GROUP ASSIGNMENT

FOREST COVER TYPE PREDICTION

MACHINE LEARNING II

PROF. ANGEL CASTELLANOS GONZALEZ

SECTION 1 - GROUP C – MBD OCT

- Edward Melgar
- Eleanor Yipei Li
- Alexandre Collot
- Pieter Van Poecke
- Karim El Chamaa
- Julian Krauth
- Ann-Charlotte Verstreken

Executive Summary:

The objective of this report is to detail the model findings from predicting the forest cover type based on geographical and geological features. After an extensive exploration of the dataset, feature engineering was performed to create meaningful combinations of the distance measurements that had a strong positive impact on the performance of the models.

Afterwards, various classification algorithms were tested before deciding that RandomForest and ExtraTrees are the best classifiers for this problem since both models achieved a base accuracy of over 80%. To further improve the performance, GridSearch was utilized to find the best hyperparameters for these models. The final model uses ExtraTreesClassifer and achieved **a cross-validated accuracy of 82.6%**.

Applying this model to the test set, the model achieved an accuracy score of 81.1%. Thus, the model developed based on 15K training data points indicates its sufficient competency in predicting the test data (560K).

Analysis:

This machine learning project aimed to create a model that classifies 7 different tree covers. The **CRISP-DM method** was leveraged to guide through the modelling process with the following steps:

1. Understanding the problem, the requirements, and the goal
2. Data collection
3. Data exploration and cleaning
4. Feature engineering
5. Model training
6. Model evaluation

1. Understanding the problem, the requirements, and the goal:

The project team was tasked to identify the forest cover type (the predominant kind of tree cover) from strictly cartographic variables. The actual forest cover type for a given 30 x 30 meter cell was the target variable, which was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The original data is unscaled and contains binary features such as wilderness areas and soil type.

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so the existing forest cover types are more a result of ecological processes rather than forest management practices.

2. Data collection:

The training set (15,120 observations) contains both features and the cover type. The test set contains only the features. To obtain the model performance, the cover type of every record in the test set (565,892 observations) needs to be computed.

With this approach, the model needs to be trained to learn the relationship between the features and the target variable, Cover_Type.

Different features are included in the dataset including:

- Distance features
- Hillshade index
- Slope and aspect
- 4 wilderness areas
- 40 soil types

To access the original features of the dataset provided, refer to annex.

The forest Cover Type to predict include:

1. Spruce/Fir
2. Lodgepole Pine
3. Ponderosa Pine
4. Cottonwood/Willow
5. Aspen
6. Douglas-fir
7. Krummholz

3. Data exploration and cleaning:

The distribution of every feature was deeply explored to obtain insights into the dataset's characteristics.

Correlation (Figure 1): hill shade features are correlated with each other, and with Aspect and Slope. This is expected as the hill shade is dependent on these natural factors. Elevation and all the horizontal distance features are positively correlated, indicating that fire points, hydrology, and roadways are mostly located in valleys.

