Calibration Task Evaluation

Team Members

Aisha Hagar

Models:

- Ali Sameh
- Hassaan Gamal
- Yara Mahfouz
- Ziad Abdlhamed

Aya Khaled

Models:

- Ali Badawy
- Mariam Seedawy
- Mohamed Ezzat
- Rana Hossny
- Ziad Mahmoud

Data insights:

Data distribution in train dataset class

data distribution in train dataset class	
XGBRegressor	177
HUBERREGRESSOR	85
LinearSVR	69
LASSO	46
QUANTILEREGRESSOR	17
ELASTICNETCV	6

Data distribution in test dataset class

data distribution in test dataset class	
XGBRegressor	44
HUBERREGRESSOR	22
LinearSVR	17
LASSO	11
QUANTILEREGRESSOR	5
ELASTICNETCV	1

Our focus:

1- **Calibration.**
2- **False Negatives**

Why? if the model predicted some class to be the best while it's not (false +ve), its not a big deal, I can test it my self and see if the results are sufficient or not, but id the model fail to predict a well preforming regression model and though it away (false -ve), how many models do I have to test in this pool to identify the good one.

Since:

- The data distribution is highly imbalanced.
- **Accuracy** matrix is less informative, misleading, especially in such imbalanced datasets, while **Precision** focuses on reducing false positives.
- **Our focus is (Calibration + False Negatives)**

We decided to choose:

Critical Metrics:

- **Recall**:
  - Recall measures how well the model identifies all relevant positive cases , reducing false negatives (e.g., failing to identify a well-performing model) are critical, ensuring that the model identifies the optimal regression model in each dataset.

- **F1-Score**:
  - A balanced metric that combines precision and recall, especially useful in imbalanced datasets.

- **Balanced Accuracy:** Balances recall across all classes, ensuring that minority classes aren't neglected and equally weighted, that helps addressing false negatives in underrepresented classes while providing a fair, balanced view of performance across classes
- **Expected Calibration Error (ECE)**
  - For understanding calibration at the class level

Performance Comparison (Test data)

Data Team Member	Recall (Weighted)	F1-Score (Weighted)	Balanced Accuracy	Total ECE
Ali Badawy	46%	45%	27%	0.998
Ali Sameh	30%	29%	16%	0.739
Hassaan Gamal	61%	58%	41%	0.607
Mariam Seedawy	6%	4%	22%	0.843
Mohamed Ezzat	55%	53%	31%	0.744
Rana Hossny	59%	58%	36%	0.764
Yara Mahfouz	61%	56%	38%	0.614
Ziad Abdlhamed	23%	17%	16%	0.451
Ziad Mahmoud	51%	47%	26%	0.749

Performance Comparison (Train data)

Data Team Member	Recall (Weighted)	F1-Score (Weighted)	Balanced Accuracy	Total ECE
Ali Badawy	88%	87%	80%	0.199
Ali Sameh	30%	29%	16%	0.904
Hassaan Gamal	96%	96%	92%	0.77
Mariam Seedawy	20%	20%	16%	0.283
Mohamed Ezzat	95%	94%	78%	0.16
Rana Hossny	87%	87%	80%	0.113
Yara Mahfouz	65%	62%	43%	0.582
Ziad Abdlhamed	23%	19%	16%	0.439
Ziad Mahmoud	89%	88%	83%	0.164

Comments on Performance Metrics

Ali Badawy's model

- On the training data, the model demonstrates strong performance, with a high recall (87.50%), F1 score (87.40%), and balanced accuracy (80.07%). On test Data, recall drops to 46%, the F1 score to 45.18%, and balanced accuracy to a low 26.78%, all pointing to the model's inability to generalize to unseen data. These oncourse are signs to significant overfitting.
- However, the total Expected Calibration Error (ECE) for the training data is 0.1985, indicating moderate misalignment between predicted probabilities and actual frequencies, though still acceptable given the high overall performance, but the test data shows a sharp decline in performance compounded by an extremely high ECE

of 0.9981, indicating the poor calibration as the model is overly confident in its predictions on the test set, even when it is incorrect

Ali Sameh's model

- The recall, F1 score, and balanced accuracy are extremely low on both the train and test data. This indicates that the model is underfitting and unable to learn the patterns of each class.

