Predicting Forest Cover Type with Classification Models

Christen Ye, Mariana Chen & Anastasia Ivanova

I. Summary

Nowadays, one of the most important global concerns is deforestation. This makes any insights into forest monitoring very useful. Understanding different types of forest cover can aid in predicting forest ecosystem response to climate change and other disturbances. It is also useful in natural resource management, such as for timber harvesting, wildlife habitat management and recreation planning.

The objective of this project is to identify the classification model that best predicts the forest cover type. As such, we decided to use the Forest Cover Type dataset from Kaggle[1]. This dataset contains a variable called Cover_Type that denotes seven cover types in the Roosevelt National Forest of Northern Colorado, including Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, and Krummholz. Our final model uses 19 feature variables out of 54 to predict Cover Type. The chosen 19 predictor variables are based on the two-way table between Wilderness_Area and Cover_Type, as well as the two-table between Soil_Type and Cover_Type.

Out of all the classification models we applied including Classification Tree, Random Forest, Naive Bayes Classifier and KNN Classifier, Random Forest has the best performance among all folds because it yields the highest out-of-sample accuracy. In addition, based on 50% and 80% prediction intervals, Lodgepole Pine (Cover Type 2) is harder to discriminate.

[1] Data source: https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset

II. Data Characteristics

The dataset called “Forest Cover Type Dataset” is taken from Kaggle. The data was provided by the Remote Sensing and GIS program at Colorado State University. The dataset contains 581,012 observations of cover types from four wilderness areas located in the Roosevelt National Forest of northern Colorado.

The original dataset includes 54 features and 1 response variable. Of the 54 features, 10 are numeric variables and 44 are categorical variables. Of 40 categorical variables, 4 are wilderness areas and 40 are Soil Types. Table 1 gives detailed descriptions of each variable.

Table 1: Variable Descriptions

Variables	**Description**
Elevation	Elevation of the observed area in meters
Aspect	Aspect in degrees azimuth - compass direction that the surface is facing
Slope	Slope of the observed area in degrees
Horizontal distance to hydrology	Horizontal distance in meters to the nearest surface water features
Vertical distance to hydrology	Vertical distance in meters to the nearest surface water features
Horizontal distance to roadways	Horizontal distance in meters to the nearest roadway

Hillshade (9 am)	Hillshade index at 9am, summer solstice - how bright or dark the area is in the 9 am sun (0 darkest, 255 brightest)
Hillshade (noon)	Hillshade index at noon, summer solstice - how bright or dark the area is with the noon sun (0 darkest, 255 brightest)
Hillshade (3 pm)	Hillshade index at 3pm, summer solstice - how bright or dark the area is with 3 pm sun (0 darkest, 255 brightest)
Horizontal distance to fire points	Horizontal distance in meters to the nearest wildfire ignition points
Wilderness area (1 to 4)	4 binary columns for wilderness area designation (4 wilderness areas with "1" for present)
Soil type (1 to 40)	40 binary columns for soil type designation (40 soil types with "1" for present)
Cover type	Our response variable - one of the 7 forest cover designations (1 - Spruce/Fir, 2 - Lodgepole Pine, 3 - Ponderosa Pine, 4 - Cottonwood/Willow, 5 – Aspen, 6 - Douglas-fir, 7 - Krummholz)

Table 2 shows the frequency of each cover type in the original dataset. As seen, there exists a quite heavy class imbalance in the dataset among the forest cover types. Type 4 takes the smallest portion of measurements (0.5%) with 2,747 observations and Type 2 takes the largest portion of measurements (49%) with 283,301 observations. Therefore, we decided to randomly select 2,700 measurements from each cover type, which undersamples the majority class to avoid

class imbalance. Then we will confirm the best model with the original dataset that has a larger sample size.

Table 2: Cover Type Frequency

Cover Type	1	2	3	4	5	6	7
Count	211840	283301	35754	2747	9493	17367	20510
Proportion of dataset	36.5%	49.0%	6.0%	0.5%	1.5%	3.0%	3.5%

Having too many features could make a model become too complex, which overfits the training data and leads to poor generalization to the new data. Additionally, it can be difficult to interpret the relative importance of each feature in making predictions. Therefore, we wanted to remove some of the soil types and wilderness areas.

