RESTAURANTS HEALTH SCORE DATA IN SAN FRANCISCO

Project Report by
Group - 1
Abhinandan Somachetty
Leela Kali Manasa Mandadi
Swathi Bommina

BAN 620 – Data Mining

—

Balaraman Rajan

DATA SOURCE

We are using Restaurants Health Score Data in San Francisco. The source of the data is from the website: https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i?row_index=0

Restaurants health score Dataset contains the basic details about the restaurant and the inspection details of the restaurant. The Health Department has developed an inspection report and scoring system. After conducting an inspection of the facility, the Health Inspector calculates a score based on the violations observed. Violations can fall into:
High risk category: Records specific violations that directly relate to the transmission of foodborne illnesses, the adulteration of food products and the contamination of food-contact surfaces.
Moderate risk category: Records specific violations that are of a moderate risk to the public health and safety.
Low risk category: Records violations that are low risk or have no immediate risk to the public health & safety.

ATTRIBUTES OF THE DATASET

Business_ID	It is a unique identifier to the restaurant.
Business_Name	Name of the restaurant.
Business_address	It is the address of the restaurant.
Business_City	City where the restaurant is located.
Business_State	State where the restaurant is located.
Business_postal_code	Postal code of the restaurant region.
Business_latitude	The latitude of the restaurant location.
Business_longitude	It’s the longitude of the restaurant location.
Business_Location	Coordinates of the restaurant’s location.
Business_phone_number	The phone number of the restaurant.
Inspection_id	Unique numbers are assigned to each inspection.
Inspection_date	Date when the inspection taken place in a restaurant.

Inspection_score	Score given to the restaurant according to the Healthy maintenance of the restaurant between 0 to 100. 100 signifies highest score and 0 being the lowest score.
Inspection_type	Type of inspection happened on a particular date. It can be Routine - Unscheduled, Reinspection/Follow-up, etc.
Violation_id	Different numbers are assigned for different types of violations. If a restaurant is violating any rules, then they will add a unique number in this column, which corresponds to the rule they are violating.
Violation_description	Description about the violations that the restaurant made.
Risk_category	This take values like Low risk, high risk and moderate risk which signifies how unhealthy the restaurant is.

DESCRIPTION:

The motivation of the project is to predict the risk factor with the running of a restaurant and various factors which contribute to risk and knowing the necessary sectors where focus and improvement is required for the Health Department to take necessary actions on the restaurants that are violating the measurements.

DATA CLEANING:

Columns with no data will be imputed with '0' in case of numeric and 'unknown' in case of the categorical variables. If risk category columns data is null then removed that row as we don't know to much category it below to (Low Risk, High Risk, Medium Risk).

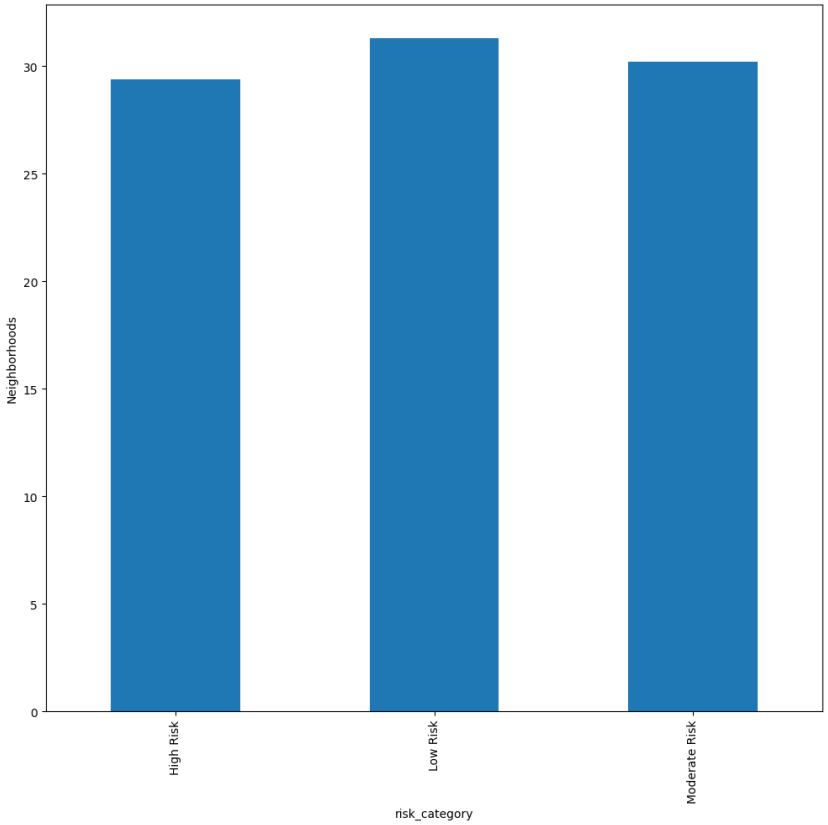
Numerical columns are standardised to get them to comparable scales and use this data to produce best model in KNN.

SUMMARY OF THE DATA SET:

- Total number of inspections done in all the restaurants in San Francisco is 53973.
- The maximum inspection score given by the health department is 100 and the minimum score showing as 45.
- The average inspection score of the restaurants is 86.22.

