BY- ISHAN DHAWAN
PUID:0029958159

IE 590:PREDICTIVE MODELLING

MIDTERM-1

1) Fitting the model

A) Data Cleaning: Initially, the N/A values were removed from the Data set and exploratory Data Analysis was performed to remove the insignificant columns from the data. The columns containing MXSD, TSNW, DX32, DT00 were removed because they contain insignificant observations and majority of the observations were 0 in the columns.

B) Finding correlation: A correlation between the different variables was plotted to find the highly correlated variables and remove them from the Dataset.

C) Model building:

- The data was split into testing and training sets for analysis.
- A General Linear Model (GLM) was used to find the significance of different variables. From the General Linear Model we observed that the variable UNEMP is highly correlated and therefore, has to be removed from the model.
- After removing the UNEMP variable from the dataset a Forward Stepwise Regression model was developed and the most important variables were selected according to the results generated. The stepwise selection procedure initially includes a model with no predictors and then adds predictors to the model, one-at-a-time until all the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. The most important variables at this step were: com.sales.adj, month, HTDD, DT90, CLDD, DT32, VISIB, WDSP, DP05, MWSPD, DP10, MMXT, MNTM.
- A simple linear model was developed taking the variables obtained from Forward Stepwise Regression and utilizing the data from the Training set. The residual plots were developed to check heteroscadasticity, and I observed that the residuals were normally distributed without any specific shape or pattern. The predictions were made using the Testing Dataset and the RMSE value obtained was: 344.1375. Now, performing k-fold cross validation (10 folds) the RMSE value reduced to 311.6678.
- The Boxcox Transformation was the next step to performed and I observed that the RMSE value for the transformed was the poorest and hence, we will not perform the Boxcox transformation.

After selecting the best subset from the stepwise regression shrinkage approach is applied which involves fitting a model involving all the predictors. The shrinkage approach has the effect of reducing the variance and also serves the purpose of variable selection for further applications in the model. For shrinkage and variable selection the two approaches used were: Ridge Regression and Lasso Regression

Ridge Regression: In Ridge Regression a shrinkage penalty is applied to the variables to make them zero. A main disadvantage of Ridge Regression is that it will include all the predictors in the final model. The penalty will shrink all the coefficients towards zero, but it will not set any of them exactly to the zero. After performing the Ridge Regression the RMSE value obtained was 360.3958. Even after performing k-fold cross-validation the RMSE value remained unchanged.

BY- ISHAN DHAWAN
PUID:0029958159

Lasso Regression: Lasso Regression overcomes the shortcomings of Ridge Regression, the shrinkage penalty shrinks coefficients towards zero. The RMSE value obtained after performing k-fold cross validation is 333.226. Even though the RMSE value for Lasso Regression is slightly larger than the RMSE of Linear Regression I would choose the variables from the Lasso Regression because insignificant variables are shrunk to zero after applying the Shrinkage penalty. The final set of variable selection from Lasso Regression is month, com.sales.adj, MMXT, DT90, DT32, DP05, VISIB, GUST, WDSP, HTDD, CLDD and PCINCOME.

Generalized additive models (GAM): The variables selected from Lasso Regression are used to make a Generalized Additive model. GAM provides a general framework for extending a standard linear model by allowing non-linear functions of each variable, while maintaining additivity. In the GAM model smoothing splines are used via an approach called backfitting. This method fits a model involving multiple predictors by repeatedly updating the fit for each predictor in turn, holding the others fixed. The best model is selected with the help of the AIC criterion and the model with the lowest AIC value is taken. The lowest AIC value in the best model obtained from GAM is: 3663.16. The predicted RMSE value for the testing set is : 294.9596. So, far GAM is the best model obtained.

Advantages of GAM:

- GAMs allow to fit non-linear models and we do not have to try different transformations.
- Non-linear fit can potentially make more accurate predictions for the response variable.
- GAMs provide useful inference due to the additivity of the model.

Disadvantage of GAM:

The main limitation is the additivity of the model because of which we may miss many important interactions between the variables.

Random Forest: Random Forests are superior to bagged trees because they employ a small tweak that decorrelates the tree. When building decision trees using Random Forest, each time a split is considered, a random sample of predictors is chosen as split candidates from the full set of predictors. The algorithm is not allowed to consider a majority of available predictors. Random forest overcomes the problem of bias by forcing each split to consider only a subset of predictors. The predicted RMSE value for Random Forest in the dataset is 276.7002 which is the least among all the algorithms.

Regression Trees: Regression trees, on the other hand, give you a piecewise linear relationship between the predictor and the predicted variables that is freed from the constraints of superimposed continuities at the joins between the different segments. The RMSE value obtained is 445.7188 which is the highest I have obtained so far. This means that the Regression Tree model is the least acceptable model.

Advantage of Regression Trees:

- They are very easy to explain. Even better than the linear regression model as they can be displayed graphically and are easily interpreted

Disadvantage of Regression Tree:

The trees can be very non-robust i.e a small change in the data can cause a large change in the final estimated tree.

