FOREST COVER TYPE PREDICTION

PROJECT BY

SUMITH GANNARAPU

Executive Summary

Concise problem statement:
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.

List of major concerns/assumptions:
The major assumption with the above problem was, we assumed that the forest cover types are a result of ecological processes rather than human intervention. The other assumption was we considered the water source that is available on land, but we did not consider the ground water at that area which might play a considerable role in determining the cover type. The other assumption was the soil.

Summary of findings:
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. By plotting the histograms on continuous variables, we found the skewness for some variables which helped in making the skew transformation. We also found that correlation exists among some predictors. Modelling process consisted of creating both random forest and logistic models. The first natural step in creating basic logistic model will help you to understand all of 55 predictor variables and this model was used as a stepping stone to other models. In random forest model we found the most significant variable in based of mean decrease in accuracy as well as contributing to reduction in node impurity. In Gradient Boosting machine learning technique, we found most significant variables using variable importance plot and from this plot we Elevation, HD_road, Soil_type101, HD_fire, HD_hydro, PythDist explained maximum variance. Decision tree model didn't help much on improving accuracy, we tried with information gain and Gini index and among which Gini index performed better than other. In Support Vector Machine, SVM works very well in high dimensional data and our data is high dimensional and Our data set is not too large, hence by doing hyper parameter hyper parameter we achieved better accuracy than decision trees.

Recommendations:
Based on our exploratory analysis and classification we found that feature engineering plays a very crucial role as our model has improved with the addition of relevant features. By predicting the forest cover type.
we could be able to address many issues including forest fires, soil erosion and maintain flora diversity. For example, if we can recommend certain kinds of plants that grow well in certain landscapes so that the land is utilized efficiently.

Problem Description:

If we know the soil composition, geographical location and the availability of water we can predict the type of trees that grow in an area. The problem consists of identifying the different kinds of forest covers from four wilderness areas located in the Roosevelt National Forest of northern Colorado. The four areas are RAWAH, NEOTA, CACHE LA POUDRE and COMANCHE PEAK each differing by mean elevation level and types of vegetation grown in that area. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices. By identifying the types of covers we could be able to address many issues including deforestation, forest fires, soil erosion and maintain flora diversity.

The actual forest cover type for a given observation (30 x 30-meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data.

Data Description:

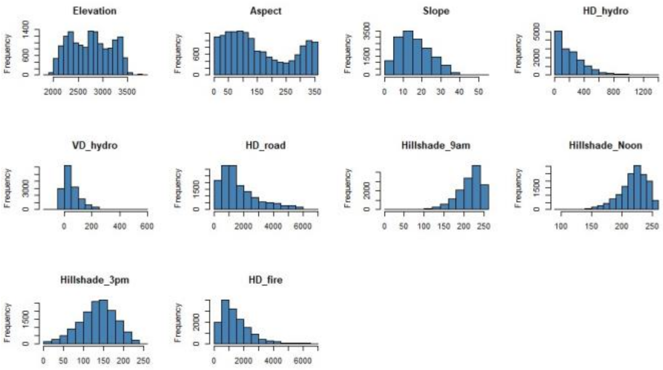
The training data consists of 10 quantitative variables, two binary class qualitative variables with a combined total of 44 binary columns and seven factor variables of predictor classes. The data doesn't contain any missing values. The Training data consists of 15,120 records and test data consists of 5,65,892 records.
Integer classification for the forest cover type. The seven types are:
1 - Spruce/Fir
2 - Lodgepole Pine
3 - Ponderosa Pine
4 - Cottonwood/Willow
5 - Aspen
6 - Douglas-fir
7 - Krummholz

The predictor variables and their descriptions are as follows:

Elevation	: Elevation in meters
Aspect	: Aspect in degrees azimuth
Slope	: Slope in degrees
Horizontal_Distance_To_Hydrology	: Horizontal Distance to nearest surface water features
Vertical_Distance_To_Hydrology	: Vertical Distance to nearest surface water features
Horizontal_Distance_To_Roadways	: Horizontal Distance to nearest roadway
Hillshade_9am (0 to 255 index)	: Hillshade index at 9am, summer solstice
Hillshade_Noon (0 to 255 index)	: Hillshade index at noon, summer solstice
Hillshade_3pm (0 to 255 index)	: Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points	: Horizontal Distance to nearest wildfire ignition points
Wilderness_Area (4 binary cols, 0 = no or 1 = yes)	: Wilderness area designation
Soil_Type (40 binary cols, 0 = no or 1 = yes)	: Soil Type designation
Cover_Type (7 types, integers 1 to 7)	: Forest Cover Type designation

Exploratory Data Analysis

Initially we examined every feature individually as well as its interactions with other predictors. The given training data consists 15,120 records. The training data does not have any missing values and training data contains equal number of cover types i.e. 2160 records of each type hence no class re-balancing is needed.
Skewness of distribution of each predictor variable is calculated. We found several attributes in Soil_Type show a large skewed hence we performed skew transformation on such data.



