Sebastian Zdarowski
DSC 540 Advance Machine Learning
June 12, 2019

Predicting Forest Cover types: Getting Lost in the Forest

Abstract: This paper is a random forest approach for multi-classification of the Covertype dataset. It will consist of preprocessing the dataset, feature section, hyperparameter tuning of the random forest algorithm. The intended purpose of the random forest prediction will be to see if there are any important attributes in predicting forest Covertypes within Roosevelt National Park. Further, I hope to provide insight to what factors contribute to tree growth in specific park locations. The random forest results prove that its able to distinguish between the multiple classes present within the dataset with 94% accuracy. The features selected within the dataset are able to predict the class labels with precision but were not able to provide any meaningful insight into what constitutes tree growth in certain areas of the National Park.

Introduction: As of 2010 the United States had 751,255,000 acres of forested lands, a number that represents 1/3 of the country. *"Between 1990 and 2010 the United states has lost an average of 949,750 acres of forest each year. The United states deforestation rate has been more than offset by the reforestation rate between 1990 and 2010, the country added 18,995,00 acres of forest land during that period. (Becker, 2016)* The trend is driven by organizations such as the US Forest Service and the Arbor Day Foundation. The goal of this study is to be able to accurately predict various forest types in Roosevelt National Forest park using machine learning. I will be focusing on the effectiveness of using the random forest algorithm to predict various classes of trees. By being able to do this, we may provide some insight into what causes trees to grow within a certain area. The dataset that I will be using is the Covertype Data set from the UCI Machine learning Repository.

Dataset: The dataset itself contains 581,012 instances and contains 54 attributes. The data was collected by observations from the US Forest Service Department. It contains 10 continuous variables Elevation, Aspect, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_roadways, Hillshade_9am, Hillsahde_Noon, Hillshade 3pm, Horizontal_distance_to_fire_points. Soil Type is broken out into 40 dummy variables based on the soil type around the area, Wilderness area is broken out into 4 types based on the location the tree is in the park. Our variable we will be predicting is Cover Type which is the 7 different tree species: Spruce/Fir, Lodgepole Pine, Pondersoa Pine, Cottonwood/Willow, Aspen, Douglas-Fir and Krummholz. There is a class imbalance present within the dataset as shown in the chart to the right. Lodgepole Pine and Spruce vastly outnumber other tree species. Wilderness area 1 and 3 contained the most samples within this dataset. (See Figure 1)

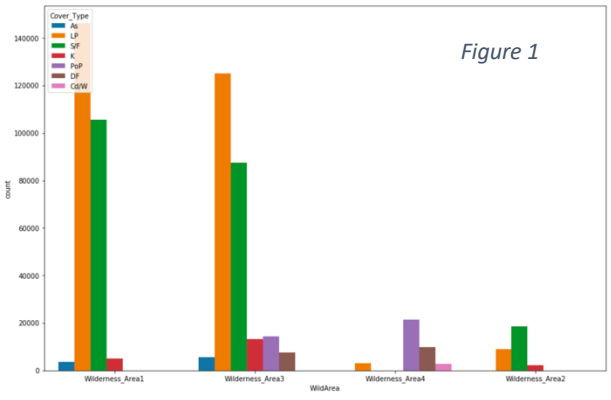


Figure 1

Lit review: *An improved random forest classifier for multi-class classification by Archana Chaudhary, Savita Kolhe, Raj Kamal.*

Within this research paper the authors present an improved random forest classifier approach for a multi class disease classification problem. The authors argue that their combined use of attribute evaluator method and instance filter method can boost the performance of a random forest algorithm. The dataset that the authors used was real-life groundant disease, which takes into account of different disease and 26 common factors attributes that contribute to plant based diseases. Authors used 3 different feature selection methods, correlation – based feature selection, symmetrical uncertainty, and gain ratio. Another tool that the authors had to utilize was simple random sampling – instance filter, since real world datasets such as the one in question, have a non-

uniform class distribution. The non-uniform class distribution considerably influences the performance of a classification algorithm in the training phase. The model was first trained without sampling and the other method applied was using simple random sampling that gives a fair sample from the original data. In addition to using random forest, the authors experimented with SVM, neural networks, and linear regression. Random forest still showed the best results out of all these methods. The improved RFC approach used a random forest algorithm and an attribute evaluator method (feature selection) and instance filter method(random sampling) The authors used 10 fold cross validation on their testing sets to see the improved performance. F measure was used, accuracy, specificity and ROC to evaluate each model. I will be using the F score as well in my model, due to AUC not being applicable to a multi class problem. After evaluation the authors make the claim that their improved RFC approach was able to obtain better results. Their approach was able to obtain 97.80% accuracy compared to the standard random forest accuracy of only 80.20%. They conclude that this approved would be a good substitute in dealing with computer aided diagnosis and multi classification problems. The random forest model that will be built for the covertypes dataset, will be a facing a multi class problem as well, this paper was an inspiration to show that multi-classification can be done using random forest models and to be able to predict with a high degree of certainty.

