Homework 5

The examination of the module *31-M29 Data Science* consists this year of a portfolio of four programming tasks and a fifth additional programming task (that replaces the exam). This is the fifth programming task. Please hand in your solution via "LernraumPlus" by March 17, 2021.

Customer Segmentation & Return Shipment Prediction (60 Points)

In the last homework you were asked to analyze the data of an online shop. Satisfied with your work the company asks you to assist them in another data science related task. With the goal to better understand their customer groups they (1) ask you to perform a **customer segmentation analysis**. Describe your findings (e.g., the different customer clusters) in a text or with graphical representations. Furthermore, the company observed that many of the ordered items are returned. In the interest of the environment (and their shareholders), the company wants to significantly reduce the amount of returns. As a first step towards this goal, they also ask you to (2) **develop a model that predicts if a customer will return an item**.

Included with this task description you will find a dataset (training.csv) that is identical to the dataset from homework 4. Furthermore, you will find a second dataset (test.csv) that contains the same features as the training dataset, but without the label. Your task is to create a prediction for each item order in the second dataset.

As part of this programming task we are also starting a **Kaggle in Class** challenge. This means that you can upload your predictions (twice a day) to Kaggle to evaluate them. Kaggle also provides a leaderboard that allows you to compare the quality of your predictions to those of your peers (Note that the use of the Kaggle leaderboard is optional).

Please also note the following additional instructions:

- Only use the dataset made available by us in the LernraumPlus. Do not use the data from the data mining cup website!
- Hand in a file that contains you predictions as well as the *.ipython notebook* that was used to generate them. The notebook should read in the provided raw data and write the predictions to a file. **When all cells are executed in sequential order ("Runtime" → "Restart and run all")** the generated prediction file should match the one that you handed in. Furthermore the notebook should also contain the code and your findings for the customer segmentation.
- To improve the prediction quality of your model you should perform hyperparameter tuning. You can use any freely available hyperparameter tuning tool you with. Comment out the tuning process in the notebook before you hand it in, but set the parameter values to the values found by the tuning process.

- Your prediction file must have the same format as the provided sample prediction (sample_prediction.csv).
- Working in groups is not permitted. Do not share your code with the other students.
- Use text cells to describe and structure your code so that each step is easy to understand.
- Generate *at least* 7 new features. Describe each feature in 1-2 sentences.
- We will award one bonus point for uploading at least one prediction to Kaggle before March 16, 2021. The 10 students that make the best predictions according to the final leaderboard will get two additional bonus points.
- Your predictions are evaluated with the *Area Under the Receiver Operating Characteristic Curve* metric (ROC AUC).
- You can access the Kaggle challenge here: `https://www.kaggle.com/t/217641dc0ce74dcb86de1329c7ecf825`

Grading will be done based on the following criteria:

- **Data preparation (4 Points):** How are the missing values handled?
- **Generated features (14 Points):** Do the generated features make sense (i.e., is it reasonable to assume that they might improve prediction performance or allow for better customer clusters)?
- **Hyperparameter tuning (5 Points):** Has hyperparameter tuning been conducted correctly (e.g., on a separate validation set)?
- **Prediction process: (5 Points)** Are all steps of the prediction process correctly implemented?
- **Prediction quality (12 Points):** How good are the generated predictions?
- **Clustering (10 Points)**: Does the clustering code work as described? Do the identified clusters make sense?
- **Code quality (5 Points):** How difficult is it to understand the code? Could the same result be achieved faster or with a more simple implementation?
- **Notebook quality (5 Points):** How well is the code/notebook structured? Are comments and text fields used to describe the content of a code block or code cell?

Late Submission Policy

Late submissions are penalized 5% per started hour after the deadline (e.g. a solution that is handed in 122 minutes after the deadline is penalized 15%). Note that these should still be handed in through LernraumPlus.