As seen from Table 3 below, the number of measurements in Wilderness_Area3 and Wilderness_Area4 take up more than 50% of all measurements so we decided to remove Wilderness_Area1 and Wilderness_Area2 as features.

Table 3: Two-Way Table for Wilderness_Area and Cover_Type

Cover_Type	1	2	3	4	5	6	7
Wilderness_Area							
1	1361	1320	0	0	1083	0	674
3	1107	1264	1105	0	1617	1183	1705
2	232	88	0	0	0	0	321
4	0	28	1595	2700	0	1517	0

As seen from Table 4 below, the total number of measurements in Soil_Type29, Soil_Type23, Soil_Type32, Soil_Type30, Soil_Type38, Soil_Type10, Soil_Type4 and Soil_Type3 take up more than 50% of all measurements, so we decided to remove the other Soil_Types as attributes.

Table 4: Two-Way Table for Soil_Type and Cover_Type

Soil_Type / Cover_Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	18	19	20
1	0	0	0	1	0	0	0	2	2	6	9	37	30	0	10	3	2	30	43
2	0	8	5	40	0	6	2	1	9	123	96	242	132	0	16	8	16	14	60
3	157	386	202	575	72	297	0	0	0	829	111	0	3	7	13	37	0	0	0
4	174	115	999	161	46	314	0	0	0	220	34	0	0	153	50	434	0	0	0
5	0	76	0	155	0	0	0	0	0	79	186	0	370	0	12	163	52	24	16
6	102	217	32	102	88	214	0	0	0	1376	79	0	97	58	42	112	0	0	46
7	0	0	0	15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Soil_Type / Cover_Type	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	14	327	465	141	1	5	3	0	557	91	156	287	224	2	13	0	0	87	88	64
2	0	70	184	81	4	13	3	14	647	183	140	308	242	16	0	0	0	6	4	7
3	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	203	15	0	37	0	5	299	624	84	146	146	8	0	0	0	0	0	0
6	0	0	3	22	0	0	0	0	0	0	8	23	78	1	0	0	0	0	0	0
7	1	19	94	35	0	0	4	0	115	25	24	112	68	8	108	9	35	799	751	477

Since most of the numeric variables are quite skewed (refer to Figure 1 in Appendix part), we have transformed the data targeting each skewness type. Additionally, the variables vary greatly in magnitude. This can result in larger-scaled variables dominating the others in models we fit. To fix this problem, we standardized all the numeric variables. We also transformed the aspect variable by converting it to radians and then taking a cosine. This

transformation helps better represent the circular nature of the variable. Variable transformations are shown in Table 5. The histogram of transformed numeric features is attached as Figure 2 in Appendix.

