Predicting the Grade of Restaurants in New York City

Sneha Sunil

2022-11-25

Contents

This is a data analysis project for DSCI 522 (Data Science workflows); a course in the Master of Data Science program at the University of British Columbia.

Contributors

- Nikita Susan Easow
- Sneha Sunil
- Edward (Yukun) Zhang
- Lauren Zung

Introduction

After the state proclaimed the restoration of indoor dining during the COVID 19 era, hundreds of new restaurants have opened throughout New York City (Eater NY, 2020). Now that things are getting back to normal as the restrictions set by government are lifted and people are returning back to dining out as the hotel sector reopens, the general safety of restaurants has taken on utmost importance in light of the current state of affairs of COVID. The standards that health inspectors use for grading will probably need to be revised because health rules have become more stringent in order to curb the pandemic. The overall plan used by health inspection is as follows, though it may vary by state:

- **GRADE A:** The restaurant is clean, up to code, and free of violations.
- **GRADE B:** The restaurant has some issues that must be fixed.
- **GRADE C:** The restaurant is a public risk and on verge of closure.

(Source: SmartSense, 2018)

As data scientists, we're curious about how we can evaluate and predict a restaurant's general level of quality so that we can provide recommendation on the right restaurants which can be dined in safely by classifying the restaurant as "good" or "poor" (in our case, Grade A vs.Grade B/C). As we have access to restaurant data for the New York City, we would like to concentrate our analysis on forecasting the grading of restaurants as Good or Poor for specific NYC locations, with plans to eventually expand to other metropolitan regions.

We believe that our effort could be useful to the residents or tourists in the NY city and this could be a one stop solution for people who are looking to dine in without having to worry about the quality.

Research question :

`Can we predict the grade for a restaurant (Grade A or F) given different metrics describing their health violations during a routine inspection?`

Besides this main research question, our analysis would also like to address some interesting sub-questions given below:

- Which cuisines are more likely to be graded A in NYC?
- Which cuisines are more likely to be graded B or C in NYC?
- Which borough in NYC seems to have the best restaurants?
- Which borough in NYC seems to have the most restaurants with the most severe violations?
- What words in a violation description contribute most to whether a restaurant is graded A or B/C?

Data

The data set that we are using in our analysis for the restaurant grading, DOHMH New York City Restaurant Inspection Results, is sourced from NYC OpenData Portal. It was obtained from the tidytuesday repository by Thomas Mock. The original data set can be found here.

Summary -

The data includes all of the violation citation from the restaurant inspections held in New York City from 2012 to 2018. Each row represents a restaurant that has undergone a health inspection which has the information about each establishment including the restaurant name, phone number, location (borough, building number, street, zip code), cuisine type, and also the details about the inspection itself which includes date, violation code, description, whether there were any violations cited, whether they were critical, etc.). The restaurants may receive an official grading of A, B, or C; alternatively, they may receive a Z or P for an evaluation that is still pending. Here is a complete dictionary of the data can be found here.

Exploratory Data Analysis

We performed the exploratory data analysis on the restaurant dataset and we noticed that the total strength of inspections were 3,00,000, out of which only 151,451 had a value filled in for the grade column that we are interested in.

Table 1.1 Counts of inspections in the training data by class.

Grades	Number of Inspections
A	87597
F	18621

As we can see from the above table, there is a significant class imbalance of which 79.8% inspections are graded as A. Hence, we've chosen to approach our research question as a binary classification problem , where the outcome will determine whether a restaurant should be graded as A(Pass) or F(Fail - clubbing the B and C grades) based on the standards that are set. We have excluded the restaurants with "PENDING" grade and will be considering in the deployment data in order to predict the grade using our model.

We performed the rest our analysis on the training data where we split the initial data set such that 75% of the data will go to our train data and the rest 25% will be for validating the performance of the model on

restaurants which hasn't been graded yet based on the inspection features that we have.

One important point to note here is that when we grouped the restaurants by `camis` feature, we could see that many restaurants were inspected more than once and we are not sure on whether the restaurants share the same name or if some restaurants have changes their name in between 2012 and 2018. Since we could not incorporate this issue while modelling, we have added it to the limitations.

