Forest Cover Types
Multiclass Classification Model

Supervised Learning Capstone
Nathalie Ries
June 2020

INTRODUCTION

CLIMATE CRISIS

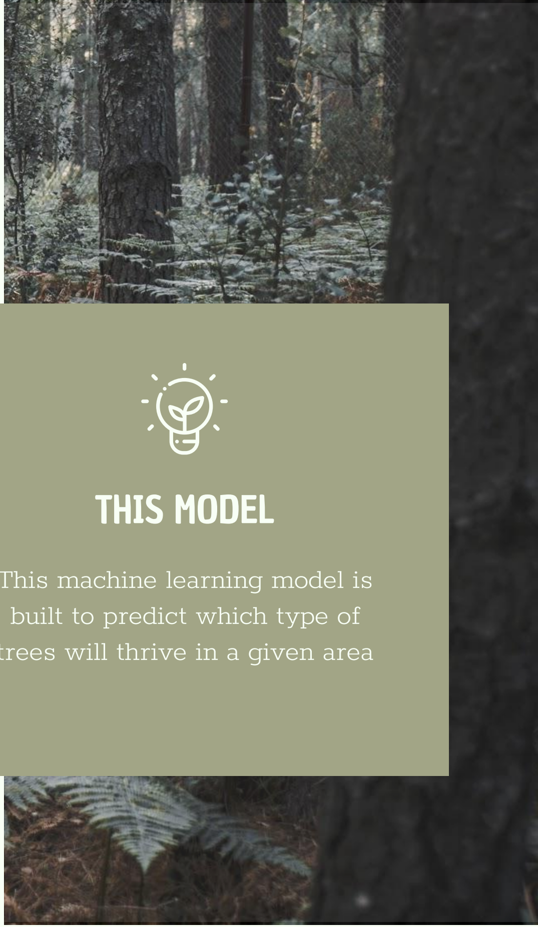
Our world's environment is facing pressing environemntal crises

