



Forest Cover Classification

W207 Applied Machine Learning
- Phil Gagnon, Kai Ding



Table of Contents

01.

Motivation

~1min

Why did we pick this dataset?

02.

Dataset EDA

~3min

Key findings & opportunities for feature engineering / selection

03.

Solution & Approach

~3min

High-level strategy and model performance

04.

Experiments

~2mins

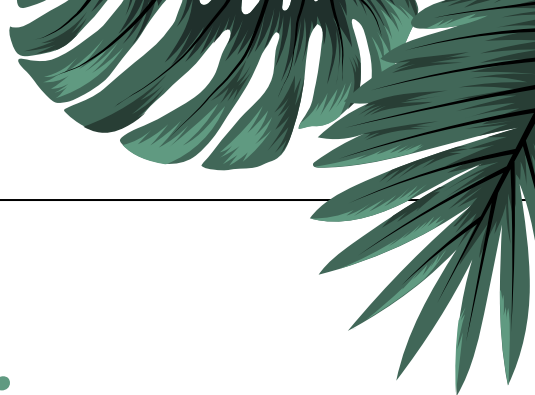
Experiments on feature engineering

05.

Conclusions & Considerations

~2mins

Summary learnings
Ethical considerations etc



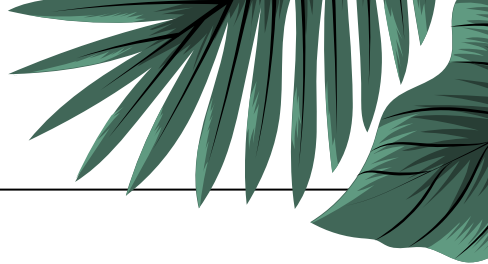


01.

Motivation

Why did we pick this
dataset?

Problem Motivation



Natural conservation requires an understanding of many aspects of ecology.

Tree Cover is one of ecology's key components. It varies as a function of terrain, luminosity, distance to water, and a plurality of other factors.

This study looks at:

- 4 wilderness areas located in the Roosevelt National Forest of northern Colorado, and the
- 7 tree cover types that compose it





Dataset EDA

Overview & key findings

02.

DataSet Overview



Features

Total 54 columns (excl. index IDs):
(overall 15,120 training examples, and
565,892 test examples)

- **One-hot encoded (44 columns):** Wilderness Area 1-4 and Soil Type 1-40
- **Grayscale Pixels (3 columns):** Hillshade 9am/noon/3pm all ranging from 0-255
- **Distance to local amenities (4 columns):** horizontal/vertical distance to hydrology, distance to roadways, distance to firepoints
- **Others (3 columns):** Elevation, Aspect, Slope

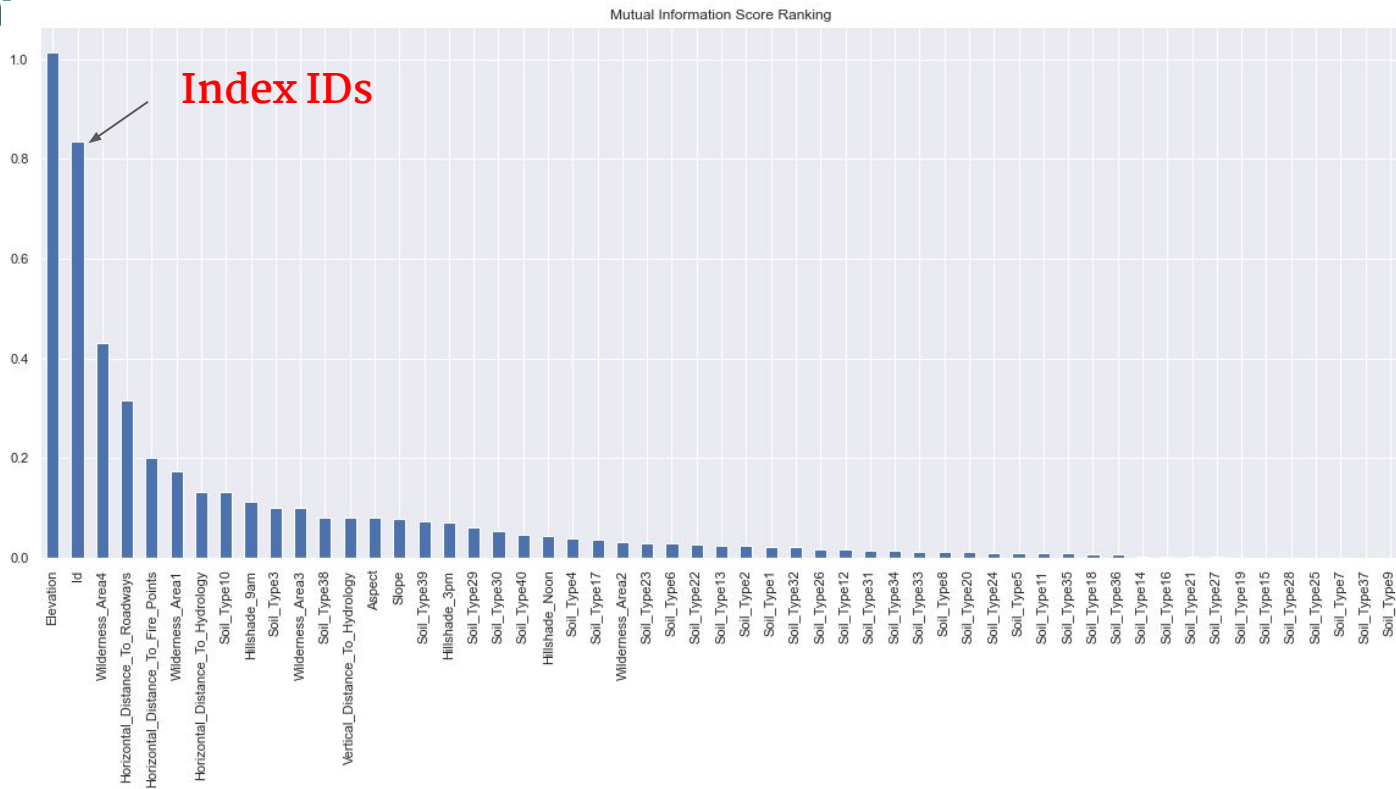
Target

Forest Cover Type from 1-7

- The dataset is perfectly balanced with evenly split cover types, each accounting for 1/7 training examples (2,160 examples each)

