Forest Fire Prediction

Final Project Report

AI-539 Machine Learning Challenges

Winter 2024

Oregon State University

SANTHOS KAMAL ARUMUGAM BALAMURUGAN

03/19/2024

1. Problem to be solved

Problem Statement:

The aim is to focus revolves around creating a predictive model to anticipate forest fire occurrences. The objective is to develop a reliable algorithm capable of estimating the likelihood of a forest fire happening in a specific location.

Users / beneficiaries of a solution:

The people and their communities living In forest and Government agencies who tend in protecting the forest for the wildlife such as flora and fauna as well as the forest Management agencies who keep on monitoring the vulnerable areas and ensuring the protection of ecosystems, properties, and lives.

2. Data set properties:

Source: The source of my dataset for my model is from Kaggle. Which provides a wide variety of datasets which are necessary for my project to do testing and prediction.The citation would be Suryanand Singh and the

URL: https://www.kaggle.com/code/surya635/forest-fire-prediction/notebook

Data set profile: number of items, class distribution, type of features, min/max/mean/mean or distribution for each feature, etc. [length depends on the number of features in your data set]

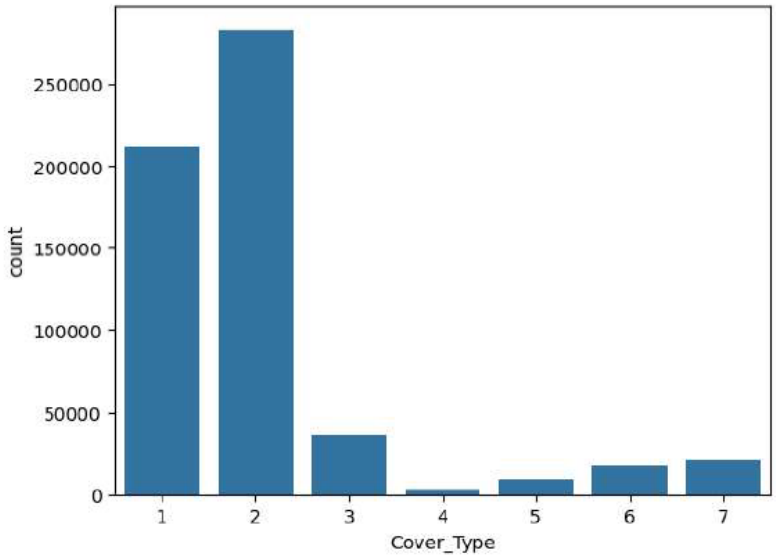
Number Of Items: (581012, 55)

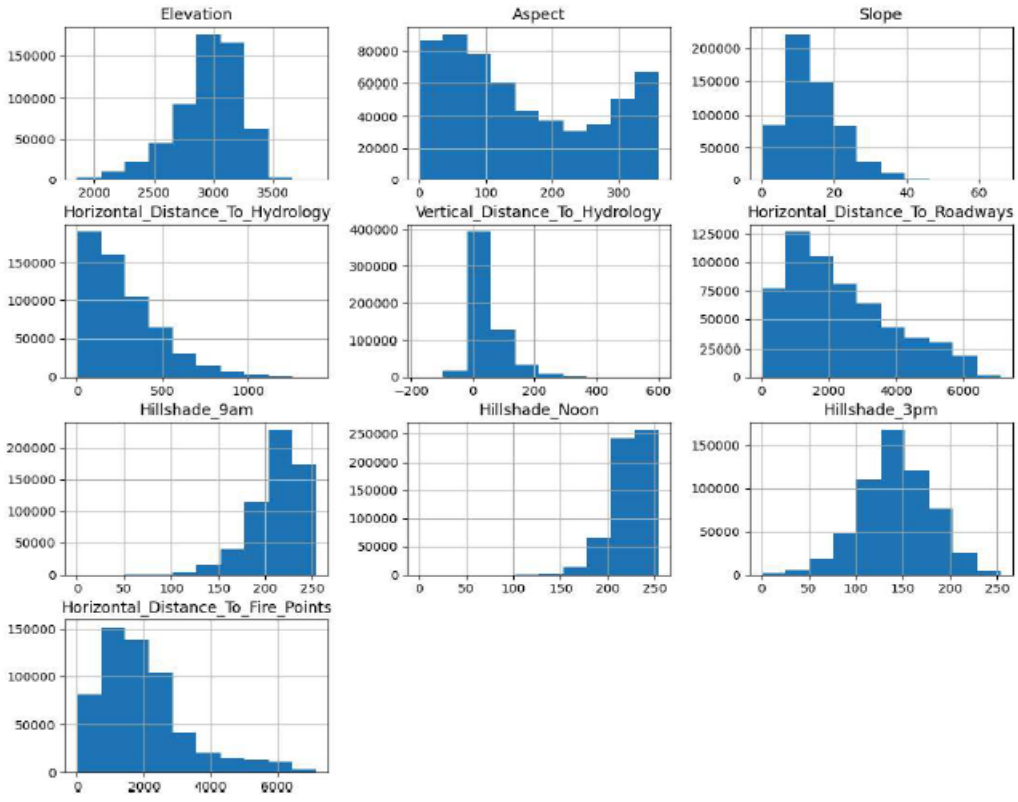
Types of Features:

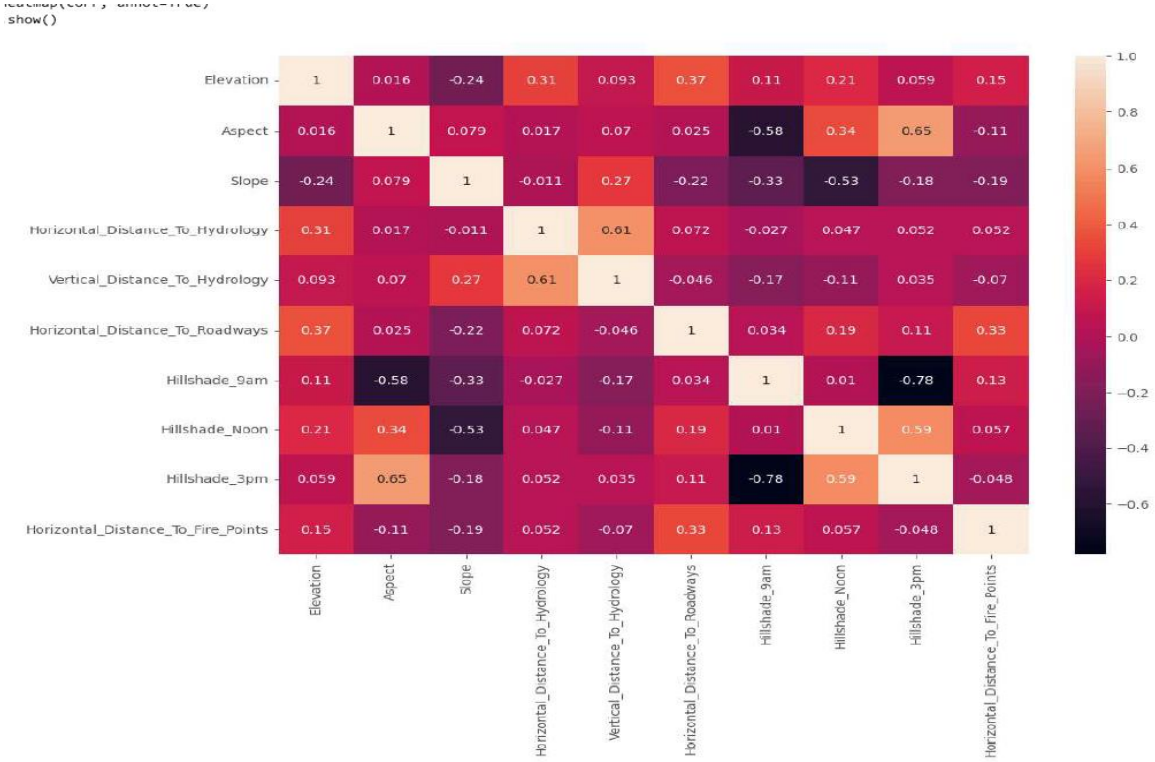
Elevation, Aspect, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology, Horizontal_Distance_To_Roadways, Hillshade_9am, Hillshade_Noon, Hillshade_3pm, Horizontal_Distance_To_Fire_Points, Wilderness_Area1,Wilderness_Area2, Wilderness_Area3, Wilderness_Area4, Soil_Type1, Soil_Type2, 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, Cover_Type.

min/max/mean/mean or distribution for each feature:

- Elevation: Min: 1859.000000, Max: 3858.000000, Mean: 2959.365301
- Aspect : Min: 0.0, Max: 360.00, Mean: 155.65
- Slope : Min: 0.0, Max: 66.00, Mean: 14.1
- Horizontal_Distance_To_Hydrology : Min: 0.00, Max: 1397.0, Mean: 269.42
- Vertical_Distance_To_Hydrology: Min: -173.0, Max: 601.00, Mean: 46.41
- Horizontal_Distance_To_Roadways: Min: 0.00, Max: 7117.0, Mean: 2350.14
- Hillshade_9am: Min: 0.00, Max: 254.0, Mean: 212.14
- Hillshade_Noon: Min: 0.00, Max: 254.0, Mean: 223.3
- Hillshade_3pm: Min: 0.00, Max: 254.0, Mean: 142.5
- Horizontal_Distance_To_Fire_Points: Min: 0.00, Max: 7173.0, Mean: 1980.2
- Soil_Type32: Min: 0.00, Max 1.00:, Mean: 0.090
- Soil_Type33: Min: 0.00, Max: 1.00, Mean: 0.077
- Soil_Type34: Min: 0.00, Max: 1.00, Mean: 0.0027
- Soil_Type35: Min: 0.00, Max: 1.00, Mean: 0.00325
- Soil_Type36: Min: 0.00, Max: 1.00, Mean: 0.00020
- Soil_Type37: Min: 0.00, Max: 1.00, Mean: 0.00051
- Soil_Type38: Min: 0.00, Max: 1.00, Mean: 0.0268
- Soil_Type39: Min: 0.00, Max: 1.00, Mean: 0.0237
- Soil_Type40: Min: 0.00, Max: 1.00, Mean: 0.0150
- Cover_Type: Min: 1.00, Max: 7.00, Mean: 2.051

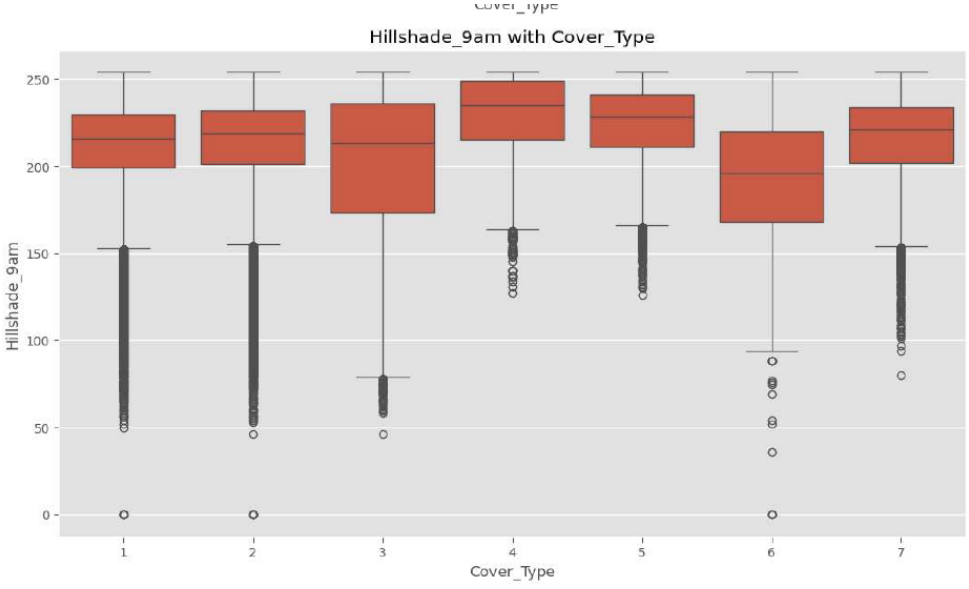

count
250000
200000
150000
100000
50000
0
1
2
3
4
5
6
7
Cover_Type


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



3. Machine learning model

The Model which I used for training my model is Random Forest for training my model and I chose it because it combines many decision trees to make predictions more dependable. It's known for being good at correctly classifying things. And it helps us figure out which factors are most important for making those predictions, which is super helpful for understanding our data better. It's also great at handling lots of different kinds of data and doesn't get thrown off by weird outliers.