Hassaan Gamal's model

- The model achieved the highest recall, F1 score, and balanced accuracy on both the train and test data.
- However, there is a huge gap between the train and test scores. For example, the balanced accuracy of the train data is 92% but for the test data it is 41%. Thus, the model is clearly overfitting.

Mariam Seedawy's model

- For the training data, the recall and F1 score are extremely low (around 20%), showing that the model fails to correctly identify the majority of the true classes. The balanced accuracy of 16.14% suggests that the model is highly biased toward dominant classes and is not effective at distinguishing between classes. For the test data, Recall and F1 score are very low (6% and 4.38%, respectively), highlighting the inability of the model to predict the true classes in unseen data. The balanced accuracy of 22.95% is also extremely poor, indicating that the model is almost guessing randomly. The model poorly on both the training and test data, meaning it fails to learn meaningful patterns from the data. This is a classic sign of underfitting,

- The total ECE of 0.2830 reveals moderate miscalibration even on the training data, suggesting that the predicted probabilities are not well-aligned with true class frequencies, while total ECE of 0.8429 underscores indicating severe miscalibration on test data, with the model being highly overconfident in its incorrect predictions on the unseen test data.

Mohamed Ezzat's model

- On the training data, the high recall (94.5%) and F1 score (93.8%) indicate that the model has learned the training patterns effectively. On test F1 score (52.97%) drop sharply, reflecting difficulty in maintaining the same level of prediction quality for unseen data. The low balanced accuracy in both points to the inability to handle

class imbalances during both train and generalization, overall the performance on the test data highlights **significant overfitting.**

- The moderately low ECE (0.1595) on the training data further confirms that the model is reasonably calibrated for predictions on known data, high ECE (0.7439) indicates that the model's predicted probabilities are poorly calibrated on the test data, making it overly confident in incorrect predictions.

Rana Hossny's model

- The model performs well in terms of recall and F1 score, suggesting good learning on the training data, but struggles to generalize to unseen test data, as indicated by the substantial drop in recall, F1 score, and balanced accuracy, which indicate overfitting, also the slightly lower balanced accuracy compared to recall and F1 hints at mild bias toward dominant classes.
- The low ECE on train data shows that the model's predicted probabilities are well-calibrated, while high ECE on test data highlights poor calibration, further undermining the reliability of the model's predicted probabilities, as it revels overly confident in its incorrect predictions.

Yara Mahfouz's model

- The model achieved the highest recall of 61% on the test data. It achieved the second-highest f1-score and balanced accuracy of 56% and 38% respectively. The model ranks third in terms of ECE.
- The model's performance is consistent on the train and test data. There isn't a huge gap between the scores. Therefore, the model is not overfitting.

Ziad Abdlhamed's model

- Although the model has the lowest ECE of 0.451 on the test data, the model's performance is extremely poor.
- The recall, F1 score, and balanced accuracy are extremely low on both the train and test data. This indicates that the model is underfitting and unable to learn the patterns of each class.

