Identifying Suitable Volunteers to Work with Alzheimer Patients

Data is a survey from people aged 15-30. It consists of 1010 rows and 150 columns (139 numerical and 11 categorical) with some missing values. Our task is to predict if a person is suitable to work as volunteers to help people with Alzheimer by predicting how empathetic he or she is. By definition, people with answers of 4,5 are considered very empathetic and with answers 1,2,3 not very empathetic. For this purpose, data were preprocessed. First, missing numerical and categorical values were filled using median and most frequent answer, respectively. Then, one hot encoding were applied for categorical data. Columns with significant difference in value range with other columns were scaled using MinMaxScalar. Finally, problem were turned into binary classification by mapping answer 4,5 of empathy as 1 and answers 1,2,3 as 0.

For this project, two models were chosen; support vector machine (SVM) and random forest (RF). Both of these models provide high accuracy. SVM is versatile, can make complex decision boundaries, and finally can work well with high-dimensional data if the sample size is not very large. Despite all the advantages, SVM is hard to visualize and interpret. Therefore, random forest were chosen as the second model since it is powerful and can provide better insight toward feature importance compared to SVM.
To evaluate success, baseline classifier defined as the most frequent class. Then, data were split into train(80%) and test(20%) portions. Models were built on training data. For tuning hyperparameters, a grid search of 10-fold cross validation (GSCV) was used on training data. For SVM (C and gamma) and RF (maximum depth and number of decision trees) combination of 35 and 64 values were tested to find optimum value for hyperparameters, respectively.

Mainly pandas and scikit learn libraries were used to fulfill this task. Having R-style dataframe, ability to handle different data types (float, string, …) at the same time, and built-in preprocessing tools are main advantages of pandas. Scikit learn covers most machine learning tasks and scales to most type of data problems. In addition, scikit learn provides convenient diverse evaluation metrics.
Accuracy of baseline classifier which returns most frequent class is 68.31% and accuracy of SVM on test data after hyperparameter tuning is 72.27%. To improve the accuracy, a feature reduction process was inspected. In this process, feature selection using select from model of scikit learn library by implementing linear SVM and L1 penalty was used. After applying feature selection, number of features were reduced to 138 from 172 and accuracy was increased to 75.24%. Random Forest was used as the second model. Optimum maximum depth and number of trees were obtained using GSCV by testing combination of 64 numbers. Accuracy of tuned RF model was obtained as 69.80%. To further improve the accuracy feature selection was implemented. In this process select from model option of scikit learn using RandomForestClassifier with threshold of median was implemented. This process reduced the number of features by half and increased the accuracy of our RF model to 74.75%. Our RF model exhibited judgement calls, compassion to animals, weight, and children as the most important features and education and smoking as the least important features.

In SVM model, examples 934 and 943 classified incorrectly and correctly, respectively. Since SVM does not provide interpretable information it is hard to find which features of this example lead to misclassification. However, by increasing gamma we could increase the importance of each example and classify the example 934 correctly. In case of RF, we can remove the error by increasing number of trees.