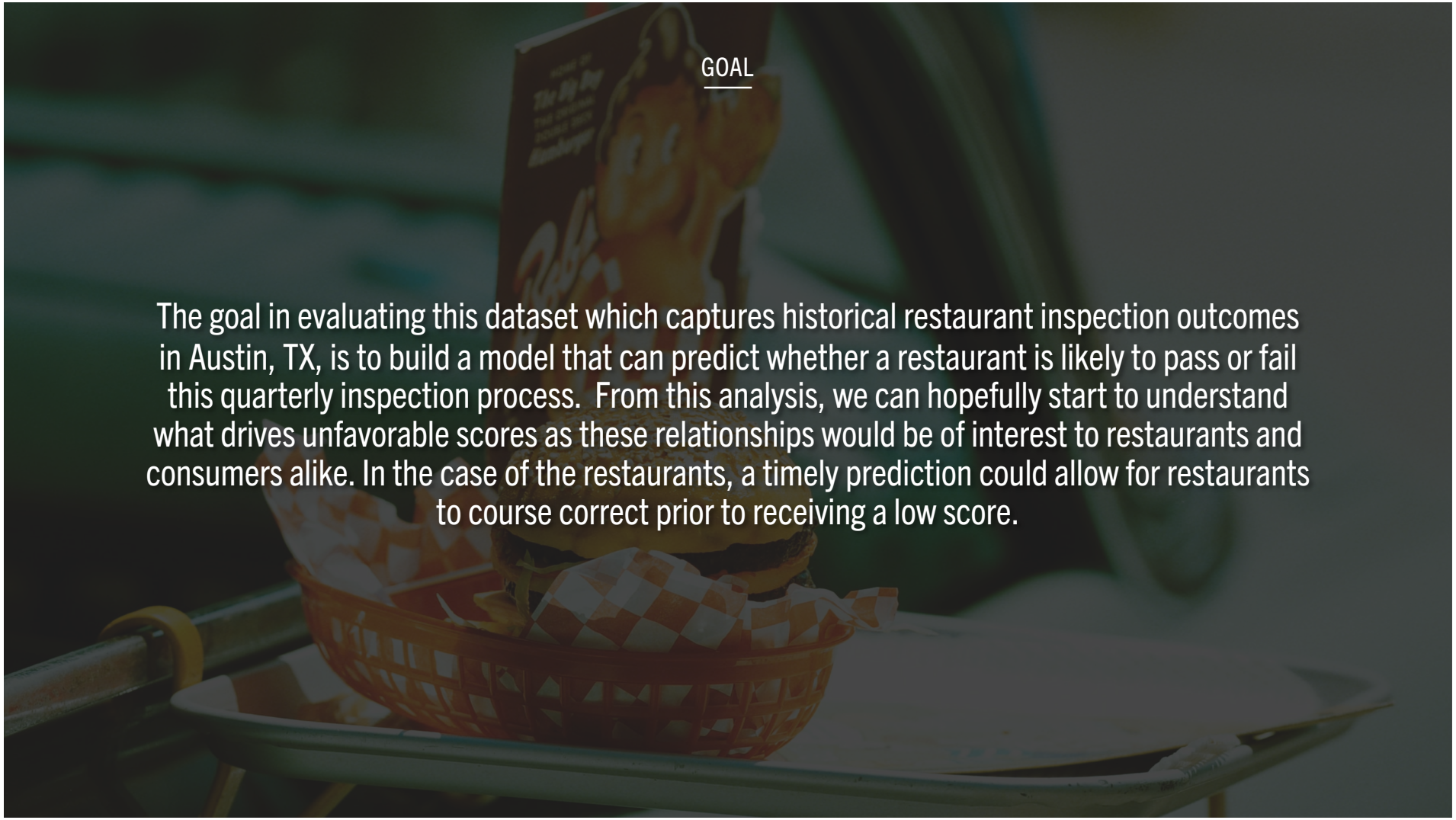
PREDICTING RESTAURANT INSPECTION OUTCOMES

CAPSTONE 2 - PRESENTATION



GOAL

The goal in evaluating this dataset which captures historical restaurant inspection outcomes in Austin, TX, is to build a model that can predict whether a restaurant is likely to pass or fail this quarterly inspection process. From this analysis, we can hopefully start to understand what drives unfavorable scores as these relationships would be of interest to restaurants and consumers alike. In the case of the restaurants, a timely prediction could allow for restaurants to course correct prior to receiving a low score.