Ziad Mahmoud's model

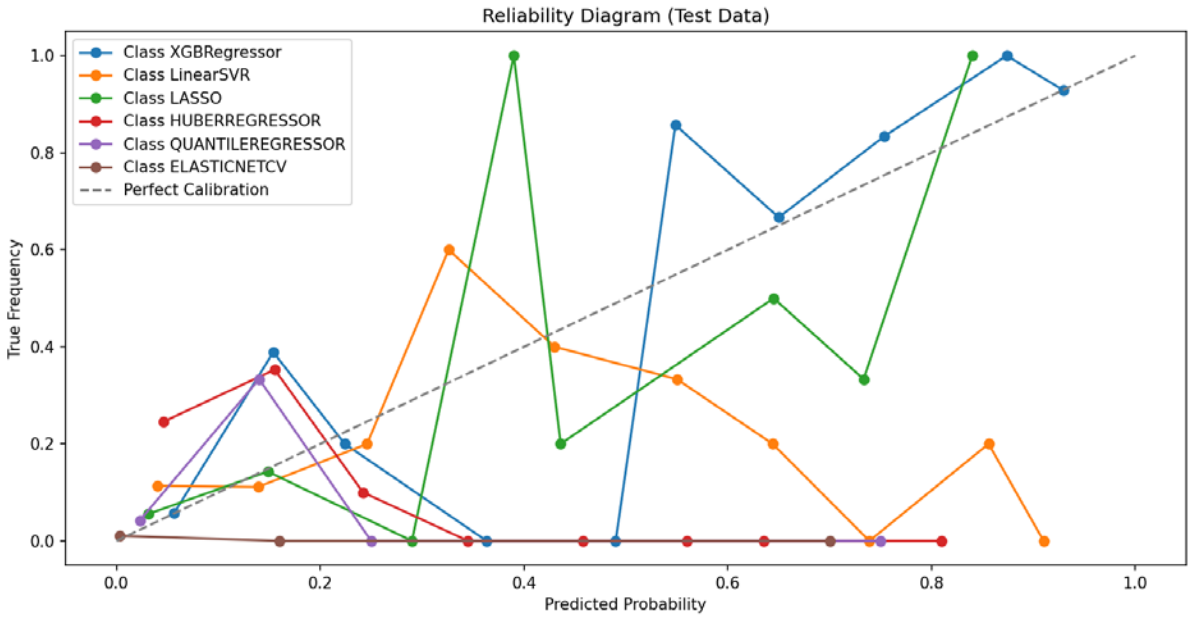
- The model shows strong performance on the training data specially balanced accuracy (83.13%). These metrics suggest the model effectively learns patterns from the training set and maintains a reasonable balance across different classes. However, the substantial drop in metrics on the test data, with recall at

51%, F1 score at 46.96%, and balanced accuracy decreased to 26.43%, indicates significant overfitting. The model struggles to handle class imbalance effectively and generalize to unseen data, likely memorizing the training data instead of capturing generalizable patterns.

- Moreover, while the calibration on the training data appears relatively good, with an ECE of 0.1641, the high test ECE of 0.7493 reveals that the model is overly confident in its incorrect predictions, further undermining the reliability of its probability estimates.

Reliability Diagrams (Test data)