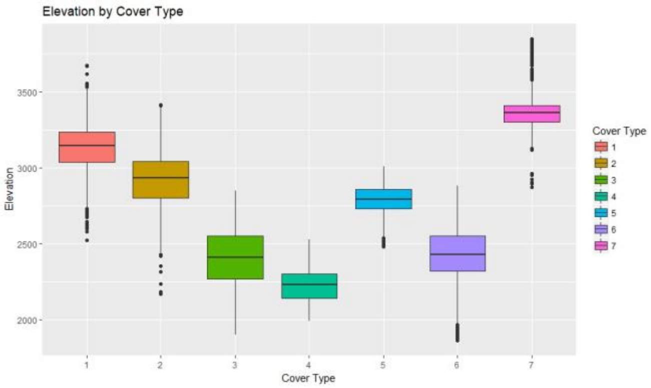
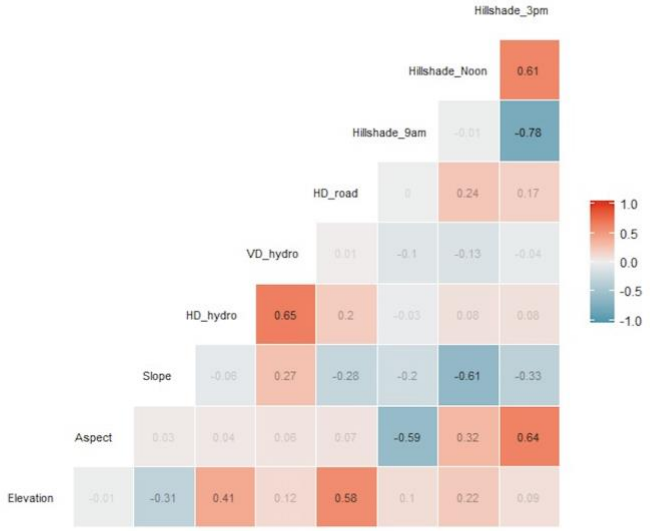


Fig 1.1 Histogram for continuous variables

Fig 1.2 Box Plot for Elevation by Cover Type

Fig 1.1 Histograms are plotted on all continuous variables. The predictor variable aspect followed a bi normal distribution. HD_hydro, HD_road, HD_fire is right skewed where as Hillshade_9am and Hillshade_Noon are left skewed. Hillshade_3pm and Slope seems to be normally distributed. The variable Elevation has a separate distribution for most classes as depicted by a number of peaks in the histogram above and the box plot Fig 1.2 of elevation by its cover type shows a clear separation for each cover type.

Correlation:



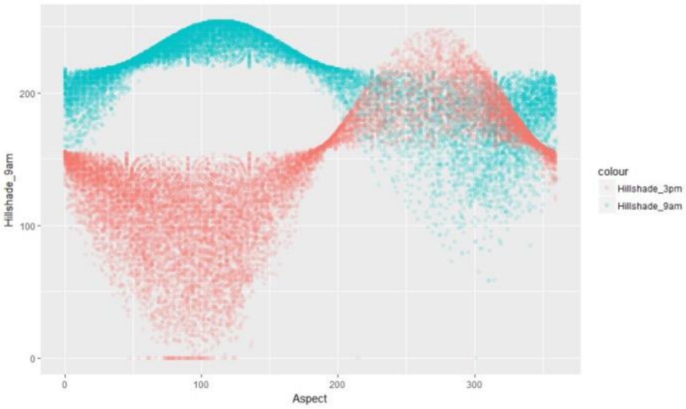


Fig 2.1 Correlation Heat Map of Features

Fig 2.2 Correlation between Hillshade_9am and Aspect

correlation is observed between the some of the pairs. We performed dimension reduction technique using principal component analysis.

Below are the some of the correlation:

Variable 1	Variable 2	Correlation
Hillshade_9am	Hillshade_3pm	-0.78
Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	0.65
Aspect	Hillshade_3pm	0.64
Hillshade_Noon	Hillshade_3pm	0.61

Principal Component Analysis:

To further investigate the data, Principle Component Analysis (PCA) is a method of reducing the dimensionality of data while, usually, maintaining most of the data's variance. All dimensions of the reduced data are linearly uncorrelated, meaning the original data is projected into a space where each component is orthogonal to the others. Figure 3.1 shows the line plot of the first 10 principle components and their associated variances that they explain. From the plot we can see that the elbow occurred at principal component 5 and then the graph flattened. It is clear that the first five PCAs explain a large portion of the variance of the data. Figure 3.2 First two components (PC1 and PC2) itself explains 51% of the variance in the whole matrix. From the plot Hillshade_9am more towards PC2, Hillshade_Noon, Slope are more towards PC1, HD_to_Fire_points, Elevation, HD_hydrology, HD_Roadways and Hillshade_3pmare participating in PC1 & PC2 components.