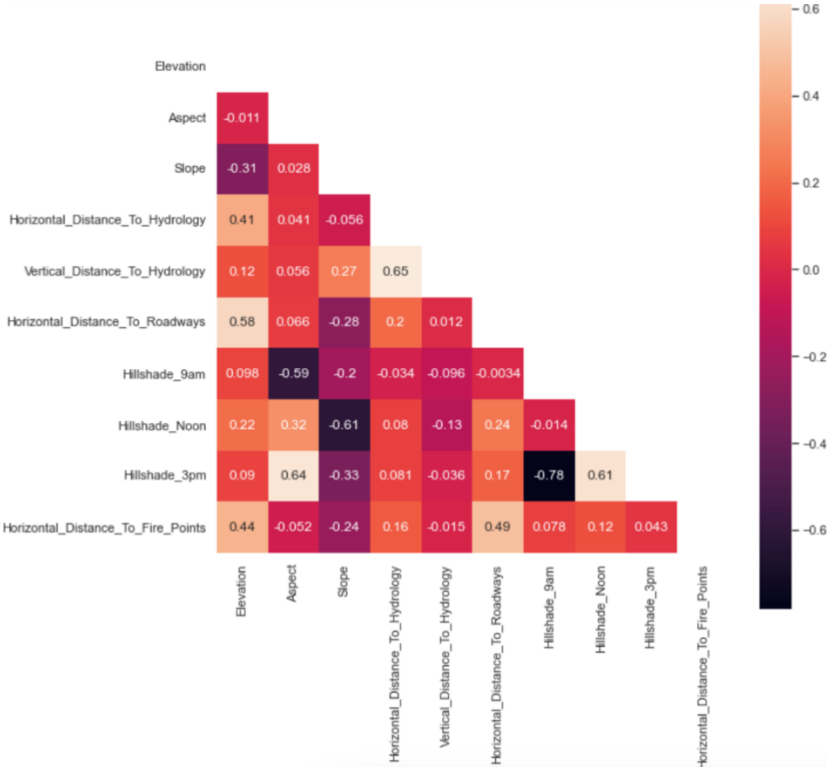


Figure 1: Correlation matrix of numerical variables

Skewness and distribution (Figure 2): the reason why some features are widely spread and have high values is that 5 out of the 10 variables are measured in meters. These variables are Elevation, all the horizontal distances variables. Features like Aspect and Slope are measured in degrees, which means that their maximum values cannot go above 360. Hill shade features can only take on a maximum value of 255 and are skewed.

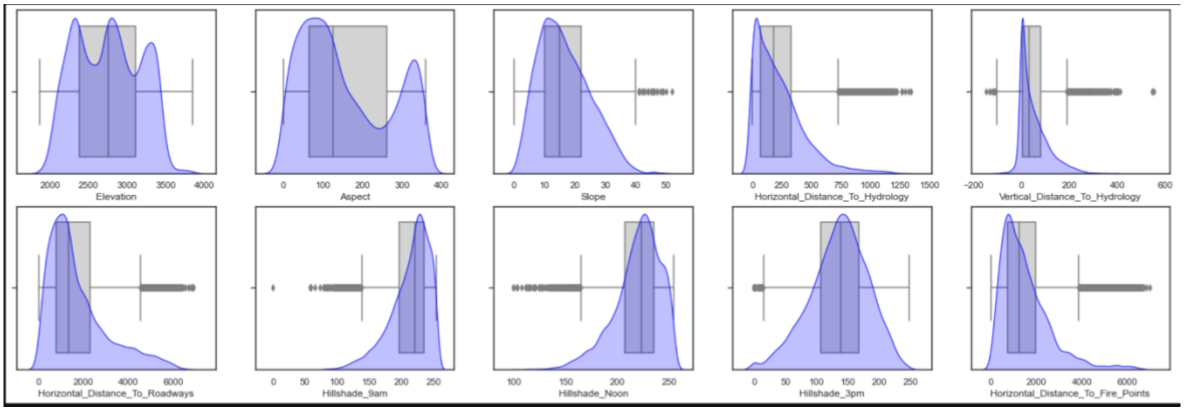


Figure 2: Distribution of numerical variables

Outliers (Figure 2): The outliers were visualized using boxplots. Although there are not many outliers present in the set, Hillshade_3pm has some 0 values that can be considered as outliers. Those values were inspected in further detail - they do not belong to a unique tree. It can be inferred that perhaps an error of measurement had occurred, and the values are not correct. Thus, it was decided that those 0 values would not be imputed since. However, it's worth noting that RobustScaler was applied to some algorithms later to handle the outliers.

Irrelevant features: There are two soil types that are not present in the training set, but have a small presence in the test set. These two variables do not bring any value to the analysis and were removed in later steps.

Missing values: Fortunately, the dataset does not include any missing values.

Expected feature importance (Figure 3 & 4): After plotting KDE to visualize the distribution of the cover type in relation to elevation, there seems to be a relationship between the two variables. In other words, the elevation is potentially a good estimator of the tree types because it separates the cover types well.

Knowing that elevation is correlated with most of the other distance metrics, deriving features from the distance metrics in the feature engineering part could potentially contribute to the prediction power of the model. In addition, the soil type and wilderness area should be good features to classify the trees. For instance, cover type 4 (Cottonwood/Willow) only exists in wilderness area 4 (Cache la Poudre).

