ADVANCED ANALYSIS

INTELLIGENT SMOKE DETECTION SYSTEM WITH AI SENSOR FUSION

Group 10

Prepared By

S15089 - SANJANI WICKRAMASINGHE
S14982 - POORNIMA DISSANAYAKE
S14953 - BUDDHIMA SENARATHNA
S15006 - PASINDU SACHINTHA

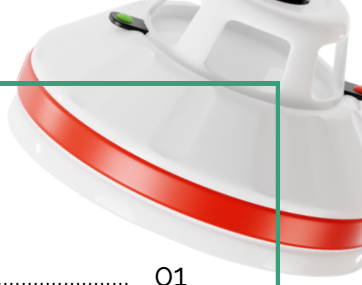

Table of Content

List of Figures

List of tables

1.0 LOGISTIC REGRESSION

Logistic Regression operates by modeling the probability of a binary outcome, providing a straightforward and comprehensible prediction. It showcases its versatility by effectively handling structured data and offering transparent insights into the significance of each feature.

Summary of the initial logistic regression model is as follows.

```
Optimization terminated successfully.
         Current function value: 0.269558
         Iterations 11
                         Results: Logit
==================================================================
Model:              Logit            Method:           MLE
Dependent Variable: Fire Alarm       Pseudo R-squared: 0.549
Date:               2023-09-16 11:44 AIC:              27035.8892
No. Observations:   50104            BIC:              27141.7514
Df Model:           11               Log-Likelihood:   -13506.
Df Residuals:       50092            LL-Null:          -29942.
Converged:          1.0000           LLR p-value:      0.0000
No. Iterations:     11.0000          Scale:            1.0000
------------------------------------------------------------------
                 Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
------------------------------------------------------------------
Temperature[C]  -0.0475   0.0015 -32.1201 0.0000   -0.0504  -0.0446
Humidity[%]     -0.0218   0.0024  -9.0005 0.0000   -0.0265  -0.0171
TVOC[ppb]       -0.0014   0.0000 -38.5551 0.0000   -0.0015  -0.0013
eCO2[ppm]        0.0020   0.0001  29.2232 0.0000    0.0019   0.0021
Raw H2           0.0067   0.0002  36.1211 0.0000    0.0064   0.0071
Raw Ethanol     -0.0088   0.0001 -59.9153 0.0000   -0.0091  -0.0085
Pressure[hPa]    0.0952   0.0015  61.7024 0.0000    0.0922   0.0982
PM1.0           11.2523   5.3934   2.0863 0.0369    0.6815  21.8232
PM2.5            0.0465   5.4470   0.0085 0.9932  -10.6294  10.7224
NC0.5           -8.3314   1.2066  -6.9048 0.0000  -10.6963  -5.9665
NC1.0           44.6279  10.4021   4.2903 0.0000   24.2401  65.0157
NC2.5          -76.5978  14.0428  -5.4546 0.0000 -104.1213 -49.0744
==================================================================
```

Figure 1: Summary of the initial logistic regression model

An R-squared (R2) value of 0.549 which is relatively good indicates that approximately 55% of the variance in the dependent variable has been explained by the logistic model. But there are some insignificant variables (at 5% level) present in the model and a new logistic model will be fitted by dropping these variables.

```
Optimization terminated successfully.
         Current function value: 0.269558
         Iterations 11
                         Results: Logit
==================================================================
Model:              Logit            Method:           MLE
Dependent Variable: Fire Alarm       Pseudo R-squared: 0.549
Date:               2023-09-16 11:44 AIC:              27033.8892
No. Observations:   50104            BIC:              27130.9296
Df Model:           10               Log-Likelihood:   -13506.
Df Residuals:       50093            LL-Null:          -29942.
Converged:          1.0000           LLR p-value:      0.0000
No. Iterations:     11.0000          Scale:            1.0000
------------------------------------------------------------------
                 Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
------------------------------------------------------------------
Temperature[C]  -0.0475   0.0015 -32.1202 0.0000   -0.0504  -0.0446
Humidity[%]     -0.0218   0.0024  -9.0007 0.0000   -0.0265  -0.0171
TVOC[ppb]       -0.0014   0.0000 -38.5551 0.0000   -0.0015  -0.0013
eCO2[ppm]        0.0020   0.0001  29.2238 0.0000    0.0019   0.0021
Raw H2           0.0067   0.0002  36.1218 0.0000    0.0064   0.0071
Raw Ethanol     -0.0088   0.0001 -59.9154 0.0000   -0.0091  -0.0085
Pressure[hPa]    0.0952   0.0015  61.7034 0.0000    0.0922   0.0982
PM1.0           11.2538   5.3902   2.0878 0.0368    0.6892  21.8185
NC0.5           -8.3310   1.2054  -6.9115 0.0000  -10.6935  -5.9685
NC1.0           44.6692   9.2040   4.8532 0.0000   26.6297  62.7087
NC2.5          -76.5999  14.0348  -5.4579 0.0000 -104.1076 -49.0923
==================================================================
```