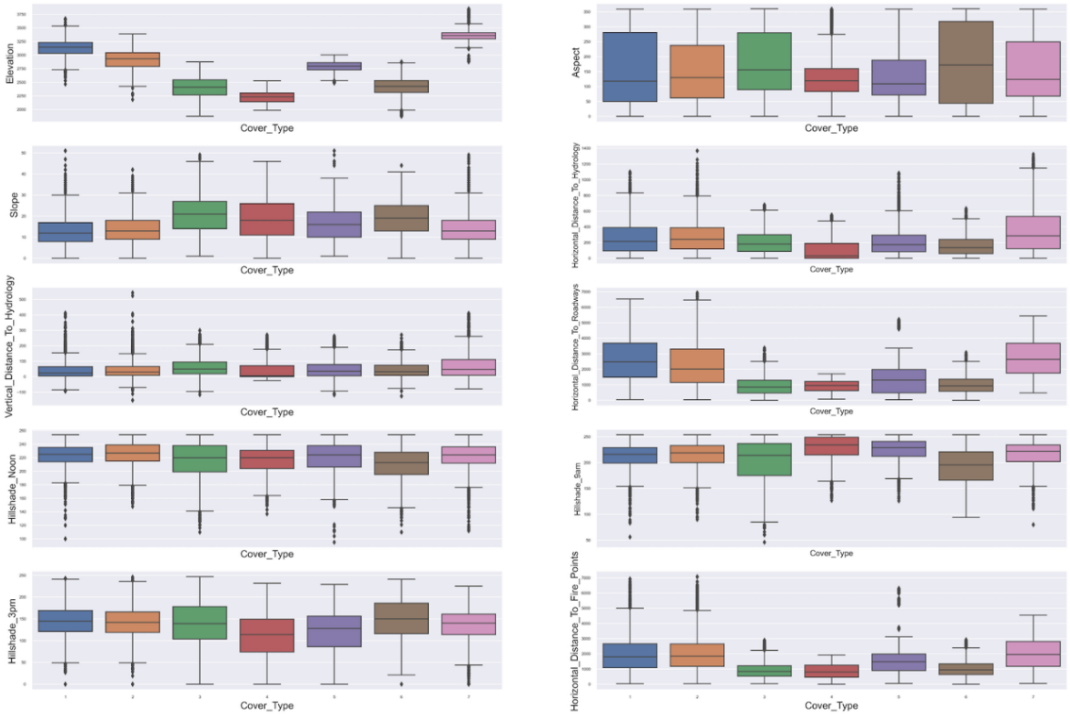
Table 5: Variables Transformation

Variables	**Transform**
Aspect	np.cos(np.radians(Aspect))
Vertical_Distance_To_Hydrology	np.log((Vertical_Distance_To_Hydrology +153) + 1)
Horizontal_Distance_To_Hydrology	np.log(Horizontal_Distance_To_Hydrolog + 1)
Horizontal_Distance_To_Roadways	np.log(Horizontal_Distance_To_Roadways + 1)
Horizontal_Distance_To_Fire_Points	np.log(Horizontal_Distance_To_Fire_Points + 1)
Hillshade_9am	*np.square(*Hillshade_9am*)*
Hillshade_Noon	*np.square(*Hillshade_Noon*)*

Figure 1 shows the side-by-side boxplot of numeric variables. As seen, each forest Cover Type appears to be significantly different in Elevation, HDTR[2], Hillshade_9am, HDTF [3]and HDTH[4]. This suggests that Elevation, HDTR, Hillshade_9am, HDTF and HDTD are probably important predictor variables since they have a stronger discriminating ability.

[2]HDTR stands for Horizontal_Distance_To_Roadways.
[3] HDTF stands for Horizontal_Distance_To_Fire_Points.
[4] HDTH stands for Horizontal_Distance_To_Hydrology.

Figure 1. Boxplot of Numeric Variables



III. Analysis and Interpretation

After transforming the variables, we first performed Classification Tree, Random Forest, Naïve Bayes Classifier and KNN Classifier with all variables. After that, we selected a subset of 10 variables to perform each model, based on the Gini importance score for Random Forest. Additionally, we noticed six selected variables worked better for Naïve Bayes than 10 selected variables so we then performed Naïve Bayes for a smaller subsets to compare model performance.

Table 6 provides an overview of the methods and Python libraries deployed in this analysis. Table 7 shows the 5-fold cross-validated accuracy comparison. For KNN Classifier, we

used RandomizedSearchCV and GridSearchCV to determine the best number of neighbors, which is 1.

Table 6. Methods and Python Libraries Deployed

Methods	**Python Libraries**
Classification Tree	sklearn.tree.DecisionTreeClassifier
Random Forest	sklearn.ensemble.RandomForestClassifier
Naïve Bayes Classifier	sklearn.naive_bayes.GaussianNB
KNN Classifier	sklearn.neighbors.KNeighborsClassifier

As shown in Table 7, Random Forest with all variables is best among all 5 folds, which has the highest average accuracy. It is also noticeable that Naïve Bayes Classifier with 6 variables including Elevation, HDTR, HDTF, HDTH, Hillshade_9am and Aspect is performing better than with all variables and with 10 variables. One of the assumptions for Naïve Bayes Classifier is the independence between predictor features. As shown in Figure 3 in Appendix, there is a strong correlation between Aspect and Hillshade_3pm as well as HDTH and VDTH[5]. After removing Hillshade_3pm and VDTH, the accuracy of Naïve Bayes Classifier has improved significantly. On the other hand, reducing variables did not help improve the performance of Classification Tree, Random Forest and KNN Classifier since these models don't have an independence assumption.

[5] VDTH stands for Vertical_Distance_To_Hydrology.

