location to ward or borough in house price data. If several wards or boroughs can be found, the average house prices is filled in. If no ward or borough can be found, the median London house price (465,000) is filled in. Figure 9 is the Sainsbury's data with house prices:

	id	name	categories	latitude	longitude	location	borough	house price
0	55bb5956498e05559fb4e8c3	Sainsbury's	Supermarket	51.492826	0.120524	Abbey Wood	Bexley, Greenwich	356000.000000
1	4fae6beb108139586b79187f	Sainsbury's Local	Grocery Store	51.508466	-0.266318	Acton	Ealing, Hammersmith and Fulham	541166.666667
2	4d6cf1ec68406ea8461d6f83	Sainsbury's Local	Grocery Store	51.509944	-0.287654	Acton	Ealing, Hammersmith and Fulham	541166.666667
3	58933663102f4705344cc0a6	Sainsbury's Local	Grocery Store	51.514967	-0.268977	Acton	Ealing, Hammersmith and Fulham	541166.666667
4	51910d33498e8a71955213fe	Sainsbury's Local	Grocery Store	51.513419	-0.073019	Aldgate	City	835000.000000

Figure 9. Sainsbury's data with house prices

3.4 Store ratings and neighbourhood data

The rating of each store is collected through Foursquare API (premium call) by store ID and added into the DataFrames of each brand. The stores with NaN rating value are dropped.

The neighbourhood venues are searched by using each stores coordinates as centroid and 500 meters as radius, and only category name of each venue is kept for future exploration. In total 12742 venues are collected for Tesco stores and 11889 venues are collected for Sainsbury's stores.

To better present these neighbourhood venues and connect them to the datasets, venues are rearranged by one-hot encoding and concatenate them to the corresponding grocery store. Because the one-hot encoded matrix will be very sparse, only the top 5 venues for each grocery store and their occurrences are recorded and an extra column named "total" is created to count the total number of venues around the store. The Tesco dataset with ratings and neighbours are presented in Figure 10:

rating	total	top 1 category	top 2 category	top 3 category	top 4 category	top 5 category	top 1 occurrence	top 2 occurrence	top 3 occurrence	top 4 occurrence	to occurren
7.4	18	Food & Drink Shop	Pub	Grocery Store	Eastern European Restaurant	Coffee Shop	2	2	2	2	
6.9	100	Coffee Shop	Hotel	Cocktail Bar	Restaurant	Pizza Place	12	12	5	4	
5.3	96	Cocktail Bar	Coffee Shop	Sandwich Place	Gym / Fitness Center	Café	7	7	6	4	
6.7	57	Italian Restaurant	Hotel	Coffee Shop	Scenic Lookout	English Restaurant	5	4	4	3	
5.3	73	Hotel	Coffee Shop	Pub	Café	Indian Restaurant	10	9	5	4	

Figure 10. Tesco dataset with ratings and neighbours

3.5 Final dataset

Finally the Tesco and Sainsbury's dataset are combined into one DataFrame with an additional column names "brand" created to record the brand of each store. The final data contains 591 rows and 21 columns, and it will be used for the subsequent data exploratory analysis.

4. Exploratory Analysis and Feature Engineering

4.1 Ratings and rating labels

Because the research is constructed as a classification problem, target labels need to be created from ratings. The distribution of ratings is presented in Figure 11 and box plot is showed in Figure 12. The 25% quantile is at 5.5 and 75% quantile is at 6.2, with the highest rating being 8.3 and the lowest being 4.5. Therefore, 5.5 and 6 are chosen as cutoff points. Ratings below 5.5 are labelled as "low", ratings above 6 are labelled as "high" and the rest are labelled as "median".

In total, low label has 160 entries, median label has 232 entries and high label has 160 entries.

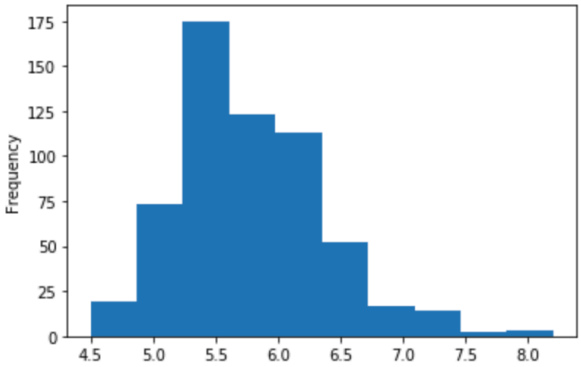


Figure 11. Histogram of ratings

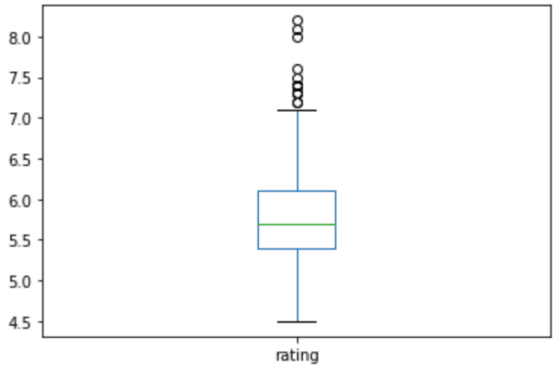


Figure 12. Box plot of ratings

4.2 Splitting training and test data

To make sure that information is not leaked out from test set, 25% of data is randomly picked as test datasets and only the training dataset is explored to form the model. The following exploratory data analysis is based on training data with 443 entries.

