Ticket to Ride Where? Predicting Commuter Ridership to Inform Transit Expansion Decisions

Nathan Dignazio [12244525], Mike Feldman [12240000], and Nguyen Luong [12245255]
GitHub Repository: https://github.com/ndignazio/transit-ml

Executive Summary

While decisions about public transportation expansions can have substantial socioeconomic and environmental consequences, the means of predicting ridership are often resource-intensive, time-inefficient, or entirely infeasible for large and small localities alike. Our analysis aims to produce a cost-effective, efficient, and generalizable model geared towards interpretable and actionable policy suggestions. We combine Census and real estate data to build regularization- and tree-based regression models that predict commuter transit ridership in Illinois Census tracts, using root mean squared error (RMSE) as our evaluation metric. We find that a Random Forest model minimizes RMSE and confirms that socio-economic attributes are important indicators of transit ridership. Using our best model, we identify 30 Census tracts in Illinois that we believe should be considered by local and state policymakers as candidates for public transit expansions. While we acknowledge that data limitations may introduce bias against rural locations in our model, we believe that this hurdle can be remedied with more comprehensive data. The results of our analysis yield grounded suggestions that policymakers with domain knowledge can incorporate into locally-informed decision-making.

Background and Overview of Solution

Decisions around public transportation have become increasingly impactful for both localities and their residents. Not only are these decisions a key factor in urban planning but they are also a means of driving environmentally-conscious urban development. Thriving public transit systems strongly benefit senior and low-income populations by connecting job-seekers with work opportunities, permitting greater mobility and self-sufficiency for the physically-impaired, and reducing household expenses (by eliminating the need to own a vehicle).[1] Public transit is an especially important consideration for smaller localities due to their higher share of aging and low-mobility populations.

Given the impactful nature of public transit implementation and expansion decisions, the demanding nature of current forecasting methods is not ideal. In larger cities, forecasting ridership typically takes the form of four-step conventional synthetic travel-demand modeling, a procedure that is resource-intensive, time-intensive, and frequently overestimates ridership.[2] [3] Smaller localities, on the other hand, often lack the resource capacity to evaluate, demonstrate, and advocate for their needs.[4] In response to these needs, we build a generalizable model of ridership prediction that is cost-effective, time-efficient, and easily interpretable to maximize

[1] https://nascsp.org/wp-content/uploads/2018/02/issuebrief-benefitsofruralpublictransportation.pdf
[2] http://dspace.calstate.edu/bitstream/handle/10211.3/212124/JanicekLyle_Thesis2019.pdf?sequence=3
[3] https://pdfs.semanticscholar.org/3849/5b499381a69cc2effbcf7a03f6c78208040f.pdf
[4] http://t4america.org/wp-content/uploads/2010/03/T4-Whitepaper-Rural-and-Small-Town-Communities.pdf

usefulness to local policymakers in both larger cities and small townships in considering a change to their current state of public transit.

The aim of this analysis is not to give a stringent directive, but rather to identify key geographic areas that would benefit from an expansion of public transit services and patterns in ridership by locality that warrant a closer look by local policymakers with domain expertise. These results can give town mayors and city councils considering public transit expansion or implementation a tool that can be used in tandem with more qualitative methods (such as surveys and town halls) to help determine where in their community transit demand might be highest, and to more generally gauge the potential of such a project. Further, given the reliance of localities on state and federal grants for transportation projects, we hope that local governments can include the output of our model in grant proposals to demonstrate the promise of public transit expansion in their communities.

Data

Our analysis draws from three main sources of data. The first data source is the American Community Survey (ACS), collected by the U.S. Census Bureau. The ACS is a current and reliable data source for local statistics on planning topics including commuting, income, and employment. Designed to yield information on small areas and population groups, the ACS covers over 35 topics from roughly 3.5 million households, supports over 300 Federal government uses, and informs "$675 billion of Federal government spending each year."[5] We use the 5-year estimates from 2014-2018 and pull 77 demographic variables for the state of Illinois. To supplement this set of demographic variables, we also use the Origin-Destination Employment Statistics (LODES) for 2017 from the U.S. Census Longitudinal Employer-Household Dynamics program. Lastly, we obtain a key feature called a transit score from an API called WalkScore, a data provider owned by Redfin.

The transit score metric is a measure of how well a location is served by public transit on a scale of 0 to 100 and is often used in a real estate context. The transit score metric allows for the consideration of differential levels of public transit connectivity across the Census tracts. Because this metric is available only for cities with public transit systems who also publish their data, 1320 Illinois Census tracts were omitted, leaving 1795 tracts for our analysis. Using a transit connectivity score limited to the census-defined place level from an alternative source called AllTransit, we find that these omitted tracts overlap with places that overwhelmingly exhibit low transit scores.[6] Given that these omitted tracts, which have very low or zero-valued transit scores, constitute a considerable share of all available tracts, we believe that imputing missing values of the transit score attribute with zero values weakens our ability to capture relationships between those tracts for which we do observe transit connectedness, imputing missing values with other statistics incorrectly captures true transit connectivity, and the bias

[5] https://www2.census.gov/about/training-workshops/2020/2020-02-12-clmso-presentation.pdf?

