Bridge Improvement Cost

DSO530 Final Project - Spring 2017

Team Member: Cheng Chen, Xi Jiang, Yu Zhang, Yuting Zheng, Hangyu Zhou

1. Problem of Interest

Bridge plays a significant role in infrastructural system. Every year, governments spend money not only on bridge improvement itself, but also on labors to conduct on-site investigation and improvement cost estimation. The traditional process to determine how much to invest in bridge improvement involves 4 stages: project planning, project proposal, preliminary feasibility study, and detailed feasibility study, with allowable deviation >=30%, <=30%, <=20% and <=10% respectively. Our goal is associated with optimizing first two stages with tolerable variability similar to the allowable deviation here. If we could build an algorithm to automate the cost estimation procedure, we can save time and money for governments and realize better budget allocation.

2. Data Description

2.1 Data Description

Dataset Name: data_NBI
Download URL: https://www.kaggle.com/broach/build-bridges-not-walls
Data Source: United States Federal Highway Administration (FHWA)
Description: This dataset, known as the National Bridge Inventory (NBI), is about the nation's bridges that are located on public roads, including both interstate and US highways, state, country roads, and publicly accessible bridge on federal land.
Data Size: 607,070 observations x 135 variables
Variables: location, owner, bypass length, year built, lanes, average daily traffic, bridge improvement cost, roadway improvement cost, total project cost, etc.

2.2 Data Preparation

• Variable Selection

Unquantifiable variables such as route number, direction suffix, existence of speed limit post and variables with many missing values are omitted. Also, some quantitative variables are recorded as 999 or 9999 if exceeding certain values. These predictors are also omitted for the analysis. Other selections are based on expertise knowledge.

• Variable Creation and Adjustment

A new variable - Age of Bridge - is created by subtracting "Year Built" from "Year of Improvement Cost Estimate". By intuition, the age of a bridge should have influence on the amount of improvement cost.

Furthermore, to make total bridge improvement costs estimated in different years comparable, we take time value into consideration and adjusted "Total Project Cost" by Producer Price Index (PPI). This approach mitigates the effect of inflation on estimated bridge improvement costs.

• Scope Reduction
Observations with "Total Project Cost" greater than $40,000 are kept for further investigation. "Year of Improvement Cost Estimate" should be no more than 8 years old from now suggested by the official guide of the data set. Thus, disqualified observations are omitted.
• Training and Test data set
70% of data are randomly selected to fit the model and remaining 30% of data are used to assess our model performance.

3. Statistical Learning Method

Linear Regression with Lasso and Ridge
Linear, Lasso and Ridge regression model are applied on the data set.
First, we do the linear regression on the all predictors. The resulting test MSE is around 194 million. Since there are too many predictors in the dataset, we then perform feature selection by two shrinkage methods, Lasso and Ridge regression for the variable selection. However, only a minor decrease in test MSE is resulted from these two shrinkage methods. Apparently, the reduction of variance is not significant by just shrinking linear coefficient estimates. We attribute this result to the nonlinearity examined from the residual diagnostic plots as shown in Figure 1. Since the plot suggests that the true relationship is far from linear, we decide to move on with nonlinear models in the following section.

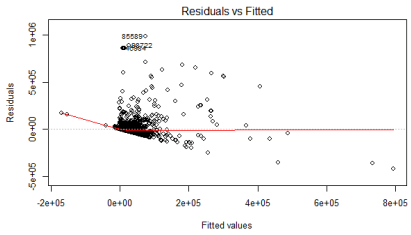

Figure 1: Residual plot from linear model

Non-Linear Model
Decision Trees Regression
Since the linear regression equation is not ideal, decision trees regression is applied to our data, in order to recognize the relationships between variables. With a post-prune method, we first grow a tree, then trim the nodes of the decision tree in a bottom-up fashion. For the original decision trees, we have 44 terminal nodes and test MSE is around 240 million. Although the MSE is large, we can see that our decision trees model chose most of the reasonable variables by looking at the variable importance (Figure 2). The most important variable is structure length followed by length of structure improvement and main structure type of design.
Using cross validation we are able to prune the original tree with only 6 terminals.The MSE is getting better but not good enough and the model is too simple.
Decision trees is a simple and interpretable model. With this model, the importance of variables can be explored. However, tree based model is not very effective at prediction and includes high-order interactions. Also, since each split is conditional on all of its previous splits, this model generates large variance. Therefore, other non-linear regression models will be considered and discussed in next part.

