National Forest District Project

Dataset contains tree observations from four areas of one national forest district. The dataset includes information on tree type, shadow coverage, distance to nearby landmarks, soil type and local topography.

EDA (Exploratory data analysis)

Data contains 581012 observations and 55 attributes.

Attribute names:

'Elevation', 'Aspect', 'Slope', 'Horizontal_Distance_To_Hydrology', 'Vertical_Distance_To_Hydrology', 'Horizontal_Distance_To_Roadways', 'Hillshade_9am', 'Hillshade_Noon', 'Hillshade_3pm', 'Horizontal_Distance_To_Fire_Points', 'Wilderness_Area1', 'Wilderness_Area2', 'Wilderness_Area3', 'Wilderness_Area4', 'Soil_Type1', 'Soil_Type2', 'Soil_Type3', 'Soil_Type4', 'Soil_Type5', 'Soil_Type6', 'Soil_Type7', 'Soil_Type8', 'Soil_Type9', 'Soil_Type10', 'Soil_Type11', 'Soil_Type12', 'Soil_Type13', 'Soil_Type14', 'Soil_Type15', 'Soil_Type16', 'Soil_Type17', 'Soil_Type18', 'Soil_Type19', 'Soil_Type20', 'Soil_Type21', 'Soil_Type22', 'Soil_Type23', 'Soil_Type24', 'Soil_Type25', 'Soil_Type26', 'Soil_Type27', 'Soil_Type28', 'Soil_Type29', 'Soil_Type30', 'Soil_Type31', 'Soil_Type32', 'Soil_Type33', 'Soil_Type34', 'Soil_Type35', 'Soil_Type36', 'Soil_Type37', 'Soil_Type38', 'Soil_Type39', 'Soil_Type40',

Label name: 'Cover_Type' with 7 distinct classes and following distribution:

Class	Count	Percent
1	211840	36.46
2	283301	48.76
3	35754	6.15
4	2747	0.47
5	9493	1.63
6	17367	2.99
7	20510	3.53
	581012	100.00

Goal of the project is to build a model that predicts what types of trees grow in an area.

Dataset is unbalanced which and will require performing balancing techniques during

preprocessing. Dataset contains sufficient amount of data (number of observations) which is suitable for performing undersampling technique on multiclass data.

All attributes are of the type int64.

There are no missing values in the data.

There are variables which do not follow normal distribution - data is right or left skewed with outliers.

Proposed algorithms for classification are not sensitive on outliers and they do not require their handling, but some experimentations with outlier removal are performed in order to check if classification performance changes.

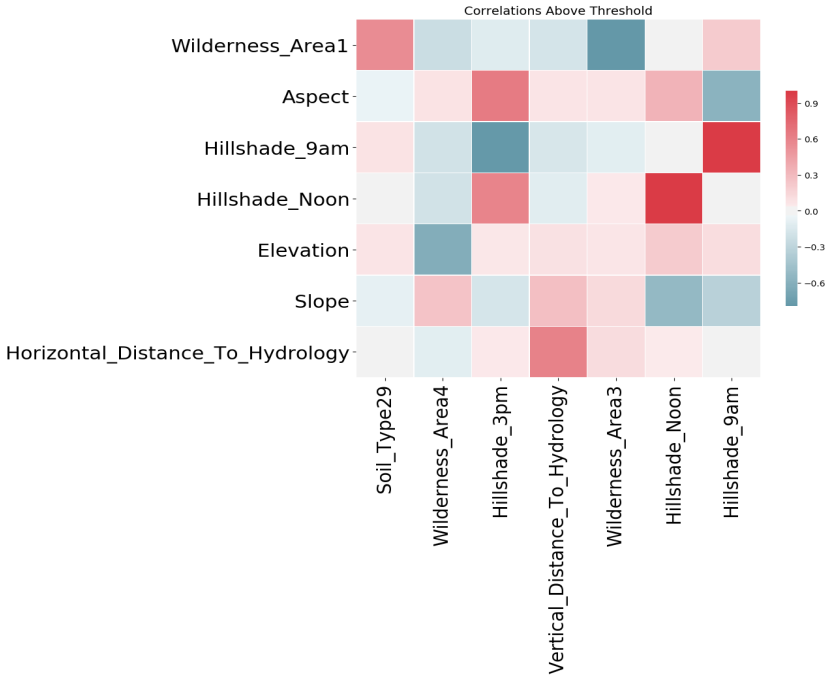
Data varies on different scales so standardization/normalization technique is performed in order to get better classification convergence.

