Classification of Food Inspections and Prediction of Energy Benchmarking Data in Chicago

Riya N. Bhutada

Northeastern Illinois University, rnbhutada@neiu.edu

ABSTRACT

I set out to examine the food inspection reports in Chicago based on Facility Types, Risk of facility that depends on how frequent the inspections should be carried, number of violations the report specifies, and other violations related binary attributes. I did this using decision trees and logistic regression to classify whether the inspection report passed or failed. I found that using decision trees, the accuracy score was 0.936 and 0.952 for logistic regression. However, the train score for decision tree came out to be 0.969, which was more than test score. I also set out to examine the Energy Benchmarking 2021 reported in 2022 data of Chicago based mainly on the Type of Property, Source EUI(Energy Use Intensity) used, Site EUI(Energy Use Intensity), Weather Source EUI, Weather Site EUI, ENERGY STAR score that assesses a property's overall energy performance and Energy Rating(the-zero-to-four-star-rating). I did this using K-Means clustering algorithm to see if there were patterns in Energy Rating. When plotted Source EUI against Site EUI, I found clusters of data points that closely represented the Rating values from 0 to 4.

1 DATASET

1.1 Chicago Food Inspections Dataset

This information is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010 to the present. Inspections are performed by the Chicago Department of Public Health's Food Protection Program. Since Food Inspection procedures changed in 2018, to avoid any change in data behavior, I deleted all the old data before 2018. The dataset contains demographic information such as Address, City, State, ZIP, Latitude, Longitude, Location and other such as DBA Name i.e. Database name, Inspection ID, License Number which I chose to discard them all because they carry no meaning in terms of data for training and classification. Below is a table of other significant data.

Table 1: Chicago Food Inspections Dataset

Attribute/Column Name	Meaning	Used in Input Set?	Comments
Name (Text)	Business Name	No	
Inspection Date (Date)		No	
Facility Type (Text)	Each establishment is described by one of the following: bakery, banquet hall, caterer, coffee shop, day care center, gas station, restaurant, etc.	Yes	Categorical
Risk (Object)	Tells how adversely public's health can be affected. The	Yes	Categorical

	frequency of inspection is tied to this risk, with risk 1 establishments inspected most frequently and risk 3 least frequently.		
Inspection Type (Object)	Type of inspection: Complaint-based, task-force, re-inspection, canvass, license	Yes	Categorical
Results	An inspection can ***pass***, pass with conditions or ***fail***.	Yes (Target attribute)	Since 'Pass with conditions' value records were found to be corrected during inspection by the business owners, it held no value for current purpose, so I decided to discard all those records.
Violations (Text)	One or more of 45 distinct violations that establishments received along with comments.	No	I extracted number of violations a row contained in this attribute's value for all records and found those violations that occurred most number of times in the entire dataset. I encoded these as new binary attributes for all records.
NumOfViolations (Binary)	Total number of violations	Yes	Derived attribute
CertifiedManagerOnSite (Binary)	If the manager was present during inspection and had a certification. 0 for No, 1 for Yes	Yes	Derived attribute
EmployeeHealthPolicyPresent (Binary)	0 for No, 1 for Yes	Yes	Derived attribute
ResponseToIllnessEvent (Binary)	Whether guidelines were recorded for response to any diarrheal event or any form of illness. 0 for No and 1 for Yes.	Yes	Derived attribute
PestControlRecordPresent (Binary)	If pests, rodents were found and pest control record was present. 0 for No and 1 for Yes.	Yes	Derived attribute

1.2 Chicago Energy Benchmarking 2021 Data Reported in 2022 Dataset

The dataset contains demographic information such as Address, City, State, ZIP, Latitude, Longitude, Location and other such as Name, Year, ID, Property Name, Reporting Status(whether building submitted its report for that year), Community Area, Exempt from Energy Rating(Discarded all rows where value is Yes, since it beats the purpose of clustering), Year Built, etc. I chose to discard them all to avoid skewing since they did not held any meaning for clustering. Other attributes were All other fuel Use (annual amount of fuel use other than electricity), Total GHG Emissions(Total Green House Gas Emissions), GHG Intensity(Emissions/Gross Floor Area), District Chilled Water Use, Natural Gas Use, Electricity Use. For these, I found no meaningful clustering other that clusters already formed using below in table fields, so I did not use them for fitting and visualizing the data.

