Forest Cover Type Prediction

-- A Classification Problem

Group Members:
Abhishek Agrawal
Nisarg Gandhi
Rohit Arora
Tyler Stocksdale

The Problem Statement

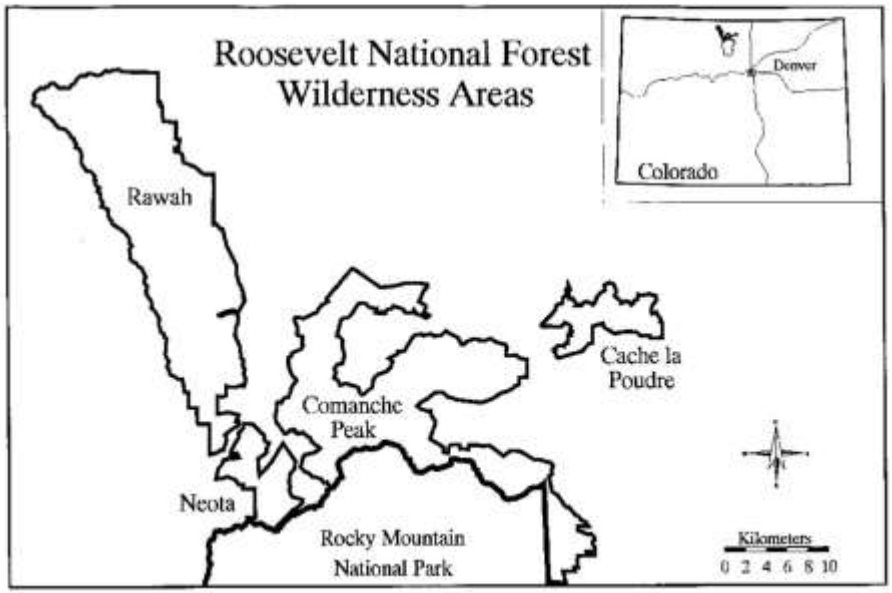
Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

