**Random Forest Classifier Model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Original (imbalanced) dataset | | | Under Sampled Dataset | | |
| Metrics | 10 Features | 20 Features | 37 features | 10 Features | 20 features | 37 features |
| Accuracy | 0.91 | 0.90 | 0.90 | 0.71 | 0.73 | 0.73 |
| Precision | 0.43 | 0.38 | 0.31 | 0.72 | 0.72 | 0.72 |
| Recall | 0.08 | 0.12 | 0.12 | 0.68 | 0.77 | 0.77 |
| F1-score | 0.13 | 0.18 | 0.17 | 0.70 | 0.74 | 0.74 |
| AUC | 0.77 | 0.78 | 0.79 | 0.77 | 0.80 | 0.80 |

**Results overview:**

**Analysis of Original (Imbalanced) dataset without under sampling:**

1. **High accuracy but low recall:**

* Across all datasets , accuracy is relatively high (around 0.90-0.91). However, recall is very low, particularly in the 10-feature dataset (0.08). This means the model is not identifying true positive cases (people with a heart disease) effectively.
* Given the context , might need to prioritize having higher recall over high accuracy , as missing cases of heart diseases can have serious consequences.

1. **Decline in Precision and F1-Score as Features Increase:**

* Precision and F1-score decrease slightly as more features are added. For instance, precision goes from 0.43 in the 10-feature dataset to 0.31 in the 37-feature dataset.
* This hints at the possibility that adding more features may be introducing noise rather than useful information, potentially because the new features may not be strongly correlated to the target variable (heart disease).
* Model could also be overfitting .

1. **AUC score:**

A line graph with different colors

Description automatically generated

* AUC values are in the same range across the datasets (ranging from 0.77 to 0.79), with a slight increase as more features are added.
* The curves are well above the diagonal baseline, which represents random guessing (AUC = 0.5), indicating that the model performs better than random chance.
* Given the ROC curve's shape and the low recall scores reported earlier, it appears the model has difficulty detecting true positive cases (heart disease cases) with high sensitivity. This limitation is consistent with the low recall and F1-scores observed in the earlier metrics.

**Conclusion**

The current models provide high accuracy but struggle with low precision, recall and overall F1-score, which is a significant drawback for heart disease prediction. Improving recall should be a priority, possibly through rebalancing techniques, feature selection etc.

**Analysis of Under sampled dataset :**

1. **Recall Significantly Improved**:

* In the original imbalanced datasets, recall was very low (0.08 to 0.12), indicating the model was missing most cases of heart disease. With under sampling, recall increased to 0.68, 0.77, and 0.77 across the three feature sets, making the model much more effective at identifying true positives.

**2. Precision and F1-score:**

* Precision increased substantially, from as low as 0.31 in the original datasets to 0.72 across all under sampled datasets. The models are now more accurate in the positive predictions they make, while also capturing more true positives.
* F1-scores also improved greatly (0.13-0.18 in the original datasets versus 0.70-0.74 in the under sampled ones), indicating that the models now have a more balanced precision and recall.

**3.Accuracy :**

* Accuracy has dropped in the under sampled datasets (from ~0.90 to ~0.71-0.73), which is expected when using under sampling. This decrease in accuracy is likely due to the balanced nature of the dataset, where fewer negative cases may lead to an increase in false positives.

**4.AUC:**

A line graph with different colors

Description automatically generated

* AUC has improved slightly (from ~0.77-0.79 to ~0.77-0.80).
* There is an increase for the 20 and 37 feature sets compared to the original dataset, suggesting an improvement in model performance due to under sampling. The 10-feature AUC remains the same (0.77), indicating that under sampling mainly benefits models with a larger feature set.

1. **Feature count impact:**

* With under sampling, the 20 and 37-feature models both achieved high recall and F1-scores.
* There is no improvement in the metric values between the 20-feature dataset and the 37-feature dataset . Thus, adding more features beyond the initial 20-feature dataset may be redundant due to low correlation to the target variable.

**Overall summary and conclusion:**

The **accuracy** for the original dataset was relatively high due to the class imbalance, as the model favoured predicting the majority class (No Heart Disease). However, high accuracy can be misleading in an imbalanced dataset, as it may mask the model's inability to detect the minority class (heart disease).

The **accuracy** of the model decreased after under sampling. This is expected since under sampling reduces the number of majority class samples, making the model's accuracy more representative of its performance on both classes. The decrease in accuracy here is not necessarily negative, as it indicates a shift towards a more balanced detection of heart disease cases.

**Precision** is lower across the models (10, 20, and 37 features) on the original dataset compared to the under sampled dataset, with values of 0.43, 0.38, and 0.31 respectively. This lower precision reflects that the model was unable to reliably predict heart disease, with a high false positive rate due to class imbalance.

Additionally, the low **recall** values (0.08 to 0.12) from the original dataset confirm that many actual heart disease cases were missed, making this model unreliable for identifying the minority class (heart disease) .

Under sampling the dataset to address the class imbalance has been effective . It has a more balanced precision and F1-score and addressed the model's original limitation of low recall, making it far better at identifying cases of heart disease.

**Feature set selection:**

The 20-feature model consistently showed the best performance balance for both original and under sampled datasets. Increasing the features to 37 did not provide substantial improvements, suggesting that 20 features may be the trade-off .

**Importance of Evaluation Metrics**:

* The **recall** metric, along with **F1-Score** and **AUC**, proved to be the most reliable indicators of performance for this application. **Recall** is critical because it indicates the model’s ability to detect true heart disease cases, which is the primary objective in a healthcare setting.
* **Accuracy** was misleading in the original dataset due to class imbalance; it was high but masked the models poor recall and precision for the heart disease class.
* The **AUC** for the under sampled datasets (0.77 to 0.80) indicates that the model has a moderate ability to distinguish between heart disease and no heart disease, and it performed similarly across feature sets.