Figure 2: Summary of the logistic regression model after dropping insignificant variables

The R2 value of the new model is the same as the previous model. It suggests that those variables were indeed not contributing significantly to the model's predictive power. But the accuracy of the test set is 0.89. The performance of logistic regression, in terms of accuracy, depends on the degree of linear separability in the dataset. In the previous analysis, we discovered that our dataset may not be linearly separated.

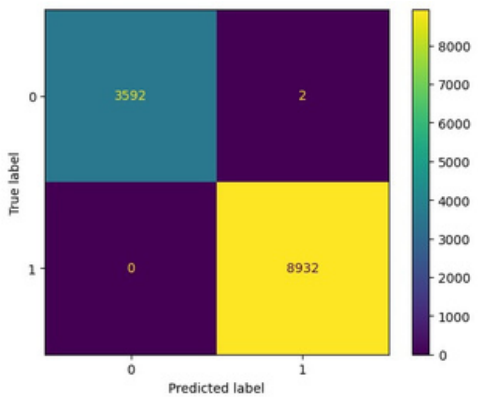

Figure 3: Confusion Matrix of the logistic regression model

As the accuracy of the model is not quite high, further improved models will be fitted.

2.0 LOGISTIC REGRESSION WITH REGULARIZATION TECHNIQUES

Logistic Regression with Regularization techniques (Ridge, Lasso, Elastic Net) was applied to the dataset to enhance the predictive performance and manage the influence of features. These techniques add penalties to the logistic regression model to prevent overfitting and improve generalization.

2.1 Logistic Regression with Ridge :

Ridge Regression introduces L2 regularization to the logistic regression model. Some of the feature coefficients of the Ridge Regression model for this dataset were reduced to exactly zero. (PM values and NC Values).
The model achieved an accuracy of 0.88 on the test dataset. The Confusion matrix for the Ridge is attached here. It helps to understand how well our model is distinguishing between the classes it's trying to predict

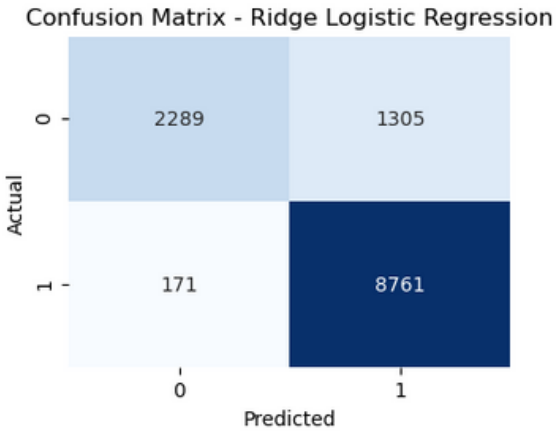

Figure 4: Confusion Matrix of the Ridge model

2.2 Logistic Regression with Lasso :

In Lasso Regression also some of the feature coefficients were reduced to exactly zero, effectively performing feature selection. The non-zero coefficients were as follows:

- TVOC[ppb]: -0.0014
- eCO2[ppm]: 0.0020
- Raw H2: 0.0069
- Raw Ethanol: -0.0088

Other coefficients for the remaining features were reduced to exactly zero. The Lasso Regression model also achieved an accuracy of 0.88 on the test dataset. Similar to Ridge, it provided precision, recall, and F1-score metrics for different classes. Lasso also showed variations in precision and recall

