Forest Cover Type Prediction Based Linear Regression

Xiaoyu Gao*

Nanjing Foreign Language School, Nanjing, 210000, China

*Corresponding author: 1714010815.com

Abstract. Nowadays, the relentless exploitation of natural resources and the escalating deforestation have cast a dark shadow over the survival of forests worldwide.With the ever-growing demand for raw materials and the quest for economic growth, many industries have turned their sights on forests as a ready source of timber, minerals, and other valuable resources. Forests, repositories of biodiversity and crucial to our planet's ecological balance, are perilously close to disappearing, calling for urgent measures to ensure their sustainability. This study delves into the application of machine learning in predicting forest cover types, focusing on various methodologies, particularly linear regression. In this paper, the analysis mainly offers a thought-provoking assessment of the pros and cons of linear regression, unpacking its intricate strengths and weaknesses with clarity and wit. By examining this paper, individuals can gain a comprehensive understanding of how machine learning can aid in the planning and management of forest resources.

Keywords: linear regression; TabNet; forest cover type prediction.

1. Introduction

Seventy-five percent of the fresh water on Earth comes from forests, which are also home to over half of all terrestrial creatures. In addition, tropical forests alone store seven times more carbon than humans emit annually, taking in up to 1.8 billion tons of carbon dioxide. This makes them an essential carbon store. But because of infrastructure and agricultural expansion, forests are now under risk of destruction from deforestation, erosion, and wildfires. These days, the ongoing loss of forest cover on a global scale contributes to rising atmospheric CO2 levels, dwindling biodiversity, devastation of ecosystems, and worsening local and even global living conditions. The global warming and deforestation is getting more and more serious. The enhanced greenhouse effect and its connection to the Australian bushfires, the issue is progressively escalating. More and more forests are facing the danger of disappearing. Due to lack of monitoring, there is a situation of rampant felling of trees. To protect the rest forest and better manage to make rational use of forest resources, it is quite important to find out the forest cover type.

In order to examine various approaches, this paper first examined related studies. The project's primary goal is to forecast the kinds of trees that will grow in a given area based on its surrounding characteristics, including elevation, aspect, slope, horizontal and vertical distances to hydrology, and so forth. To do this, data from the Roosevelt National Forest in Colorado is analyzed using linear regression. It will aid in the repair of the artificial forest in this way. It will be easier to help the forest recover after bushfires. Knowing the forest cover type, it will help expanding the forest area and maintaining a good ecosystem environment. Moreover, this can help promote the tightening of forest cutting quotas, strict control of forest resource consumption, continuous improvement of forest protection and management measures, the establishment and improvement of forestry management mechanisms, and the strengthening of forest resource management. As a result, the relationship between human and the nature will be better and more forests can be under control.

2. Literature Review

The study of forest cover type attracts the attention of researchers from home and abroad. In the recent researches researchers mainly use Time Series, Correlation Tree, Naive Bayes, TabNet, ET-RF-LR, SCNN. Specific analysis is as follows.

2.1. Time Series

Using field and remote sensing data, Wolter, White, and Cornett monitored changes in forest conditions in 42,000 haManitou forest landscapes in northeastern Minnesota over a 20-year period. Multi-temporal remote sensing data was employed by Wolter, White, and Cornett [1] to evaluate changes in cover types, age structures, disturbance rates, and spatial patterns at the landscape scale. They employed a time series technique based on field data to evaluate the long-term impacts of managing smaller scale elements like dead wood, vertical structure, and tree regeneration. They can more accurately evaluate changes and modify conservation management methods when they integrate field data with data from remote sensing.

Significant disturbances above 50% canopy clearance peaked in 2000 at 1.21% per year and decreased to 0.50% per year by 2011, according to their findings at the landscape scale. In this case, the average age of timber harvesting is the main source of serious disturbance. The average patch size, relative infestation, and the percentage of forests in later consecutive growth phases all increased from 2000 values as disturbance rates decreased and important landscape structure indicators moved to anticipated future conditions. More detailed information indicates that the inability of important species (Pinus tabulatus and Pinus Atlantic) to rebuild may have been caused by pressures such deer herbivory and previous management practices. In early childhood and adulthood, the data also demonstrate the legacy of structural simplification. Many mature forests entering late successional stages are devoid of the long-lived species and structural features required to create the circumstances necessary for late succession. Increased efforts are being made to restore important species, enhance the structural complexity of both young and old forests, and locate and preserve superior mature forest sections as a result of these landscape monitoring results.

