Forest Cover Prediction

Data Analysis & ML Modelling

Bhushan Ingale

26/09/2023

Project ID: PRCP-1005-ForestCoverPred | **Team ID**: PTID-CDS-SEP-23-1652

Project Objectives

- The goal of the Project was to predict seven different *Cover Types* given four different *Wilderness Areas* of the Roosevelt National Forest of Northern Colorado with the best accuracy.

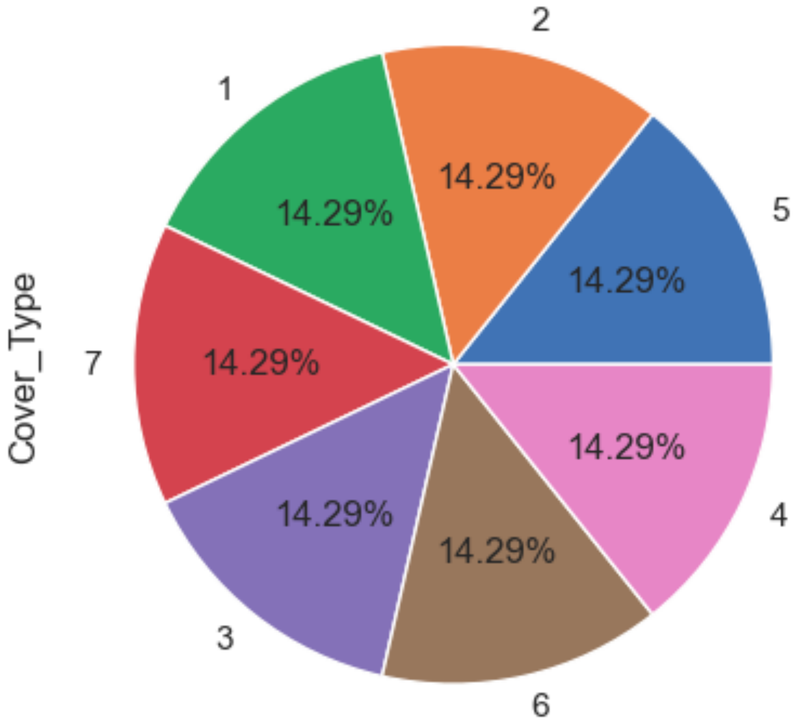
- It was a Multi-label Classification Problem.

- Task 1: Performing Data Analysis on the Forest Cover Dataset.