4.3 Rating vs brand

To explore the relationship between rating and brand, box plot is presented in Figure 13. The key findings are :

1. Tesco(5.6) on average has lower rating than Sainsbury(6.0)
2. Tesco and Sainsbury have similar lower bound in rating, at 4.5
3. Tesco(6.8) rating upper bound is lower than Sainsbury(7.4)
4. Tesco(7.5) max rating is smaller than Sainsbury(8.5) max rating

Therefore, brand is considered as a useful feature as it differentiates data.

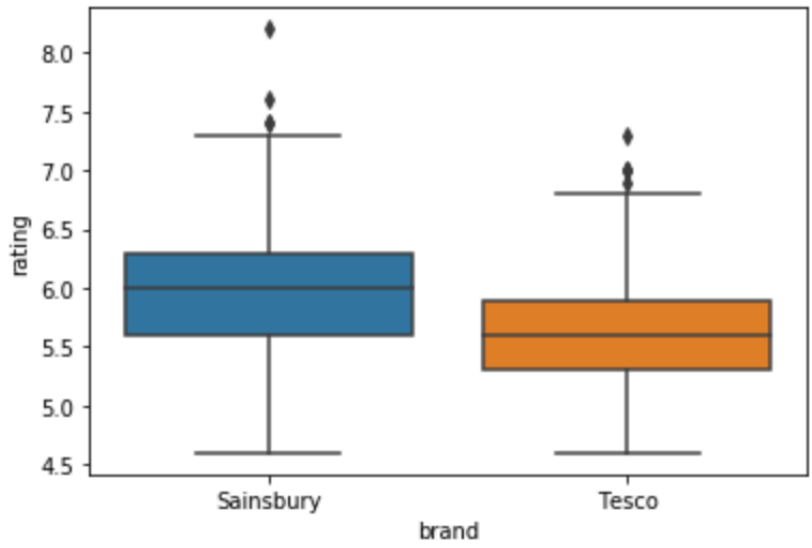


Figure 13. Rating vs brand

4.4 Rating vs name

To explore the relationship between rating and name, stacked histogram is presented in Figure 14 and box plot is exhibited in Figure 15. The key findings are :

1. No significant differences between sub-categories (names)
2. Sainsbury's have a higher upper bound
3. Within Tesco, Tesco Express and Tesco Extra make a big difference
4. Within Sainsbury's, Sainsbury's is better than Sainsbury Local

Therefore, it's worth reclassifying column "name" into four new categories: Tesco Express, Tesco Extra, Sainsbury's, and others. The new feature is named "new name" and it added into the DataFrame.

