

A night-time photograph of the Dallas skyline, featuring several illuminated skyscrapers and the Reunion Tower with its glowing red sphere. The city lights are reflected in a body of water in the foreground. The title text is overlaid on the upper half of the image.

Predicting Food Inspection Scores in Dallas

Brandon Greenspan, Data Scientist

Problem Statement

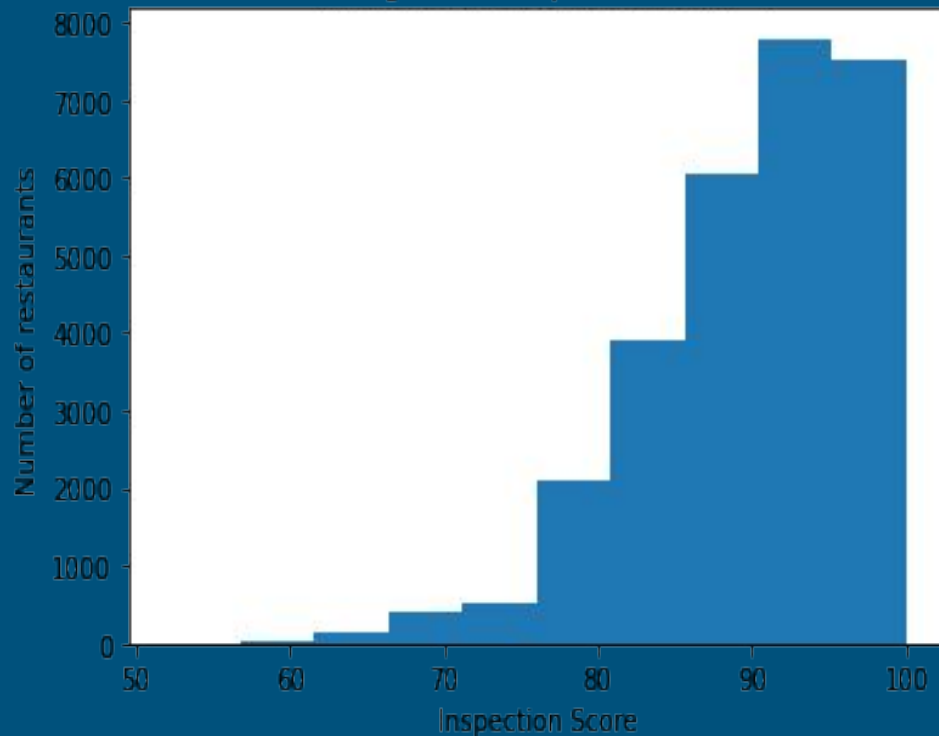
- An article in the Dallas Observer unearthed a massive problem in the city's ability to follow up on restaurants requiring re-inspection due to a low grade upon original inspection.
- Classification problem (re-inspection requirements fall into grade letters- A, B, C, D, F) utilizing accuracy as our metric with a goal to beat the baseline (64%).
- We will utilize NLP to evaluate inspector notes and grades to predict the next inspection score for a restaurant.
- Models used will be Logistic Regression, Multinomial Naive Bayes, Decision Tree with PCA, and a Neural Network.

Project Workflow

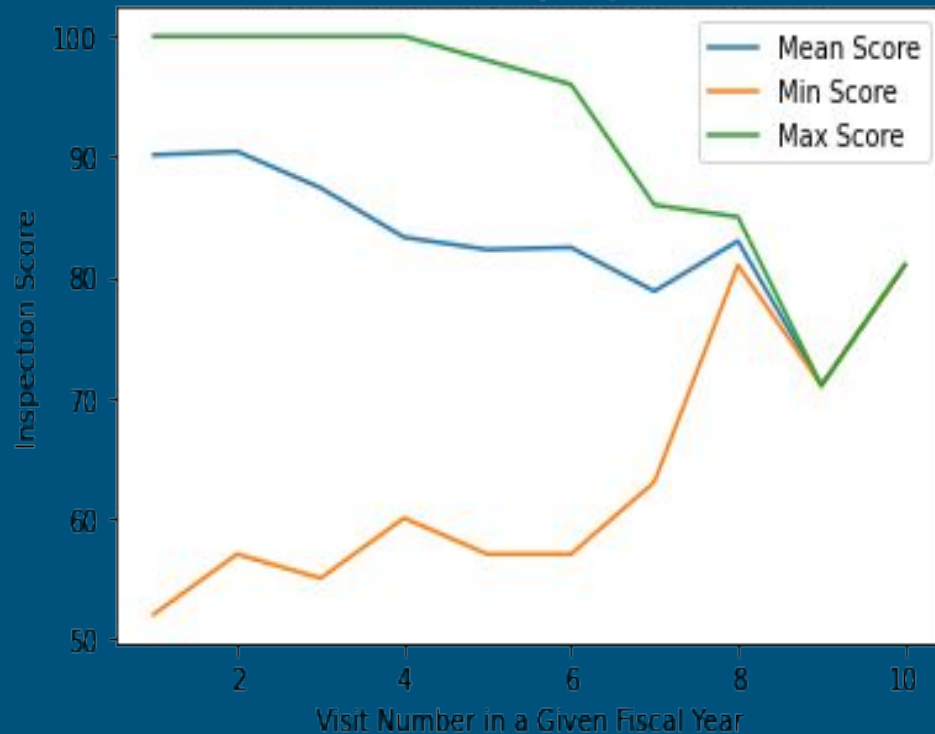
- Gathered data from <https://www.dallasopendata.com/City-Services/Restaurant-and-Food-Establishment-Inspections-Octo/dri5-wcct/data>
- Data cleaning involved imputing restaurant name nulls, calculating inspection scores based on points lost and replacing the scores with that calculation.
- Had to shift our data in order to create target column, which led to dropping some data.
- EDA looked at possible correlated features to the next inspection score.
- After running CountVectorizer, we did a train test split, stratifying our data.
- Ran Logistic Regression, Multinomial Naive Bayes, Decision Tree with PCA, and a Neural Network.
- Evaluate results on best model, looking at strong positive and negative coefficients.