BY- ISHAN DHAWAN
PUID:0029958159

2) Describing the Final Model

The Final Model according to me that will give the best predictions will be the Random Forest Model.

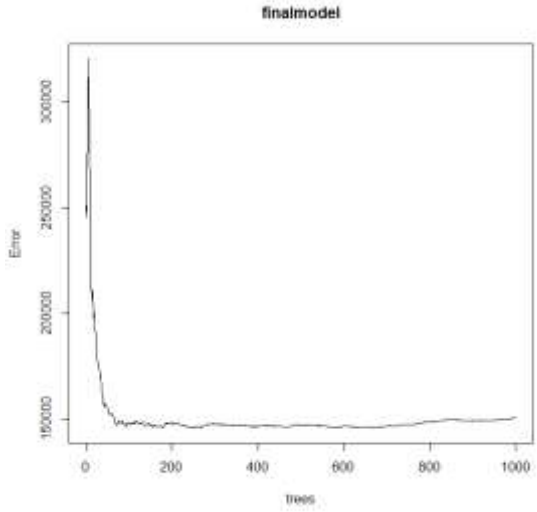
Since, the Random Forest model is a black box and it takes mean of the different trees generated so it is not feasible to generate only one tree as it will not be a true representation of the final model.

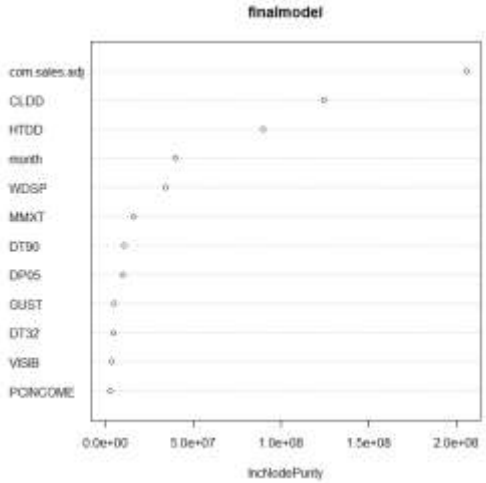
Instead of that I have plotted the variable importance plot which gives the importance of the each variable and we can compare the most important variables in the model.

Another plot, which shows that the error decreases when the numbers of trees are increased, is also important.

The final equation of the model will be:

```
finalmodel<-randomForest(res.sales.adj ~  com.sales.adj+month+MMXT+DT90+DT32+DP05+VISIB+WDSP+GUST+
                         HTDD+CLDD+PCINCOME, data=training_set_Midterm,ntree=1000,importance=T)
```





BY- ISHAN DHAWAN
PUID:0029958159

3) Reasons for selecting the model

The Random Forest model was selected because of numerous reasons:

- The RMSE value was the best for the Random Forest Model(RMSE=267.7002)
- There was a slight improvement in the RMSE value of Random Forest in comparison with Generalized additive models and hence, I selected Random Forest as the Final Model.
- The accuracy obtained is better in the case of Random Forest
- The variance is less because we are using multiple trees and the chance of stumbling across a predictor that doesn't perform well is reduced.
- The GAM model is simply additive and doesn't include the interactions between the different predictors whereas in Random Forest the interactions between predictors is also observed.
- Even though GAM is easier to interpret than Random Forest has better predictive modelling ability and I would not want to compromise on the predictive modelling capability as compared to the descriptive capability so, I would prefer Random Forest over GAM.
- Random Forests don't require any input preparation and they perform implicit feature selection and provide a pretty good indication of feature importance.
- Random Forests are versatile in nature. They can be used for variety of modelling tasks, they work well for regression tasks as well as for classification tasks.

<u>Disadvantages of Random Forest:</u>

- Due to high complexity Random Forests are not easily interpretable.
- Also since it is a black box and it is difficult to figure out what exactly is the algorithm doing and how it is doing it.

From the Variable Importance plot for Random Forests we can see that the most important variables on which res.sales.adj variable depends are com.sales.adj, HTDD, CLDD, WDSP and month.
I find that the dependence of res.sales.adj on the highly important variables is accurate.
The res.sales.adj depends on the month as the electricity sales will be the highest in the hottest and the coldest months due to excessive consumption of electricity to run air conditioners and heaters.
Res.sales.adj depends on the heating degree days because higher the higher the number of days with more heat higher will be the consumption of electricity.
Res.sales.adj also depend on the cooling degree days as higher the number of cool days there will be more consumption of electricity to run the heaters.
The windspeed (WDSP) also plays a role in influencing the electricity sales as we may have fluctuations in temperature due to windspeed which may increase or decrease the residential electricity sales.
A change in the consumption of commercial electricity sales may mean that the industries are sunning overtime or their outputs are contributing to an increasing GDP. Increasing GDP may mean higher salaries for employees and in turn the people may increase their residential electricity consumption due to buying new appliances.