Business Report for Machine Learning Group Assignment

Masters in Big Data & Business Analytics

Course Name:

MACHINE LEARNING II

Supervised by:

Prof. DANIEL GARCIA HERNANDEZ

Submitted by:

Manuel Marina Herrera

1. Goal

The goal of this report is to solve the Forest Cover Type Classification Problem. The assignment involved the prediction of the Forest Cover Type (the predominant kind of tree cover), using cartographic variables. The study area was situated in the Roosevelt National Forest of northern Colorado and contains four wilderness areas. This integer classification problem consisted of seven Forest Cover Types (target variable), which were Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz.

The dataset contains 15.120 observations (training set) with the features and the target variable, while the test set contains only the features. Each observation corresponds with a 30x30 metres patch of the forest.

Findings:
The final model selected for the predictions was a XGBoost model, which generated the highest accuracy, recall and precision compared to the different models tested to predict the cover type. The model was very successful to predict the cover types. The most significant features used to predict were those related to area types, soil types in addition to all the features related to elevation.
The final model's prediction accuracy against the test was 89.2%, against the kaggle test set was 79,1%

2. Data Exploration and Explanation

The Dataset contains 56 features, that can be divided in 5 dimensions:

- Identification fields : It consist of a unique identification number for each row
- Numeric Variables: It consists of features with numeric type that contains information about the characteristics of the 30x30 patch such as Elevation, Distance to water, Slope, etc.
- Area: There are four areas in the forest and each patch belongs to one. This feature is categorical hence transformed to oneHot encoding.
- Soil: There are 40 soil types overall, and further OneHotEncoded.
- Target: The target variable that is Cover Type

For the data exploration part the execution of a pairplot with Seaborn helped to have an overview of the marginal distributions of the numerical variables (Data Visualisation 2) and as well to have a look at relationships between the variables. Above this, a Correlation Matrix has been applied as well (Data Visualisation 4). In the following section every feature will be explained more in detail.

The marginal distribution of the feature Elevation shows 3 clear peaks. Those peaks show that there are 3 different heights within the forest. In this way, this feature looks as a good feature for our model because it differentiates the Cover Types (Data Visualisation 1). This constatation could also be retrieved from the violin plot that plots the target feature Cover Type against the explanatory variables.

The strange distribution for the feature Aspect can be explained by the fact that values are ranging between 0 and 360 and they represent the direction. Here, the majority of the slopes are pointed

towards the north or east. The second peak could be cut and pasted before the first one and then it was identified as a normal distribution.

The features Hill shades for 9am, noon and 3pm differentiate the shadow of the hill depending on the time of the day.

The variable Slope seems to have a normal distribution with values ranging between 0 and 50. On the contrary, when Slope and Aspect are plotted against the three different hillshades, they show an odd relationship. This can be interpreted by the fact that when Slope is zero, according to the hill shade the scatter plot looks differently because the values are depending on the presence or absence of the sun. The line where Slope is zero can be pulled through to understand the feature Aspect in the relationship with hill shade and in this way, you can observe where the sun is present or not.

40 Soil Types and the 4 Wilderness Areas features have the appearance of being significant as well for differentiating the Cover Types. The 4 Areas consist of Rawah, Neota, Comanche Peak and Cache la Poudre Wilderness Area. A violin plot showing Area against some features proves that Elevation can explain the difference between the different Areas, while the other distributions look rather similar. When Area is plotted against Cover Type, it can be observed that several Cover Types only occur in some of the Areas (Data Visualisation 6). The 40 Soil Types are all named by families, some of them appear several times with different characteristics. The families can be differentiated by the kind of soil it has for example stony and rubbly and those traits occur in different degrees. To understand in general the Soil and Area columns heatmaps had been used with Seaborn to get an overview of their relationship with Cover Type.

Finally, there are some distance-related features as well that represent the horizontal or vertical distance to hydrology, roadways or fire points.

3. Data Cleaning

With the execution of the option shape on the dataset, it was immediately identified that there were no null values involved in the dataset. In this way, it was also observed that Soil Type 15 and Soil Type 7 do not have any values but they are included in the dataset (they are present in the test set). To facilitate later use of the features, it was divided into four different dimensions as mentioned before. The first dimension is the target variable which is Cover Type, the second is four Wilderness Areas, third is with the Soil Types and the last one with the numerical features.

