CAPP 30254 Spring 2020
Prof. Nick Feamster

Predicting Transit Service and Local Area Growth for Public Transit System Investment

Final Project Report

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

UChi GitLab Repository Link

Executive Summary

In order to address declining ridership and fulfill their mission for Transit Oriented Development in the city, the Chicago Transit Authority must utilize a data-driven strategy to direct their efforts and funds towards areas with the highest growing need for transit resources. Using block group-level information about local business growth, transit options, and demographics, our team employed machine learning methods to select a model that would best predict ridership growth in the next year, and identify which factors are most important for these predictions. We found that the age of residents and business growth were the most consistently important factors in predicting ridership.

Background and Overview

Policy Problem

Transportation systems are an important part of a region's economic health.[1] The Chicago Transit Authority (CTA), a public entity charged with running the bus and heavy rapid transit train system (the "L") within the city of Chicago, holds a mission that reflects this view: "We deliver quality, affordable transit services that link people, jobs and communities." The CTA has also published studies specifically seeking to implement a Transit Oriented Development approach to regional planning.[2]

Transit Oriented Development is a methodology that seeks to build communities around existing transit hubs, incorporating mixed use building plans, walkability and transit accessibility in order to drive economic growth.[3] Under this urban development philosophy, sustainable community growth occurs hand-in-hand with increased ridership.

The CTA provided over 400 million rides last year. Despite that impressive number, ridership has been slowly declining and resources, whether tax dollars, fares or capital allocations to infrastructure, are limited. Clearly the CTA must prioritize service to the areas it matters most. What areas should the CTA target for investment?

Proposed Solution

Given the theoretical link between successful transit delivery and community growth offered by Transit Oriented Development, areas experiencing business growth should ideally also experience increases in ridership. If such a relationship can be established through machine learning, it becomes possible to predict where transit resources should be prioritized by studying the growth of community areas.

Audience and Actions

The Chicago Transit Authority itself is the main audience for this report. The proposed analysis will produce a model that would best provide the CTA with predictions of future ridership trends in Chicago, while also identifying which factors are most important for predictions. We feel findings can offer insight into how the CTA might engage in planning efforts. Our model may also allow them to monitor progress in supporting regional economic development goals more transparently.

Data Used

At its most basic level, Transit Oriented Development tells us, done right, an increase in the number of businesses in a concentrated area should be strongly correlated with increased transit ridership.

To model this we primarily used two datasets: the CTA's own ridership data[4] and a registry of active business licences published on the City of Chicago Open Data Portal.[5] To account for possible confounding factors, we also incorporated American Community Survey (ACS)[6] demographic data. To have enough variation in the data and as much local specificity as possible, we aggregated by US Census block groups and annual per month estimates for both ridership and active business counts. There is an overlapping dataset that includes 2,186 unique block groups for four complete years from 2015 to 2018.

Table 1 - List of Data Fields Used	
Features	
Geographic Feature	- Census Block Group
Time Features	- Year - Month
Business Features	- Active businesses in that area that month. - Percent change in active businesses from that month in the prior year. - New businesses in that area that month. - Percent change in new businesses from that month in the prior year.
Demographic Features*	- Estimated population. - Percent change in population from prior year. - Median income. - Percent change in Median Income from prior year. - Average age. - Percent change in average age from prior year.
Transit Accessibility Features	- Number of bus routes intersecting block groups geographically. - Percent change in bus routes from prior year. - Number of workers reporting they use public transit to get to work.* - Percent of workers using public transit to get to work.* - Percent change from the prior year for both count of workers and percent of workers using transit to get to work.*

Target	
Ridership	Combined bus and train ridership numbers per area per month.

* starred attributes are from the ACS and only available annually.

Dataset Assumptions

Fitting source data into month/block group units required some specific decisions and assumptions.

Since there was no stop-by-stop ridership totals available for bus routes, we used route path geography to average ridership over intersecting block groups weighted by each block group's population.

The opposite was true for "L" train ridership, where stop data was available. For train ridership, stop data was assigned to block groups within a 2 km (1.24 mi) radius.