7 Cover Types

The Problem Statement

Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

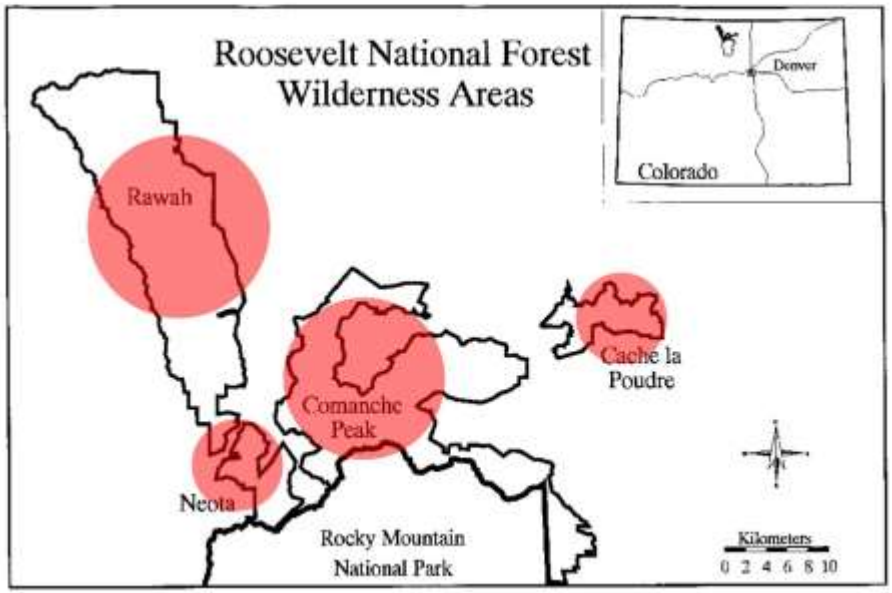
7 Cover Types



The Problem Statement

Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

7 Cover Types



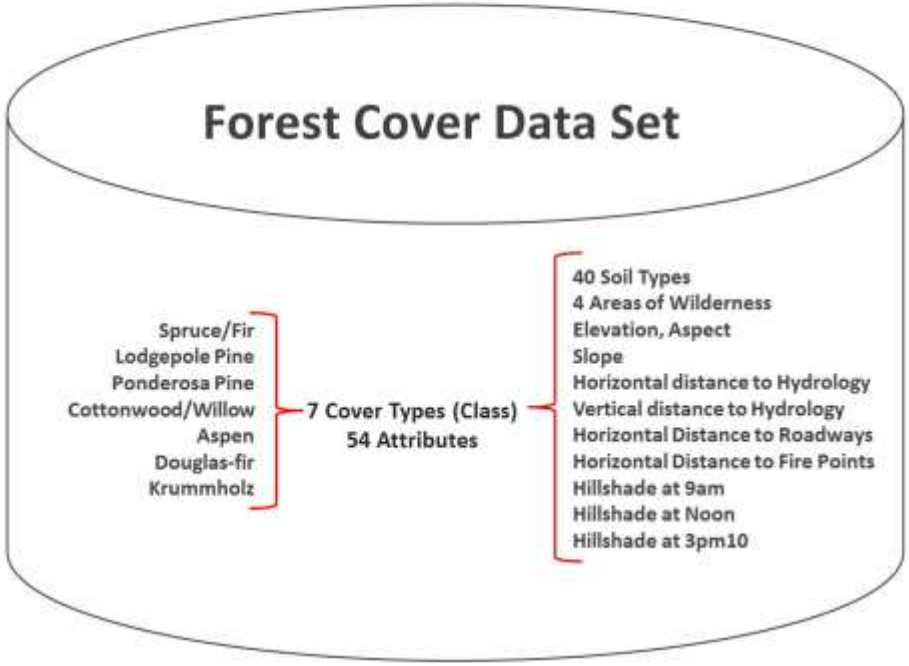
The Data Set

Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

7 Cover Types (Class)
54 Attributes

40 Soil Types
4 Areas of Wilderness
Elevation, Aspect
Slope
Horizontal distance to Hydrology
Vertical distance to Hydrology
Horizontal Distance to Roadways
Horizontal Distance to Fire Points
Hillshade at 9am
Hillshade at Noon
Hillshade at 3pm10