[6] Roughly 80% of the missing tracts fell below the 10th percentile of transit score and roughly 90% fell below the 25th percentile. See Appendix 1.

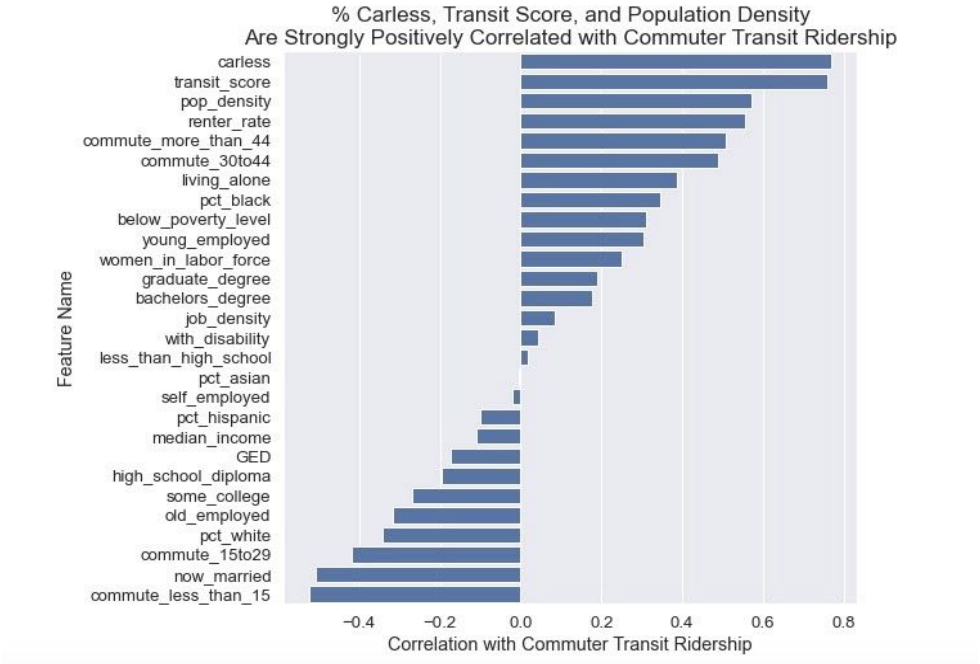
introduced is not significant. Further discussion of the implications of our data omission takes place in the Modeling and Ethical Considerations sections.

The consolidation of these data allow us to carry out our goal of using demographic data for Illinois Census tracts to predict public transit ridership. Our unit of analysis is a single Census tract in the state of Illinois and our outcome label is the proportion of total commuters who take public transit to work. In exploring this demographic data, observe the following distributions for our target and select features in our dataset:

	commuting_ridership	transit_score	carless	median_income	pop_density	pct_white	pct_black
count	1795.000000	1795.000000	1795.000000	1794.000000	1795.000000	1795.000000	1795.000000
mean	0.167428	44.074652	0.092390	66810.863434	4244.255876	0.579734	0.250422
std	0.145940	23.241258	0.112493	36035.932895	5008.674506	0.299826	0.332813
min	0.000000	0.000000	0.000000	3558.000000	9.168791	0.000000	0.000000
25%	0.047487	27.000000	0.012250	41287.750000	1262.264135	0.402264	0.024442
50%	0.120577	43.000000	0.041739	58406.500000	2576.257997	0.663640	0.069164
75%	0.260351	64.000000	0.137177	84944.000000	5759.929627	0.827443	0.353356
max	0.740741	100.000000	0.675926	250001.000000	101928.125283	0.992210	1.000000

Note that while the median transit score is 43 units, the median commuting ridership is only 12.05% of the commuting population over all Census tracts.

In examining the relationship between our features and our target, we observe that the variables most strongly positively correlated with commuter transit ridership are the percent of commuters who do not own a car, the transit score, and population density. See Appendix 2 for distributions of the most positively correlated variables and visualizations of their relationship with commuter transit ridership.



Modeling

Further Data Considerations

All of the data gathered from the ACS, LODES and WalkScore API is numeric data.[7] Originally, our intention was to maximize granularity of the unit of analysis, keeping in mind the needs of our policy question. Although Census blocks allow for analysis on the most granular level, the typical population of a Census block is around two thousand people, while block groups, the next level up in the geographic hierarchy, generally have up to four thousand people. Block groups ensure a high degree of granularity with a greater degree of statistical validity due to increases in sample size. Having settled on block groups as the primary census geography to use in our analysis, we discovered the existence of data suppression by the ACS for several features we wished to include at the block group level.[8] Because of data suppression, we conduct our analysis at the Census tract level, the next highest level from block groups in the geographic hierarchy. All data pulled from the ACS came in the form of population counts. Most data were normalized into probability levels between 0 and 1 by dividing every feature by the total population count, with the exception of population and employment densities, explained below.