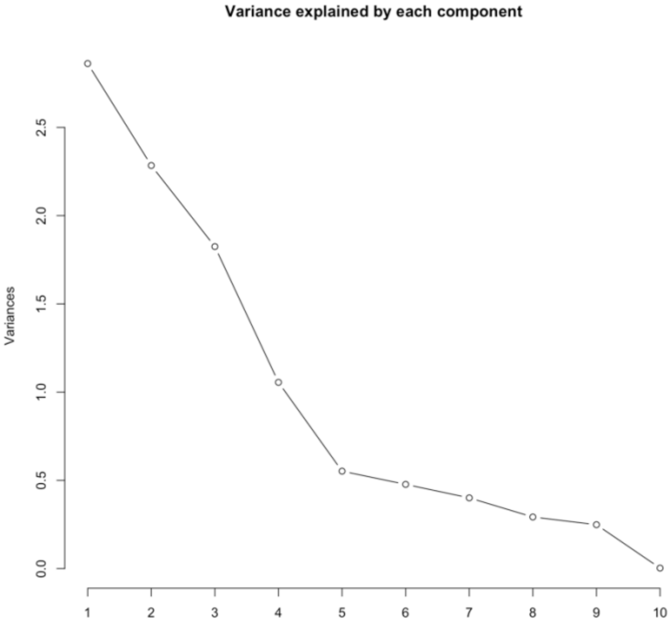


Fig 3.1 Line Plot of First 10 PCAs

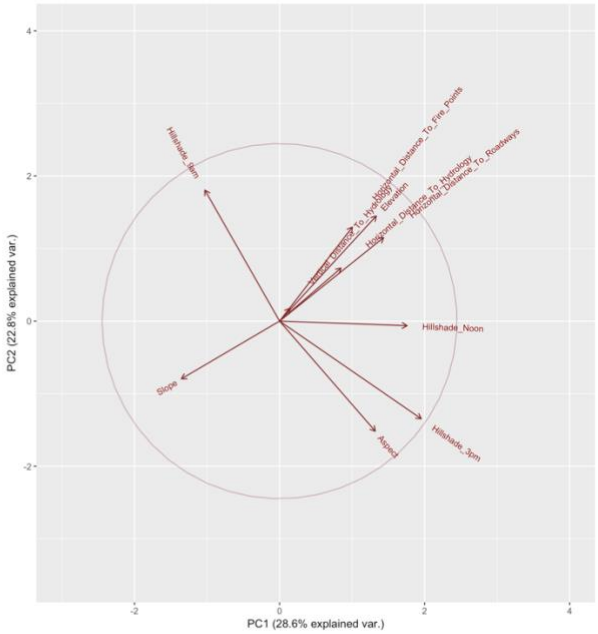


Fig 3.2 Plot of PC1 and PC2

Analysis Plan:

Explanation of Modeling Choice:

The first natural step in creating basic logistic model will help you to understand all of 55 predictor variables and this model was used as a stepping stone to other models. In random forest model we found the most significant variable in based of mean decrease in accuracy as well as contributing to reduction in node impurity. We selected Gradient boosting technique as it helps us in finding variable importance plot which will help us to see the best predictor variables.

We split the training data into two parts with the ration of 70 and 30, first part is to train the model and second part of the data is to validate the model

Modeling Technique and its Strength, Weakness:

Logistic regression helps to understand the coefficients and significance of the variables. The exponential of coefficients corresponds to odd ratios for the given factor. Logistic regression requires that all variables are independent of each other if there is any correlation between variables and then the model will tend to overweight the significance of those variables

Random forest runs fast, and it is good at dealing with unbalanced data. To do classification with Random Forest, it's cannot predict beyond the range in the training data, and that there may be chances of over-fit datasets that are particularly noisy. But the best test of any algorithm is how well it works upon your own data set. We also chose this model because it helps us in finding the feature importance plot which is bases for the understanding the features and coming up with performing feature engineering.

We tried other various models like Gradient Boosting, Support Vector Machine, Decision Trees so that we can compare CV performance metrics with each of the models for the better accuracy.

Data Pre-Processing:

- **Missing Values**
  There are no missing values in the dataset
- **Skew Transformation**
  Fig 1.1 Skew transformation has been applied for HD_hydro, HD_road, HD_fire which are right skewed and Hillshade_9am and Hillshade_Noon which are left skewed
- **Outlier Analysis & treatment**
  After plotting scatter plot for continuous variables, we identified outliers for Vertical Distance to Hydrology and Hillshade at 9AM predictors and we also verified using the Grubb's test, a multi variate approach. Since we have very few outliers in our dataset, we removed outlier records.
  Below are the Outlier plots for distance to hydrology and hillshade 9 AM and 1804,1893,11939 outliers for distance to hydrology and 2721, 2790, 4928, 12807 is an outlier for hillshade 9 AM

