Assignment 1 : Predict 454

Rahul Sangole[1]

[1] Northwestern University

Assignment 1 : Predict 454

Overview of Methodology

The approach used for this assignment, modeling the "Two month's Salary" dataset is as follows:

1. Data are split into train and test datasets before further consumption
2. Exploratory Data Analysis (EDA) was conducted at a univariate, bivariate and multivariate level using numerical and graphical summarizations
3. The EDA influenced the data preparation stage in terms of transforming the response variable, and adding new features to the predictor matrix
4. Models are built on two sets of data preparation: (i) with no interaction terms, and (ii) with interaction terms. The types of models investigated are linear regressions, variable selection procedures on regression (best subsets, forward & backward selection, lasso & ridge), tree models (recursive partition trees & conditional inference trees) and random forest models
5. Two staged of model building are conducted. In the first stage, models are built using individual R commands using each model's respective package (for example, `lm`, `regsubsets`, `party` and `partyKit` etc). This allowed a deep dive into the residuals and diagnostics for each model, which influenced the modeling strategy. Once the strategy is finalized, the `caret` package is used to build all models again using the `caret::train` interface. `caret::trainControl` is used to perform 10-fold cross validation to perform hyperparameter tuning and to determine the variation of performance for each model.
6. RMSE, MAE and RSquared are the metrics used to judge the performance of each model.
7. Once all the models are built using `caret`, each model is applied to the test dataset.
8. Model performance is compared on both the training and testing data, which reveals

which model has generalized the best.

