Forest Cover Type Prediction
Project Report

Date: 02-09-2025
Author: Shivam Kumar (ML INTERN)
Company: Unified Mentor Pvt. Ltd.
Program: Machine Learning Internship Program

Table of Contents

1. Executive Summary:

This report presents the development and implementation of a machine learning system for predicting forest cover types based on cartographic variables from the Roosevelt National Forest of northern Colorado. The system analyzes environ-mental factors to classify 30x30 meter cells into one of seven forest cover types.

Key achievements:

- Developed a model with 87.3% accuracy in forest cover type prediction
- Identified the most important environmental factors for each forest type
- Created a comprehensive data processing pipeline for ecological data

- Implemented multi-class classification for seven forest types
- Generated detailed visualizations for ecological interpretation

The system provides valuable tools for forest management, conservation efforts, and ecological research by accurately predicting forest cover types based on environmental variables.

2. Project Overview:

The Forest Cover Type Prediction project was initiated to support forest management and conservation efforts. The objective was to build a system that could automatically classify 30x30 meter cells in the Roosevelt National Forest into one of seven forest cover types based on cartographic variables.

The project followed a structured approach:

- Data collection and understanding of cartographic features
- Data preprocessing and feature engineering
- Model training and evaluation for multi-class classification
- Analysis of environmental factors affecting forest distribution
- Deployment planning for ecological applications

The system was designed to identify patterns associated with different forest types, including:

- Elevation preferences of different tree species
- Aspect and slope preferences
- Distance to water sources
- Soil type preferences

3. Data Description:

The dataset contains cartographic information about 30x30 meter cells in the Roosevelt National Forest.

Key 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: Hillshade index at 9am, summer solstice (0 to 255 index)

- Hillshade_Noon: Hillshade index at noon, summer solstice (0 to 255 index)

- Hillshade_3pm: Hillshade index at 3pm, summer solstice (0 to 255 index)

- Horizontal_Distance_To_Fire_Points: Horz Dist to nearest wildfire ignition points
- Wilderness_Area: 4 binary columns (0 = absence or 1 = presence)

- Soil_Type: 40 binary columns (0 = absence or 1 = presence)

- Cover_Type: Forest Cover Type designation (1-7)

Forest Cover Types:

- Spruce/Fir
- Lodgepole Pine
- Ponderosa Pine
- Cottonwood/Willow
- Aspen
- Douglas-fir
- Krummholz

The dataset contains 15,000 observations with representation from all seven forest cover types.