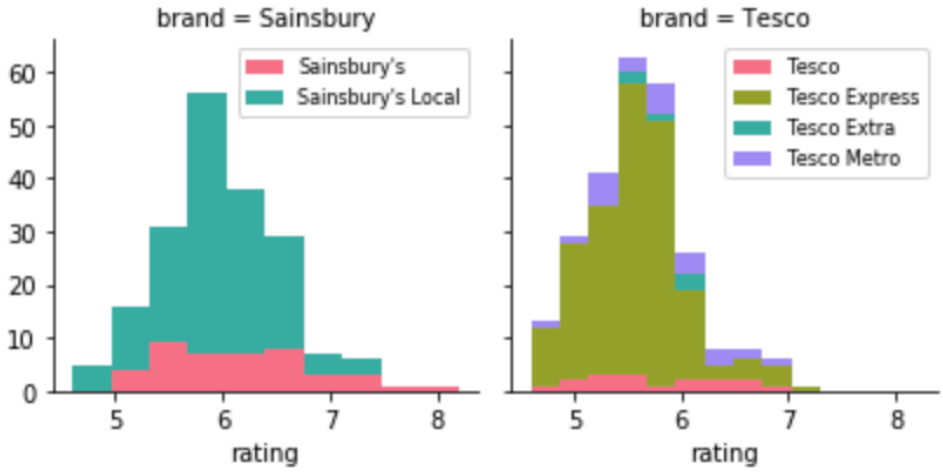


Figure 14. Rating vs name: stacked histogram

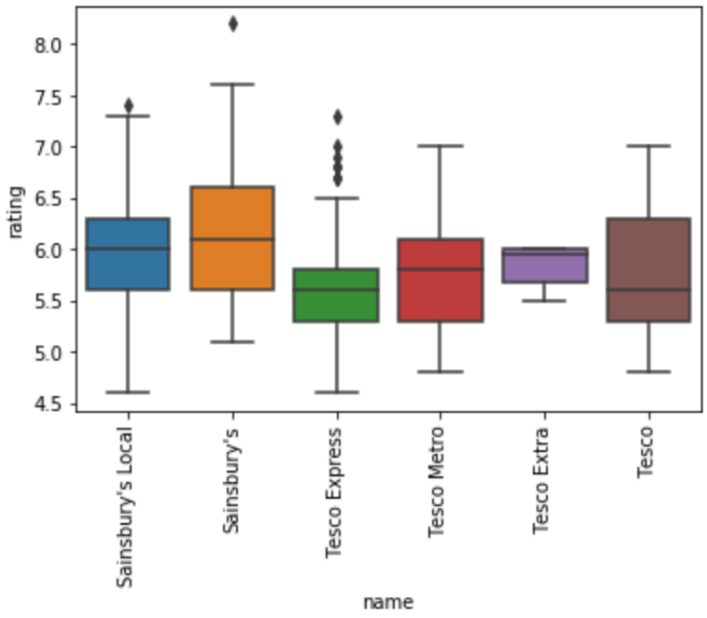


Figure 15. Rating vs name: box plot

4.5 Rating vs categories

To explore the relationship between rating and categories, stacked histogram is presented in Figure 16 and box plot is exhibited in Figure 17. Supermarkets generally has a higher rating than grocery stores. It's a useful feature, however, the previous new name column already contains the information of classification of supermarket, so it is not used as an input for the model.

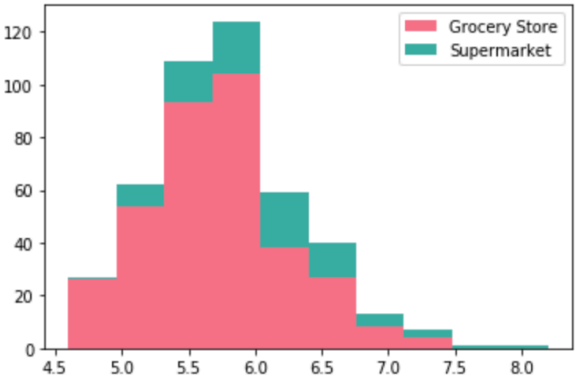


Figure 16. Rating vs categories: stacked histogram

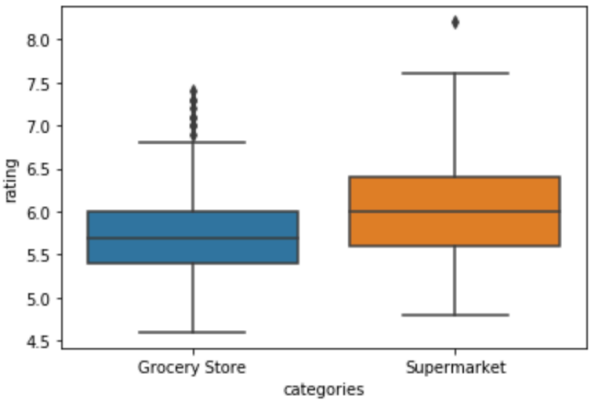


Figure 17. Rating vs categories: box plot

4.6 Rating vs borough

The box plot in Figure 18 show the distributions of ratings in each borough. Although the relationships are not obvious, the following 10 boroughs — 'Richmond upon Thames' , 'Kensington and Chelsea', 'Hammersmith and Fulham', 'Lambeth, Wandsworth', 'Barnet, Brent, Camden', 'Redbridge', 'Newham', 'Kensington and Chelsea\nHammersmith and Fulham', 'Enfield', 'Ealing, Hammersmith and Fulham' — exhibit relative high ratings than others.

Therefore, a new column named "new borough" is created such that boroughs in the 10 above are renamed as "high", and the rest are renamed as "other".

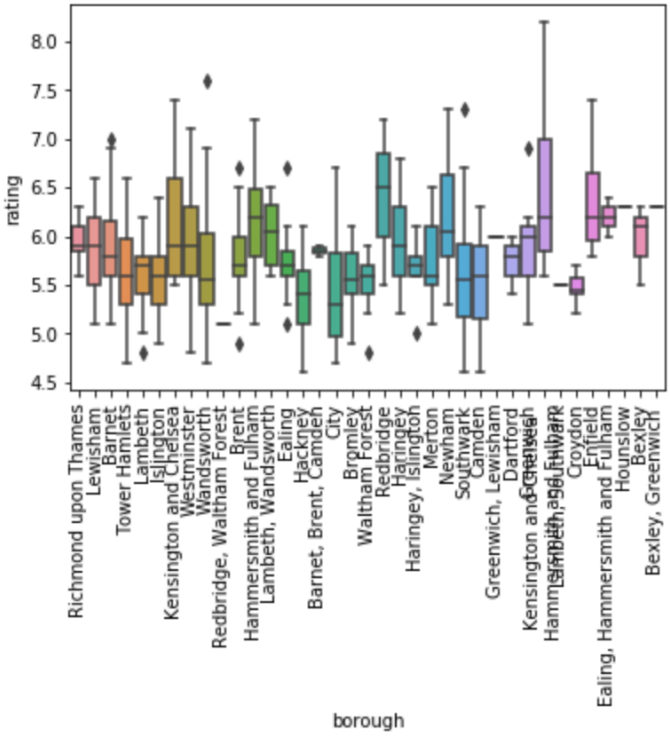


Figure 18. Rating vs borough

4.7 Rating vs house price

To explore the relationship between rating and house price, a scatter plot with regression line is presented in Figure 19. Additionally, the correlation between rating and house prices is at 0.19. So, the key conclusions are :

1. There is weak positive relationship between house price and rating
2. A significant difference can be found in low bounds between stores where house price above 1100000 and those below 11000000

Therefore, a new feature named “house price label” is created such that house prices below 1100000 are called “normal” and those above are called “luxury”.