The target variable of the model is a proportion: the number of commuters in a Census tract whose primary means of transportation to work is public transit relative to the total number of commuters in a Census tract. We recognize that commuter ridership is not precisely equivalent to total ridership. Furthermore, the target variable does not take into account the seasonal and annual fluctuations in transit ridership by geography. Instead of a panel model that takes into account time series, our model is cross-sectional and the target variable is averaged over 5 years of survey results from the ACS. In other words, the goal of the model is not to forecast ridership next month – the goal of the model is to forecast average ridership rates over the course of years as the result of potential long-term investment at the municipal and state level in public transportation in local communities across Illinois.

One way to frame our policy question is: which Census tracts in the state of Illinois would benefit most from increased investment in public transportation? We define "benefit" as the projected percentage increase in commuter ridership in a tract given a 10-point increase in transit score from the WalkScore API. The process is to perform a grid search to find the best model and the best hyperparameters that most effectively explain the target variable of commuting ridership, to train this model on the entire training set and test the model on the entire test set, and then to generate a baseline of predicted commuting ridership levels based on the actual transit score reported from the WalkScore API in each Census tract. The projected percentage increase in ridership is generated by adding 10 points to the transit score of every tract in the dataset and calculating the increase in projections as a result of the constant 10-point increase.

Our analysis will produce a rank order of Census tracts with the highest projected benefit given these criteria. While the target variable is quantitative, the ultimate policy

---

[7] See the GitHub repository for a dictionary of ACS table ID's used to generate features: https://github.com/ndignazio/transit-ml

[8] https://www.census.gov/programs-surveys/acs/technical-documentation/data-suppression.htm

recommendations will be qualitative: that is, we will recommend the state prioritize investment in the specific tract areas with the highest projected increase in ridership. These communities likely stand to benefit the most from increased access to public transportation, while state and local government may expect a reasonable return on investment based on high usage and resultant revenue from ticket fares. Since our goal is not to classify tracts in need of intervention but rather to identify tracts receiving the largest potential benefits of transit expansion, we judged it reasonable to keep the features and target numeric and frame the question as a regression problem.

Feature Selection

In order to identify potential features that have the highest effect on transit ridership, we reviewed academic papers that perform meta-analyses of the transit ridership forecasting literature.[9] Taylor and Fink note that factors influencing ridership can be separated into internal and external factors.

External factors are largely exogenous to the system and its managers, such as service area population and employment. Such factors typically function as proxies for large numbers of factors thought to affect transit demand. Internal factors, on the other hand, are those over which transit managers exercise some control, such as fares and service levels. They note that most studies consider internal and external factors separately, and that external factors often exert an outsize influence on ridership.

External factors can be related to socio-economic, spatial, and public finance factors. Although incorporating more information about land use would have been interesting and potentially helpful to our model, we chose to stick with socioeconomic indicators retrieved from the ACS. Employment density data are obtained from LODES, and transit score is obtained from the WalkScore API. These indicators include data about **disability status** ('with_disability'), **income and poverty levels** ('below_poverty_level', 'median_income'), **automobile access** ('carless'), **employment status by age group and sex** ('young_employed', 'old_employed', 'women_in_labor_force', 'self_employed'), **educational attainment** ('high_school_diploma', 'GED', 'some_college', 'bachelors_degree', 'graduate_degree', 'less_than_high_school',) **racial makeup** ('pct_white', 'pct_black', 'pct_asian' 'pct_hispanic'), **average commute time** ('commute_less_than_15', 'commute_15to29', 'commute_30to44', 'commute_more_than_44'), **marital status** ('now_married'), **household characteristics** ('living_alone', 'renter_rate'), **employment and population density per square kilometer** ('job_density', 'pop_density'), and **transit score** ('transit_score'). Note that transit scores are obtained given a specific latitude and longitude. As a result, an assumption of our model is that the transit score of the centroid of a Census tract is sufficiently representative of transit access in the tract as a whole.

[9] "The Factors Influencing Transit Ridership" by Brian D. Taylor and Camille N.Y. Fink, http://www.reconnectingamerica.org/assets/Uploads/ridersipfactors.pdf

Model Selection

We tested which regression model would perform best using five-fold cross validation. Five folds were selected given the total training set size of 1436 records with an 80-20 train/test split. We expected that using ten folds would generate unnecessarily small training and test sets in the process of cross-validation that would result in a high degree of variance in model performance.

We selected linear regression, regularization models Lasso, Ridge, Elastic Net, and tree-based regression models (decision trees and Random Forest) as models to test. We tested polynomial feature expansions of degrees 1 to 3, but initial best-performing models were consistently of degree 1 or 2 and the computational expense of running regressions on third-degree polynomial expansions was unwarranted, especially in tree-based models. Our selection of models ensures a broad range of methods with varying levels of sophistication and computational complexity. All data was standardized to fit a standard normal distribution in the linear regression, Lasso, Ridge, and Elastic Net models.