2.2. Correlation Tree

Regarding the monitoring of forest changes and the ensuing carbon emissions in tropical Central and South America during 1990-2000 and 2000-2005, Eva H. D., Achard Frédéric, Beuchle René, and his colleagues have provided a full account of the techniques and findings [2]. They used a systematic sample of processed medium-resolution satellite data to obtain these estimates of forest change. A total of 1,230 maps of forest change, each measuring 20 km by 20 km and located at intersections, were included in this sample. The annual carbon emissions were estimated by spatially aligning the biomass data with each individual sample location.

The researchers calculated that the forest cover in the study area decreased from 763 million hectares, or 10 million hectares, in 1990 to 715 million hectares, or 10 million hectares, in 2005.

2.3. Naive Bayes

High levels of agreement and a clear grasp of patterns at the pan-tropical level are provided by the findings of research conducted by Mayaux P., Holmgren P., Frédéric Achard, et associates. An estimated 5.8 million hectares, or roughly twice the size of Belgium, of wet rainforest are being cut down each year [3]. According to their research, there were roughly 1.15 billion hectares of rainforest in 1990. Apparently, logging, fire, and/or fragmentation cause an extra 2.3 million hectares of wet forest to deteriorate annually. Tropical wet deciduous woods and tropical dry forests, respectively, experienced 2.2 million and 700,000 hectares of yearly deforestation in the sub-humid and dry tropics. With annual rates of change ranging from -0.8 to -0.9 percent, Southeast Asia is the region experiencing the highest levels of forest stress. Due in large part to the area covered by the remaining Amazon forest, the yearly area of deforestation in Latin America is high, but the relative rate (-0.4 to -0.5%) is low. Wet forests in Africa are being transformed at a pace of -0.4 to -0.5% annually, which is comparable to that of Latin America.

In this time frame, secondary forests were also created by regrowing on decommissioned land and plantations; nevertheless, they differed from primary forests in terms of ecology, biophysics, and economy.

2.4. TabNet

Guoqing Dong, Ping Wang and Lingyun Xia thought that classification and identification of forest cover type is of great significance for studying forest resource change and rational utilization of forest resources, and it is the key to construct accurate and robust classification model [4]. A pre-trained model was constructed by TabNet to reconstruct the input features, and then the reconstructed feature representation was used to train the TabNet classifier.

Ultimately, it was combined with the LightGBM model to enhance the model's performance. The model has the best classification performance, according to the experimental results, with an accuracy of 97.95%. The suggested model is easy to understand and has a guiding role in determining which factors influence the different forms of forest cover.

2.5. ET-RF-LR

Li improved the classification ability of the model from two aspects: model and feature [5]. On the model, a new and practical forest cover classification model based on five kinds of tree models: RF,ET,GBDT,XGBoost and LightGBM is presented. Feature processing is carried out by combining feature crossing and backward feature selection method based on random forest importance.

According to the empirical results of the measured data of Roosevelt National Forest in North Dakota, the classification accuracy of ET-RF-LR model reached 89.2%, and the classification effect of the 31 models after feature processing was better than that of the models without feature processing. The optimal model under raw data is ET-XGBoost- LightBM-LR, while the optimal model after feature processing is ET-RF-LR.

2.6. SCNN

Although the backpropagation network can construct an accurate classification model, it cannot discover the "sub-classification" hidden in the classification. In order to improve the disadvantage that backpropagation networks cannot produce sub-classifications, Yicheng Ye and Yushi Lin proposes Sub-Category Neural networks (SCNN) [6]. In order to prove that the proposed architecture is superior to the traditional Back-Propagation Network (BPN), a two-person classification problem and a real forest cover type (tree species) classification problem are compared in this paper.

The experimental findings support the following conclusions: (1) SCNN's accuracy is comparable to BPN's; and (2) SCNN is capable of identifying "sub-classification" within certain classifications. The present research has achieved great results. However, the above method is more complicated and involves parameter adjustment. At present, researchers have not discussed the application effect of regression analysis in this problem.