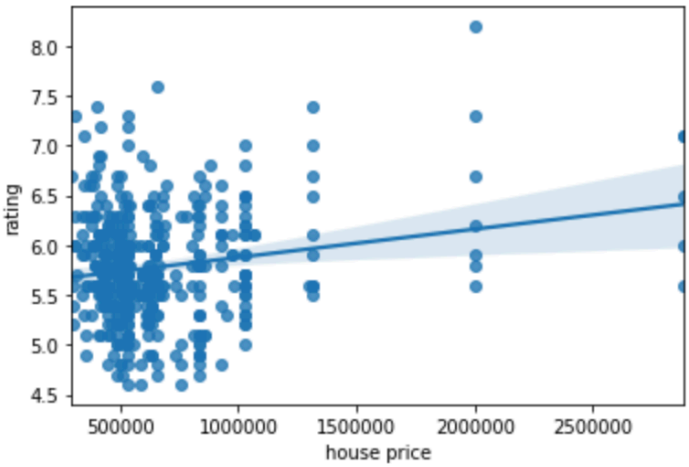


Figure 19. Rating vs house price

4.8 Rating vs total and venue occurrence

Two scatter plots with regression lines are plotted for rating against total and top 1 occurrence respectively in Figure 20 and 21. One can easily argue that there is no significant relationships between rating and these two features. The same conclusions can be drawn for the other 4 venues occurrence columns (no plots presented here) as well.

As a result, these 6 features should not be considered as inputs for the final model.

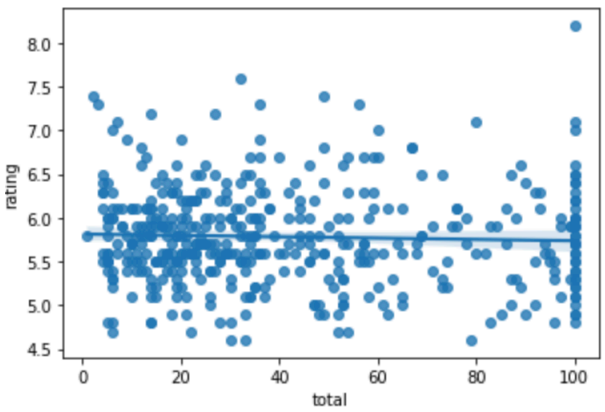


Figure 20. Rating vs total

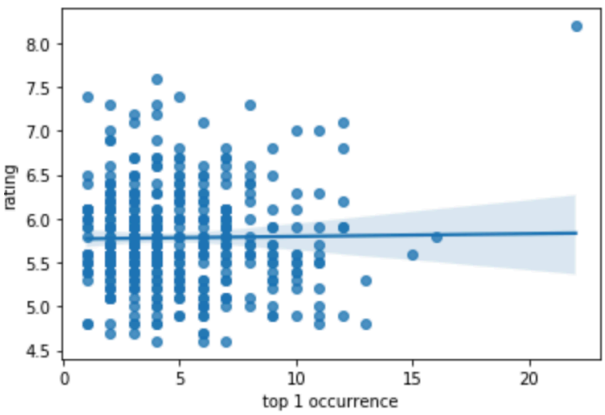


Figure 21. Rating vs top 1 occurrence

4.9 Rating vs top 1 concentration

Business activities around a grocery store is likely to be a valid factor that influence its customer ratings. One way to measure the business activity is to calculate the number of the most popular venue as a percentage of the total venues around the store. The new feature column "top 1 concentration" is created by the method described above and a scatter plot with regression line is presented in Figure 22. As the chart shows, a weak positive relationship exists for rating and top 1 concentration, and a correlation of 0.14 confirms this relationship. Therefore, top 1 concentration could be an interesting feature for modelling.

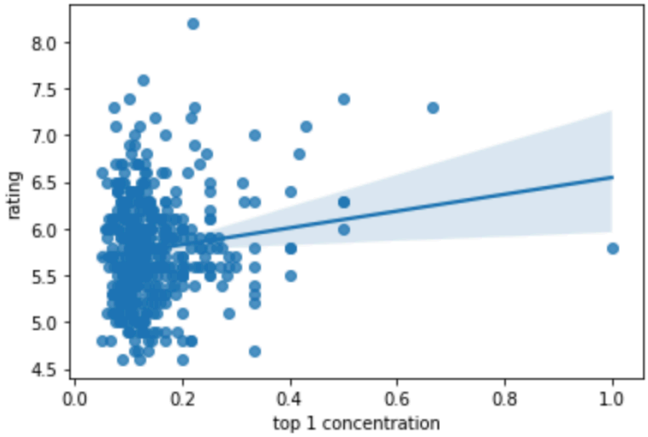


Figure 22. Rating vs top 1 concentration

4.10 Rating vs business diversification

Another way to measure business activity is to calculate the sum of squared percentages of each 5 popular venues to total numbers of venues (Sum((occurrence/total)^2)). The new feature column "business divers" is created by the formula and the regplot chart is exhibited in Figure 23. A very similar relationship can be found as top 1 concentration. Since business diversification uses more information, it is chosen as the input feature in stead of top 1 concentration.