MOTIVATION AND CHALLENGES

MOTIVATION:

- GREAT OPPORTUNITY TO DO SOME ANOMALY DETECTION
- THE DATA IS HIGHLY CLASS IMBALANCED

CHALLENGES:

- THE DATA IS HIGHLY CLASS IMBALANCED ☺
  (WITH ONLY 0.9% OF INSPECTIONS RESULTING IN A FAILING SCORE)
- MANY RESTAURANTS HAVE FEW OR NO REVIEWS
- RELATIVELY SMALL NUMBER OF INSPECTIONS

THE DATA

SOURCES:

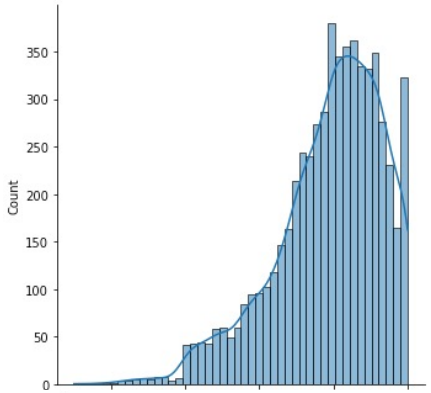
- AUSTIN RESTAURANT INSPECTION SCORES | AUSTIN PUBLIC HEALTH
- RESTAURANT REVIEWS | YELP

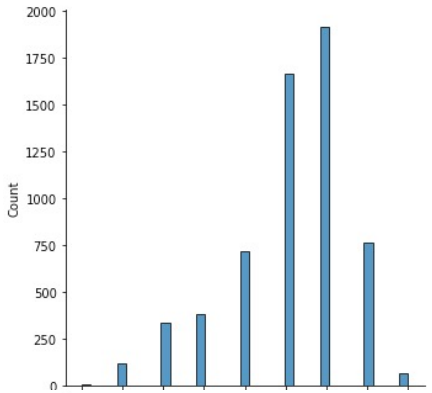
HIGH-LEVEL STATS:

- 3 YEARS OF INSPECTIONS HISTORY
- INSPECTIONS OCCUR QUARTERLY
- 5,825 TOTAL INSPECTIONS
- ALLOWING A 60-DAY LOOKBACK FOR REVIEWS-BASED FEATURES
- ASSSUMES THAT NON REVIEWS-BASED FEATURES ARE STATIC OVER THE 3 YEARS

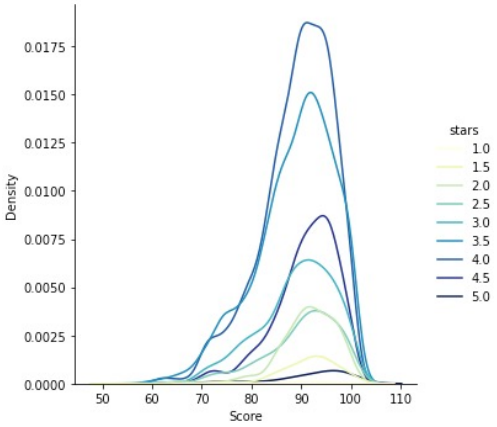
HOW IS THE DATA DISTRIBUTED?