Key Findings 1 – Weak Predictors

Index IDs

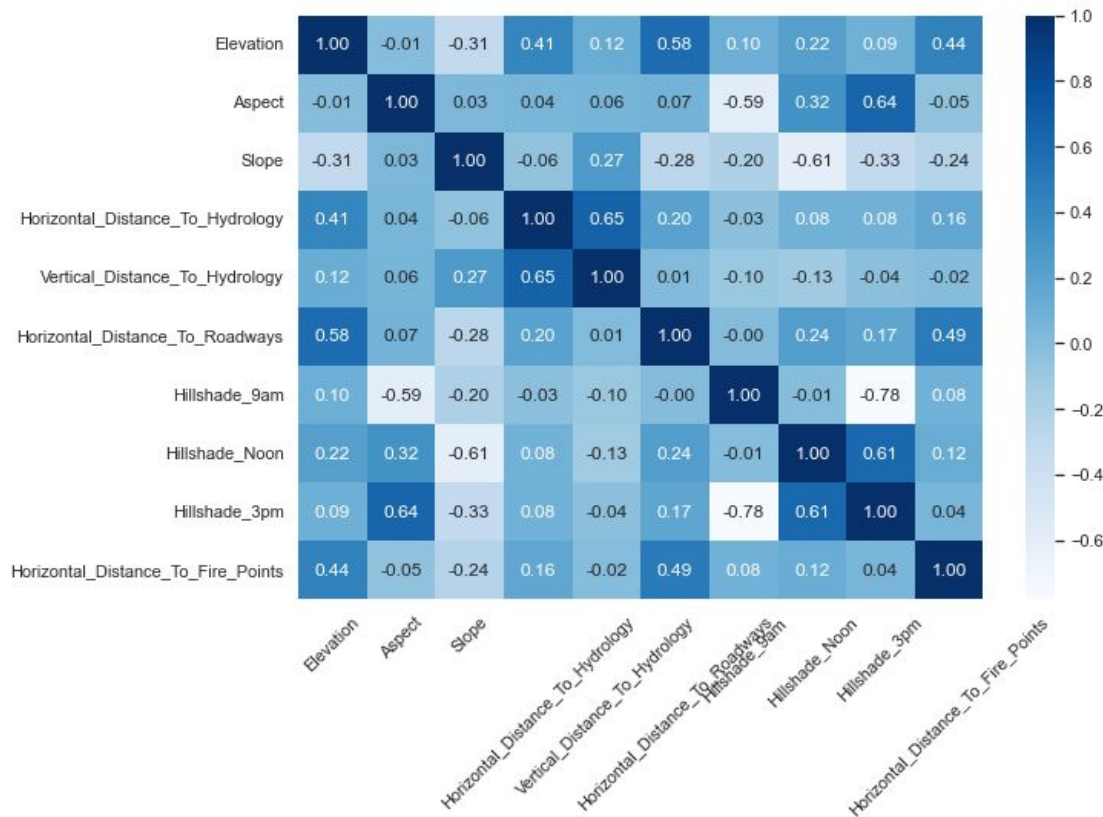


Feature importance ranking with MI score showed that index IDs ranked as #2, but index IDs contain no information about target variables. This means, with the exception of Elevation, other individual variable by itself has very weak predicting power

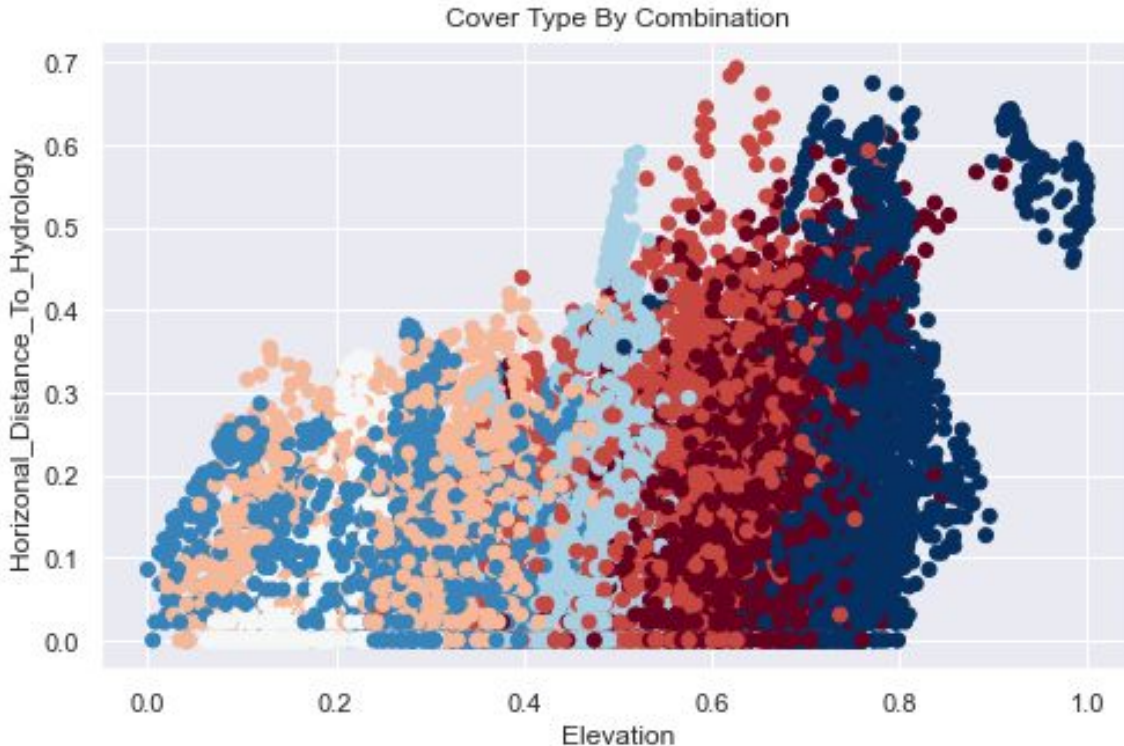
Key Findings 2 – High Collinearity in Numeric Features

Correlation matrix of numerical variables identified multiple pairs of features with high correlation. This indicates potential opportunities for feature engineering. For a few examples:

- Vertical / Horizontal distance to hydrology (0.65)
- Hillshade 9am and 3pm (-0.77)
- Notably, Elevation as the #1 ranked variable has significant corr with several:
 - Distance to hydrology (0.41)
 - Distance to roadways (0.58)
 - Distance to firepoints (0.44)