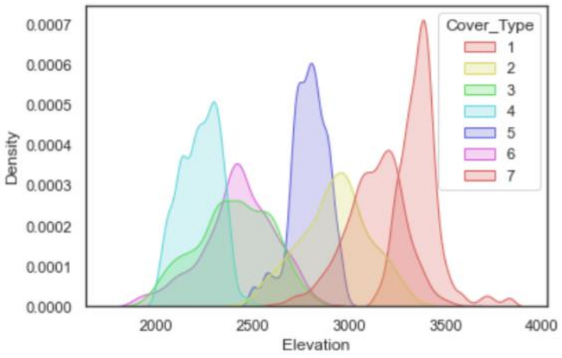


Figure 3: Kernel density estimate plot of Elevation by Cover_Type

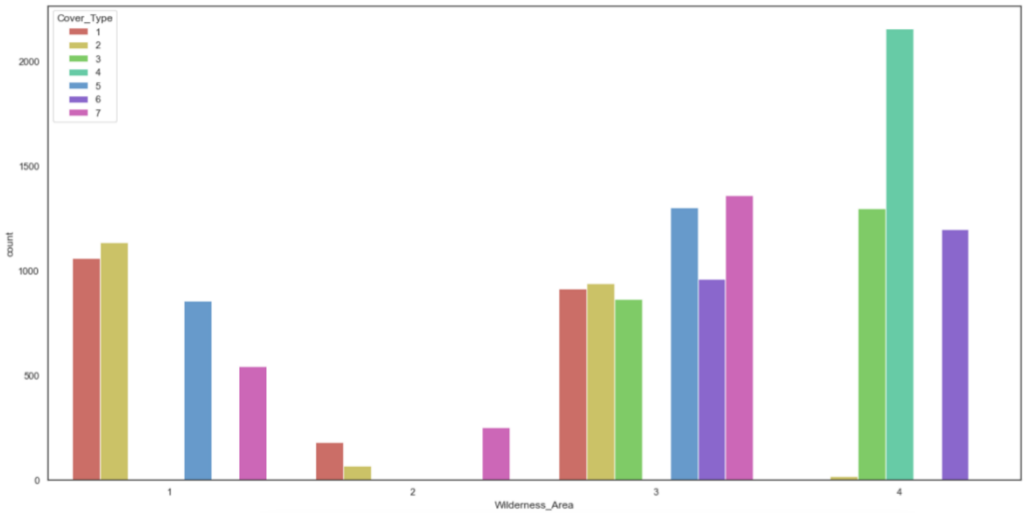


Figure 4: Count of records in each wilderness area by tree cover types

Balance: The training dataset is very balanced. The number of observations for the different trees type is identical (2160 each), which is good for the machine learning process.

4. Feature engineering:

Using the knowledge gained in data exploration, some new variables were created. Encoded variables were reverted to one categorical column only for tree-based machine learning algorithms as they are not needed to perform efficiently, and were kept as encoded for the other algorithms.

Distance-based features: Additional features were created using all distance features. Different combinations were created by adding and subtracting every vertical feature with each other, and this process was repeated for the horizontal distances, generating 8 new features:

1. Elevation – Vertical Distance to Hydrology
2. Elevation + Vertical Distance to Hydrology
3. Horizontal Distance to Hydrology + Horizontal Distance to Fire Points
4. Horizontal Distance to Hydrology - Horizontal Distance to Fire Points
5. Horizontal Distance to Hydrology + Horizontal Distance to Roadways
6. Horizontal Distance to Hydrology - Horizontal Distance to Roadways
7. Horizontal Distance to Fire Points + Horizontal Distance to Roadways
8. Horizontal Distance to Fire Points - Horizontal Distance to Roadways

Circular features: the Aspect feature has a minimum of 0 and a maximum of 360, thus representing a circle. The sine of Aspect was initially explored and added as a feature, but after testing, this feature affected the model performance negatively.