Exploratory Data Analysis

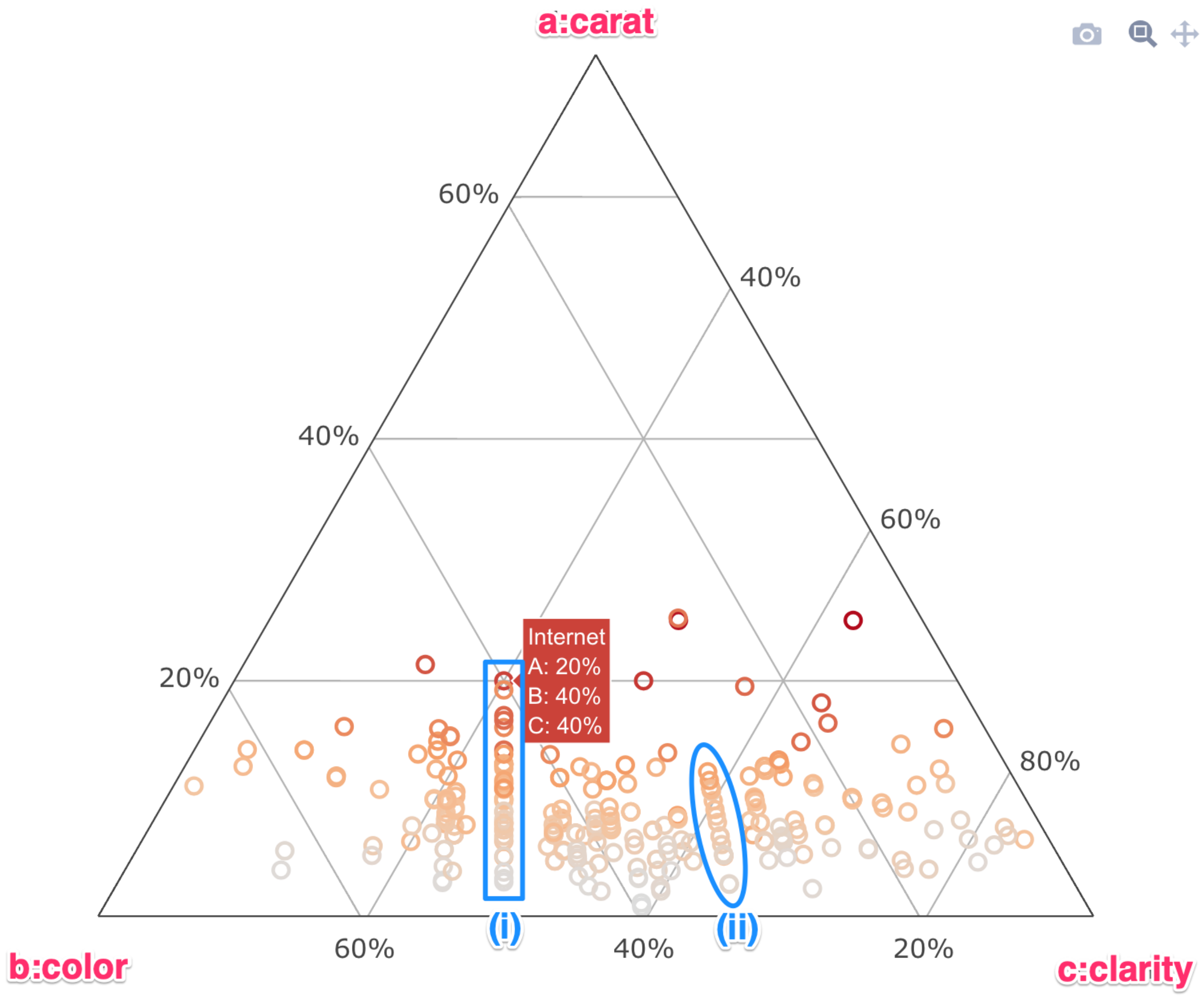


Figure 1. Ternary Plot

