Predicting Underserved Growth Areas for Public Transit System Expansion

Project Team

Name	CNET ID
Liza Marie Soriano	lizamarie1218
Ronald Kwan	rkwan
Kelsey Anderson	kjanderson

Project Summary

- What is the problem?

  Although ridership data is generally widely available for public transit, route patterns that may under-serve the places residents need to travel are more invisible. In order to suggest areas for increased route capacity or expansion, our group proposes to build a machine learning model that predicts how much ridership should be expected to grow for specific areas based on local business growth and connected residential populations. By finding areas of high business growth where the predicted growth of ridership is stunted by limited accessibility, we hope to identify underserved locations.

- Why is it an important policy problem?

  According to the National Academies of Science, Engineering and Medicine's Transportation Research Board[1], the top 12 issues facing the United State's transportation infrastructure include both ensuring equity as transportation systems develop and serving shifting population demographics.

  Increasing public transit access to local businesses and social/governmental service providers operates on two important policy levels.

  First, it helps people get where they need to go for employment, services and shopping. Public transit creates opportunities for populations that do not own a personal automobile, have unreliable access to cars or more workers residing in the home than cars owned. These are largely issues related to socioeconomic mobility and equity.

  Second, increasing the ease of taking transit to local businesses can lower barriers to using mass transit as an alternative to single occupancy vehicles. This can help relieve road congestion and reduce air pollution.

- Who is the audience for your report?

  The Chicago Transit Authority (CTA) would be the primary target audience for our project. However, if our model proves useful, it might be of interest and replicable for other regional transit agencies.

- What kinds of actions could be taken based on your results, and who is equipped to take those actions?

  The CTA could supplement current route scheduling and expansion plans by incorporating under-served areas highlighted by our model to close any identified service gaps.

- How will you validate whether your results might be relevant to your intended audience?

  There are currently four bus and two L-train expansion projects listed on the CTA website[2]. There are also three discarded studies and a few capacity reports. It will be interesting to see if our model's suggestions align with what the CTA is actively pursuing (or not interested in pursuing).

  Since our preliminary data analysis shows that both total population and ridership have been slowly decreasing over the past five years, the CTA likely cares deeply about identifying potential future ridership growth areas. We hope our analysis will accomplish finding these.

Data

- Describe the data you have and the data you'll need to collect.

  We will use business license data from the Chicago Data Portal to identify trends in businesses opening and closing and identify areas of business growth. We will use L train and bus route/ridership data from the CTA to identify both features for accessibility and target ridership numbers. We will also use U.S. Census Bureau data for local population estimates, and affordable housing development data from the Chicago Data Portal to derive locations from which residents travel. This list of datasets may expand if we discover something that fits our needs, but other than geographic lat/long locations to supplement the census data, we believe we have what we need to proceed.

- Include some descriptive stats that show you have enough to solve the problem.
    - Chicago Transit Authority L-train data[3]:
        - 145 Train Station Names with daily ride counts for years 2015 to 2019
        - 109 stations with known latitude and longitude locations (lat-long)
        - Annual numbers of train boardings:
            - 2015: 197,807,452
            - 2016: 195,555,726
            - 2017: 188,665,453
            - 2018: 185,146,121
            - 2019: 179,071,205
    - Chicago Transit Authority bus route data[3]:
        - 185 routes with daily ridership totals for between 2001 and 2019.
        - 127 routes with lat-long data
        - Annual number of riders:
            - 2015: 272,122,558
            - 2016: 257,917,007
            - 2017: 248,107,289
            - 2018: 240,995,859
            - 2019: 235,636,572
    - 1,002,443 recorded business licenses from 2003 to present[4].
        - 58,510 currently active business licenses across 77 precincts in Chicago.
            - 45,363 renewals
            - 12,124 new businesses
            - 1,023 changes to previous businesses
        - 52,860 active businesses with lat-long data.
    - 2010 U.S. Census population count data[5].
        - 46,298 block-level population totals within Chicago.
            - 2,695,598 total persons