The Data Set

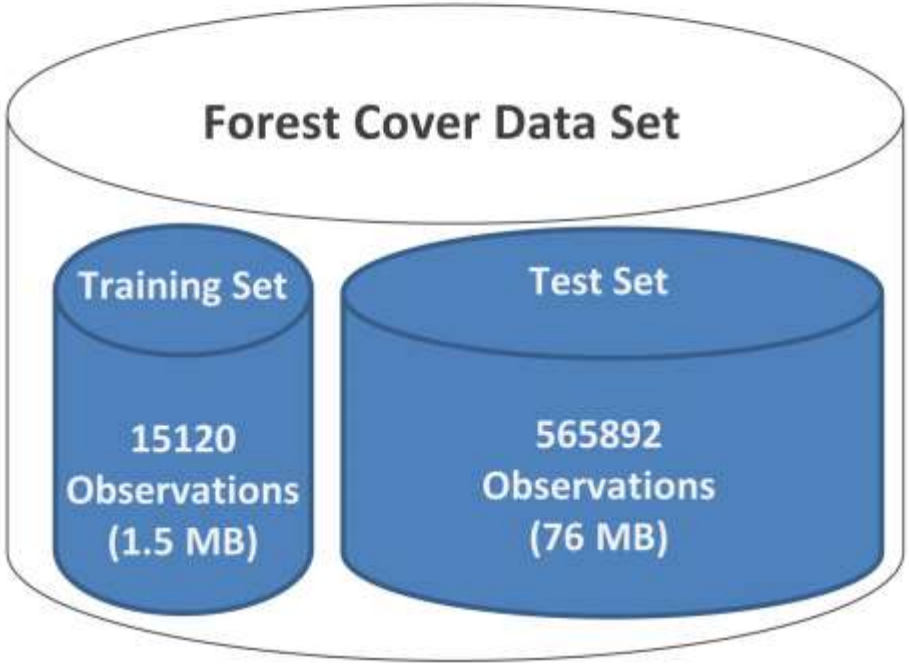


kaggle

UCI Machine Learning Repository

USGS science for a changing world

The Data Set

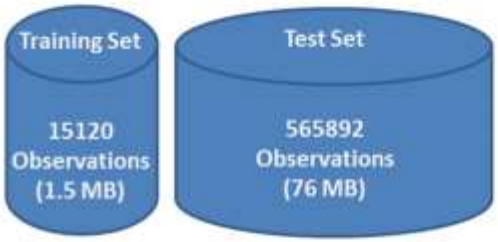


kaggle

UCI Machine Learning Repository

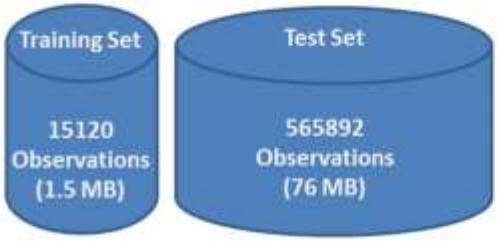
USGS science for a changing world

Constraints



Big Difference in Training and Test Data Size

Constraints



Big Difference in Training and Test Data Size

40 Soil Types
4 Areas of Wilderness
Elevation, Aspect
Slope
Horizontal distance to Hydrology
Vertical distance to Hydrology
Horizontal Distance to Roadways
Horizontal Distance to Fire Points
Hillshade at 9am
Hillshade at Noon
Hillshade at 3pm10

Different Attribute Types