Key Findings 3 – Feature Engineering Opportunities on Elevation



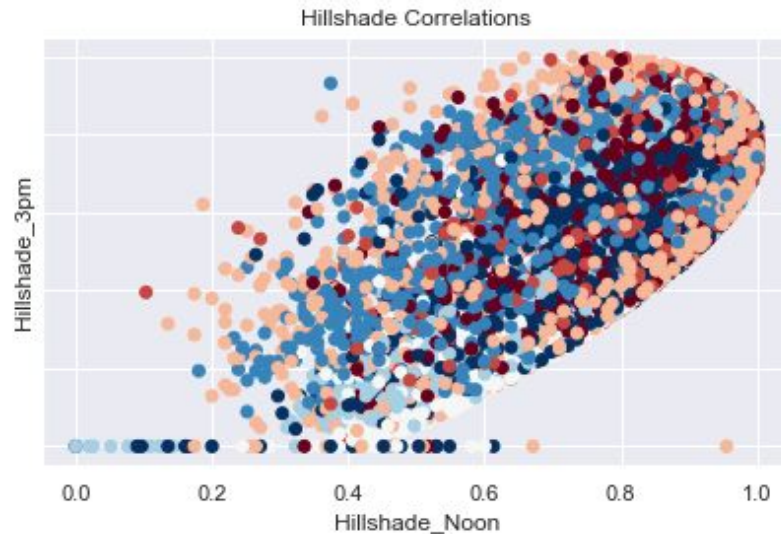
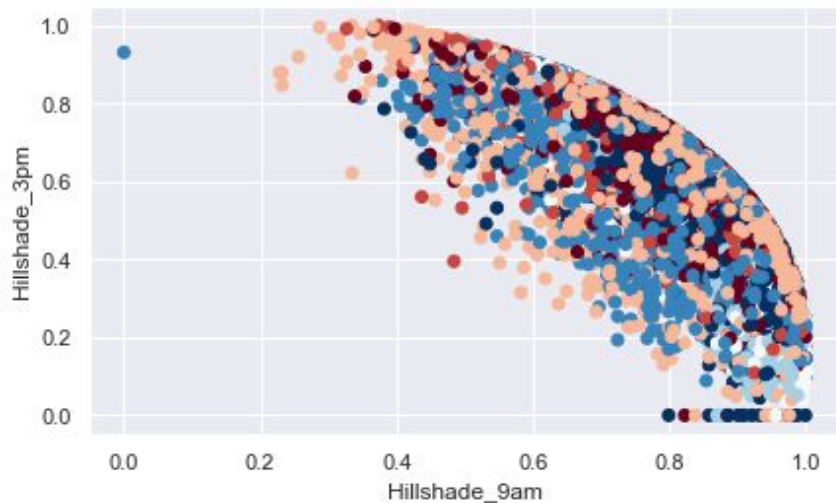
Elevation is strongly correlated with 3 variables:

- We chose to focus on Elevation-related opportunities because Elevation is #1 ranked in feature importance; therefore feature engineering based on Elevation is more likely to create important new features
- Take Horizontal_Dist_To_Hydrology as an example (0.41 correlation). There is an opportunity to create linear combination of it with Elevation so that the new feature has a linear relationship with target variable - This potentially helps improve tree algorithm performance

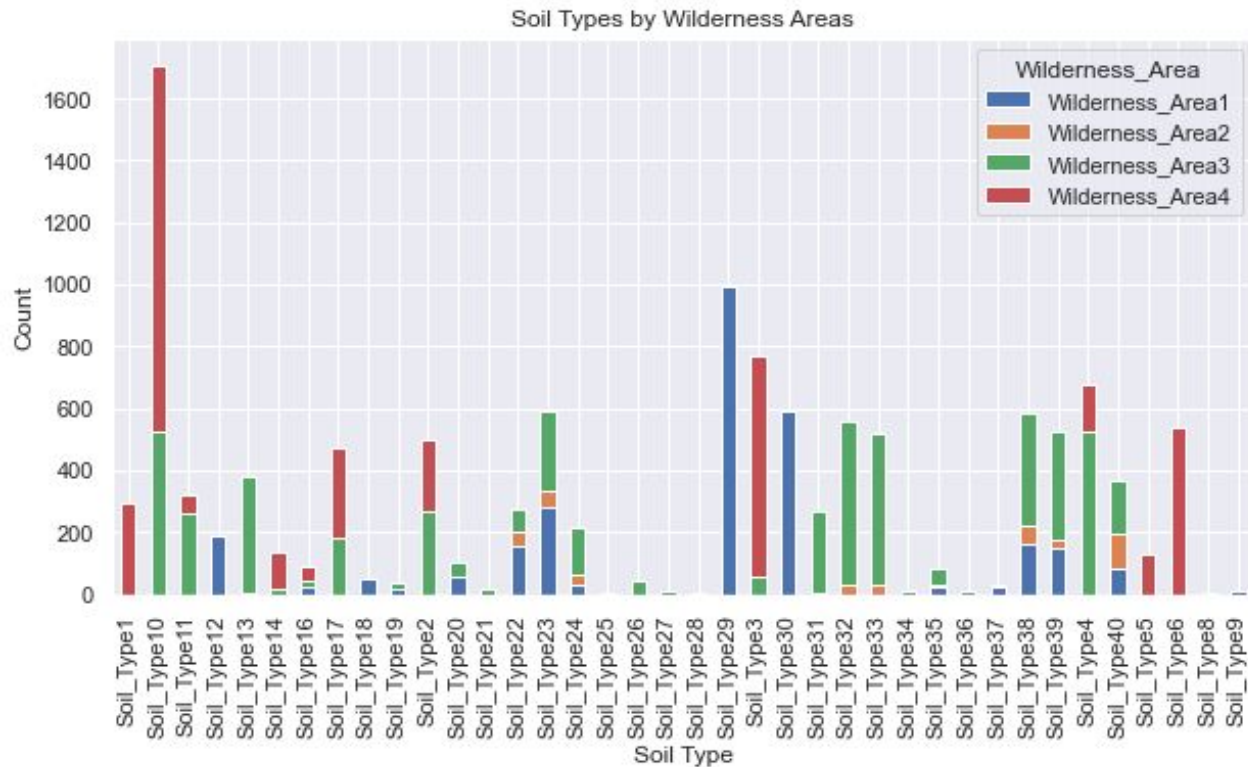
Key Findings 4 - Hillshade Features Highly Collinear

Hillshade 9am/Noon/3pm are encoded as 1-255 grayscale pixel values:

- Hillshade 9am/3pm and Hillshade Noon/3pm are two highly collinear pairs, as indicated in the charts
- Curiously, Hillshade 3pm has lots of 0 values, potentially missing values or indicate the side of hill that is not exposed to sunlight?