Feature selection:

For this purpose Feature Selector tool is used: https://github.com/WillKoehrsen/feature-selector

It contains wide range of functions for qualitative datasets analysis such as identification of:

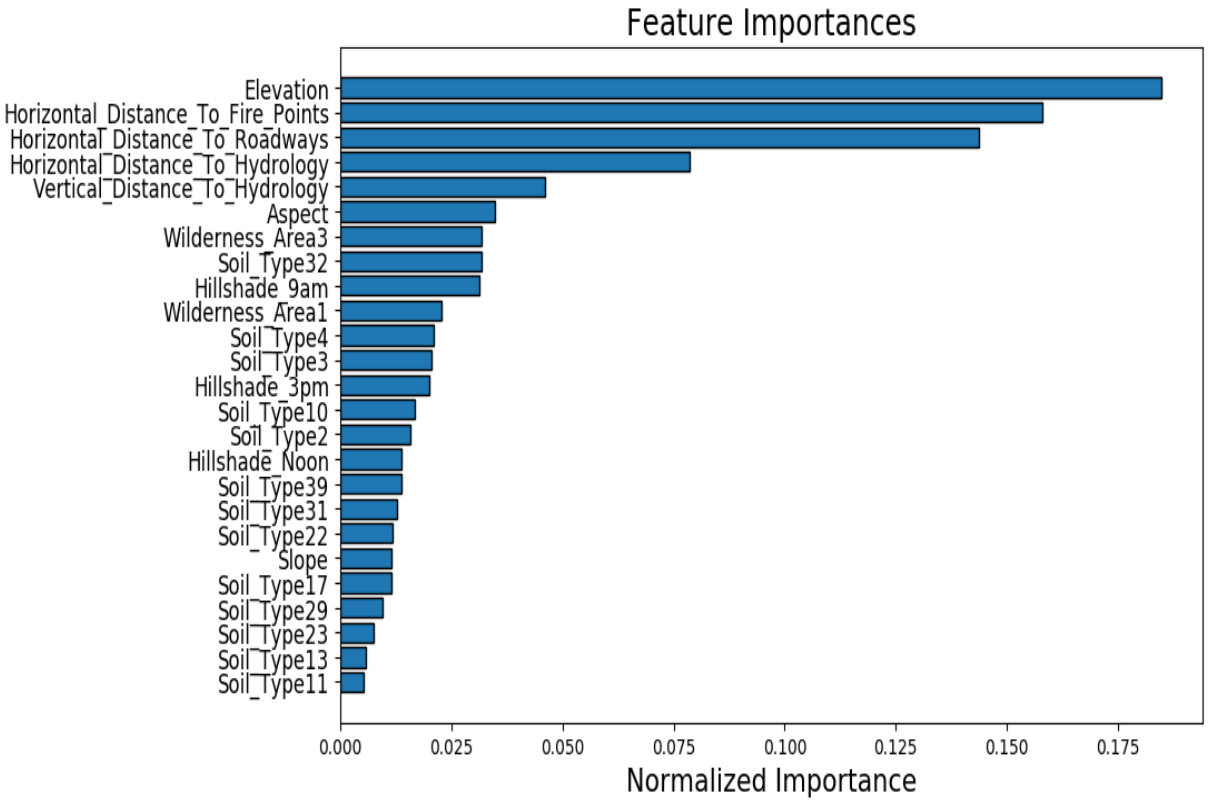
1. Missing values above particular threshold - set on 0.5
    - 0 features
2. Single unique values
    - 0 features
3. Colinear features above particular threshold - set on 0.5
    - 7 features - 'Vertical_Distance_To_Hydrology', 'Hillshade_9am', 'Hillshade_Noon', 'Hillshade_3pm', 'Wilderness_Area3', 'Wilderness_Area4', 'Soil_Type29'



drop_feature	corr_feature	corr_value
Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Hydrology	0.60624
Hillshade_9am	Aspect	-0.57927
Hillshade_Noon	Slope	-0.52691
Hillshade_3pm	Aspect	0.64694
Hillshade_3pm	Hillshade_9am	-0.7803
Hillshade_3pm	Hillshade_Noon	0.59427
Wilderness_Area3	Wilderness_Area1	-0.79359
Wilderness_Area4	Elevation	-0.61937
Soil_Type29	Wilderness_Area1	0.55055