Table 7. 5-Fold Cross-Validated Accuracy

	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Classification Tree – full model	78.3%	78.4%	78.4%	78.8%	76.8%	78.1%
Classification Tree – 10 variables	77.4%	76.3%	75.7%	76.1%	76.9%	76.5%
Random Forest – full model (*)	86.0%	86.0%	86.0%	86.7%	85.6%	86.0%
Random Forest – 10 variables	84.2%	85.1%	84.9%	85.3%	84.4%	84.8%
Naïve Bayes Classifier – full model	45.1%	44.3%	43.9%	44.0%	44.0%	44.3%
Naïve Bayes Classifier – 10 variables	49.9%	50.3%	49.1%	50.1%	49.8%	49.8%
Naïve Bayes Classifier – 6 variables	62.7%	63.0%	63.2%	62.7%	63.1%	63.0%
KNN Classifier – full model	79.7%	79.9%	79.0%	80.4%	79.6%	79.7%

KNN Classifier – 10 variables	77.2%	78.6%	78.0%	78.1%	77.8%	77.9%

As shown in Table 8 below, Forest Cover Type 1 and 2 have lower accuracy across all models, while Forest Cover Type 4, 5 and 7 have higher accuracy across all models. This implies that it is harder to discriminate between Cover Type 1 and 2. In Appendix, Figure 4 demonstrates the image of each Cover Type. Based on research, Spruce/Fir (Cover Type 1) and Lodgepole Pine (Cover Type 2) are both coniferous trees that belong to the Pinaceae family and they both have needle-like leaves. Also, they both produce cones and they can often be found growing together in mixed coniferous forests. As such, it makes sense that Cover Type 1 and 2 have lower accuracy and it's harder to discriminate them.

Table 8. Out-of-sample Accuracy for 50% and 80% PIs

	Out-of-Sample	**Out-of-Sample**	**Out-of-sample**	**Out-of-sample**	**Out-of-sample**	**Out-of-sample**	**Out-of-sample**
Method	Accuracy on type 1	Accuracy on type 2	Accuracy on type 3	Accuracy on type 4	Accuracy on type 5	Accuracy on type 6	Accuracy on type 7
Classification Tree 50% PI	62.5%	12.0%	74.5%	81.2%	81.7%	76.6%	94.8%
Classification Tree 80% PI	92.5%	93.8%	93.9%	81.2%	86.4%	76.6%	94.8%
Random Forest 50% PI	80.2%	83.3%	83.5%	99.2%	96.7%	91.4%	97.5%

Random Forest 80% PI	96.5%	94.4%	95.9%	99.4%	99.8%	97.5%	99.9%
Naïve Bayes Classifier (6 variables) 50% PI	59.6%	42.4%	51.0%	87.5%	84.1%	73.4%	91.6%
Naïve Bayes Classifier (6 variables) 80% PI	87.3%	80.3%	84.5%	95.1%	93.1%	87.8%	97.6%
KNN Classifier 50% PI	61.0%	55.2%	74.8%	94.7%	90.9%	81.6%	92.7%
KNN Classifier 80% PI	61.0%	55.2%	74.8%	94.7%	90.9%	81.6%	92.7%

Based on the Gini importance scores from Random Forest, Table 9 shows that Elevation, HDTR, HDTF and HDTH are the most important features in predicting Cover Type. On the other hand, Soil Types and Wilderness Areas are less useful to distinguish different Cover Types.

Table 9. Gini Importance based on Random Forest

Feature	***Gini Importance***	***Feature***	***Gini Importance***
Elevation	3127.5	Slope	390.8

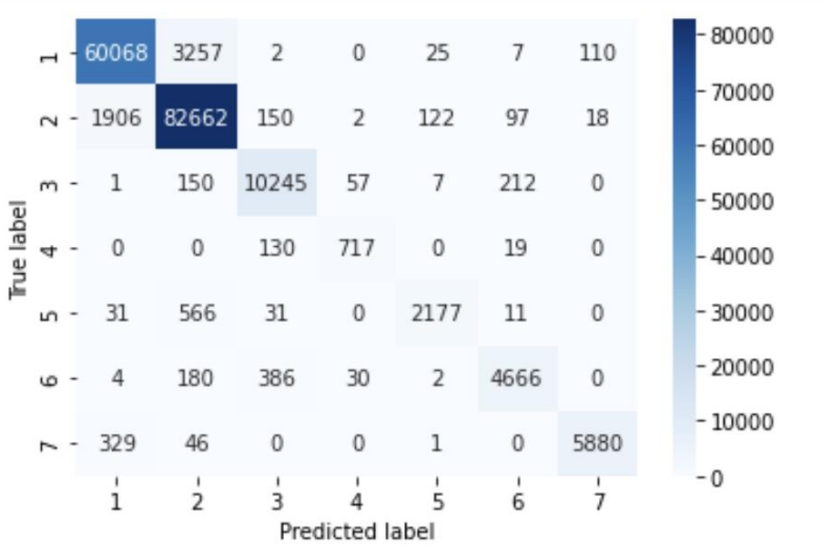
Horizontal_Distance_To_Roadways	1122.0	Soil_Type10	274.5
Horizontal_Distance_To_Fire_Points	888.8	Wilderness_Area3	248.0
Horizontal_Distance_To_Hydrology	748.2	Soil_Type3	215.5
Vertical_Distance_To_Hydrology	615.1	Soil_Type4	145.6
Aspect	570.3	Soil_Type30	109.6
Hillshade_9am	568.8	Soil_Type29	75.1
Hillshade_3pm	485.6	Soil_Type32	65.7
Wilderness_Area4	592.9		