To Tune or Not to Tune the Number of Trees in Random Forest By Philipp Probst, Anne_Laure Boulesetix.

This research paper is about selecting the optimal amount of trees in a random forest. The authors go through multiple examples of using binary classification, regression classification and multi classification. For the purpose of this paper, we will focus on their research on the multiclass optimal tree count. The overall goal of this stud was to determine the magnitude of the frequency of non-monotonous patterns of error rate in read data settings. The authors focused on the OOB (Out of bag error) rate, AUC, logarithmic loss and number of trees in the model. For each dataset that they used in the study, they started out with 2000 trees as the default setting. They chose this because in the preliminary study on the subset of datasets, they observed that the convergence of the OOB curves was reached within 2000 trees. Within the study they noticed that the biggest performance gain the in OOB curves can be observed with the first 250 trees. Setting number of trees from 10 to 250 provided an average decrease of .0306 error rate and an increase of .0521 on the AUC, for multiclass the improvement is bigger with an average improvement error rate of .0739 OOB, AUC .0057 for using 2000 trees compared to 250 trees. But the authors do make a point that the improvement is minimal and up to the users interpretation and what they expect to get from the results. Through their studies they have concluded that tuning is not recommendable in the case of classification, since non monotonous patterns are observed only with some performance measures such as error rate and AUC, for regression increasing the number of trees to an optimal point proves to be beneficial in reducing the error rate, but if you include too many trees, the performance may get worse. This paper inspired me to take a look at the number of decision trees that would end up within the final model, the goal will be to reach an optimal amount of trees to provide a good balance between prediction and runtimes.

Robust feature Selection Using ensemble Feature Selection Techniques by Yvan Saeys, Thomas Abeel, and Yves Van de Peer

This research paper talks about the robust feature selection using ensemble (voting) feature selection methods to improve techniques on classification performance. Feature selection is an important step in preprocessing for machine learning applications. It is used to maximize the model performance and gives the ability to build simpler and faster models only using a subset of features. There a 3 feature selection approaches the author claims, filter methods which directly operate on the dataset, wrapper methods that perform a search in the space of feature subsets which are guided by the outcome of the model, they report better results that filter methods but are computational costly and embedded methods which uses internal information of the classification model. Within ensemble feature selection "*a collection of single classification or regression models is trained, and the output of the ensemble is obtained by aggregating the out- puts of the single models, e.g. by majority voting in the case of classification, or averaging in the case of regression". (Saeys, pg 316)* The authors make the claim that ensemble feature selection can be used to improve the robustness of feature selection techniques, due to ensemble voting, it can reduce the risk of choosing an unstable subset of features. 6 datasets were used in the experiment in total, which combined all feature selection methods and used 10 fold

cross validation to access the accuracy as the performance measure. The best feature selection method used was SVM followed by random forest ensemble method. Authors state trade off using these 4 methods depends on the dataset on hand and using a robust feature selection method will gain importance in the future, due the vast information it can provide in optimizing model performance. For this paper, I will be doing 3 different feature selection techniques, I will be creating a wrapper based feature selection, low variance filter, and using stepwise feature selection for the random forest. I will compare all results from the 3 different feature selection methods and evaluate which would be best for the model performance.

Random Forests for land cover classification by Pall Oskar Gislason, Jon Atli Benediktsson , Johannes R. Sveinsson

