Problem

The issue this project addresses is socioeconomic mobility amongst minority communities across America.

Specifically attempting to answer the following questions:

- What attributes about a given community have the biggest impact on the socioeconomic mobility of its children?
- Given data on those attributes, can we predict the level of socioeconomic mobility of children from that community?
- What key visualizations can we create to portray this on a community-level in the US?

Conclusions will provide insight into geographical variation, trends amongst communities, and important factors to children rising out of poverty, thus informing potential targeted solutions for minority communities and future policy reform.

Attempts made previously

Namely two studies from Opportunity Insights (where we drew our data from), tackled the issue in different ways. They used traditional statistical analysis and regression to investigate the impact of community and race on economic mobility.

Studies:
- Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States
https://opportunityinsights.org/paper/land-of-opportunity/
- Race and Economic Opportunity in the United States: An Intergenerational Perspective
https://opportunityinsights.org/paper/race/

I applied Machine Learning models for prediction and feature engineering to provide additional insight, while keeping in mind explainability given the social importance of the subject. I also approached the problem with Classification methods instead to add interpretability: given a specific community or area of communities, what level of economic mobility can we predict the children to have (low, medium, high)? In addition, my project focuses on only minority communities, and I've added interactive visualizations to portray the geographical variations.

Data

The data is from Opportunity Insights (https://opportunityinsights.org/data/), a Harvard non-profit focused on the issue, which has a great library of data on socioeconomic and educational factors by geographic level across America.

Datasets:
- Neighborhood Characteristics by Commuting Zone ('CZ_neighborhoodcharacteristicsbycsv.csv')
- Geography of Mobility: Commuting Zone Characteristics - Definitions and Data Sources ('online_data_tables-8.xls')

Data Cleaning and Pre-processing

The first dataset is a CSV of neighborhood characteristics, from which we grab certain racial share data and apply a filter for just the minority communities, grouping by Commuting Zone, a unique numeric identifier for communities ranging across the entire United States.

Our second dataset is an XLS file from which we import the two sheets: Online Data Table 5 and Online Data Table 8. We filter and clean the sheets for the tables and features we care about, which span a range of educational, social, economic, and community attributes. Examples include - racial segregation, commuting times, fraction middle class, local tax rates, student teacher ratio, school expenditure per student, teenage labor force participation rate, violent crime rate, fraction of children with single mothers, etc.

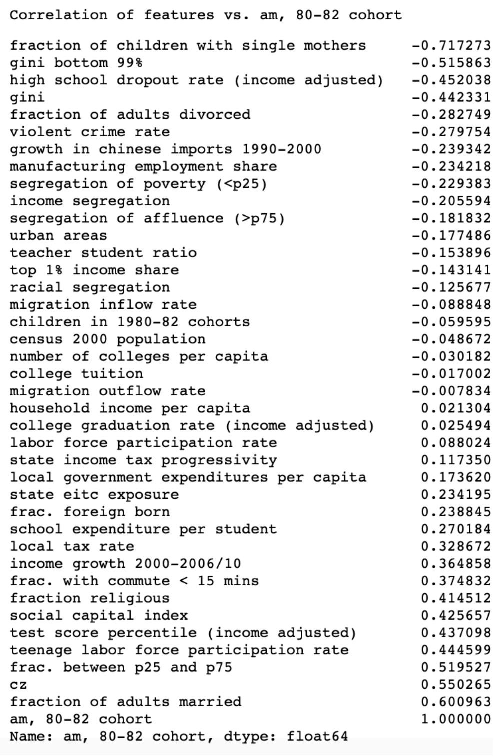
Finally we merge the two datasets on Commuting Zone, resulting with a dataset of 40 features and 500 entries.