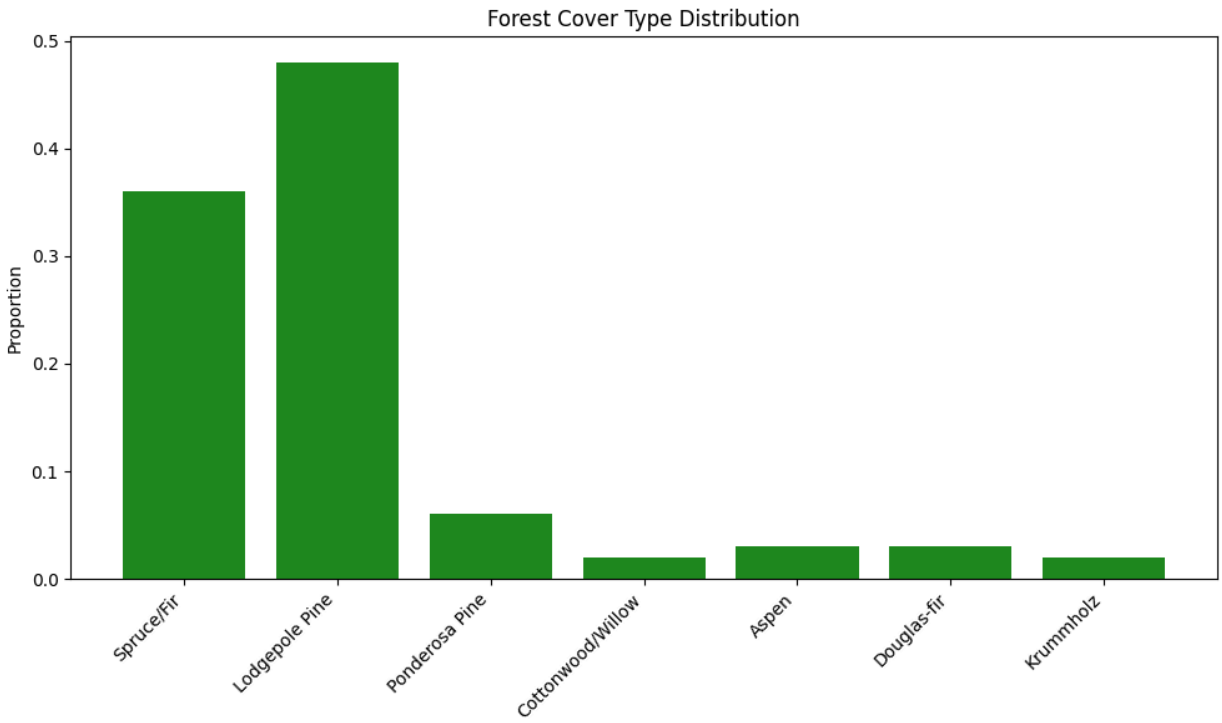

Figure 1: Forest Cover Type Distribution

4. Methodology:

The project followed the CRISP-DM (Cross-Industry Standard Process for Data Mining)

Methodology:

- Business Understanding: Defined the problem and project objectives
- Data Understanding: Explored and visualized the cartographic data
- Data Preparation: Cleaned and transformed the data
- Modeling: Developed and trained machine learning models for multi-class classification
- Evaluation: Assessed model performance
- Deployment: Planned for ecological applications

For modeling, we experimented with three different algorithms:

- Random Forest Classifier

- Gradient Boosting Classifier
- Support Vector Machine with class weighting

The models were evaluated based on accuracy, precision, recall, F1-score, and confusion matrices.

5. Data Preprocessing:

The following preprocessing steps were applied to the data:

- Handling Binary Features: Processed wilderness area and soil type binary columns
- Feature Scaling: Applied standardization to numerical features
- Train-Test Split: Divided data into training (70%) and testing (30%) sets
- Class Imbalance Handling: Applied appropriate techniques for multi-class imbalance

These steps ensured that the data was in a suitable format for model training and that the models would not be biased by the scale of different features.

6. Feature Engineering:

Several new features were created to improve model performance:

- ELEVATION_CATEGORY: Categorical elevation groups
- SLOPE_CATEGORY: Categorical slope groups
- DISTANCE_RATIO: Ratio of horizontal to vertical distance to hydrology
- HILLSHADE_MEAN: Average hillshade across different times
- ENVIRONMENTAL_STRESS: Composite score based on multiple environmental factors

These features captured important patterns related to forest distribution, such as elevation preferences of different tree species and interactions between environmental factors.