Within this paper, the authors talk about using random forests for land cover classification. The most widely used ensemble methods are bagging and boosting. The random forest classifier uses bagging or bootstrapping aggregating to form an ensemble of classification. The authors will be comparing the accuracy of the random forest classifier to other ensemble methods on a multisource remote sensing and geographic data. In ensemble classification serval classifiers are trained and their results are combined through a voting process to output the best results. The most popular are bagging and boosting ensemble methods. Boosting is more computationally demand than bagging, but offers better predictive results. *"Random forest is a general term for ensemble methods using tree-type classifers*" *(Oskar pg. 295)* In the training period RF creates multiple trees, which each being trained on a bootstrapped sample of the training data. For classification the random forest casts a vote on the most popular class, the output of the classifier is determined by the majority of tree votes. The authors used four different datasets to test the accuracy of the random forest, we will be focusing on the first dataset that they used which was a multi class problem. The dataset was labeled Landsat Multispectral Scanner (MSS). The dataset contained 10 classes, 1 being water the rest being forest covertypes. Similarly to what I am working with the forest cover type dataset. Two similar variables that they have in their dataset to predict trees, is evaluation and slope, which are two variables that I have in my dataset. The model that was built used 500 trees and split on 3 split variables. The random forest algorithm had no issues in predicting the class water, but had challenges when dealing with classes 2-9 which are all forest cover types. Processing time were pretty fast for the random forest algorithm. The authors used 10 fold cross validation on the dataset for feature selection. Authors ran between 500 to 10,000 trees, and noticed that 500 trees were the most optimal model in the random forest classifier. Random forest classifier performed well and out performed in terms of processing time against bagging and boosting ensemble methods. The main benefit the author argues about random forest is that it's not susceptible to overfitting and doesn't require much guidance although its parameters can always be tweaked.

Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran by Omid Rahmati a, Hamid Reza Pourghasemi b, Assefa M. Melesse c

In this article the authors will be studying the application of a random forest and maximum entropy models for ground water potential mapping in Iran. Water is one of the most valuable natural resources to communities, before machine learning ground water is discovered by drilling and the methods are usually time consuming with limited results. The authors propose using random forests and maximum entropy using remote sensing data to potentially map groundwater within a region. RF and ME models are new to the area of ground water potential mapping compared to other methods, the reason why is because there hasn't been any studies to prove these methods. This article intends to do just that. The OOB error was used to find the optimal amount of trees. The authors took a look at the feature importance of the dataset, and were able to find the most influencing conditioning factors for groundwater, which were altitude, drainage density, lithology and land use. After setting up all the parameters of the models, it seemed towards the end that the maximum entropy model outperformed the random forest model AUC - 87.7% compared to 83.1%. The authors make the claim that these models are much more efficient than using neural networks, and can provide useful information for interpretation of the GIS dataset. One of the drawbacks would be the time related to set up time for analysis of the random forest and ME models. These results can be useful for comprehensive evaluation of groundwater exploration development. This paper provided inspiration on using the random forest algorithm within this

dataset, since random forests are relatively new to the area of geographic data, The use of random forests within the GIS space is growing in popularity with data scientists. *"The RF and ME models are new in the area of ground-water potential mapping compared to other methods"(Rahmati, pg 361).* It will be interesting to see how the algorithm performs within this dataset using the cartographic variables provided.

Preprocessing: Before anything is done regarding prediction, the dataset has to be explored for any errors. The dataset itself had no null values, so nothing needed to be changed. Normalization was attempted, but with model building it didn't improve overall accuracy so it was not used. Correlation amongst variables was also looked at. The only variables that were correlated with hillshade_3PM, Hillshade_Noon and Hillshade_9am which makes sense since they are dependent on the sun shining within the area. The paper regarding improved forest classifiers talked about class imbalance being an issue "Real world datasets, as groundnut disease dataset have non uniform class distributions, this non – uniformity of class distributions considerable influences the performance of a classification algorithm within the training phase". (*Chaudhary*, *pg 216).* ,but within this dataset the random forest model had no issues with it. The data will be split into training 65% and 35% testing after feature selection is completed.

Methods/Modeling building: Random forest are an ensemble learning method used for various purposes, but for this paper we will be studying its effectives with multiclass labels. Another thing I will be looking into is what are the best features for the dataset on hand, can we derive results from the 44 features present within the dataset? Would they be able to tell us anything important in classifying cover types? 10 fold cross validation will also be performed on all training models before settling on a final model.
Tree settings: Optimizing the number of trees in the model will also be a focus point. *"T is one of several important parameters which have to be carefully chosen by the user. Some of these parameters are tuning parameters in the sense that both too high and too low parameter values yield sub-optimal performances"(Probst, pg 1).* I will be testing 10, 50 and 200 trees and evaluating which would be best for the model performance, taking into consideration accuracy, F scores and runtimes.