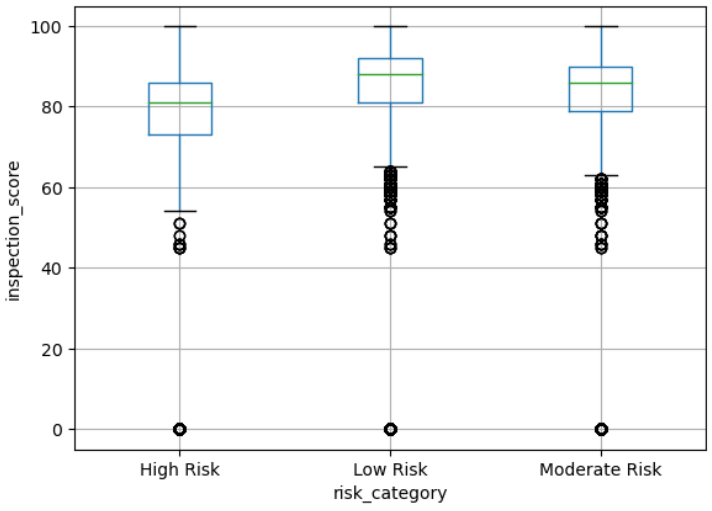
RISK CATEGORY VS NEIGHBOURHOOD

As the graph shown below, the number of low risk category restaurants are more that the moderate risk and high risk categories.



RISK CATEGORY VS INSPECTION SCORE:

Low risk category has the smallest interquartile range with the highest medium of all the categories. It means the inspection score is directly related to the risk category. The outliers represent the number of violations. This indicated that the low-risk category is higher (with more number of violations). The moderate risk category has the smallest IQR with (2^{nd}) highest number of violations. The violations in low-risk category can be rectified by the restaurants with doesn't have serious impact. However, the moderate risk has some severe impact which can make a customer sick. Finally, the high-risk category has very few outliers. This category is the most dangerous to human health. Its IQR is large. Also, the maximum value shows as 100 which means even a restaurant with 100 inspection score has a chance that they had a high-risk violation.

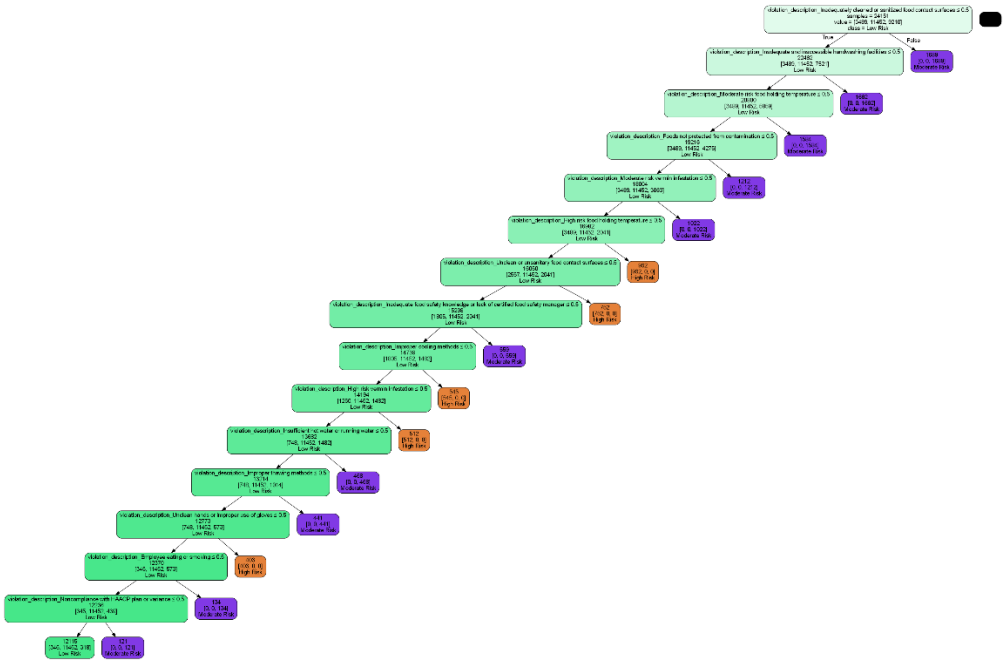


DECISION TREE:

The classification as shown below in done on the basis of violation type.

Each category of the risk score has different descriptions of violation.

Example: violation description – vermin infestation >= 0.5 – then it falls under High risk category



CONFUSION MATRIX:

Considering ‘0’ as low risk, ‘1’ as moderate risk and ‘2’ as high risk:

The model classified 2113 as low risk correctly but 223 as moderate risk incorrectly.

The model correctly classified 7660 moderate risk categories with no incorrectly predictions.

The model predicted correctly for 5903 high risk category but incorrectly predicted 203 as moderate risk.

The prediction accuracy is 97.35%.

K-NEAREST NEIGHBORS (kNN):

The best k value is at k=1 with an accuracy of 99.89%. However, it also tends to be sensitive and noise.

Hence, considering the best k as 3 with an accuracy of 99.85%

	k	accuracy
0	1	0.998944
1	2	0.998696
2	3	0.998510
3	4	0.997888
4	5	0.997764
5	6	0.997205
6	7	0.996460
7	8	0.996212
8	9	0.996212
9	10	0.995466
10	11	0.995280
11	12	0.995156
12	13	0.994721
13	14	0.994597
14	15	0.994411
15	16	0.994038
16	17	0.993727
17	18	0.993603
18	19	0.993293
19	20	0.993044
20	21	0.993044
21	22	0.992796
22	23	0.992361
23	24	0.991492
24	25	0.991430
25	26	0.991119
26	27	0.991119
27	28	0.990312

RECOMMENDATION:

Though the number of low-risk scored restaurants are higher compared to the high risk and moderate risk, the difference is not too much due to which we can assume that higher number of restaurants are with unhealthy environment and the high-risk category restaurants have a score of 100 which is more concerning as the high risk category is for very severe reasons. The accuracy on both the models is high which shows that there are more than enough restaurants that are historically persistent violators of the San Francisco Department of Health rules. This can help the health department with their inspections on the restaurants that are violating the code.