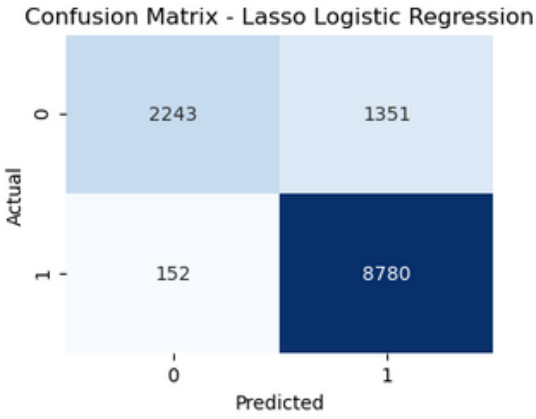

Figure 5: Confusion Matrix of the Lasso model

2.3 Logistic Regression with ElasticNet :

Elastic Net Regression combines the strengths of both Ridge and Lasso by adding both L1 and L2 penalties to the loss function. In this model, some coefficients were reduced to zero, similar to Lasso.
The Elastic Net Regression model achieved an accuracy of 0.88 on the test dataset, the same as Ridge and Lasso. To understand how well our model is distinguishing between the classes it's trying to predict, here shows the confusion matrix for the Elastic Net model as well.

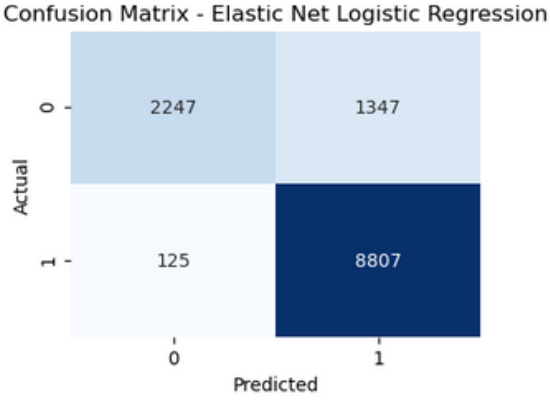

Figure 6: Confusion Matrix of the Elastic Net model

3.0 RANDOM FOREST

Random Forest is an ensemble learning technique that combines the predictive power of multiple decision trees. It excelled in achieving high accuracy while effectively mitigating overfitting, making it a robust choice for our task. The Random Forest model demonstrated its versatility, effectively handling structured data and providing insights into feature importance. Furthermore, its ability to capture complex relationships within the data proved invaluable for our predictive objectives.

First, a random forest classifier was implemented for the whole training dataset and the results are as follows. (Red values represents the test data)

Class	Precision	Recall	F1-score
0	1.0 1.0	1.0 1.0	1.0 1.0
1	1.0 1.0	1.0 1.0	1.0 1.0

Table 1:Evaluvation of Random Forest model

As we can observe, all the observations in both the training and testing datasets were correctly classified. This model has used all twelve predictor variables.

3.1 Feature Selection

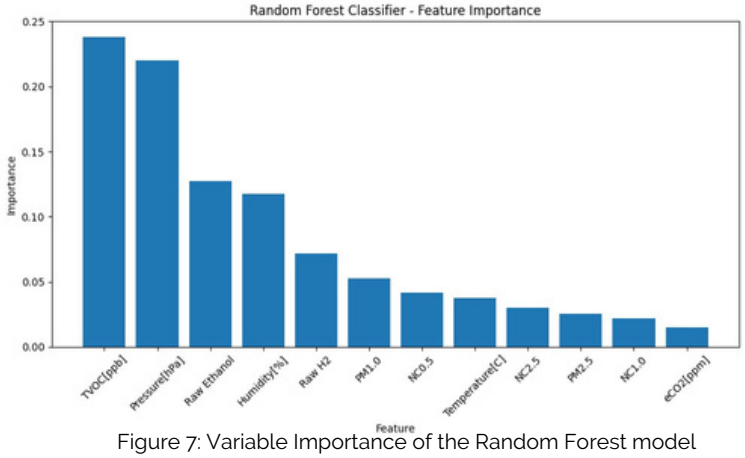

Figure 7: Variable Importance of the Random Forest model

When considering the feature importance graph of the random forest classifier model, it suggests that the TVOC, pressure, raw ethanol and humidity are the top four features that are contributing the most to the model's predictions. This suggests that these features have the highest importance or influence in determining whether a certain outcome or class is predicted by the model.