Random Forest tree #	Acc	F1 score	RF Runtime	10 CV Runtime
10	0.94(+-0)	0.90 (+-.01)	9.13	108.02
50	0.95 (+-0)	0.92 (+-.01)	45.54	518.6139
200	0.95(+-0)	0.92(+-.01)	184.27	2103.42

Looking at the tree performance it's pretty consistent across 10, 50 and 200 trees. There isn't that much of a model performance boost by increasing the number of trees. Acc is staying between .94 -.96, f scores are only going up by .01 not much of a difference. In the 10 fold cross validation that we do for the training set, we do see a large increase in run times 10 trees compared to 200 trees. With all the other metrics being pretty consistent besides runtimes, I will go ahead with building a random forest consisting of 10 trees.
Feature selection: For the feature selection, I will be using 3 different methods. Low variance filter, stepwise feature selection and wrapper select method from random forest. Here we see the features that were selected by the model based on the feature selection.

Low Variance Filter	**Wrapper**	**Stepwise**
Elevation	Elevation	Elevation
Aspect	Aspect	Horizontal_Distance_To_Hydrology
Slope	Slope	Vertical_Distance_To_Hydrology
Horizontal_Distance_To_Hydriology	Horizontal_Distance_To_Hydrology	Horizontal_Distance_To_Roadways
Vertical_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points
Horizontal_Distance_To_Roadways	Horizontal_Distance_To_Roadways	
Hillshade_9am	Hillshade_9am	
Hillshade_Noon	Hillshade_Noon	
Hillshade_3pm	Hillshade_3pm	
Horizontal_Distance_To_Fire_Points	Horizontal_Distance_To_Fire_Points	
	Wilderness_Area1	
	Wilderness_Area4	

Across the 3 methods, Elevation, Horizontal & Vertical distance to Hydrology, Horizontal Distance to roadways & horizontal distance to fire points are the 5 variables that are consistent across the model.

Feature Selection	Acc	F Score
Low Variance Filter	0.93	.89(+-.01)
Wrapper	0.94	.89(+-.01)
Stepwise	0.87	.83(+-.01)

After running these models though, Wrapper selection was the best method in obtaining the highest accuracy and f-score, so I will go ahead with that method in the final model. There isn't much difference across all feature selection method in performance. We see that the top 5 variables that are consistent across all feature methods are the attributes that drive the model prediction. The features that were selected are interesting, because I originally thought that soil types and wilderness area would be more important variables in predicting labels. Distance to hydrology makes sense for tree cover types, since trees require water to flourish, so it would make sense for that to be an important variable.

Findings: The final model was built using 10 trees, using cross entropy as the criteria, min_samples split was = 3, no max depth selected, rand st set to 1 to ensure models stayed consisted across testing. The model accuracy was 93.929%(+-.01), the f score was .8946(+-.01) it used 10 fold cross validation. From the confusion matrix it shows that it classified Spruce 94.71%, Lodgepole Pine 95.8%, Ponderosa Pine 94.4%, Cottonwood 82.3% Aspen 73.4%, Douglas-Fir 84.0% and Krumholz 92.7 % correctly.

Confusion Matrix-test	Spruce	Lodgepole	Ponderosa	Cottonwood	Aspen	Douglas	Krumholz
Spruce/ Fir	69605	4033	4	0	40	10	121
Lodgepole Pine	4162	94893	233	4	159	135	28
Pondersoa Pine	3	280	11798	55	10	317	0
Cottonwood/Willow	0	3	174	733	0	33	0
Aspen	53	805	54	0	2373	10	0
Douglas-fir	10	261	703	44	3	5056	0
Krumholz	552	47	0	0	0	0	6551

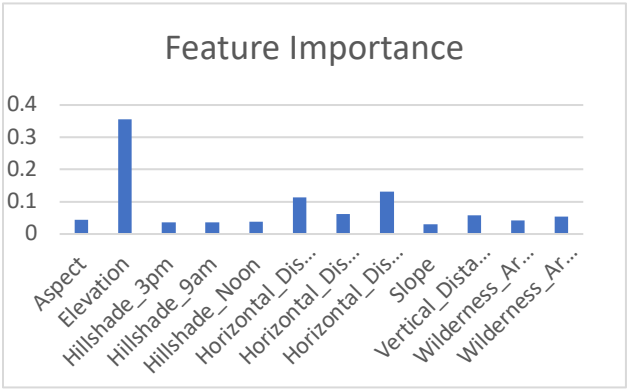
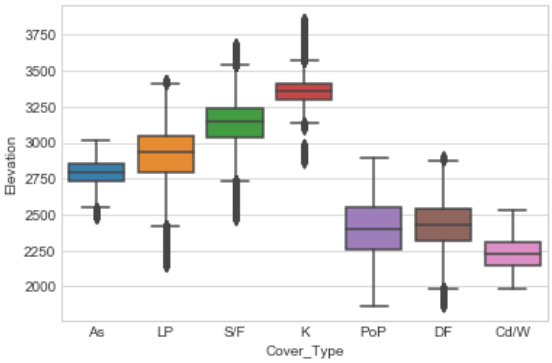