From the preliminary analysis above, Random Forest with all features is the best model to predict Cover Type since it has the highest cross-fold accuracy. As for the next step, we wanted to validate this result using the full dataset. Due to the class imbalance in the full dataset, we used Weighted Random Forest to avoid bias towards larger classes. The weights are inversely proportional to the number of observations in the class.

As seen from Figure 2, the out-of-sample prediction accuracy on the full dataset when fitting a weighted random forest model is 94%. Again, we can see that cover types 1 and 2 are

harder to discriminate. As for the out-of-sample accuracy for 50% and 80% prediction intervals, it was 95.8% and 99.5% respectively.

Figure 2. Out-of-Sample Confusion Matrix



IV. Conclusion

Overall, Random Forest model with all features is the best classification model to predict Cover Types. This can be concluded from reasonably high 5-fold cross-validated accuracy and out-of-sample accuracy based on 50% and 80% prediction intervals.

Furthermore, Weighted random forest model performed even better on the full dataset, with 94% out-of-sample prediction accuracy. Based on the Random Forest feature importance, Elevation, HDTR, HDTF and HDTH are the most useful features in predicting Cover Type. The model is confusing Cover Type 1 and 2 which is expected because these types belong to the coniferous tree group and live in similar conditions.

V. Contributions

Our team was formed via Piazza. Most of our discussions took place on Facebook, we also held few in-person meetings to enhance team collaboration.

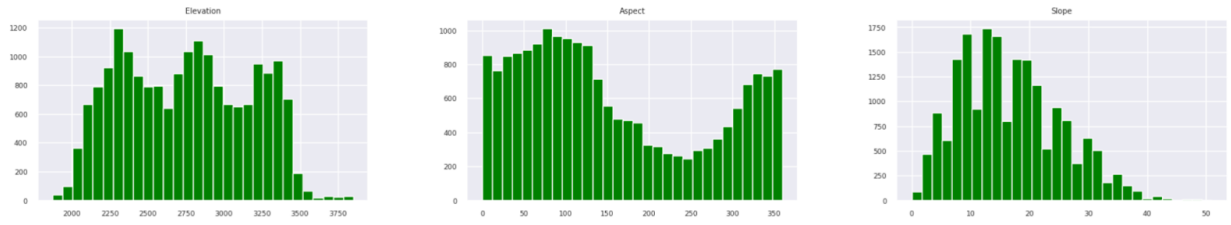
Chen, Changling: Prepared presentation 1 and presentation 2 slides. Found the dataset and came up with the ideas of our project. Contributed to the code for exploratory data analysis, 5-fold cross-validation performance for all methods and out-of-sample performance based on 50% and 80% prediction intervals. She also contributed to the report writing.

Ye, Christen: Code optimization, quality check, contributed to report writing and formatting, contributed to presentation 1 and 2 slides.

Ivanova, Anastasia: Exploratory data analysis and data transformation. Model validation on the full dataset and report writing.

