Applied Machine Learning Homework 2

Due 2/19/20 1pm.

Please create the submission using github classroom with the following link:
https://classroom.github.com/a/gaSP8UdG
and as a single PDF via gradescope.

Please submit Task 1 and Task2 as separate Jupyter notebooks. Clearly mark which part of the notebook fulfils which task. Make sure to not commit any data to github classroom. Please ensure that your jupyter notebooks do not have too much spurious output.

Task 1 Classification on the 'credit-g' dataset (40 points)

You can download the dataset with 'fetch_openml('credit_g')' and see it's description at https://www.openml.org/d/31

1.1 Determine which features are continuous and which are categorical.

1.2 Visualize the univariate distribution of each continuous feature, and the distribution of the target.

1.3 Split data into training and test set. Do not use the test set until a final evaluation in 1.4. Preprocess the data (scaling, treatment of categorical variables) without using a pipeline and evaluate an initial LogisticRegression model with an training/validation split.

1.4 Use ColumnTransformer and pipeline to encode categorical variables (your choice of OneHotEncoder or another one from the categorical_encoder package, or both). Evaluate Logistic Regression, linear support vector machines and nearest neighbors using cross-validation. How different are the results? How does scaling the continuous features with StandardScaler influence the results?

1.4 Tune the parameters using GridSearchCV. Do the results improve? Evaluate only the be model on the test set.
Visualize the performance as function of the parameters for all three models.

1.5 Change the cross-validation strategy from 'stratified k-fold' to 'kfold' with shuffling. Do the parameters that are found change? Do they change if you change the random seed of the shuffling? Or if you change the random state of the split into training and test data?

1.6 Visualize the 20 most important coefficients for LogisticRegression and Linear Support Vector Machines using hyper-parameters that performed well in the grid-search.

Task 2 Regression on Sydney Dataset (60 points)

You can load the Sydney housing dataset from https://www.kaggle.com/shree1992/housedata where you can also find a description. The goal is to predict the 'price' column. For this assignment you can ignore the date. Please don't make any kernels public on Kaggle before the assignment ends.

2.1 Determine which features are continuous vs categorical. Drop rows without a valid sales price.

2.2 Visualize the univariate distribution of each continuous feature, and the distribution of the target. Do you notice anything? Is there something that might require special treatment?

2.3 Visualize the dependency of the target on each continuous feature (2d scatter plot).

2.4 Split data in training and test set. Do not use the test-set unless for a final evaluation in 2.6. For each categorical variable, cross-validate a Linear Regression model using just this variable (one-hot-encoded). Visualize the relationship of the categorical variables that provide the best R^2 value with the target.

2.5 Use ColumnTransformer and pipeline to encode categorical variables (your choice of OneHotEncoder or another one from the categorical_encoder package, or both). Impute missing values using SimpleImputer. Evaluate Linear Regression (OLS), Ridge, Lasso and ElasticNet using cross-validation with the default parameters. Does scaling the data (within the pipeline) with StandardScaler help? Use the preprocessing that works best going forward.

2.6 Tune the parameters of the models using GridSearchCV. Do the results improve? Visualize the dependence of the validation score on the parameters for Ridge, Lasso and ElasticNet.

2.7 Visualize the 20 most important coefficients of the resulting models. Do they agree on which features are important?