    - U.S. American Community Survey 1-Year Estimate Data[6].
        - 17 combined neighborhood areas in Chicago (Public Use Microdata Areas - PUMA) with population estimates for years 2012 to 2018
            - 2012: 2,650,479
            - 2013: 2,660,636
            - 2014: 2,667,304
            - 2015: 2,658,370
            - 2016: 2,638,501
            - 2017: 2,649,402
            - 2018: 2,641,907
    - The City of Chicago subsidized affordable housing development locations[7].
        - 428 housing developments
        - 388 with lat/long data

- What analysis do you plan to perform or show on this data, to help inform your understanding of the data, as well as your design and selection of Machine Learning Models.

  We will need to do extensive data cleaning and engineering to transform it into usable features. Overlapping date ranges of our datasets will need to be identified. Locational data will have to be found and matched to Census records. Distance thresholds and areas to use for analysis will need to be determined. Data transformations rely on assumptions that need to be decided on and documented.

  For example, unlike L-trains, CTA bus ridership data does not include where people board, only daily totals for each bus route. So it will be hard to know how much traffic each neighborhood is getting from any single route. We propose using population estimates to "weight" bus route ridership totals for each area that a route traverses through to solve this problem.

  The nature of our data and our policy question both inform our selection of machine learning models. Since the ridership question lends itself to quantitative analysis, we anticipate using regression. To form a list of top areas/routes to recommend for improvements, we will either rank outcomes purely based on our regression results, or if time allows, try to train a more complex model.

Machine Learning

- What type of machine learning problem is this?

  Our question is in two parts: First we want to predict what will happen with ridership numbers over time based on business growth. Second, we want to compare the trends in business numbers and ridership/route numbers to classify an area as adequately, under or over-served.

  We plan on training regression models on ridership numbers and then creating a ranking based on the predicted ridership numbers to classify any discrepancy between an area's business growth and route/ridership growth to identify under-served areas.

  In the end we hope to use a combination of prediction and either manual grouping by calculated rankings or a more formal classification algorithm to make recommendations based on an analysis of learned outcomes.

- What types of models will you apply?

  We will need to continue to work with our gathered data to shape the details of our design, but our preliminary ideas for the feature table / targets are:

  - Features:
    - **Area**
    - **Year**
    - **Business Features:** growth over prior year, totals, special types
    - **Housing Features:** area population, growth over prior year, area affordable housing
    - **Transit Accessibility Features:** connected routes, population connected to the area by transit, connected population growth over prior year
  - Targets:
    - **Change in Ridership for Relevant Routes Next Year** (target of our prediction model)
    - **Area Accessibility Ranking**
      (classifier for business growth compared to accessibility growth)

  For the question of ridership, provided we can group our data in small enough areas to have a large number of records, we anticipate using grid search with cross-validation to find a suitable regression model.

With accessibility rankings we are interested in finding bus or train routes with ridership access trends that diverge from area business trends. The simplest way we see of doing this would be to take regression outputs for expected ridership growth and compare them to expected business growth. This could be accomplished with a simple formula.

Our primary concern is getting ridership predictions correct, but given time we may try various classification models, like k-nearest neighbors or decision trees, with the end goal of identifying routes that pass through the most areas with accessibility rankings and demographics indicating changes are needed. Our goal with ranking is to identify priorities for transit resources, so if a simple calculated function accomplishes this we may not utilize machine learning in our second step. Attempting a second machine learning task for classification is open for revision.

Once we have these rankings for areas,, since routes can be modeled as groups of areas, we could compare the approaches of aggregating by:

- grouping areas with similar rankings then identifying common routes, or
- grouping areas connected by routes and measuring the similarity of rankings for connected areas.

Evaluation

- Describe your process for evaluating the models.

We plan to start with the regression on predicted ridership using temporal holdouts and a grid search with cross validation over a number of folds that is appropriate to our final data set size. Once we have our features built, graphing business features against ridership features should help us determine if anything more involved than polynomial basis expansion might need to be tested.