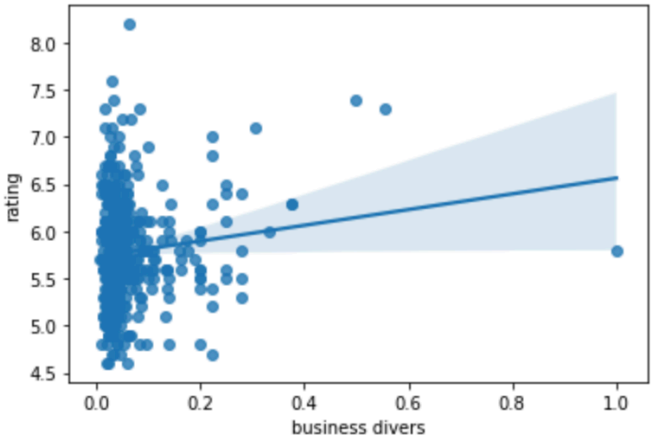


Figure 23. Rating vs business diversification

4.11 Rating vs top 1 category

To explore the relationship between rating and top 1 category, a box plot is presented in Figure 24. Since some categories have small counts, only those with count larger than 5 is examined. Three classes of venues can be identified by grouping ratings:

1. high: 'Grocery Store', 'Hotel' 'Italian Restaurant', 'Clothing Store' 'Supermarket', 'Portuguese Restaurant'
2. low: 'Coffee Shop', 'Café'
3. median: 'Pub', 'Platform'

Therefore, a new columns named "top 1 reclass" is created by the corresponding categories mentioned above. The box plot of these new categories is presented in Figure 25.

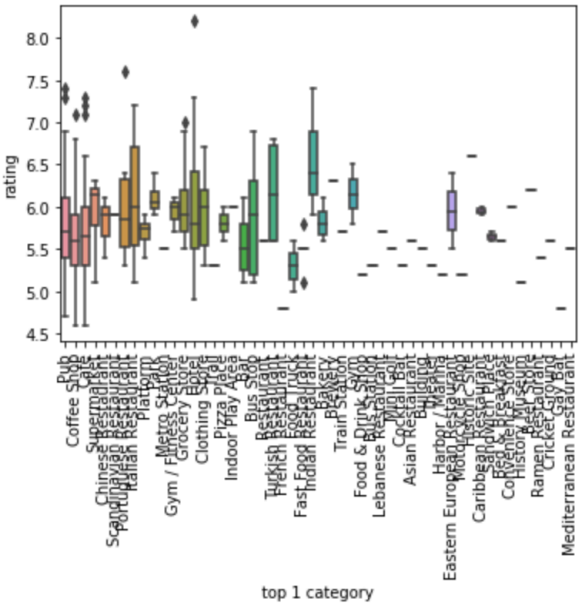


Figure 24. Rating vs top 1 category

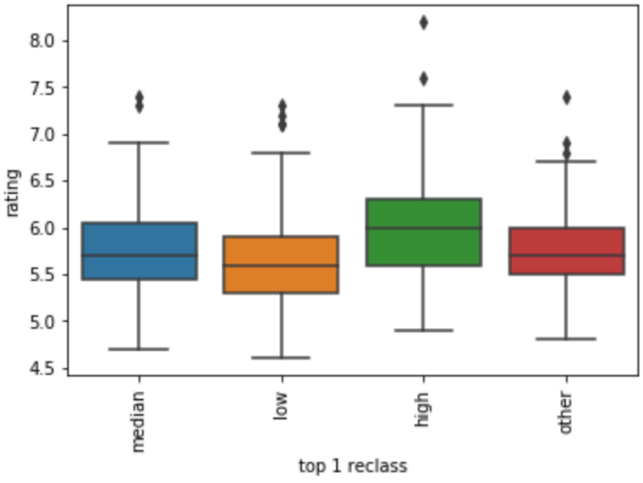


Figure 25. Rating vs top 1 reclass

4.12 Re-examine rating labels with selected features

So far, several relationships with ratings are examined, and below features are found useful:

1. new name (categorical)
2. new borough (categorical)
3. house price label (categorical)
4. business diversification (numerical)
5. top 1 reclass (categorical)

Since the target is rating label, a re-examination of these features with rating labels is necessary to confirm the relationships.

Five stacked bar charts are illustrated from Figure 26 to Figure 30. The observation of these five charts can confirm that the desired relationships still exist with rating labels, and these five feature can be utilised to develop a machine learning model.