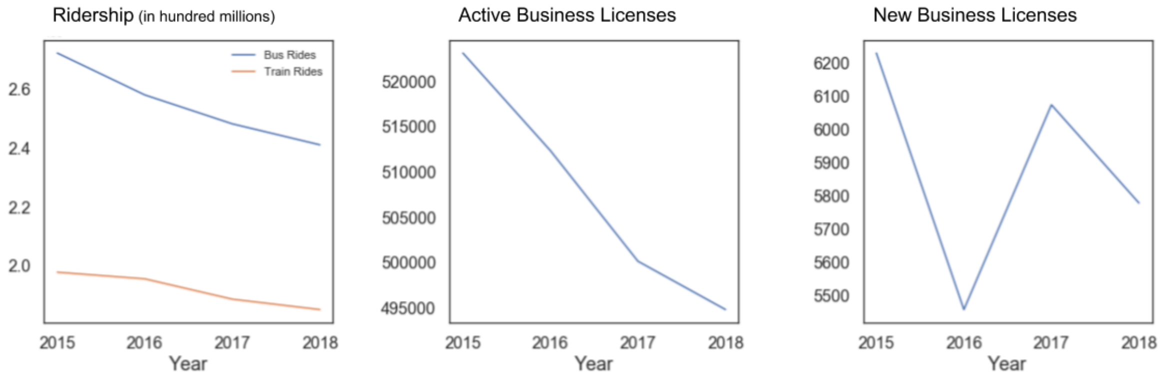
Business data was represented per license application (whether a renewal or new issue) and account numbers could apply for multiple business licenses for multiple locations. We counted one active business for each unique address where a license is active during a given month-year and counted a new business if that unique address's license was newly issued during that month-year.

Initial Data Exploration

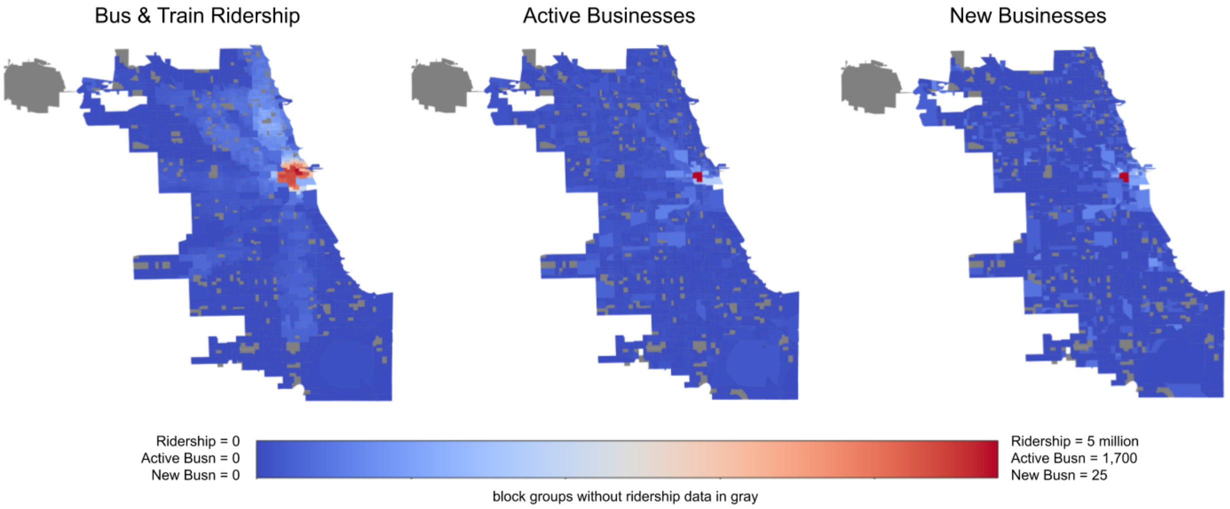
Both overall public transit ridership and overall active business licenses show a declining trend. New business licences, however, show greater annual variation and there is geographical variation for all three variables.

Beyond a downtown "loop" epicenter outlier visible in the choropleth maps above, there are no immediately obvious correlations between business data and ridership data when plotted against each other.