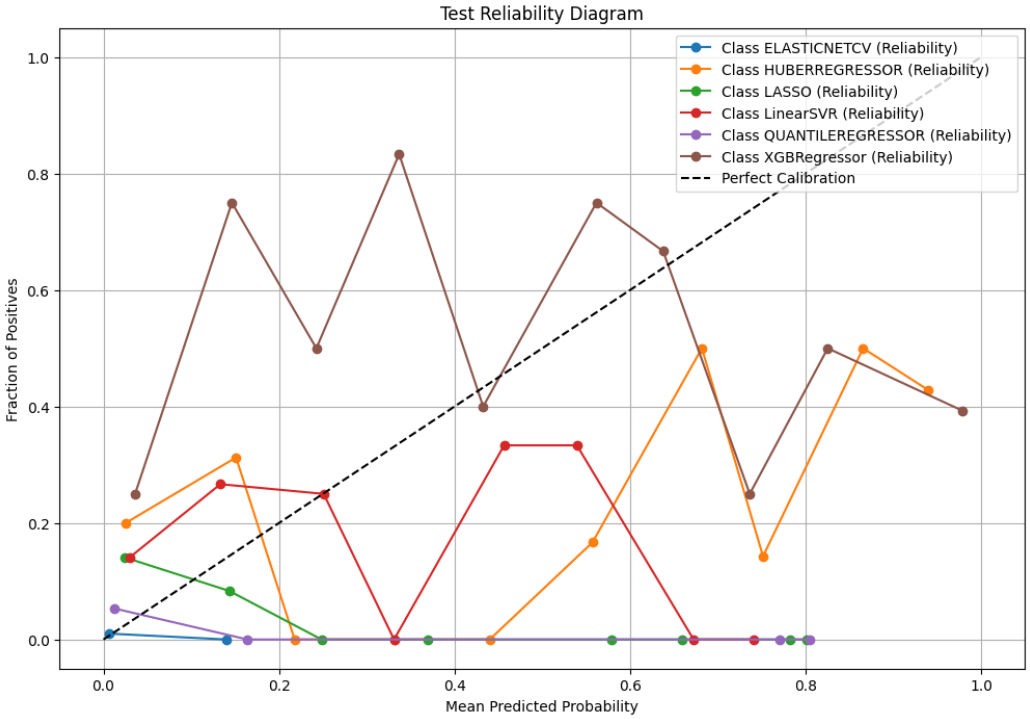
Ali Badawy's model



Ali Badawy's Reliability Diagram

- The reliability diagram reveals significant calibration issues, with most classes deviating substantially from the diagonal line, indicating that the model's predicted probabilities on test data are unreliable. Classes like XGBRegressor and LinearSVR show cerratic and overconfident predictions, while others like HUBERREGRESSOR and QUANTILEREGRESSOR seem underrepresented. This poor calibration aligns with the high Expected Calibration Error (ECE) for test data and suggests the model struggles to generalize and accurately represent probabilities, leading to unreliable predictions in real-world scenarios.

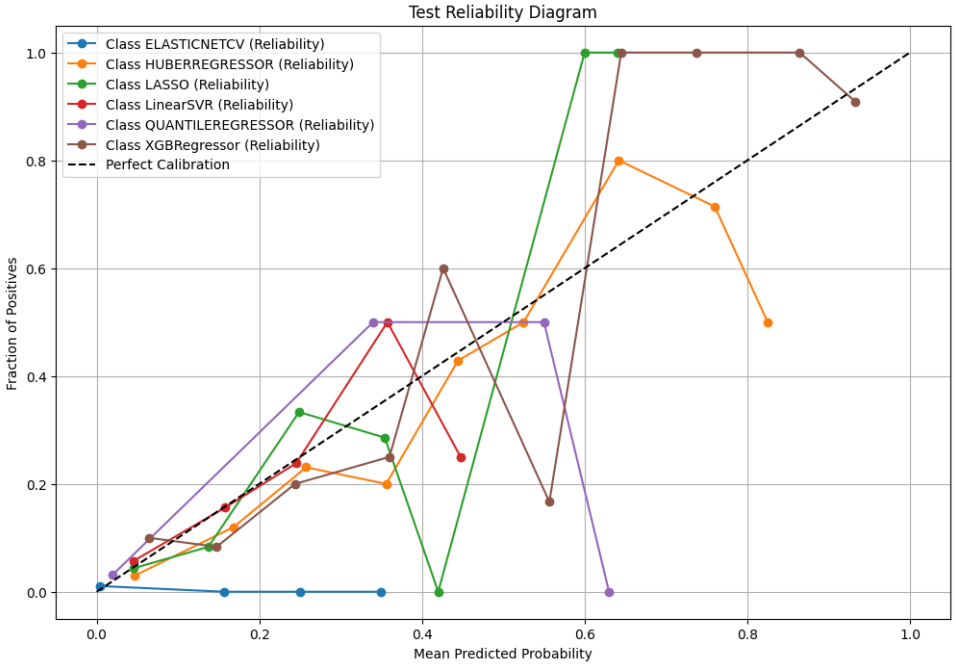
Ali Sameh's model



Ali Sameh's Reliability Diagram

- From the reliability diagram, the model is overconfident when predicting *ELASTICNETCV*, *QUANTILEREGRESSOR*, and *LASSO* classes.
- None of the classes are well-calibrated.