1. Zaro importance features
    - There are identified 19 importance features - 'Soil_Type8', 'Wilderness_Area4', 'Soil_Type37', 'Soil_Type36', 'Soil_Type5', 'Soil_Type34', 'Soil_Type33', 'Soil_Type7', 'Soil_Type19', 'Soil_Type20', 'Soil_Type15', 'Soil_Type28', 'Soil_Type27', 'Soil_Type26', 'Soil_Type25', 'Soil_Type24', 'Soil_Type18', 'Soil_Type21', 'Soil_Type9'
    -
2. Low importance features
    - There are identified 23 low importance features - 'Soil_Type40', 'Soil_Type1', 'Soil_Type14', 'Soil_Type18', 'Soil_Type24', 'Soil_Type25', 'Soil_Type26', 'Soil_Type27', 'Soil_Type28', 'Soil_Type15', 'Soil_Type20', 'Soil_Type19', 'Soil_Type34', 'Soil_Type33', 'Soil_Type21', 'Soil_Type5', 'Soil_Type36', 'Soil_Type37', 'Wilderness_Area4', 'Soil_Type8', 'Soil_Type16', 'Soil_Type7', 'Soil_Type9'

In total 31 features are selected for model building in the folowing order:



	feature	importance	normalized importance	cumulative importance
0	Elevation	38.8	0.18476	0.18476
1	Horizontal_Distance_To_Fire_Points	33.2	0.1581	0.34286
2	Horizontal_Distance_To_Roadways	30.2	0.14381	0.48667
3	Horizontal_Distance_To_Hydrology	16.5	0.07857	0.56524
4	Vertical_Distance_To_Hydrology	9.7	0.04619	0.61143
5	Aspect	7.3	0.03476	0.64619
6	Wilderness_Area3	6.7	0.03191	0.6781
7	Soil_Type32	6.7	0.03191	0.71
8	Hillshade_9am	6.6	0.03143	0.74143
9	Wilderness_Area1	4.8	0.02286	0.76429
10	Soil_Type4	4.4	0.02095	0.78524
11	Soil_Type3	4.3	0.02048	0.80571
12	Hillshade_3pm	4.2	0.02	0.82571
13	Soil_Type10	3.5	0.01667	0.84238
14	Soil_Type2	3.3	0.01571	0.8581
15	Hillshade_Noon	2.9	0.01381	0.87191
16	Soil_Type39	2.9	0.01381	0.88571
17	Soil_Type31	2.7	0.01286	0.89857
18	Soil_Type22	2.5	0.01191	0.91048

19	Slope	2.4	0.01143	0.92191
20	Soil_Type17	2.4	0.01143	0.93333
21	Soil_Type29	2	0.00952	0.94286
22	Soil_Type23	1.6	0.00762	0.95048
23	Soil_Type13	1.2	0.00571	0.95619
24	Soil_Type11	1.1	0.00524	0.96143
25	Soil_Type38	1	0.00476	0.96619
26	Soil_Type35	1	0.00476	0.97095
27	Wilderness_Area2	1	0.00476	0.97571
28	Soil_Type30	1	0.00476	0.98048
29	Soil_Type12	1	0.00476	0.98524
30	Soil_Type6	0.9	0.00429	0.98952

Presentation and visualization of results

This is Multiclass classification problem since we are predicting the probabilities of the cover type label which contains 7 different values.

Metrics for evaluating models:

- F1-Score and Accuracy
- ROC and AUC curves
- Confussion matrix

2. Model: SVM Classifier

SVM Classifier

Initial run of SVM model with parameters: kernel=linear and C=1:

- Accuracy=0.79, F1-Score=0.78

SVM model after tuning with *GridSearchCV* : C, gamma, kernel and degree.

Best parameters selected by *GridSearchCV*: C=1, gamma=1, kernel=rbf, degree=1

- Accuracy=0.84, F1-Score=0.84

3. Explore ensemble model: XGBoost

XGBoost

Initial XGB model parameters

xgb.XGBClassifier(learning_rate=0.1, n_estimators=1000, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, objective='multi:softmax', nthread=4, num_class=7, seed=27)

- F1-Score: 0.84

XGB model after tuning with *GridSearchCV* : max_depth, min_child_weight and reg_alpha Best parameters/mean selected by *GridSearchCV*:

{'best_mean': 0.8651985002262362, 'best_param': {'max_depth': 9, 'min_child_weight': 1}})

XGBoost model pays high attention on the Soil Type + Wilderness Area 1/2 + Elevation variables.

- XGBoost Accuracy: 0.8642745709828393 XGBoost F1-Score (Micro): 0.8642745709828393

Final XGBoost model is selected since it gives higher F1-score and accuracy. Performances can be evaluated with further tuning parameter values for the model.

Model Deployment

Pickle files for both models are uploaded for further deployment and testing