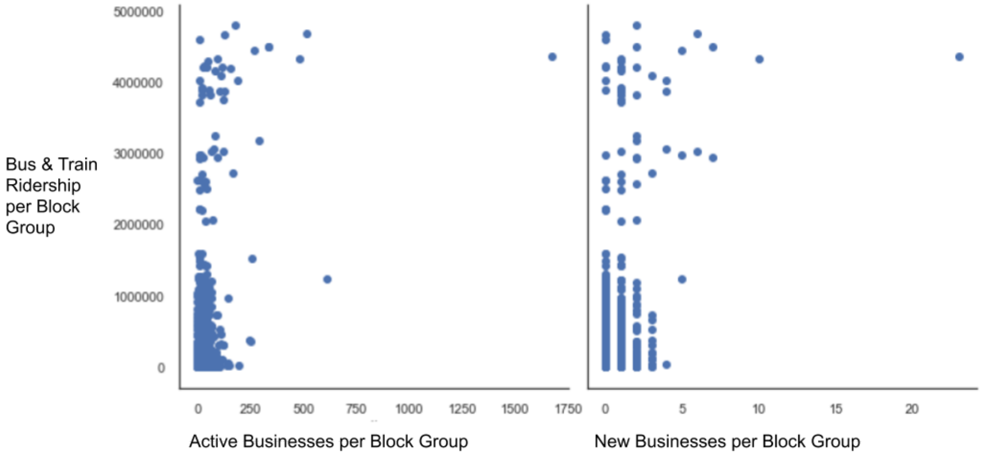
Annual Counts for Ridership and Business Licenses



January 2018 Block Group Counts



January 2018 Ridership Compared to Active and New Business Counts



Machine Learning and Details of Solution

Machine Learning Problem

In order to understand the predictive relationship between business growth and ridership, we used continuous numerical values and regression-style learning models. We chose regression rather than classification as a means to understand this policy problem because we are interested in modeling how to predict ridership numbers, rather than dividing areas into classes of ridership (e.g. top 10% in the city).

Machine Learning Models

To find the best regression model, we used the following estimators from scikit-learn's open source python machine learning repository.

Table 2 - List Models Evaluated

Models	Tuned Hyperparameters	
Support Vector Machine - Regression (SVR)	**kernel**	**C**
	rbf, poly, linear	0.1, 1, 10, 100, 1000, 10000
Linear Models:	**Polynomial Degree**	**alpha**
LinearRegression	1, 2	n/a
Ridge	1, 2	0.1, 1, 10, 100
Lasso	1, 2	0.1, 1, 10, 100

Feature Selection

Initial test regressions included the full range of features from our cleaned source datasets. Numerical features were all normalized by standard scalar (standard deviations from the mean) so coefficient magnitude in the regression results can be meaningfully compared. 0.6% of data was missing, and these were filled with the feature's mean value. Categorical features were one-hot encoded (to create dummy variables).

We decided to drop primary neighborhood dummy variables out of the final model. Including them gave higher predictive power (R^2 = 0.91), but the full set of most important features became neighborhoods. We determined that having neighborhoods as the most important feature set may cause overfitting. While neighborhood importance does suggest further investigation (to unobserved characteristics of high ridership neighborhoods), it is less helpful in generalizing our model.

In the end, we used all the remaining features with polynomial features of degree 2 to allow interactions and behavior that is not strictly linear.

Evaluation and Results

Training, Testing and Evaluation

We used temporal holdouts to cross-validate the learning models, enumerated in table 2, such that 2015 data was used to predict 2016 data and so on. Hyperparameters were tuned based on minimizing Mean Absolute Error (MAE) and Mean Squared Error (MSE) and maximizing explained variance (R^2).

The models that performed the best without a neighborhood feature are listed in table 3.

Table 3 - Model Selection: Evaluation Results

Model	Parameters	MAE	MSE	R^2

LinearRegression	year=2016,
	no_neigh,
	poly=2
		290,067.7	228 x 10^{10}	0.597

LassoRegression	year=2016,
	no_neigh,
	poly=2,
	alpha=100
		289,354.0	229 x 10^{10}	0.595

LassoRegression	year=2015,
	no_neigh,
	poly=True,
	alpha=100
		299,355.8	242 x 10^{10}	0.586

LinearRegression	year=2015,
	no_neigh,
	poly=True
		300,347.4	242 x 10^{10}	0.585

For our final predictive test, we used a pool of 2015-2016 data to train the classifiers to predict 2016-2017 ridership, then tested these models on 2017 data, predicting 2018 ridership.

