Final Project

Supervised Learning - AY 2021-2022 - January 12th 2022

The goal of this project is to analyze a dataset representing potential audience people for an advertising campaign. We want to build a model which is able to classify all people (potential audience) into 4 segments (A, B, C, D), in order to perform personalized advertising. The provided dataset includes the following information:

- `ID`: Unique ID
- `Gender`: Gender of the customer
- `Ever_Married`: Marital status of the customer
- `Age`: Age of the customer
- `Graduated`: Is the customer a graduate?
- `Profession`: Profession of the customer
- `Work_Experience`: Work Experience in years
- `Spending_Score`: Spending score of the customer
- `Family_Size`: Number of family members for the customer (including the customer)
- `Var_1`: Anonymised Category for the customer
- `Segmentation`: Customer Segment (*target feature*)

In order to build the desired predictive model, develop the following tasks and answer the following questions.

Questions and Tasks

1. Load and explore the dataset. Eventually perform data engineering (handlying missing values, encode cathegorical values).
2. Train a **Softmax Regression** model able to predict the Segmentation class.
   (a) Perform features pre-processing if necessary. Discuss your choices and the performed actions.
   (b) Train a regularized model by applying ℓ_2 regularization (default regularization when you perform multinominal `LogisticRegression` on `sklearn`): tune the hyperparameter **C** in order to optimize the generalization performances of the model. What happens if you increase the value of **C** ?

   (c) Evaluate the trained model on the provided test set: produce a confusion matrix comparing the true target test values to the predicted target values; calculate Precision, Recall and f-score for each class. Discuss the obtained results.

3. Train a **kNearestNeighbor** model able to predict the Segmentation class.

   (a) Perform features pre-processing if necessary. Discuss your choices and the performed actions.

   (b) Train a k-Nearest Neighbour classifier on the training set, eventually tuning the model's hyperparameters. Specify which hyperparameter requires a tuning procedure, and how does the model performs with different hyperparameter values.

   (c) Evaluate the trained model on the provided test set: produce a confusion matrix comparing the true target test values to the predicted target values; calculate Precision, Recall and f-score for each class. Discuss the obtained results.

4. Train a **DecisionTree** model able to predict the Segmentation class.

   (a) Perform features pre-processing if necessary. Discuss your choices and the performed actions.

   (b) Use grid search with cross-validation (with the help of the `GridSearchCV` class) to find good hyperparameter values for the `DecisionTreeClassifier`: make a choice on the hyperparameters you might tune and provide comments on your choice. Specify which hyperparameter might require a tuning procedure, and which is the effect of the tuning procedure on the final model.

   (c) Evaluate the trained model on the provided test set: produce a confusion matrix comparing the true target test values to the predicted target values; calculate Precision, Recall and f-score for each class. Discuss the obtained results.

5. Train a **Random Forest** model able to predict the Segmentation class.

   (a) Perform features pre-processing if necessary. Discuss your choices and the performed actions.

   (b) Use grid search with cross-validation (with the help of the `GridSearchCV` class) to find good hyperparameter values for the `RandomForestClassifier`: make a choice on the hyperparameters you might tune and provide comments on your choice. Specify which hyper-parameter might require a tuning procedure (concentrate on the hyperparameters related to the ensemble, and skip the ones discussed above and related to the DecisionTree).

   (c) Which are the 2 most important features for the trained model?

   (d) Evaluate the trained ensemble model on the provided test set: produce a confusion matrix comparing the true target test values to the predicted target

values; calculate Precision, Recall and f-score for each class. Discuss the obtained results.

6. Compare the performances of the previously trained classifiers and discuss the results.

7. *(OPTIONAL)* Train a **Voting Classifier** model able to predict the Segmentation class.

   (a) Combine the models trained above into an ensemble, using a soft or hard voting classifier.

   (b) Evaluate the trained model on the provided test set: produce a confusion matrix comparing the true target test values to the predicted target values; calculate Precision, Recall and f-score for each class. Discuss the obtained results.

   (c) How much better does the voting classifier perform compared to the individual classifiers?

8. *(OPTIONAL)* Train various classifiers (different from the ones trained above): such as Naive Bayes classifier, Extra-Trees classifier, AdaBoost classifier, GradientBoost classifier, XGBoost classifier. Include the trained models in the *Voting Classifier* trained above and evaluate again the model.