Key Findings 5 - One-Hot Encoded Columns Most Soil Types Primarily in One Wilderness Area

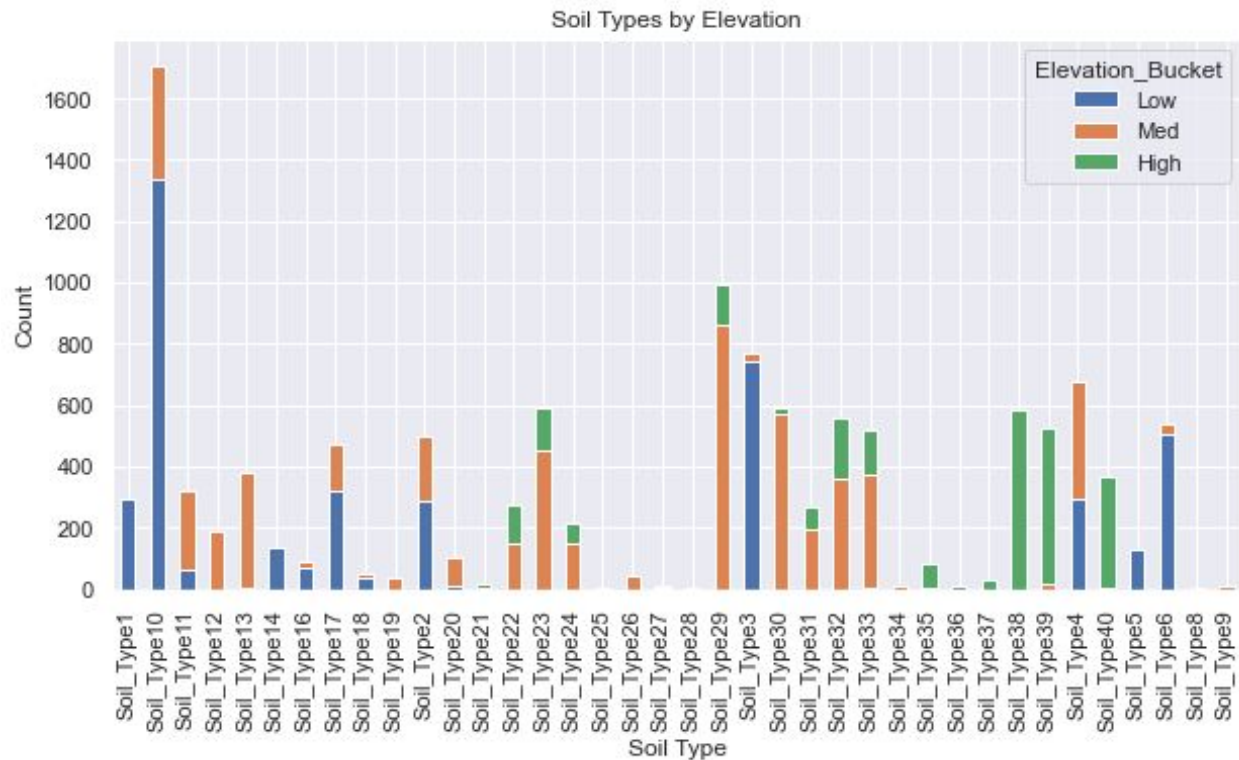


We cannot simply compute correlation between these variables, but plotting soil types by wilderness area suggest the two are highly correlated:

- Some soil types such as Type 29 and 30 are present exclusively in Area 1; Type 6 only in Area 4
- While some soils are present in multiple areas, such as Type 10 and 11, they have 1 pre-dominant color (Area)
- Only a few soils are evenly split (i.e Type 2 and 23)

One-Hot Encoded Columns

Soil Types and Elevation Also Correlated



If we bucket Elevation into 3 categories “Low/Mid/High”, most soil types are present primarily in one elevation bucket.



03.

Solutions & Approach

Strategy and model
performance



Our Approach



Feature Engineering

- Create Euclidean distance or ratio from horizontal/vertical distance:
- Linear combination of highly-correlated features (esp. for top-ranked features such as Elevation)
- PCA to combine soil types



Classifiers Selection

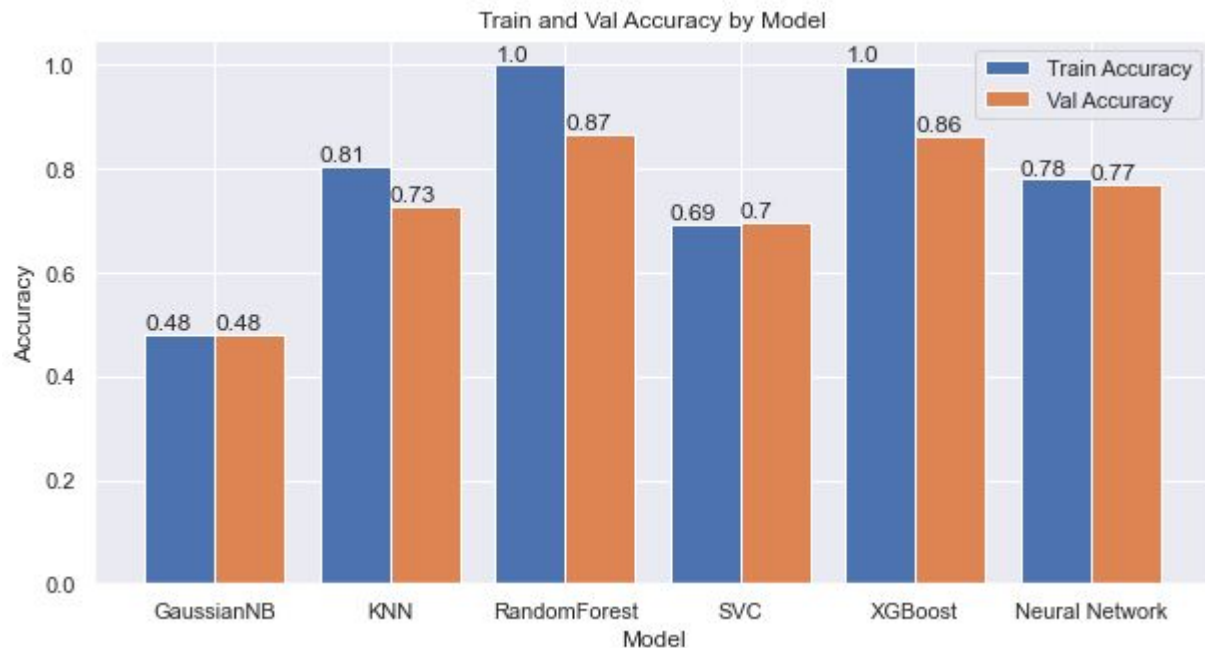
- Multi Classification with non-linear decision boundaries
- Accommodate both numeric and categorical variable types
- Potentially help us with feature selection