Constraints

Big Difference in Training and Test Data Size

Different Attribute Types

Missing Remote Sensed Data

Related Work

Blackard, Jock A. and Denis J. Dean, 1999

	Prediction Accuracy %
Feed-Forward Artificial Neural Network (ANN architecture: 54-120-7, Learning Rate: 0.05, Momentum Rate: 0.5 and Learning Algorithm: Backpropagation)	**70.58%**
Linear Discriminant Analysis	**58.38%**

Preprocessing

Reduction and Transformation

	1	2	3	...
1	0	0	0	1
2	0	0	0	0
3	1	0	0	0
...	0	1	0	0

	Soil Type	Area Of Wilderness	Cover Type
1	S40	W2	C5
2	S26	W2	C4
3	S1	W4	C7
...	S2	W1	C1

Preprocessing

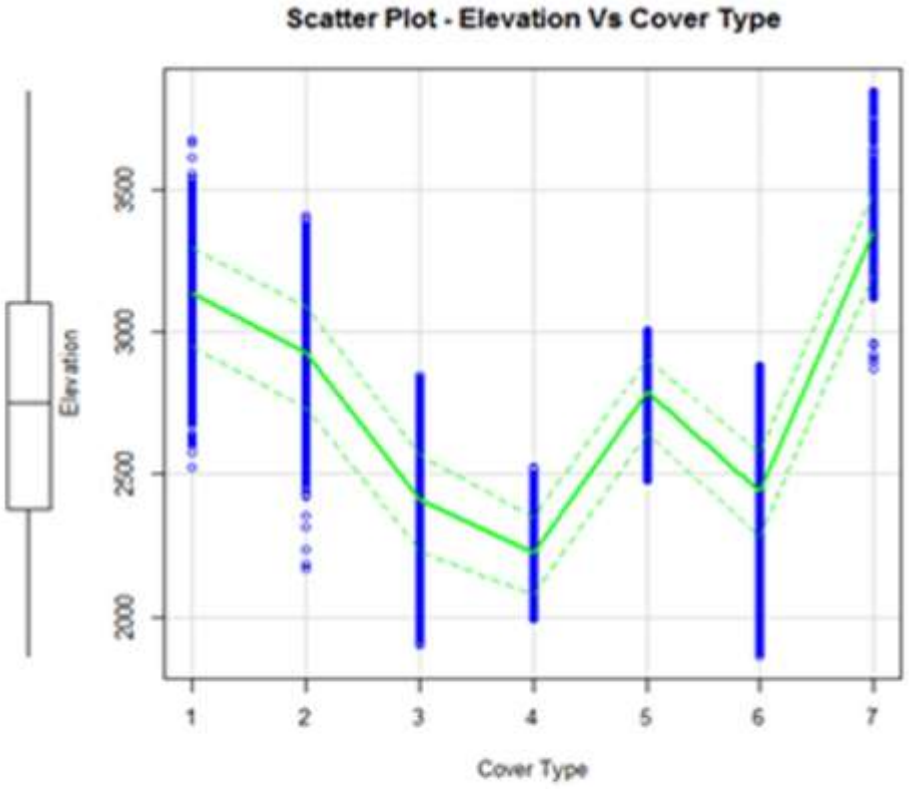
Standard Normalization:
$x' = (x - \mu)/\sigma$