After finding the best regression model, if we use classification to find a set of routes for which changes can be suggested, such models can be evaluated both on group purity (are routes with the most similar accessibility rankings placed together by the model) and against manual estimates for how we think routes priorities might be selected (accuracy against test/train grouping data).

- How will you validate the correctness of the models?

  We are hopeful that training and testing across the data divided by year in cross-validation folds will help us not overfit our models and get a good idea of what our out-of-sample errors may be.

  For the ridership regression, we will use real values from test data to validate against predicted ridership growth numbers for the current year. We may use scoring methods along with precision metrics to make sure we are predicting correctly the areas with highest ridership growth, for instance. We will also test regularization techniques like ridge, lasso and elastic net in order to balance variance in the model against bias in overfitting.

  For classification of routes, we will primarily use purity measures for groups of routes that are high priority or low priority (or somewhere in between) based on area rankings. In addition to numerical measures (like Gini purity), since recommendations are somewhat subjective, we would want to compare any potential model outputs to our own manual selections on a small group of testing data to see if the model makes sense by human logic.

  One concern is a possible lack of heterogeneity in data. If all areas are experiencing similar trends, this could result in imbalanced classes and muddy results. If this is the case, we may have to look into rebalancing the data for the purposes of model selection.

Ethics

- Briefly discuss the ethical implications of your work.

  The primary assumption of our underlying model is that counts of active business licenses provides a good indicator of areas that should have high public transit ridership. Further we assume that trends in business licenses should be mirrored by trends in ridership.

  We may also need to consider including supplemental data or weights to areas based on some characteristics. For example, we may wish to avoid prioritizing a cluster of new questionably socially optimal businesses (like a dense group of liquor stores and minimarts) over an older group of critical businesses or services (like grocery stores, clinics and hospitals).

  Our models will also use a number of geographic distance measures to group businesses, routes and populations into areas. These measures will be somewhat arbitrary because our model does not have the ability to “see” access barriers other than distance. For example, if there is a large freeway with no overpass between a local block and a bus route, but it’s still within our distance radius, that route might be deemed accessible in error.

Recommended outcomes using these assumptions could lead to missing accessibility barriers that exist in the real world for some people. It also could favor decreasing residential-to-residential routes in favor of residential-to-business routes, which could negatively impact people primarily working in other people’s homes.

Most importantly, because our work would be recommending expansion or increasing transit resources towards routes and areas experiencing business growth, there is a good chance that our models may imply funneling resources towards already affluent areas. This would be very problematic and introduce issues of fairness if most of the proposed areas end up being already well-resourced neighborhoods.

References

1. “Critical Issues in Transportation 2019.” Critical Issues in Transportation 2019: Policy Snapshot. The National Academies Press. Accessed May 3, 2020. https://www.nap.edu/resource/25314/criticalissues/.

2. “Planning & Expansion Projects.” Chicago Transit Authority. Accessed May 3, 2020. https://www.transitchicago.com/planning/.

3. Chicago Transit Authority. “CTA - List of CTA Datasets: City of Chicago: Data Portal.” *Chicago Data Portal*, 17 May 2017, data.cityofchicago.org/Transportation/CTA-List-of-CTA-Datasets/pnau-cf66. Accessed May 3, 2020.

4. Chicago, City of. “Business Licenses: City of Chicago: Data Portal.” Chicago Data Portal, 4 May 2020, data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr. Accessed May 3, 2020.

5. “Total Population Survey, Table: P1.” Data.census.gov/Cedsci/, Decennial Census: 2010, 7 Apr. 2020, data.census.gov/. Accessed May 3, 2020.

6. “Total Population Survey, Table: B01003.” Data.census.gov/Cedsci/, American Community Survey Years: 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 7 Apr. 2020, data.census.gov/. Accessed May 3, 2020.

7. Chicago, City of. “Affordable Rental Housing Developments: City of Chicago: Data Portal.” Chicago Data Portal, 12 July 2019, data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi. Accessed May 3, 2020.