DMG Assignment 2

Kaggle competition link:
https://www.kaggle.com/t/cb719aca29ab4452a3c048eec64e0608

Dataset: (uploaded on Kaggle)

Given Dataset size - 20,41,687 rows X 9 columns
Target variable - 'T'

Aim: Build a classification model with maximum performance
Evaluation Metric: AUC
Deadline: 15th October 11:59 PM

Instructions:

1. Use only the username you provided for submission on Kaggle.
2. Mention all assumptions if any in the report.
3. Report plus code in .py format should be submitted in the classroom in a zip folder with name 'A2_RollNumber_Name'.
4. Use any classifier from KNN, Rules Set or Decision Trees.
5. Deep learning and Transfer learning techniques are not allowed.
6. Include one runner function in code which takes to_predict.csv as input and produces result.csv. All preprocessing to be done on data before applying the model should be present in the runner function.
7. Save your top 3 models. Upload it on drive with the corresponding result.csv and share the link in the report. Make sure the drive folder is not private.
8. No restrictions on which libraries to use.
9. Top 20% teams on private leaderboard at the end of competition will get 10 marks, next 20% will get 9 marks and so on.
10. Some students will be randomly picked for demo of assignment 2. So write the code on your own, make sure you don't cheat. If you can't answer the questions during your demo, 50% of your marks will be deducted.
11. We will run submitted code at our end also. It will be to confirm that you have submitted the result file generated by a model out of Rules Sets, Decision tree, and KNN and not others.
12. Both team members should submit in the classroom.

Kaggle Instructions:

1. Two datasets are given on Kaggle. One is ‘given_dataset’ where samples are labelled i.e. ‘T’ variable is specified. Second is ‘to_predict’ for which your model predicts the ‘T’ variable.
2. On Kaggle, you need to upload a .csv file. It should contain two columns and 875006 rows + 1 header row (total 875007 rows). Header row contains column labels ‘id’ and ‘T’. ‘id’ values in submission should match the ‘id’ values of ‘to_pred_dataset’. One sample submission is also there for your reference.
3. Make a team of 2 on Kaggle. Team name should be roll numbers of both members: RollNo1_RollNo2
4. In case of doubts, comment on the classroom and not on Kaggle discussion forum.
5. Maximum daily submission limit is 10.
6. Do not share the competition link.

Following should be included in the Report:

1. Explain your methodology: approach and reason clearly in the report.
2. Visualize skewness of data before and after preprocessing.
3. Plot Training and Testing accuracy w.r.t. hyperparameters of the model.
4. Drive link for the saved top 3 models and corresponding output .csv files.
5. Make a section “Learning”, which describes your learning in doing this assignment.

Suggestions for participants:

1. Handle missing values if any
2. Handle class imbalance
3. Select relevant features
4. Does normalization or standardization help?
5. Get rid of outliers if any
6. Make sure your model is not overfitting
7. After splitting train and test data, are there enough instances of each class?
8. Parameter tuning

Resources:

https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/
https://docs.python.org/2/library/pickle.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_csv.html
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score