Statistical features: having the hill shade at 3 different points of the day (9am, 12am, and 3pm), we decided to calculate the mean as a metric for total hill shade throughout the day. The skew of the hill shade was not corrected as it did not improve the accuracy for the selected models.

5. Model Training:

5.1. Models Training & Selection:

Three separate training sets were prepared to be able to test the different classification algorithms with their different requirements (numerical/categorical handling).

X_scale: Some algorithms are sensible to rescaling (i.e KNN) and ordinal encoding for categorical features. In this way, X_scale numerical columns were transformed using Robust Scaler and reapplied One Hot Encoding for categorical features (leaving out soil types with no observations).

X_nb: The Naïve Bayes classifier cannot handle negative values. Thus, following the same steps as for X_scale, the Min Max Scaler was applied instead of Robust Scaler.

X_tree: Tree-based models do not require feature transformation and support ordinal encoded categorical features. No transformation for numerical features nor categorical ones was done (having one column for Wilderness Areas and one for Soil Types).

X_pca: PCA was applied on the X_scale using 11 components (after iterating over accuracy with LDA), to be able to test LDA with dimensionality reduction.

Without any hyperparameter tuning and setting the same seed, 11 different algorithms were trained, and their cross-validation accuracy scores were measured. All 11 models were training using the data transformation they required:

Ranking	Model	Accuracy CV=5	Training data
1	**Extra Trees**	**82.2%**	**X_tree**
2	**Random Forest**	**80.8%**	**X_tree**
3	XGBoost	80%	X_tree
4	Decision Tree (bag)	78.9%	X_tree
5	Decision Tree	72.5%	X_tree
6	SVM	70.6%	X_scale
7	KNN	69.3%	X_scale
8	Logistic Regression	61.9%	X_scale
9	LDA	60.3%	X_scale
10	LDA with PCA	59.7%	X_pca (n=11)
11	Naïve Bayes	56.9%	X_nb

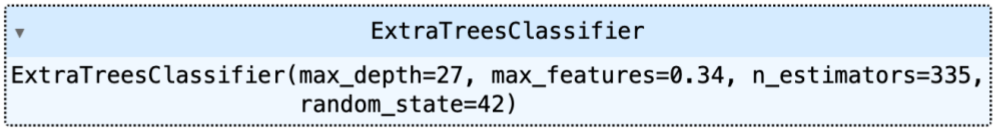
Overall, the tree-based models are the best performers in terms of accuracy, Extra Trees having the highest score. Tree-based methods are intuitive, based on decision rules and consider the variable interactions. SVM has an average accuracy of 70%, so it performed quite well even though categorical features were encoded using One Hot Encoder. This is because SVM can handle large feature spaces and even non-linear feature interactions. LDA combined with PCA did not perform well. There were no highly correlated features, thus PCA was not useful in this situation. Logistic Regression was also trained as a base model to compare with other classifiers. Naïve Bayes is the worst performing model.

5.2. Models Tuning:

Extra Trees and Random Forest were the best performing models and therefore were then selected for further hyperparameter tuning to improve the cross-validation-score for the final model.

Algorithm	Grid Search	Best Parameters	Accuracy CV=5
Extra Trees Classifier	max_depth max_features min_samples_leaf n_estimators	30 depth 34% of total 1 sample per leaf 335 estimators	**82.6%**
Random Forest	max_depth max_features min_samples_leaf n_estimators	37 depth 14% of total 1 sample per leaf 130 estimators	**80.8%**

Extra Trees Classifier was chosen for its accuracy (Figure 5). This model is a bit different than Random Forest (Bagging) in the sense that it does not perform bootstrapping (sub-samples of dataset) and the split is made at random rather than using the best split.



```
▾                    ExtraTreesClassifier
ExtraTreesClassifier(max_depth=27, max_features=0.34, n_estimators=335,
                     random_state=42)
```

Figure 5: Final model

The model has been tuned with the following hyperparameters:

- N_estimators: number of created trees
- Max_depth: the maximum depth of the tree
- Max_features: the maximum number of features used by the tree
- Min_sample_leaf: the minimum number of observations corresponding to the rule.