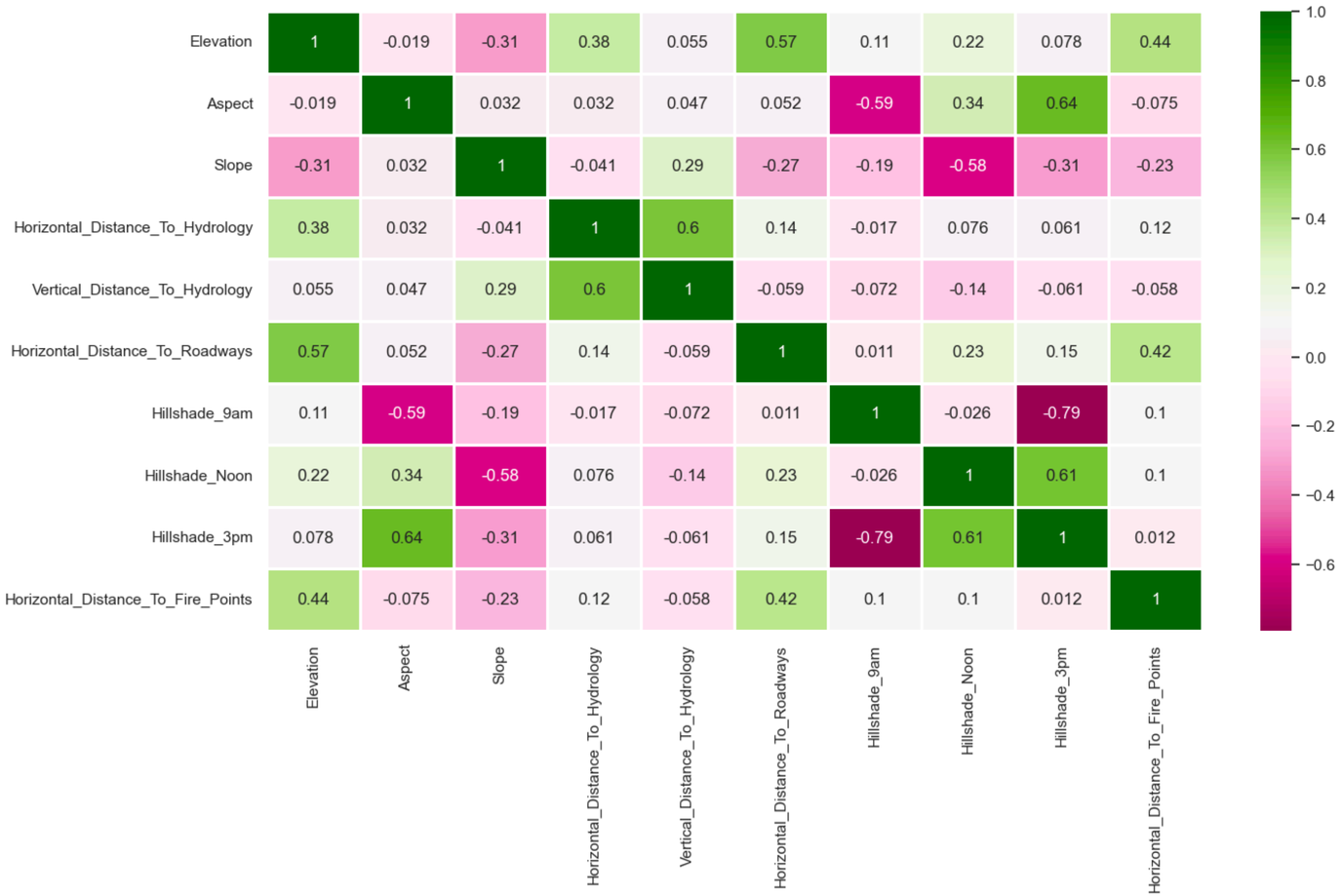
- Task 2: Developing a Machine Learning Model with best possible Accuracy.