Figure 2

	Feature Score
Elevation	0.355030
Horizontal_Distance_To_Roadways	0.130446
Horizontal_Distance_To_Fire_Points	0.113900
Horizontal_Distance_To_Hydrology	0.061202
Vertical_Distance_To_Hydrology	0.057217
Wilderness_Area4	0.054485
Aspect	0.044731
Wilderness_Area1	0.041422
Hillshade_Noon	0.038127
Hillshade_9am	0.036801
Hillshade_3pm	0.035505
Slope	0.031136

Figure 3

The final model did well overall on the testing set. Even with a major class imbalance, the random forest algorithm was able classify the 7 classes with high accuracy. The feature importance of the top 5 variables led the prediction rates for the random forest as seen in this chart. The elevation feature obtained a score of .355(see figure 2 & 3) which the model heavily to predict the class labels. Looking at the box plot for the distribution of elevation across classes, there is clearly a relationship for each tree regarding elevation. Spruces and lodgepole pines seem to grow in higher altitude areas while Ponderssoa, Douglas-Firs, and Cottonwood/willows seem to grow in lower altitudes. Horizontal distance to roadways & fire points were also contributors in predicting the class labels. When I thought of trees growing in an area, I didn't think that proximity to fire points or roadways would be a major contributor. It was interesting to see that wilderness areas or soil types were not picked up as major factors in predicting tree types. I made the assumption that those would be important features, but the random forest disproved that hypothesis.



Conclusion & Future Work: The random forest was able to predict/solve the multi class problem that was present within the dataset but I wasn't able to find a satisfactory answer in regards to tree growth factors. The dataset was interesting to explore, but wasn't fitted to the original problem proposed. Regarding predicting the tree types, elevation seemed to be the most important. It seemed that certain tree types grow in certain altitudes. Distance to hydrology was also an important feature noted was trees need for water to enhance growth. The other 2 variables distance to roadways and fire points don't seem to promote the idea of trees growing in a certain area. I was able to find variables that were able to predict the class of the tree types but I was unable to find any meaningful attributes that contributed to tree growth besides elevation and distance to hydrology. In the future I would like to work on a similar project like this, but have other feature attributes present in the dataset. Using gird search and out of bag(OOB) error would be another consideration for this dataset as the papers that I read explored these methods to find the optimal model. Perhaps using those additional tools in the future could offer greater insight and better model explainability. Exploring these kinds of datasets fascinates me in addition to using this prediction algorithm in the future. Since the dataset was provided by the US forest service, it would be interesting to see if they include different variables in other national parks. If I attempt a similar project again I would look to see if I could find longitude and latitude coordinates to discover whether tree species grow next to each other or if they form their own clusters within the wilderness area. With that information, it would be more applicable to the original goal intended for this project.

Sources:

Rahmati, Omid, et al. "Application of GIS-Based Data Driven Random Forest and Maximum Entropy Models for Groundwater Potential Mapping: A Case Study at Mehran Region, Iran." *Catena*, vol. 137, 2016, pp. 360–372., doi:10.1016/j.catena.2015.10.010.

Gislason, Pall Oskar, et al. "Random Forests for Land Cover Classification." *Pattern Recognition Letters*, vol. 27, no. 4, 2006, pp. 294–300., doi:10.1016/j.patrec.2005.08.011

Saeys, Yvan, et al. "Robust Feature Selection Using Ensemble Feature Selection Techniques." *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science*, 2008, pp. 313–325., doi:10.1007/978-3-540-87481-2_21.

Probst, Philipp, and Anne-Laure Boulesteix. "To Tune or Not to Tune the Number of Trees in Random Forest." *To Tune or Not to Tune the Number of Trees in Random Forest*, no. Journal of Machine Learning Research 18 (2018) 1-18, Apr. 2018, pp. 1–18.

Chaudhary, Archana, et al. “An Improved Random Forest Classifier for Multi-Class Classification.” *Information Processing in Agriculture*, vol. 3, no. 4, 2016, pp. 215–222., doi:10.1016/j.inpa.2016.08.002.

Becker, Andrea. “Rates of Deforestation & Reforestation in the U.S.” *Education*, 29 Sept. 2016, education.seattlepi.com/rates-deforestation-reforestation-us-3804.html.