Therefore, a new random forest model will be fitted using these four features and the results of the new model are as follows

Class	Precision	Recall	F1-score
0	1.0 1.0	1.0 0.9994	1.0 0.9997
1	1.0 0.9998	1.0 1.0	1.0 0.9999

Table 2:Evaluvation of Random Forest model after removing some features

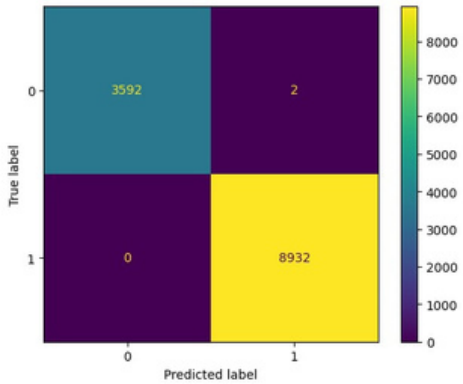

Figure 8: Confusion Matrix of the Random Forest model

The accuracy of the test dataset of the new model is 0.99984 which is quite a high accuracy. As the new model is simpler and provides similar or slightly lower accuracy compared to a more complex model, it may be a better choice due to its interpretability. Therefore, this parsimonious model will be used for further analysis.

3.2 Model Interpretation

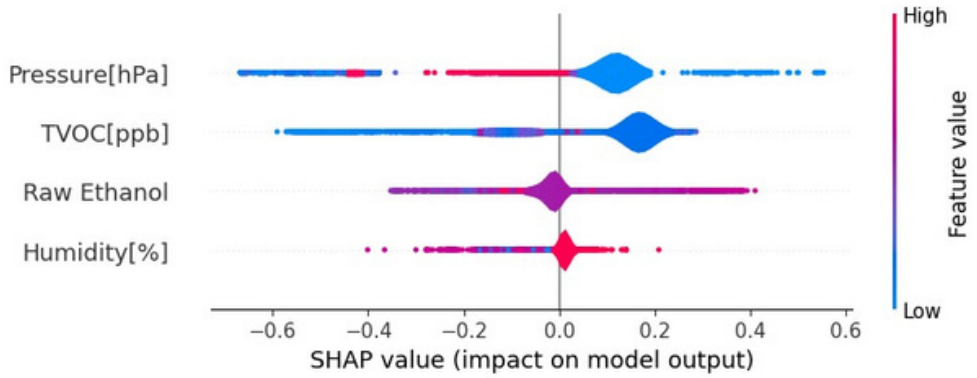

Figure 9: Summary plot of SHAP values

The SHAP (SHapley Additive exPlanations) summary plot with a violin plot type is a visualization that helps to interpret the contributions of different features to individual predictions made by the model. It provides insights into how each feature influences model predictions across the dataset.

When considering the pressure variable, Instances with low pressure values are associated with an increased probability of being classified as 1. In other words, when pressure levels are low, the model is more likely to predict the presence of a fire alarm. Additionally, higher humidity levels also have increased probability of being classified as 1.

This figure plot provides a visual representation of how different features contribute to the predicted value or output for the first instance in the test dataset. It's true class is 1.

Now E[f(x)] = 0.715 gives the average predicted log odds across all observations in the test data. That is the log odds of a positive (1) prediction.

