Problem Set 5

The fifth problem set focuses on classification tasks. Your solution should be composed of a well-structured R script which should provide the required analyses. Besides the functions the code should be directly runnable or at least sufficiently well documented (working directory, path settings) to be executed.

Furthermore, provide a documentation (in text format or powerpoint) where you document and illustrate your general approach, document the qualitative tasks and provide instantiations of your graphs. Provide some intuition of what your plots show and what you hypothesize.

Grading will reflect your performance on both the coding as well as the documentation and interpretation tasks – however, there is no fixed grading scheme between the two categories.

This problem set is **due on July 2nd by 12.00** through the wuecampus submission functionality.

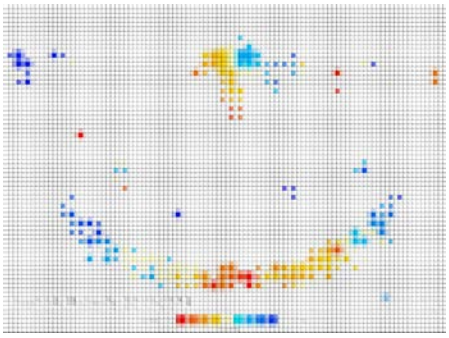
Cooperation with other groups is not permitted and will lead to severe credit deductions.

1. Again some US sports analytics, this time basketball. You are provided with the data sets basketball_train and basketball_test which provide labelled data of ~25,000 basketball shots taken by a single player. Training data is labelled and provides you with information if the shot was made or not, test data does not provide this information.
   [The data set was created with Excel therefore you need to specify sep=";" in import statement]

   The variables are as follows:
   - combined_shot_type – what kind of shot was taken (jump shot, lay-up, dunking)
   - loc_x, loc_y – x-y coordinates of the shot relative to the basket – this allows you to identify shot distance as well as the relative position of the shot
   - minutes_remaining, seconds_remaining – how much time was left in the current period [this is not the shot clock]
   - period – which period of the game
   - playoffs – was the game a playoff game
   - season – which season was the shot taken
   - shot_type – was this a 2 point shot or a 3 point shot (longer distance)

   All model training should be performed on the first 16,000 rows of the training data set and then reporting prediction results on the remaining rows. The test data set is only for the challenge task e.

a. Training data exploration
   - Compile season descriptive statistics: 2 vs 3 pointer success rates in regular season vs playoffs, can you see
   - Spatial visualization: Illustrate shot frequency (e.g., size of circles) and success rate (e.g., color) over the court as depicted on the right (you do not need to overlay the court schematics)



   - Break down the game periods into minutes / 30 second segments / 15 second segments (three different to check robustness) and illustrate the shot counts over these bins. Assuming equal sampling of the shot data how would you interpret your findings with respect to resting periods (player is on the bench) and tendency to take important shots (end of period / end of game).

b. kNN classification
   - Use kNN classifiers based on shot location and time remaining in the game with varying values of k to predict shot success of the test data.
     - Why does kNN classification fail if you predict solely from the location of the shot?
     - Experiment with different scaling of the remaining time variable. Why is this crucial for the kNN classifier?
   - Discuss the result of your classification factoring in the baseline shot rates from 1.

c. Logistic regression
   - Train and compare at least five different logistic regression models for the classification task, discuss your feature selection considerations and interpret the results. Some ideas include,
     - Coordinates converted to polar coordinates
     - Attacking from the left or from the right of the basket
     - Crunch time (remaining time < X)
     - Aging curve – replacing season dummies through continuous variables which are incorporated as ~year + year^2
   - Pick the most promising model (lowest AIC value) for predicting the training data

d. Train a random forest model on the data.
   - Compare your within sample prediction accuracy with the former classifiers.
   - You will most likely have encountered very similar and not overwhelming prediction accuracy across the different classifiers. Try to explain the underlying reason.

e. Challenge Task: Predict the probability of a successful shot on the test data set. Scoring of this task will be based on the following formula:

$$\sum_{x \in obs} (x - p_x)^2$$

where x are the true labels (unknown to you) of the test data set and p_x your predictions. Explain the rationale behind this metric and submit the predictions from the model of your choice. Why did you choose this model?

The UCI data set archive provides an interesting collection of data sets for training your machine learning skills, often accompanied by interesting papers. The last two tasks take advantage of this source.

2. The https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients data set revisits the credit card default theme from the lecture.
   In the companion paper by Yeh and Lien different classification approaches are applied to predict the occurrence of default events. Furthermore, in Section 4. they introduce a novel comparison approach for the methods.

   a. Using a 25:5 split train kNN, logistic regression, decision tree and random forest classifiers and provide a table akin to Table 1 in the paper.

   b. Explain the Sorting Smoothing Method introduced in Section 4 and develop the graphs in your own words and apply it to your classifiers from a. How would you asses the performance of your random forest (not present in the paper) vis-à-vis the methods in the paper?

3. Remaining in the banking sector the https://archive.ics.uci.edu/ml/datasets/Bank+Marketing looks predicting the success of marketing activities. The data set is offered as large and as a small variant. The small data set can be helpful for faster generation of models. As the open data set is not completely equivalent to the data used in the paper we cannot directly compare our findings.

   a. Adapting a 2:1 split of the data train a logistic regression and a random forest classifier. Extract the ROC characteristic for the two classifiers. [compare Figure 2]

   b. In practice we want to use machine learning to improve business operations. In this case by means of data-driven marketing: If we cannot contact all the customers, we want to contact those where we think the probability of success is highest. This is explored in Figure 3 in the paper. Replicate this analysis with your classifiers and interpret the results. [Note: You do not have to train any new models!]