Target Variable

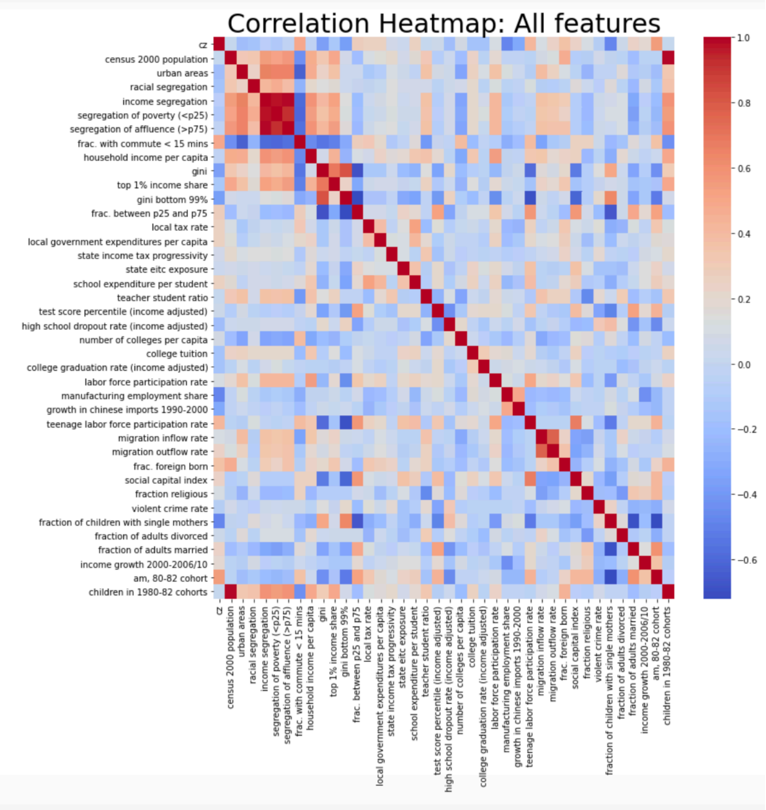
The target variable is the metric we use to measure socioeconomic mobility, deemed “Absolute Upward Mobility”, engineered from the paper (**https://opportunityinsights.org/paper/land-of-opportunity/**). It is the mean rank (in the national child income distribution) of children whose parents are at the 25th percentile of the national parent income distribution. The paper goes into great comprehensive detail of this ranking method, as well as data sources used such as Census Data and IRS tax filings, and adjustments for robustness of the metric.

EDA

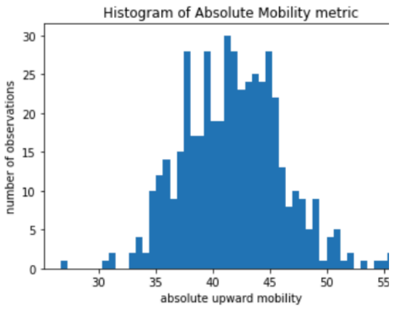
Correlation tables and heat maps are printed for all features vs. the target variable (labeled ‘am, 80-82 cohort’). Immediately we see the biggest positive and negative correlations, including features such as fraction of children with single mothers, racial shares, high school dropout rate, fraction of adults married, fraction of middle class families, teenage labor force participation rate, etc.