Regularization models were tested with alpha levels of 0.0001, 0.001, 0.01, and 0.1, with the best-performing model among linear regression and regularization models consistently being a second-degree Lasso regression with alpha equal to 0.001. Tree-based models were tested with maximum depths of 5, 10, 15, and 20 for decision trees and 5, 10, and 15 for Random Forest models, splitting criteria of mean squared error, Friedman mean squared error, and mean absolute error for decision trees and mean squared error and mean absolute error for Random Forests, and number of iterations of 100, 200, and 300 for Random Forests.[10]

The resulting best model was a Random Forest model with maximum depth of 15, 300 iterations, and degree 2 with a splitting criterion of mean absolute error. Although these parameters were on the higher end of our selection of hyperparameters, we decided to accept it as the best model given the computational expense of Random Forest models and requisite investment of time for diminishing returns on performance. We also considered the policy context of our recommendations: the aim is not to create an automated decision-making system with minimal human interference. If that were the case, it would be reasonable to go further in minimizing the error because of the repercussions of inaccurate predictions. Rather, the aim is to create an interpretable tool that local policymakers and community members can use to make their own decisions about where to allocate resources with regard to transportation infrastructure, given their knowledge of their local communities.

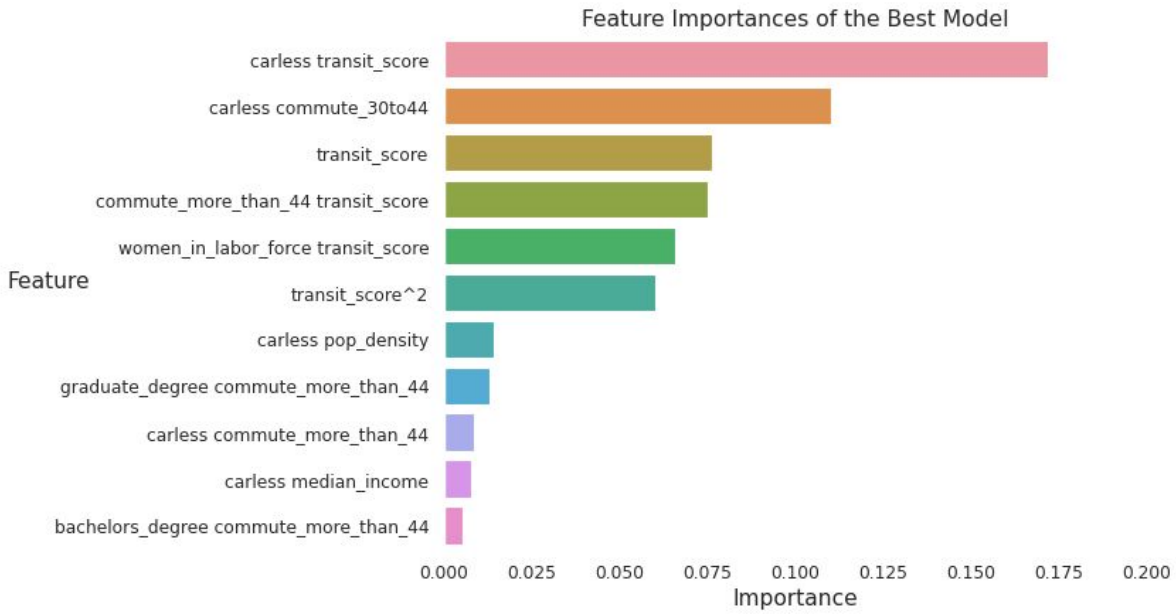
Evaluation

The evaluation metric used to assess the performance of each regression model is the root mean squared error (RMSE) metric. RMSE represents the standard deviation of errors in predicted commuter ridership relative to the actual commuter ridership in a Census tract. It is

[10] Given the size of our feature space and the presence of polynomial expansions, Random Forest models took the longest by far to run at several hours.

reported in the units of the target variable, which, as mentioned before, is a proportion between zero and one. RMSE was the best choice because of its clarity and ease of interpretation when evaluating models. The mean RMSE score in cross-validation was 0.051. When the best model was fitted to the entire training set and tested on the entire testing set, the RMSE score was 0.05. In addition, the model's R2 score on the entire test set was 0.882, meaning that it explained slightly more than 88 percent of the variance in commuting ridership. We judged this model's performance to be strong enough to use it to generate policy recommendations in the next section.

As the best model has 28 explanatory variables and a polynomial expansion degree of 2, it contains 435 features. Below are eleven of the most significant features of the model with the most explanatory power:



After 11 features, the relative importance of features begins to flatten. The plot above illustrates that various transportation-related characteristics are often the most significant predictors of commuter ridership. The exceptions in the top eleven feature importances are the interaction between the proportion of women in the labor force with transit score and between various levels of educational attainment and commute time longer than 44 minutes.