Fig 1 :

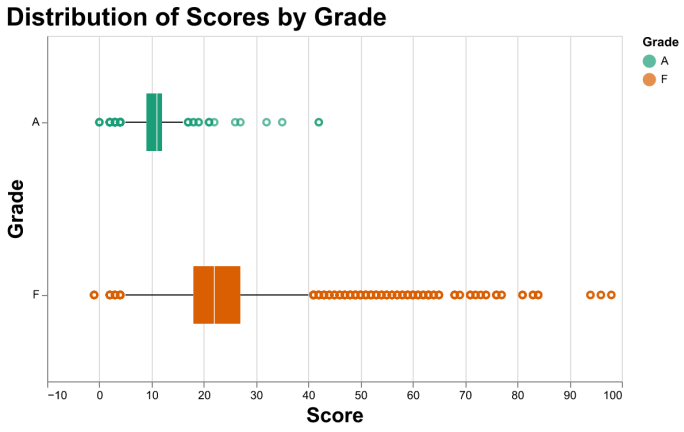

Figure 1: Figure 1. Distribution of Scores by Grade

From the above plot, we can see that the grade F restaurants are associated with higher scores on an average when compared to that of the graded A restaurants even though some of them have low scores. We can conclude that the scores are higher for more critical health violations, but we cannot generalize as we do not see a hard cut off for when a restaurant is graded A or not.

Fig 2 :

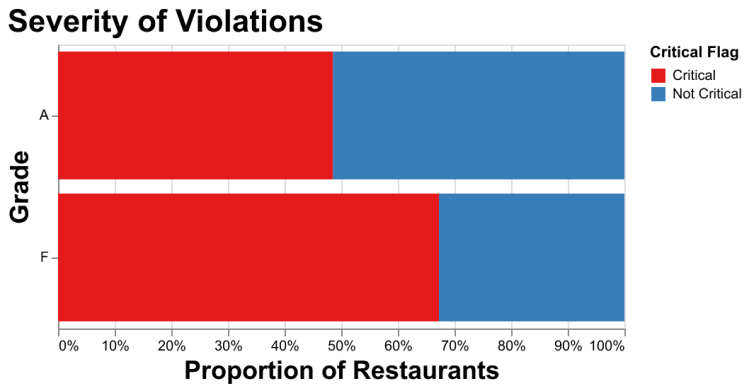

Figure 2: Figure 2. Severity of Violations

In the above figure, the plot suggests that the Grade F restaurants receive proportionally more red flags related to violations than Grade A restaurants do, but it is interesting to see that even grade A restaurants have had some critical violations. It will be intriguing to see if our model can determine whether the seriousness of a violation actually counts for grading because it is unclear what the threshold for a "major" violation is.

Fig 3 :

We should be able to dine in any neighborhood of NYC because all of the boroughs have a majority of Grade

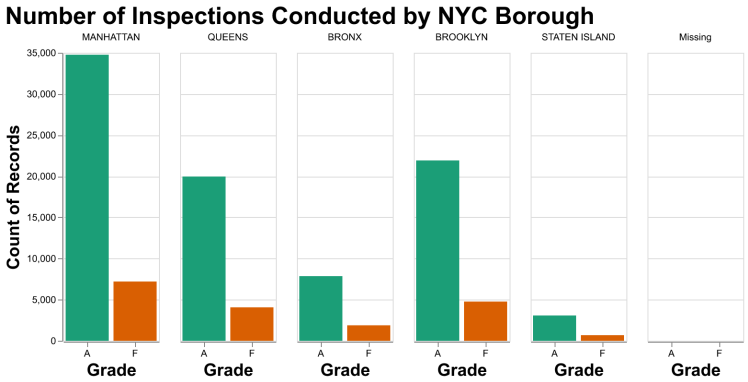

Figure 3: Figure 3. Number of inspections conducted by NYC Borough

A restaurants. It is clear that the majority of the inspections took place in Manhattan, which also has the highest concentration of restaurants receiving a Grade F rating among the other boroughs.

The complete EDA including the above figures and tables can be found here.

Interpretation of the Results & Discussion

Considering our research question is a binary classification model, we picked two different models to perform training using our data - logistic regression (balanced and imbalanced) and support vector machines (balanced and imbalanced). In addition to this, our trained models are also compared against a baseline model dummy classifier ' so we could measure how our models performed.

Note - We downsampled our data in order to reduce the training time.

As we can see from the below table, all models are performing better than the baseline model. In light of our data set's class imbalance, we expected the balanced models to perform better in cross validation. Comparing the scores of each model, we can see that the balanced are doing much better than their counterparts. In addition to that, the validation scores of the logistic regression model is higher than that of the support vector model. Hence we choose the balanced logistic regression as our classifier to train the dataset. Additionally, the logistic regression model's validation scores are higher than that of the support vector model. Hence, we chose the balanced logistic regression as our classifier to train the data set.