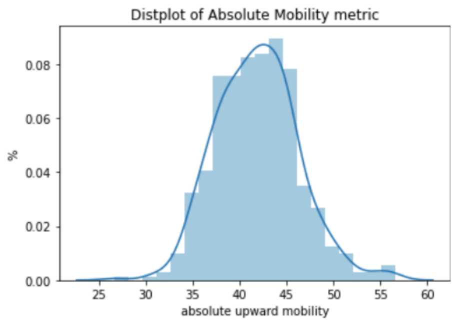


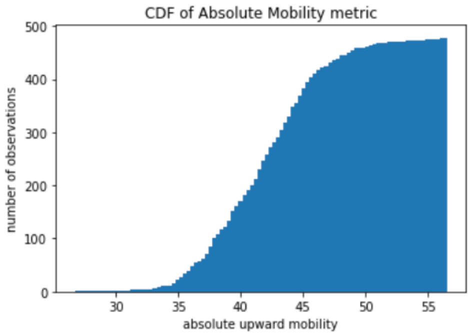
```
Correlation of features vs. am, 80-82 cohort

fraction of children with single mothers      -0.717273
gini bottom 99%                               -0.515863
high school dropout rate (income adjusted)    -0.452038
gini                                          -0.442331
fraction of adults divorced                   -0.282749
violent crime rate                            -0.279754
growth in chinese imports 1990-2000           -0.239342
manufacturing employment share                -0.234218
segregation of poverty (<p25)                 -0.229383
income segregation                            -0.205594
segregation of affluence (>p75)               -0.181832
urban areas                                   -0.177486
teacher student ratio                         -0.153896
top 1% income share                           -0.143141
racial segregation                            -0.125677
migration inflow rate                         -0.088848
children in 1980-82 cohorts                   -0.059595
census 2000 population                        -0.048672
number of colleges per capita                 -0.030182
college tuition                               -0.017002
migration outflow rate                        -0.007834
household income per capita                    0.021304
college graduation rate (income adjusted)      0.025494
labor force participation rate                 0.088024
state income tax progressivity                 0.117350
local government expenditures per capita       0.173620
state eitc exposure                            0.234195
frac. foreign born                             0.238845
school expenditure per student                 0.270184
local tax rate                                 0.328672
income growth 2000-2006/10                     0.364858
frac. with commute < 15 mins                   0.374832
fraction religious                             0.414512
social capital index                           0.425657
test score percentile (income adjusted)        0.437098
teenage labor force participation rate         0.444599
frac. between p25 and p75                      0.519527
cz                                             0.550265
fraction of adults married                     0.600963
am, 80-82 cohort                               1.000000
Name: am, 80-82 cohort, dtype: float64
```



We also investigate the distribution of the target variable with visualizations, noting a relatively normal distribution.



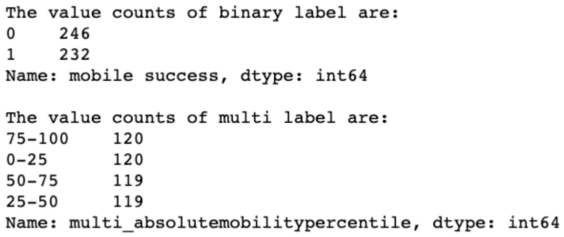




We create target variable labels from the Absolute Upward Mobility metric for both Binary and Multi-label Classification.

For binary classification, 'am, 80-82 cohort’ is split in half by its numeric mean for labels 1 and 0, success being 1 and failure being 0, representing good or bad mobility. For multi-label classification, 'am, 80-82 cohort’ is split into quartiles 0-25%, 25-50%, 50-75%, and 75-100% - respectively representing low, medium, high, and excellent mobility.

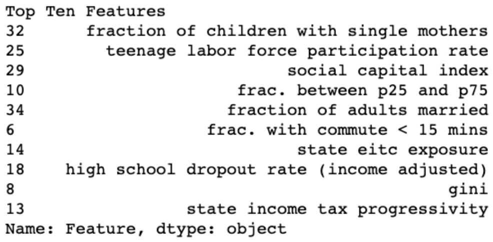
Note that Classification should not suffer from imbalanced classes given the distribution and engineering of the target variable labels.



```
The value counts of binary label are:
0    246
1    232
Name: mobile success, dtype: int64

The value counts of multi label are:
75-100    120
0-25      120
50-75     119
25-50     119
Name: multi_absolutemobilitypercentile, dtype: int64
```

Feature Selection

I create a function for Mutual Information Classification to create feature rankings for binary and multi-label Classification and print top features. These results are consistent with the correlation EDA from before.



