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

This paper investigates whether implementing a recommendation algorithm that helps immigrants to choose their initial settling location improves expected income of those immigrants. The algorithm presents information on a small set of recommended locations (with the highest expected earnings for that immigrant), based on individual characteristics of the immigrants, the locations, and the immigrant's preferences. 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) or have heard of before. By referring to the availability bias originating from behavioural economics and psychology, the authors nicely embed their problem in the literature and build a convincing hypothesis why their recommendation system would improve on the current situation. More generally, I appreciate how the authors position themselves in the literature and convey the social implications of their research.

After describing the goals and intentions of the papers, the authors continue to describe the model design and assumptions. The data-driven approach generates recommended settling locations for immigrants by creating models that predict expected earnings of immigrants for each economic zone they can move to. To predict these values, the models take individual characteristics of the immigrants and locations into account. Historical data of both Express Entry and Non-Express Entry immigrants are used to train the models. More specifically, the modelling strategy entails fitting a gradient boosted tree model for each location, with historical earnings for immigrants as dependent variable, and individual/location characteristics as explanatory variables. A regularized squared error loss function is used to optimize the model parameters.

In general, the validation of the models for each location could be presented in a clearer way. The recommendation system now mainly derives its strength from the simulation study (which we will get into shortly). For the individual fitted models, the (aggregated) R2 metric is provided in a footnote. Also, in the appendix there is not much information available on model validation, except for optimal hyperparameters. What would give more insight though, is to visualise the R2 metric across locations. The authors repeatedly mention that current immigration flows tend to concentrate 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 would be interesting to look into and get a better sense of how the models perform for each location. This information could be perfectly displayed together with the variable importance measures for each location. Overall, the variable importance visualizations are very well done and greatly contribute to the policy relevance of 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 models 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, increasingly more attention is given to probabilistic forecasts, which provide confidence bounds on predictions in addition to (or instead of) point estimates. 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 give more context than a mere point estimate of their expected earnings. Even for gradient boosting models that do not give prediction intervals generally (only point predictions), 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). This would, thus, be a relatively easy extension of the current approach, with potentially large impact.

After explaining how the recommendation system is trained and makes recommendations, the authors set up a simulation study to assess the impact on the economy and society. In doing this, the authors assume that a percentage of people will comply with the recommendation, and for those people, the extra earnings over the status quo are quantified. To account for the status quo, the (predicted!) expected earnings of the location where the immigrant actually landed are taken (the simulations are run as a "back test"). It is not entirely clear to me why the authors opt for the predicted expected earnings at the true settlement location of the immigrants. Since the simulations are performed as back tests, the realized earnings are also available in the data. This would potentially make for a better quantification of the impact of the system, since the realizations will be different from the predicted earnings. Nonetheless, in this section the authors do a great job at evaluating the consequences of their assumptions on compliance rates, clearly conveying how their results should be interpreted by policy makers. I also appreciate the comment on introducing a randomized experiment to properly assess the impact, before blindly deploying the algorithm.

Separately from using the realized earnings over predicted earnings, I see another way to improve the quantification of the value of the recommendation system for policy makers. The data-driven recommender algorithm can be compared to the results of a recommender system that outputs purely (or partly) random recommendations. 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 equivalently to the proposed system. 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 the status quo, 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. Moreover, the comment about potential issues using non-Express System immigrants in the historical train data, while making predictions for Express System immigrants is appreciated. Due to data availability, there is not much the authors can do about this, but it is transparent to discuss this issue. 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 (sending all similar immigrants to the same location), 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

Ferwerda, J., Adams-Cohen, N., Bansak, K., Fei, J., Lawrence, D., Weinstein, J. M., & Hainmueller, J. (2020). Leveraging the power of place: A data-driven decision helper to improve the location decisions of economic immigrants. *arXiv preprint arXiv:2007.13902*.

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).