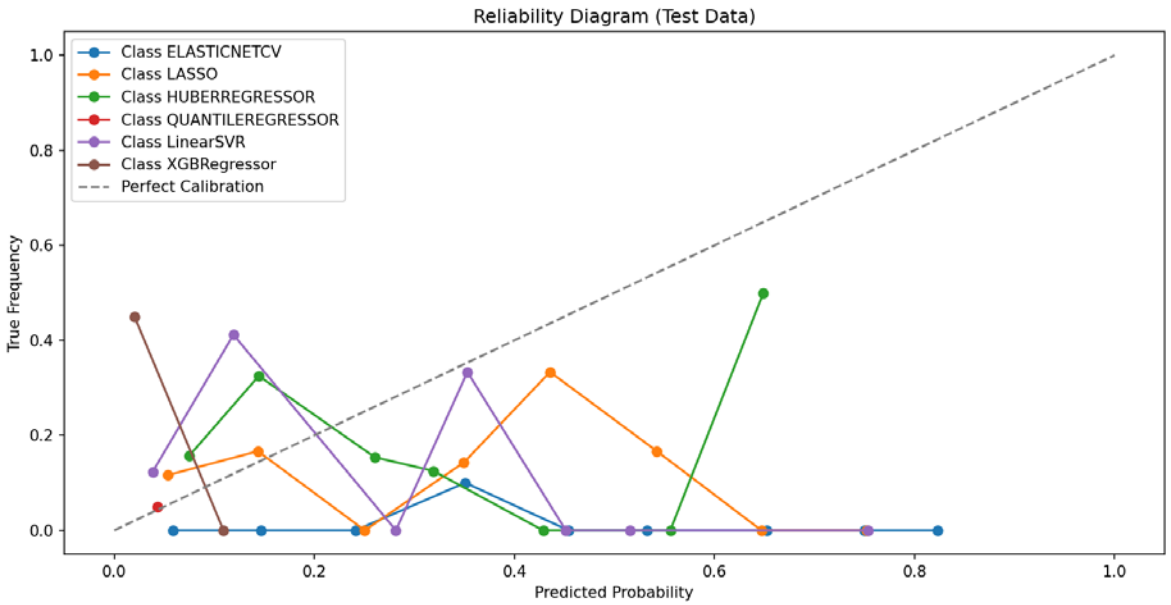
Hassaan Gamal's model



Hassaan Gamal's Reliability Diagram

- From the reliability diagram, the model is overconfident when predicting *ELASTICNETCV* class.
- The model is almost well-calibrated for *HUBERREGRESSOR* class. However, the model is uncalibrated over the rest of the classes.

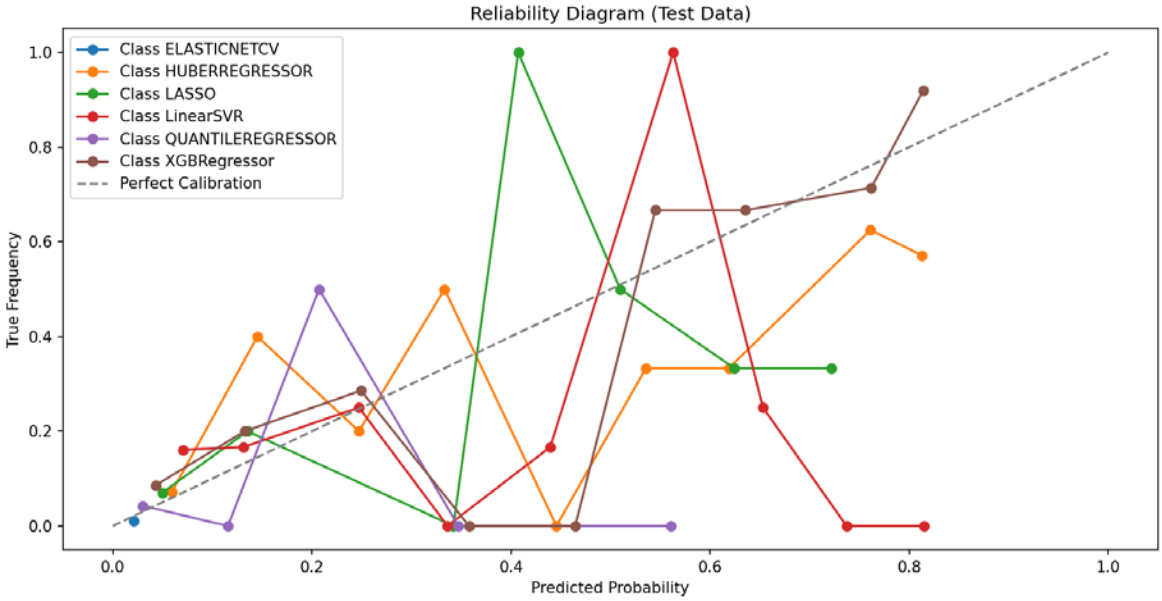
Mariam Seedawy's model



Mariam Seedawy's Reliability Diagram

- The jagged nature of the curves suggests that the model struggles to produce consistent probabilities across all bins, which is symptomatic of **underfitting** or inadequate training. Although some classes (e.g., QUANTILEREGRESSOR) exhibit slightly better calibration (closer to the diagonal in some bins), but the overall picture is one of poor reliability. The poor alignment with the diagonal reflects that the model is poorly calibrated on the test data, confirming the high Total ECE (0.8429) observed in the metrics.