Random Forest presents significant strengths for forest fire prediction. Its ensemble learning approach, combining multiple decision trees, enhances generalization and diminishes overfitting, crucial for modelling the complex relationships inherent in environmental features. However, despite its strengths, Random Forest poses challenges in interpretability, as understanding individual decision trees within the ensemble can be intricate, particularly with a large number of trees. Moreover, hyperparameter tuning demands time and computational resources, involving the optimization of parameters like the number of trees and their depth, which can impact the model's effectiveness and efficiency in forest fire prediction. For analysis n_estimators is the hyperparameter has been used and I have used the default value as 100 to train my model.

4. Evaluation:

Performance Measure: I used a number of measures, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared Error, to assess how well the Random Forest classifier performed in predicting the kind of forest cover. The response variable is denoted as y_test and it represents the target variable in the dataset. These measures provide information on how precise and accurate the model's predictions are.

Experimental Methodology: I used datasets for training and testing, I splited the data into two parts: one for teaching the model and the other for testing how well it works. This helps ensure the model learns from one set of data and is tested on another set to see if it can generalize well to new examples. To deal with the issue of imbalanced classes in the dataset, where some classes are much more common than others, I used

techniques like SMOTE and Random UnderSampler to balance out the classes. This prevents the model from being biased towards the more common classes and helps it learn from all types of examples.

Baseline Approach: For Baseline approach by using the "do nothing" strategy, I essentially trained the Random Forest model on the original dataset without making any adjustments. This approach serves as a baseline against which i can compare the performance of other strategies.The accuracy which I obtained (94%) indicates how well the model performs when no specific actions are taken. By comparing this accuracy with the accuracies achieved using other strategies, such as resampling techniques or cost-sensitive learning and more, I can evaluate the effectiveness of these strategies in improving the model's predictive capabilities.

5. Challenges:

Class Imbalance: Class imbalance refers to the situation where the distribution of classes in a classification dataset is not uniform, meaning that one class has significantly more instances than the other class or classes. In other words, there is an unequal representation of different classes in the dataset.

Feature Scaling: Feature scaling is a preprocessing technique used in machine learning to standardize or normalize the range of independent variables or features in the dataset. It involves transforming the values of features to a similar scale to ensure that no single feature dominates the learning algorithm due to its larger magnitude. Feature scaling is essential for many machine learning algorithms, particularly those based on distance metrics or gradient descent optimization, as it helps these algorithms converge faster and perform better.

Outliers: Outliers are data points that significantly deviate from the rest of the observations in a dataset. They can be unusually high or low values compared to the majority of the data and may represent measurement errors, experimental anomalies, or genuine but rare events. Outliers can have a substantial impact on statistical analyses and machine learning models, potentially skewing results and leading to inaccurate predictions or conclusions if not properly addressed.

Alternative Strategy

Class Imbalance:

- Do Nothing: To handle class imbalance, use the original dataset without making any changes.
- Resampling Techniques:
  Oversampling: Create synthetic samples to increase the number of examples in the minority class.

Under sampling: Remove samples at random to reduce the number of occurrences in the majority class.

- Cost-sensitive Learning: Divide the costs of incorrectly classifying occurrences of various groups, highlighting the significance of accurately anticipating the minority class. I have just given classes weight assigning from higher weight to minority class. And initializing Random forest with class weight.

Feature Scaling:

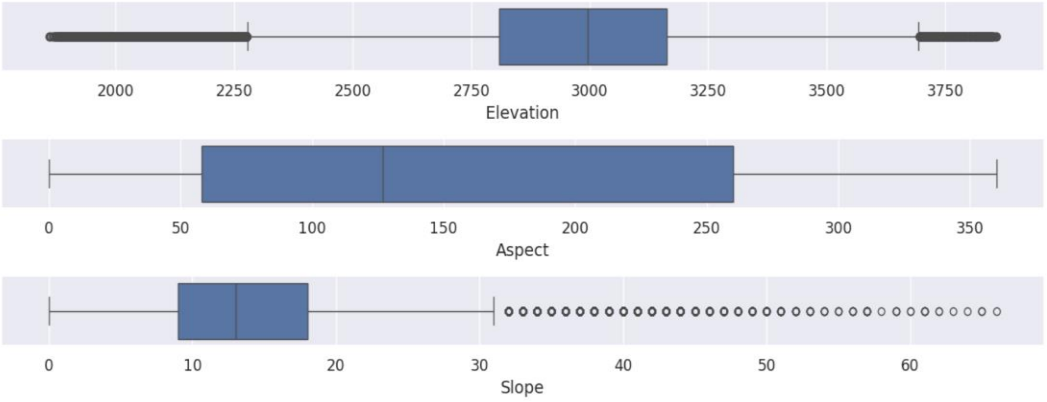
- Min-max scaling - Scale feature values to a range between 0 and 1.
- Standardization - Transform feature values to have a mean of 0 and a standard deviation of 1.
- Do Nothing - Scale features using statistics that are robust to outliers, such as the interquartile range.