4. Baseline Model

Before moving on to the feature engineering part of the analysis, first a baseline model needs to be created. This model will test the accuracy of the dataset before adding, transforming, or selecting the best feature to use. Later, different models using different machine learning algorithms will be created, their results will be compared to the baseline model. The model indicating the highest improvement in accuracy for predicting the target variable will be selected as the final model.

The baseline model selected was Decision Tree Classifier, a model effective for large amounts of data that needs classification for categorical attributes. Some of the benefits of this model are that it requires minimal data adjustments or normalisation to be able to predict the results. This model does not necessitate any parameter setting to predict the target variable. Some of the disadvantages of this model must be noted, Decision Tree Classifier is not good for dealing with continuous variables, and overfitting can also be a problem. Therefore, feature engineering will be required to ensure only the best features will be selected in the final prediction model.

The baseline model accuracy score was 71%.

5. Outliers Handling

Outliers are the data points that are isolated from the remainder of the points. Detecting outliers for the numeric variables is one of the first steps in data manipulation (Data Visualisation 3). There are several techniques that can be used in handling outliers. For this model, three different outlier detection methods were used:

- **Isolation Forest:** an isolation method used to detect the outliers
- **Elliptic Envelope:** this method can identify the outliers by creating a boundary that restricts the shape of the data points that are relevant
- **Local Outlier Factor:** this method identifies the outliers by measuring the deviation of each point in respect to its neighbours.

To test the impact of the outliers on the model, the baseline model was rerun after taking out the outliers detected using each of the three methods. Based on the accuracy results of each model after removing the outliers, there was no significant improvement in the models accuracy, in two of the methods used it had a negative impact on the accuracy score. For this reason, the outliers will be kept in the model and will not be removed.

6. Feature Creation

Feature creation is the process of creating new features using the existing data in the model. The newly created features will then be used to train the different machine learning models. This step is very important since the algorithms used can only be improved based on the data provided to it. Generating new appropriate features and variables is crucial to improve the accuracy of the model.

The two main types of feature creations that can be done are transformations and aggregations. Feature creation using transformations are done using at least one of the attributes available in the data. Aggregated features creations are done using the data available in the dataset in addition to other data which can be used to explain the data at hand.

The features created to improve this model are the following:

- **Area Number:** refers to the relationship between the elevation and the horizontal distance to roadways
- **Area & Soil:** since both area and soil type are believed to be significant drivers for predicting cover type, creating categorical variable linking both attributes could possibly improve the model
- **Target Soil:** this feature was added based on additional information available regarding the dataset which enables to group the soil types into 5 different groups instead of the initial 40 types.
- **Soil Families:** this feature was created to better explain the family for which each soil type is associated with
- **Hill shade name & slope sun:** are two features created driven from the slope attribute and the hill share attribute. These new categorical attributes will help us understand how the hill shade and the sun impact the cover types.
- **Distance to hydrology** is another feature created based on the numerical variables horizontal and vertical distance to hydrology. This feature is the hypotenuse, creating a direct connection between the location of the tree and the water.
- **Cardinal & cardinal numbers** are categorical features which were created based on the numeric attribute aspect. This feature groups the data at hand into six different categories related to the direction at which the sun faces the trees. The main difference between them one is label with a name, the other is the label with a number
- **Absolute Vertical Distance to Hydrology** is a numerical feature created to indicate the absolute distance between each tree and the water. This feature converts the negative values, those below the sea level to a positive number without making any additional changes
- **Above Below water** is a feature which indicates whether the location of each tree is located above or below sea level. This categorical variable was created based on the vertical distance to water attribute.
- **Other numerical features** additional numerical features were created driven from the elevation, distances to hydrology, means of distances between different numerical attributes were created.

After the creation of the additional features, the numerical features were checked for correlations using a confusion matrix. Following this step, the groups of correlated variables were put in a principal component analysis method. This process computes the principal components and uses them to reduce the high correlation between the different attributes by creating new attributes.

After testing this method, it did not improve the prediction results and was later neglected.

7. Feature Selection

Before running the different models to predict the cover type, feature selection needs to be conducted. This is the process of selecting the most relevant attributes which will then be used to predict. Different feature selection methods and feature validation criteria were used to select the most relevant and significant attributes to improve the models predictions.

