**Homework 3**

**Problem 1**

1.We see that we have no missing values. There are many outliers especially for the following variables.

|  |  |
| --- | --- |
| Monthly Income | Years At Company |
| Performance Rating | Years In Current Role |
| Stock Option Level | Years SinceLastPromotion |
| Total Working Years | Years WithCurrManager |
| Training Times Last Year |  |

2.To deal with them I considered as outlier the values that are out of mean +,- 3 \* standard deviation and got rid of them for some variables with the highest number of outliers.

We have also many correlated variables. As monthly income 94% correlated with Job Level and as Monthly income has many outliers, I delete that variable. I also deleted “Years at Company” and “Performance Rating”, ”Job level” to have no variables with more than 75% correlation.

I also tried to scale data with mean u standard deviation but the results became worse, so I chose the models on unscaled data.

3.

|  |  |  |
| --- | --- | --- |
|  | Logit | CART |
| Best values for model  hyperparameters | Logit for unscaled data  “C”= 3.3334  “penalty”= l1 | Tree for unscaled data  'max depth'=8, 'min samples leaf'=69,  'class weight'= 'balanced' |
| ROC AUC for Train | 79.45 | 73.94 |
| ROC AUC for Test | 79.06 | 69.88 |
| Recall for Train | 81.92 | 68.36 |
| Recall for Test | 77.78 | 61.11 |
| Accuracy for Train | 77.82 | 77.63 |
| Accuracy for Test | 79.94 | 75.92 |
| Average cross-validation ROC  AUC | 83.5 | 74.6 |

4. I select as winner Logistic Regression on unscaled data with results in the table. It has the highest roc\_auc as well as it doesn’t overfit, moreover logistic regression is more interpretable than decision tree.The variables that have p\_value <0.05 are the following: Age, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, NumCompaniesWorked, RelationshipSatisfaction, TrainingTimesLastYear, WorkLifeBalance, YearsInCurrentRole, YearsSinceLastPromotion, BusinessTravel\_Travel\_Frequently, BusinessTravel\_Travel\_Rarely, Gender\_Male, JobRole\_Laboratory\_Technician, MaritalStatus\_Single, OverTime\_Yes.

Maybe by deleting the variables step by step the importance can change.

5.I have also calculated the feature importances for the logistic regression and plotted them which shows that the most important features are overtime, age, distance from home, total working years etc.

**Problem 2**

Raymond Reddington (known as well as Red) is a data scientist who aims to predict whether a given individual will conduct another crime or not, based on his/her criminal behavior history. He is using a basic Classification and Regression Tree (CART) algorithm, which achieves 74.56% ROC AUC on the training set and only 59.83% ROC AUC on the test set. What would you advice Mr. Red to do to improve his model performance? Please, provide at least 3 recommendations with grounded explanation

The difference between train and test means overfitting. To overcome overfitting, they can.

1. To use cross validation, which allows to have different test sets from the whole data and will help to have better results for unseen data
2. To stop the Tree from growing and going to deep in the train data on some point by tuning hyperparameters like maximum depth, minimum samples size, etc.
3. To clean the data before running the model, for example dealing with outliers, which will reduce the noise and consequently overfitting.