Policy Recommendations

Using our Random Forest regression model, we produce 30 Census tracts that we believe should be considered by local and state policymakers for public transit expansion, should the opportunity arise (see figure below). To arrive at these 30 tracts, we

1. Predicted commuting_ridership with the current feature set (including the reported transit score for each tract);

2. Increased transit_score for each tract by 10 points to simulate an expansion of transit, keeping all other features constant;[11]
3. Predicted commuting_ridership with the modified feature set (including the increased transit score);
4. Subtracted value (3) from value (1) to calculate the predicted change in commuting_ridership with an increase in transit_score by 10 points; and
5. Chose the 30 tracts with the largest predicted increases from (4) to be the ones recommended for review.

Tracts Recommended for Transit Project Review [12] [13]

	tract_id	commuting_ridership	pred_chg_commuting_ridership	transit_score	carless	below_poverty_level	pct_white	pct_black
576	17031710800	0.075682	0.083189	64	0.076303	0.245096	0.000844	0.986079
23	17197881900	0.031297	0.082063	43	0.109091	0.339978	0.305221	0.437483
1553	17113000200	0.011189	0.076871	49	0.100000	0.901373	0.786537	0.169184
1646	17163500400	0.120181	0.074644	48	0.181406	0.390873	0.054894	0.927249
138	17097862401	0.029292	0.070643	41	0.080027	0.249545	0.552153	0.127350
1515	17031150502	0.084926	0.069333	55	0.073326	0.167091	0.730014	0.018045
1473	17031020702	0.066748	0.068742	55	0.061361	0.220932	0.571149	0.087568
692	17031827100	0.115346	0.068665	52	0.090271	0.353175	0.085894	0.810481
1001	17031806004	0.059443	0.067940	48	0.020136	0.133400	0.690839	0.019928
1060	17031200300	0.104255	0.067334	59	0.074468	0.325688	0.514271	0.018858

We recognize the approach above might lack some precision (in the colloquial sense). For example, our assumption to keep all features constant in (2) is likely not to hold in reality Increases in transit connectivity may have effects on attributes like median income and the poverty rate, and could potentially change the demographic composition of the neighborhood, altering the rate of commuter transit ridership in ways our model does not capture. Yet we are also not strongly prescriptive in our recommendations. We realize that our model cannot synthesize all of the local context that drives transit usage, and that decisions around transit projects are often politically fraught, particularly given the typical high costs of American transit operations.[14] However, we do believe that the tracts our model recommends would respond particularly well to additional transit options in the form of increased ridership based on their demographic composition and current levels of transit connectivity. . We are agnostic to the form of public transit that might be best for these localities; we do not prescribe one over another.

---

[11] We choose a 10 point increase as our standard, given that RedFin revealed in 2019 that the highest transit score increase by a top-ranked city from 2018 to 2019 was +4. We hypothesize that Census tracts, which are much smaller in area than a city, have the potential for greater increases.

[12] Only 10 tracts are pictured here for space purposes. Note that all 30 tracts can be found in Appendix 3.

[13] Note that 'pred_chg_commuting_ridership' is the model-predicted change in the target, commuting_ridership, if transit_score were increased by 10 percentage points. For example, the first tract, 17031710800 is predicted to see its percentage of commuters riding public transit to increase by eight percentage points, from 7.6% to 15.8%, if its transit score were to increase by 10 points.

[14] https://www.vice.com/en_us/article/884kvk/why-the-us-sucks-at-building-public-transit

Nevertheless, we do implore local and state policymakers and activists to review this list and consider whether an expanded transit offering—whether more frequent service, an additional bus line, or a larger project—might be prioritized for any of the tracts included.

We have also enclosed below a list of tracts that most underperform our model's predictions. These are the tracts that, based solely on demographic makeup and transit connectivity, we would expect to have much higher commuter transit ridership than they actually do. For these tracts, we recommend that local policymakers and transit experts, who have substantial contextual knowledge of their locality, determine whether there might be low-cost interventions to increase transit ridership among commuters. For example, the first tract on the list, number 17031842900, is located in the Illinois Medical District of Chicago. About 44 percent of the tract lives below poverty, and 34 percent are carless; why then, do only 20 percent of commuters use public transit, 19 percentage points lower than our model-predicted ridership? Can this be improved?

Underperforming Tracts Recommended for Further Scrutiny[15]

	tract_id	diff_actual_and_model_pred	commuting_ridership	model_pred_ridership	transit_score	carless	below_poverty_level	pct_white	pct_black
1447	17031842900	-0.193355	0.204433	0.397788	68	0.344828	0.436466	0.141186	0.765631
1666	17163504600	-0.190295	0.074681	0.264976	57	0.250460	0.469286	0.019923	0.977864
526	17031283800	-0.166803	0.112139	0.278943	69	0.078014	0.217753	0.458530	0.393065
1256	17031081202	-0.155910	0.152174	0.308084	83	0.291667	0.042020	0.860386	0.021349
649	17031081800	-0.142579	0.140933	0.283512	99	0.261533	0.026212	0.853614	0.025305
1523	17031190402	-0.138739	0.056769	0.195507	60	0.064845	0.073223	0.731156	0.015255
1221	17031825802	-0.134582	0.021268	0.155850	31	0.033669	0.083752	0.066094	0.877855
1194	17031670700	-0.131570	0.080000	0.211570	66	0.056856	0.337687	0.111940	0.830224
1259	17031081403	-0.114872	0.163134	0.278006	77	0.399409	0.108253	0.715387	0.030077
1359	17031835700	-0.108303	0.000000	0.108303	76	0.000000	0.110577	0.000000	1.000000