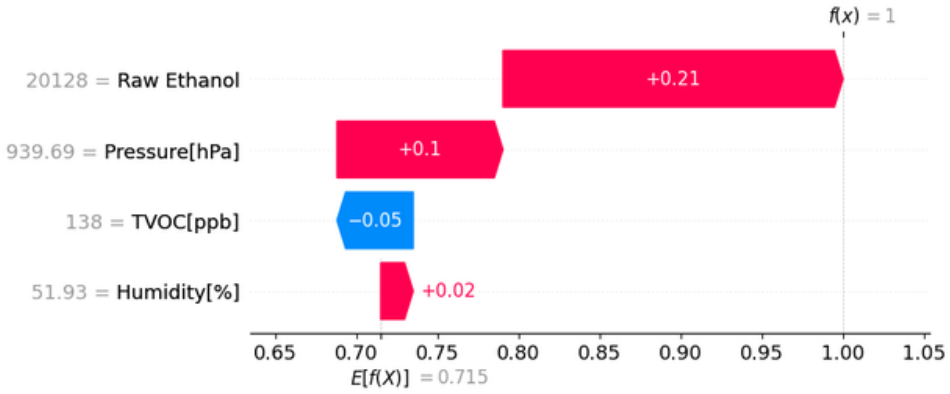

Figure 10: Waterfall Plot of first test observation

For this specific observation, the model predicted a probability of 1 that it belongs to class 1. The SHAP values give the difference between the predicted log odds and the average predicted log odds. Positive SHAP values increase the log odds. For example, raw ethanol level of 20128 increased the log odds by 0.21. In other words, this feature has increased the probability that the model will predict class 1. Similarly, negative values decrease the log odds.

4.0 BEST MODEL SELECTION

In our pursuit of finding the most effective model for our classification task, we will conduct a thorough assessment using various metrics and ROC curves of the test dataset. These metrics include accuracy, precision, recall, F1-score, and the confusion matrix, each offering a unique perspective on the model's performance. Additionally, the ROC curves will provide visual representations of how well the models differentiate between positive and negative cases at different classification thresholds. Our model selection process will consider a combination of these factors to ensure that the chosen model not only excels in accuracy but also aligns with the specific requirements of our application, ensuring its reliability and effectiveness.

Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
Logistic Regression	0.89	0.94	0.87	0.65	0.98	0.77	0.93
Logistic Regression with Ridge	0.87	0.93	0.86	0.58	0.98	0.72	0.91
Logistic Regression with Lasso	0.86	0.93	0.84	0.54	0.98	0.68	0.91
Logistic Regression with ElasticNet	0.86	0.93	0.84	0.54	0.98	0.68	0.91
Random Forest	0.9998	1.0	0.99	0.99	1.0	0.99	0.99