EDA- Visualizing Data

Histogram of Inspection Scores



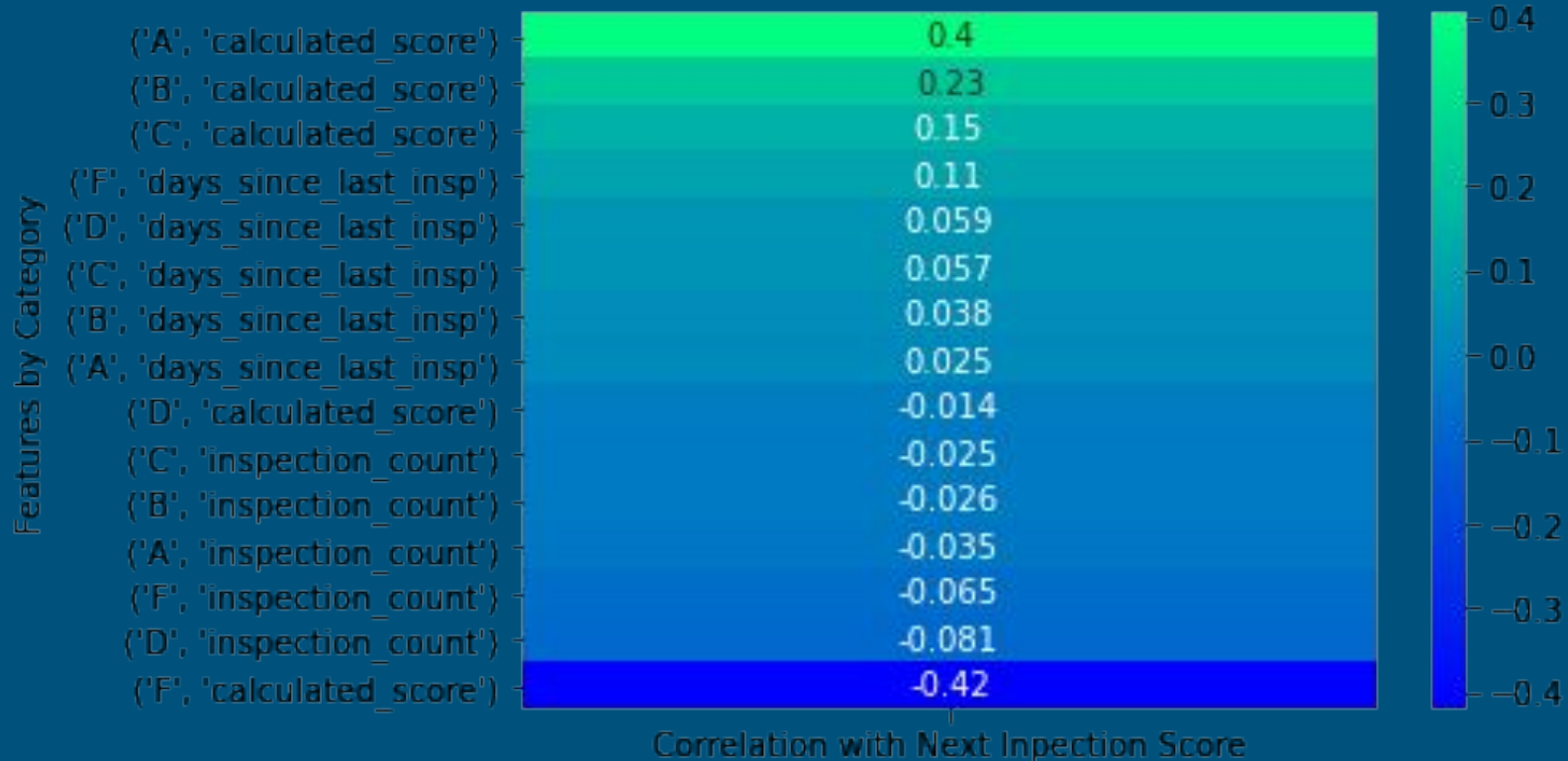
Mean & Min Score by Inspection Count



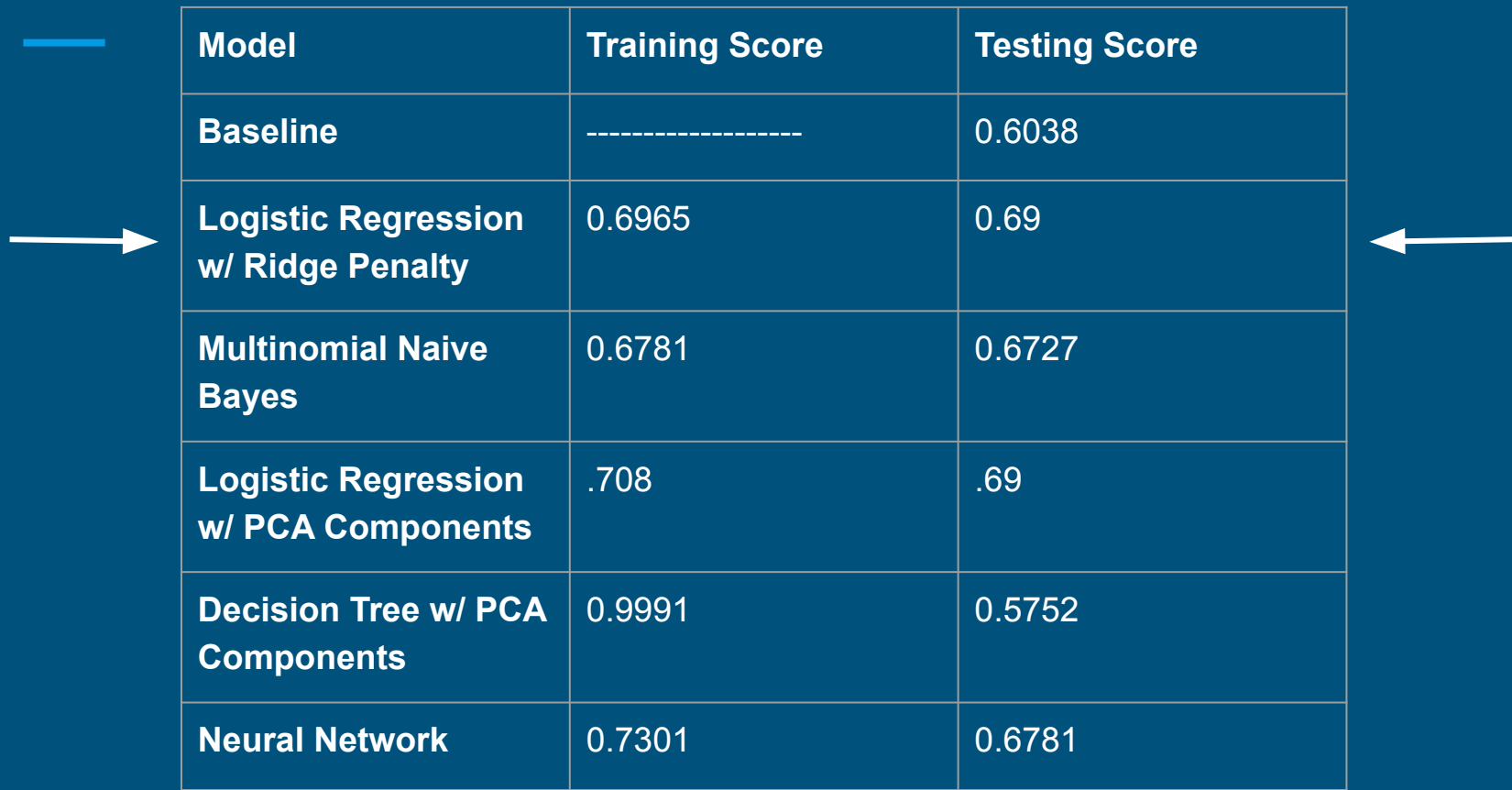
EDA- Correlation

	calculated_score	next_inspection_score	days_since_last_insp	inspection_count
next_inspection_score	0.554437	1.000000	0.105646	-0.088987
calculated_score	1.000000	0.554437	0.078654	-0.079388
days_since_last_insp	0.078654	0.105646	1.000000	-0.376266
inspection_count	-0.079388	-0.088987	-0.376266	1.000000

EDA- Correlations by Inspection Grade



Model Selection



A diagram illustrating model selection. A blue horizontal line is positioned above the table. A white arrow points from the left towards the 'Logistic Regression w/ Ridge Penalty' row. Another white arrow points from the right towards the same row. The table contains seven rows of model performance data.