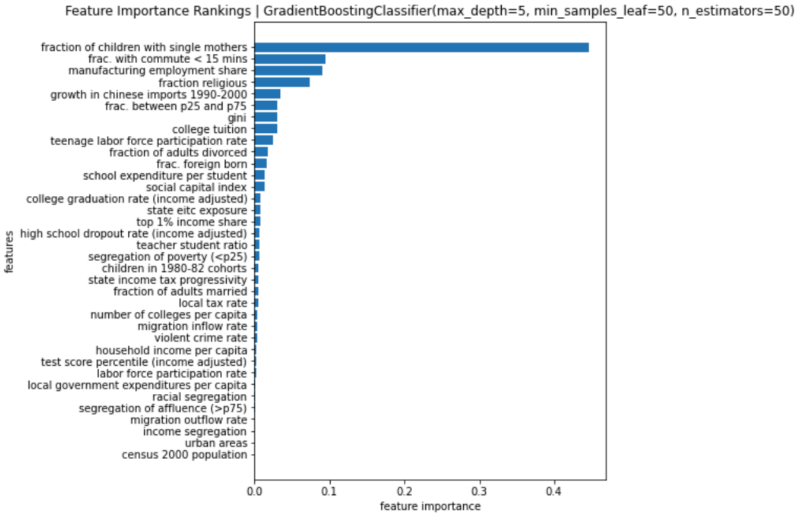
```
Top Ten Features
32     fraction of children with single mothers
25       teenage labor force participation rate
29                          social capital index
10                      frac. between p25 and p75
34                     fraction of adults married
6                     frac. with commute < 15 mins
14                           state eitc exposure
18    high school dropout rate (income adjusted)
8                                           gini
13                 state income tax progressivity
Name: Feature, dtype: object
```

Model Selection and Results

Binary Classification

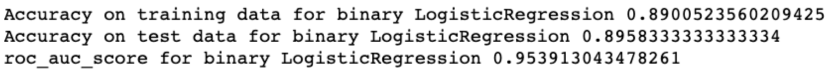
Started with a simple LogisticRegression model which showed poor performance. Ensemble methods RandomForestClassifier and GradientBoostingClassifier showed extremely high scores on training data (~95%) but much lower on test data (~80%), signaling overfitting. These models are likely suffering from over-cardinality and multi-collinearity, for example the features ‘fraction of adults married’ and ‘fraction of children with single mothers’ are naturally closely related. Also, given the modest size of the dataset, complicated models are prone to over-fitting.

Here is an example of the Feature Importances from the GradientBoostingClassifier. The highest one is again fraction of children with single mothers, but other important features not identified before include manufacturing employment share, fraction religious, growth in Chinese imports. Interesting insight.



Given the need for regularization, I scaled the data and employed LogisticRegressionCV with elastic-net regularization, 5-fold cross-validation, and hyper-parameter tuning ranges of Cs = np.logspace(-10,10,50) and L1 ratios = np.arange(0,1,.05).

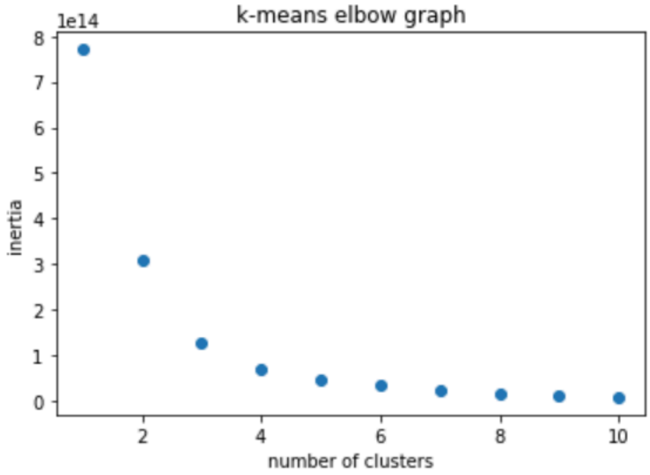
Training and test set scores both drastically improved to ~90%, with roc_auc_score consistently above 90% as well.



```
Accuracy on training data for binary LogisticRegression 0.8900523560209425
Accuracy on test data for binary LogisticRegression 0.8958333333333334
roc_auc_score for binary LogisticRegression 0.953913043478261
```