Table 2: Chicago Energy Benchmarking Dataset

Attribute/Column Name	Meaning
Property Type (Object) Categorical.	Office, Hospital, Multifamily Housing, K-12 School, Hotel, Restaurant, Retail Store, Senior Living Community, Grocery Store, Mixed Use Property, etc.

Attribute/Column Name	Meaning
Chicago Energy Rating (Float) Categorical values – 1, 1.5, 2, 2.5, 3, 3.5 and 4	The zero-to-four-star Chicago Energy Rating assigned to the building. A building with zero stars did not submit a report, or did submit a report but was missing required information. All other buildings receive between one and four stars, with four stars reflecting the highest performance.
Chicago Energy Score (Int)	1-100 rating that assesses a property's overall energy performance, based on national data to control for differences among climate, building uses, and operations. A score of 50 represents the national median.
Site EUI (kBtu/sq ft) (Float)	Site Energy Use Intensity is a property's Site Energy Use divided by its gross floor area. Use is the annual amount of all the energy consumed by the property on-site.
Source EUI (kBtu/sq ft) (Float)	Property's Source Energy Use divided by its gross floor area. Source Energy Use is the annual energy used to operate the property, including losses from generation, transmission, & distribution.
Weather Normalized Site EUI (kBtu/sq ft)	WN Site Energy divided by its gross floor area (in square feet). WN Site Energy is the Site Energy Use the property would have consumed during 30-year average weather conditions
Weather Normalized Source EUI (kBtu/sq ft)	WN Source Energy divided by its gross floor area. WN Source Energy is the Source Energy Use the property would have consumed during 30-year average weather conditions.

2 METHODOLOGY

I examined the Food Inspection reports dataset based on Facility Types, Risk of facility (that depends on how frequent the inspections should be carried), number of violations the report specifies, and other violations related binary attributes (calculated from Violations column) as mentioned in Dataset table. I used Scikit library's **DecisionTreeClassifier** Class and **LogisticRegression** class to classify the Results target attribute into Pass (i.e. 1) or Fail (i.e. 0) category. I also set out to examine the Energy Benchmarking 2021 reported in 2022 data of Chicago based on the Type of Property, Source EUI(Energy Use Intensity) used, Site EUI(Energy Use Intensity), Weather Source EUI, Weather Site EUI, ENERGY STAR score and Energy Rating. I did this using Scikit library's **KMeans** clustering algorithm to find patterns on how the data distributes across the energy parameters. For preparing and handling the data, I used **Pandas** and **Numpy** libraries. Since DecisionTreeClassifier and LogisticRegression learners work with binary or numerical data, I chose ordinal encoding to handle all categorical data. The models predicted better with ordinal encoding over one-hot encoding. For data visualization, I used plotly, matplotlib and seaborn.

3 RESULTS

	MeanAbsolute Error	***Accuracy Score***	***CLF Score***
Decision Tree	0.063	0.936	Train 0.969, Test 0.936
Linear Reg	0.047	0.952	Train 0.956, Test 0.952

Output Label	***Precision***	***Recall***	***F1score***	***Support***
Fail (0)	0.86	0.73	0.79	219
Pass (1)	0.95	0.98	0.96	1130
Fail (0)	0.98	0.72	0.83	219
Pass (1)	0.95	1.00	0.97	1130

From above table, Linear Regression shows a fair accuracy score and CLF score on both train and test data. For out of total 1349 test records, 1130 are positives. Recall score for both classifiers is 0.98 and 1.0, which means they predicted 97% of positive records correctly with a precision of 0.95 for positive class. Overall, LinearRegression model predicted values slightly better than DecisionTree.

The clusters predicted by K-means model show similarity with the Energy Score distribution with respect to the Energy Use Intensity features. From the right figure, for a Site and source EUI of 50kBtu, the bottom most cluster represents an Energy Score of close to 100. We can conclude, lower the Energy Use, higher is the Energy Score and better is the performance.