Using the Gini Importance method of Random Forest Classifier which calculates each feature's importance by dividing the total number each feature is applied over the number of splits. This method will enable the ranking and identification of the most relevant attributes by selecting the attributes with the highest Gini Important score.

The sklearn method used was Select from Model, that chooses the features that have a higher importance than a given threshold. With a cross validation strategy various cut off points were tried and the best threshold was features that had an importance above 0.9 times the mean. Since random forest is used as the feature selection algorithm, only the ordinal encoding and not the one hot encoded column will be used in this method.

The resulting columns where: Elevation, Horizontal_Distance_To_Roadways, Hillshade_9am, Horizontal_Distance_To_Fire_Points, Area, Soil, Area_soil, Distance_To_Hydrology, Elev_to_Horizontal_Hyd, Elev_to_Horizontal_Road, Elev_to_Verticle_Hyd, Mean_Horizontal_Dist, Mean_Fire_Hydro.

The feature validation process is then conducted. The comparison of the accuracy of the model prior to and after the feature engineering is computed to test the impact of the new features models. The comparison was done with a Random Forest with a baseline value of 84,7%

The final list of attributes selected in the model have indicated that there is a significant improvement increasing the accuracy of the model to 86.9%. The total number of attributes that will be used to predict the target variable is 13, of which 11 attributes were creation or transformations done in the feature engineering part.

8. Model Selection

Different Machine Learning algorithms are used depending on the type of information and the target variable (continuous or categorical). As this is a classification problem where the categorical variable “Cover_Type” needs to be predicted, the following models have been considered to be executed on the data and further hyper-parameter tuning can be performed to improve the accuracy of the model.

- Logit Regression
- SVM
- Random Forest
- XGBoost

Data Preparation for Modelling:

To test the performance of the model, the data was split between train and test in the 80:20 ratio. However, as few of the above models are sensitive towards the distance (SVM, Logit Regression, etc), it is necessary to transform the features on the same scale. The data has few categorical features which have been previously transformed using OneHotEncoder () and lie within 0 and 1. Therefore, it is necessary to bring continuous variables on the same range. Hence, it was suitable to use MinMaxScaler() that helps data to scale within 0 & 1.

9. Model Selection and Hyperparameter Tuning

After the data preparation phase, above-mentioned models were executed to compare the relative value of different statistical models and determine which one is the best fit for the observed data.

1. Logit Regression

Logit regression helps to predict the probability of a certain class or event taking place. But this model can also be used when there are multiple classes in the given dataset. To handle this, the following two parameters have been passed to the LogisticRegression() model: 1. multi_class=ovr 2. solver=saga (as ovr is supported by saga solver). Along with this, additional parameters like random_state and max_iter are passed to the model. To get the best hyper-parameters, GridSearchCV() was used for parameters like C, penalty with different values. With this model, max accuracy of “70.0%” was achieved.

2. SVM

Support vector machines(SVMs) is a supervised learning method used for classification. As SVM as well predicts from binary classes, it is important to pass special parameters to handle such situations. In the built model, decision_function_shape = ovo was introduced in the model for multiclass strategy. In this model too , GridSearchCV() was used for parameters like C, gamma, kernel with different values. With this model, max accuracy of “75.1%” was achieved.

3. Random Forest

Random forest is a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. and It is a very robust technique for classification as it greatly boosts the performance of the final model. In this model as well, GridSearchCV() was used for parameters like max_depth, n_estimators, min_samples_leaf with different values. With this model, max accuracy of “87.2%” was achieved.

4. XGBoost

XGBoost is another very robust algorithm for classification problems. It combines the estimates of a set of simpler, weaker models. In this model as well, GridSearchCV() was used for parameters like max_depth, n_estimators, learning_rate , subsample with different values. Along with this, additional parameters like objective = multi:softmax, num_class, eval_metric are passed to the model. With this model, max accuracy of “87.4%” was achieved.