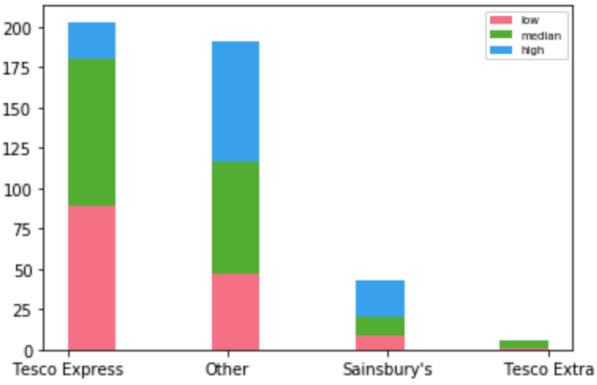


Figure 26. Rating label vs new name

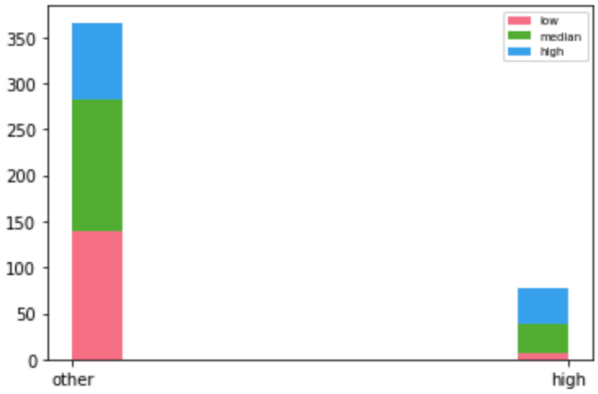


Figure 27. Rating label vs new borough

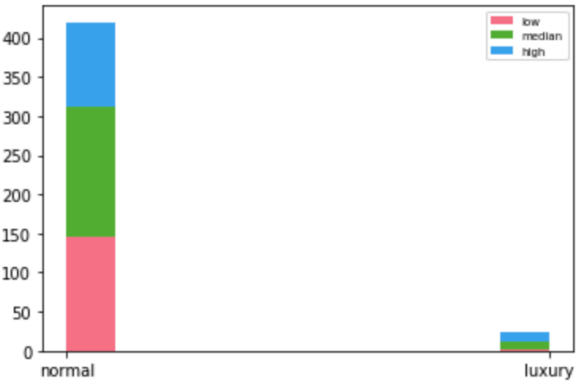


Figure 28. Rating label vs house price label

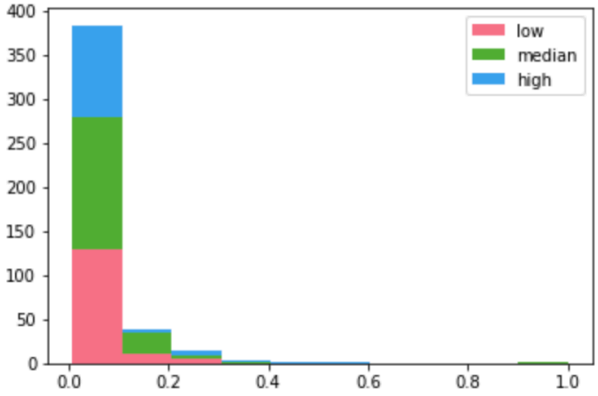


Figure 29. Rating label vs business diversification

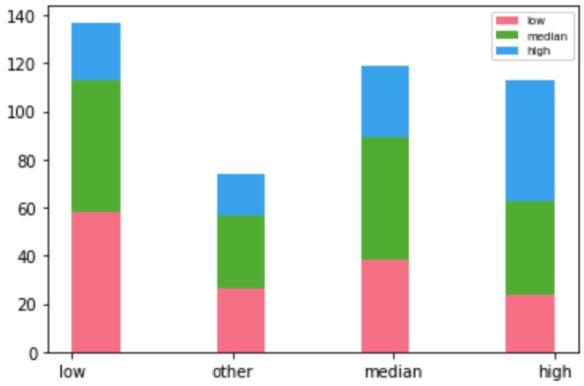


Figure 30. Rating label vs top 1 reclass

5. Machine Learning Models

5.1 Splitting data

In order to form a valid input matrix for model training, all new created categorical features are one-hot encoded. Then, 4 columns of the encoded matrix are dropped because n dummies only need n-1 input features. Finally, the selected features are extracted to form input matrix X, with 443 rows and 9 columns, and rating labels are used as target y, with 443 rows and 1 column. The top 5 rows of X is exhibited in Figure 31:

	business divers	name_Sainsbury's	name_Tesco Express	name_Tesco Extra	borough_high	price_luxury	top1_high	top1_low	top1_median
0	0.043084	0	0	0	1	0	0	0	1
1	0.056122	1	0	0	0	0	0	0	1
2	0.071111	0	1	0	0	0	0	1	0
3	0.081597	0	0	0	0	0	0	1	0
4	0.096939	0	1	0	0	0	0	1	0

Figure 31. Model input data

5.2 Baseline models and untuned models

Before the development of machine learning models, two baseline models are created as the benchmarks for the final model evaluation. These baseline models are rather intuitive dummy classifiers, as the first one randomly guessing labels with equal 33.3% probability and the second model randomly guessing labels with corresponding label frequencies in training data. A model that can consistently outperform these two dummy classifiers must possess some predictive capability of the rating labels.

Beside the two baseline models, five machine learning models, which are decision tree, k nearest neighbours, support vector machine, random forest and XGBoost, are imported from SKLearn and XGBoost packages without parameter-tuning. The 7 models are trained and 5-fold cross validated by the training data, and the validation results with mean accuracy and standard deviation of the predictions are presented in Figure 32.