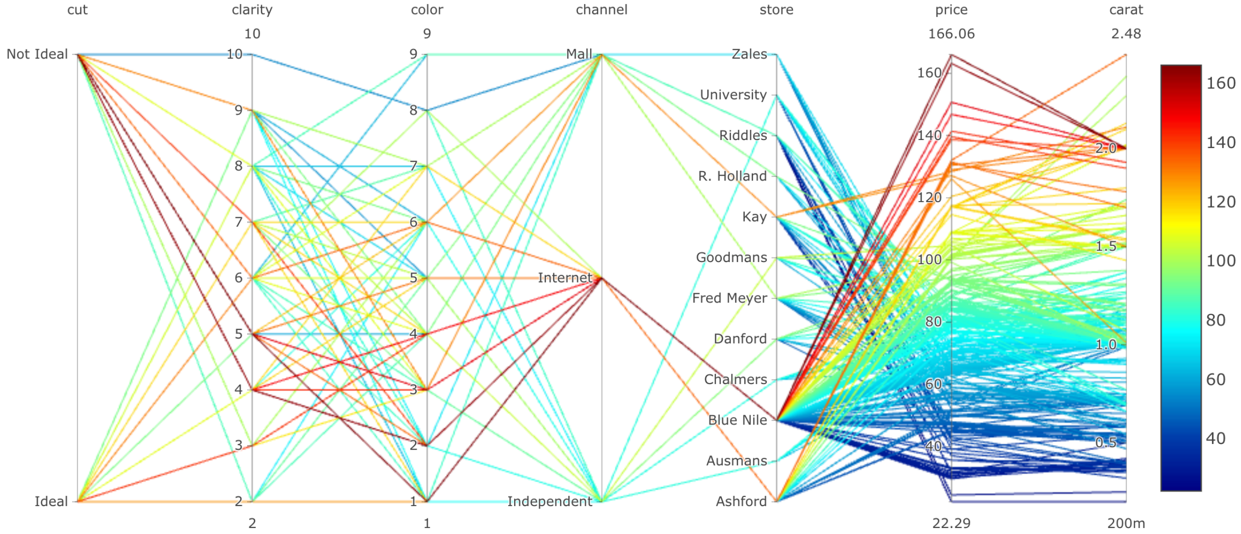
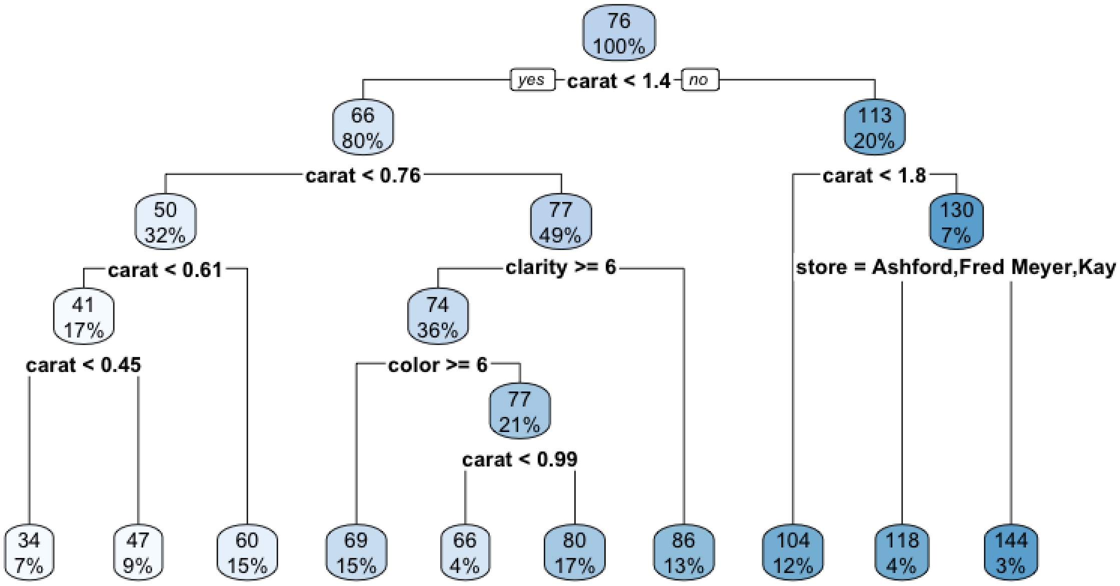