Best Parameters

The best parameters found in the grid search for each model where:

Model	Score	Best Parameters
Logistic Regression	0.700893	C: 1, multi_class: multinomial, penalty: none, solver: saga
Support Vector Machine	0.751406	C: 30, decision_function_shape: ovo, gamma: 0.1, kernel: rbf
Random Forest	0.872272	max_depth: 30, min_samples_leaf: 1, n_estimators: 1000, random_state: 42
XGBoost	0.874587	learning_rate: 0.01, max_depth: 30, n_estimators: 2000, subsample: 0.66, tree_method: hist

10. Final Model

The models that performed above 80% accuracy(XGBoost, RandomForest) were selected to perform the final validation. Both models were fitted with the training set, and predicted the unseen test set.

The results were 89,1% for Random Forest and 89,2% for the XGBoost. Although both models performed similarly, the XGBost performed better in the Kaggle's test set when fitted with all the data and was our final model.

The results of the XGBoost in our test data set was Accuracy: 89,2% , Recall: 89,2% and Precision 89,1%. Overall our model is very good, the weakest point is the prediction between class 0 and 1. Because those 2 types have very similar values in almost every variable in our dataset. For the Kaggle test set the score was 79,12%.

Finally the variables that influenced the model the most where from 2 categories:

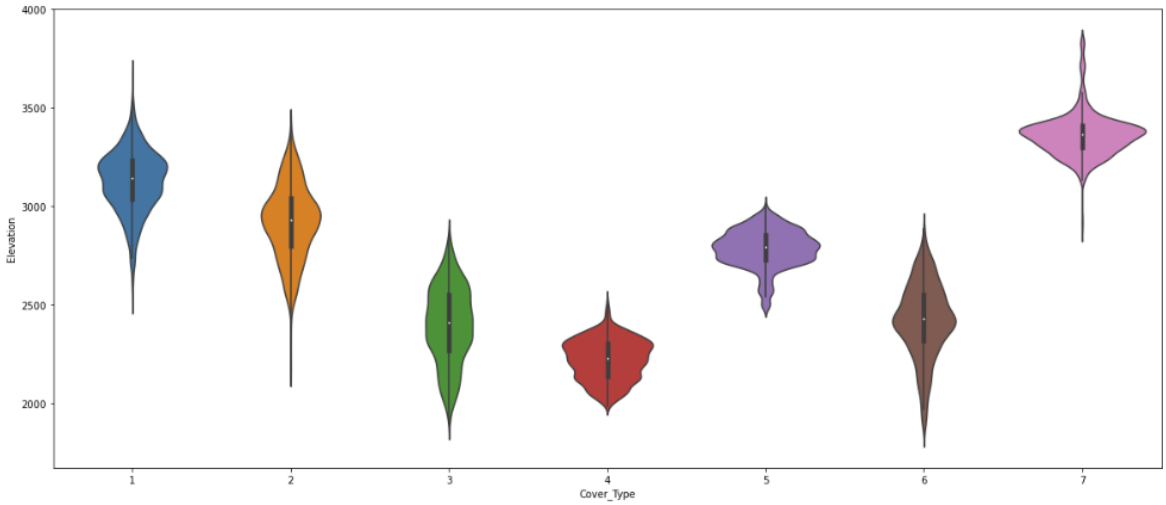
Area and soil related variables: For predicting the cover type the area, soil and most importantly the combination of both really helped the model to differentiate between classes

Elevation related variables: As mentioned before, elevation was key to differentiate some of the cover types as can be shown in the visualisation 1. Also, other calculations containing elevation were influential.