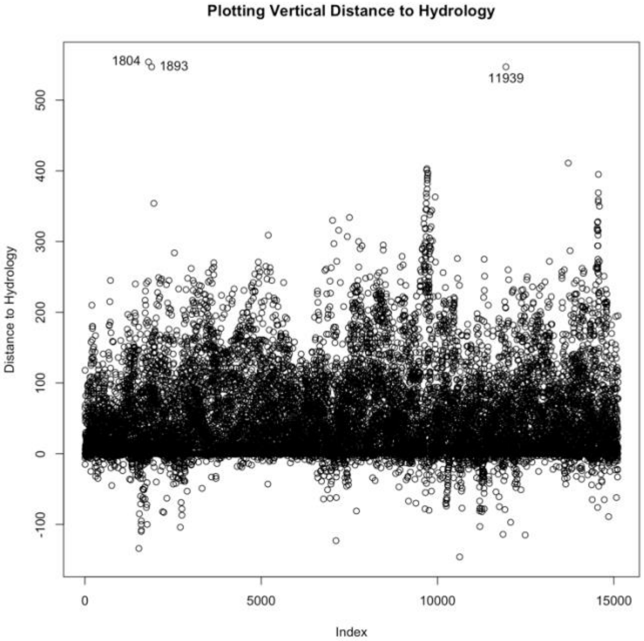


Fig 5.1 Outlier plot for Distance to Hydrology

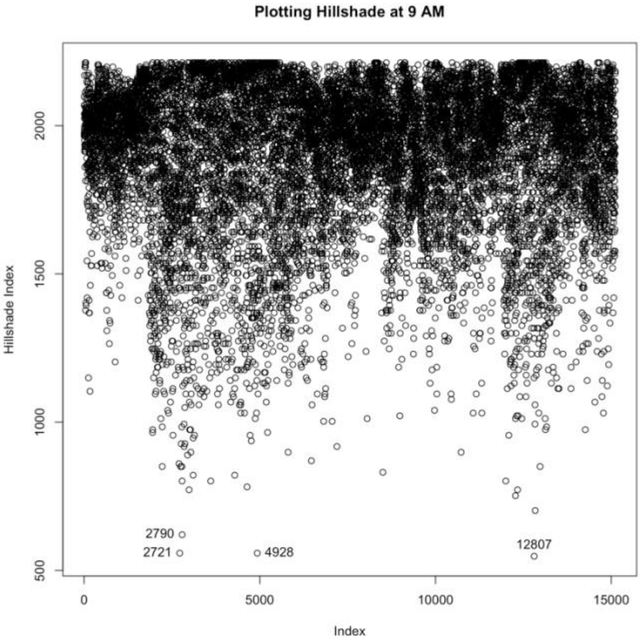


Fig 5.2 outlier plot for Hilshade at 9 AM

Feature Engineering:

Feature engineering is a very important in getting better predictions. Existing features cannot always explain why certain types of trees are grown in a particular area. Also, linear combination of features cannot explain the nonlinear relationship between features. Subject Expertise plays a very important role in feature Engineering After studying and understanding the features we came up with some new features that we think are crucial for the prediction of cover type

- **Euclidian Distance to water source:**
  Since the availability of water source effects, the type of trees grown, and the roots of trees don't always spread just horizontally and vertically, we created a new feature that measures the Pythagoras distance between the water source and tree based on the horizontal distance and vertical distance
- **Feature Removal**
  Not all features are important in the model performance. We calculated the variance of the predictor variables and removed two variables that has zero variance i.e. soil type 7 and soil type 15
- **Mean Hill Shade**
  We created new feature that gives the mean of the shades at three different time i.e. 9am, noon and 3 pm
- **Interaction between the Hill Shades at different times**
  We created new features by considering the linear interactions between the three hill shade features
- **Azimuth Angle transformation**
  We crated new variables by transforming degree azimuth features (Aspect and Slope) to numerical features. We did this because we wanted to keep all the features in the same units. We took the sine and cosine transformations of the Aspect and Slope

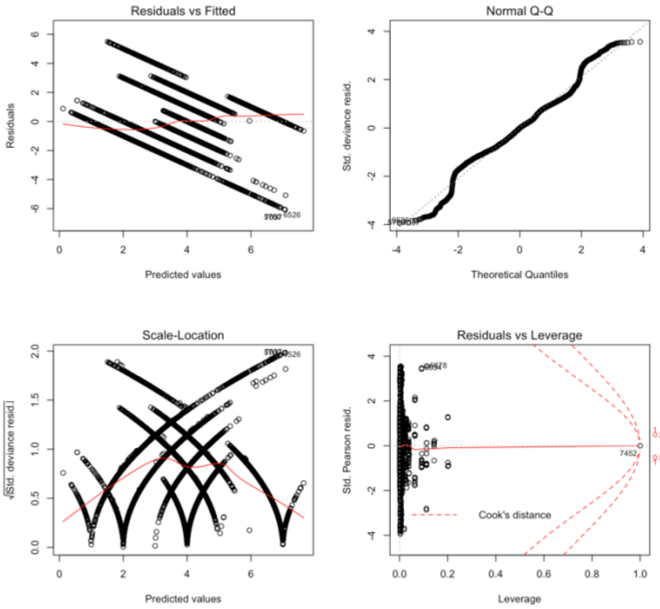
Modeling:

Basic GLM without Hyper Parameter Tuning:

The first natural step in creating model from scratch was to create a model that will help you to understand all of 55 predictor variables. We feel this basic model is the base for other models

Coefficient Understanding:

• The coefficient of Elevation is significant and negative(-7.987e-04). Elevation value indicates that the expected decrease in the log-odds of being defaulted for a unit increase in the Elevation value holding all other predictors constant is e^-7.987e-04
• The coefficient of HD_fire is also significant and positive. HD_fire value indicates that the expected increase in the log-odds of being defaulted for a unit increase in the HD_fire value holding all other predictors constant is e^1.659e+00
#NULL DEVIANCE is deviance for the empty model (8598.5 on 10583 degrees of freedom)
#RESIUDAL DEVIANCE is a deviance based on actual basic model (4962.1 on 10533 degrees of freedom)



