Stat 220 Final Lab

Project Description

Project Description: You are hired as data scientists by Mashable, an online news platform that generates buzz through shares of its posts. Your task is to build a model to predict the number of shares a news article will receive based on its characteristics.

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

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 (EDA):

1. Plot the target variable. Determine if the target variable seems appropriate or if any transformations are needed.
2. Build a linear regression model without higher-order terms and identify the most significant predictors.
3. Build a regression tree to identify important predictors.

4. Select several significant features from steps 2 and 3. Create visualizations or tables to explore the relationships between these features and the target variable.

5. Write an EDA section in your technical report. Report the results of the initial models and include figures or tables that show the target variable and its relationship with potentially significant predictors.

6. Use appropriate methods to remove insignificant variables from the model.

Linear Regression Modeling: Build and tune a linear regression model with high predictive power, explaining to Mashable which features most influence the number of shares.

1. Split the data into training and testing sets. Use the training set for model fitting and the testing set to check for overfitting and predictive performance.

2. Explore transformations of the target and other variables.

3. Explore higher-order terms.

4. Reduce the model using the following methods:

   - Stepwise model evaluation methods to remove insignificant variables.
   - LASSO regression to fit the full model and remove insignificant variables. Tune the model to find the best α.

5. Write a section in your technical report that reports the out-of-sample performance of the models. Discuss the most significant predictors and evaluate the model?s usefulness for predicting future shares.

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

1. Use the same training and testing sets 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, using cost-complexity pruning for the individual trees.

4. Write a section in your technical report that reports the out-of-sample performance of the models. Discuss the model's usefulness for predicting future shares.

Conclusion: Compare each model's predictive accuracy on the test set. Choose the best-performing model as the final predictive model. Write a concluding section in your technical report that addresses Mashable?s business concerns and presents your final model along with your confidence in its predictions.