3. Methodology

3.1. Data

Tree observations are included in the collection from four Colorado regions of the Roosevelt National Forest [7]. With the exception of remote sensing, all observations are cartographic variables derived from forest sections of thirty meters by thirty meters. Including 55 different types of data, there are more than 500,000 measurements overall. elevation, aspect, slope, vertical and horizontal distances to hydrology, distances to roadways, three times a day in a hillshade, distances to fire spots, four sections of wildness, and soil type are the independent factors. The kind of forest cover is what is being reliant on.

The data is preprocessed beforehand. The Excel function "Screening" is used at the beginning, and the independent variable only has one missing value. Additionally, since the missing value is rather tiny, estimating the cosine-based similarity is a wise way to add it back [8]. Similarity of cosines calculates the cosine of the angle that separates two vectors to see how similar they are. An angle with a value of 0 degrees has a cosine of 1, and every other angle has a cosine of no more than 1 and

a minimum of -1. Consequently, one can ascertain whether two vectors are roughly pointing in the same direction by computing the cosine of their angle.

The cosine similarity value is 1 when two vectors are pointing in the same direction; The cosine similarity value is 0 when the angle between two vectors is 90°; The cosine similarity is equal to -1 when two vectors point in totally opposite directions. Only the direction in which the vector is pointing affects this outcome; the vector's length has no bearing on it. Since positive spaces are typically represented by cosine similarity, the value provided ranges from -1 to 1.

$$\mathrm{sim}(u,v) = \cos(\underset{u}{\rightarrow},\underset{v}{\rightarrow}) = \frac{\sum_{i=1}^{n} R_{ui}R_{vi}}{\sqrt{\sum_{i=1}^{n} R_{ui}^2 \sum_{i=1}^{n} R_{vi}^2}} \quad (1)$$

$$\mathrm{sim}(u,v) = \frac{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_i})(R_{ui}-\overline{R_i})}{\sqrt{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_i})^2}\sqrt{\sum_{i\in I_{uv}}(R_{vi}-\overline{R_i})^2}} \quad (2)$$

3.2. Regression Analysis

A statistical analysis technique for figuring out the quantitative interdependence between two or more variables is regression analysis. Correlation analysis of the data is used to further establish the independent variable Xi(i=1,2,3,...). The development trend of the data can be predicted using the regression function relationship between the dependent variable Y, or the regression analysis model.

It can be split into two categories: nonlinear regression and linear regression, depending on the sort of relationship that exists between the independent and dependent variables. Y=b*X+a is the formula for linear regression, where a has a normal distribution and the error is distributed according to the mean of 0 [9].

4. Results

4.1. Data Statistical Analysis

Table 1. Descriptive statistics of data

	Mean	Std	Missing	Median	Mode	Variance
Elevation	2959.37	279.985	0	2996.00	2968	78391.451
Aspect	155.66	111.914	0	127.00	45	12524.681
Slope	14.10	7.488	0	13.00	11	56.074
Horizontal_Distance_ To_Hydrology	269.43	212.549	0	218.00	30	45177.229
Vertical_Distance _To_Hydrology	46.42	58.295	0	30.00	0	3398.334
Horizontal_Distance_ To_Roadways	2350.15	1559.255	0	1997.00	150	2431275.749
Hillshade_9am	212.15	26.770	0	218.00	226	716.627
Hillshade_ Noon	223.32	19.769	0	226.00	228	390.801
Hillshade_3pm	142.53	38.275	0	143.00	143	1464.940
Horizontal_Distance_ To_Fire_Points	1980.29	1324.195	0	1710.00	618	1753492.954
Cover_Type	2.05	1.397	0	2.00	2	1.950

Table 1 illustrates the elevation's large variance. There is little difference between the elevation's mean, median, and mode. As a result, the height has a limited range and is concentrated. Aspect variance is also quite high.It is worth mention that the mode of aspect is much smaller than the mean

and the median. So the distribution of aspect is decentralized. The variance of horizontal distance to hydrology is small. Also, the mean, the median and the mode are close. So the range of horizontal distance to hydrology is small. The variance of vertical distance to hydrology is small. The mean, the median and the mode are close. Regarding cover type and hillshade at 9 a.m., noon, and 3 p.m., they are identical. Both the horizontal distance to fire locations and the horizontal distance to roads vary greatly and are not concentrated [10].