Model	Training Score	Testing Score
Baseline	-----	0.6038
Logistic Regression w/ Ridge Penalty	0.6965	0.69
Multinomial Naive Bayes	0.6781	0.6727
Logistic Regression w/ PCA Components	.708	.69
Decision Tree w/ PCA Components	0.9991	0.5752
Neural Network	0.7301	0.6781

Model Evaluation - Confusion Matrix

	Actual A	Actual B	Actual C	Actual D	Actual F
Predict A	3744	1290	88	15	0
Predict B	550	1190	186	41	1
Predict C	0	4	1	0	2

Strongest Positive Coefficients By Class

1

Feature	A Coefficient	A Odds
calculated_score	3.156263	23.482674
days_since_last_insp	1.54628	4.693978
days	1.234078	3.435211
need	0.996654	2.709202
provide	0.909066	2.482003

Feature	B Coefficient	B Odds
walls	0.769392	2.158453
replace	0.750669	2.118417
criteria	0.669006	1.952297
fy2019	0.666853	1.948097
kitchen	0.618624	1.856372

Feature	C Coefficient	C Odds
food	1.770606	5.874414
shall	1.176586	3.243282
self	0.882943	2.418006
drain	0.859599	2.362214
228	0.850446	2.340691

Feature	D Coefficient	D Odds
food	1.562919	4.772732
cooling	0.723023	2.060652
raw	0.716474	2.047202
shall	0.579666	1.785443
228	0.575524	1.778062

Feature	F Coefficient	F Odds
food	0.237149	1.26763
sources	0.198695	1.21981
waste	0.163792	1.177969
time	0.162398	1.176328
water	0.139536	1.14974

Strongest Negative Coefficients By Class

Feature	A Coefficient	Not A Odds
food	-2.969037	19.473161
shall	-2.365283	10.647055
228	-1.75842	5.80326
approved	-1.594548	4.926102
raw	-1.475344	4.37254

Feature	B Coefficient	Not B Odds
need	-0.84387	2.32535
2018 fy2018	-0.757339	2.132593
wash	-0.736959	2.089572
exposed splash	-0.710789	2.035597
report	-0.680064	1.974004

Feature	C Coefficient	Not C Odds
calculated_score	-1.656572	5.241315
provide	-0.72251	2.059596
days	-0.542778	1.720781
food containers	-0.528401	1.696217
avoid	-0.457575	1.580238

Feature	D Coefficient	Not D Odds
calculated_score	-1.303184	3.680997
days_since_last_insp	-0.743219	2.102693
clean	-0.41399	1.512842
provide	-0.319318	1.376189
hair	-0.280309	1.323539

Feature	F Coefficient	Not F Odds
days_since_last_insp	-1.241805	3.461858
calculated_score	-0.297675	1.346724
hair	-0.085362	1.089111
light	-0.066791	1.069072
restraints	-0.055349	1.056909

Conclusion

- Our Logistic Regression model with a Ridge penalty and usage of TFIDF was able to predict the next inspection grade of a restaurant with 69% accuracy, which did exceed our goal of defeating the baseline of 60.38% accuracy.
- However, unbalanced classes led to ZERO predictions below a C, which makes it very difficult to predict possible food & safety hazards. Out of the 7 C predictions, 2 were actually F's. The rare C prediction should signal that a restaurant requires further scrutiny.
- In order to cut cost and lighten inspector workload, we should look to decrease the cadence of inspections for places that consistently score an A.

Next Steps

- Since this was NLP based, we should look to accommodate all 800,000 features. This requires a lot more time and computing power, likely requiring the use of cloud computing.
- With added computing power, we may want to use bootstrapping to see if we can improve predictions for low inspection grades.
- Further time series analysis. Maybe the model can predict the next score of a restaurant more accurately if it takes into account every previous inspection, not just the most recent.

References

[Restaurant and Food Establishment Inspections \(October 2016 to Present\) from Dallas Open Data](#)

[Failing Grade: Dallas Policies Protect the City's Filthiest Restaurants from Health Inspectors by Brian Reinhart from Dallas Observer](#)