

Expedia Hotel Ranking Competition Process Report

Dongqi Pu^[2687343], Wouter Korteling^[2526876], and Edeline Contempre^[2636499]

Vrije Universiteit Amsterdam, DMT Group 84

1 Project Schedule

Table 1 shows the project's schedule and the contribution of each team member for this project. First, we understood and analyzed the goals and tasks of the project on April 20th. Initially, we hoped to complete data processing, model construction, and prediction in the Kaggle notebook. However, in the actual practical stage, the memory provided by Kaggle was not enough to meet our tasks, so we could only clean the data and build the model on the local machine. Therefore, the code update and maintenance became a challenge. We used SourceTree and a github repository to solve this problem. Dongqi and Edeline contributed more in exploratory analysis of data, while Wouter and Dongqi contributed more in feature generation, data cleaning, and final prediction models. In the report writing, Dongqi first wrote preliminary documents. Wouter and Edeline assisted in polishing and modifying the documents.

Table 1. Project Schedule

Date	Description	Member
April 20	Assignment setup, task understanding and analysis.	All
May 1-3	Data cleaning, deal with missing values and start of feature engineering.	Dongqi, Wouter
May 2-3	Data exploration and visualization.	Dongqi
May 4	Feature selection base on the improtance and build a preliminary model.	Dongqi, Wouter
May 6	Optimize the model further by using stacking method.	Dongqi
May 7-8	Memory optimization, parameter tuning.	Wouter
May 14-15	Parameter tuning testing.	Wouter
May 15-20	Assignment report writing.	All
May 18	Process report writing.	Dongqi, Edeline
May 21	Final report review and submission.	All

The first week of May was the most burdensome period because the analysis and processing of data accounts for more than 70% of the total task volume. We first explored the original data set and drew some plots to understand data distribution and correlation check. After cleaning the data, we proceeded to building our model. Initially, candidate models included random forests, neural

networks, and XGBoost. Due to its higher accuracy and faster training speed, XGBoost became our preferred choice. To further improve accuracy, we used the stacking method and checked for overfitting in the final model.

2 Performance Criteria

Table 2 shows a star rating system evaluating each person's contribution. Throughout the project, all team members were actively involved and provided a non-negligible role for the smooth progress of the project.

Table 2. Performance Rating

Rating \ Name Criteria	Wouter Korteling	Edeline Contempre	Dongqi Pu
Availability	*****	*****	*****
Response Time	*****	****	*****
Willingness	*****	*****	*****
Knowledge Contribution	*****	*****	*****

The most time-consuming part of the project are data processing and model optimization. Due to the large amount of data, the model needs to wait for a long time before each training to obtain results, and the iterative optimization of the model poses a great challenge. In this process, we discussed many potential optimization methods, such as SMOTE method, downsampling method and some other new ideas. Although some methods cannot improve the final score, we have learned a lot of new knowledge and skills. Overall, this recommendation ranking project has enabled us to improve our code experience and overall insight into data in data mining techniques.