51. **Correct Cross-Validation Approach for Feature Selection in Predicting Medicare Eligibility**

**Abstract**

This paper investigates the impact of proper cross-validation techniques when predicting Medicare eligibility using logistic regression models. We specifically focus on implementing feature selection inside the cross-validation process, an approach that mitigates the risk of data leakage and overfitting. The results highlight the critical importance of correctly integrating feature selection within cross-validation loops to achieve reliable model performance assessments. The findings suggest that failure to do so can lead to overly optimistic estimates of model accuracy, whereas correctly executed cross-validation reflects a more realistic model performance.

**1. Introduction**

Predicting Medicare eligibility is a complex task influenced by multiple interacting factors, including patient demographics, diagnoses, treatment types, and cost-related variables. Cross-validation is a fundamental technique for evaluating model performance, but its implementation can significantly affect the reliability of the results. In particular, it is essential to ensure that feature selection occurs within each fold of the cross-validation process to prevent information from leaking from the test set into the training set.

This research uses a logistic regression model to predict Medicare eligibility, incorporating a 10-fold cross-validation procedure with proper feature selection inside each fold. The dataset includes various patient-related variables, such as age, gender, diagnosis types, treatment modalities, and costs.

**2. Methodology**

**A. Dataset Overview**

The dataset includes 11 variables related to patient characteristics and their Medicare eligibility status:

* **Patient\_ID**: Unique identifier for each patient.
* **Age**: Patient's age.
* **Gender**: Categorical variable indicating male or female.
* **Diagnosis**: Type of diagnosis (e.g., Hypertension, Arthritis, Diabetes, Cancer).
* **Treatment\_Type**: The type of treatment received by the patient (Medication, Surgery, Therapy).
* **Medicare\_Eligibility**: Binary outcome indicating whether a patient is eligible for Medicare.
* **Medicaid\_Eligibility**: Binary variable indicating Medicaid eligibility.
* **Medicare\_Coverage**: Percentage of treatment costs covered by Medicare.
* **Medicaid\_Coverage**: Percentage of treatment costs covered by Medicaid.
* **Total\_Treatment\_Cost**: Total cost of the medical treatment.
* **Out\_of\_Pocket\_Cost**: Amount paid out-of-pocket by the patient.

**B. Cross-Validation with Feature Selection**

A 10-fold cross-validation approach was employed to assess the performance of the logistic regression model. The feature selection was conducted inside each cross-validation fold to prevent data leakage. This method ensures that the feature selection process does not use information from the test data, leading to a more unbiased estimate of model performance.

**3. Results**

**A. Cross-Validation Accuracy Analysis**

The cross-validation results are summarized in the plot showing the accuracy of the logistic regression model for each fold.

* **Correct Cross-Validation (Feature Selection Inside CV):** The accuracy fluctuates between approximately 0.47 and 0.52 across the 10 folds. This variation reflects the model's sensitivity to different data subsets and demonstrates the importance of proper feature selection within the cross-validation process.
* **Interpretation of Results:**
  + **Folds with Lower Accuracy (Fold01, Fold03, Fold08):** These folds exhibit accuracy rates around 0.47-0.48, suggesting that the model struggled to generalize well to these particular subsets of the data. This may be due to unique patterns or noise within these folds that are not captured effectively by the logistic regression model.
  + **Folds with Higher Accuracy (Fold05, Fold06, Fold09, Fold10):** The accuracy for these folds reaches approximately 0.51-0.52, indicating better performance on these data subsets. However, the overall accuracy remains relatively low, highlighting that the model may not be sufficiently complex to capture all the nuances in the dataset.

The results suggest that while the model's average accuracy is slightly above random chance, it lacks robustness and generalizability. The incorporation of feature selection within each fold of the cross-validation process leads to a more realistic estimate of the model's true predictive performance.

**4. Discussion**

**Implications of Correct Cross-Validation**

The analysis underscores the critical importance of implementing proper cross-validation techniques to avoid overly optimistic performance estimates. By incorporating feature selection inside each cross-validation fold, we ensure that the model does not indirectly "see" information from the test data during training, which would otherwise inflate its performance metrics.

**Insights on Model Performance**

* **Feature Selection:** The modest accuracy observed suggests that the logistic regression model, even with proper feature selection, struggles to capture the complex relationships between patient characteristics and Medicare eligibility.
* **Model Complexity:** The findings indicate that the logistic regression model may be overly simplistic for this dataset. The low cross-validation accuracy suggests that more sophisticated models, such as ensemble methods or neural networks, could potentially offer better predictive performance by capturing non-linear relationships and interactions between variables.

**Recommendations for Future Research**

1. **Model Enhancement:** Future studies could explore more advanced modeling techniques, such as Random Forests, Gradient Boosting Machines, or Support Vector Machines, to improve predictive accuracy.
2. **Feature Engineering:** Additional work could focus on engineering new features that better capture the underlying patterns in the data, such as composite indices of patient health status or interaction terms between variables.
3. **Validation Strategies:** Expanding the validation strategies beyond simple cross-validation to include bootstrapping or nested cross-validation could provide more robust estimates of model performance.

**Conclusion**

This research highlights the importance of correct cross-validation procedures in predicting Medicare eligibility. Proper feature selection within cross-validation folds prevents data leakage, leading to more accurate performance estimates. While the logistic regression model shows limited predictive power, the study provides a foundation for further exploration using more advanced modeling approaches and enhanced feature engineering.