To compare the candidate models based on their performance metrics (Accuracy, Recall, F2 Score, and AUC), we can analyze each metric individually and then consider the overall implications of each.

Metrics Analysis:

1. **Accuracy**:
   - All three models achieve an accuracy of approximately 0.87 to 0.91.
   - This indicates that all models are performing similarly in terms of overall prediction accuracy.

2. **Recall (True Positive Rate)**:
   - **Neural Network**: Recall is around 0.999, which is almost perfect.
   - **Gradient Boosting**: Recall is also high, with values close to 1.
   - **Logistic Regression**: Recall is moderate, at approximately 0.9988.
   - **Recall** is crucial in scenarios where the cost of missing a positive instance is higher than incorrectly classifying a negative instance.

3. **F2 Score**:
   - **Neural Network**: F2 Score is around 0.971, which is slightly lower than Gradient Boosting's ~0.704.
   - **Gradient Boosting**: F2 Score is the highest at approximately 0.704, indicating a good balance between precision and recall.
   - **Logistic Regression**: F2 Score is moderate, around 0.961.

4. **AUC (Area Under the ROC Curve)**:
   - **Neural Network**: AUC is around 0.695, which is lower than both Gradient Boosting and Logistic Regression.
   - **Gradient Boosting**: AUC is higher at ~0.703, suggesting better discrimination between classes.
   - **Logistic Regression**: AUC is moderate, around 0.683.

Comparison and Implications:

- **High Accuracy**: All models perform similarly in terms of overall prediction accuracy. This suggests they are effective at classifying instances correctly on average.

- **Gradient Boosting**:
  - **Top Recall**: The superior recall makes it a strong candidate, especially if precision is also important or if the cost of missing positive instances is high.
  - **Lower F2 Score vs AUC**: While Gradient Boosting has a higher F2 Score compared to others, its lower AUC suggests it may have more issues with distinguishing between classes.

- **Neural Network**:
  - Performs best in terms of Recall but struggles overall due to low AUC. It could be suitable if the dataset is complex and patterns are intricate.
  - The relatively high precision (for Gradient Boosting) might make it a good choice for scenarios requiring balanced predictions.

Visual Comparison:

To better understand these results, plotting each metric against each other or creating a performance matrix can help visualize how the models perform relative to each other. This would provide insights into which model excels in specific aspects and complement the metrics analysis.

Conclusion:

- **Top Performers**:
  - Gradient Boosting leads with high Recall, F2 Score, and AUC.

- **Considerations**:
  - If precision is a critical factor, consider looking at Precision for each model as well.
  - For understanding class distribution or handling imbalanced datasets, accuracy might be insufficient; alternative metrics like precision-recall curves would be more informative.

By evaluating these models using their specific needs and the context of their application, one can make an informed decision on which model to choose.