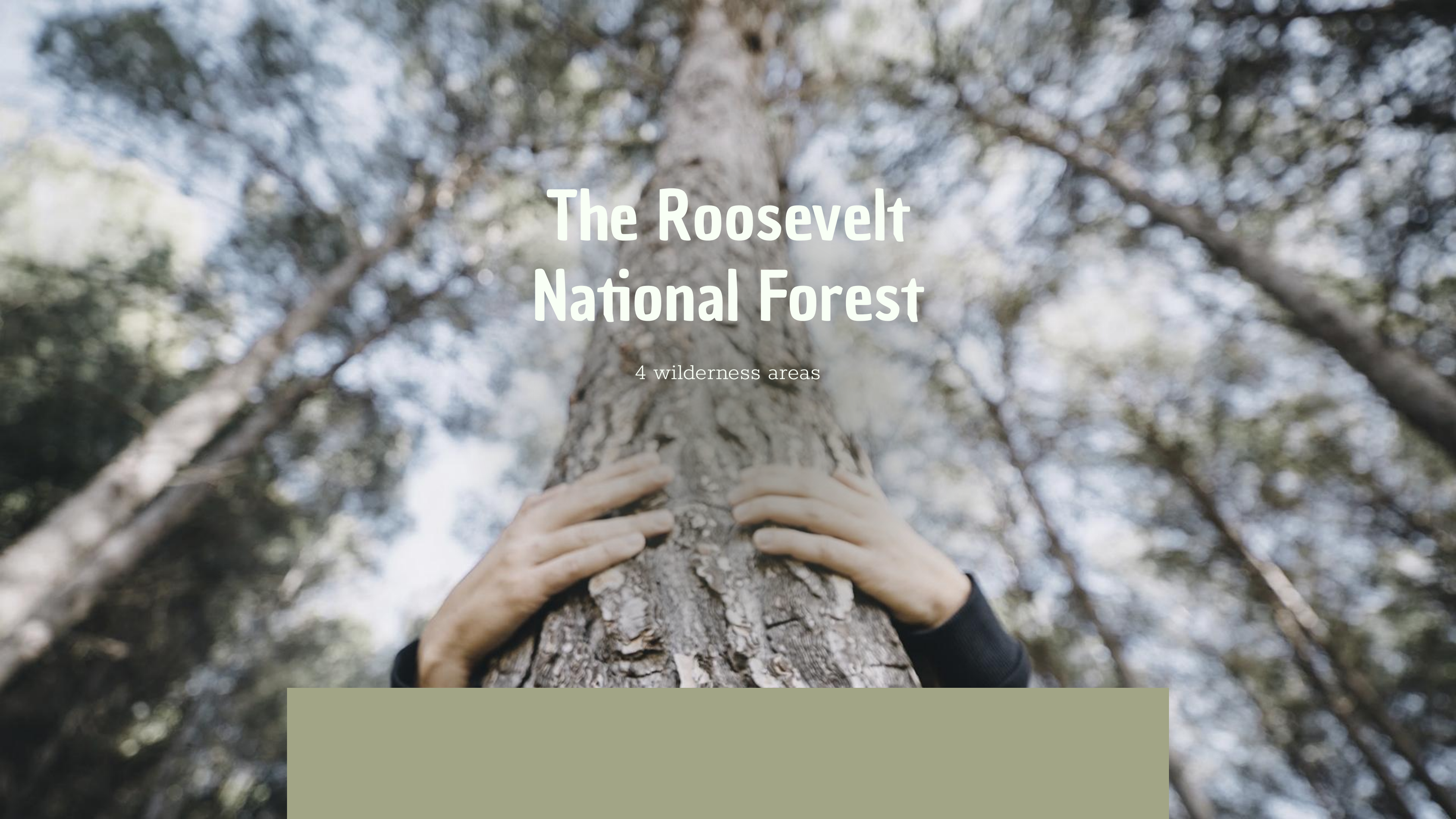
DATA SCIENCE CAN HELP

Data Scientists have lots to offer to help the environmental sciences as they struggle with explosion of big data

THIS MODEL

This machine learning model is built to predict which type of trees will thrive in a given area





The Roosevelt National Forest

4 wilderness areas

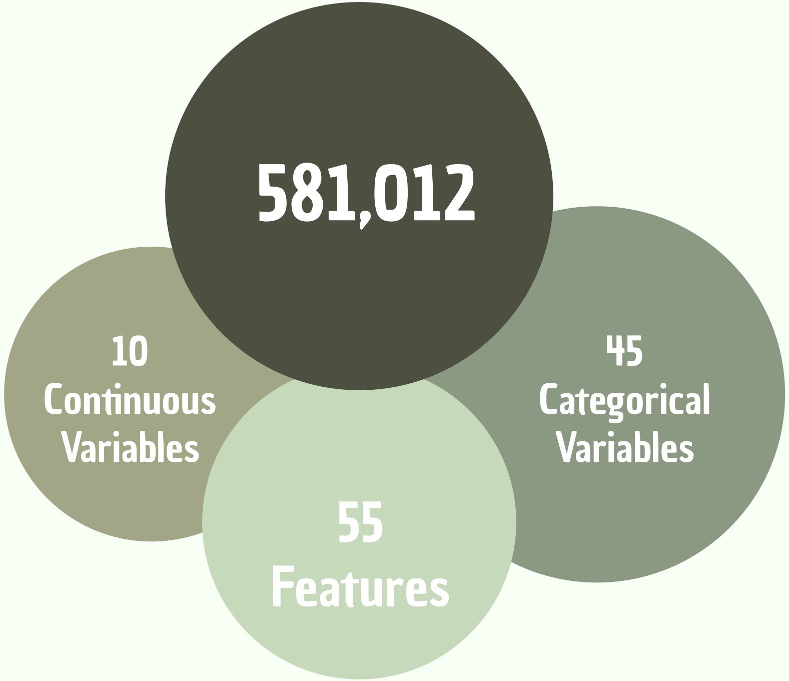
The Data Set

TOPOGRAPHY

Elevation
Slope
Aspect
Sun

DISTANCES

Vertical and Horizontal
Distances to Water
Distance to Roads
Distance to fire



581,012

10 Continuous Variables

45 Categorical Variables

55 Features

WILDERNESS

4 Wilderness Areas:
Rawah
Neota
Comanche
Cache La Poudre

40 SOIL TYPES

Specific geological attributes:
Ex: Cathedral family - Rock outcrop complex, extremely stony.