Figure 2. Ternary Plot



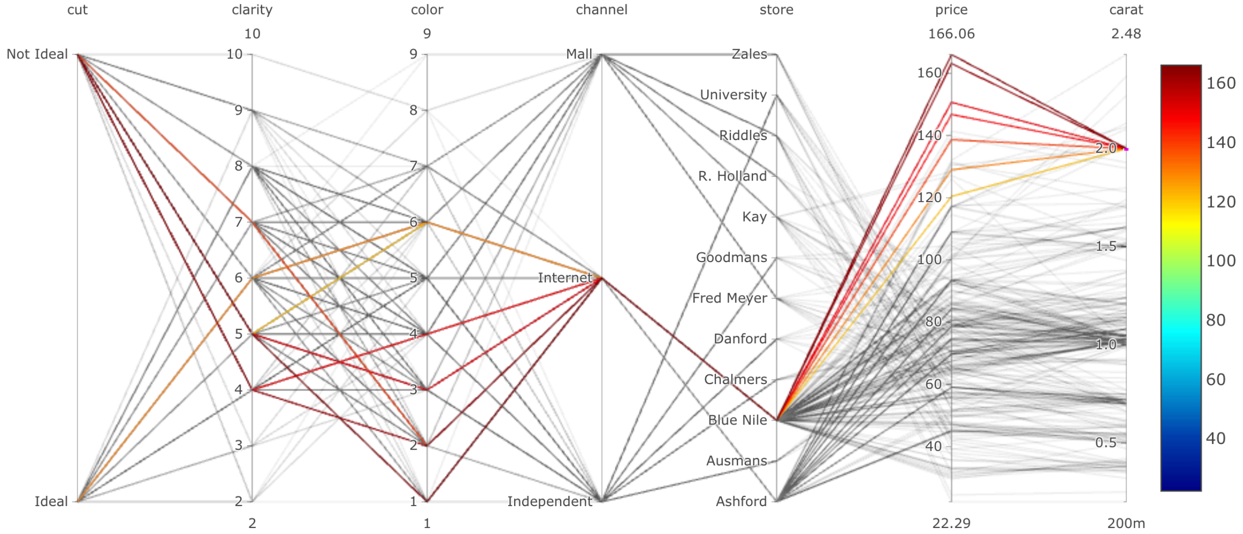
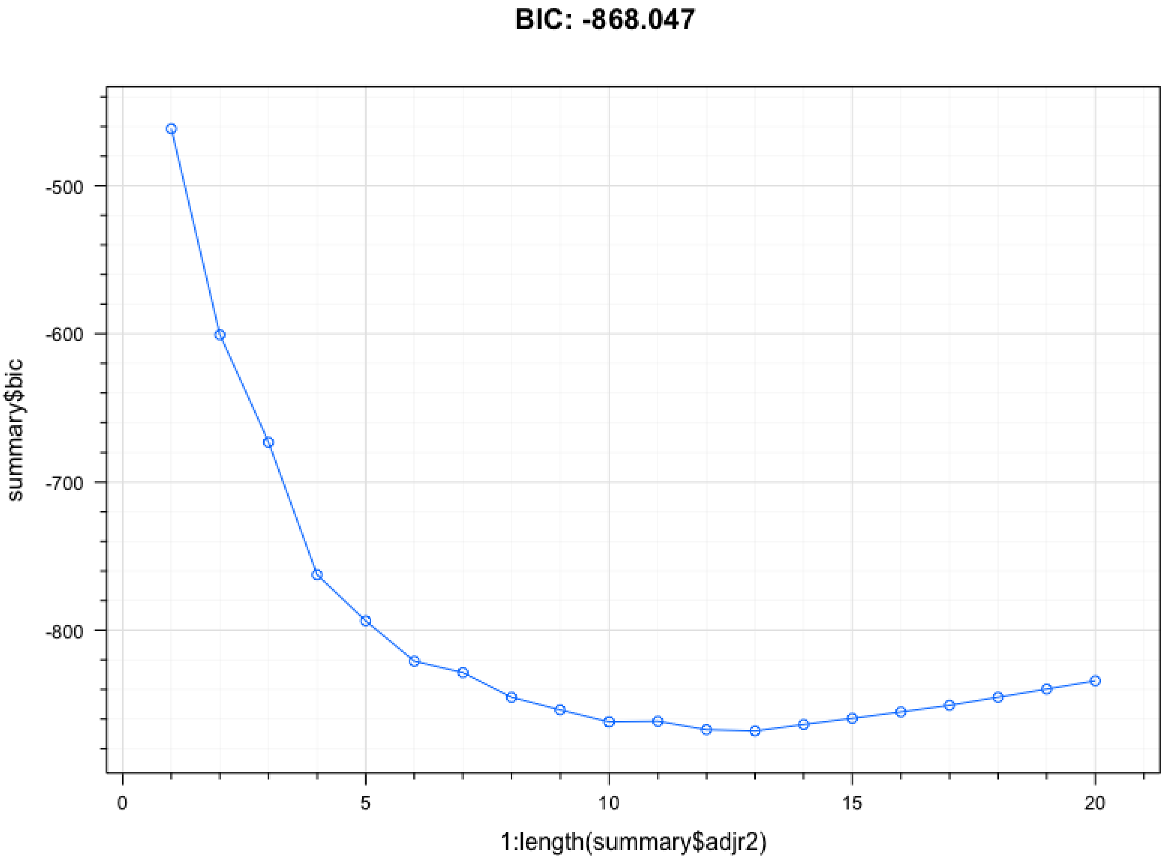
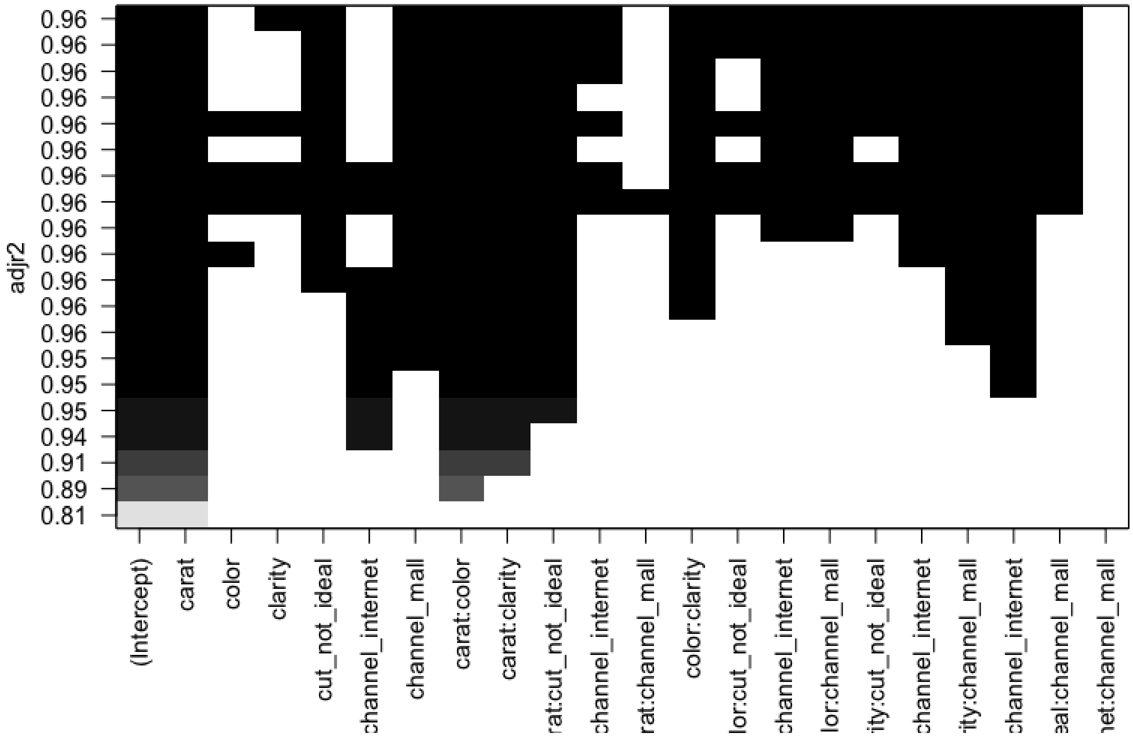
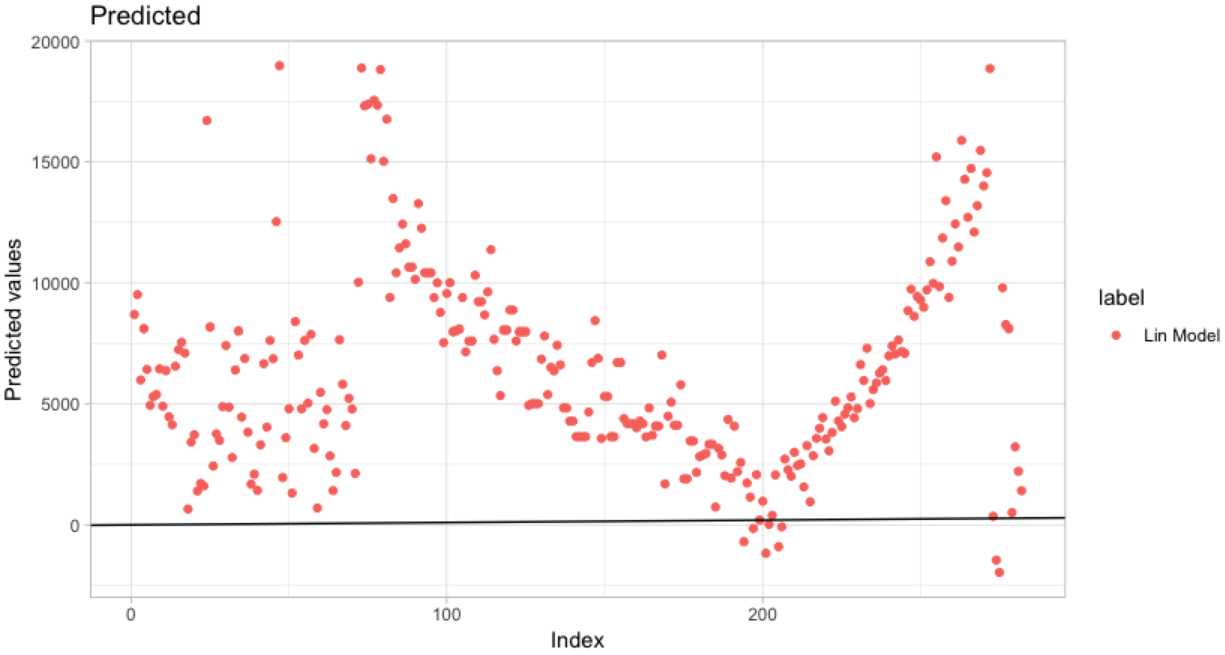
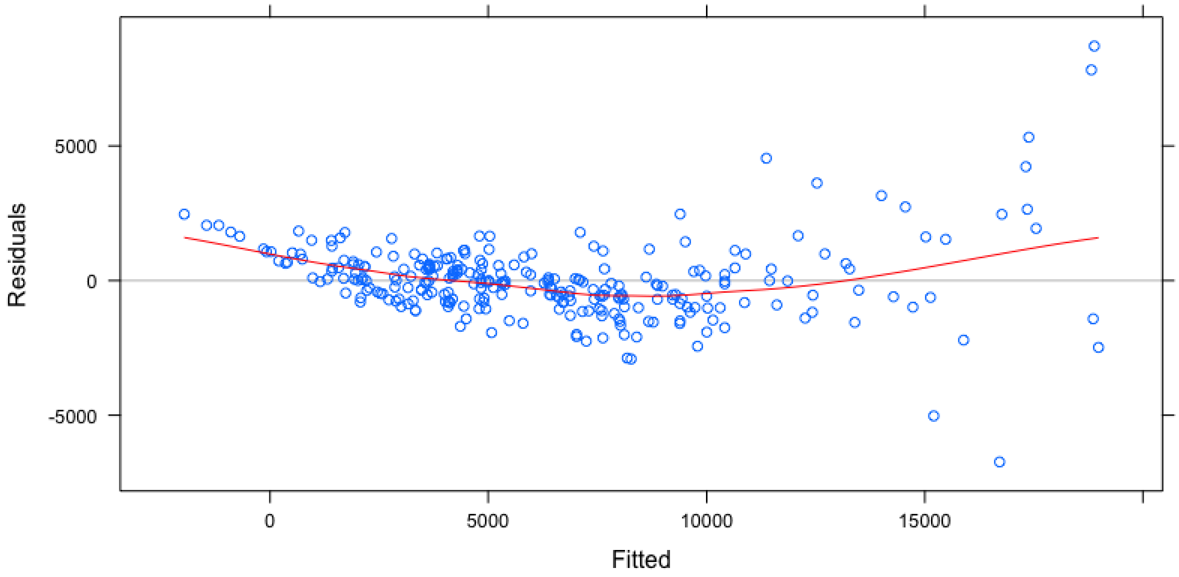


Figure 3. Ternary Plot




adjr2
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.96
0.95
0.95
0.95
0.94
0.91
0.89
0.81
(Intercept)
carat
color
clarity
cut_not_ideal
channel_internet
channel_mall
carat:color
carat:clarity
rat:cut_not_ideal
channel_internet
rat:channel_mall
color:clarity
lor:cut_not_ideal
channel_internet
lor:channel_mall
ity:cut_not_ideal
channel_internet
ity:channel_mall
channel_internet
al:channel_mall
et:channel_mall


Predicted
Predicted values
20000
15000
10000
5000
0
0
100
200
Index
label
Lin Model



Data Preparation

Modeling

Comparison

Challenges and learnings

R Packages Used