Hyperparameter Tuning

- Perform hyper-parameter tuning on Neural Network and Random Forest

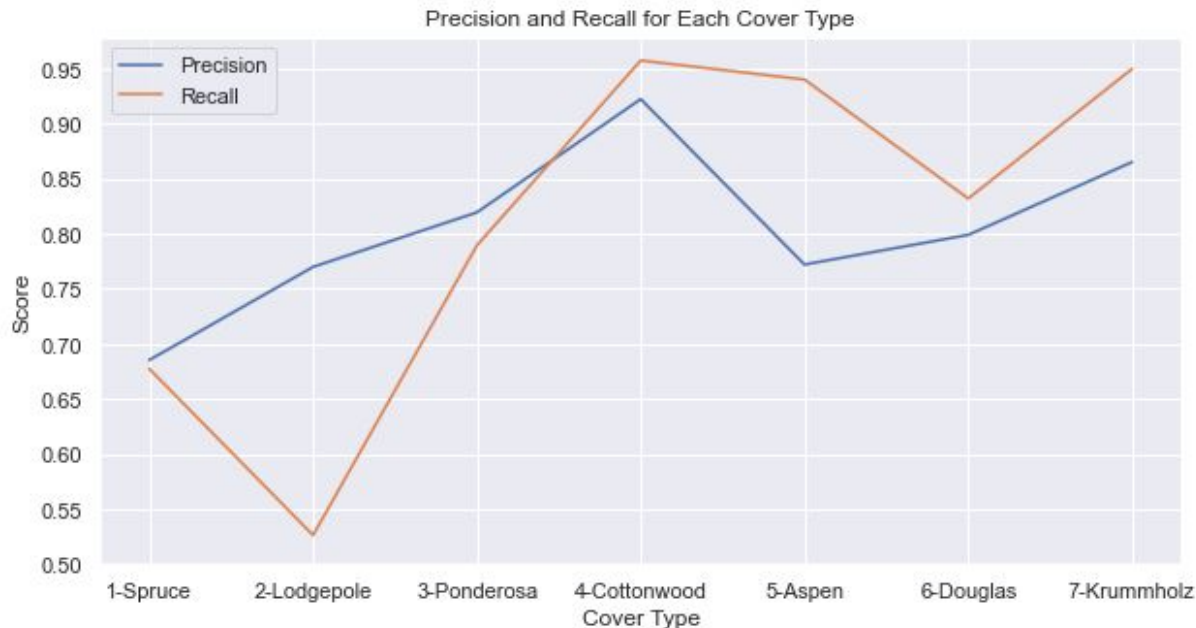
Baseline model comparison



- While RF and XGBoost are tied at #1 for validation accuracy at ~0.86–0.87, they significantly overfit with training accuracy at 1.0. In hyperparameter tuning, we need to add stronger regularization for these models
- KNN and Neural Network produced more balanced results

Tuned model performance – RandomForest

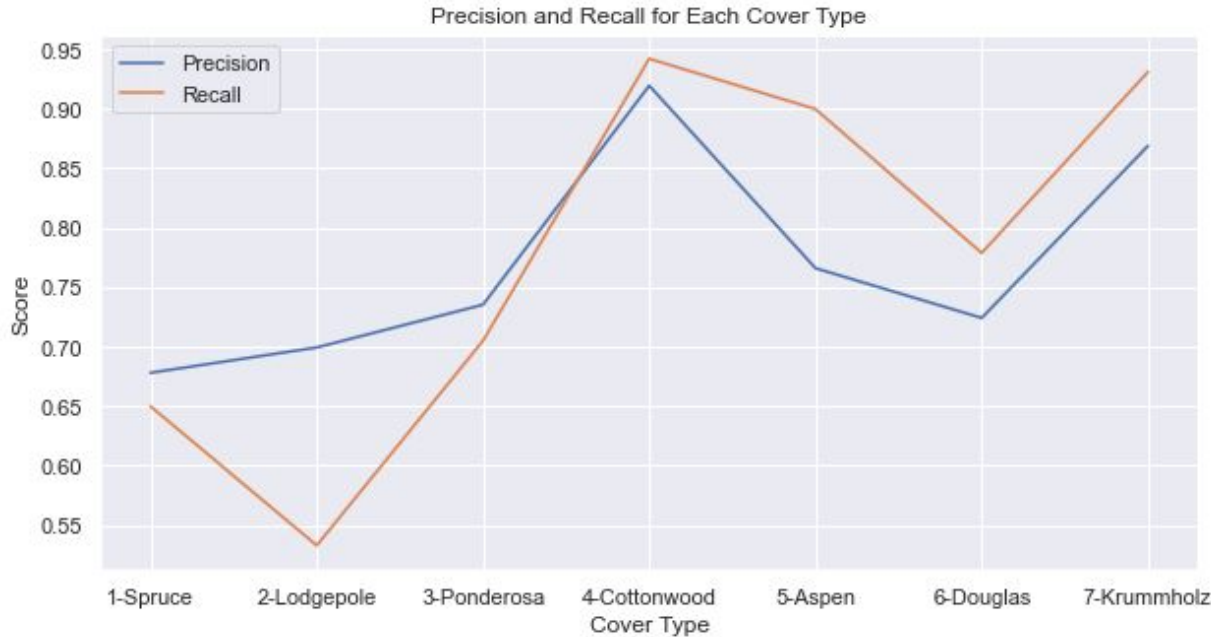
Best performance is training accuracy of **0.86** and val accuracy of **0.81** based on 5-fold CV. Hyperparameter tuning focused on 1) **Max_depth** 2) **Max_features** since the model is prone to overfit.



Precision & Recall by cover type

- Doing well on #4 and #7
- Performed the worst on #1 and #2 precision
- Also not doing well on #3 and #6

Tuned model performance – Neural Network



Precision & Recall by cover type

- Compared with RF, actually performed even better on #4, the top-accuracy cover type for both models
- Performed worse for #2 and #6

Neural Net – Hyperparameter Tuning



Data	Epochs	Hidden Sizes	Activation	Optimizer	LR	# Param	Accuracy	Validation Accuracy
PCA_adj	15	(1024, 128, 32)	relu	Adam	0.01	170,408	0.7880	0.7857
No_PCA	15	(1024, 128, 32)	relu	Adam	0.01	170,408	0.7858	0.7738
PCA_adj	30	(1024, 128, 32)	relu	Adam	0.01	170,408	0.8225	0.7656
PCA_adj	15	(128, 32)	relu	Adam	0.01	8,744	0.7716	0.7725
PCA_adj	15	(8, 2)	relu	Adam	0.01	314	0.6675	0.6885
PCA_adj	15	(1024, 128, 32)	tanh	Adam	0.01	170,408	0.6463	0.6448
PCA_adj	15	(1024, 128, 32)	relu	SGD	0.01	170,408	0.6460	0.6786
PCA_adj	15	(1024, 128, 32)	relu	Adam	0.001	170,408	0.7830	0.7778