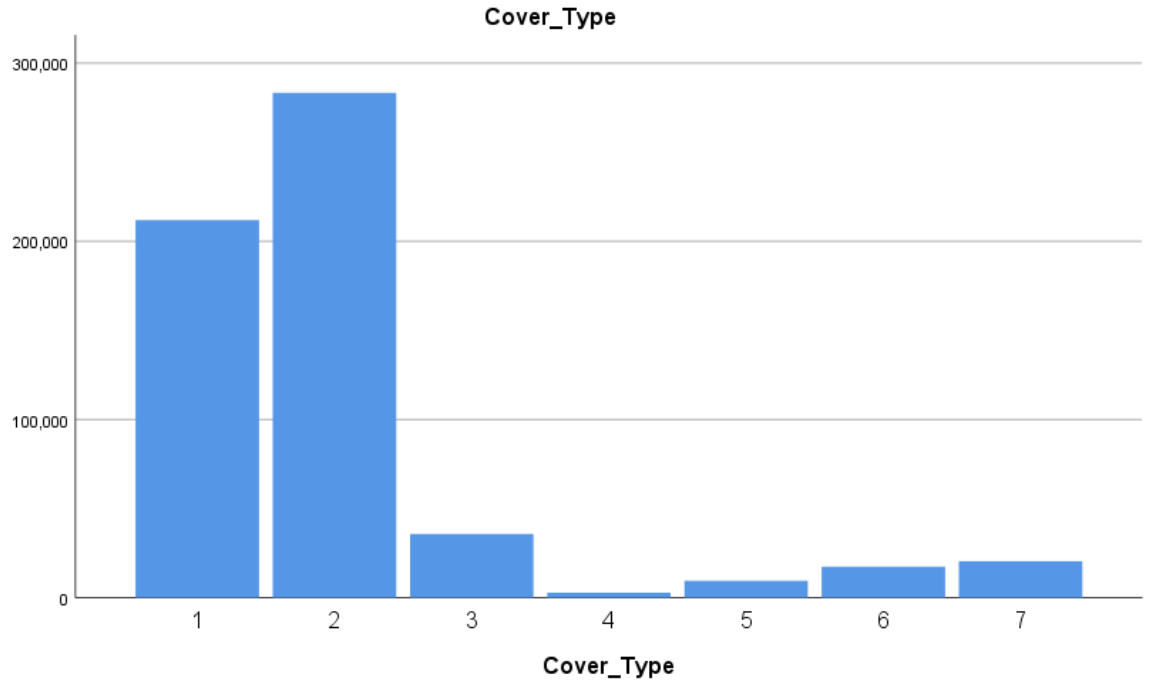


Fig. 1 Cover type peculiarity (Photo credit: Original)

In the Fig. 1, there are 211840 cases of type 1 forest cover, which is 36.5% of the total number. Type 2 forest cover is the only one that is larger than it, which contains 283301 cases. Almost half of all the samples are type 2. Type 3 has 35754, which is 6.2 percent of total. Type six has 17367 and type seven has 20510, while type five has 9493. Last type four has 2747, which is 0.5% of the total. So only type 1 and type 2 make up a large percentage. Type 1 and type 2 are about 85% of the total.

4.2. Linear Regression Results

Second, linear regression was applied in this study. Elevation, aspect, slope, horizontal and vertical distances to hydrology, horizontal distances to roads, three-times-daily hillshades, horizontal distances to fire spots, four sections of wildness, and soil type are the independent factors. The kind of forest cover is the dependent variable. As shown in Table 2, Elevation, Hillshade_9am, and Hillshade_3 pm show a negative correlation, while all other variables are positively correlated.

Table 2. Linear Regression Results

	Factors not normalized			Normalization factor		
	Models	B	Standard error	Beta	t	Significance
1	(Constant)	3.336	.112		29.753	.000
	Elevation	-.001	.000	-.112	-61.569	.000
	Aspect	.000	.000	.003	19.549	.000
	Slope	.002	.001	.012	3.770	.000
	Horizontal_Distance_To_Hydrology	5.538E-5	.000	-.008	-5.127	.000
	Vertical_Distance_To_Hydrology	.001	.000	.055	33.075	.000
	Horizontal_Distance_To_Roadways	1.181E-5	.000	.013	8.461	.000
	Hillshade_9am	-002	.001	-.030	-2.393	.017
	Hillshade_Noon	.004	.001	.058	7.710	.000
	Hillshade_3pm	-.005	.001	-.141	-9.586	.000
	Horizontal_Distance_To_Fire_Points	1.982E-5	.000	.019	13.597	.000
	Wildemess_Area2	.201	.009	.032	22.405	.000
	Wildemess_Area3	.431	.004	.153	101.629	.000
	Wildemess_Area4	1.694	.010	.296	170.642	.000