OUR TARGET VARIABLE

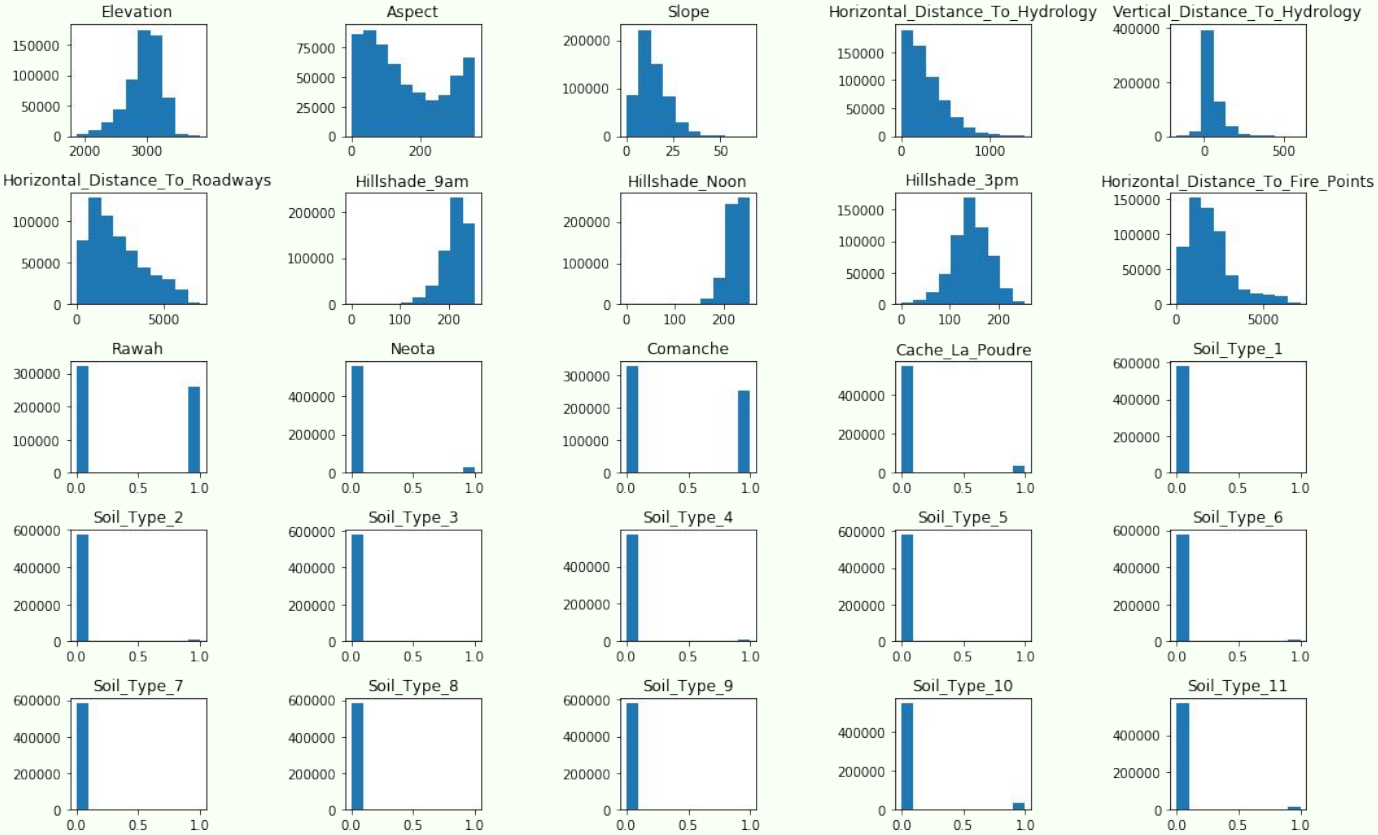
Forest Cover Type:
1 Spruce/Fir
2 Lodgepole Pine
3 Ponderosa Pine
4 Cottonwood/Willow
5 Aspen
6 Douglas-fir
7 Krummholz