Ethical Considerations

Given that our recommended intervention involves infusions of funds and infrastructure into localities, we recognize that our model has the potential to reinforce existing inequities. Ready access to public transit can be a lifeline to low-income and mobility-impaired individuals, and can also be the cornerstone for the growth and flourishing of neighborhoods. Yet access to public transit is far from equal, with gaps falling along familiar lines of race and class.[16] [17] We recognize this disparity between want and need of public transit among historically marginalized communities and do not wish to exacerbate this disparity through our work. On the contrary, precisely because of this gap between want and need, our model supports the case for expanding transit in under-resourced communities. The tracts identified by our model as the best candidates

[15] Only 10 tracts are pictured here for space purposes. Note that all 30 tracts can be found in Appendix 4.

[16] https://www.demos.org/research/move-thrive-public-transit-and-economic-opportunity-people-color

[17] https://www.chicagotribune.com/opinion/editorials/ct-editorial-south-side-transit-desert-20190618-2wk3fnq2mffvljivogml47upce-story.html

for transit expansion, at the median, have a median income that is $25k lower than the other tracts in our analysis, a poverty rate 14 points higher, job density 100 jobs lower, and % college completion 14 points lower.

Do we thus believe that our model is free from bias? We do not. Due to limitations with our data, our model may underweigh the needs of rural communities, particularly those with substantial senior and mobility-impaired populations that may have high need for, and low access to, public transit. In training our model, we were only able to consider as our target the rate at which commuters use public transit, which omits groups that do not work, but for which transit is high need. Also, our measure of transit connectivity, transit score, is only available for localities whose transit agencies publish data in machine-readable format. These localities are more likely to be urban centers, where transit is more widely available. In omitting localities that do not publish transit data, we may be missing sparsely-populated areas with relatively strong transit offerings, data which could have led our model to more strongly represent rural tracts in its 'recommendations' output.

To audit our model for bias against rural tracts, we would first need a dataset with complete transit score data for urban, suburban, and rural tracts across Illinois. Then, we would need to group tracts into 'rural', 'suburban', and 'urban' buckets and generate commuter ridership predictions for all of the tracts in each bucket. With those predictions we could compare the mean values across the three buckets and determine whether any statistically significant differences may be detected. We could interpret statistically significant differences across buckets, particularly those that persist over time, as prima facie evidence of bias in our model (particularly if it was against rural tracts, as expected).

Limitations, Caveats, Suggestions for Future Work

As discussed above, the prime limitations in our work stem from limitations in the transit data we were able to collect.

Our target, while highly reliable due to its source, is at best an imperfect proxy of overall transit ridership, which was our desired target. Because it is limited to measuring solely the transportation habits of commuters, our target omits groups such as seniors and youths, who are often prime consumers of public transit. As a result, we opted not to use as features attributes such as 'percent of population above 65', which is traditionally a strongly positive correlate with transit ridership (instead, we used percent of workers who are 60 or above).

Also, while transit score measures the proximity of a tract to transit options, it does not measure transit quality. Transit frequency, reliability, and safety are all vital to a strong transit system, and are entirely absent from our transit score measure.