Outliers:

- Trimming:

  Removing extreme values from the dataset is known as trimming. So in my model I have removed the entire row from the dataset with the values ranging from lower to upper thresholds. By this I can cutoff point at which certain data points constitute errors, and subsequently eliminating them from the dataset.

- Interquartile Range (IQR):
  A statistical measure of dispersion called the interquartile range (IQR) is computed as the difference between the data's upper and lower quartiles, or the 75th and 25th percentiles. This approach is popular for identifying outliers as it is resilient to extreme values.



- Do Nothing:
  Sometimes, outliers may contain valuable information or may not be truly indicative of errors or anomalies in the data. In such cases, you might choose to leave the outliers as they are and proceed with the analysis or modelling without any adjustments.

6. Results :

Class Imbalance

Random Forest	Do Nothing	Resampling Technique	Cost- Sensitive Learning
Accuracy	94.82	94.7835	94.81318
Mean Squared Error	0.26884	0.26966	0.27286
Mean Absolute Error	0.09243	0.09307	0.09309
R-Squared Error	0.86163	0.86120	0.85956

In Class Imbalance the "Resampling Technique" was the most effective method for addressing class imbalance in this scenario. Although the "Cost-Sensitive Learning" approach achieved slightly higher accuracy compared to the "Resampling Technique," it resulted in higher mean squared error, mean absolute error, and lower R-squared error. On the other hand, the "Resampling Technique" achieved competitive accuracy while maintaining lower errors across all metrics, indicating its effectiveness in mitigating the impact of class imbalance without significantly sacrificing model performance.

Feature Scaling

Random Forest	Do Nothing	Min-max scaling	Standardization
Accuracy	94.82	94.7890	94.8482
Mean Squared Error	0.26884	0.27059	0.26750
Mean Absolute Error	0.09243	0.09308	0.09206
R-Squared Error	0.86163	0.86072	0.86231

In Future Scaling the “Standardization ” was the most effective method for addressing the feature scaling. While on the other hand Do nothing and Min max scaling tend to perform bit low when compared to standardization of the Accuracy, mean squared error, mean absolute error and the r squared error.

Outliers

Random Forest	Do Nothing	Trimming	Interquartile Range (IQR)
Accuracy	94.82	94.9302	94.8544
Mean Squared Error	0.26884	0.26559	0.26533
Mean Absolute Error	0.09243	0.09116	0.09171
R-Squared Error	0.86163	0.8633	0.86343

In Outliers methods such as do nothing, trimming, and the Interquartile Range (IQR) approach applied to a Random Forest model, it's evident that Trimming emerges as the most effective method across various performance metrics. Trimming exhibits the highest accuracy score, indicating its superior ability to correctly classify data points, closely followed by doing nothing and then the IQR method.

Best Strategy:

Among the strategies employed - addressing class imbalance, feature selection, and handling outliers - the most effective one appears to be outlier handling through the Interquartile Range (IQR) method. This method not only yielded the highest accuracy but also led to the lowest mean squared error, mean absolute error, and highest R-squared error compared to the alternatives. Outliers can significantly skew the model's performance and prediction accuracy, hence effectively managing them results in better model performance overall.

7. Reflection:

Surprising Results: I got better results which I was not expecting in case of calculating the mean squared error, mean absolute error and r squared error for the model evaluation of Class imbalance, Future Scaling and Outlier.

Classmate Feedback: When I showed my results to my classmate sai neelie Balaji he found the results of outlier had a better accuracy rate when compared to other results such as class imbalance and Feature selection and even do nothing.

Changes from original plan: Initially I planned missing values as one of the strategy by on further implementation I found that there was no missing value in my dataset so after then checking my dataset and then I came up with another strategy which is feature scaling and I used as one of my challengeable strategies.

Additional Investigation: As I mentioned in the presentation I would do further more improvements in my data set like adding the temperature, wind speed, vegetation and Humidity or air moisture, And also by trying different model such as Recursive feature elimination and Principal component analysis which would make my predictions even more effectively.