VI. Appendix

Figure 1. Histogram of Original Numeric Variables



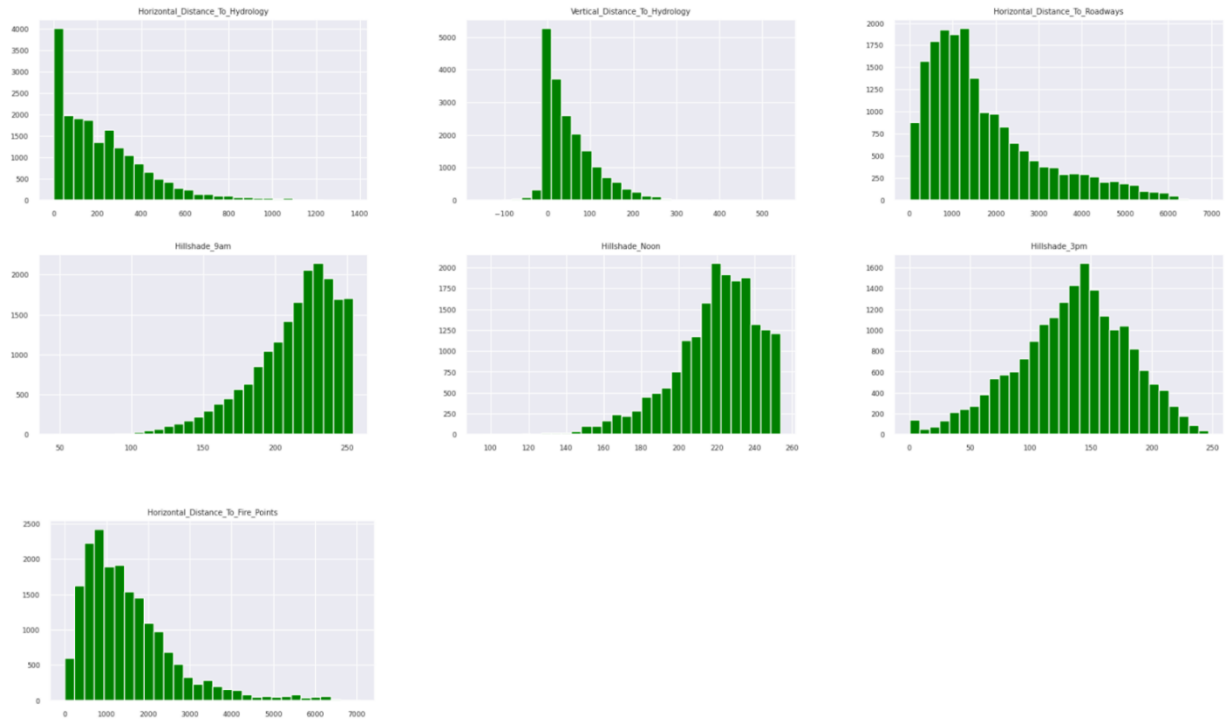
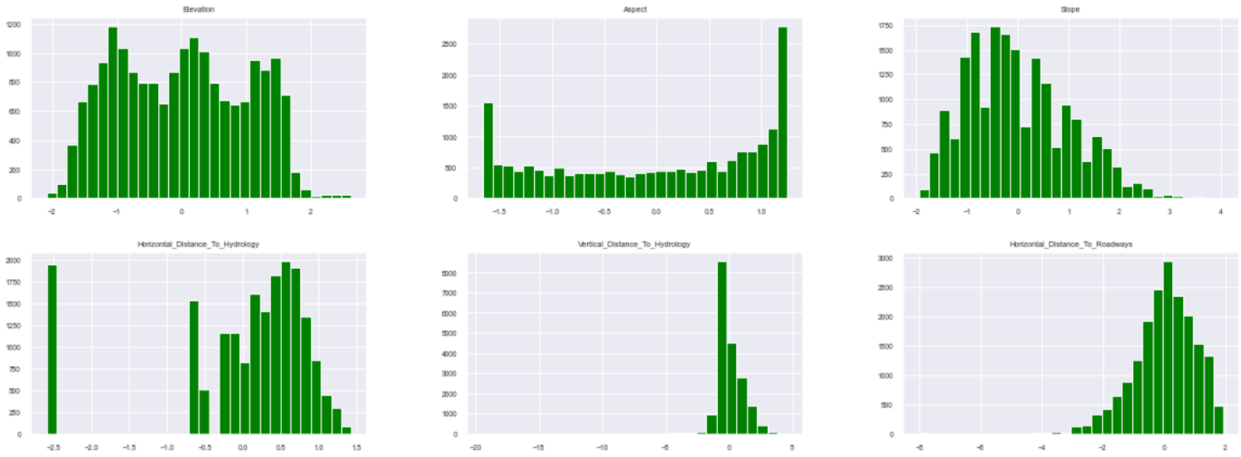