01

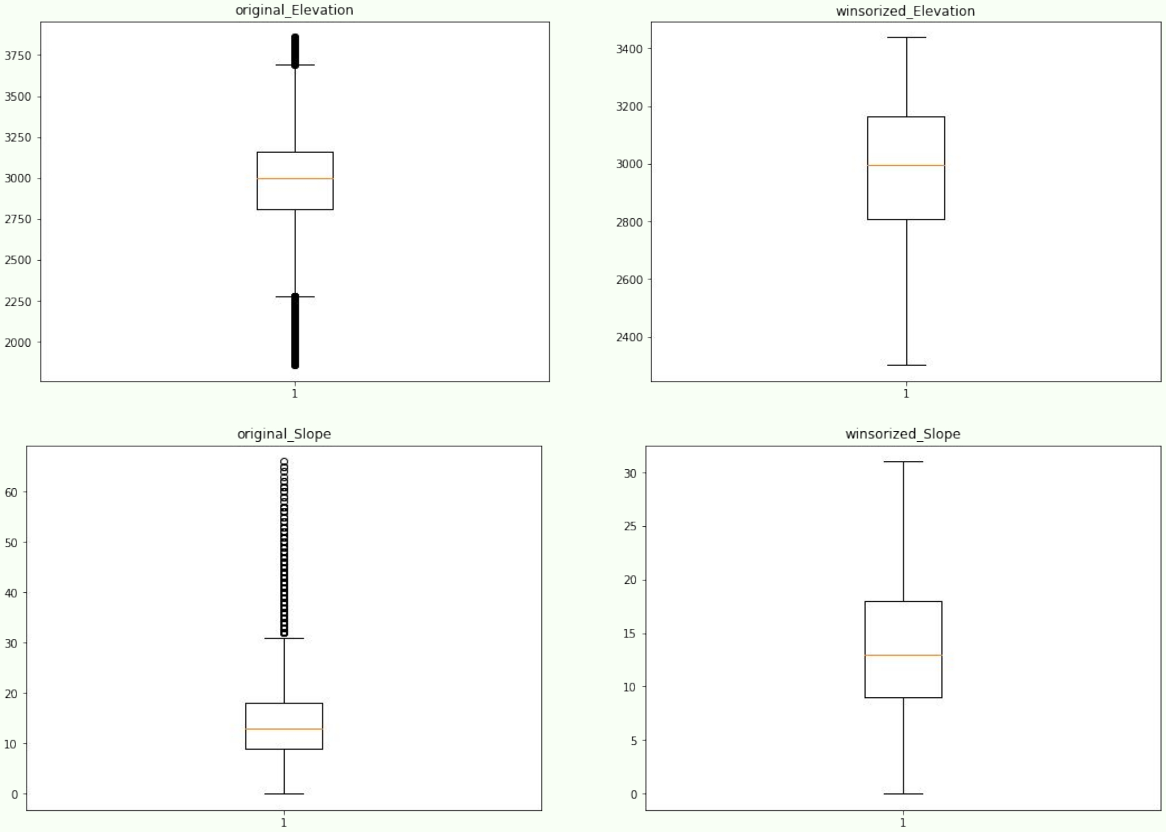
EXPLORATORY DATA ANALYSIS

VARIABLE DISTRIBUTION

Mostly non normally distibuted continous variables



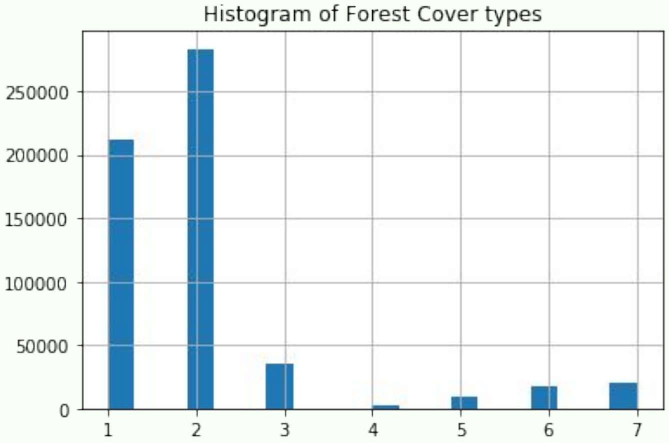
OUTLIERS



Identified small quantity of outliers in 9 out of the 10 continuous variables

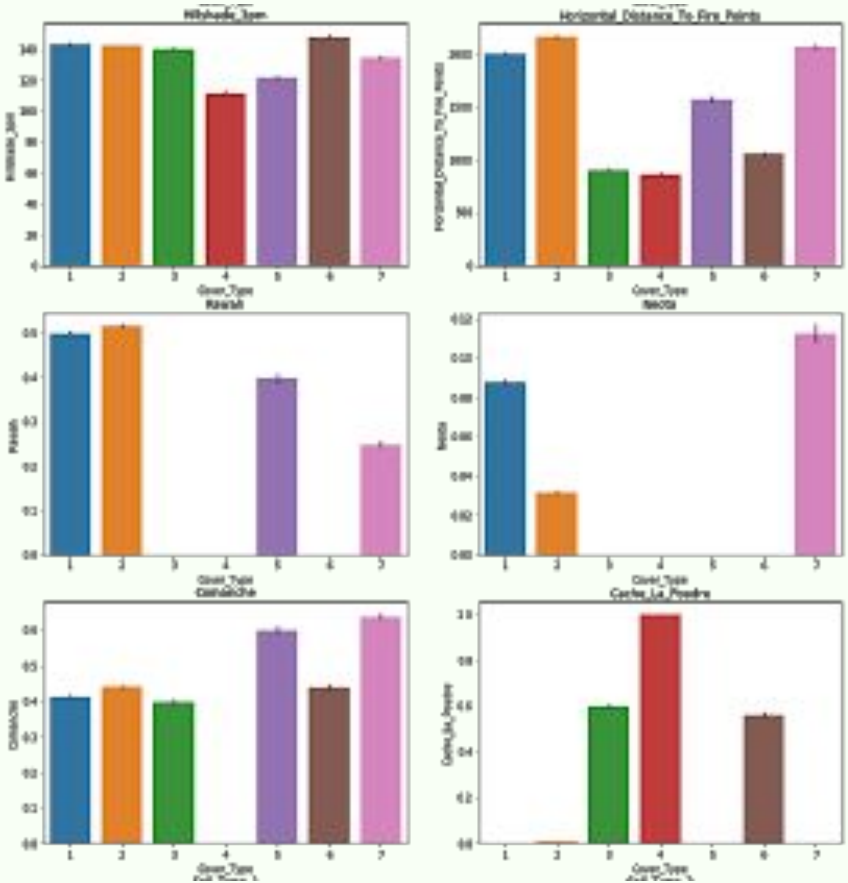
Addressed them by using the winzorisation technique