The results suggest three key conclusions:

1. All machine learning models outperformed the baseline random guessing models. It means that input features and machine learning models do provide extra information to explain the store ratings
2. KNN provides best accuracy with smallest standard deviation of predictions
3. The differences between machine learning models are small

	mean accuracy	std accuracy
KNN	0.478652	0.0296194
Decision Tree	0.465169	0.0485704
SVC	0.460444	0.0378738
XGBoost	0.45383	0.0371145
Random Forest	0.451532	0.0481365
Base Model 2	0.336313	0.0389415
Base Model 1	0.329699	0.052785

Figure 32. CV results of untuned models

5.3 Fined tuned models

The next step is fine-tuning these models by grid search. Decision tree model is tuned by 3 parameters with total 6 combinations, KNN model is tuned by 1 parameter with 18 difference values, SVC model is funded by 2 parameters with total 20 combinations, random forest model is tuned by 4 parameters with total 360 combinations, and XGB model is tuned by 5 parameters with total 1800 combinations. Each combination is tested by 5-fold cross validation and the final results are shown in Figure 33.

	mean accuracy	std accuracy
XGBoost Tuned	0.514888	0.067094
Random Forest Tuned	0.50572	0.0284465
KNN Tuned	0.487666	0.0459845
KNN	0.478652	0.0296194
Decision Tree Tuned	0.471859	0.0299297
Decision Tree	0.465169	0.0485704
SVC	0.460444	0.0378738
SVC Tuned	0.458248	0.0208889
XGBoost	0.45383	0.0371145
Random Forest	0.451532	0.0481365
Base Model 2	0.336313	0.0389415
Base Model 1	0.329699	0.052785

Figure 33. CV results of fine-tuned models

There are four conclusions from the results:

1.Almost all models improves their performances, except for SVC

2.XGB model has the best cross validation score but also the highest standard deviation

3.Random forest shows slightly worse score than XGB, and it has much smaller standard deviation

4.Fine-tuned random forest model is the best model by far

5.4 Feature importance and feature selection

Since three out of the five models are tree-based models, feature importances can be extracted from the model results for feature selection purposes. The two most accurate models, random forest and XGBoost are used and their feature importance are plotted in Figure 34 and Figure 35.

The charts suggest following findings:

1. The primary feature of random forest model is business diversification
2. The primary feature of XGB model is name_Tesco Express
3. borough_high and top_high are both important features for both models
4. name_Tesco Extra, price_luxury and top1_median do not contribute much to the models

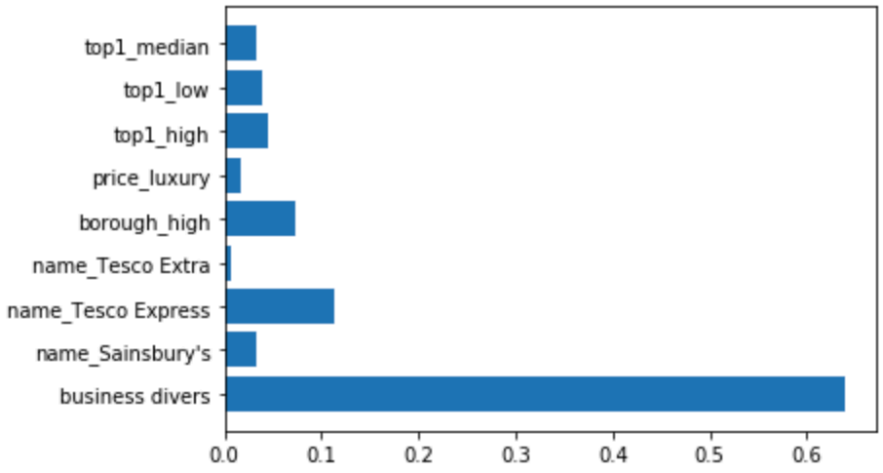


Figure 34. Feature importance of random forest model

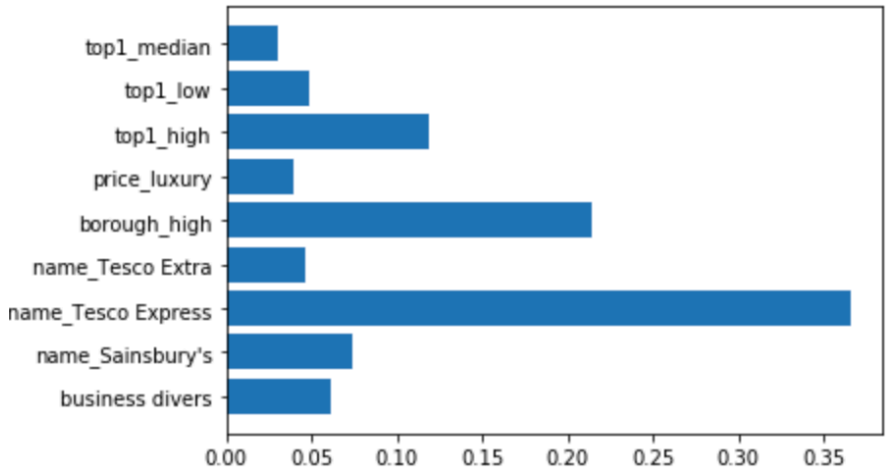


Figure 35. Feature importance of XGB model

5.5 Retaining models with less features

As feature importance results suggest, it is worth dropping the insignificant features and retraining models to compare results. Therefore, columns "name_Tesco Extra", "price_luxury" and "top1_median" are deleted and the five models are retrained. The cross validation results are named as "model Tuned 2" and are shown in Figure 36.