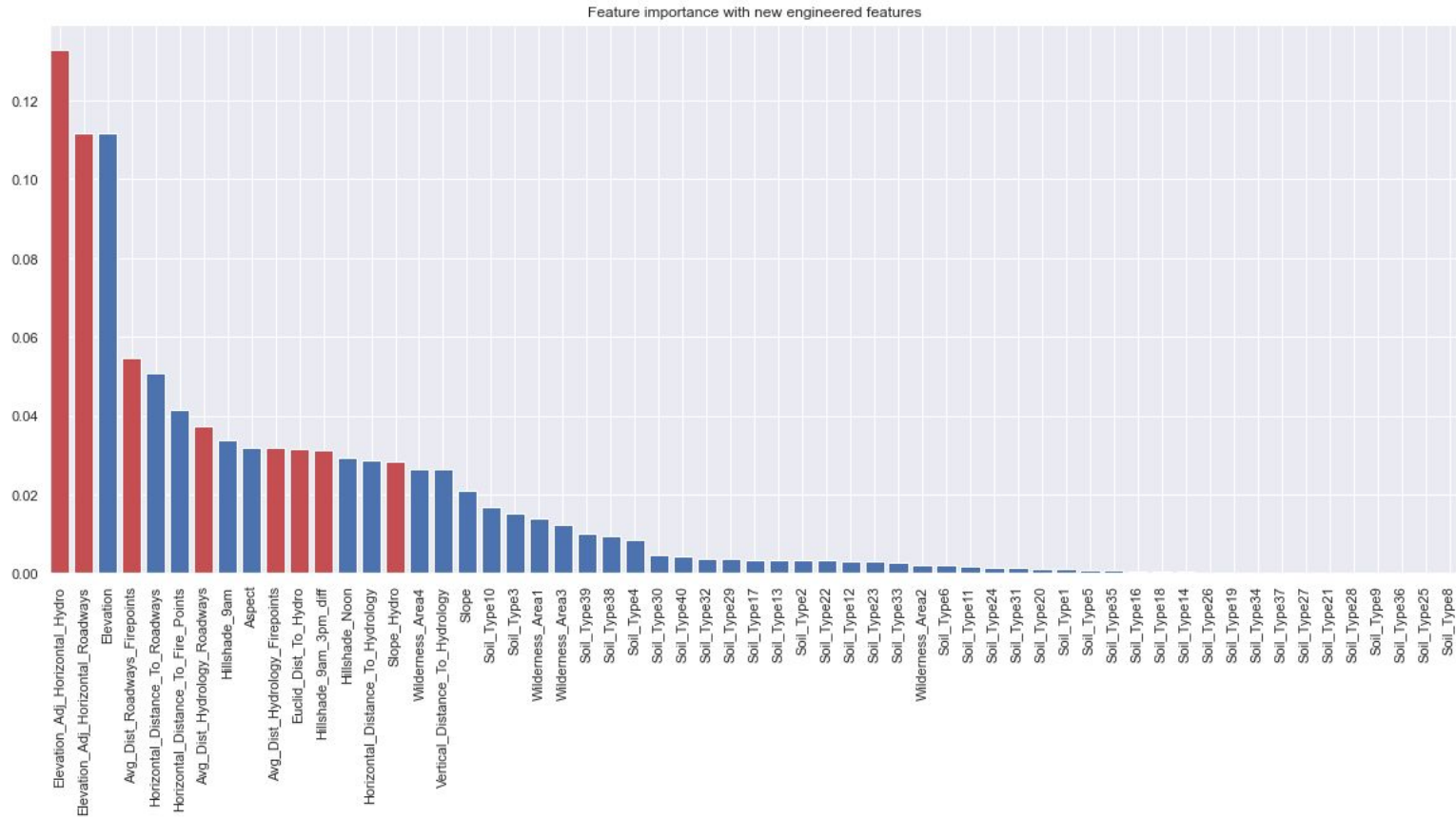
The slide features a white background with a thin black rectangular border. In the four corners, there are stylized green leaves. The leaves on the left and right are large and have a distinct pattern of holes, resembling Monstera leaves. The leaves at the top are smaller and more delicate, resembling fern fronds. The main text is centered within the white area.

Experiment

Feature Engineering Experiments

04.

Feature Importance Ranking with Engineered Features (red colored)



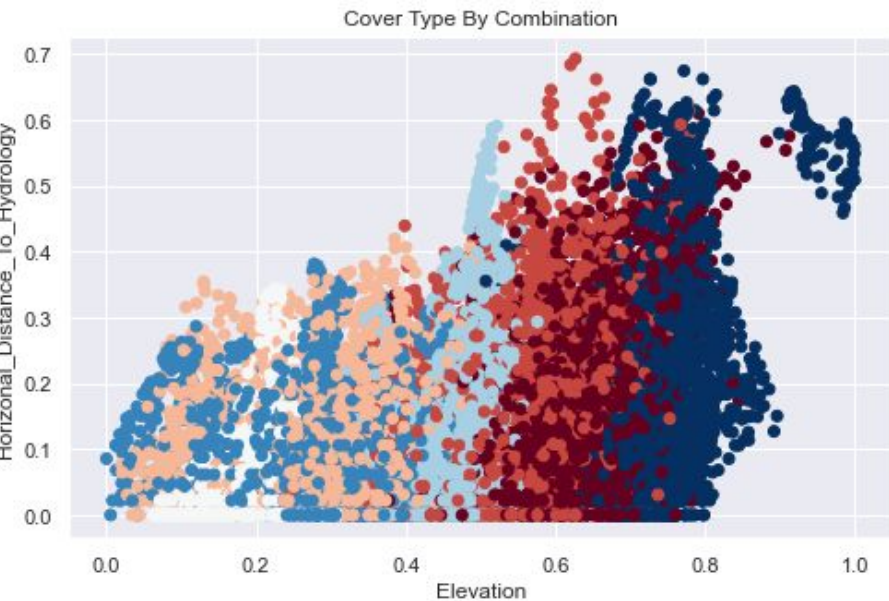
Feature Engineering - New Features Overview

We created 8 new features, 5 of which ranked among Top 10 important features by MI score

