Constructing Explainable Optimal Decision Trees for Taxi Demand Prediction

Elif Arslan – e.arslan-1@tudelft.nl
Department of Transportation and Planning, Delft University of Technology, Delft 2628 CN, the Netherlands

The dial-a-ride problem (DARP) is a vehicle dispatching problem aiming to serve customers with certain pick-up and drop-off locations. DARP solutions determine which vehicles serve which customers while allowing for ride-pooling and aiming at minimizing total passenger waiting times. Determining optimal solutions to DARP, known to be NP-hard even if all ride requests are known in advance ([1]), and therefore is especially challenging when solved online. However, finding such a solution is paramount as deviations from the optimum affects both passenger convenience and service providers operating costs due to failing to account for future ride requests.

However, how to predict the future taxi demand remains unclear in DSDARP literature. While in studies a particular prediction method is used, the reason for this selection is not necessarily detailed. One can refer to taxi demand prediction literature to determine which prediction method to use, however these studies mostly focus on the level of prediction accuracy. Considering the significance of establishing explainability in a dial-a-ride system to gain the trust of the service owners, it becomes crucial to seek for a balance in accuracy and explainability in developed/evaluated methodologies.

Due to their graphical structure, regression trees are deemed to be one of the explainable models. In addition, they do not have the assumption that each feature holds equal significance across all parts of the sample space unlike classical statistical models ([2]). It means that they have the potential to categorize the historical taxi demand well enough to achieve high accuracy while facilitating explanation generation. However, a regression tree's level of explainability should not be taken for granted as recent studies shows that only a subclass of regression trees provides succinctness in generated explanations ([3]). This finding highlights the importance of defining what the explainability of a regression tree is and how it can be measured. Only then, it can be constructed for accuracy and explainability maximization.

Determining the tree algorithm may also play a key role in this process. Although the CPU time requirement of greedy tree algorithms may be advantageous for large tree constructions, their greedy nature may result in disregarding the global explainability level of the regression tree. Optimal regression trees on the other hand, are promising to combine the accuracy and explainability metrics of the regression tree as one can define mixed integer linear programming formulation of a tree with the objective of maximizing tree metrics.