Reflection Paper

Project: Forest Cover Type Prediction
Owner: Sumith Gannarapu

Project Process Summary:

The problem consists of predicting the Cover Type of the forest area based on the cartographic variables collected from the forest area. This problem is a multiclass classification problem with 7 target class variables each depicting a different cover type and 55 predictor variables each depicting the variables relating to sun light, water, soil, elevation, forest fire point distance and distance to road. The variables consist of numerical variables, categorical variables of binary class and quaternary class.

We used the **CRISP-DM** (Cross industry Standard process for data mining) approach for the project. We spent a good amount of time on understanding the project. We did a thorough research on the purpose of our project and its real-world applications which serves as a motivation for the project and make the most of it. We developed the necessary domain knowledge for the project which helped us later in the feature engineering stage. We also came up with a rough project execution plan during this stage.

The next step was data understanding. We studied the variables in our dataset by plotting them individually and with comparing to other variables. The reason was we are looking for the distributions that these variables follow and the permitted domains these variable values can take.

The next important stage we went to was data preparation. We looked for things like correlation, missing values, outliers and dealt with these cases by identifying the outliers using Grubbs test and also detecting and identifying through plots and removing the outliers and removing the highly correlated variables. We also remove the variables with zero variance. We also did PCA (principle component analysis) to come up with new features. We checked for skewness in the variables and transformed the variables with high skewness using SYMBOX. We did feature construction by coming u with five new features which we used for the models.

The next stage we went to was modelling stage. We held out 30% of our data for validation and constructed our models on 70% of our data. Here we started off with basic GLM model. We constructed various other models subsequently. The list of models we created are as follows

1. Basic GLM
2. Decision Trees with Information Gain
3. Decision Trees with GINI index

4. Support Vector Machines
5. Random Forest
6. Gradient Boosting method
7. K Means Clustering

we did hyper parameter tuning for the models 2 to 6 as these are the advanced models and improved on their accuracy and formulated the best models in a table.

We also did clustering on the data set as we already knew about the number od target classes that we need to cluster our data on.

Summary of Feedback from the Partner:

As we took care of all the steps that have been taught in our course, the summary of feedback for our project from our partner did not point to anything that we missed in the project. The Feedback mentions that the problem description and introduction was very clear.

One of the comments mentions that why we held 30% of our data for validation and not use NxK fold cross validation as it makes the most of our training data and training will be done in a robust way.

The feedback also mentioned that the feature selection and analysis was very clearly explained, and coefficient analysis was very clear and good.

The feedback paper also mentions that "It would be more informative to show the comparison of the results provided by the Logistic Model and the Random Forest Model in the same table". the feedback also asks about the metrics that we used other than Accuracy to compare distinct models. It also asks us about the error metrics that we are using to decide the best fit.

Discussion on feedback:

The most valuable feedback was it asks us about the other metrics we used other than accuracy to compare the models and decide on what is best fit. It asked us to compare the results of different models which we did not do because we do not have enough metrics during the first draft submission to compare the models.

The comments I discarded or disagreed with are it asks us why we did not use K fold cross validation with N repeats. The reason we discarded this comment was we used the k fold cross validation with N repeats which he might have missed during the review. We held 30% of the data as well as the we did the cross validation.

Reflect On iterations:

We started by constructing various basic models with the available variables. We observed that our models did not show much performance. We then revisited the feature engineering stage and came up with various new features and improved on our model performance. The other thing we did to improve our performance was cross validation. We tuned using various range of parameters for each of the models like complexity parameter for decision tree model, number of trees and mtry for random forest etc.

The things that led us to make these changes was our understanding of the models and thorough literature review that we did during the project understanding phase. The other things we did to improve on our model was we tried constructing a wide variety of models and compared with different metrics like accuracy, Kappa, AUC, F1 score.