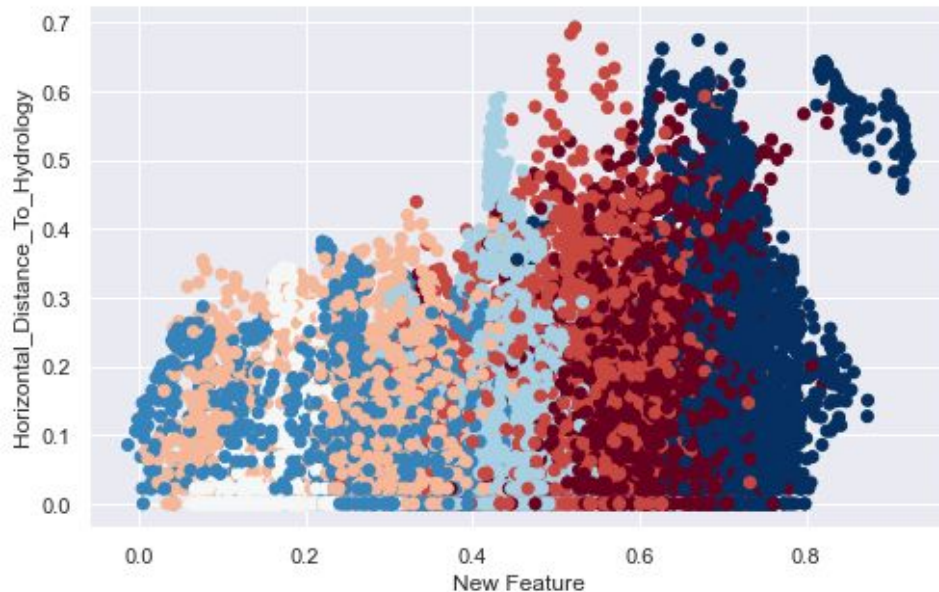
	New Feature	Details	Rank
Euclidean /Ratio	Euclid_Dist_to_Hydro	$=\text{sqrt}(\text{Vertical_Dist_to_Hydro}^2 + \text{Horizontal_Dist_to_Hydro}^2)$	11
	Slope_Hydro	$= \text{Vertical_Dist_to_Hydro} / \text{Horizontal_Dist_to_Hydro}$	15
Linear combinations	Elevation_Adj_Horizontal_Hydro	$= \text{Elevation} - 0.15 * \text{Horizontal_Dist_to_Hydro}$	1
	Elevation_Adj_Horizontal_Roadways	$= \text{Elevation} - 0.05 * \text{Horizontal_Dist_to_Roadways}$	2
	Avg_Dist_Roadways_Firepoints	$= \text{Avg}(\text{Horizontal_Dist_to_Roadways} + \text{Dist_to_Firepoints})$	4
	Avg_Dist_Hydrology_Roadways	$= \text{Avg}(\text{Horizontal_Dist_to_Roadways} + \text{Dist_to_Firepoints})$	7
	Avg_Dist_Hydrology_Firepoints	$= \text{Avg}(\text{Horizontal_Dist_to_Hydro} + \text{Dist_to_Firepoints})$	10
	Hillshade_9am_3pm_diff	$= \text{Hillshade_9am} - \text{Hillshade_3pm}$	12

Top-Ranked New Features Rationale

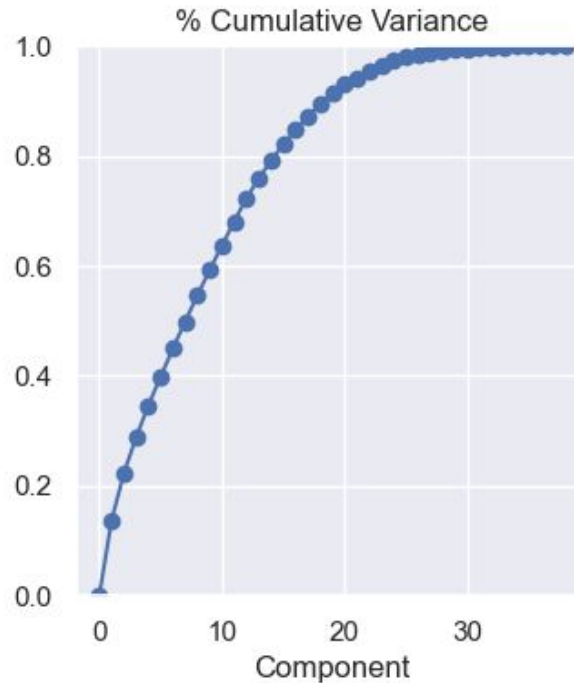
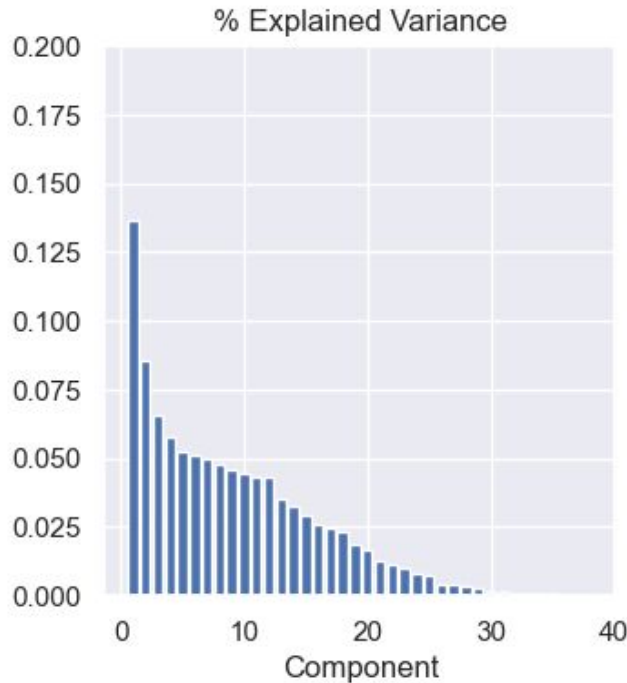
Original features