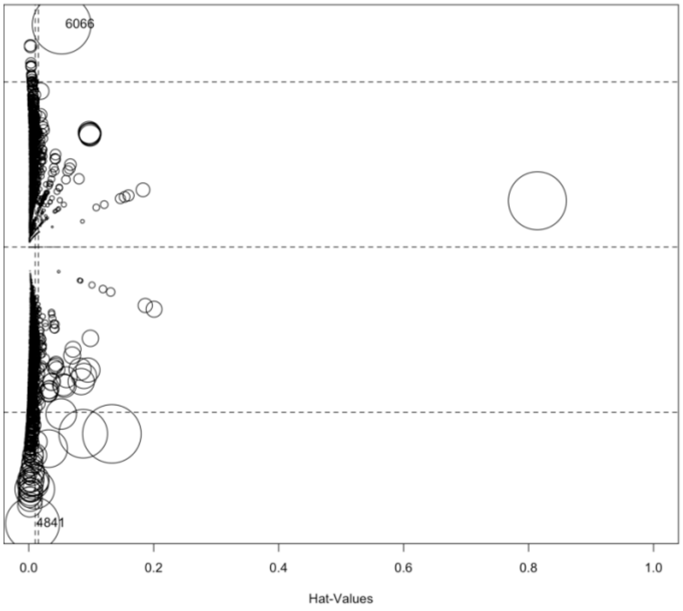


Fig 5.1 Residuals Plot

Fig 5.2 Influence Plot

• The Residuals vs Fitted plot hard to interpret for logistic regression. The Normal Q-Q plot of the residuals are not normally distributed. The Scale-Location indicates that the model is heteroscedastic in nature. The Residuals vs Leverage having few outliers

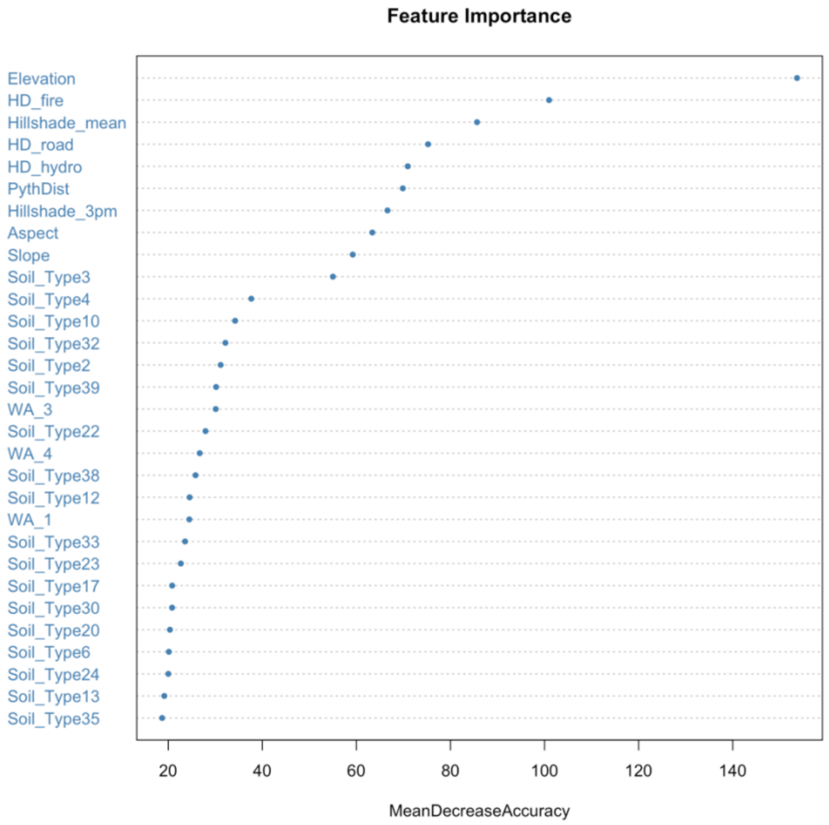
• From Influence Plot, we can see that 6066, 8165 are the outliers and size of the data points related to the cook's distance

Random Forest:

Random forests are an ensemble method which is based on decision trees. It is implemented by a number of decision trees during training time and outputting the mode or mean prediction of the trees. Random forest can reduce the overfitting problem of decision trees at a large level. Random Forest can lower variance without much of an increase in bias

We used tuneRF to find Optimal value (with respect to Out-of-Bag error estimate) for Number of variables available for splitting at each tree node which is called mtry and Number of trees to grow Which is called ntree. From the Elevation clearly shows up as the most significant variable both in terms of mean decrease in Accuracy as well as contributing to reduction in node impurity.

The Variable Importance plot indicates the importance of the predictors in the model. From this plot we see that the Elevation, HD_fire, HD_hydro and new feature variable Hillshade_mean explains maximum variance in the data and the accuracy for our model is 0.8512 for the validation dataset.