OUR TARGET VARIABLE

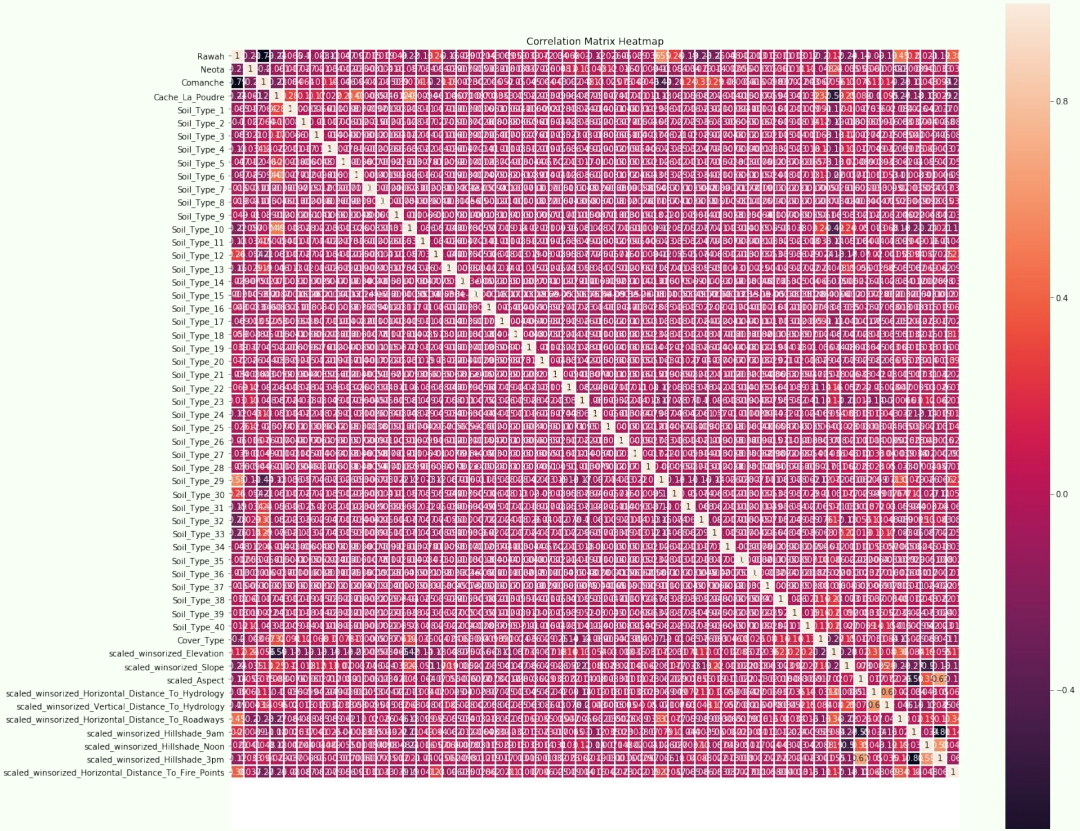


Significant class imbalance

Meaningful differences between features



CORRELATION ANALYSIS



Used a function to identify highly correlated features (>90%).