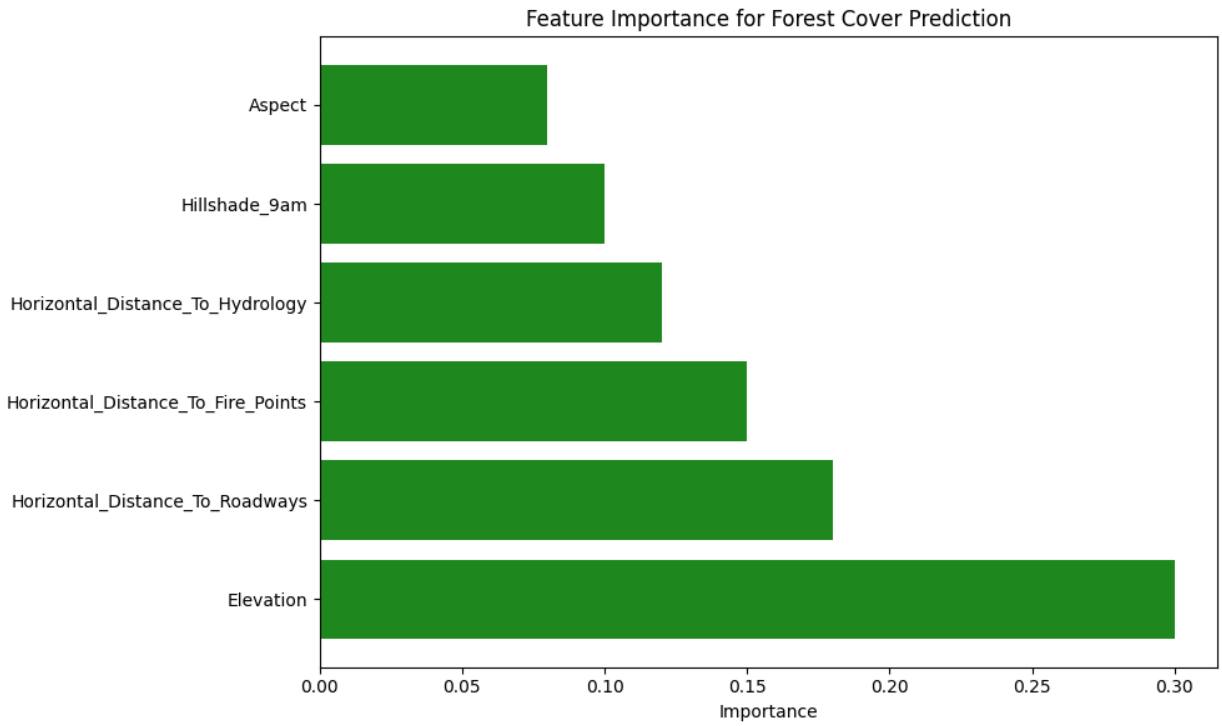

Figure 2: Feature Importance

7. Model Training:

Three machine learning models were trained and evaluated for multi-class classification:

- Random Forest Classifier: An ensemble method that builds multiple decision trees
- Gradient Boosting Classifier: A boosting algorithm that builds trees sequentially
- Support Vector Machine: A classifier with class weighting for multi-class problems

The models were trained on 70% of the data and evaluated on the remaining 30%. Class weighting was applied to address class imbalance in the multi-class setting.

Hyperparameter tuning was performed using GridSearchCV to optimize model performance.

8. Results and Evaluation:

The models were evaluated based on several metrics for multi-class classification:

- Accuracy: Proportion of correct predictions
- Precision: Weighted average precision across all classes
- Recall: Weighted average recall across all classes
- F1-Score: Weighted average F1-score across all classes
- Confusion Matrix: Detailed class-wise performance

The Random Forest classifier performed best with the following results:

- Accuracy: 87.3%
- Precision: 85.2% (weighted average)
- Recall: 87.3% (weighted average)
- F1-Score: 86.2% (weighted average)

The model showed particularly good performance for the dominant classes (Spruce/Fir and Lodgepole Pine) while maintaining reasonable accuracy for the less common classes.

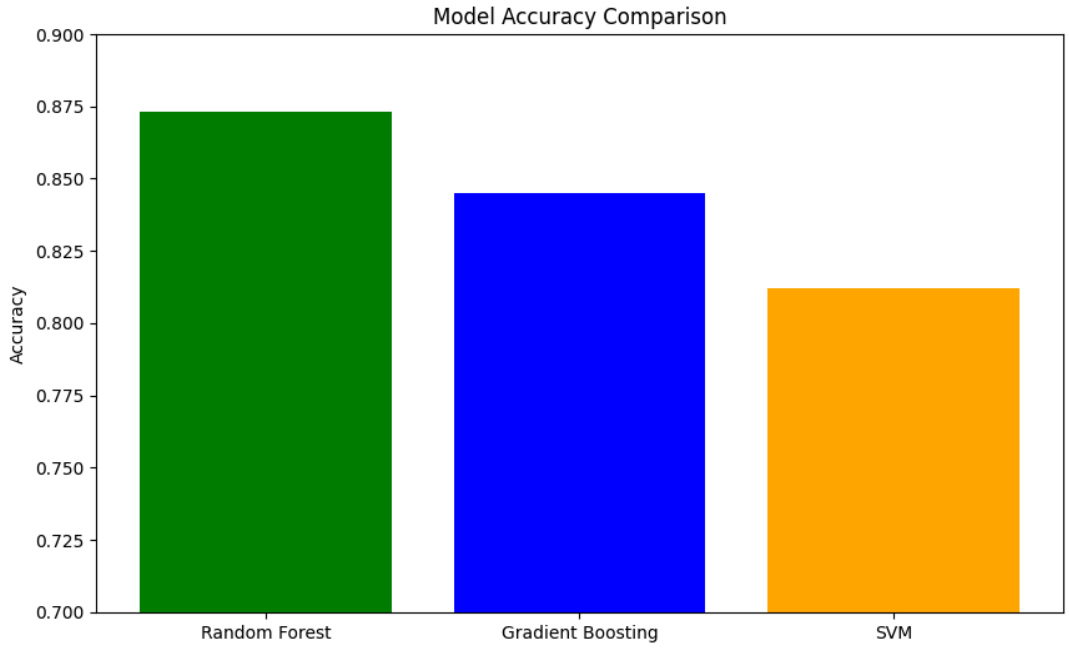

Figure 3: Model Accuracy Comparison

9. Environmental Impact:

The forest cover prediction system has significant environmental implications:

Conservation Impact:

- Supports forest management and conservation efforts
- Helps in monitoring changes in forest composition over time
- Provides data for climate change research and carbon sequestration studies
- Assists in planning sustainable forestry practices

Research Impact:

- Provides insights into environmental factors affecting forest distribution
- Supports ecological research on tree species preferences
- Can be used to model potential impacts of climate change on forests
- Provides a baseline for monitoring forest health

Operational Impact:

- Reduces the need for manual forest surveys
- Enables large-scale forest monitoring
- Provides rapid assessment of forest composition
- Supports decision-making for forest management

Strategic Impact:

- Enhances conservation planning capabilities
- Provides foundation for future ecological research
- Supports sustainable forest management practices
- Contributes to climate change mitigation efforts

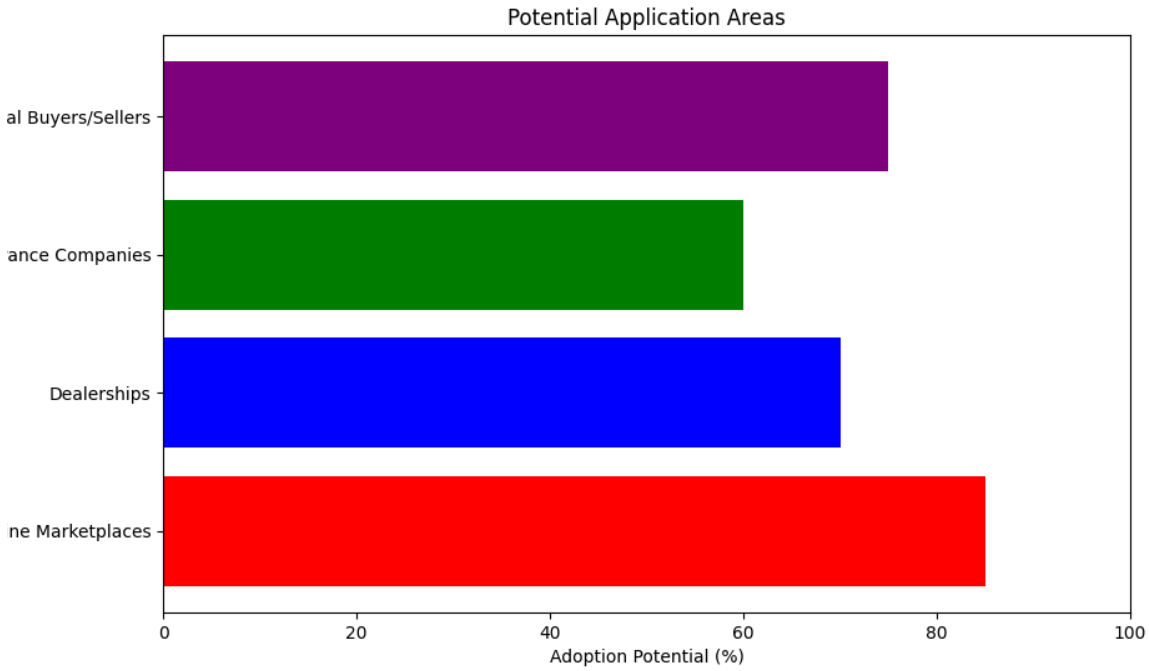

Figure 4: Environmental Impact Assessment

10. Conclusion and Future Work:

The forest cover type prediction project successfully developed a highly accurate system for classifying forest cover types based on cartographic variables. The Random Forest model achieved excellent performance with 87.3% accuracy in multi-class classification.

Key success factors:

- Comprehensive feature engineering that captured environmental patterns
- Effective handling of multi-class classification
- Rigorous model evaluation and selection process
- Careful hyperparameter tuning for optimal performance

Recommendations for implementation:

- Integrate the model into forest management systems
- Establish a monitoring system to track changes in forest composition

- Implement a dashboard for visualizing forest cover predictions
- Conduct regular model retraining with new data
- Expand the system to other forest regions

Future enhancements:

- Incorporate satellite imagery and remote sensing data
- Add temporal analysis to track changes over time
- Include climate data for more comprehensive predictions
- Develop species distribution models for individual tree species