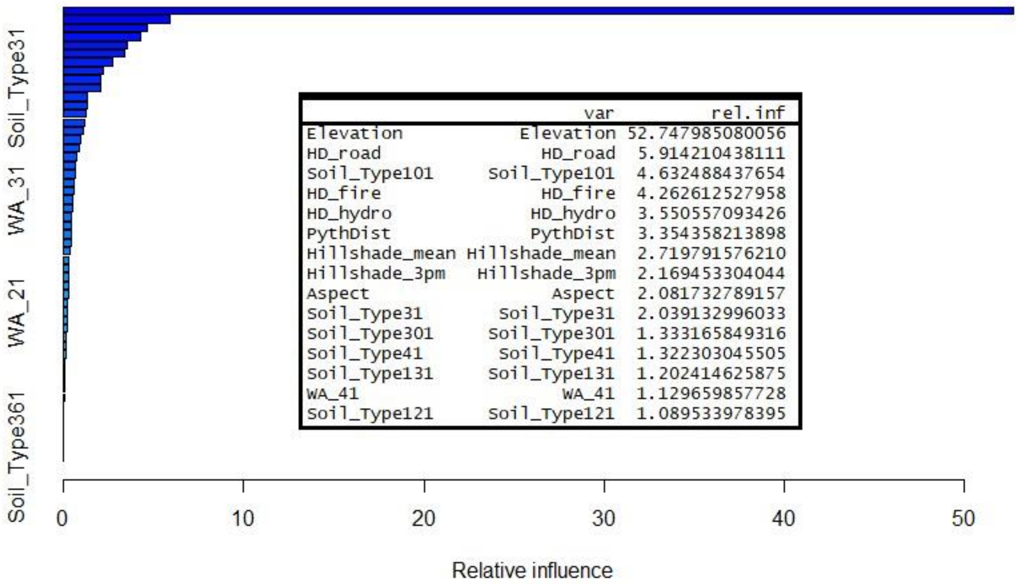
Plot 6 Variable Importance Plot

Gradient Boosting Method:

Gradient Boosting Method is a machine learning technique for classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The purpose of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers. It inherits all the good features of trees (Variable selection, mixed predictors, missing data) and improves on the weak features such as prediction performance.

Variable Importance plot indicates the importance of the predictors in the model. From this plot we see that the Elevation, HD_road, Soil_type101, HD_fire, HD_hydro, PythDist explains maximum variance in the data and the accuracy for our model is 0.8004 for the validation dataset.



```
                        var         rel.inf
Elevation         Elevation 52.747985080056
HD_road             HD_road  5.914210438111
Soil_Type101   Soil_Type101  4.632488437654
HD_fire             HD_fire  4.262612527958
HD_hydro           HD_hydro  3.550557093426
PythDist           PythDist  3.354358213898
Hillshade_mean Hillshade_mean  2.719791576210
Hillshade_3pm   Hillshade_3pm  2.169453304044
Aspect               Aspect  2.081732789157
Soil_Type31     Soil_Type31  2.039132996033
Soil_Type301   Soil_Type301  1.333165849316
Soil_Type41     Soil_Type41  1.322303045505
Soil_Type131   Soil_Type131  1.202414625875
WA_41                 WA_41  1.129659857728
Soil_Type121   Soil_Type121  1.089533978395
```

Plot 7 Variable Importance plot for GBM Model

Support Vector Machine:

Support Vector Machine is a supervised machine learning algorithm used for both classification or regression challenges. Here we plot each data item as a point in n-dimensional space, n being the number of features we have with the value of each feature being the value of a coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes. When the number of dimensions in the data increases, we use kernels which project the lower dimensional data in to a higher dimension where we can separate the data easily using hyperplanes.

The reason we selected SVM was it works very well in high dimensional data and our data is high dimensional. It also uses a subset of values for decision function, so it is much efficient in dealing with memory. Our data set is not too large, so it will be not being taking much time for us to run svm which usually takes very high time for huge data sets. With support vector machine we achieved 82% accuracy.

On this Fig 8, we can see that the darker the region is the better our model is we can narrow down our search for cost and gamma values to and try further tuning if required.