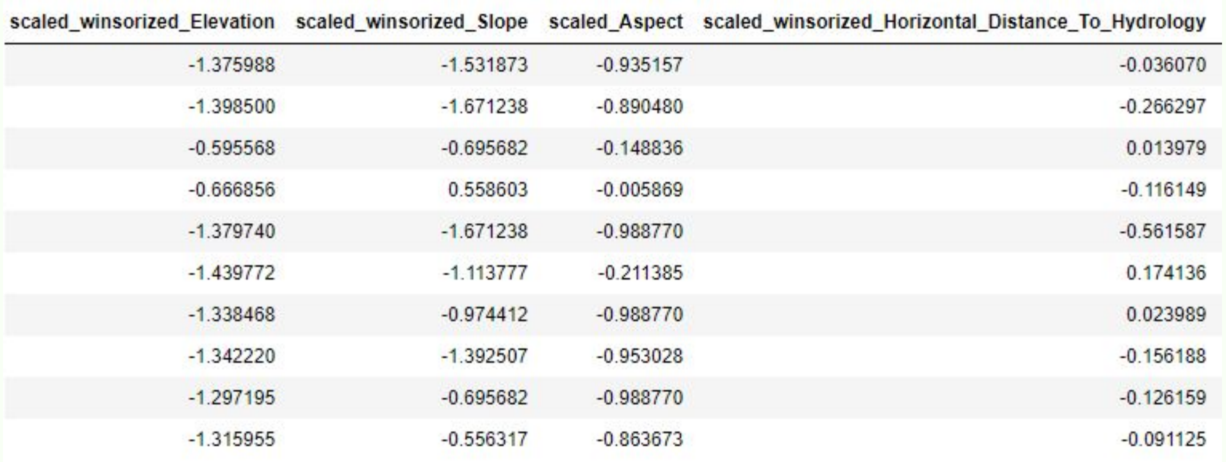
Retained all variables

SCALED CONTINUOUS FEATURES

Evaluated for normality using Jarque Bera and Normal tests

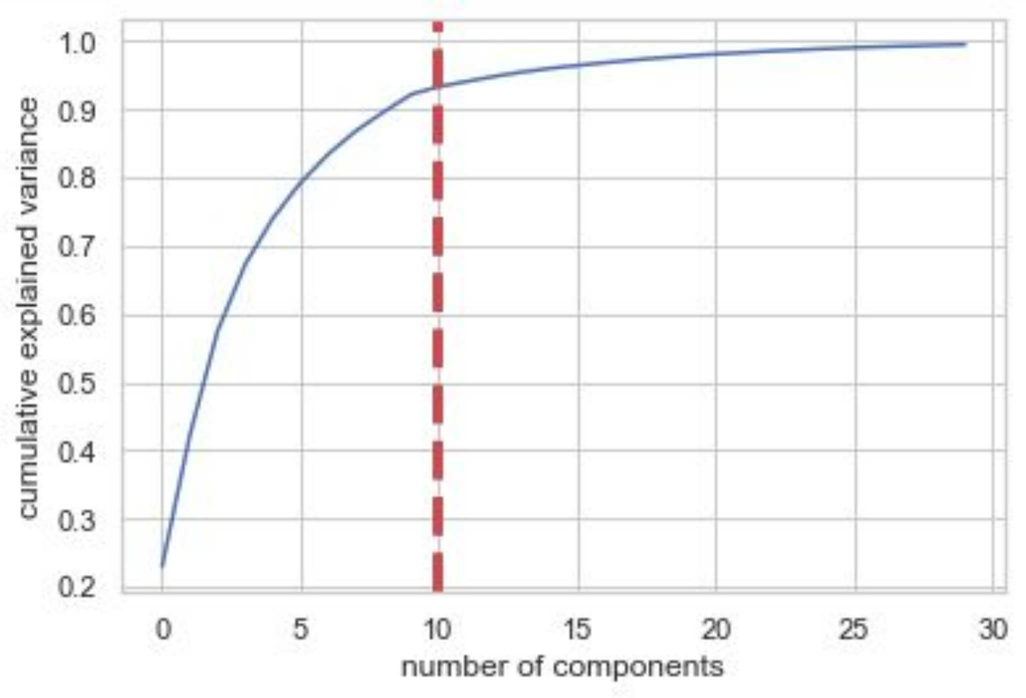
Continuous Variables weren't normally distributed

Applied scale



scaled_winsorized_Elevation	scaled_winsorized_Slope	scaled_Aspect	scaled_winsorized_Horizontal_Distance_To_Hydrology
-1.375988	-1.531873	-0.935157	-0.036070
-1.398500	-1.671238	-0.890480	-0.266297
-0.595568	-0.695682	-0.148836	0.013979
-0.666856	0.558603	-0.005869	-0.116149
-1.379740	-1.671238	-0.988770	-0.561587
-1.439772	-1.113777	-0.211385	0.174136
-1.338468	-0.974412	-0.988770	0.023989
-1.342220	-1.392507	-0.953028	-0.156188
-1.297195	-0.695682	-0.988770	-0.126159
-1.315955	-0.556317	-0.863673	-0.091125

APPLIED PCA



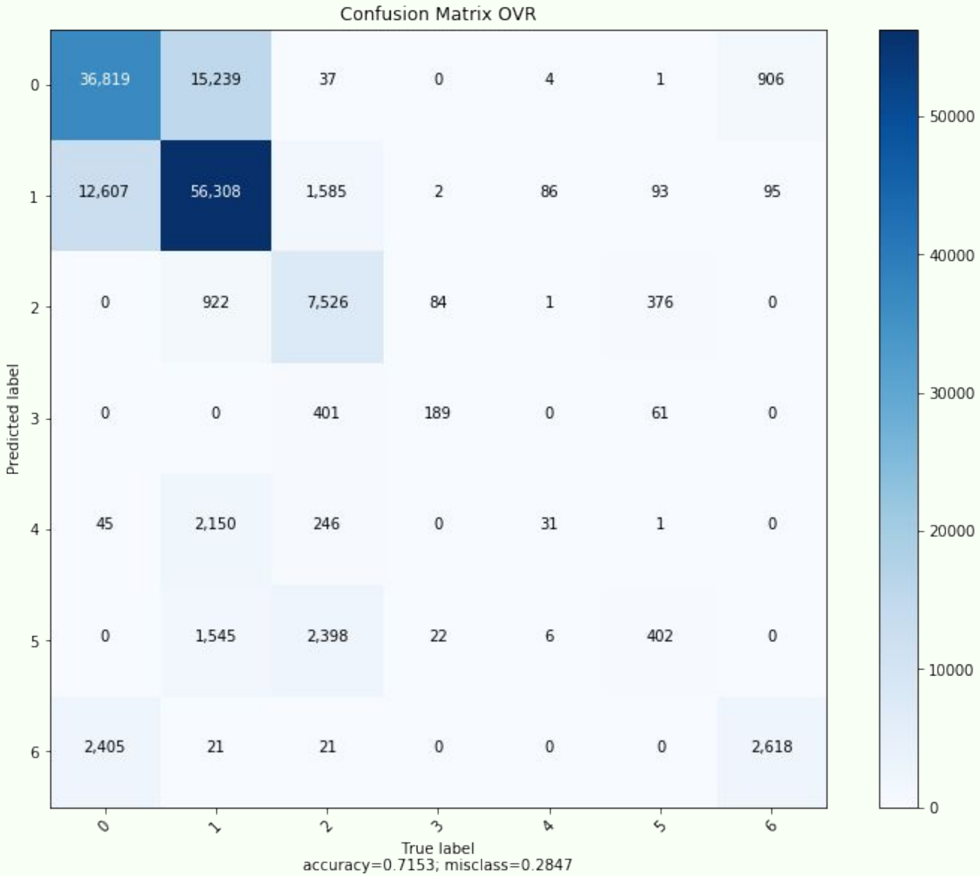
Applied PCA to improve the performance of the gradient boosting model