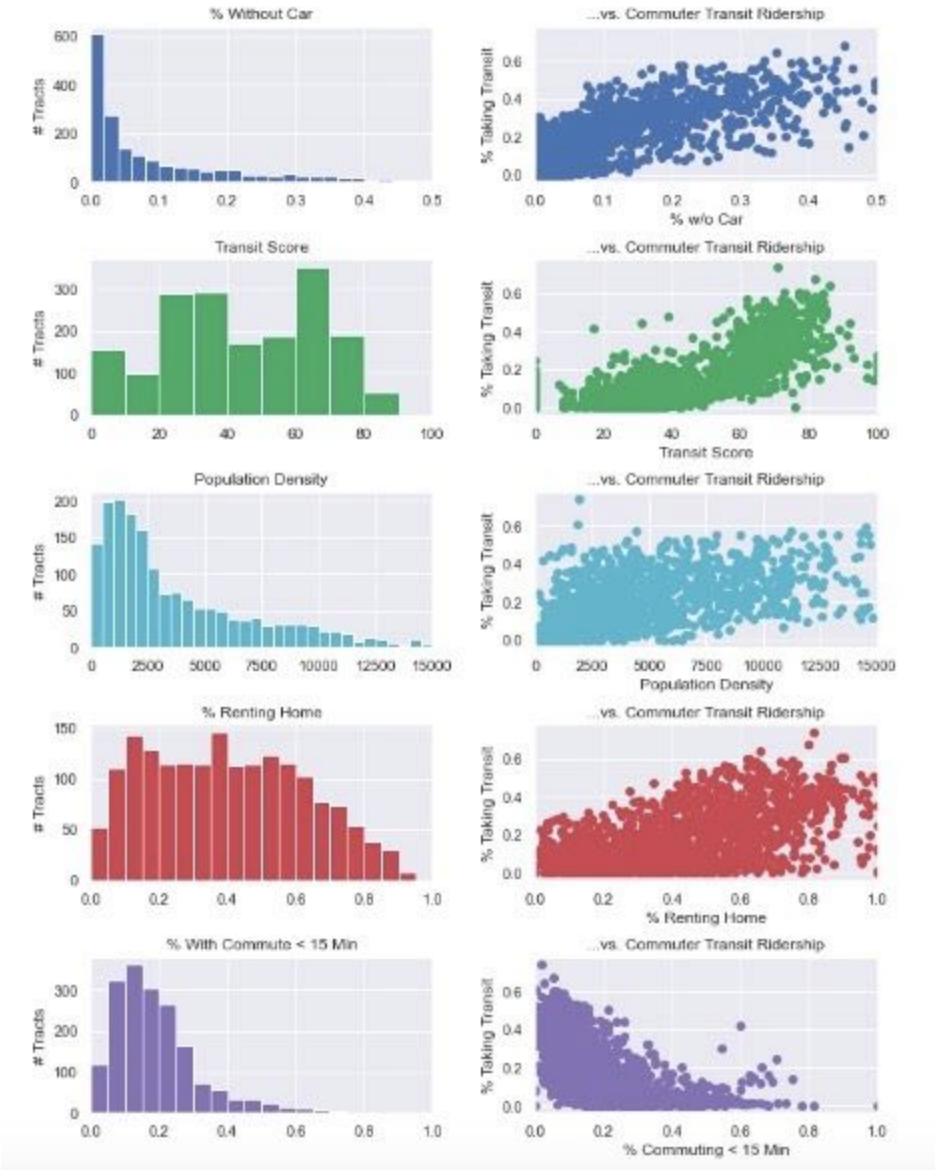
Future work might include obtaining a transit connectivity feature (or features) that better captures transit quality and is readily available for all tracts in Illinois. A target that more closely aligns with overall ridership would also likely yield predictions that are more highly-valued by policy-makers, and so would also likely be a fruitful avenue for future work.

Appendix 1 - Distribution of Transit Scores for Omitted Tracts

	communiting_ridership	alltransit_performance_score
count	1320.000000	1320.000000
mean	0.020771	1.469968
std	0.033757	1.605242
min	0.000000	0.000000
5%	0.000000	0.000000
10%	0.000000	0.000000
20%	0.000000	0.000000
25%	0.000000	0.000000
30%	0.000000	0.000000
40%	0.003241	0.021333
50%	0.007465	1.000000
60%	0.014426	1.900000
70%	0.023087	2.276667
75%	0.029773	2.666667
80%	0.037926	2.950000
90%	0.057558	3.755000
95%	0.077789	4.500000
max	0.504065	9.100000

Note: alltransit_performance_score has a range of 0-10 and is strongly correlated with the transit score metric used in our analysis

Appendix 2 - Distribution of Key Variables and Relationship with Target



Appendix 3 - Full List of Tracts Recommended for Transit Project Review

	tract_id	commuting_ridership	pred_chg_commuting_ridership	with_disability	below_poverty_level	carless	young_employed	old_employed
576	17031710800	0.075682	0.083189	0.213879	0.245096	0.076303	0.138052	0.042420
23	17197881900	0.031297	0.082063	0.154903	0.339978	0.109091	0.295930	0.027262
1553	17113000200	0.011189	0.076871	0.016637	0.901373	0.100000	0.402099	0.000000
1646	17163500400	0.120181	0.074644	0.257275	0.390873	0.181406	0.130054	0.047844
138	17097862401	0.029292	0.070643	0.119939	0.249545	0.080027	0.221987	0.046934
1515	17031150502	0.084926	0.069333	0.108497	0.167091	0.073326	0.110746	0.087463
1473	17031020702	0.066748	0.068742	0.091710	0.220932	0.061361	0.222476	0.050827
692	17031827100	0.115346	0.068665	0.141844	0.353175	0.090271	0.184974	0.038992
1001	17031806004	0.059443	0.067940	0.126925	0.133400	0.020136	0.110168	0.107339
1060	17031200300	0.104255	0.067334	0.088175	0.325688	0.074468	0.274834	0.035099
485	17031491100	0.328250	0.065827	0.226759	0.182075	0.183733	0.086082	0.069033
920	17031807100	0.066327	0.064637	0.094978	0.049928	0.004535	0.129239	0.144310
736	17031827000	0.151467	0.063902	0.135085	0.313256	0.095163	0.171020	0.032916
1627	17163501200	0.070312	0.063867	0.254677	0.312333	0.092448	0.115769	0.054788
693	17031829302	0.037139	0.063092	0.088007	0.430489	0.022039	0.268917	0.042656
1726	17143000900	0.077778	0.061083	0.178600	0.744349	0.218391	0.167815	0.065122
1563	17113001600	0.021739	0.058991	0.280528	0.371287	0.058219	0.241117	0.030457
1635	17163501300	0.077371	0.057765	0.233577	0.293431	0.074875	0.127181	0.073490
1789	17019005401	0.042129	0.057171	0.098562	0.321973	0.026447	0.357302	0.120959
91	17097864601	0.090575	0.056720	0.062992	0.014732	0.000000	0.087100	0.148136
111	17097862902	0.022562	0.056642	0.175777	0.137363	0.134677	0.286054	0.033793
927	17031814100	0.110950	0.053933	0.042687	0.176250	0.078148	0.225212	0.034462
1550	17031301803	0.107776	0.053639	0.082818	0.283377	0.070032	0.223202	0.047046
1686	17089850103	0.020877	0.053545	0.048948	0.015948	0.000000	0.159678	0.072996
677	17031824300	0.065758	0.053380	0.275774	0.411116	0.056135	0.133753	0.027275
450	17031100200	0.104286	0.052246	0.064776	0.096269	0.057747	0.144995	0.118454
659	17031808100	0.060000	0.052017	0.259596	0.097643	0.044351	0.050366	0.071449
1231	17031813500	0.085326	0.051853	0.071649	0.114409	0.046072	0.203887	0.077350
370	17031300700	0.168345	0.050736	0.078373	0.326661	0.172260	0.161090	0.034646
157	17097865600	0.111699	0.050365	0.044102	0.058448	0.007643	0.072040	0.183666