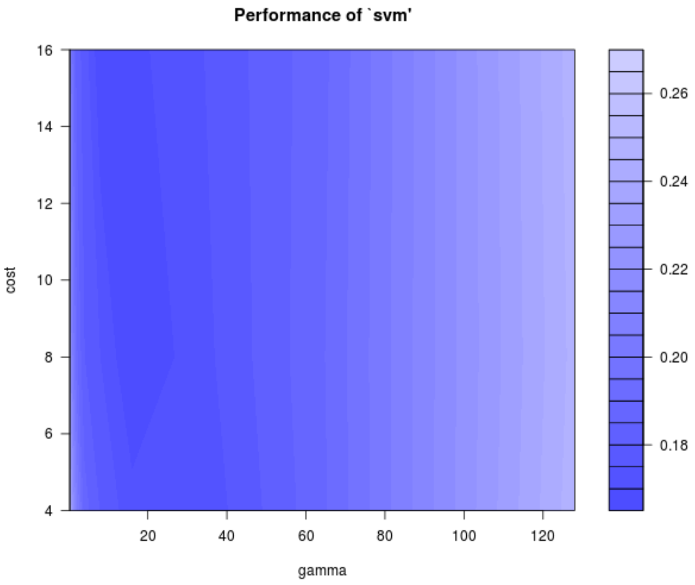


Fig 8 Performance of SVM

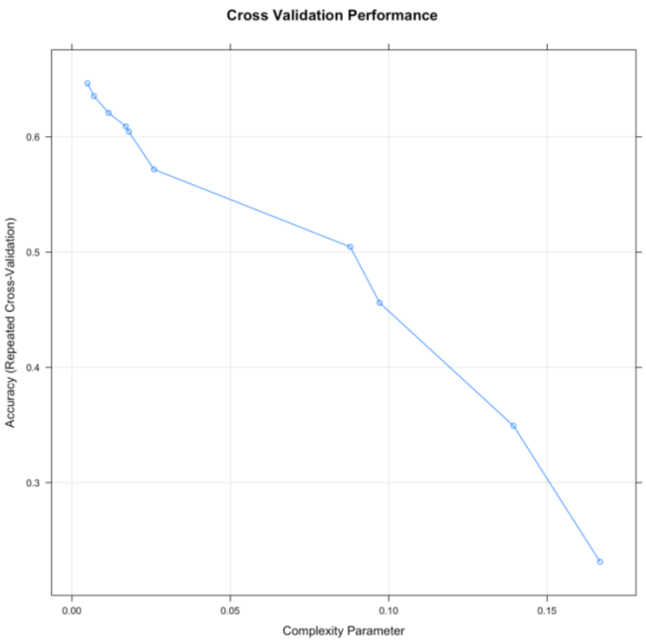


Fig 9 Cross Validation Performance for Decision Tree Model

Decision Tree:

The essential thought associated with choice trees is to separate a complex decision into a union of several simpler decisions. Furthermore, to boost, the training events which were misclassified have their weights increased, and another tree is formed. This technique is then repeated for the new tree. this way, many trees are developed. Information Gain is a measure that quantifies the change in the entropy before and after the split. It's an elegantly simple measure to decide the relevance of an attribute. Gini Index is more suitable to continuous attributes and entropy in case of discrete data. Also, it works well for minimizing misclassifications.

From the Fig 9, we can clearly observe that, repeated cross validation for the best model at complexity parameter is 0.0047

K-Means

we have considered 7 predictor continuous variables to draw bi-variate cluster plot. Bi-variate cluster plot in K-means uses PCA to draw the data. It uses the first two principal components to explain the data. In the below fig 10.1 first two components explain 55.9% of variance.

Hierarchical methods use a distance matrix as an input for the clustering algorithm. This metric will influence the clusters shape, as some elements may be close to one another according to one distance and farther away according to another. We use the Euclidean distance as an input for the clustering algorithm. Ward's minimum variance criterion minimizes the total within-cluster variance

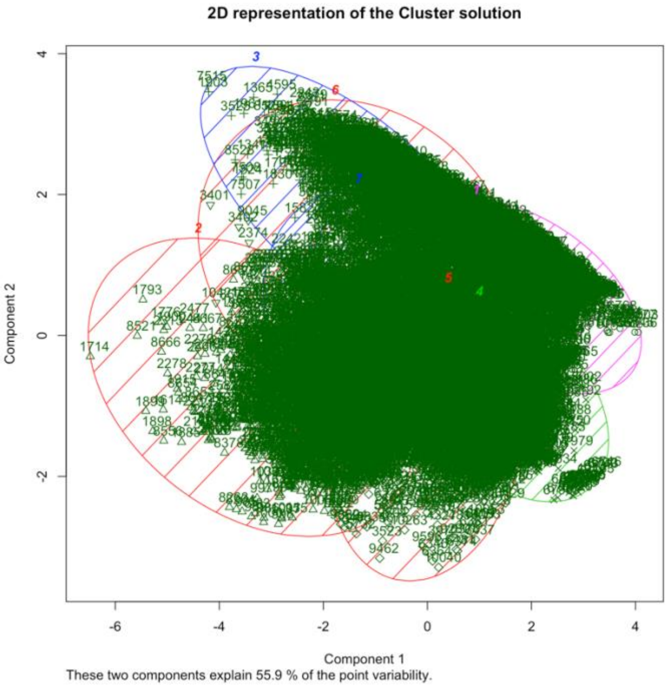


Fig 10.1 2D Representation of 7 clusters

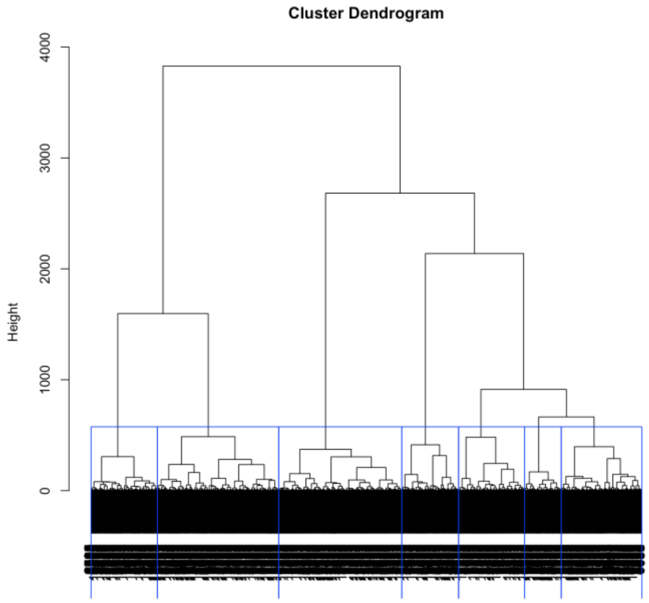


Fig 10.2 Cluster Dendrogram

Metric Evaluation for Machine Learning Models:

