Adam Kiehl

4/23/2023

STAA 578

Homework 2 (Loan Prediction)

The goal of this modeling exercise was to predict whether a small business loan should be issued using various business metrics and descriptive predictors. A min/max scaler was applied to all numeric predictors and and factor level encoding was applied to all categorical predictors. Ultimately, 44 predictors were used for modeling. The training data were split into training and validation sets to tune and select models for final consideration. All models were evaluated using F1-scores due to imbalances in the response classes.

Before any deep learning techniques were employed, K-Nearest Neighbors, Random Forest, and Gradient Boosted models were fit as a baseline. The decision trees performed impressively, achieving validation F1-scores of 0.865 and 0.871, respectively. An MLP structure with five hidden layers (512, 256, 128, 64, and 32 neurons) and an early stopping criterion based on F1-score was constructed but performed poorly. This design was extended to include normalizing penalty and dropout regularization techniques and was tuned over a variety of hyperparameter values. The normalizing penalty (L2 with penalty 0.025) model performed curiously with unchanging evaluation metrics across epochs and an ultimate validation F1-score of 0.809. The dropout (rate 0.1) model performed better with a validation F1-score of 0.848. Based on validation evaluation, the Gradient Boosted tree and dropout neural network models were ultimately chosen for final testing.

The Gradient Boosted tree model chosen used a learning rate of 0.1 and and a maximum depth of one. Loan length and volume, current number of employees, and numbers of jobs created and retained by the loan were identified as some of the most important predictors of loan repayment. A test submission was uploaded to Kaggle and achieved an F1-score of 0.798. A second test submission was generated using the dropout-regularized neural network and early stopping after 21 epochs. This submission only achieved an F1-score of 0.712 which was considerably lower than the validation F1-score used to select the model. This may be due to poor generalizability, chance, or an inconsistency between the F1-score metrics provided by Keras and Tensorflow. Ultimately, the Gradient Boosted tree model outperformed all deep learning methods and was chosen for the final submission.