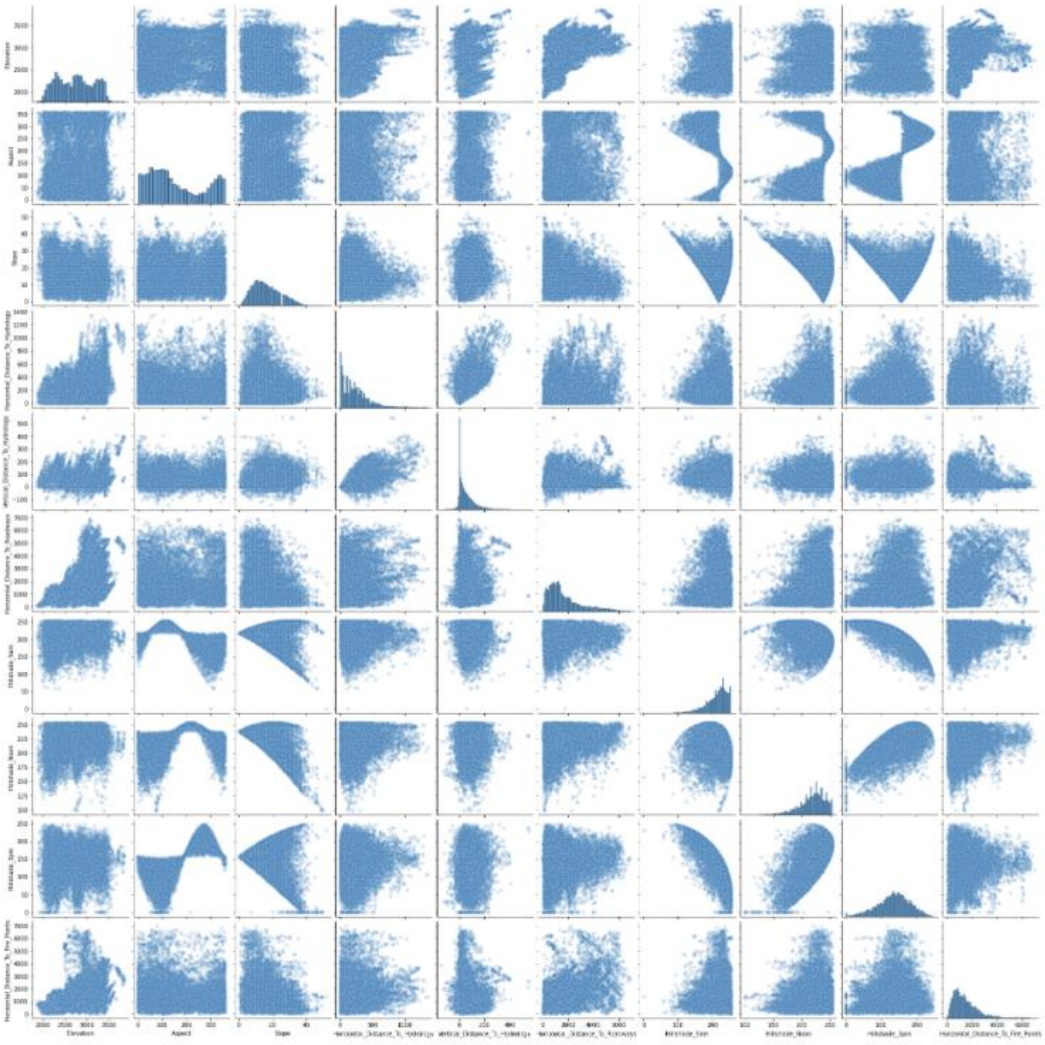
Those most important variables were plotted in a horizontal bar chart against their features importance (Data Visualisation 5) as well as the Confusion Matrix (Data Visualisation 7).

