Predicting prices on FourSquare venues

Surya Prerapa
7th April 2020

Contents

- Introduction
- Data sources
- Assumptions
- Approach
- Results and Discussion
- Conclusion

Introduction: Background

- Knowing the price range of an eatery/watering-hole at a glance is very useful for customers to make quick and informed decisions
- Travel, holiday and tourism service providers (and aggregators), home delivery platforms, etc would benefit from such a platform
- TripAdvisor.co.uk and FourSquare (via premium calls) are some of the existing players I know of providing this functionality
- This project is to assess how effective a simple intuitive model would work in reality and recommend improvements for future

Introduction: About this project

- Continue with Toronto venues dataset from Capstone project.
- Enrich it with additional datasets that are openly available which contain indicators for a given area's disposable income.
- Find 3 clusters in the disposable income indicator data → Cluster neighbourhoods as *Bargain*, *Medium* or *Premium* priced.
- Apply these labels to Toronto venues and compare against some real-world prices.

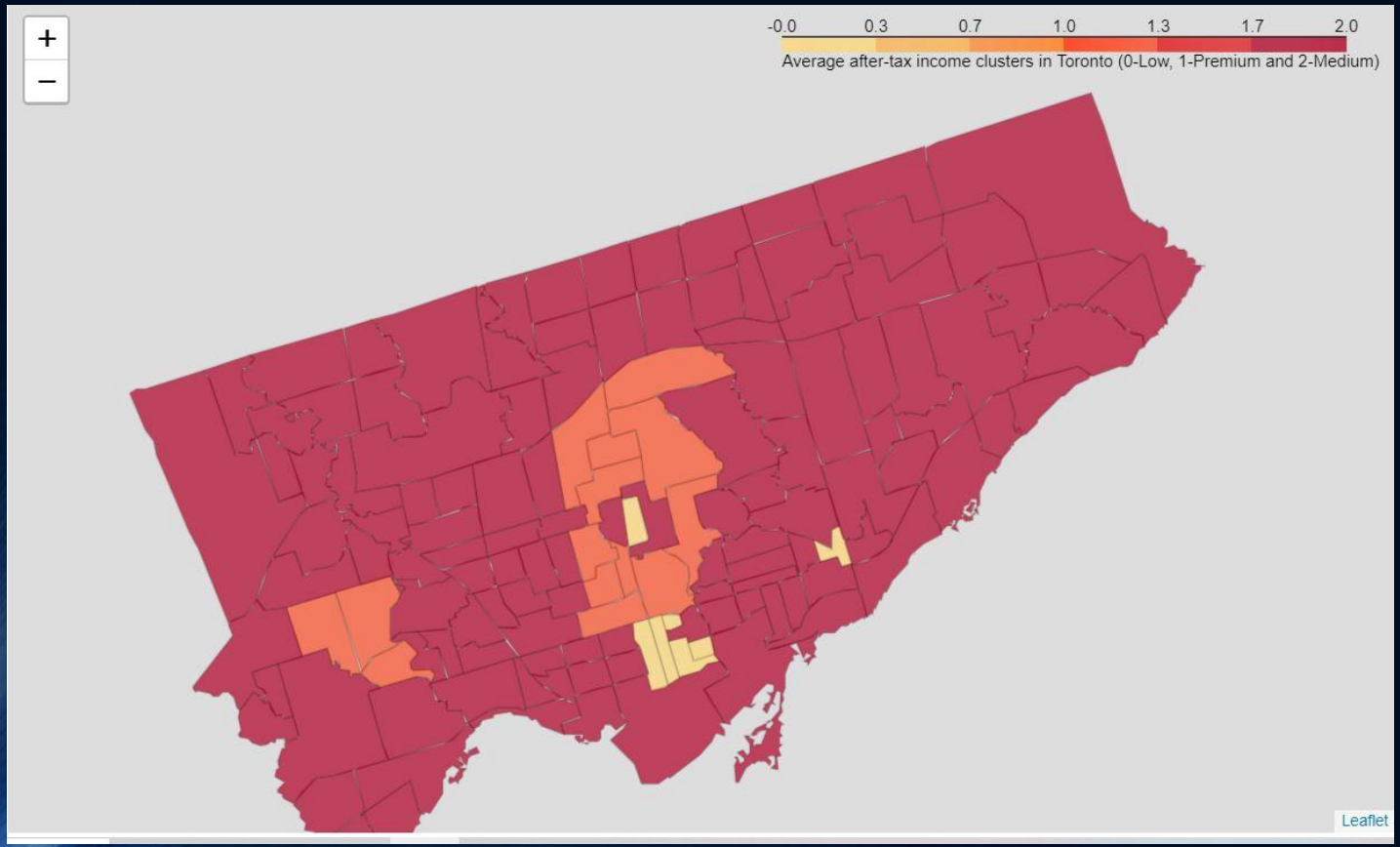
Data sources

1. Neighbourhoods - containing Toronto neighbourhoods WGS84 coordinates in CSV. GEOJSON file for Choropleth map
2. Neighbourhood demographics - containing ethnicities, age-groups, incomes, education, benefits, etc.
3. FourSquare regular API call data on the neighbourhoods described above
4. Picked average after-tax income and population density from dataset #2 as indicators of disposable income.

Assumptions

- Average tax-free income is a good indicator of disposable income.
- Affluent area has low population density and high average tax-free income and vice-versa.
- Establishments cater to local area closely and their prices reflect affordability of locals
- Effects of transient population such as tourists and commuters ignored
- Fixed priced models of local and international chains ignored.
- Each venue has it’s neighbourhood cluster label associated.

Approach: Clusters identified



Approach: FourSquare API calls

- 100 venues in a radius of 500M queried on each neighbourhood
- Filtered 703 establishments with categories falling in restaurant, pub, cafe, coffee and dessert shops.
- Cluster label of neighbourhood from previous plot applied to these venues.

Approach: Cluster label applied to venues

Bargain

Medium

Premium

Approach: Analysis

- 20 samples were drawn at random from the filtered dataset
- TripAdvisor labels were associated with these 20 samples
- TripAdvisor labels were compared against computed labels.
- Out of 20, 11 mismatched and only 2 of the mismatches were drastic.
- More thorough analysis could not be performed as it is very labour intensive.
- But nevertheless important conclusions could be drawn from this small scale study

Results and discussion

- Handle chains and tourist/commuter hotspots separately.
- Median (not average) tax-free income might be a better indicator of disposable income.
- Living costs must be removed from income to quantify disposable income accurately.
- Other demographic data such as age-group, ethnicity, etc could be used for accurate price calculations.

Conclusion

- Attempted to address an important business challenge with reasonable success using a rough and ready model.
- However, this challenge can be addressed these days by real businesses by procuring readily available data such as FourSquare Premium.
- But where such data is not available (for example in developing countries or rejuvenated neighbourhoods), techniques such as the one presented above could be immensely helpful.