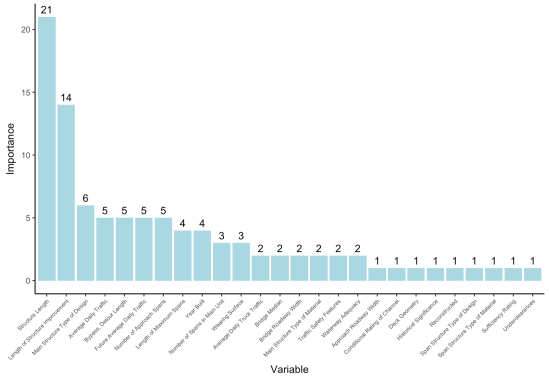

Figure 2: Variables Importance

Other Non-linear Regression Models

Several regression models are applied in order to compare if the test MSE could be significantly decreased by non-linear models. Results show that Neural Network performs better than Gradient Boosting and Random Forest. However, there are two problems associated with those regression methods. First, the model might end up predicting some negative estimated improvement cost which makes the result hard to be interpreted. Second, deviation between predicted values and true values requires metrics to evaluate the performance of each model.

Generally there are two solutions for those problems. Firstly, algebra transformation can force the predicted value to be positive. However, mean square error tends to get even worse by doing this. The second option is to use classification instead of regression - by dividing the whole range of the improvement costs into several continuous intervals and enables us to build a multi-class classification model instead of the traditionally regression model. There is a trade-off in this approach. The classification model in this case suffers from extra prediction inaccuracy in this model. However, there will be no more negative predicted improvement cost and we can directly measure the performance of our model based on the multi-dimension confusion matrix. What's more, the project focuses on the project planning stage and project proposal stage, which means the target is not constrained by predicting exact accuracy and around 30% deviation from the true value is totally acceptable.

Therefore, we move on with classification rather than regression. The classification creates 30 geometric continuous intervals that cover the range of improvement cost. Each interval represents one class and the upper limit is 1.4 times the lower limit. For evaluating the performance of the model, we need to consider not only the exact match but also one-class deviation from the true class, or even two-class deviation as well. In terms of the inaccuracy, there will be an 8.3% error on average generated by the exact-match situation if it is assumed that all the records in this interval follow a continuous uniform distribution. For one-class deviation situation, this number would be approximately 27%. And for two-class deviation situation, this number would be approximately 52%.

Neural Network Classification

Neural Network is applied for the this multi-class classification problem. The training structure uses three hidden layers and the result is shown in the Figure 3.

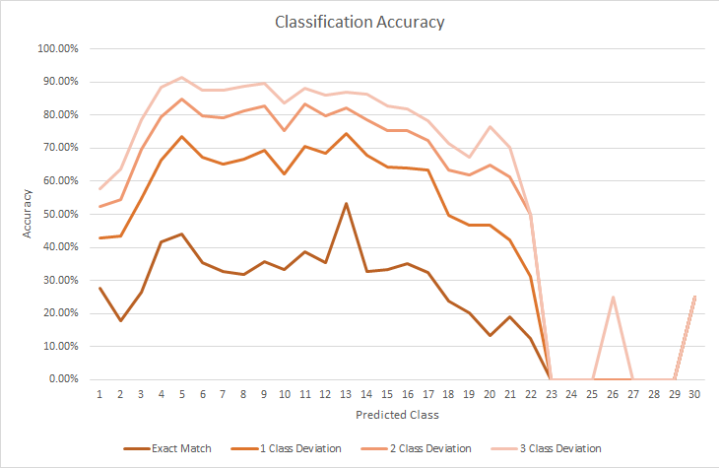

Figure 3: Classification accuracy

Figure 3 shows that the neural network perform well from class 3 to class 17. Exact match accounts for approximately 35% of accuracy in these classes while one-class deviation has an accuracy of around 70%. Nearly 95% of records comes from these classes. Generally, it means that for each record, if the predicted improvement cost lies between class 3 and class 17, one might have confidence that this prediction is quite convincing.

4. Result Summary

Using classification instead of regression makes the results interpretable. The model could predict acceptable improvement costs for around 70% of cases among those classes that contain the majority of data. And it is almost certain that given better variables, the model could generate better results which satisfy the accuracy for Project Planning Stage, Project Proposal Stage, and even Preliminary Feasibility Study Stage as well.

5. Insight + Limitation

Insights:
From the result, we can clearly see that there is a mapping between independent variables and dependent variable. Because the model does not overfit on the training set, we can believe that by adding more variables might make bigger improvement in accuracy than by adding more data. So, finding more expert variables or building more interactions is beneficial.

The model generally has already got decent accuracy for the Project Planning Stage and Project Proposal Stage. It can not only save cost for expertise, but also realize instant estimation. We are confident that given more expert variables, the model will improve a lot and even work for Preliminary Feasibility Study Stage, which could further reduce the consulting fees paid for engineers' expertise.

Limitations:
1. In the original dataset, some important metrics have too many missing values.
2. In the dataset, the reliability of the dependent variable is poor. The variable itself is an estimated number. And the definition of the improvement is not clear. We have no idea whether it is a long-time or short-term, overall or local improvement.
3. This is a national dataset, and therefore the records come from different states, agencies and companies, which might generate discrepancy of recording the data.
4. More expert variables are needed, for example the concrete and steel price at the time the estimation proposed, the size of the bridge beam and so on.

Acknowledgements:
We gratefully thank Ying Li and Yinghui Dong for their academic support, and professor Xin Tong for his great comments.