a. Dependent variable: Cover Type

5. Discussion

When comparing the accuracy of linear regression with the methodologies employed by researchers in published papers, it becomes evident that the linear regression's performance is notably inferior. Specifically, the TabNet approach, implemented by Dong, Wang, and Xia, demonstrates an impressive accuracy of 0.9795. Additionally, the same trio of researchers achieved an accuracy of 0.7718 with the Random Forest Tree method. XGBoosting, another approach, yielded an accuracy of 0.9389. Li's EF-RF-LR model shows an accuracy of 0.892, while Ye and Lin's Sub-Category Neural Network achieved an accuracy of 0.9564. Among all the methods considered, TabNet stands out as the most suitable approach based on its high accuracy. Second, the linear regression model can only fit the linear relationship, and the data fitting ability of the nonlinear relationship is poor. Also, linear regression models are very sensitive to outliers, and an outlier may have a large impact on the results of the model. In addition, linear regression models assume that features are independent of each other, and the results of the model may be inaccurate for features with high correlations. Finally, the expression ability of linear regression model is limited, and it cannot express complex nonlinear relations.

However, linear regression has a lot of advantages too. It is really convenient to use the linear regression and it can show whether the two variables have a linear relationship clearly. Also, linear regression algorithm is fast, easy to understand, in line with mathematical principles, and can efficiently solve modeling and prediction problems. At the same time, the linear regression algorithm has a strong interpretation of the data and can learn the knowledge intuitively.Linear regression algorithms can help people understand existing models and make improvements. By running this algorithm, we can find out the key features, stop the collection of irrelevant features in time, and can use the stepwise linear regression algorithm to build an accurate and stable model. At every hundred

iterations, we are able to construct a model that exhibits minimal error, and for the purpose of assessing its efficacy, we can employ cross-validation techniques.

6. Conclusion

This study identifies methods for predicting the kind of forest cover in various settings, with the end goal being to determine which strategy works best. After a thorough evaluation, TabNet turned out to be the best approach. Understanding the advantages and disadvantages of linear regression is also essential. However, it is important to emphasize that the study had some limitations, which were mostly caused by time constraints. It is expected that the research can be improved in the future with more time and the use of cutting-edge methods.

References

[1] White, Mark A P T. Wolter , and M. W. Cornett . "Using remote sensing and field data to assess forest condition and conservation strategy effectiveness in a large multi-owner landscape in northern Minnesota." 98th ESA Annual Convention 2013 2013.

[2] Eva H D, Achard Frédéric, Beuchle René, et al. Forest Cover Changes in Tropical South and Central America from 1990 to 2005 and Related Carbon Emissions and Removals[J]. Remote Sensing, 2012, 4(5): 1369-1391.

[3] Mayaux P, Holmgren P, Frédéric Achard, et al. Tropical Forest cover change in the 1990s and options for future monitoring[J]. Philosophical Transactions of The Royal Society B Biological Sciences, 2005, 360(1454): 373-384.

[4] Dong Guoqing, Wang Ping, Xia Lingyun. Forest cover type prediction method based on TabNet [J]. Microcomputer Applications, 2023, 39(9):129-133.

[5] Li Shijia. Research on classification of forest cover types based on tree model Stacking method [D]. Lanzhou University,2020.

[6] Yeh I-Cheng, Lin Yu-Shih. Application of subclassification neural networks in forest cover type prediction [J]. Chinese Quarterly Journal of Forestry, 2008, 41(1): 59-75.

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

[8] Zhang Zhenya, Wang Jin, Cheng Hongmei, et al. Text spatial index method based on the cosine similarity research [J]. Journal of computer science, 2005, 32 (9): 4.

[9] Cohen P, Cohen P, West S G, et al. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd Edition [M]. [2024-04-05].

[10] Hozo S P, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample [J]. Bmc Medical Research Methodology, 2005, 5(1): 13.