Retained 20 features
(98 % of the variance)

02

MODELING

TESTED AND EVALUATED 10 MODELS



	0	1	2	3	4	5	6
0	36,819	15,239	37	0	4	1	906
1	12,607	56,308	1,585	2	86	93	95
2	0	922	7,526	84	1	376	0
3	0	0	401	189	0	61	0
4	45	2,150	246	0	31	1	0
5	0	1,545	2,398	22	6	402	0
6	2,405	21	21	0	0	0	2,618

accuracy=0.7153; misclass=0.2847

Using different parameters:

Logistic Classifier

KNN

Decision Tree

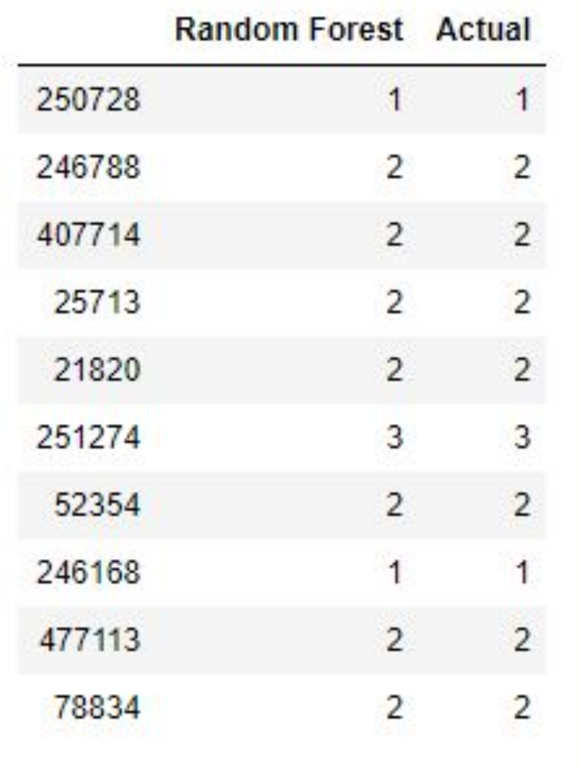
Random Forest

Gradient Boosting

SUMMARY OF MODEL PERFORMANCE

	Accuracy (Test)	Accuracy (Training)	Run Time (In Seconds)
Logistic Classifier (OVR)	71.5%	71.7%	45
Logistic Classifier (Multinomial)	72.4%	72.6%	22
KNN (k=5)	92.4%	95.3%	219
KNN (K=5, distance)	92.9%	100%	209
KNN (k=7, distance)	92.7%	100%	260
KNN (k=3, distance)	93.1%	100%	143
Decision Tree	87.6%	95.4%	2
Random Forest	**94.9%**	**99.9%**	**30**
Gradient Boosting	80.2%	80.6%	1016
Gradient Boosting with PCA	76.5%	77.2%	1720