K-Means Clustering

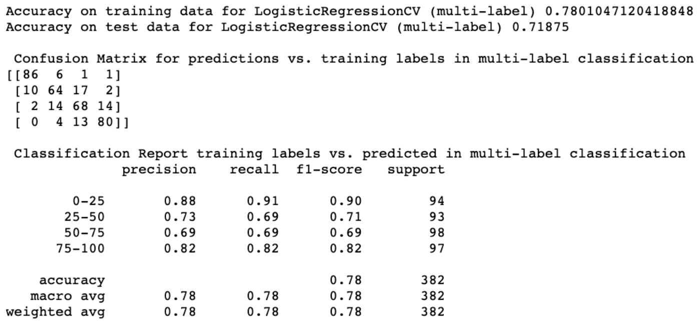
Although not originally a clustering problem, I investigated the data with the K-Means Clustering method to find the optimal number of clusters to be around 4. This is theoretically consistent with our splitting of target variable labels into 4 groups for multi-label classification. I also appended cluster labels to the dataset as a feature in multi-label classification.



Multi-Label Classification

Using the same hyper-parameter tuning and regularization with 10-fold cross-validation, LogisticRegression is giving accuracy scores of ~80% on training and ~70% on test data.

Gradient Boosting overfit again.

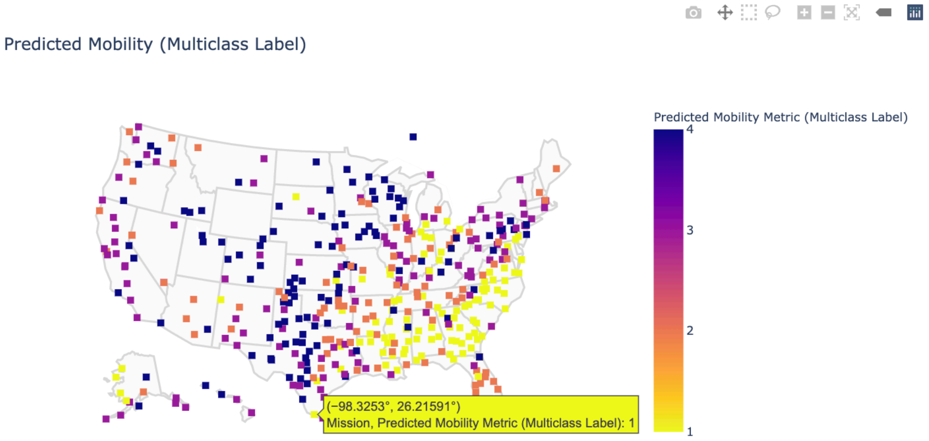


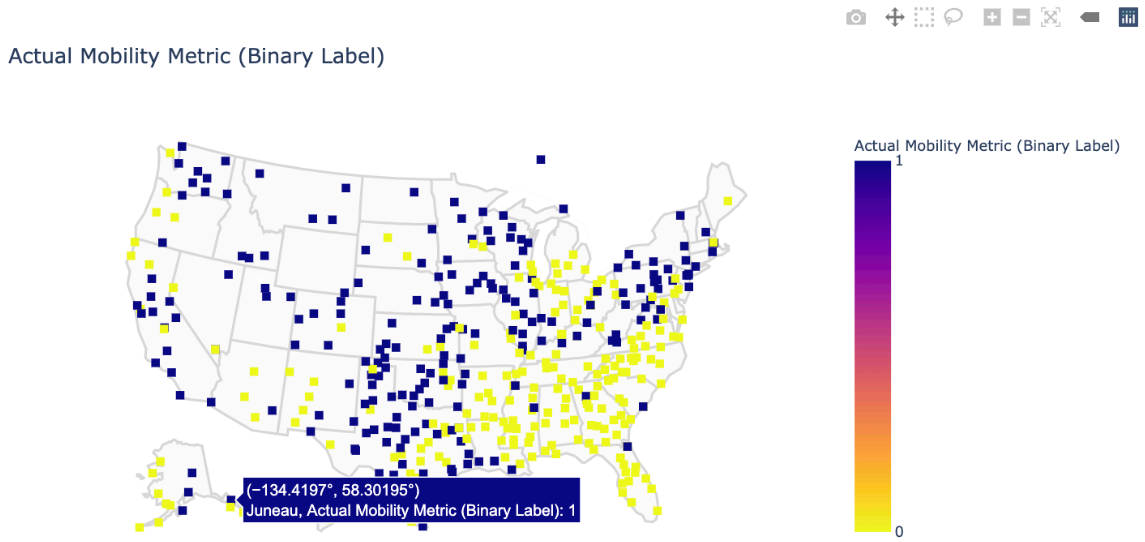
```
Accuracy on training data for LogisticRegressionCV (multi-label) 0.7801047120418848