Table 3 :Evaluvation of all models

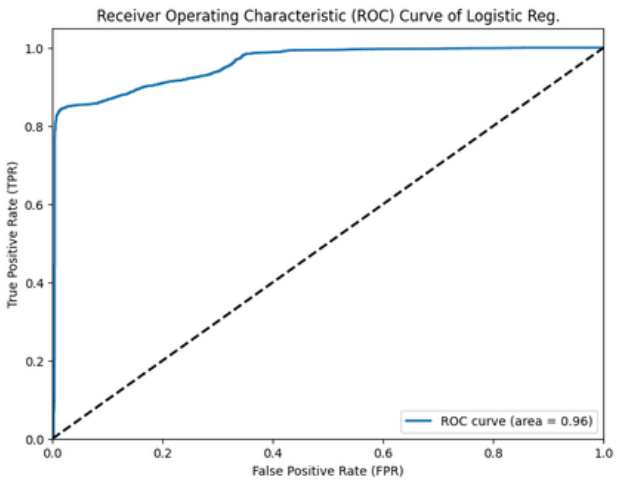

Figure 11: ROC Curve of Logistic Regression

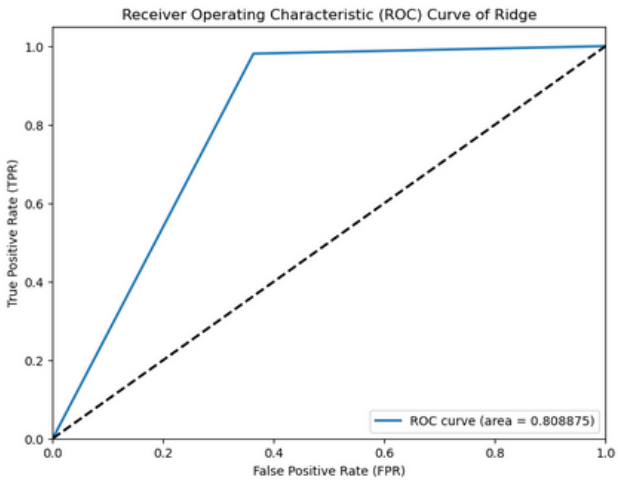

Figure 12: ROC Curve of Logistic Regression with Ridge

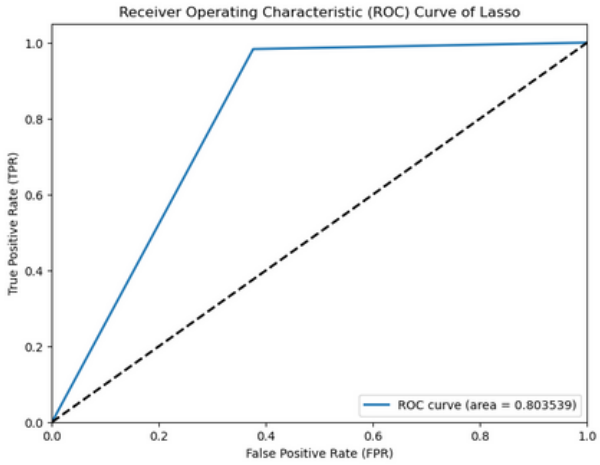

Figure 13: ROC Curve of Logistic Regression with Lasso

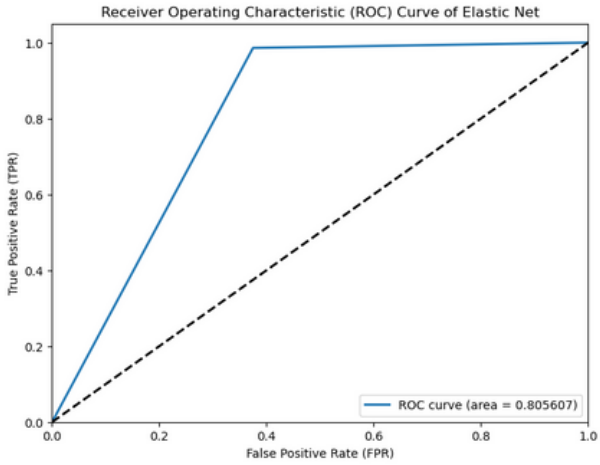

Figure 14: ROC Curve of Logistic Regression with Elastic Net

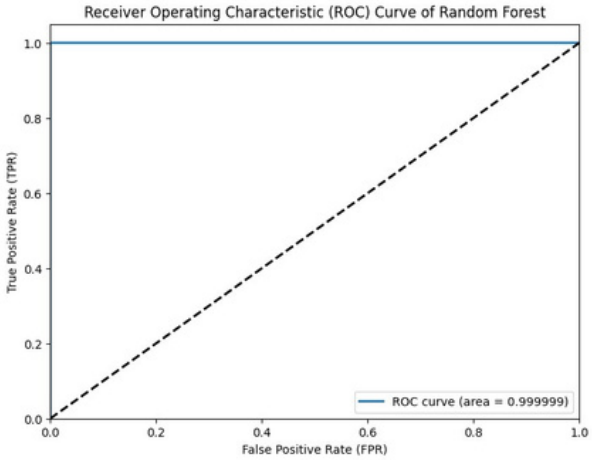

Figure 15: ROC Curve of Random Forest

We will select the Random Forest model as our final classification model due to its outstanding performance across various critical metrics. This choice is substantiated by its highest accuracy rate compared to other models, alongside excellent precision, recall, and F1-score values. The ROC curve, a pivotal indicator of model effectiveness, also illustrates the model's proficiency in distinguishing between positive and negative cases. This strong overall performance, especially in the ROC curve, solidified our decision to employ the Random Forest model as the cornerstone of our predictive solution, ensuring we meet and exceed our project objectives. Hence, a random forest model with the same parameters will be fitted to the whole dataset obtained by combining the training and testing dataset and will be used as the final model. This final model will be used for further predictions.

Appendix : Python Codes for Advanced analysis on Smoke Detection