New feature = Elevation - 0.15 * Horizontal Dist To Hydrology



PCA to combine and simplify Spoil_Types

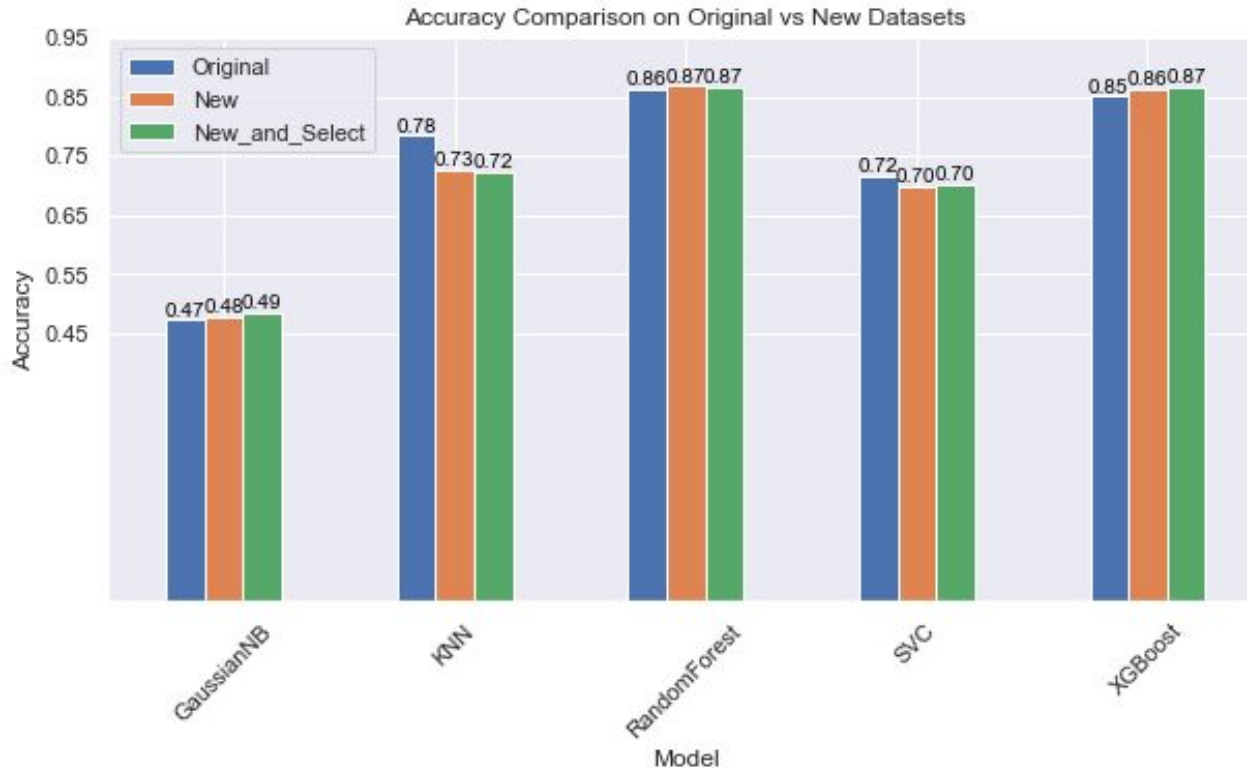


Using PCA to combine 40 Soil Types, we discovered that:

- Top 10 synthesized soil types explained over 60% of the total variance
- Top 20 synthesized soil types explained over 90% of the total variance

We ended up feeding Top 20 synthesized soil types into the Neural Network model.

Feature Engineering & Selection on model performance



- New features did not significantly improve accuracy
- Feature selection seems to be more effective; removing non-essential features (feature_importance score < 0.01) did not decrease accuracy

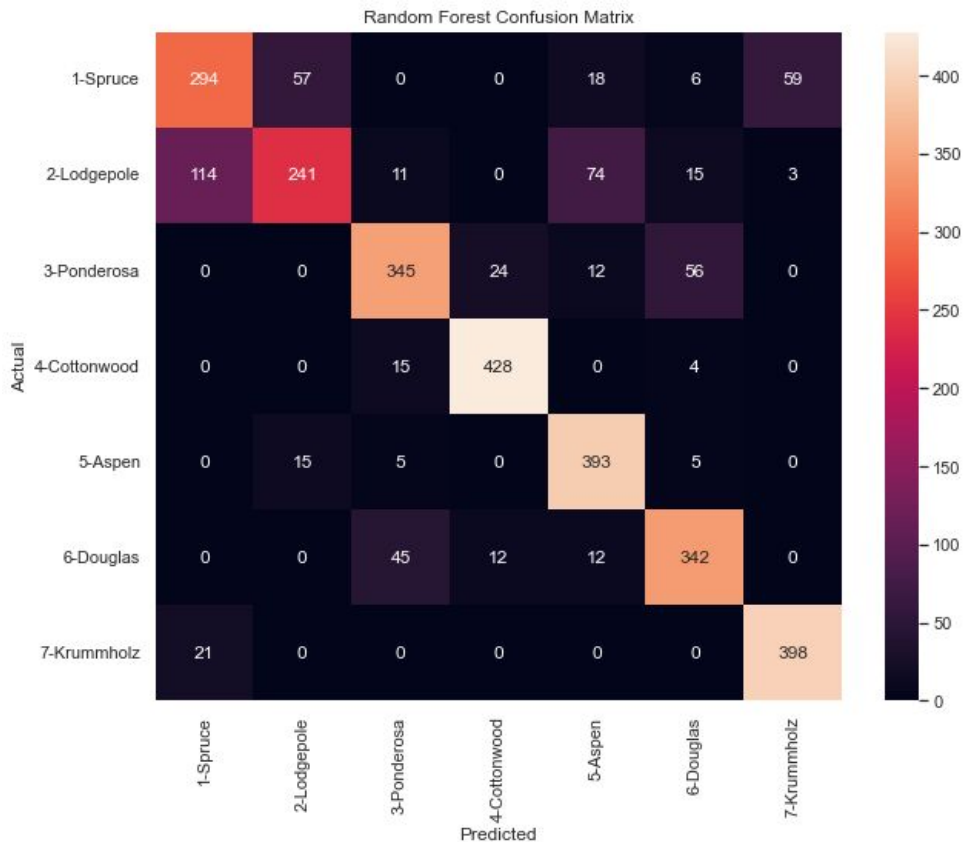
The slide features decorative green leaves in the corners. On the left, there is a large monstera leaf and a palm frond. On the right, there is a monstera leaf and a palm frond. At the top right, there is a palm frond. The central content is enclosed in a thin black rectangular border.

Conclusions

Summary learnings

05.

Confusion matrix insights



RF Confusion matrix (similar for Neural Network matrix) reveals the following insights

- It's challenging to classify between cover #1 and #2, with most of the mis-classifications between the two types
- It's also challenging to classify between type #3 and #6

To further improve model performance, these could be the major focus areas.



06.

Other Considerations


Keeping things clean and green

ESG Angle



If this study can serve any purpose apart from our own learning experience, we would hope it is in respect of ecology and sustainable management of forests.

Some considerations:

- i) Prevent model usage towards rent-seeking, destructive behavior towards natural habitats (e.g. illegal logging)
 - ii) If there was a tree cover type that's particularly in danger, we can optimize model performance to support user cases in relation to that specific type
- 

A decorative border of various tropical leaves, including palm fronds and monstera leaves, in shades of green, framing the central text.

Q&A