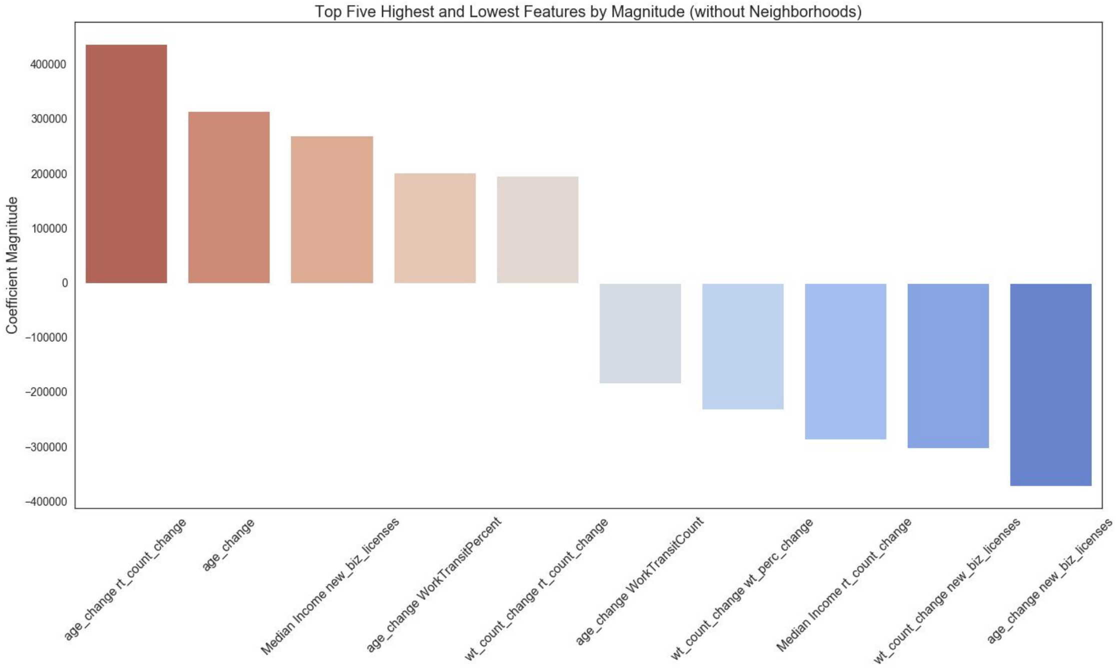
Using this final test, we selected a Lasso regularized regression without neighborhood dummy variables, with polynomial features degree 2, and an alpha level of 100 as it gave the best balance between error and R^2 without evidence of overfitting.

Feature Importance

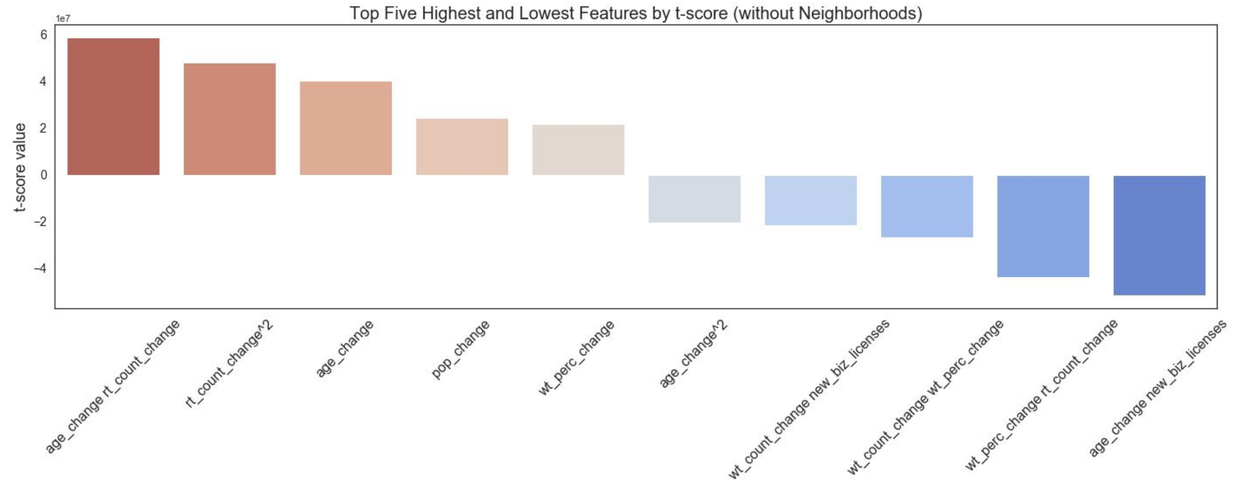
The part of our model most critical to our policy question is feature importance.

