Kari Palmier
DSC 540 – Final Project

Cervical Cancer Prediction

Dataset from UCI Machine Learning Database
Data Taken At 'Hospital Universitario de Caracas' in Caracas, Venezuela

Dataset Features

- 32 Risk Factor Features
  - Demographic and Health History
- 4 Possible Diagnosis Test Result Targets
  - Binary targets (0 or 1)
  - Hinselmann, Schiller, Citology, Biopsy

Feature Preparation

- Dealt with missing values
  - Removed features with too many missing
  - Removed missing rows from categorical variables
  - Replaced missing rows from continuous variables with mean

Target Preparation

- Selected Biopsy to use as one target
- Created a combination target from all targets
  - Row value set to 1 if any targets were 1

Dataset Splitting

- Split dataset into 80% train, 20% validation
- Validation not used during modeling

Dataset Sizes

- 30 Risk Factor Features
- 2 Targets
  - Biopsy and Combination
- Training data size
  - 580 rows
- Validation data size
  - 146 rows

Target Imbalance

- Biopsy Target Imbalance
  - 540 Value 0
  - 40 Value 1
- Combo Target Imbalance
  - 506 Value 0
  - 74 Value 1

SMOTE Oversampling

- Biopsy SMOTE Sampling
  - 100%: 540 Value 1 and 540 Value 0
  - 30%: 162 Value 1 and 540 Value 0
- Combo SMOTE Sampling
  - 100%: 506 Value 1 and 506 Value 0
  - 30%: 151 Value 1 and 506 Value 0

Random Under Sampling

- Biopsy Random Under Sampling
  - 100%: 40 Value 1 and 40 Value 0
  - 30%: 40 Value 1 and 133 Value 0
  - 10%: 40 Value 1 and 400 Value 0
- Combo Random Under Sampling
  - 100%: 74 Value 1 and 74 Value 0
  - 30%: 74 Value 1 and 246 Value 0
  - 10%: 74 Value 1 and 370 Value 0

Model Information

- Used AUC to evaluate model performance
  - Used because of imbalance
  - Reported accuracy as well
- Models Used:
  - Random Forest, Gradient Boosting, Neural Network, SVM

Model Information Cont'd

- Performed all model types on both targets
- Ran models multiple times
- Set random state flag to ensure comparable results

Model Evaluation

- Performed 5 fold cross-validation on training data
  - Reported score mean and standard deviation values
- Trained validation model with entire training data
- Scored validation data using validation model

Feature Selection

- Performed on base models of all model types
- Selected the best performing 1-2 models of each model type
  - Did for both targets

Feature Selection Types

- Low Variance Filter
  - Thresholds 0.4, 0.5, 0.6
- Simple Model Wrapper
  - Using Sklearn SelectFromModel and get_support
  - Not used for Neural Network and SVM

Feature Selection Cont'd

- Stepwise Recursive
  - K: 5, 6, 7, 8, 9, 10
  - Not used for Neural Network and SVM
- Chi-Squared Univariate
  - K: 5, 6, 7, 8, 9, 10
- Mutual Information
  - K: 5, 6, 7, 8, 9, 10

Model Optimization

- Performed grid search on best models from feature selection
  - Performed for best 1-2 models of all model types and targets
  - Searched space was list of values for several different model parameters

Model Selection

- Chose models with best validation and training AUC
- Several models had similar performance validation
  - Used training standard deviations, number of features, and interpretability to chose final models

Final Biopsy Model

- Model Information
  - Random forest
  - 30% SMOTE oversampling
  - Chi-Squared Feature Selection
    - 10 Features Selected
- Scores
  - Validation AUC: 0.574
  - Training Cross-Val Mean AUC: 0.874
  - Training Cross-Val Std Dev AUC: 0.089

Final Combo Model

- Model Information
  - Random forest
  - 100% SMOTE oversampling
  - Low Variance Filter Feature Selection
    - 8 Features Selected
- Scores
  - Validation AUC: 0.547
  - Training Cross-Val Mean AUC: 0.974
  - Training Cross-Val Std Dev AUC: 0.067

Conclusion

- Low validation AUC for both final models
  - Models had hard time predicting target 1 values
- Sampling improved training but not validation
- Future – try more feature selection and model parameters