Project2 Report

N01589087

Anjali Prakash

1. When you prepare the data for training the models, did you discover any attribute to remove or any new attribute to add? If you did, discuss the choices.

   When I prepare the data for training the models, I removed the column Name because it is a categorical attribute. The TAR column, the target variable was changed to the integer format for compatibility with ML models. Also missing values from 3P% were imputed using mean imputation to prevent data loss.

2. Normalizing (a.k.a., scaling) features is desirable for distance-based models, e.g., k-nearest neighbors. Did you try feature normalization for some of the models? If so, talk about if any improvement.

   Feature scaling was performed using StandardScaler to normalize numerical features. This was beneficial for models like KNN and ANN. For ANN, CV F1 score improved to 0.7637 and test F1 score as 0.7811. The improvement in F1 scores suggest that normalization helped to stabilize and prevented bias toward certain attributes.

3. Regularization is a common practice to battle overfitting. How is varying the penalty parameter in logistic regression affect the performance F1 score on testing? (The logistic regression penalty parameter may be ’none’, ’l1’, ’l2’ or ’elasticnet’.)

   Regularization was applied using penalty parameter (None, l1, l2, elasticnet).
   Penalty = None F1 Score- 0.8011
   Penalty = L1 F1 Score- 0.8011
   Penalty = L2 F1 Score- 0.8056
   Penalty = ElasticNest F1 Score- 0.8011
   L2 regularization gave the best F1 score on the test set. L2 regularization helped in reducing overfitting and maintaining generalization.

4. These models have hyperparameters. When training, experiment using GridSearch to select hyperparameters for your models. What are the best hyperparameters among those you tried?

   GridSearchCV was used to optimize hyperparameters for Random Forest. The best parameter was n_estimator=200 which improved the CV F1 score to 0.7413, suggesting that increasing the number of estimators helped in improving the performance. For Logistic Regression, Optimal penalty was l2 with test F1 score as 0.8056.

5. Which model you experimented with gives the best F1 score on testing?

   Logistic Regression (l2 penalty) gave the best F1 score on the test case.