General Observations

Cover Types Distribution in Dataset



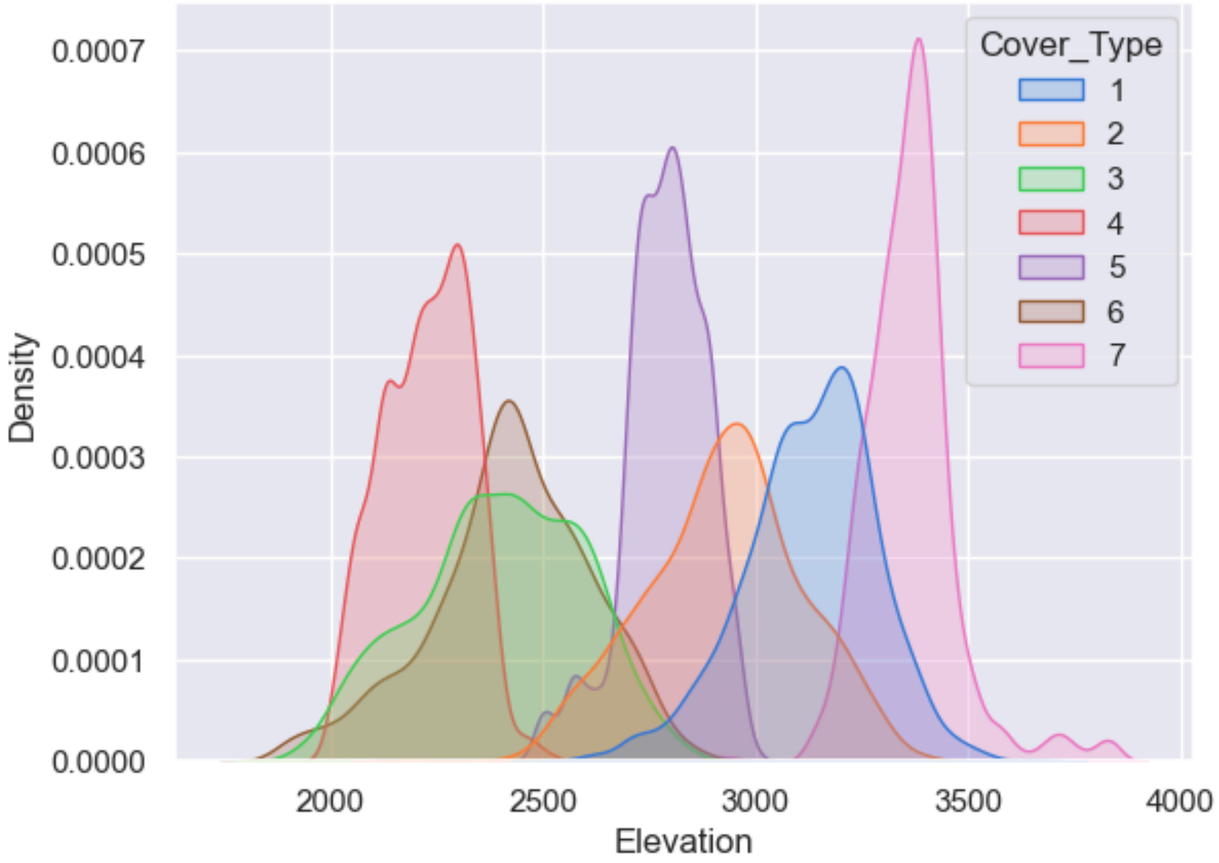
Heatmap



- The Dataset was Balanced