Logistic Regression:

Logistic Regression is used to classify Twitter profiles as either human or non-human, making it a binary classification task. To handle the imbalance in the dataset (with more non-human profiles than human profiles), the parameter class_weight='balanced' is applied. This automatically adjusts the weights of each class to give more importance to the minority class (humans). Additionally, hyperparameter tuning is performed using GridSearchCV to find the optimal regularization parameter (C). This ensures that the model generalizes well by balancing underfitting and overfitting. To further validate the robustness of the model, cross-validation is applied to assess the model's performance across different data splits.

SMOTE (Synthetic Minority Over-sampling Technique):

SMOTE is applied to address the class imbalance in the dataset, where there are more non-human profiles than human profiles. SMOTE generates synthetic samples for the minority class (humans), creating a more balanced training set. This helps the model to learn the distinctions between the two classes (human and non-human) more effectively and improves its ability to generalize.

Linear Regression:

Although Linear Regression is traditionally used for regression tasks, it can be adapted for binary classification by interpreting its continuous output as a probability. The output of the Linear Regression model is a continuous value between 0 and 1, and we apply a threshold (0.5) to convert these outputs into binary classifications: 1 for human and 0 for non-human. Like with Logistic Regression, the features are standardized using StandardScaler, ensuring that features with different scales are treated equally.

Limitations of Linear Regression for Classification:

Although Linear Regression can be adapted for classification, it is not naturally suited to this task. Unlike Logistic Regression, which directly models the probability of a binary outcome, Linear Regression predicts continuous values, making it less robust in handling imbalanced data or ensuring well-calibrated probabilities.

Metrics Used:

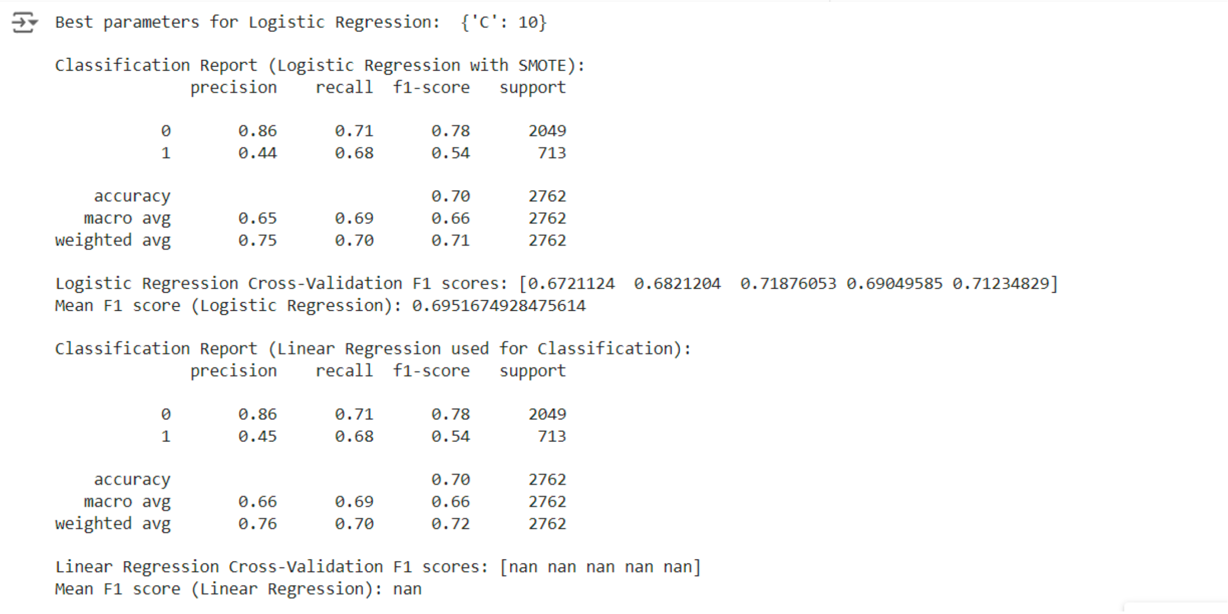
Precision: The ratio of true positive predictions to the total predicted positives, indicating how many of the profiles predicted as human were.

Recall: The ratio of true positives to all actual positives, indicating how many of the true human profiles were identified.

F1-score: A balance between precision and recall, which is useful when we need a single metric to account for both false positives and false negatives.

Cross-validation F1-scores: Provide insight into how well the models generalize across different data splits.

RESULTS



```
Best parameters for Logistic Regression:  {'C': 10}

Classification Report (Logistic Regression with SMOTE):
              precision    recall  f1-score   support

           0       0.86      0.71      0.78      2049
           1       0.44      0.68      0.54       713

    accuracy                           0.70      2762
   macro avg       0.65      0.69      0.66      2762
weighted avg       0.75      0.70      0.71      2762

Logistic Regression Cross-Validation F1 scores: [0.6721124  0.6821204  0.71876053 0.69049585 0.71234829]
Mean F1 score (Logistic Regression): 0.6951674928475614

Classification Report (Linear Regression used for Classification):
              precision    recall  f1-score   support

           0       0.86      0.71      0.78      2049
           1       0.45      0.68      0.54       713

    accuracy                           0.70      2762
   macro avg       0.66      0.69      0.66      2762
weighted avg       0.76      0.70      0.72      2762

Linear Regression Cross-Validation F1 scores: [nan nan nan nan nan]
Mean F1 score (Linear Regression): nan
```

Logistic Regression Performance:

Logistic Regression performs well with an overall accuracy of 70% and a mean F1-score of 0.695. It shows a good balance between precision and recall, particularly for the majority class (non-human). However, the precision for the minority class (human) is lower, meaning the model produces more false positives when predicting human profiles.

The cross-validation F1 scores are consistent across different splits, confirming that the model generalizes well to unseen data.

Linear Regression Performance:

When used for classification, Linear Regression produces similar results to Logistic Regression, with an accuracy of 70% and comparable F1-scores across both classes. It even slightly outperforms Logistic Regression in terms of precision for the human class (0.45 vs. 0.44).

However, the cross-validation results for Linear Regression show that it is not well-suited for classification, as evidenced by the NaN values. This is because Linear Regression is designed for predicting continuous outcomes, not binary classification tasks. The inability to handle class imbalance effectively also limits its performance.

CONCLUSION

Logistic Regression is the preferred model for this binary classification task, as it is specifically designed to handle classification problems and performs well, even with class imbalance, when combined with SMOTE.

Linear Regression can be used as a comparison model, but it is not recommended for classification tasks due to its design for continuous prediction. Although it produces similar results in this scenario, its limitations are highlighted by the failure in cross-validation.