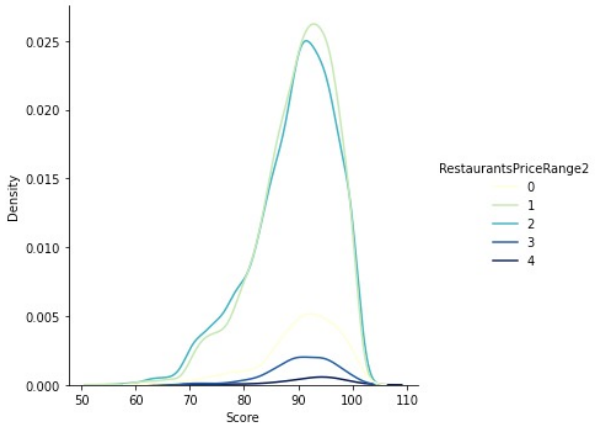
- INSPECTION SCORES AND RATINGS ALIKE ARE LEFT-SKEWED AND TEND TO BE HIGH





- INTERESINGLY AND ENCOURAGINGLY INSPECTIONS SCORES FOLLOW A SIMILAR DISTRIBUTION ACROSS RATINGS AND PRICE RANGE





THE SCORING IS HIGHLY IMBALANCED

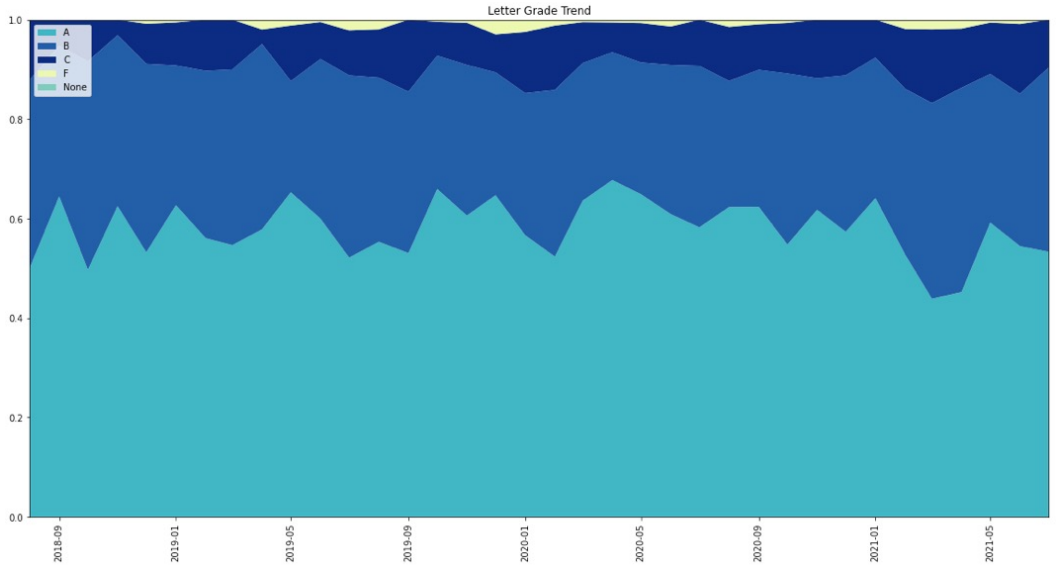
A: 58%

B: 32%

C: 10%

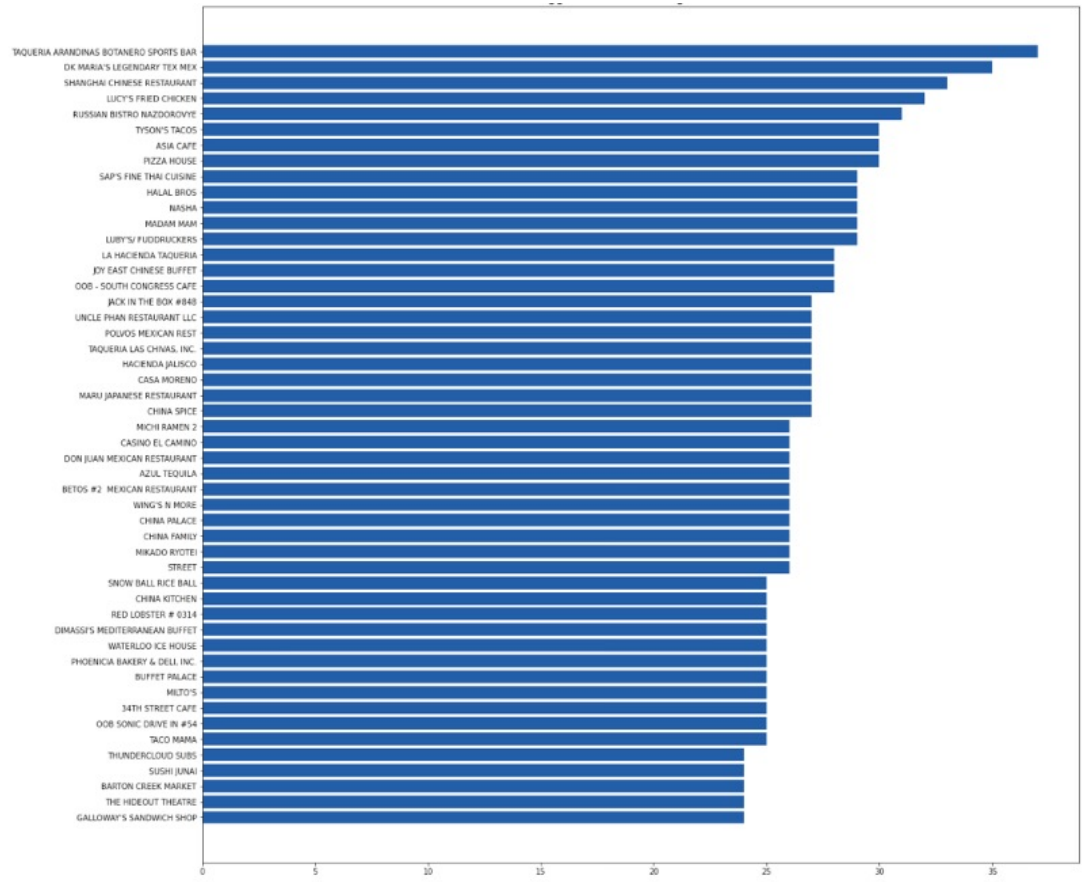
F*: 0.9%

* 70 OR BELOW



THE SCORING IMBALANCE APPEARS TO BE RELATIVELY STATIONARY OVER TIME

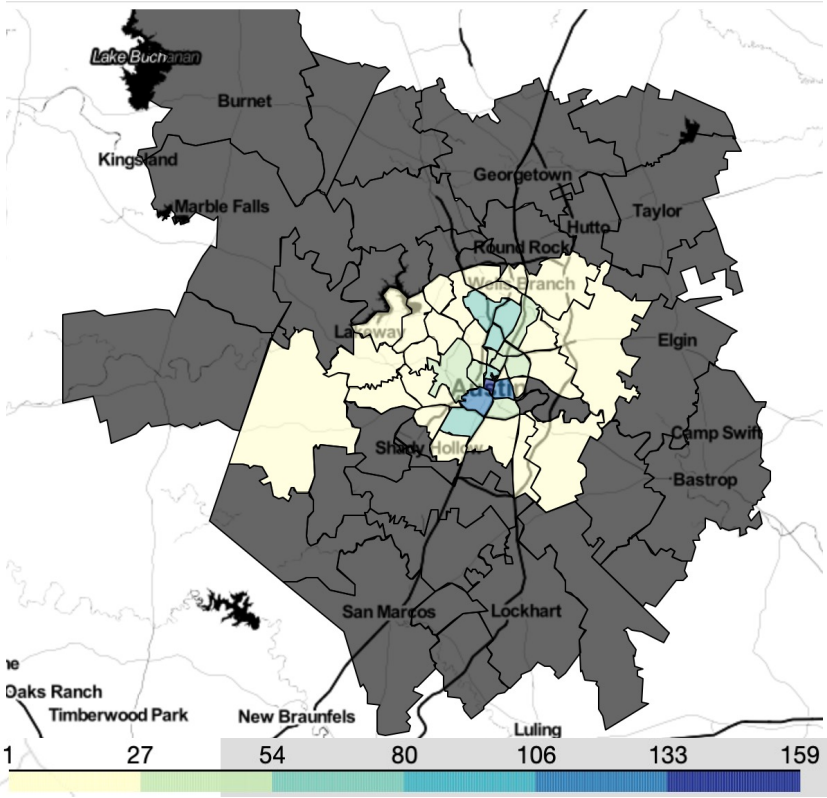
BUT AT THE RESTAURANT LEVEL WE DO SEE VARIATION OVER TIME