Appendix 4 - Full List of Underperforming Tracts Recommended for Further Scrutiny

	tract_id	diff_actual_and_model_pred	commuting_ridership	model_pred_ridership	with_disability	below_poverty_level	carless	young_employed
1447	17031842900	-0.193355	0.204433	0.397788	0.208148	0.436466	0.344828	0.192593
1666	17163504600	-0.190295	0.074681	0.264976	0.162701	0.469286	0.250460	0.184792
526	17031283800	-0.166803	0.112139	0.278943	0.104022	0.217753	0.078014	0.221006
1256	17031081202	-0.155910	0.152174	0.308084	0.091833	0.042020	0.291667	0.117413
649	17031081800	-0.142579	0.140933	0.283512	0.027019	0.026212	0.261533	0.257353
1523	17031190402	-0.138739	0.056769	0.195507	0.062455	0.073223	0.064845	0.105155
1221	17031825802	-0.134582	0.021268	0.155850	0.081808	0.083752	0.033669	0.151995
1194	17031670700	-0.131570	0.080000	0.211570	0.085821	0.337687	0.056856	0.074324
1259	17031081403	-0.114872	0.163134	0.278006	0.028337	0.108253	0.399409	0.271897
1359	17031835700	-0.108303	0.000000	0.108303	0.000000	0.110577	0.000000	0.224390
1047	17031611300	-0.107220	0.079241	0.186461	0.072194	0.258257	0.058493	0.244944
648	17031081600	-0.101555	0.198465	0.300020	0.031250	0.079500	0.525950	0.252522
557	17031440300	-0.100477	0.080978	0.181456	0.189862	0.162552	0.047334	0.141777
716	17031380500	-0.096093	0.350168	0.446262	0.125866	0.497691	0.360269	0.293814
522	17031808002	-0.095320	0.023061	0.118381	0.117512	0.058652	0.000000	0.107653
957	17031150200	-0.092056	0.177363	0.269419	0.088809	0.130115	0.097641	0.141890
825	17031242200	-0.091153	0.309231	0.400384	0.031073	0.060950	0.144615	0.307872
722	17031411200	-0.090215	0.198701	0.288916	0.068343	0.199870	0.281984	0.114286
558	17031440600	-0.086919	0.082603	0.169522	0.131710	0.096846	0.026283	0.088933
597	17031230200	-0.084393	0.184080	0.268473	0.090740	0.207073	0.160804	0.207224
475	17031390300	-0.081297	0.276982	0.358279	0.135524	0.382638	0.177650	0.185970
933	17031320100	-0.078957	0.171629	0.250585	0.056325	0.075161	0.388730	0.172157
1390	17031241100	-0.078450	0.320559	0.399009	0.032671	0.234358	0.181637	0.358032
1058	17031190300	-0.074434	0.123039	0.197473	0.090464	0.167299	0.081312	0.158610
842	17031691400	-0.074101	0.270007	0.344108	0.070665	0.201115	0.337322	0.156912
657	17031807900	-0.073207	0.026696	0.099903	0.096989	0.065153	0.000000	0.071925
900	17031242900	-0.070968	0.226463	0.297431	0.080396	0.033813	0.115646	0.240781
1329	17031831600	-0.070368	0.134761	0.205129	0.118146	0.087371	0.070933	0.195105
1014	17031710400	-0.069143	0.199332	0.268475	0.170298	0.288720	0.104120	0.148718
457	17031180100	-0.068242	0.097147	0.165389	0.097836	0.185179	0.049695	0.158453