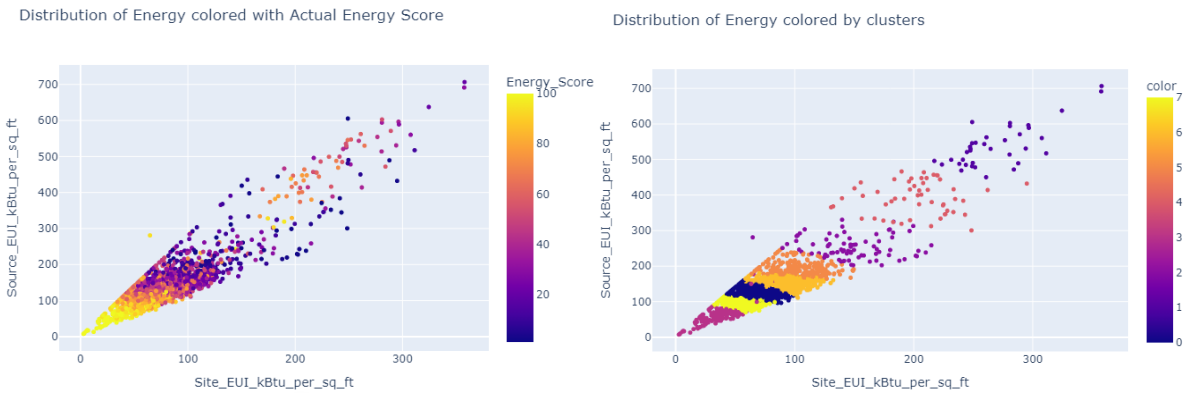

Figure 1: Site EUI VS Source EUI, Energy Score on left plot VS Clusters on right plot.

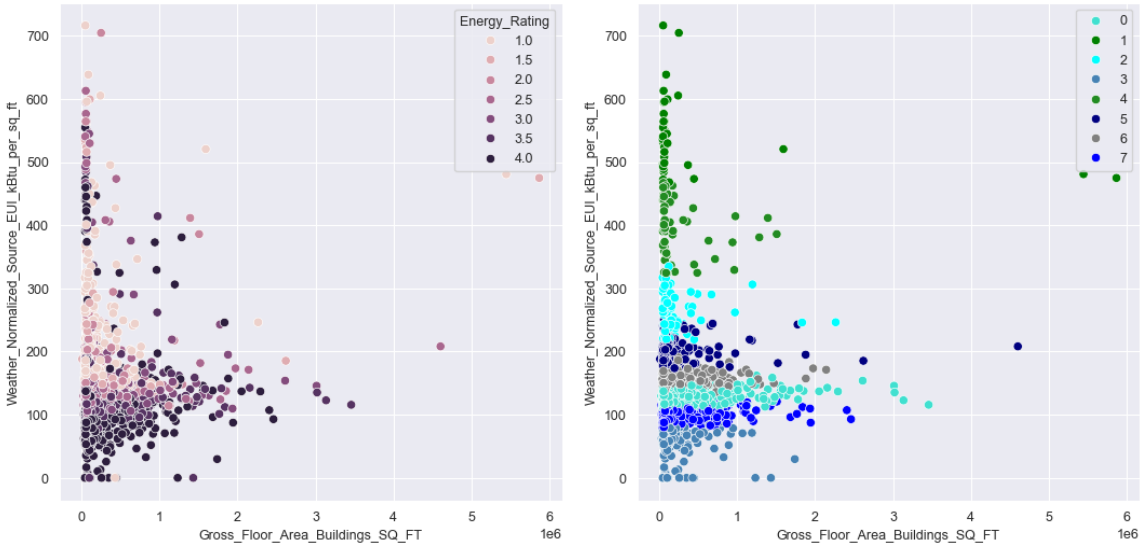

Figure 2: Floor Area VS Weather Normalized Source EUI, Energy Score on left plot VS Clusters on right plot

Fig 2 also shows clusters formed could predict Energy Rating. And that clusters with low Gross Floor Area and low Weather Normalized EUI tend to have high Energy Rating.

4 CONCLUSION

The lowest cluster from Fig 1 seems to contain all the buildings with low energy use, high Energy Score and therefore highest Energy Rating of 4.0. The uppermost cluster seems to be formed of high Site EUI, high Source EUI and lowest Energy Score. But, the left diagram shows some points with high energy use and close to high Energy Score. These could be outliers with other energy data affecting its Score. Mostly, all clusters seem to fit the Energy Score ranges.

Both classifiers learned and predicted the positive records pretty well. This could also be because of the total positive records present in the data set being much more than the negative class samples. Decision Tree classifier seemed to overfit the data a little since the test score was less than training score. Logistic Regression performed better with overall good metrics. To improvise on the metrics, more data of class ‘Fail’ could be collected, so that the model learns well and increases Precision and Recall Score.