NetApp Data Science Project

Ivan Hom

Question 2a: What were the average SCOREs of inspections for each of the FACILITYPEs “Restaurant” and “Food Stand”?

Restaurant mean score = 95.73

Foodstand mean score = 96.87

Steps:

-used R, connected to db to read restaurants_table and inspections_table

-Join on HSISID

-Selected rows that have FACILITYTYPE == “Restaurant” and “Food Stand”

-Computed summary of SCORE column

Question 2b: Does SCORE vary depending on INSPECTOR performing the inspection?

Yes, the mean score varies for each inspector.

Steps:

Performed one-way ANOVA test

H0: the mean scores for each inspector are the same

H1: the mean scores for each inspector are different

P-value = 2.2e-16, which is far less than .05 significance level

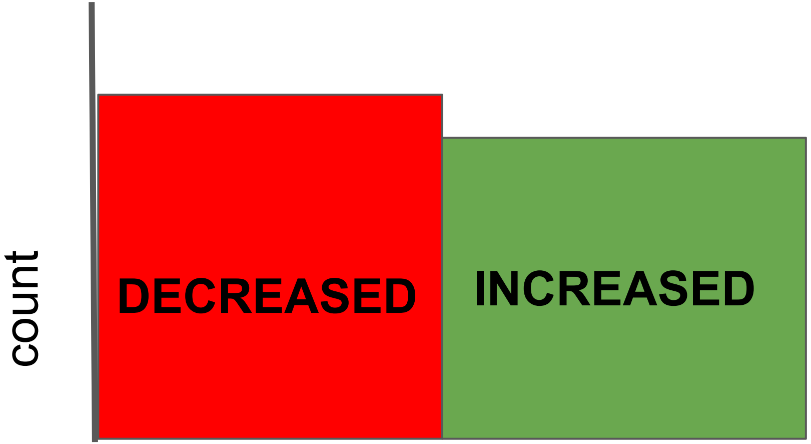
⇒ reject H0

Question 3: Relationship between an HSISID's inspection SCORE and its most recent prior inspection SCORE

No clear relationship between an inspection SCORE and most recent prior

Steps:

-spot checking of HSISID's SCORE and most recent score



Question 3: Prediction model for P(SCORE<93)

Create classification model with accuracy 90%

Steps:

-split data into 75% training and 25% test

-transform SCORE into binary categories: below_93, equal_above_93

-CART model is based on a decision tree

-loss function is Gini index (how pure the leaf nodes after the split)

Question 4: Enhancements with more time

-Better understanding of the business implications of False Positives/False Negatives so the model can be tuned

-Perform stratified sampling to overcome the severe 10%-90% class imbalance in the data

-Use Naive-Bayes model to incorporate the free text in the classifcation, similar to an spam email detection system

-Explore using a random forest or deep learning