						CV Performance			
Method	Package	Parameter	Range of Tuning	Selection	CV Method	Accuracy	Kappa	AUC	F1 Score
Decision Tree	Caret Rpart	Complexity parameter	CP at Information Gain	0.0047	5 Repeats 10-Fold CV	0.7152	0.6852	0.7513	0.7082
Decision Tree	Caret Rpart	Complexity parameter	CP at Gini Index	0.0041	5 Repeats 10-Fold CV	0.7341	0.6981	0.7565	0.7254
SVM	e1071	Cost gamma	4:16 0:120	10 18	10-Fold CV	0.8218	0.7882	0.8671	0.7940
GBM	Caret	ntree	c(500,750,1000,1500,2000)	1500	5 Repeats 10-Fold CV	0.8004	0.7761	0.8711	0.8142
Random Forest	RF	ntree mtry	100 to 1600 10 to 60	500 20	5 Repeats 10-Fold CV	0.8512	0.8264	0.9080	0.8523

Conclusion

The forest cover dataset is an extremely rich dataset for anyone who wishes to implement multi-class classification algorithms across a wide range of methods. In this project, we present an accurate approach to predict the forest cover type with limited geological information. Through feature engineering, we extract new features and transform existing data into more model-sensitive features. Then, we use different methods to train model and predict the cover types. The goal of predict cover of forest based on cartographical data, and obtain good predictions. Overall, best results are obtained using Random forest, which gave 85% accuracy on test set.

References:

[1]. W. N. Venables, D. M. Smith *an Introduction to R.*

[2] Bache, K. Lichman, *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science (2013).

[3] Max Kuhn, Kjell Johnson. *Applied Predictive Modelling*

[4] G. James, D. Witten, T.Hastie, R. Tibshirani, *An introduction to statistical learning.*

[5] T.Hastie,R. Tibshirani, J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition*

[6] https://archive.ics.uci.edu/ml/datasets/covertype

[7] https://www.kaggle.com/c/forest-cover-type-prediction

Appendix:

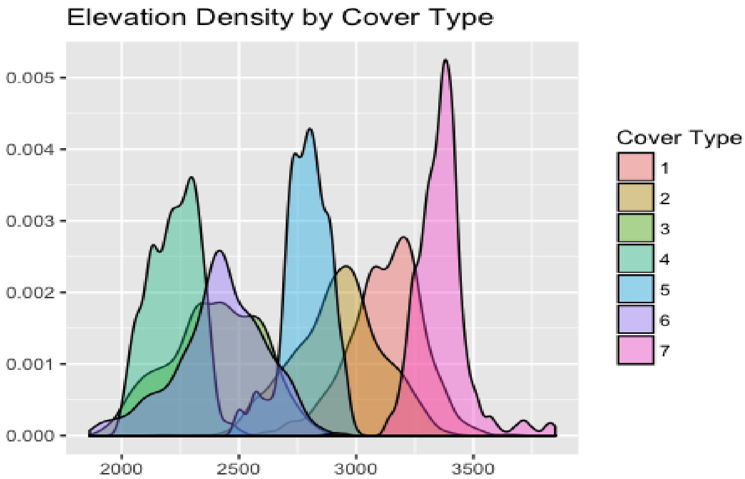


Fig 13 Elevation Density by Cover Type

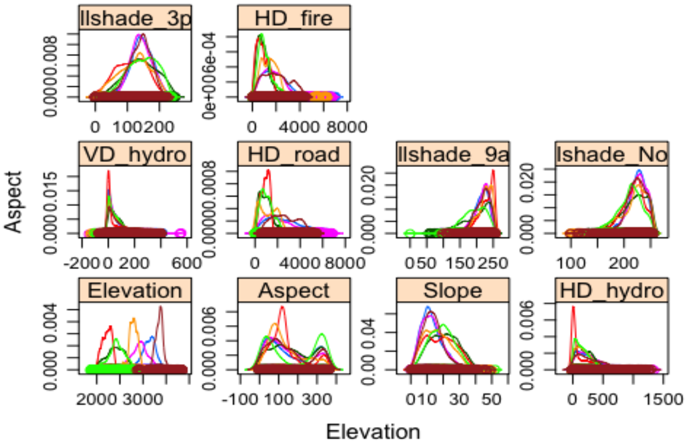


Fig 14 density of numeric features by covertype

Two best model's precisions, recall and F1 scores:

Random Forest for each class type:

	pred.class	presi	rec	F.score
1	1	0.768608414239482	0.733024691358025	0.750394944707741
2	2	0.763668430335097	0.66820987654321	0.71275720164609
3	3	0.831269349845201	0.829984544049459	0.830626450116009
4	4	0.930408472012103	0.950540958268934	0.940366972477064
5	5	0.879310344827586	0.944444444444444	0.910714285714286
6	6	0.828614008941878	0.860681114551084	0.844343204252088
7	7	0.922734026745914	0.958333333333333	0.940196820590462

GBM for each class type:

	pred.class	presi	rec	F.score
1	1	0.718354430379747	0.700617283950617	0.709375

2	2	0.688356164383562	0.621329211746522	0.653127538586515
3	3	0.726172465960666	0.744186046511628	0.735068912710567
4	4	0.926315789473684	0.950617283950617	0.938309215536938
5	5	0.852941176470588	0.895061728395062	0.873493975903614
6	6	0.750769230769231	0.755417956656347	0.753086419753086
7	7	0.9209726443769	0.935185185185185	0.92802450229709