DAT200 – Compulsory Assignment 2

(Deadline: 23 March, 23:59h – both on Canvas AND Kaggle)

Background

You will participate in an In-Class Kaggle (www.kaggle.com) competition (only students enrolled in DAT200 course are allowed to participate), where you will build classification models for the Kobe Bryant shot selection data.

Link to the competition

https://www.kaggle.com/c/dat200-ca02

Link to enter the competition (please don't forward this link to others outside our DAT200-course, since we want to keep the competition In-class)

https://www.kaggle.com/t/2d1e45abdd60491d85a44d87f36d57b7

Kobe Bryant is a former star player of the National Basketball Association (NBA). The data contain information on most of his shots taken during his 20-year career, a total of 25 697 shots.

Your training data will consist of 20 697 shots and you will train various classifiers that shall predict the outcome of the remaining 5000 shots (randomly selected over his 20-year career). Possible outcomes are 0 (missed shot) or 1 (shot made). This means that your classifier shall produce a vector of dimension (5000×1) holding predictions 1 or 0, indicating whether the shot went in or not.

Rules and context of the competition

1. Please find the suggested workflow in file 'DAT200_CA02_workflow.pdf'. Use the workflow as a guide for how to train your classifiers, make predictions and upload the predictions to Kaggle.

2. You learned only a handful of classifiers in DAT200 so far. Therefore, you have only these available for prediction. You can use the following classifiers implemented in scikit-learn:
   - Perceptron
   - Logistic regression
   - Support vector classifier (SVC kernel='linear')
   - Support vector classifier (SVC kernel='rbf')
   - Decision trees
   - Random forests
   - K-Nearest Neighbours (K-NN)

3. You have not yet learned all details of cross validation and how to apply it in scikit-learn. Therefore, we will continue our practice of using many train_test_splits (of the training data) while searching of the best parameters of each classifier as we have done so far in the lectures.

4. You can compete in teams of two or alone.

Deliverables / submissions

To have the compulsory assignment approved you need to deliver the following:

1. You or your group must appear on the leaderboard of our own Kaggle competition, which means that you or your group must submit at least one prediction (link to leaderboard: https://www.kaggle.com/c/dat200-ca02/leaderboard)
2. Submit a Jupyter notebook + PDF of Jupyter notebook on Canvas with the code for training of your best classifier (please don't submit code for all seven classifiers) and the computation of the prediction. Please make short comments throughout your notebook/code on what you are doing and how you choose the parameters of your final best classifier.
3. If you use an alias in the Kaggle leaderboard, you must provide your Kaggle alias AND your real name at the beginning of your Jupyter notebook

4. If you compete as a group, both students must submit their Jupyter notebook + PDF separately. Please also mention at the beginning of your Jupyter notebook who your group partner is.
5. Remember that you can get your compulsory assignment approved during the exercises, where you provide an oral discussion of your work.

Administrative information, inspiration and help

1. In order to be able to participate at the competition, you must register at Kaggle.com. Please decide on a user name you are comfortable with. The leader boards will be visible to all and if you don't want to be identified through your real name, use an alias.

2. There is also a page answering frequently asked questions (for students) https://www.kaggle.com/about/inclass/faqs

3. Find this compulsory assignment, the workflow document and helper-code files on Canvas in folder "Files/V2018 compulsory assignments/CA_02"

4. The Kobe Bryant data set has been used before in an earlier Kaggle competition. There are plenty of discussions on the Kaggle webpages on how to model the data. Have a look around if you need ideas or inspiration.

5. Many fine tips and tricks for handling data with Pandas are provided in section "Data Wrangling" on Chris Albon's webpage https://chrisalbon.com/#articles

Good luck! Help and learn from each other. Come to the exercises if you are stuck.