- IT'S A MIXED BAG IN TERMS OF RESTAURANTS SCORE OVER TIME

- THE LARGEST VARIATION (IN ABSOLUTE TERMS) OVER THE 3 YR PERIOD WAS 35 POINTS

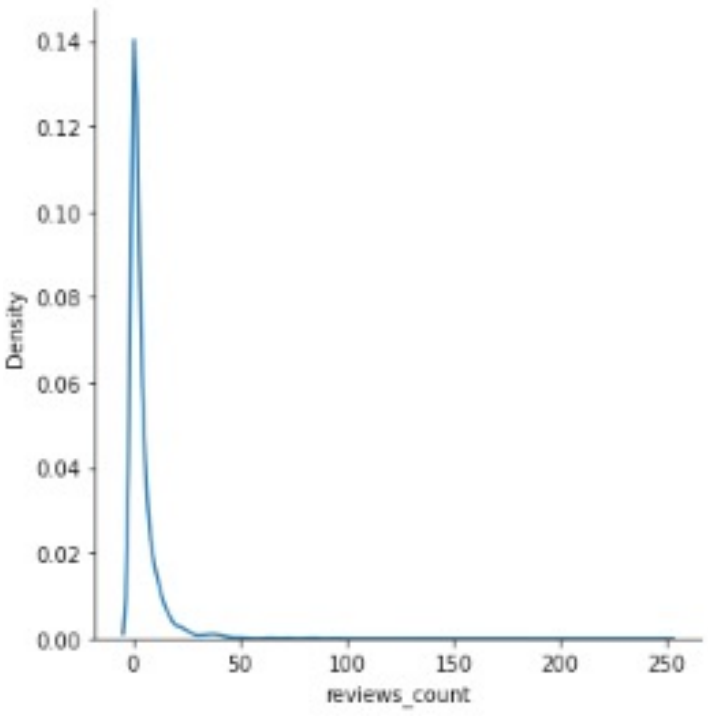
ANY INTERESTING SPATIAL RELATIONSHIPS?