Table 2.1. Mean train and validation scores from each model.

	dummy	logreg	svc	logreg_bal	svc_bal
test_accuracy	0.82	0.99	0.99	0.99	0.99
train_accuracy	0.82	0.99	0.99	0.99	0.99
test_precision	0.00	1.00	1.00	0.99	1.00
train_precision	0.00	1.00	1.00	0.99	1.00
test_recall	0.00	0.92	0.95	0.95	0.96
train_recall	0.00	0.93	0.95	0.96	0.96
test_f1	0.00	0.96	0.97	0.97	0.98
train_f1	0.00	0.96	0.97	0.97	0.98

Hyper parameter tuning results on logreg -

Table 2.3. Mean train and cross-validation scores (5-fold) for balanced logistic regression, optimizing F1 score.

rank_test_score	mean_train_score	mean_test_score	param_logisticregression__C	param_columntransformer__countvectorizer__max_features	param_columntransformer__onehotencoder__max_categories
1	0.976266	0.974667	0.024947	130	47
2	0.976985	0.974572	0.065358	155	26
3	0.976694	0.974478	1.552264	40	12
4	0.976896	0.974471	0.254745	114	19
5	0.976118	0.974468	0.023420	119	40
6	0.976860	0.974378	0.610391	159	16
7	0.977217	0.974291	0.086647	181	48
8	0.976930	0.974175	0.384748	65	21
9	0.976740	0.974097	22.219381	177	17
10	0.974403	0.973671	0.011290	69	13
11	0.977036	0.973505	90.791419	313	22
12	0.973506	0.973072	0.003570	77	13
13	0.973132	0.972273	0.002901	279	23
14	0.976662	0.972095	0.244492	254	46
15	0.976174	0.971798	0.292110	107	32
16	0.976121	0.971600	0.317785	2	37
17	0.975967	0.971120	0.427990	225	35
18	0.975226	0.969137	229.263532	272	37
19	0.975082	0.969098	20.649135	99	48
20	0.975158	0.968804	157.708444	256	49

By using Random Search CV, we optimized the hyper parameters of the balanced logistic regression model to the following :

- **C** - 0.024947
- **max_features** - 130
- **max_categories** - 47

Train/validation scores from the best model -

After performing cross validation on the training set using our optimized hyper parameters for our model, we got a good f1 score of 0.975. Both the precision and recall scores for the validation set are high, indicating that the model is accurate about its prediction (whether a restaurant will receive an F grade or not).

Table 2.4. Mean and standard deviation of train and validation scores for the balanced logistic regression model. Parameters: C = 0.024946842198440632, max_features = 130, max_categories = 47

	mean	std
test_accuracy	0.991000	0.001000
train_accuracy	0.992000	0.000000
test_precision	0.999000	0.001000
train_precision	0.999000	0.000000
test_recall	0.951000	0.007000
train_recall	0.954000	0.001000
test_f1	0.975000	0.003000
train_f1	0.976000	0.001000

Classification report from the best model on the test set -

Table 2.5. Classification report on the test set.

	precision	recall	f1-score	support
A	0.988870	0.999879	0.994344	8264.000000
F	0.999392	0.946429	0.972189	1736.000000
accuracy	0.990600	0.990600	0.990600	0.990600
macro avg	0.994131	0.973154	0.983267	10000.000000
weighted avg	0.990697	0.990600	0.990498	10000.000000

Confusion matrices from the best model on train and test set

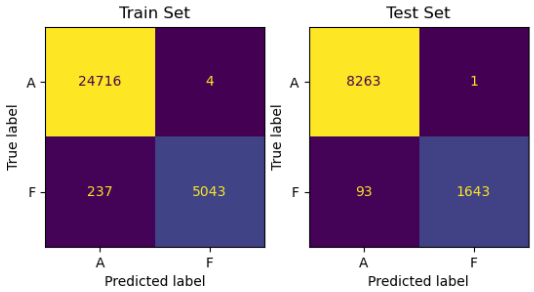

Figure 4: Figure 4. Confusion matrices from the best model on train and test set

PR curve from test set -

The below PR curve depicts that if we keep our new threshold (after balancing), we have an optimum solution with high precision and high recall value. If this threshold is lowered, the recall score also could get lesser and we may not be successful in classifying the restaurants correctly to the GRADE F class.