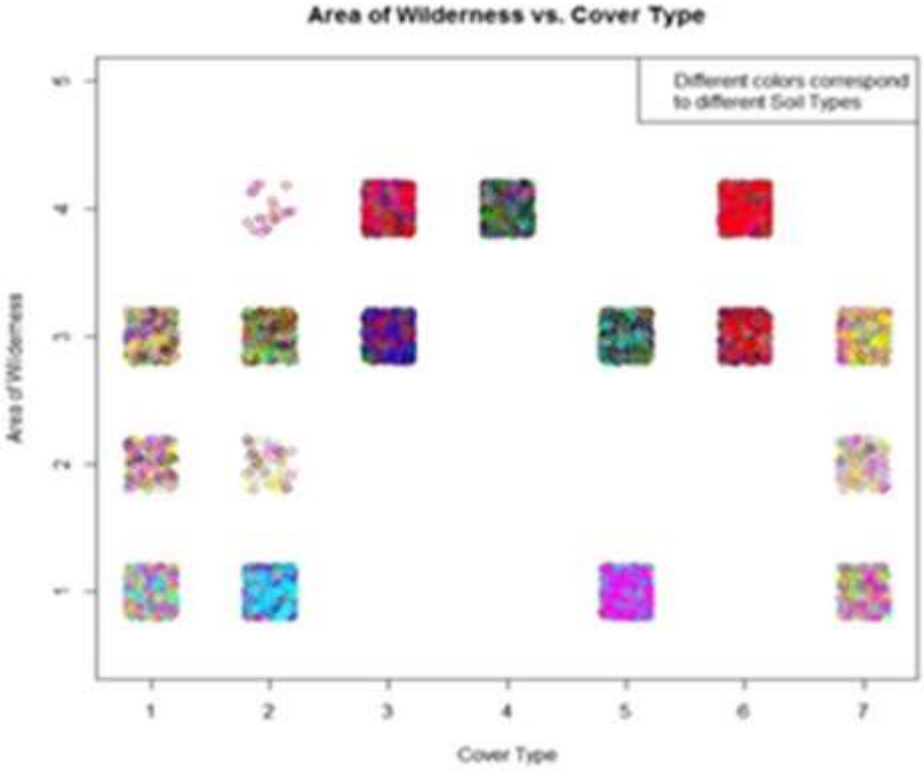
First attempt at normalizing data

Min/Max Normalization:
$x' = (x - min)/(max - min)$

Better normalization due to differences in test and training data set

Exploratory Data Analysis

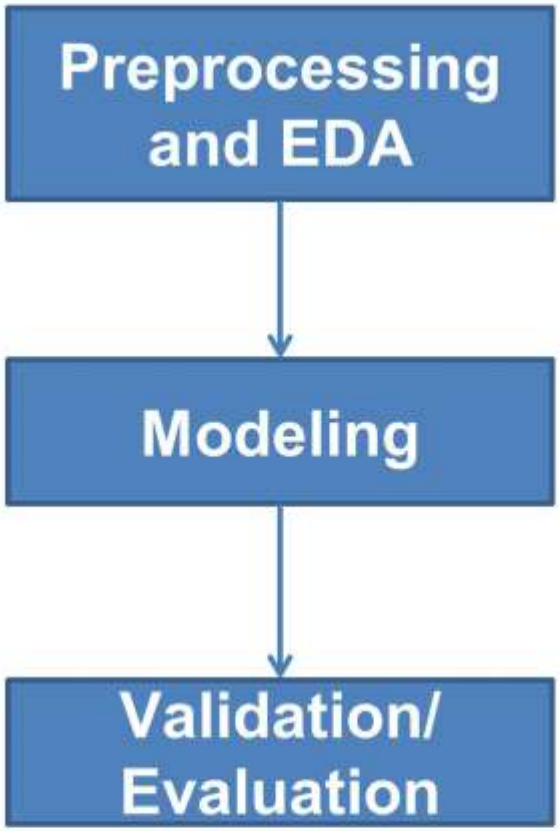




Classifiers

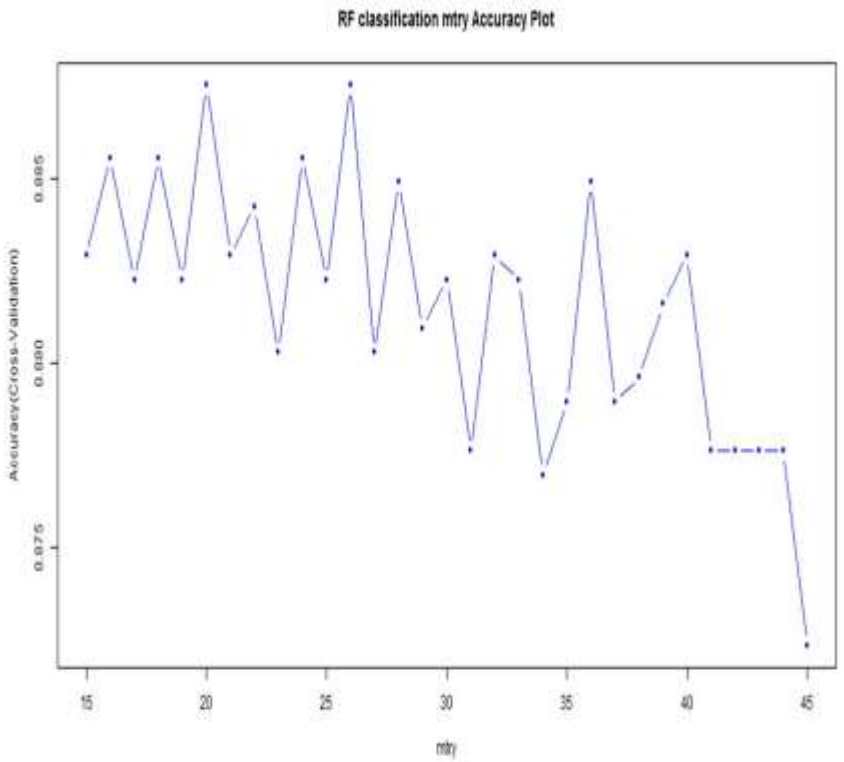
Classifiers we explored:

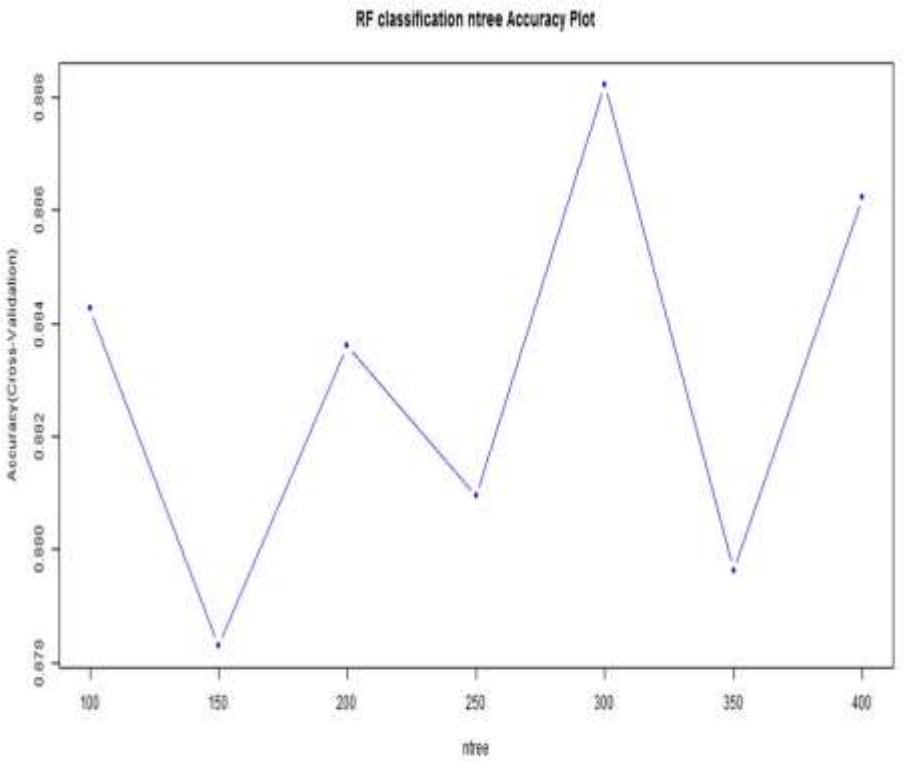
1. Decision Tree
2. SVM
3. k-NN
4. Random Forests
5. Gradient Boost Model
6. Naive Bayesian
7. Rule Induction