Figure 2. Histogram of Transformed Numeric Variables



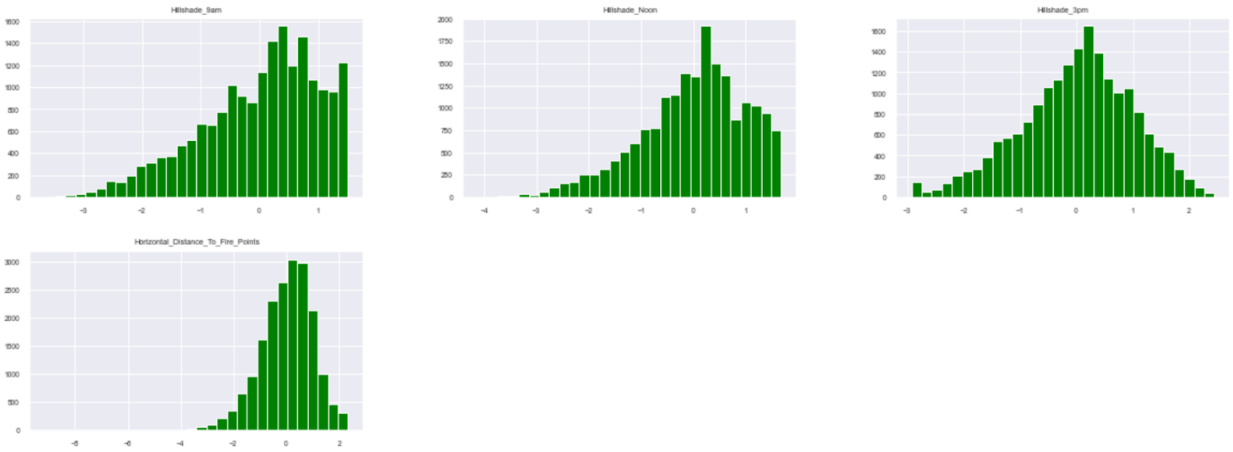


Figure 3. Correlation Plot

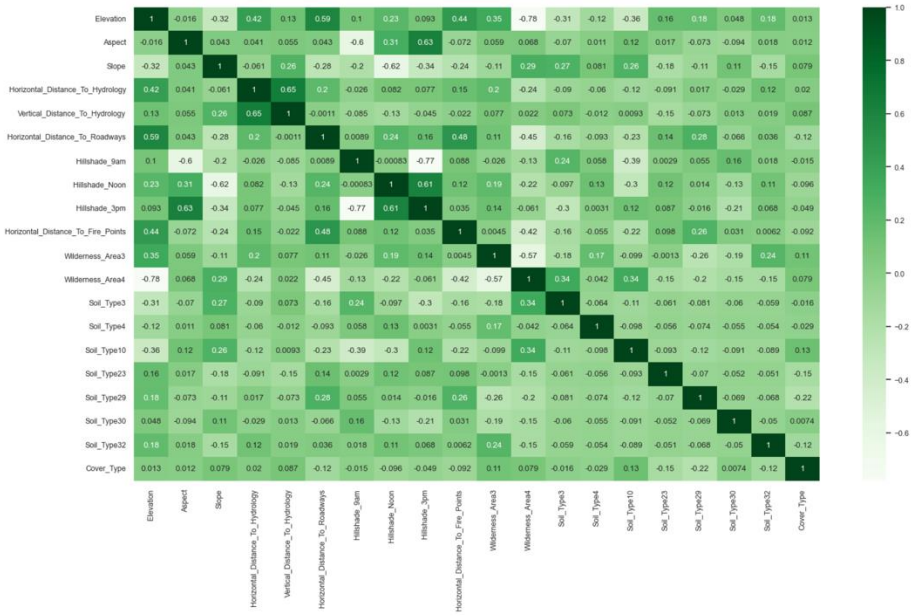


Figure 4. Cover Type

Cover Type 1: Spruce/Fir

Cover Type 2: Lodgepole Pine

Cover Type 3: Ponderosa Pine

Cover Type 4: Cottonwood/Willow

Cover Type 5: Aspen

Cover Type 6: Douglas-fir

Cover Type 7: Krummholz