but it had least Correlation among the features

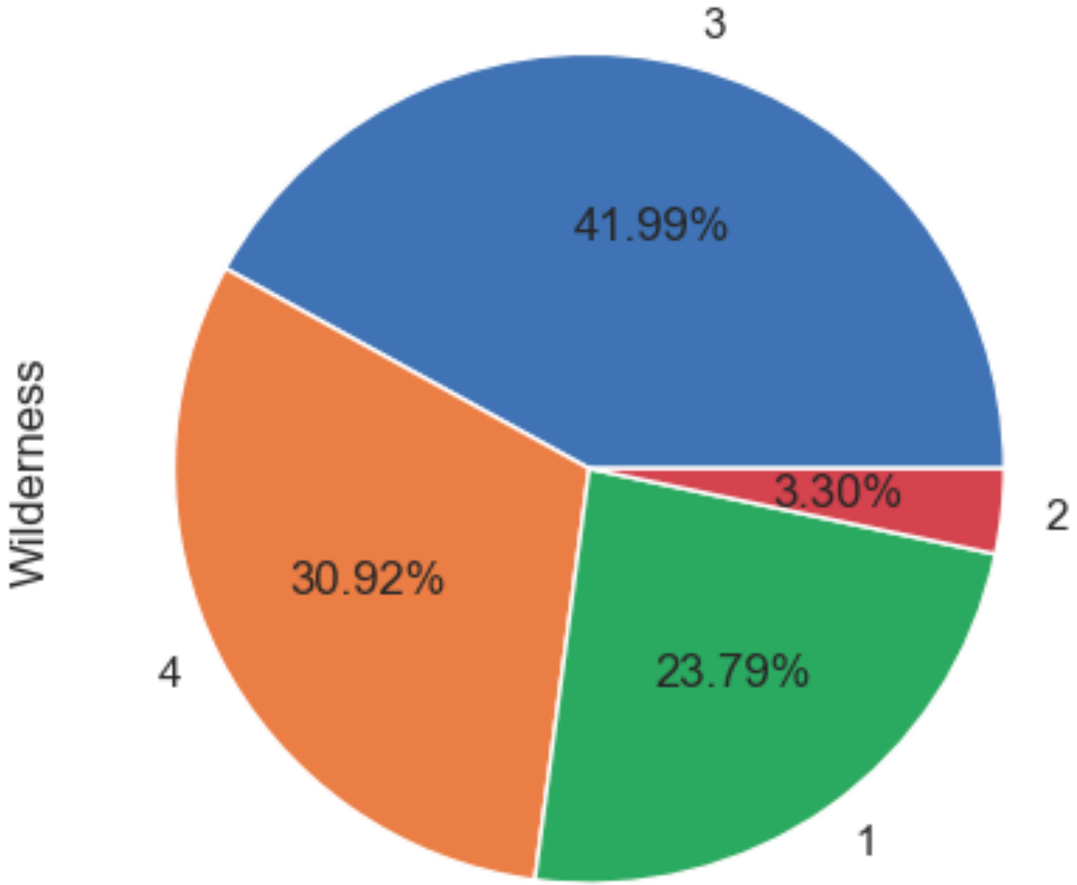
Cover Types Density for different Elevation Ranges

