DM Assignment 2 - Process Report

Victoria Peterson[1 [2871118]], Marvin Frommer[2 [2727106]], and Max Bohle[3 [2728282]]

Vrije University Amsterdam

1 Schedule

8 May 2025
Set up repository, project plan, map out and assign issues; Gather information on assignment, data and competition Github Issues Completed: Preliminary Assignment Exploration

13 May 2025
Selecting Data attributes; Explored data analysis techniques; Identified cleaning opportunities

Github Issues Completed: #2 Completed Business Understanding, #3 #9 Data set Statistical Summaries, #11 Data Distributions, #10 Missing value analysis, #12 #14 Outlier Analysis and handling

15 May 2025
Discussed progress; Decided on prediction algorithm approaches; Split up parts of the report

Github Issues Completed: #4 #13 Missing value handling, #15 exploration of data transformations, #17 Engineered hotel comparisons for specific search IDs, #18 Explored and selected two models

17 May 2025
Final adjustments on prediction algorithms; Choosing algorithm based on preliminary Kaggle results; Finalized pipeline for best possible results

Github Issues Completed: #18 Basic KNN model implementation prototyped, Data aggregation for KNN model implemented, #16 Split training data for validation, Gradient Descent Model final implementation

18 May 2025
Kaggle submission(s); Report progress discussion

Github Issues Completed: #19 Tuned hyperparameters for Gradient Descent Model, #21 Generated and submitted results for Kaggle competition

20 May 2025
Re-evaluation of approach based on other (winning) teams; Further splitting up of report; Finalizing some report distributions, continuing work on evaluation of the algorithms

Github Issues Completed: #23 Ethical bias handling features added, statistical features added to improve prediction results, #20 Evaluation data collected

25-26 May 2025
Final report discussion and writing

Github Issues Completed: #7 Write the Scientific Report, #8 Write the Progress Report, #22 Discuss Ethical AI and Deployment

2 Contributions

Victoria worked on EDA, data cleaning, feature engineering, implementing a LightGBM based prediction algorithm, general code structure and organization and those sections of the report.
Marvin worked feature engineering as well as a KNN based prediction algorithm and the relevant parts of the report.
Max worked on Business and Data understanding, some EDA and writing those parts of the report.

3 Reflection

Generally we are happy with our collaboration and results. While we did not win the Kaggle competition we arguably learned more than the more successful teams with more experience in the field.
We were able to communicate well, stay organized and meet frequently to discuss the progress and (re-)assign tasks.
In our first meeting we decided to set up a GitHub project which allowed us to get a clear overview of the tasks, as well as who should be responsible for which. Later we were able to reference these tasks in our pushes to the repository. This orderly approach allowed us to keep the big picture in mind which made a big difference in ensuring we make progress on the entirety of the assignment and not get too stuck on individual aspects.
Obviously a point of improvement would have been having more time, however, considering the amount of time we had, and especially the other commitments of each team member, we are content with our progress and result. That being said, we could have improved on our submission(s) by re-evaluating our data preparation, especially feature engineering. Our approach involved dropping a lot of the variables with high missing value amounts which could potentially have been used had we had enough time to process them.