Unfortunately, the results are mixed. The best two models' accuracies are not improved by dropped insignificant features, even though the rest three models are improved slightly. Variances are reduced across models. Because there is no significant increase in model accuracy, the original input will be used for the final model and prediction.

	mean accuracy	std accuracy
XGBoost Tuned	0.514888	0.067094
Random Forest Tuned	0.50572	0.0284465
Random Forest Tuned 2	0.503447	0.0154755
XGBoost Tuned 2	0.501251	0.0548969
KNN Tuned 2	0.49216	0.0345727
KNN Tuned	0.487666	0.0459845
KNN	0.478652	0.0296194
Decision Tree Tuned 2	0.471859	0.0299297
Decision Tree Tuned	0.471859	0.0299297
SVC Tuned 2	0.469484	0.0315936
Decision Tree	0.465169	0.0485704
SVC	0.460444	0.0378738
SVC Tuned	0.458248	0.0208889
XGBoost	0.45383	0.0371145
Random Forest	0.451532	0.0481365
Base Model 2	0.336313	0.0389415
Base Model 1	0.329699	0.052785

Figure 36. CV results of retrained models

5.6 Ensemble model

Considering that all fine-tuned models generates good performances and the differences between models are not huge, the five models are put together to form a final ensemble model. Simple hard voting classifier is applied and the cross validation result shown in Figure 37.

Voting classifier beat all other models with smaller variance in prediction. Ensemble method smoothes out the predictions and improves model performance. The voting classifier is the final model to be used for prediction.

	mean accuracy	std accuracy
Voting	0.530516	0.0170639
XGBoost Tuned	0.514888	0.067094
Random Forest Tuned	0.50572	0.0284465
Random Forest Tuned 2	0.503447	0.0154755
XGBoost Tuned 2	0.501251	0.0548969
KNN Tuned 2	0.49216	0.0345727
KNN Tuned	0.487666	0.0459845
KNN	0.478652	0.0296194
Decision Tree Tuned	0.471859	0.0299297
Decision Tree Tuned 2	0.471859	0.0299297
SVC Tuned 2	0.469484	0.0315936
Decision Tree	0.465169	0.0485704
SVC	0.460444	0.0378738
SVC Tuned	0.458248	0.0208889
XGBoost	0.45383	0.0371145
Random Forest	0.451532	0.0481365
Base Model 2	0.336313	0.0389415
Base Model 1	0.329699	0.052785

Figure 37. CV results of voting classifier

6. Predictions

6.1 Transforming test data

The test dataset need to be transformed by same methods as applied for training set. The transform test input features is.a 148x9 matrix and test label is a 148x1 matrix.

6.2 Prediction results

The voting classifier, five fine-tuned models and the two baseline models all make their predictions, and the final out-of-sample results are exhibited in Figure 38.

The key conclusions are:

1. Voting classifier, as the final model, still generates the best out-of-sample performance with accuracy close to cross validation performance
2. Random forest and XGBoost are surprisingly beaten by other three methods, which indicates the overfitting of the two models in the parameter-tuning process
3. All machine learning models outperform the baseline models

	accuracy
Voting	0.513514
KNN	0.493243
SVC	0.486486
Decision Tree	0.472973
Random Forest	0.452703
XGBoost	0.439189
Base Model 2	0.398649
Base Model 1	0.337838

Figure 38. Prediction results

7. Conclusions and Discussions

To answer the question of what drives the customer ratings of a grocery store, five distinct features are created and utilised, and an ensemble classification model consisted of five different underlying machine learning models are developed. Based on the feature importances of the models and observations from exploratory analysis, the following conclusions are presented:

1. A grocery store located in a neighbourhood with high business diversity tends to have low ratings, or, surrounding commercial activity hurts grocery store ratings.
2. Tesco Express has relatively low ratings comparing to other Tesco stores or Sainsbury's.
3. A grocery store located in Richmond upon Thames, Kensington and Chelsea, Hammersmith and Fulham, Ealing and Redbridge tends to have high ratings.
4. A grocery store that is close to its competitors, i.e. grocery stores and supermarkets, tends to have higher ratings.

5. A neighbourhood with very high house prices tends to have positive impact on grocery store ratings.
6. A voting classifier with decision tree, SVC, random forest, KNN and XGBoost can utilise the features well and produce reasonably good out-of-sample results.

For the management of grocery store brands like Tesco and Sainsbury's, the above conclusions can be incorporated into their process of choosing a store location to improve the level of customer satisfaction. Conversely, the management team may consider stripping out the impact of these exogenous factors when evaluating a store's or a store manager's performance so that the assessment is more justified.

8. Future Works

Although several exogenous factor are uncovered to explain the customer ratings of a grocery store, more works can be done for this topic. Other exogenous factors such as ratings of neighbours, distance to nearest competitors, and a higher level classification of venues (e.g. group all kinds of restaurants as one class) could be potentially predictive for the ratings of grocery store. Endogenous factors such as completeness of goods of a store, service level of staff, and internal rating of store manager could also be powerful determinants of the ratings. Researchers who have access to these data can try to build a better model based on the suggested factors.