- *Type 7* Cover is densely populated in the Highest Elevation Range. (3100m-3500m approx.)
- *Type 4* Cover is densely populated in the Lowest Elevation Range. (2000m-2500m approx.)
- *Type 4*, *Type 5* & *Type 7* Covers almost never Overlap. They exist in separate Elevation Ranges.
- *Type 3* & Type 6 are populated at almost same Elevation Range.



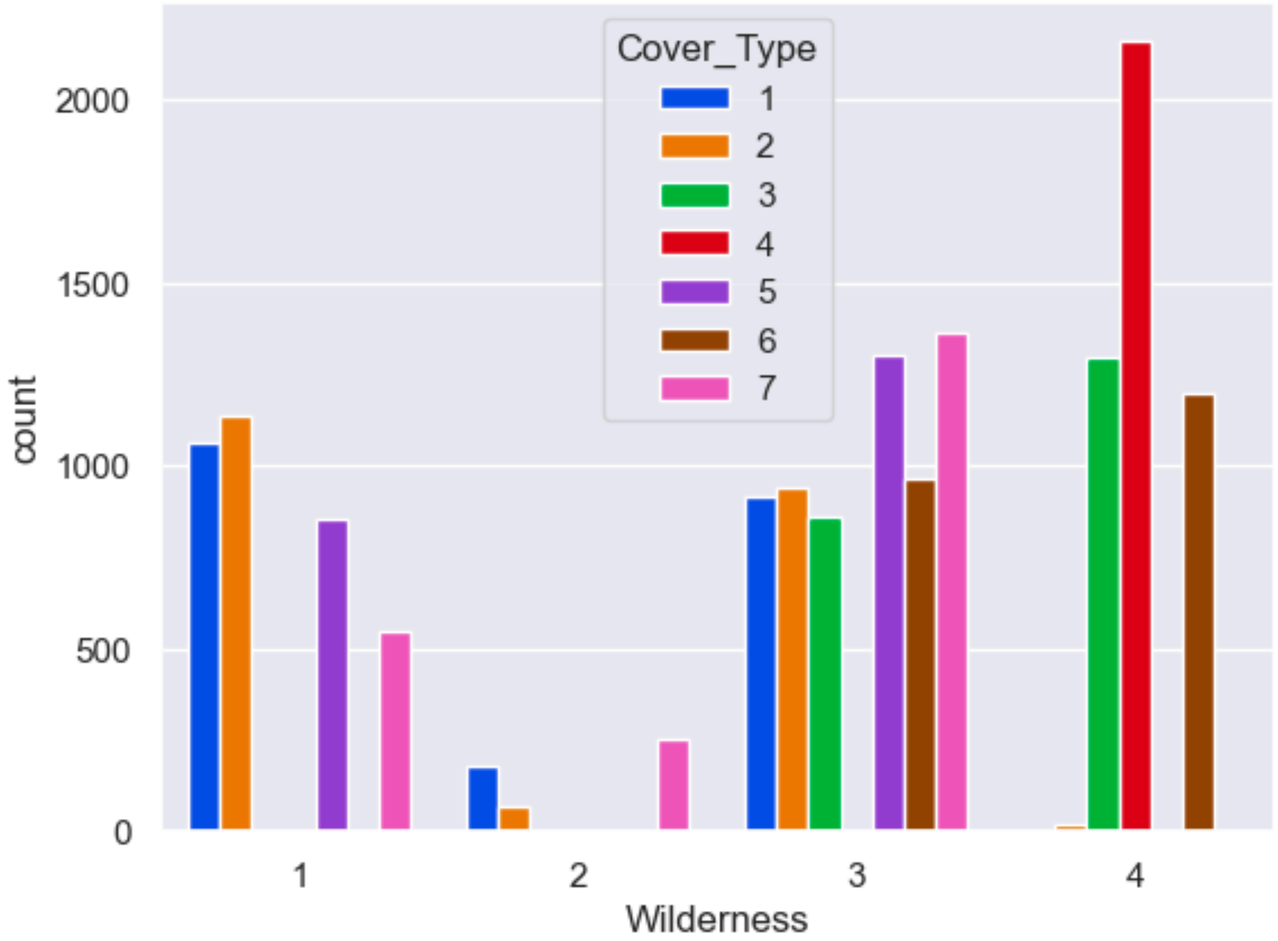
Wilderness Areas

- *Wilderness Area 3* had the highest presence in the Dataset. (42%)
- *Wilderness Area 2* had either very low representation in the Dataset or the Area does not occur frequently. (3.3%)



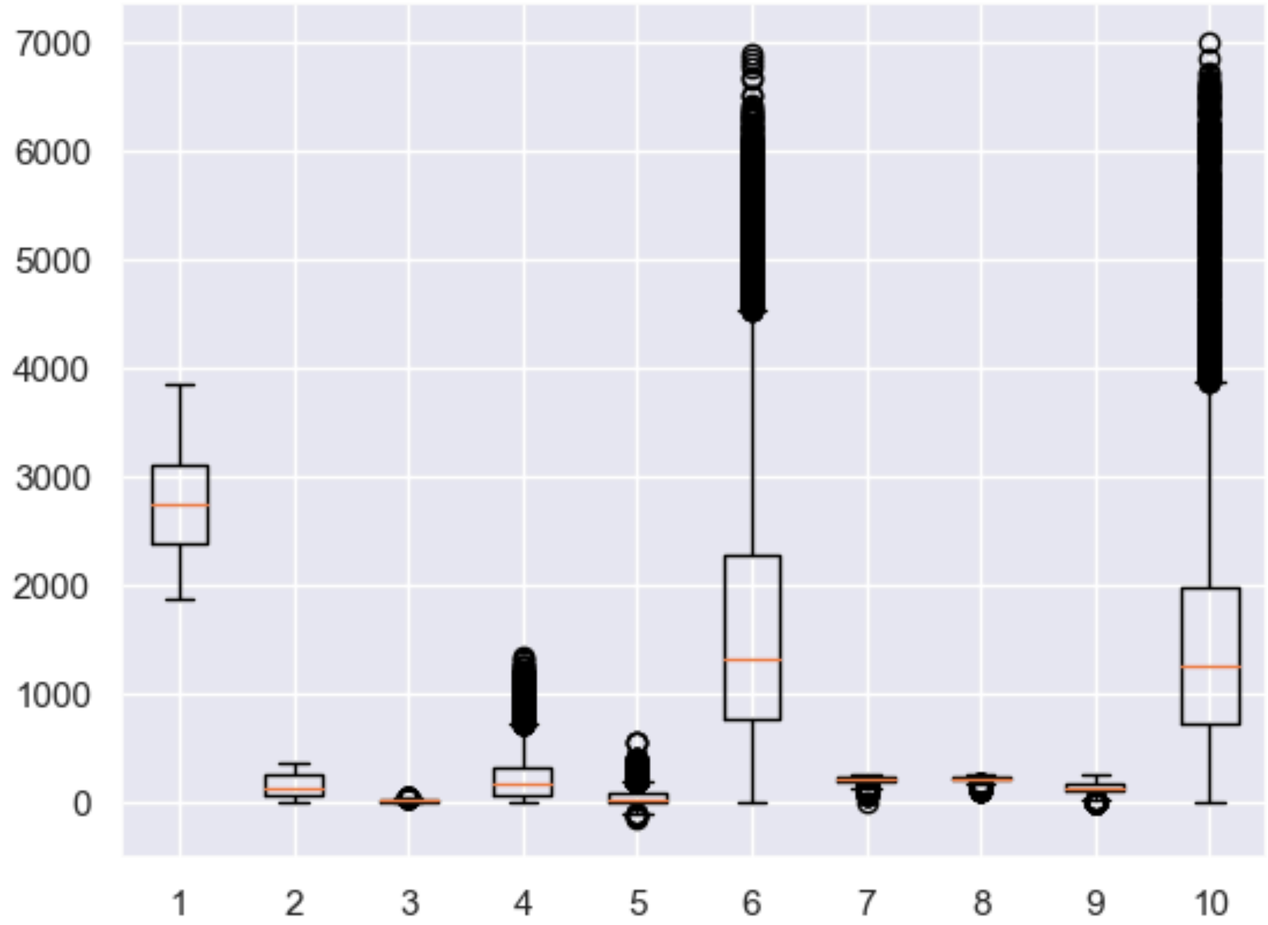
7 Cover Types across all Wilderness Areas