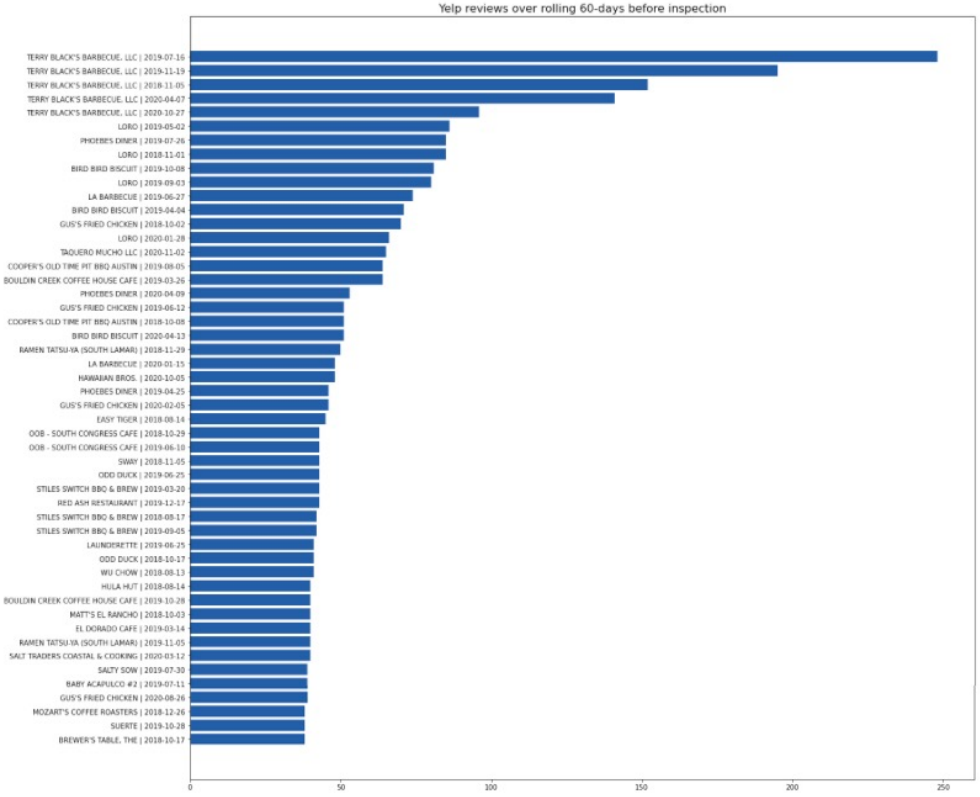


VISUALLY THERE DOESN'T APPEAR TO BE ANY MISMATCH BETWEEN CITY/RESTAURANT DENSITY AND INSPECTION DENSITY THAT WOULD SUGGEST A SPATIAL BIAS IN THE DATA

WHAT CAN WE LEARN FROM THE YELP DATASET??

HOW DO REVIEWS FIGURE IN?





MOST RESTAURANTS DO NOT HAVE VERY MANY IF ANY REVIEWS DURING THE 60-DAY LOOKBACK WINDOW

THE COVERAGE IS NOT GREAT, BUT IT IS A RICH DATASET SO LET'S DIG A BIT DEEPER

SENTIMENT ANALYSIS

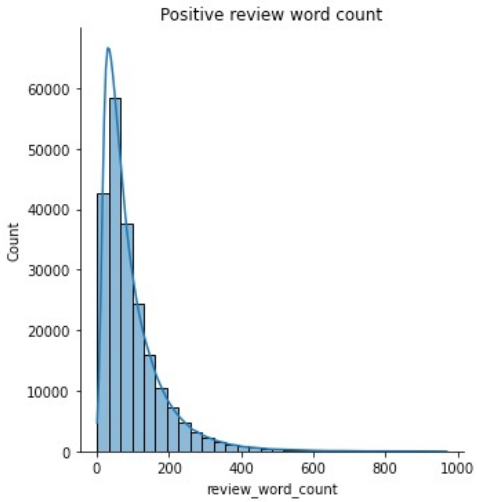

```
positive['text'][1:2].values
```

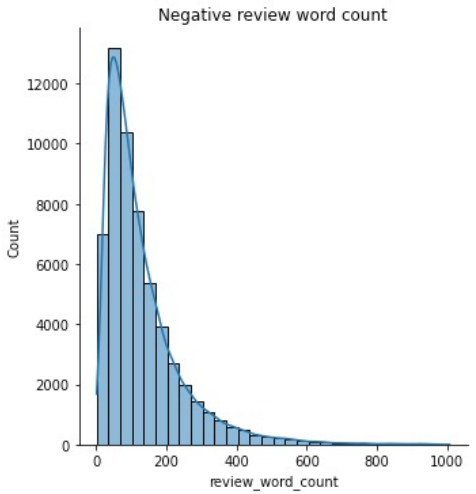
```
array(['Definitely one of my favorite places for vietnamese grub! Very authentic and although they
h, the people here are really nice. Will always recommend others to try it.'],
```

```
negative['text'][3:4].values
```

```
array(["I have been to this place a couple of times and each time, I never ceased to be disappointed.
par at best and the place is dirty.  One time a friend of mine went there and found a lizard (yes, that
ard) in his vermicelli bowl.  The service staff is not very helpful either.  They are hard to understan
em to be in a bad mood.  So if you enjoy paying too much for a can (yes, CAN) of soda with a cup of ice
ards in your food, I would recommend Pho Van."],
```

- THIS ANALYSIS PRODUCED SOME INTEREST LEARNINGS BUT IT WOULD NEED TO BE REVISITED A BIT MORE DEEPLY FOR ADDITIONAL FEATURE EXTRACTION AS WELL AS TO GAUGE THE POTENTIAL FOR INFORMATION GAIN BEYOND WHAT'S PROVIDED BY YELP STARS
- REVIEW_LENGTH AND REVIEW_WORD_COUNT ARE A COUPLE OF FEATURES WHICH WOULD SEEM TO BE INDEPENDENT OF STARS