Classifiers

Random Forest Parameter Tuning

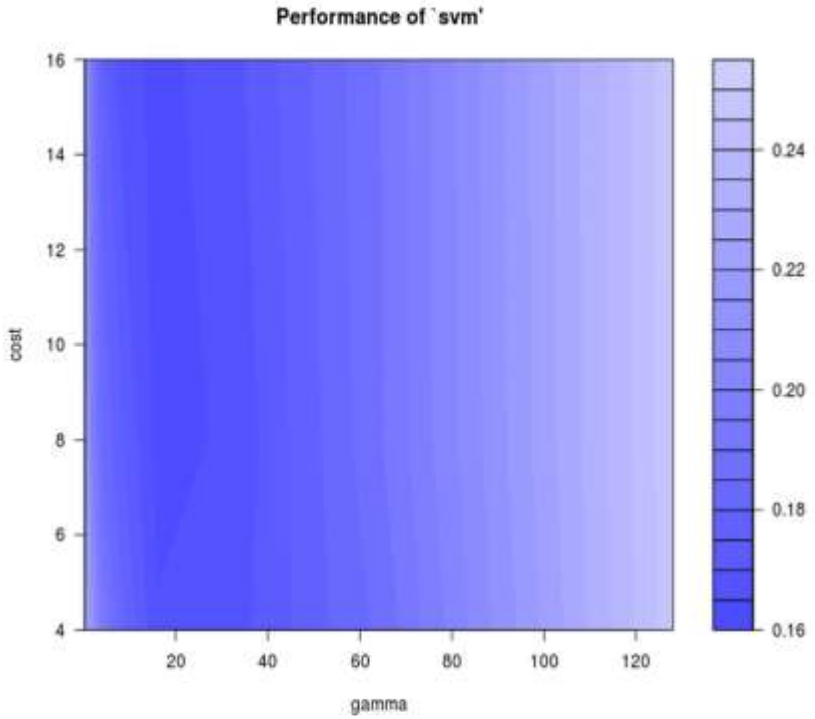




Best Parameter for Random Forest, mtry = 20, ntree = 300

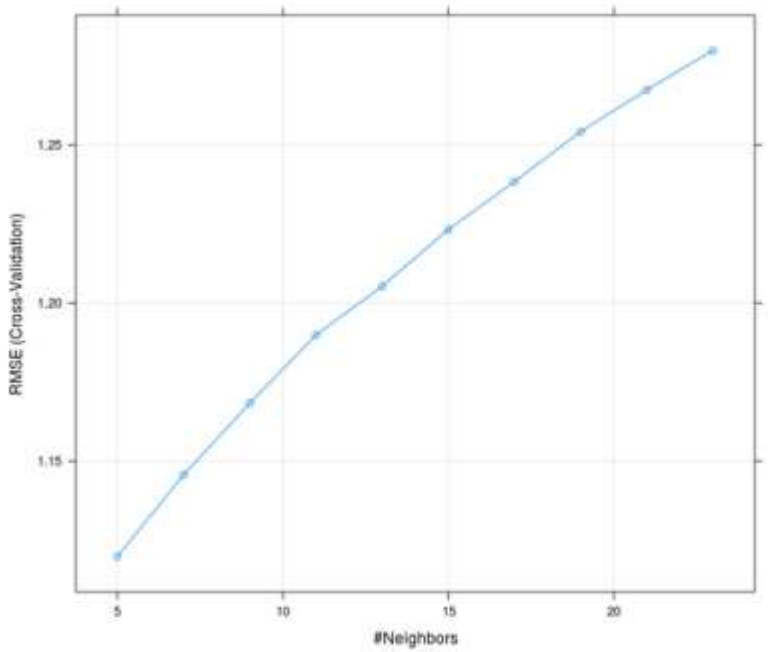
Classifiers

SVM Parameter Tuning



Best Parameter for RBF Kernel, Gamma = 10, Cost = 8

KNN Parameter Tuning

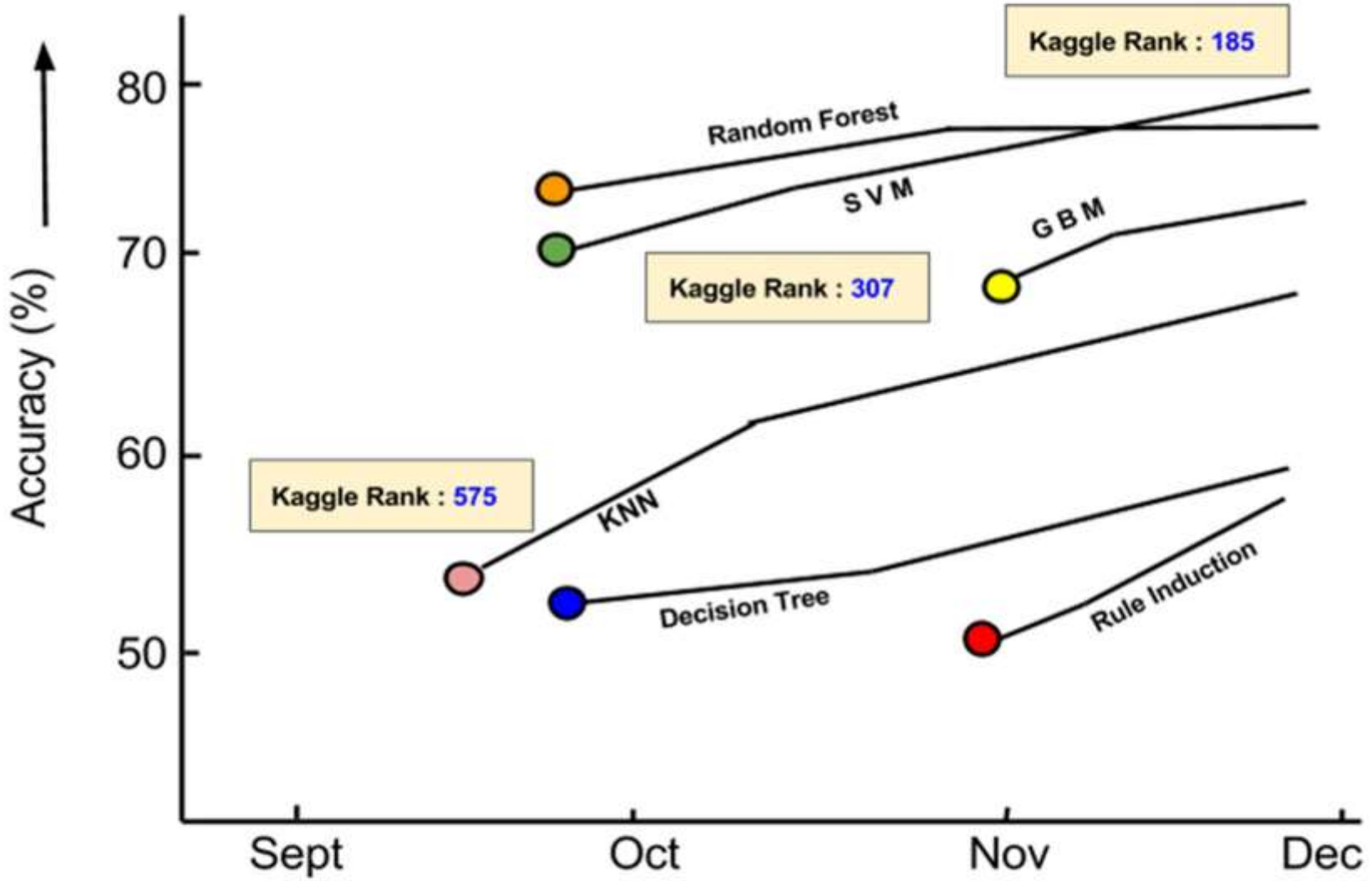


Best Parameter for KNN, K = 5

Results

Classifier	Tool	Accuracy	Top Recall	Top Precision	Test Data Accuracy(Kaggle)
KNN	RapidMiner, R	83.31%	4 > 7 > 5 > 6	7 > 3 > 4 > 5	69.67%
SVM	R, RapidMiner	84.80%	4 > 7 > 5 >6	7 > 4 > 5 >3	76.69%
Decision Tree	R, Weka	78.28%	4 > 7 > 5 > 6	7 > 4 > 5 >3	58.89%
Naive Bayesian	R	66.20%	7 > 4 > 5 > 1	7 > 4 > 5 >3	Not Submitted
Random Forest	R	88.23%	4 > 7 > 5 > 3	7 > 4 > 5 > 6	75.60%
Gradient Boost Model	R	87.12%	4 > 7 > 6 > 5	7 > 3 > 4 > 5	69.82%
Rule Induction	R	76.41%	4 > 7 > 5 > 6	4 > 7 > 5 > 6	58.32%

Timeline


Kaggle Rank : 185
Random Forest
SVM
GBM
Kaggle Rank : 307
Kaggle Rank : 575
KNN
Decision Tree
Rule Induction
Accuracy (%)
80
70
60
50
Sept
Oct
Nov
Dec

Future Work

- **Semi Supervised Learning** Methods to increase the training data size
- Using **two-way classification** approaches to distinguish between majority class groups with minority
- **Feature Engineering** using Principal Component Analysis
- Apply of **advanced classifiers** and boosting methods.

References

[1] https://archive.ics.uci.edu/ml/datasets/Covertype

[2] Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables (2000) by J. A. Blackard and D. J. Dean. In: Computers and Electronics in Agriculture 24(3), pp. 131-151.

[3] http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf "A Practical Guide to Support Vector Classification".

[4] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm "Random Forests Leo Breiman and Adele Cutler", Random Forests. Web. 16 Nov. 2014.

[5] http://vimeo.com/71992876 "Using GBM for Classification in R".

[6] https://en.wikipedia.org/wiki/Random_forest "Random Forest" Wikipedia. Wikimedia Foundation, 14 Nov. 2014. Web. 16 Nov. 2014.