Wood for Guitar

1. Introduction

Guitar industry has witnessed exponential growth in the past 5 years and is a billion dollar industry presently. Over the next five years it is anticipated to continue its trend and keep growing. The most important part of a guitar is its Fret board which is made with different type of woods and its quality and integrity is the first thing a guitarist seek for. Type of woods depend on the forest (cover) they belong to. Gibson, one of the leading guitar manufacturing companies, wants to build a model to predict the type of forests cover given geographical, physical and coordinate information about the place. Cartographic data belongs to US Forest Services Resource Information system and US Geological Survey.

2. Problem Statement

We are provided with training and testing data sets by Kaggle. In the training data set (15120 observations) we have information regarding predictor variables: elevation, aspect, slope, distances to water bodies (horizontally and vertically), hill shade angles, wilderness area it belongs, soil type of the forest and dependent variable: forest cover type. Test data has 565892 observations of the predictor variables which needs to be predicted to know the forest cover they belong to, so that the company can exploit the required wood for their guitar.

Data set: UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Covertype)

3. Methodology

I used Weka Prediction API and R to build a Naïve Bayes model to predict the dependent variable. As the forest type of test data is unknown, a cross-validation was performed to know the estimated error in the predicted values, which gave me an accuracy of 66.44% with Weka prediction model and 66.48% with R. The result of prediction for test data set in Kaggle were uploaded which evaluated the prediction to be 53% accurate.

Better results were achieved other prediction models, best of which being: Support Vector machine (74%), K means (71.016%) and combination (71.77%) of SVM, Naïve Bayes, and K-nearest neighbors. The accuracy table for different methodologies applied is given below.

4. Evaluation

The preliminary findings results in 74% accuracy for test data and 80% while cross validating in SVM. I believe that results are good so far as we have 565892 observation but it can get better with feature selection, Bagging and Random forest.