Stat 220 Lab 11

Project Description

Project Description: You are hired as data scientists by Mashable, an online news platform that generates buzz through shares of their posts. You are tasked with the responsibility of building a model that will be able to predict the number shares a news article will get based on characteristics of the article.

Data: The data comes from Mashable.com and the source is the UC Irvine Machine Learning repository: `https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`. You can download the data from there or at `https://richardson.byu.edu/220/OnlineNewsPopularity.csv`. There are 61 variables total. A description of the variables can be found at `https://richardson.byu.edu/220/ONPvariables.txt`. The target variable is the number of shares a news article receives and it is the last variable in the data set.

Deliverables: Your work will culminate in two key deliverables:

1. A script or notebook containing all analyses and modeling steps.
2. A technical report for Mashable, written according to the instructions below.

Project Details

Exploratory Data Analysis:

1. Plot the target variable. Determine if the target variable seems appropriate or if any transformations should be made.
2. Build a linear regression model without any higher order terms and determine what predictors are most significant

3. Build a regression tree and use important features to determine predictors that are most significant.

4. Choose several significant features from steps 2 and 3 and create visualizations or tables that explore the relationships between the target and your selected features.

5. Write a section of your technical report called Exploratory Data Analysis that reports the results of the initial models built and provides figures or tables that show the target variable as well as relationships between the target variable and some potentially significant predictors.

6. Use methods to remove insignificant variables from the model

Linear Regression Modeling: Build and tune a linear regression model that has high predictive power and can be used to explain to Mashable what features are most important in influencing the number of shares.

1. Split the data into a train and test set. Use the train set to fit the models and use the test set for checking for overfitting and good predictive power.

2. Explore using transformations of the target variable and other variables

3. Explore using higher order terms

4. Reduce the model in the two following ways:

    - Use stepwise model evaluation methods to remove insignificant variables from the model
    - Use LASSO Regression to fit the full model and have the model remove insignificant variables. Tune the model to find the best α.

5. Write a section of your technical report reporting the out of sample model performances for the models you have built. There will likely be many significant predictors. You don't need to name them all but discuss what seem to be the most significant or important ones. Discuss how useful you feel this model will be for predicting future shares.

Linear Regression Modeling: Build and tune a regression tree model.

1. Use the same train and test groups as above.

2. Use cost complexity pruning and cross validation to find a model that fits well on out of sample data.

3. Fit a Random Forest regression model to the data using cost complexity pruning for the individual trees.

4. Write a section of your technical report reporting the out of sample model performances for the models you have built. Discuss how useful you feel this model will be for predicting future shares.

Conclusion: Compare each model's predictive accuracy on the test set. Choose the model that performs the best as the best predicting model. Write a concluding section of your technical report that addresses the business concerns of Mashable and presents your final model and your confidence in it.