FEATURES

NUMERIC

- REVIEW LENGTH
- REVIEW WORD COUNT
- # OF REVIEWS
- TOTAL STARS
- # OF USEFUL REVIEWS
- # OF 5 STAR REVIEWS
- # OF COOL REVIEWS
- # OF FUNNY REVIEWS
- # OF 3 STAR REVIEWS

CATEGORICAL/DUMMY

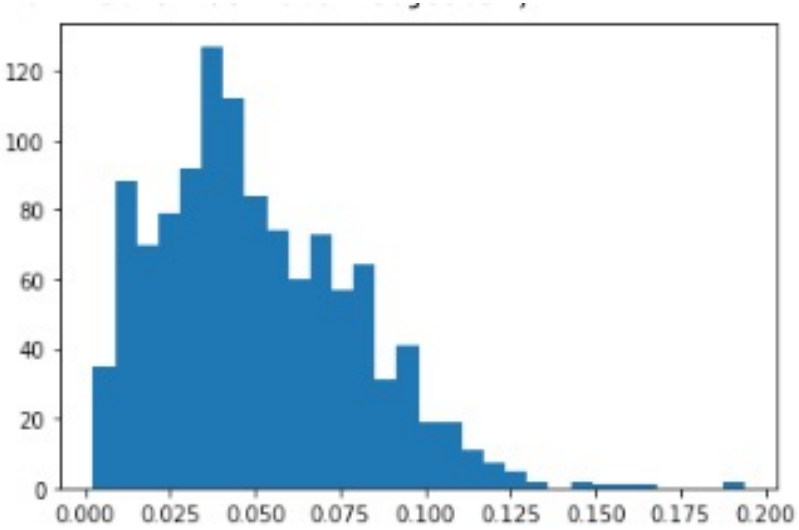
- BUSINESS ACCEPTS CREDIT CARDS
- ALLOWS SMOKING
- HAS TAKEOUT SERVICE
- IS GOOD FOR GROUPS
- IS GOOD FOR KIDS
- HAS BIKE PARKING
- HAS A PARKING LOT
- HAS TV
- HAS DELIVERY SERVICE
- IS CASUAL
- HAS AVG NOISE LEVEL
- HAS OUTDOOR SEATING
- IS OPEN FOR LUNCH
- HAS CATERING SERVICE
- IS OPEN FOR DINNER
- HAS WIFI
- IS WHEELCHAIR ACCESSIBLE
- IS CLASSY
- HAS TABLE SERVICE
- TAKES RESERVATIONS
- IS A BAR
- IS A NIGHTLIFE SPOT
- ALLOWS DOGS