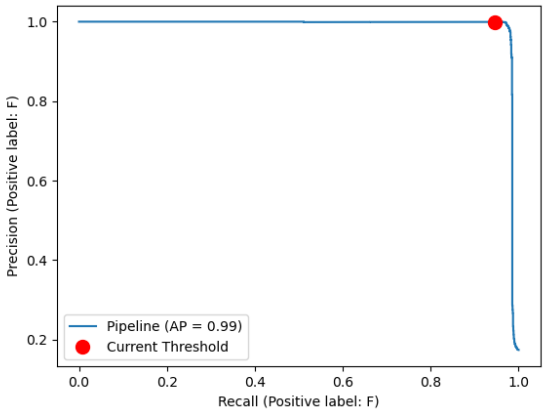

Figure 5: Figure 5. PR curve from test set

ROC curve from test set -

The ROC curve is a plot between the False Positive Rate and the True Positive Rate. Through this graph we find the area under the curve (AUC) is 1.00. This is the optimum value for an AUC and tells us that the predictions from our model are 100% correct.

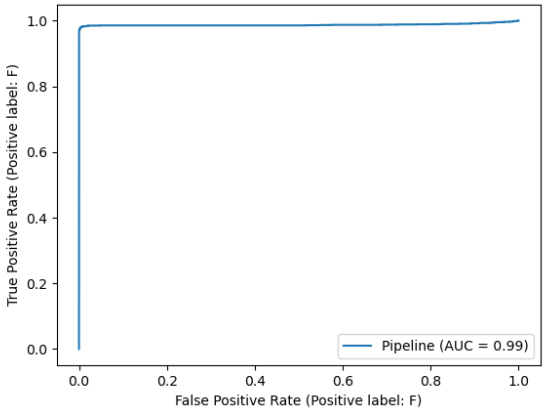

Figure 6: Figure 6. ROC curve from test set

NOTE on high f1 score -

We are aware of the fact that f1 precision and recall score of our model on the train, validation and test sets are quite high. This may be because there are underlying linear relationships between different features and the target.

Assumptions

In our data analysis, we are making the following assumptions -

- The restaurants' data set that we have is a good enough representative sample of the restaurant population in New York.
- The data is sourced ethically and is collected in a fair manner.
- There is no bias in the data that is collected.
- Grading is not affected by any external factors during the inspection.

Limitations

The EDA shows that many of the restaurants have undergone inspections more than once, with the help of the `camis` feature. But, it is unclear whether some restaurants share the same name, or if some restaurants have changed their name between 2012 and 2018. Unfortunately, we were unable to parse the data by this feature to ensure that the restaurants that are reviewed more than once are not included both in the training and validation/test sets. Since we aren't sure whether the models are in fact learning the 'features' or the specific restaurant examples, there might be some discrepancies in the prediction results. We also had to downsample our training data set in order to reduce the training time as we are in short of computational resources.

Statement of Future Directions

- Going forward, we would be considering only unique restaurants for training the model by dropping all duplicate inspections having matching `camis` features and performing predictions using the new model for other cities.
- As we progress with the research, we intend to use several models to compare different metrics and choose the best-suited model for our research problem.

- Additionally, we are also planning to make use of the deployment data and predict the grade of the restaurants using our model and compare this to the actual grading by the inspection team.
- In addition to New York, we could also generalize our model to predict the restaurant grading across other metropolitan cities.
- Lastly, we would like to incorporate certain feature engineering techniques like text-based engineering so that our model can be trained with the best features and yield better results.

References

Mock (2022) Pedregosa et al. (2011) Cortes and Vapnik (1995) Van Rossum and Drake (2009) McKinney et al. (2011) de Jonge (2018) Keleshev (2014) Xie (2014) "Anaconda Software Distribution" (2020) Software Distribution Pérez and Granger (2007) VanderPlas et al. (2018) Cox (1958) Hunter (2007)

"Anaconda Software Distribution." 2020. *Anaconda Documentation*. Anaconda Inc. https://docs.anaconda.com/.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97.

Cox, David R. 1958. "The Regression Analysis of Binary Sequences." *Journal of the Royal Statistical Society: Series B (Methodological)* 20 (2): 215–32.

de Jonge, Edwin. 2018. *Docopt: Command-Line Interface Specification Language*. https://CRAN.R-project.org/package=docopt.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Keleshev, Vladimir. 2014. *Docopt: Command-Line Interface Description Language*. https://github.com/docopt/docopt.

McKinney, Wes et al. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." *Python for High Performance and Scientific Computing* 14 (9): 1–9.

Mock, Thomas. 2022. "Tidy Tuesday: A Weekly Data Project Aimed at the r Ecosystem." https://github.com/rfordatascience/tidytuesday.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Pérez, Fernando, and Brian E. Granger. 2007. "IPython: A System for Interactive Scientific Computing." *Computing in Science and Engineering* 9 (3): 21–29. https://doi.org/10.1109/MCSE.2007.53.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057. https://doi.org/10.21105/joss.01057.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.