Random Forest Model



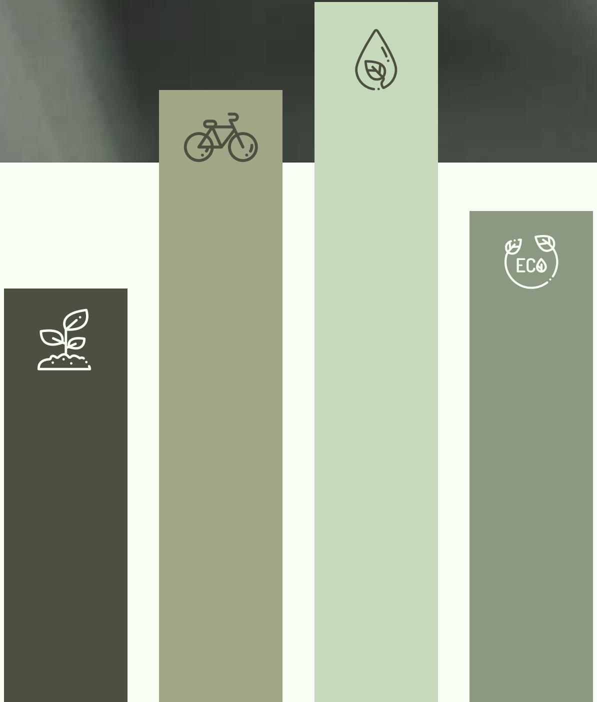
	Random Forest	Actual
250728	1	1
246788	2	2
407714	2	2
25713	2	2
21820	2	2
251274	3	3
52354	2	2
246168	1	1
477113	2	2
78834	2	2

Cross Validation Mean 94.2%

FEATURE IMPORTANCE

H DISTANCE TO HYDROLOGY
6%

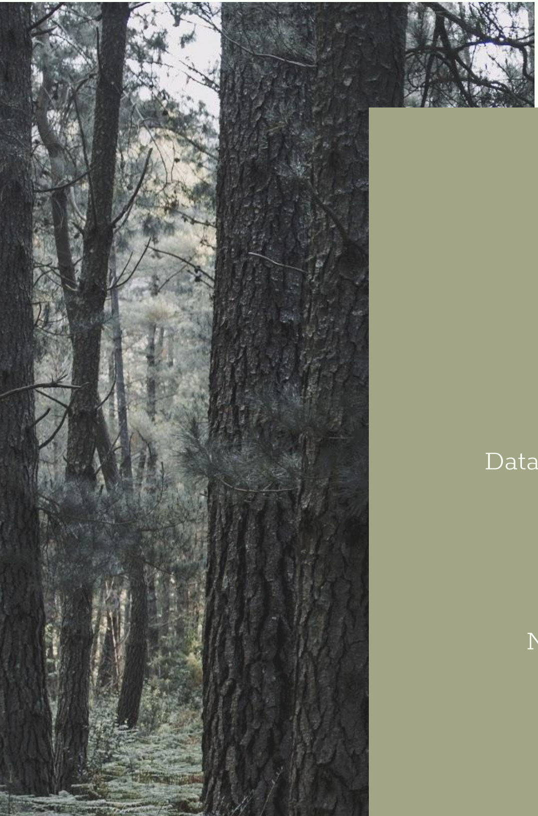
DISTANCE TO ROADWAYS
13%



ELEVATION
25%

DISTANCE TO FIREPOINTS
12%

WHAT'S NEXT?



LIMITATIONS

Data only on the Roosevelt National Forest

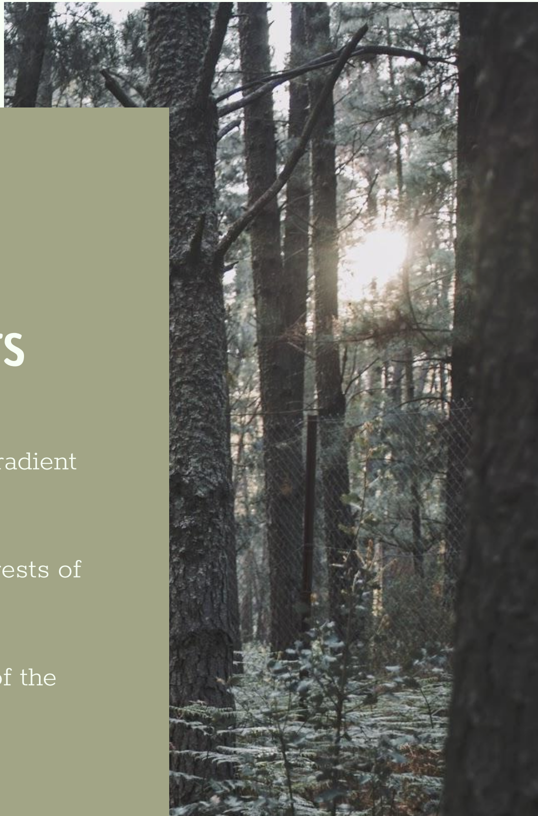
No Time Dimension

DEVELOPMENTS

Parameter Tuning for Gradient Boosting Model

Similar Data on other forests of the world

Expand the variables of the dataset



THANKS!

Does anyone have any questions?

ries.nathalie560@gmail.com

CREDITS

- Jock A. Blackard, Dr. Denis J. Dean, Dr. Charles W. Anderson, of the Colorado State University for the data set
- Presentation template by Slidesgo
- Icons by Flaticon
- Images & infographics by Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik.com
- Big image slide photo created by Freepik.com