After having run the final model with the best parameters, the feature importance was plotted (Figure 6). Understanding which variables play a more important role in the model helps with model explainability and decision-making.

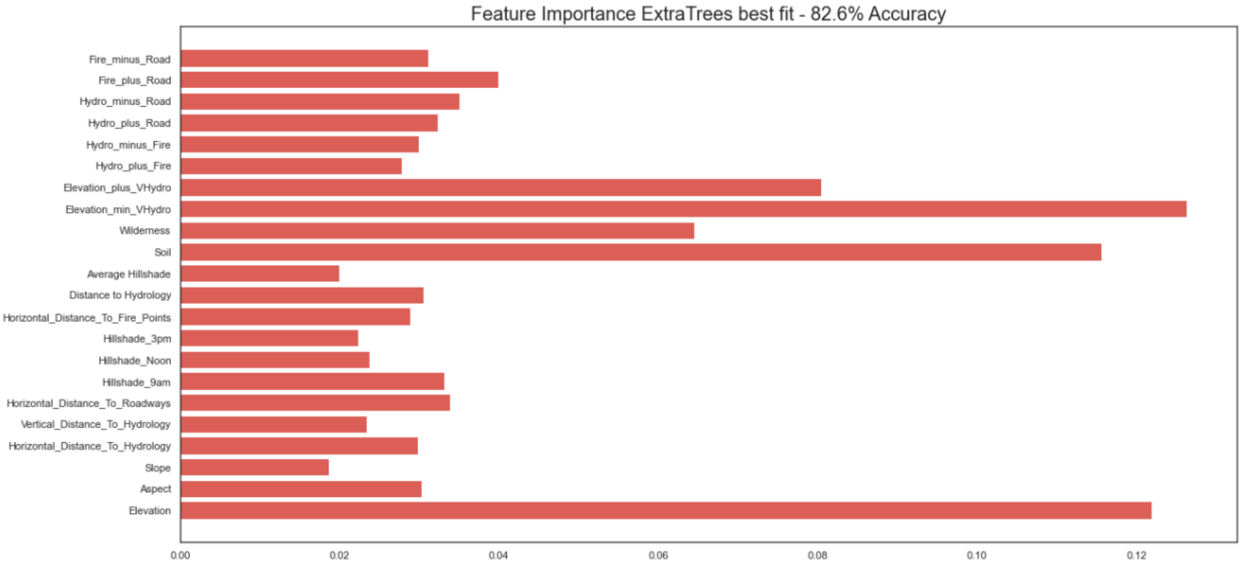


Figure 6: Feature importance from the final model

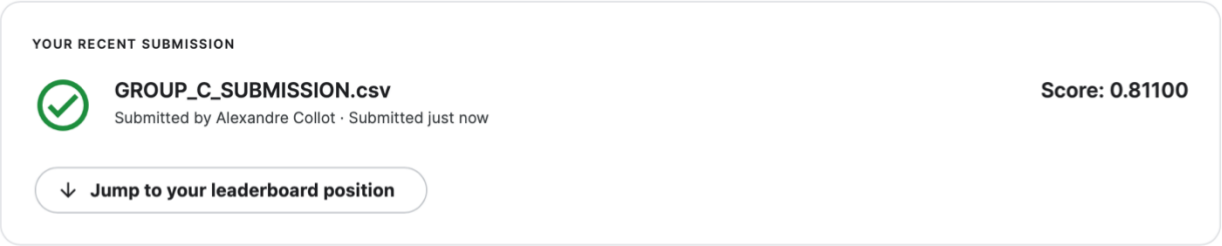
As expected, the Elevation (also in combination with hydrology), the Wilderness Area, and Soil Type have the highest importance in the final model. It means that those features mainly explain how the algorithm produces the classification between the 7 different tree types.

The model has been tuned to optimize accuracy. It is crucial to evaluate this model's performance on the testing set with the submission on Kaggle. This will validate whether or not the model will perform well on unseen data, and if it was over/underfitted.

6. Model Evaluation:

To be able to make predictions on the test set, the data was transformed by adding the features that were created in the training step. Using the model on the transformed test set, predictions can be made for each of the 565,892 observations.

