ASNA 2020 Case Competition
Round 0 Report

Team #1

Oct 1, 2019

Process Outline

Using R, these steps were taken in the data cleaning process:

1. Grouped all data sheets into one and added a column specifying the worksheet year we pulled it from
2. We deleted all completely duplicate data entries
3. We found a correlation between the 4^{th} to 11^{th} characters of the bridge ID and the year built
    a. In cases with 2475 of the same Bridge ID in the same worksheet year, we deleted all the entries where the bridge ID and year built did not match up
    b. In the case with 237 same Bridge ID entries in the same worksheet year, 4 had entries where bridge ID and year built matched. We then used the mode for each specific column and left the one Bridge ID entry that fulfilled this the best.
4. Constructed a conditional table with bridge ID vs. true or false for logical assumptions made where the last column counts the number of false entries
    a. 5 False -> 565 entries
    b. 4 False -> 23609 entries
    c. 3 False -> 24333 entries
    d. 2 False -> 10414 entries
    e. 1 False -> 1575 entries
5. We deleted all the data with 5 and 4 false entries
6. Deleted data points when inspection date is NA and estimated cost is less than 0
7. Inspection date = "99/99" was deleted
8. Added bridge lifetime (worksheet year – year built)
    a. Deleted negative lifetimes
9. Changed the categories that came up as false to NA
10. Used cart (classification and regression tree) from the library mice in R to input replacement values into the NA entries
    a. Error with year built = 1; deleted all these values
11. For the duplicate Bridge IDs: mode was used on the road type, year built, number of lanes, operational status, structure material, deck material, bridge design, number of spans, road width and deck width to find the one to keep
12. Manually inputted structural improvement needed values, given that the assessment value
    a. Assessment = 1-3 => Structural improvement needed = 0
    b. Assessment = 4 => Structural improvement needed = 1
    c. Assessment 5 => Structural improvement needed = 2
    d. Assessment 6 => Structural improvement needed = 3
    e. Assessment 7 => Structural improvement needed = 3
13. Input random number into N/A cases for Average Daily Traffic with a range of the existing data
14. Deleted columns: length of improvement, estimated improvement cost and year of estimated improvement cost (previously did not clean these columns)
    a. We assumed these would not affect the need for structural repair

15. Deleted Bridge Design, Number of Spans and Structural Material after using an ANOVA test

Using Python to develop the model:

16. Split the data into training and validating data with train_test_split() from sklearn.model_slection
17. Used logistic regression to fit the model using training data so the model can predict structural improvement is needed based off other variables
18. Use this model to predict the structural improvement needed in the validation data set
19. Compare the values from model to the actual values and a score on the accuracy (from 0 to 1)
20. We set a number of combinations for values of the variables by determining how we divide the range within the categories of road type, number of lanes, average daily traffic, operation status, bridge length, assessment and lifetime. These decreased the number of values possible for each category
    a. Road type: 5 possible values (1-5)
    b. Number of lanes: four possible values (2,6,10,14 lanes)
    c. Average Daily traffic: 4 possible values (100K, 300K, 500K, 700K)
    d. Operational status: 5 possible values (out of 5)
    e. Bridge length: 5 possible values (1000, 2000, 3000, 4000,5000)
    f. Assessment: 4 possible values (1, 3, 5, 7)
    g. Lifetime: 8 possible values (10, 25, 40, 55, 70, 85, 100, 115)
21. With these possible combinations, a probability was given for each possible combination by the model

Logical assumptions made

1. Assessment cannot improve without either structural improvement or reconstruction, ignored if the last improvement date is N/A
    a. Rationale: a bridge cannot automatically improve in condition without structural improvement or reconstruction
2. Road type and year built should not change even with structural improvement or reconstructions
    a. Rationale: the road type does not depend of the bridge, but instead the location
    b. Rationale: a bridge can only be built once, therefore the year built cannot change
3. Operation status is new (equals 5) if the year built equals the year of the sheet it came from, unless there was structural improvement or reconstruction
    a. Rationale: a bridge can only be new once
4. Given that the assessment is 1-3, the structural improvement needed will be 0
    a. Rationale: if the bridge's condition is from excellent to fair, no structural improvement should be needed
5. Date of inspection must be greater than the year built
    a. Rationale: the bridge cannot be inspected before it is built
6. Improvement cost must be greater than equal 0

    a. Rationale: a cost cannot be negative and improvement cannot cost nothing
7. Year of reconstruction and inspection cannot be less than year built
    a. Rationale: there cannot be reconstruction or inspection before the bridge is built
8. Year of reconstruction and inspection cannot be greater than year of the sheet (tab year)
    a. Rationale: assuming the year of the sheet is the database created in that given year, it is not possible to have future information on reconstruction and inspection
9. The month of the inspection date must be between 1 and 12
    a. Rationale: 12 months in a year
10. Year built must be less than or equal to the tab year
    a. Rationale: lifetime of a bridge cannot be negative
11. Bridge length, road width and deck width must be greater than 0
    a. If a bridge length, road width and deck width is 0; the bridge does not exist

Conclusion

A shortcoming of data was that since the entire database was hacked, it was difficult to choose to trust any of the values given because we were unsure about the extent that the data was corrupted or deleted. This made it very difficult to clean data because it was impossible to entirely clean the hacked data. Therefore, the cart function in mice from R did input values that did not necessarily follow our logical assumptions. We were also unsure about the correlation between bridge design, deck width, road width and other factors with the structural improvement needed, which made it a challenge to create more assumptions without compromising the integrity of the data. Due to this, we did not consider these values when finding the Since we created a truth table based off the logical assumptions we made and deleted the data with 4 or 5 false flags, this decreased our sample size significantly and the variance of the estimation increased. The categories: length of improvement, estimated improvement cost and year of estimated improvement cost, also had too much data missing which made them difficult to incorporate into our results. This made it very complex to take those categories into consideration, so we chose not to use them in the final model. Average daily traffic was unobservable so for N/A values, we had to use a random number generator within the possible average daily traffic range.

For the Python component, our probabilities only reflected a limited number of combinations (64,000 in total) which we selected in order to create a more easily usable Excel document.

In our Excel file, structural improvement needed is the last column and is based off the categorical variables we chose.