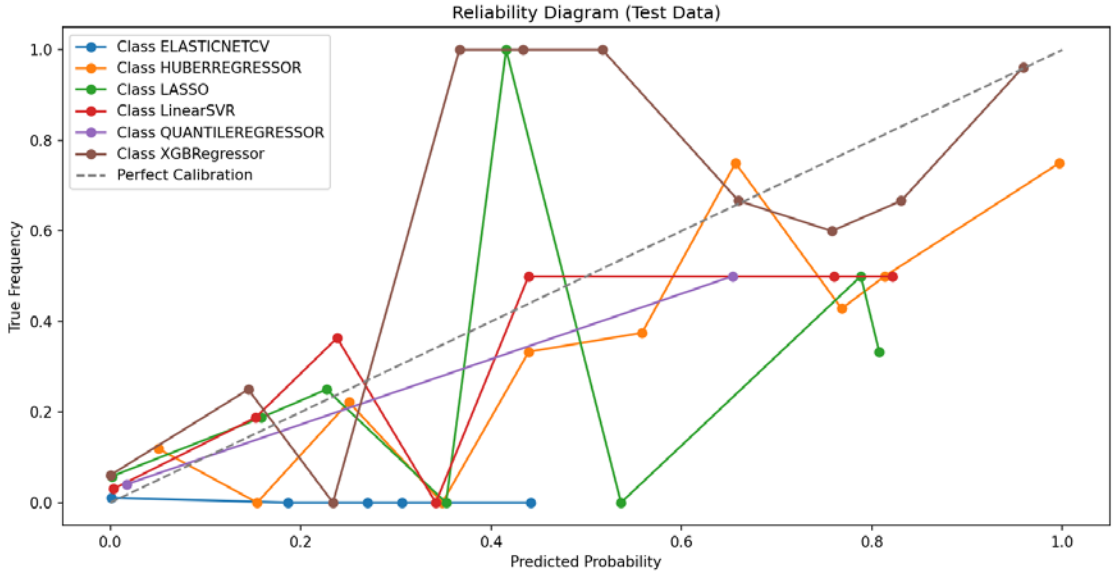
Mohamed Ezzat's model



Mohamed Ezzat's Reliability Diagram

- This reliability diagram for the test data highlights significant calibration issues across the classes. Most lines deviate substantially from the diagonal (representing perfect calibration), indicating that the model's predicted probabilities do not align well with the true frequencies for many classes.
- Overall, the diagram underscores poor calibration, which aligns with the high ECE observed in the test data.