The final prediction was submitted to Kaggle and obtained an accuracy of 81.1%, placing this model as Top 100 on the Kaggle leaderboard (Figure 7). We can then conclude that our model has not overfitted the training set as the accuracy scores are similar, which is very positive.



GROUP_C_SUBMISSION.csv

Submitted by Alexandre Collot · Submitted just now

Score: 0.81100

Figure 7: Kaggle result

Conclusion & Recommendations:

We have used this opportunity to extensively explore the data and experimented with different combinations of features and models. The final model using Extra Trees was optimized and provides strong predictive capabilities, shown by an 81.1% accuracy with the test set.

As a final remark, we think that a deeper domain knowledge – especially with geological features such as soil type - could help improve the model even further.

Annex

List of features:

- Elevation - Elevation in meters
- Aspect - Aspect in degrees azimuth
- Slope - Slope in degrees
- Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features
- Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features
- Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway
- Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice
- Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice
- Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice
- Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points
- Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation
    - 1 - Rawah Wilderness Area
    - 2 - Neota Wilderness Area
    - 3 - Comanche Peak Wilderness Area
    - 4 - Cache la Poudre Wilderness Area
- Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation
    - 1 Cathedral family - Rock outcrop complex, extremely stony.
    - 2 Vanet - Ratake families complex, very stony.
    - 3 Haploborolis - Rock outcrop complex, rubbly.
    - 4 Ratake family - Rock outcrop complex, rubbly.
    - 5 Vanet family - Rock outcrop complex complex, rubbly.
    - 6 Vanet - Wetmore families - Rock outcrop complex, stony.
    - 7 Gothic family.
    - 8 Supervisor - Limber families complex.
    - 9 Troutville family, very stony.
    - 10 Bullwark - Catamount families - Rock outcrop complex, rubbly.
    - 11 Bullwark - Catamount families - Rock land complex, rubbly.
    - 12 Legault family - Rock land complex, stony.
    - 13 Catamount family - Rock land - Bullwark family complex, rubbly.
    - 14 Pachic Argiborolis - Aquolis complex.
    - 15 unspecified in the USFS Soil and ELU Survey.
    - 16 Cryaquolis - Cryoborolis complex.
    - 17 Gateview family - Cryaquolis complex.
    - 18 Rogert family, very stony.
    - 19 Typic Cryaquolis - Borohemists complex.
    - 20 Typic Cryaquepts - Typic Cryaquolls complex.
    - 21 Typic Cryaquolls - Leighcan family, till substratum complex.
    - 22 Leighcan family, till substratum, extremely bouldery.
    - 23 Leighcan family, till substratum - Typic Cryaquolls complex.
    - 24 Leighcan family, extremely stony.
    - 25 Leighcan family, warm, extremely stony.
    - 26 Granile - Catamount families complex, very stony.

    - 27 Leighcan family, warm - Rock outcrop complex, extremely stony.
    - 28 Leighcan family - Rock outcrop complex, extremely stony.
    - 29 Como - Legault families complex, extremely stony.
    - 30 Como family - Rock land - Legault family complex, extremely stony.
    - 31 Leighcan - Catamount families complex, extremely stony.
    - 32 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
    - 33 Leighcan - Catamount families - Rock outcrop complex, extremely stony.
    - 34 Cryorthents - Rock land complex, extremely stony.
    - 35 Cryumbrepts - Rock outcrop - Cryaquepts complex.
    - 36 Bross family - Rock land - Cryumbrepts complex, extremely stony.
    - 37 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
    - 38 Leighcan - Moran families - Cryaquolls complex, extremely stony.
    - 39 Moran family - Cryorthents - Leighcan family complex, extremely stony.
    - 40 Moran family - Cryorthents - Rock land complex, extremely stony.

To predict:

- Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation
    - 1 - Spruce/Fir
    - 2 - Lodgepole Pine
    - 3 - Ponderosa Pine
    - 4 - Cottonwood/Willow
    - 5 - Aspen
    - 6 - Douglas-fir
    - 7 - Krummholz