Progress Report

Assignment 2: *Learning to Rank from a search history data set originated from the Expedia hotel booking platform.*

Group 130 - Baptiste Avot *(VU 2637357)* **- Hinrik Snær** *(UVA 12675326)* **- Marcin Michorzweski** (VU 2688837)

The assignment was worked on throughout four consecutive weeks (commencing April 27th). The team worked together to find a way to evenly distribute the workload in a way that would allow us to work consistently towards the deadline. It is important to point out that a lot of time was spent on trying to deal with the large amount of data, that made very simple steps very computationally intensive which contributed significantly to the amount of time spent on the assignment.

The work was divided into individual steps that each team member would contribute to. The steps were research, data processing, data modelling, results gathering and paper writing. The steps included the following work:

- Create several datasets with different replacement values for NaN values (mean values, constants…)
- Research statistical components of features in data (means, medians, standard deviations, correlation, distribution over different groups).
- Look at ways to handle outliers that might result in a less well performing model.
- Search for features to add to the dataset that were relevant (standard deviation of some features for specific groups, means over specific groups, log transforms…).
- Decide on a scheme to compute rank score (target value) from click_bool and booking_bool.
- Decided to remove from dataset features that were present only in training dataset and did not provide sufficient information on the rank score.
- Finding the best ranking model for the provided dataset

The first week (April 26th to May 2nd), every individual did some research for a given task that they decided to take upon themselves. Hinrik decided to research multiple different ranking models that would be well suited for the provided dataset. Baptiste and Marcin researched which preprocessing steps would be well suited for our dataset along with research on feature engineering. The step included the following work:

- Perform different combinations of feature selection methods on different data sets.
- Search which feature selection method might be best for the kind of data we were provided with.
- RFECV (Recursive Feature Elimination with Cross Validation)
    - Looked for ways to score during cross validation.
    - Estimated optimal number of parameters

- F tests
- Mutual Information (information theory)
- Undersampling (using imblearn library) in order to reduce the class imbalance that was relatively important in the initial dataset.
- Research feature correlation.
- Removed features from the original dataset that contained little information due to high number of missing values and similar missing values distribution across training, test and only clicked holets datasets.

During the second week (May 3rd to May 9th), Marcin and Baptiste implemented various different preprocessing methods to generate multiple different variations of testing and training sets. After research on methods that support the “Learning To Rank” case, the team agreed to use the LambdaRank model which satisfied the needs. Hinrik integrated an existing LambdaRank method using an existing library so that it would be able to train on the datasets provided by Marcin and Baptiste.

During the third week (May 10th to May 16th), the team worked together and analyzed the results. Using various different best practices for Data Mining, the team adjusted the model with various different datasets to get the best possible results. After analysing the results, the team concluded that the current model implementation would not be sufficient enough to generate a satisfactory Kaggle score. Marcin introduced the team to the library “lightgbm” that was able to use gradient boosting with LambdaRank. This method with default parameters provided the team with a score that they found to be satisfactory.

The final days (May 17th to May 22nd) were spent on writing the paper, analysing the final results and generating images from the results. Every member of the group contributed to writing the paper according to the tasks they were involved in. After choosing the best model and dataset, the team tried to slightly improve the score by making small adjustments to the parameters. During that time, many helpful scripts that accelerated the way the model was trained and saved RAM were developed:

- Splitting test data into chunks, so results come in many files.
- Predicting values on small portions of data.
- Merging result files into the final one.

Throughout this period, the team communicated regularly and effectively through text channels and regularly scheduled video calls (Zoom).

The team members are overall satisfied with the performance of each member, every member communicated well throughout the assignment while also working well independently when it comes to research and presenting possible solutions to the group. The team also agreed that the organization was really good and contributed significantly to a satisfactory distribution of work.