Data Mining Techniques - Process Report Gr185

A.J. Hazenberg[2649192], D. Pop[2618006], and T.J. Siebring[2683240]

Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands

This report presents the timeline of actions taken towards attempting the second assignment of the course Data Mining Techniques. Additionally, the contribution of each member can be seen in the second section, while, at the end of the report, one can see the reflection on the overall cooperation within the team.

Schedule

- April 26: We set up the working environment (reports' templates), researched approaches of other participants of the competition on a high level, and started with data exploration, more precisely, the shape of the data and its available features.
- April 28: We researched the methodology of other participants in more detail, as well as the offline Learning to Rank field, including its three approaches (i.e. pointwise, listwise, pairwise) and the most commonly used, open-source libraries which facilitate the implementation of the techniques in Python.
- April 29: We wrote the Business Understanding section.
- May 4: We performed an extensive exploratory data analysis, inspecting the types, missing and unique values, ranges, and distributions of each feature in the data
- May 9: We worked on data processing, discussing the best approaches to deal with variables with missing values and deriving additional features based on our intuition.
- May 10: We expanded the data processing part by implementing ideas from the research performed on competition winning approaches and one-hot-encoded the categorical features to prepare the data for a Machine Learning task.
- May 11: We implemented our first working model, a LightGBM Regressor with default parameters predicting the position of an instance using the original set of features.
- May 13: We improved our model by feeding it processed data.
- May 14: We tried our model using different target variables, namely *click_bool* and *target_score* (i.e. $click\_bool + 5 * booking\_bool$).
- May 15: We reduced the number of dummy features resulted after one-hot-encoding multi-categorical features by using the built-in functionality of the LightGBM module called *cat_features*.
- May 16: We tuned the hyperparameters of the regressor model using a random search technique with a 4-fold Cross-Validation.

– May 19: We managed to implement a listwise approach through the LGBMRanker estimator.
– May 23: We improved our model by tweaking its hyperparameters and started writing the reports.
– May 25: We finished the reports.
– May 26: We reviewed the reports and made the last edits.

Contributions

Julia was responsible for exploring the dataset, designing the strategy to deal with missing entries, deriving most of the extra features, preparing the dataset for a Machine Learning task (one-hot-encoding of categorical features and evaluation splitting of the train data), and she helped Dragos with generating the first working model, namely the LightGBM Regressor. She also reported her work by writing the respective parts of the report.

Dragos' responsibilities included helping Julia with the feature engineering, specifically with the creation of additional variables, and researching Learning to Rank algorithms. After generating the first working model together with Julia, he attempted improving the model by testing out different variants in target variable (*click_bool* and *target_score*), trying out distinct sets of features, and implementing a random search method to find better parameters. Lastly, he programmed the LGBMRanker model and documented his process, findings, and struggles in the part about modeling and the conclusion section.

Tom set up the working environment, organized documents, and arranged our meetings. He had a central role in researching other methodologies, the offline Learning to Rank field, and the available tools to implement the approaches in Python. Tom wrote the Business Understanding part entirely and part of the Data Processing section, as well as this report. Finally, he reviewed the final text and corrected typos.

Reflection on overall cooperation

All the group members agreed that the cooperation was excellent. Outside the free education week, the members met weekly to discuss the progress and plan tasks. The response time for matters occurring between these meetings was short, and the members displayed willingness to help whenever needed. The division of tasks went smoothly, the effort was uniform across the group members, and each contributed with their specific knowledge to enhancing the quality of this project.