Business growth appears in four variables: count of new business licences, increase in new business licences over prior year, count of active business licences and increase in active business license over prior year. Of these, the count of new business licences showed up in the most important features three times: interacted with median income for a positive correlation with ridership, and interacted with changes in both age and the count of commuters using transit for a negative correlation with ridership.

The most predictive features of ridership change for a block group are: changes in **median age**, change in the **number of bus routes**, change in the **number/percentage of commuters using public transit** and the **number of new business licences**.



Using t-tests for statistical significance, the most likely features to be predictive of ridership changes and differentiable from noise in the model are: change in **median age,** changes in the **number of bus routes**, changes in **population**, change in the **number/percentage of commuters using public transit** and the **number of new business licences**.



Of particular interest: increase in median income alone has a negative impact on ridership, but when interacted with new business growth, the impact is positive (and in our top five features according to magnitude). Age alone has a positive correlation with ridership, but age interacted with new business licenses becomes negative. Similarly, the number of commuters using public transit interacted with new business licenses has a negative correlation with ridership.

One possible interpretation is that as an area experiences income growth and business growth, more people are drawn there by transit. While as an area becomes older on average, fewer people commute outside of their block group for work or to visit businesses, especially if there is business growth nearby.

Policy Recommendations

Using our machine learning model (Lasso normalized regression with polynomial features, alpha=100), it is clear demographics like the age and income of a block group are strongly tied to predicting trends in ridership. Things within the realm of policy change, such as access to bus routes, commuter utilization of transit and business growth do have predictive power. We feel our model has use in predicting ridership, though there is room for refinement (discussed below).

A tangible policy recommendation could be to de-emphasize areas with aging populations for transit service, provided there are adequate businesses within a walkable vicinity.

Also, as suggested by the Transit Oriented Development framework, an area experiencing both business growth and increases in income is likely to experience more ridership. These areas should be prioritized for transit services.

Ethics

While our findings confirm a correlation between business growth and ridership growth, we are concerned machine learning models (and the TOD framework in general) may imply funneling resources towards already affluent areas. Continuously supporting areas of high business growth with services that bolster business growth could become a vicious cycle.

Instead, perhaps areas identified with high business activity and low median incomes could be prioritized more, in line with the spirit of sustainable growth. Even this recommendation is liable to create gentrification and displacement for current residents if coordination with affordable housing development is not carefully considered.

Limitations, Caveats, Suggestions for Future Work

There are a few model limitations that are imposed by the dataset worth mentioning:

- Population is tied to ridership for bus routes in a way that may artificially skew data.
- Using a radius of 2 kilometers for train ridership may be too far for walkability. In fact, Transit Oriented Development focuses on quarter mile walking distances.
- The model does not have the ability to “see” access barriers other than distance.

For future refinement, elaborations on the theme established here might include:

- Adding features: for example, alternative forms of transportation like taxis or bikes.
- Development of an impact metric to identity areas that could most benefit from ridership access in order to combat potential ethical blindspots.
- Disaggregating business licence data based on type of business to pull out business more likely to employ workers, mixed use areas aligned to TOD philosophy and/or including social service locations such as hospitals, government agencies and schools.

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. Cervero, Robert; et al. (2004). Transit Oriented Development in America: Experiences, Challenges, and Prospects. Washington: Transit Cooperative Research Program, Report 102. ISBN 978-0-309-08795-7.

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. 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.

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