11. Data Visualisation

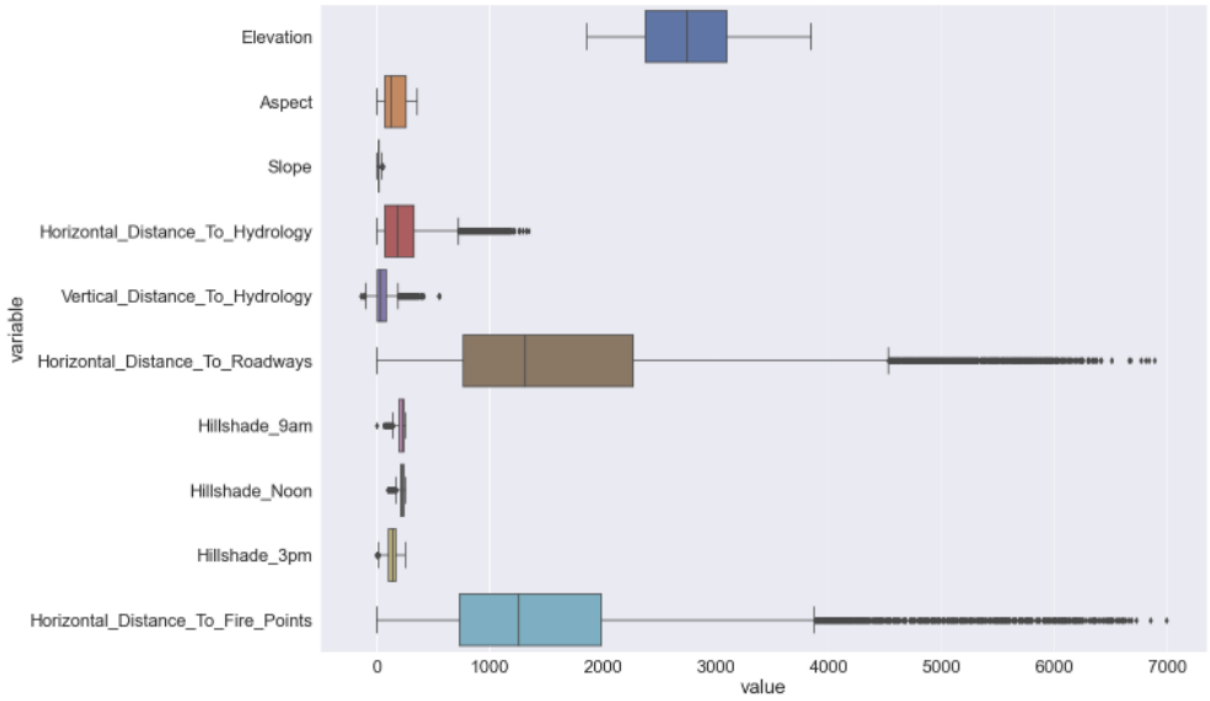
1. **Elevation vs cover type**



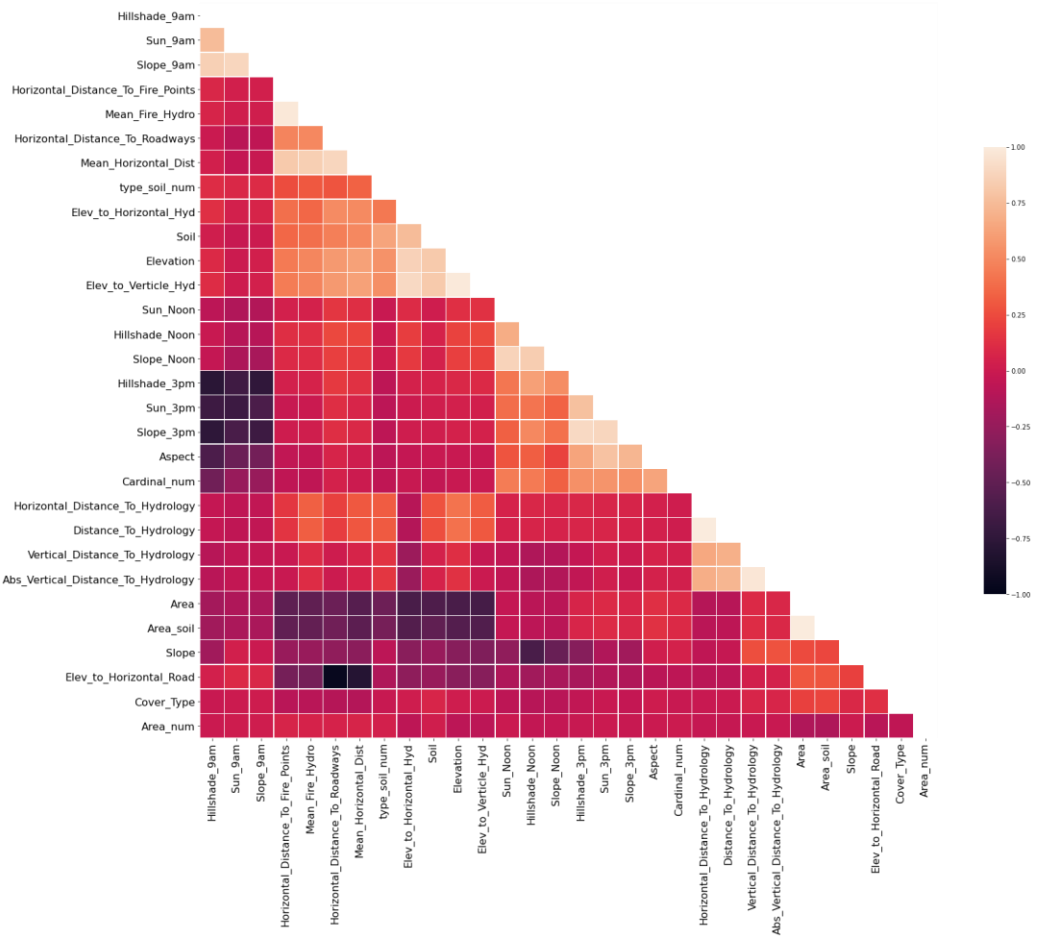
2. Pairpot numeric variables



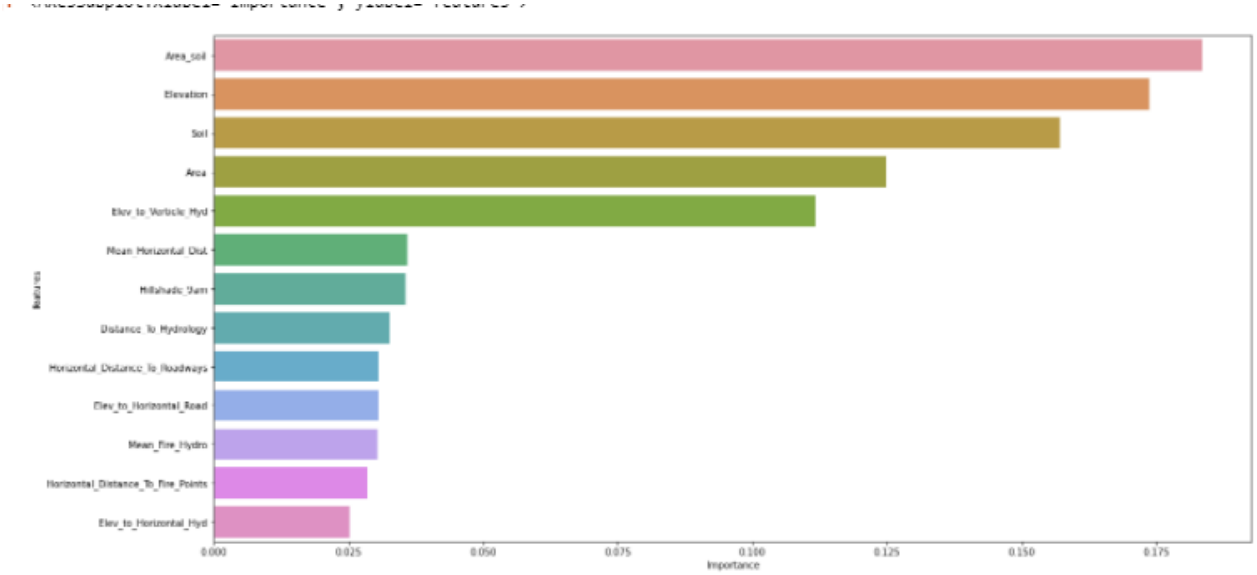
