Introduction

1. Exploratory Data Analysis
   * Variables Selection (region vs state, subtrim vs fuel, X)
     1. region vs state: region is more informative than states as individuals (economy, weather conditions, etc.)
        1. While a state might be physically close to one another, their conditions might vary
        2. Overfitting problem; insufficient data point for states
2. Linear Regression
   * Interaction (*Plots*): region\*condition, mileage\*displacement, year\*condition, trim\*mileage
   * Multiple Linear Model
   * Subset
   * Forward Backward?
   * Ridge
   * Lasso
3. Knn (Curse of Dimensionality): best k = 10 (10-fold cross validation)
4. Tree Models (Discover non-linear relationship)
   * Tree: choose best alpha
   * Random Forest: 60 trees best mtry = 15 from cross validation *Plot for diff ntrees*
   * Boosting: n.trees=5000, interaction.depth=10,shrinkage=0.001

Summary

**Rankings**

[1,] "kknn" "11096.6776827371"

[2,] "ridge" "8215.71376441645"

[3,] "lasso" "8568.11424790337"

[4,] "tree" "7654.43831665766"

[5,] "boost" "7260.87711749964"

[6,] "rf" "7012.37672045767"

Predictive Models Group Project: Pricing Cars

Mark Babbe, Camryn Callaway, Siqi Chen, Korawat Tanwisuth, Zhiyi Yang

Introduction

Our goal is to construct a predictive model in order to make predictions on car prices based on

given features. To create and validate our models, we randomly split the dataset into training

and test sets, each containing 70% and 30% of the data respectively. We use the training set to

fit proposed models and use the test set to validate the performance of each model regarding

prediction accuracy. For each model, we use cross validation within the training set to select the

parameters that give the lowest cross-validated RMSE. In the end, we make comparisons among all models and select the one with the best prediction accuracy. The following three categories of models are proposed: linear regression models, K-nearest neighbor, tree-based models.

Exploratory Data Analysis

We first examine the data to see if there is any missing values or unrelated variables. The dataset Cars.csv contains 17 variables and 29466 observations. We notice that variable X is simply the index of each observation which is unrelated for predicting price, hence it’s removed from the

dataset. Out of the remaining 15 variables other than the response variable price, 3 are

numerical and 12 are categorical. For numerical predictors mileage and year, scatter plots

versus price show that there might be some polynomial relationship, which could be further

examined in linear regression models. For categorical predictors, there exist many observations

with the factor unsp. which we initially believe that we could treat them as missing values, however,

box plots for each of the categorical predictors versus price indicate that unsp itself as a category

could be informative for price, so we keep it as a factor level.

We also found two pairs of perfectly correlated predictors: (subTrim, fuel), (state, region).

There are only two levels in subTrim: hybrid and unsp, while hybrid is already included in fuel

as a fuel type. Also, region can be expressed as a linear combination of state since each state

corresponds to a specific region. However, region is defined based on employment status,

population composition, weather conditions, etc. which could be more informative than state.

Furthermore, some levels in state contains only few data points, which could result in biased

prediction. Therefore, we decide to remove subTrim and state from the predictors.

Linear Regression Models

We first fit a multiple linear regression model with all 14 predictors and examine the residual

plot for necessary variable transformation. The plot shows that there’s an increasing trend in

residual for higher price values. Thus we perform a log transformation on price. The resulting

residual plot shows a relatively random pattern, with significant reduction in RMSEo

. We then

use stepwise regression to examine the significance of mileage2 and year2 as suggested in EDA.

Both results from forward and backward selection include the two second order terms which

indicate the significance. The next step is to examine the potential two-way interactions, since

the predictors consist a large portion of categorical variables. The interaction plots of following

nine pairs of predictors indicate significant interactions: (mileage,condition), (mileage,trim),

(mileage,displacements), (year,condition), (year,displacement), (year,trim),

(condition,region), (condition,trim), (condition,color).

This led us to utilize shrinkage methods to reduce variance in coefficients, which ultimately

Reduce RMSEo

. We fit all the predictors including second order terms and interactions in ridge

and lasso regression, and perform 10-fold cross validation within training set to determine the

best α. The regression results shows significant reduction in RMSEo

Comparing (compared) to multiple

linear regression model, and lasso outperforms ridge with slightly better prediction accuracy.

K-Nearest Neighbors

In general, KNN regression performs poorly comparing (compared) to linear regression in high dimensions, as

a result of curse of dimensionality. In order to examine the actual performance, we first

standardize 3 numerical predictors to reduce bias. We then perform 5-fold and 10-fold cross

validation to determine the best K. Cross-validated RMSE from both models suggest K = 6 for

the lowest prediction error. The resulting RMSEo

from 6-nearest neighbors regression is much

higher than linear regression models as expected.

Tree-Based Models

The algorithm of tree- based models captures potential non-linearity and interactions, thus we fit

all 14 predictors to three tree-based models: single tree, random forest and boosting. We

anticipate that tree-based models will outperform the models we have attempted so far as they

are more perceptive to non-linear relationships, and to no surprise, RMSEo

is further reduced in

all three models.

For single tree model, we perform 10-fold cross validation to determine the best tree size (number

of terminal nodes) for the lowest RMSE. For random forest model, we perform 5-fold cross

validation to determine the best combination of number of predictors (mtry) and number of trees

(n.trees). For boosting model, we also perform 5-fold cross validation to determine the best

combination of number of split (interaction.depth), number of trees (n.trees) and λ

(shrinkage). As expected, RMSEo

from single tree model is slightly higher than random forest

and boosting models, since the latter two models reduces bias through resampling and stepwise

residual-fitting respectively.

Summary

best model. variable importance!