- *Wilderness Area 2* has only 3 Cover Types.
- *Wilderness Area 3* has 6 Cover Types that are almost balanced.
- Cover Type 4 exists almost entirely in *Wilderness Area 4*.
- Cover Types 3 and Cover Type 6 are only found in *Wilderness Area 3* & *Wilderness Area 4*.

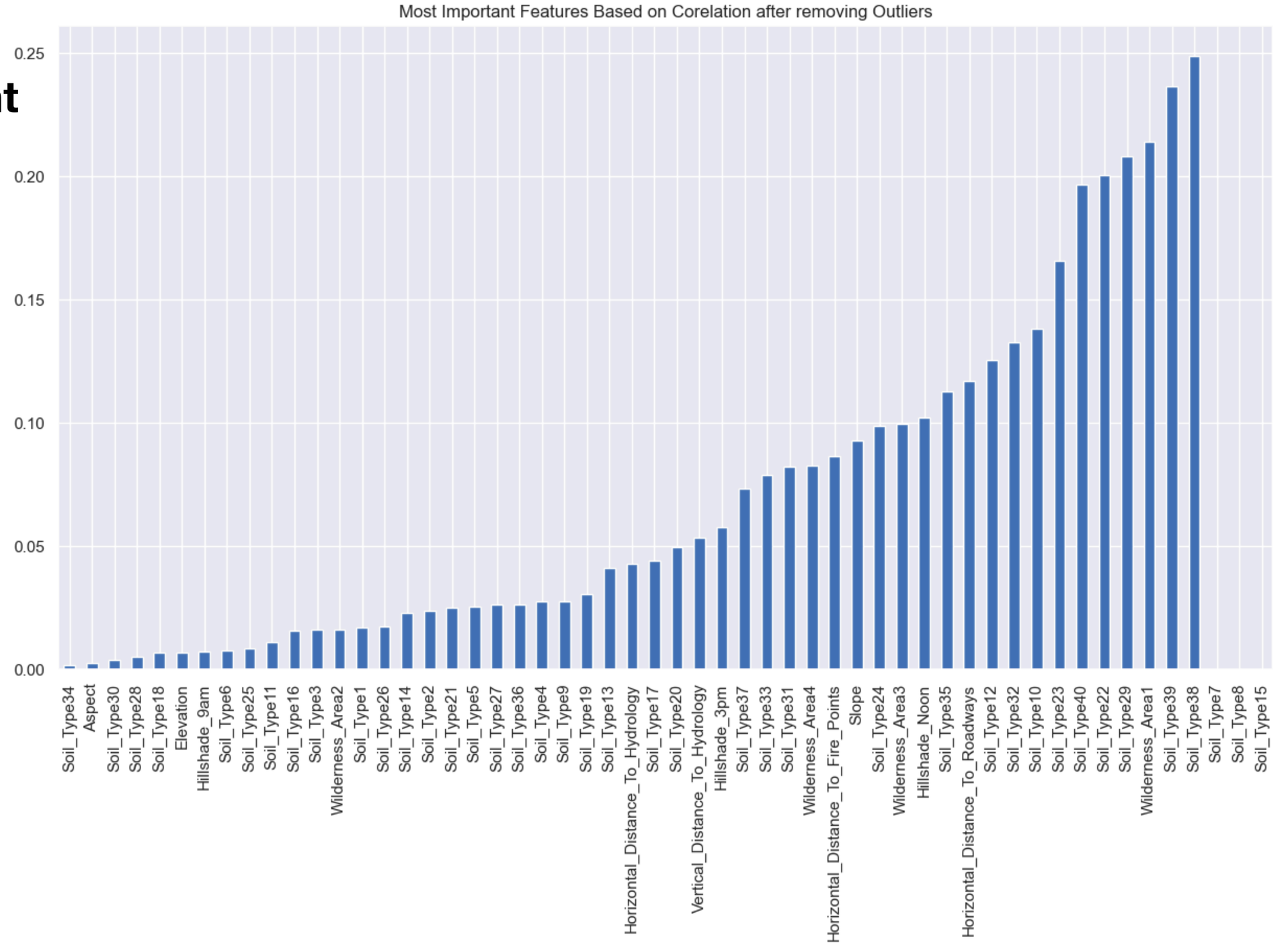


Outliers

- The Dataset consisted of 4.25% Outliers.
- *Elevation* and *Aspect* did not have any outliers.
- All other features had significant outliers.



Most Important Features



Model Selection

Data Selected: Pearson Correlation Selected Data. It performed well for most Classifiers.

Model Selected: Random Forest Classifier

Model Accuracy: 82%

Reasons for Selecting Random Forests:

- I selected this model because Random Forest does not always require the Data to be Scaled, since it is a tree based algorithm.
- Random Forest also does not enforce Normal Distribution and since our data was highly skewed and highly un-correlated, it becomes a better choice.
- Random Forest provided the best Accuracy among the tested Algorithms.

Model Comparison Report

Model	Accuracy on Raw Data	Accuracy on Scaled Data	Accuracy after Removing 5 Features (Pearson Correlation)
LinearSVC	0.519637	0.589673	0.337270
DecisionTreeClassifier	0.770393	0.727547	0.768470
LogisticRegression	0.663279	0.644328	0.673167
GaussianNB	0.581159	0.581434	0.585004
RandomForestClassifier	0.855809	0.825597	0.858830
GradientBoostingClassifier	0.793189	0.750893	0.795111
KNNeighborsClassifier	0.792365	0.724801	0.792365
Voting Classifier ‘Soft Voting’	0.815984	0.779181	0.817632
Voting Classifier ‘Hard Voting’	0.836034	0.797033	0.834660

Challenges Faced

- When Random Forest gave the highest Accuracy it came at a cost because the model was clearly overfitted. It had a Training Accuracy of 100%.

- In order to make a more generalised model, without losing any further training accuracy, I decided to prune the trees and further tune the Hyperparameters.

- The challenge was to get better accuracy while maintaining a Generalised Model. So I settled at **82%** Test Accuracy and 91% Training Accuracy, with 655 misclassifications. This was the best accuracy where the difference between Train & Test Accuracies was minimum.

- The model seemed to lose accuracy when trained on less features that are not correlated. Top 49 features out of 54 performed the best.

Conclusion

Forest Cover Prediction Dataset was both tricky and easy in many ways because it had All-Numeric data with most of the features One-Hot-Encoded and No-Missing values. But it also had very Skewed features with 4.25% Outliers(IQR) and Least Correlated Data.

With these datapoints, Random Forest Classifier achieved 82% Accuracy when trained on most important 49 features based on Pearson-r value based selection.