Ferwerda et al: Leveraging the Power of Place: A Data-Driven Decision Helper to Improve the Location Decisions of Economic Immigrants

This paper researches whether implementing an algorithm that helps immigrants to choose their initial settling improves expected income of those immigrants. Specifically, the paper focuses on immigrants in the Canadian Express Entry system.

What is novel about this approach, is that no data driven approaches have been proposed in order to help immigrants settle in their new country. Immigrants often rely on heuristics and settle in places that they either know (from being the only place they know, or because they know people that moved there). In this way, the authors nicely embed their problem in the literature, by referring to the availability bias originating from behavioural economics and psychology. More generally, I appreciate how the authors position themselves in the literature and convey the social implications of their research.

The authors generate recommended locations for immigrants to move to by creating models that predict expected earnings of immigrants for each economic zone they can move to. These models take individual characteristics of the immigrants and of locations into account to predict these values. Historical data of both Express Entry and Non-Express Entry immigrants is used to train the models.

In general, the validation of the models for each location could be presented in a clearer way. The strength of the recommendation system is now mainly derived from the simulation study. For the individual fitted models, the (aggregated) R2 metric is given in a footnote. Also in the appendix there is not much information, except on optimal hyperparameters. What would give more insight though, is to visualise the R2 metric across locations. The authors repeatedly mention that current immigration patterns tend to be concentrated in only a few locations. That means that there will be much more data for those locations, potentially making the predictions of expected income better. This information could be perfectly displayed together with the variable importance measures for each location, which gives more policy relevance to the model (policy makers also get insight in what drives predictions).

Building on this idea of varying model quality, a second improvement could make the insights in the model better and might even improve recommendation adherence. Immigrants currently get a list of locations with highest predicted expected earnings. However, there are no uncertainty quantifications added to these predictions. In the literature, more attention is given to probabilistic forecasts, that in addition (or instead of) a point estimate also provide confidence bounds. Referring to the previous argument, regions that have historically known less immigration might also have worse prediction performance and hence the predictions can be trusted less. When confidence bounds are also presented to immigrants, this might also give more context than a mere point estimate of their expected earnings. Even for boosting models that do not give prediction intervals generally, this can easily be implemented with Conformal Prediction, a distribution free method that wraps around any black-box ML algorithm (Lei et al., 2018; Shafer & Vovk, 2008).

Lastly, in assessing the impact of the recommendations on society, the authors assume that a percentage of people will comply with the recommendation, and for those people, the extra earnings are quantified. However, in the introduction and discussion of the other literature, it is argued that a more uniform distribution of migrants might improve integration and foster growth. In this spirit, it would be interesting to add a component to the analysis that randomly recommends a

location to immigrants (hence nudging immigrants to spread evenly over the country). The effect of such a recommendation system can then be quantified in a similar way. The interesting interpretation would then be to see how much the data-driven recommendation system is better than a random recommendation system. This does then not only describe what the added benefit of a recommendation system is over not having one, but also what the authors' model adds on top of a recommender system in general.

The paper nicely discusses limitations as well as potential risks of implementing this recommender system in practice. I very much like how they disentangle the difference between the final goal of successful immigration and the short term metrics that the recommendation system focuses on. I would maybe add one more potential risk, being the endogeneity of the recommender system. When the system would mainly recommend similar locations to similar people, this might introduce segregation among immigrants, which could be considered as a negative effect.

Overall, the authors write a strong paper and show well how a data driven recommendation system to aid settling decisions could benefit economic growth. The paper does not only consider the technical details of developing such a model, but also gives a good sense of limitations and risks involved with such a system.

References

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094-1111.

Shafer, G., & Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, *9*(3).