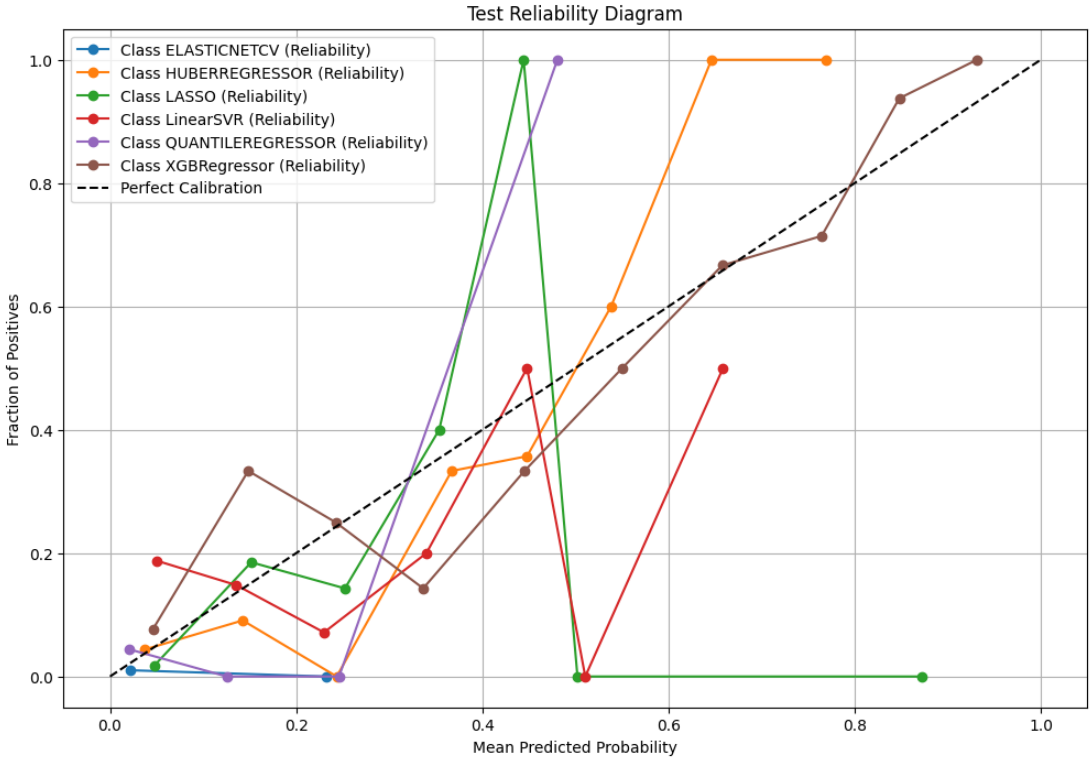
Rana Hossny's model



Rana Hossny's Reliability Diagram

- The reliability diagram highlights significant calibration issues, with most classes deviating noticeably from the diagonal, indicating that the model's predicted probabilities do not align well with the actual outcomes. While QUANTILEREGRESSOR shows relatively better calibration, other classes like HUBERREGRESSOR and ELASTICNETCV show severe overconfidence in incorrect predictions. This poor calibration is consistent with the high ECE observed for the test data, emphasizing that the model struggles to generalizing to unseen data.