3. Boxplot looking for Outliers



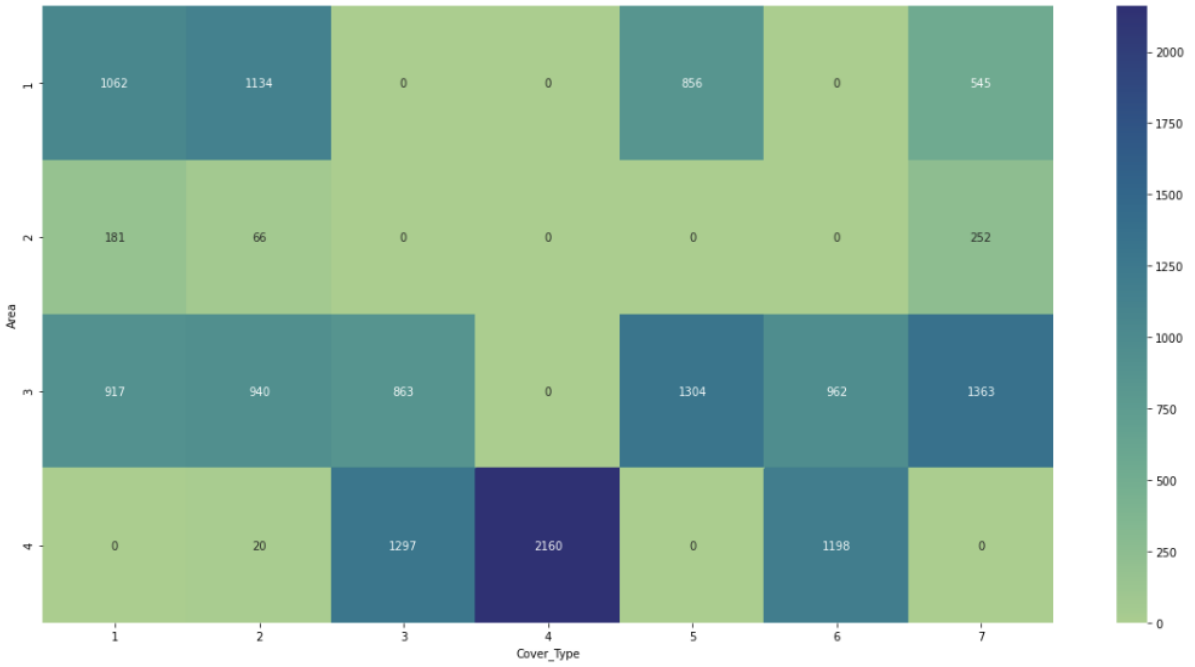
4. Correlation



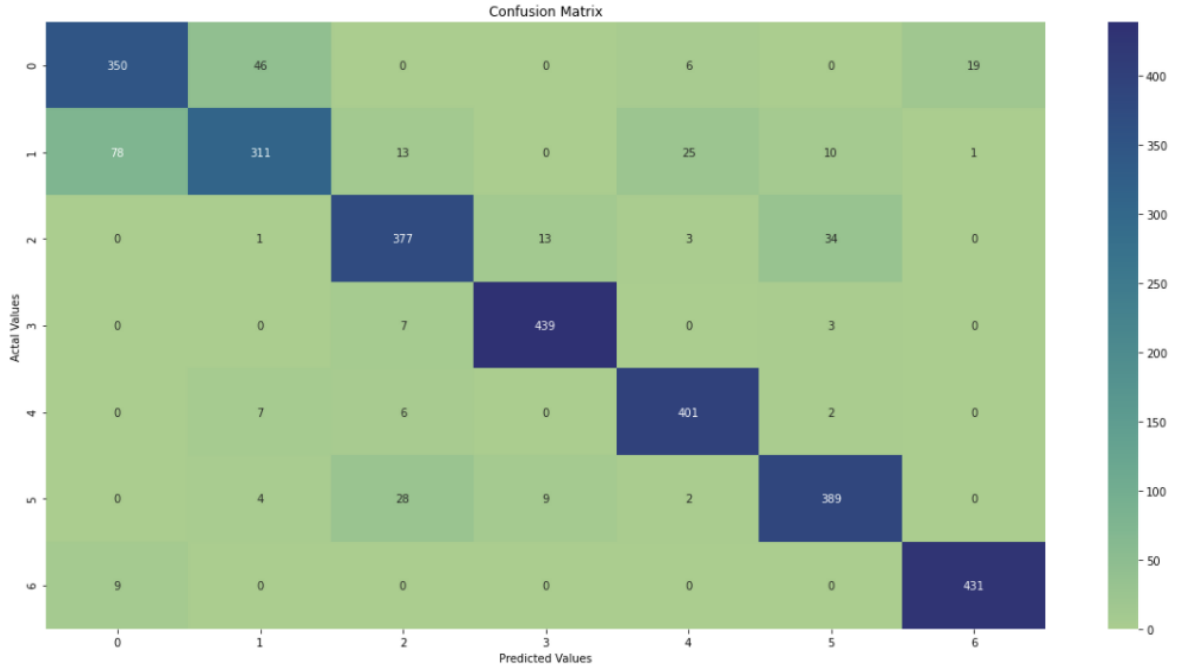
5. Most important Features



6. Area



7. Confusion Matrix


Confusion Matrix
Actal Values
Predicted Values
0 1 2 3 4 5 6
350 46 0 0 6 0 19
78 311 13 0 25 10 1
0 1 377 13 3 34 0
0 0 7 439 0 3 0
0 7 6 0 401 2 0
0 4 28 9 2 389 0
9 0 0 0 0 0 431
400 350 300 250 200 150 100 50 0