BUILDING THE MODEL

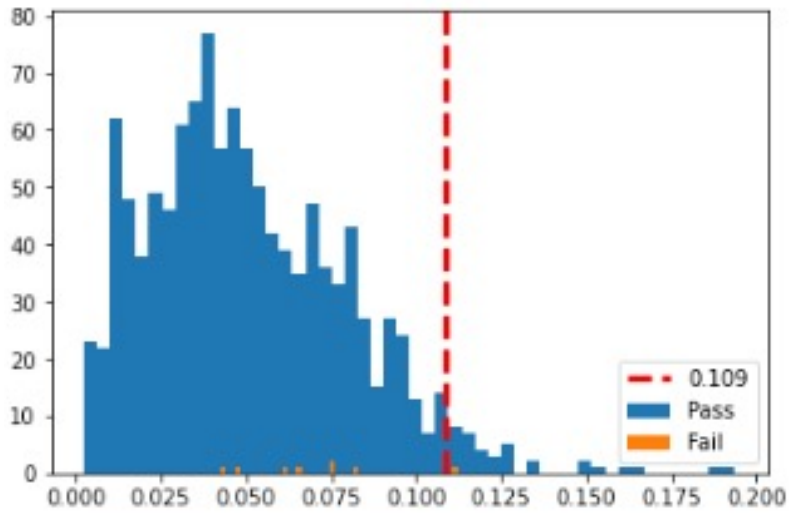
AUTOENCODER FINDINGS

DISTRIBUTION OF RECONSTRUTION ERROR

PASSED INSPECTIONS

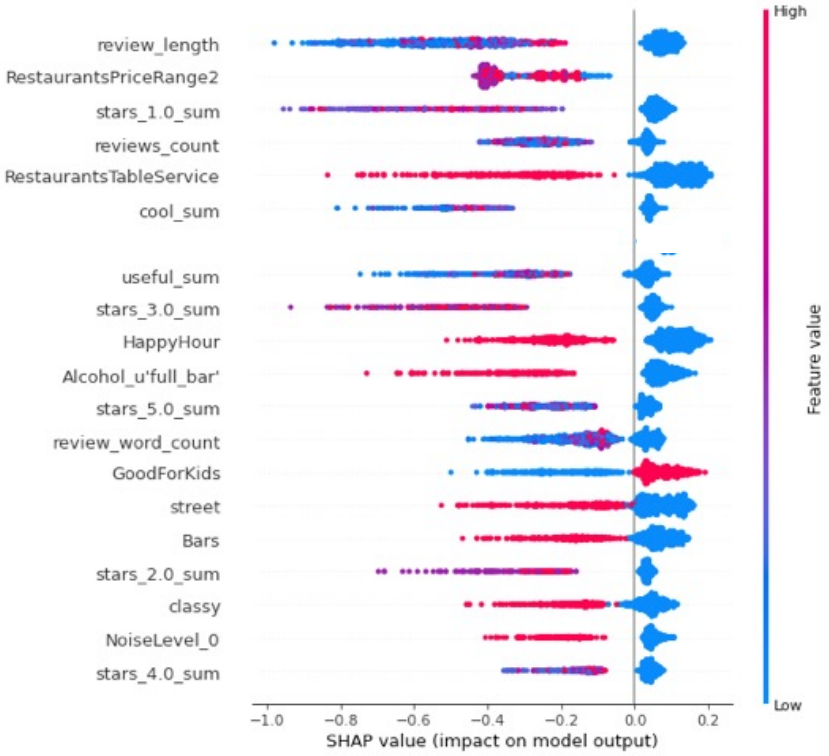


PASSED & FAILED INSPECTIONS



- THE AUTOENCODER WAS TRAINED ON OBSERVATIONS OF PASSED INSPECTIONS AND USED TO PREDICT FAILED INSPECTIONS
- THE GOAL HERE WOULD BE TO SEE A SHIFT IN THE MEAN REDCONSTRUCTION ERROR FOR FAILED INSPECTIONS, BUT THAT SHIFT DIDN'T CLEARLY SHOW UP HERE
- MODEL PARAMTERS: DECODER WITH 2 HIDDEN LAYERS USING A RELU ACTIVATION FUNCTION; ENCODER 2 HIDDEN LAYERS AND AN OUTPUT LAYER WITH SIGMOID ACTIVATION FUNCTION, RAN FOR 40 EPOCHS (W/ EARLY STOP ON MEAN SQUARED ERROR LOSS FUNCTION, IT RAN FOR 32), BATCH SIZE = 20

ISOLATION FOREST FINDINGS



- UNLIKE THE AUTOENCODER, THIS MODEL WAS TRAINED ON DATA WITHOUT ANY SCALING SINCE IT'S ROBUST TO VARIANCE/NOT LEVERAGING DISTANCE-BASED CALCULATIONS
- THE MODEL CORRECTLY PREDICTED 87.5% OF THE FAILED INSPECTIONS
- MODEL PARAMETERS: MAX SAMPLES: 100, N_ESTIMATORS = 20

TAKE-AWAYS

- CLASS IMBALANCE IS A COOL, COMMON AND CHALLENGING PROBLEM
- PERHAPS, EXPLORE MORE HISTORY/TIME-SERIES BASED FEATURES WHICH INHERENTLY HAVE COMPLETE COVERAGE FOR THE DATASET
- IF AVAILABLE CONSIDER PULLING INTO STATE-WIDE DATA TO INCREASE THE SIZE OF THE DATASET