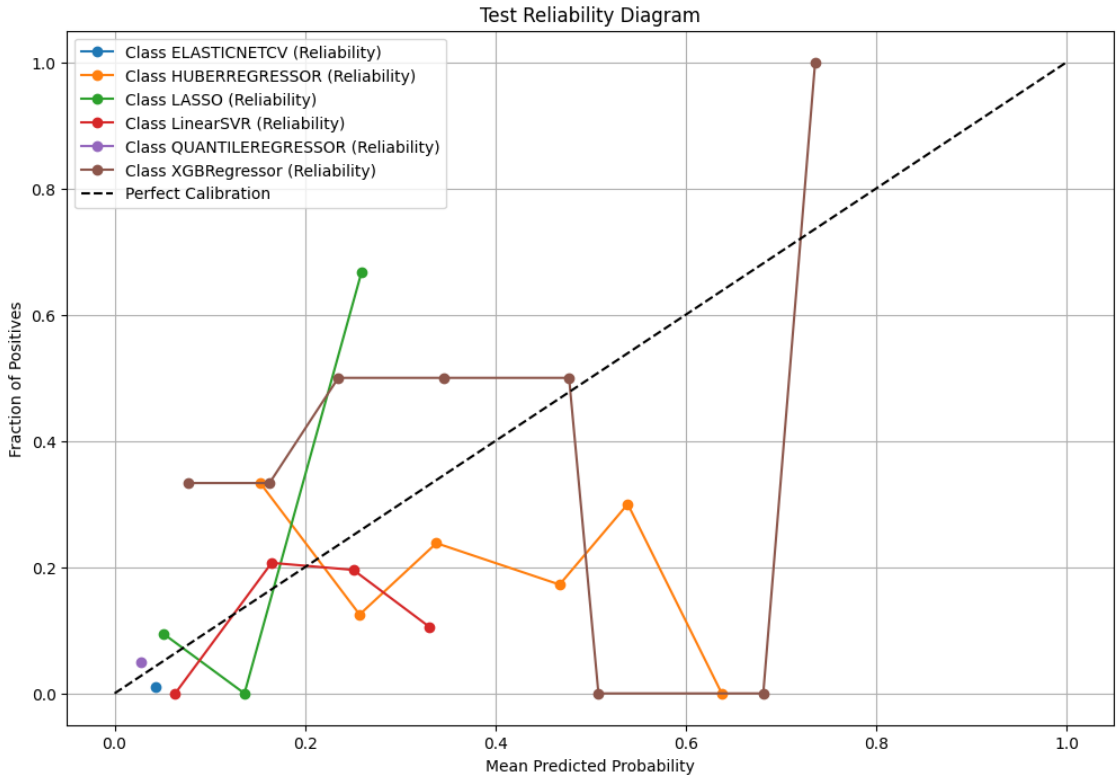
Yara Mahfouz's model



Yara Mahfouz's Reliability Diagram

- The model is almost well-calibrated for *XGBRegressor* class.
- However, the model is uncalibrated over the rest of the classes.
- The model is overconfident when predicting *ELASTICNETCV* and *LINEARSVR* classes.

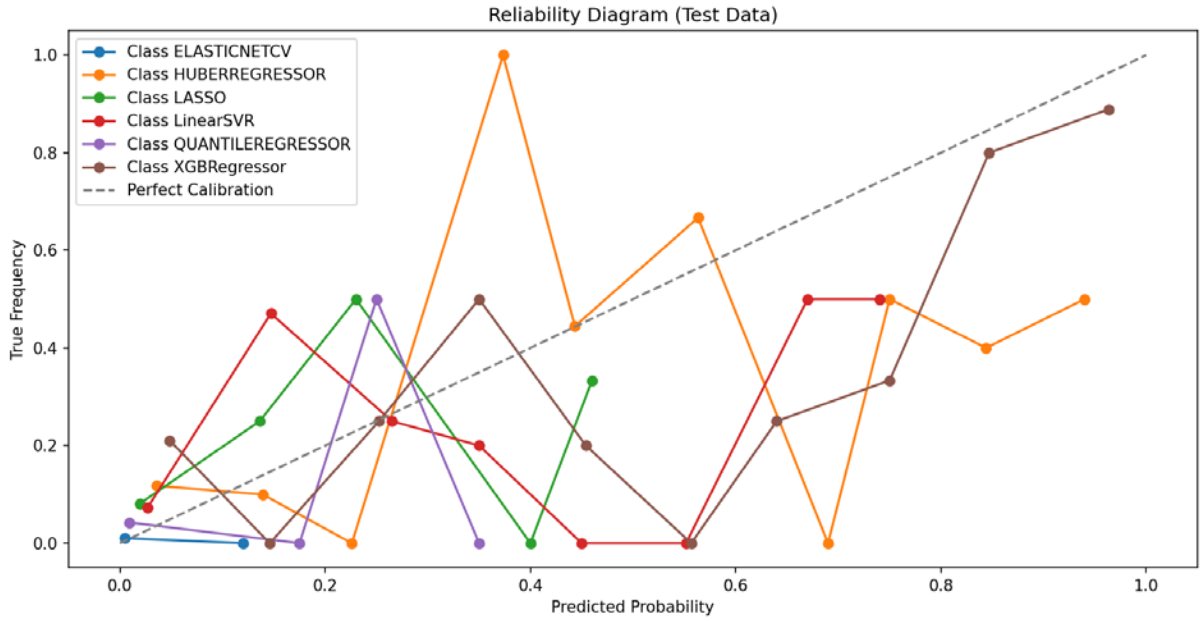
Ziad Abdlhamed's model



Ziad Abdlhamed's Reliability Diagram

- The model is uncalibrated over all the classes.

Ziad Mahmoud's model



Ziad Mahmoud's Reliability Diagram

- Some classes, like QUANTILEREGRESSOR, show better alignment with the diagonal in parts, indicating relatively better calibration in specific probability ranges. However, overall, the model fails to maintain consistent calibration, with many classes showing erratic or extreme deviations from the ideal line. This inconsistency aligns with the high ECE value observed in the metrics for test data, further emphasizing that the model's predictions on unseen data are unreliable and poorly calibrated.

Conclusion

These results collectively emphasize the need for better regularization, improved calibration, and enhanced handling of class imbalance to generalize well to new data.