Accuracy on test data for LogisticRegressionCV (multi-label) 0.71875

 Confusion Matrix for predictions vs. training labels in multi-label classification
[[86  6  1  1]
 [10 64 17  2]
 [ 2 14 68 14]
 [ 0  4 13 80]]

 Classification Report training labels vs. predicted in multi-label classification
              precision    recall  f1-score   support

        0-25       0.88      0.91      0.90        94
       25-50       0.73      0.69      0.71        93
       50-75       0.69      0.69      0.69        98
      75-100       0.82      0.82      0.82        97

    accuracy                           0.78       382
   macro avg       0.78      0.78      0.78       382
weighted avg       0.78      0.78      0.78       382
```

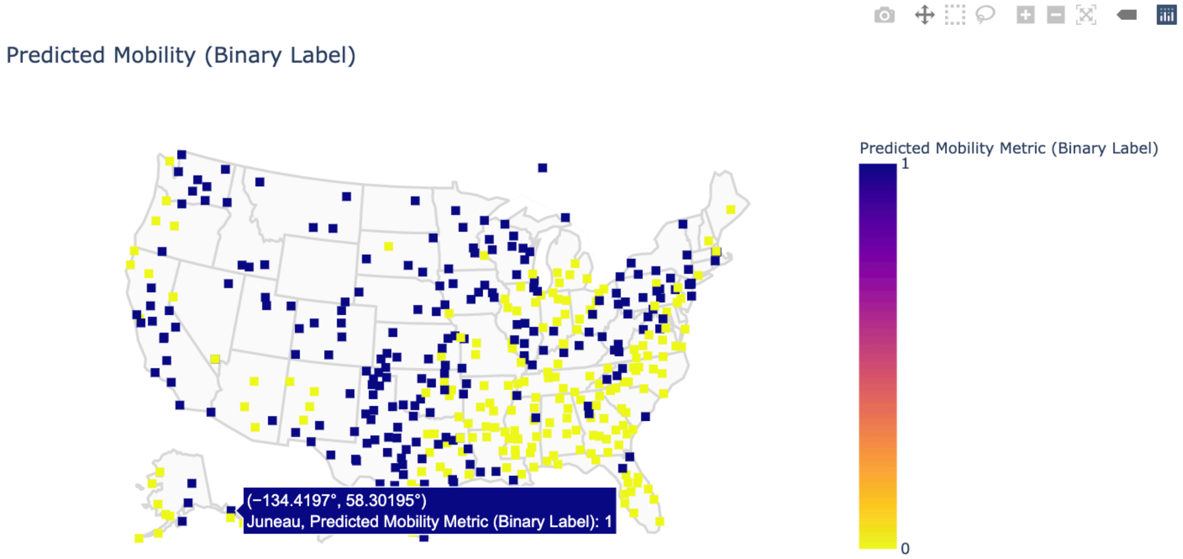
Visualizations

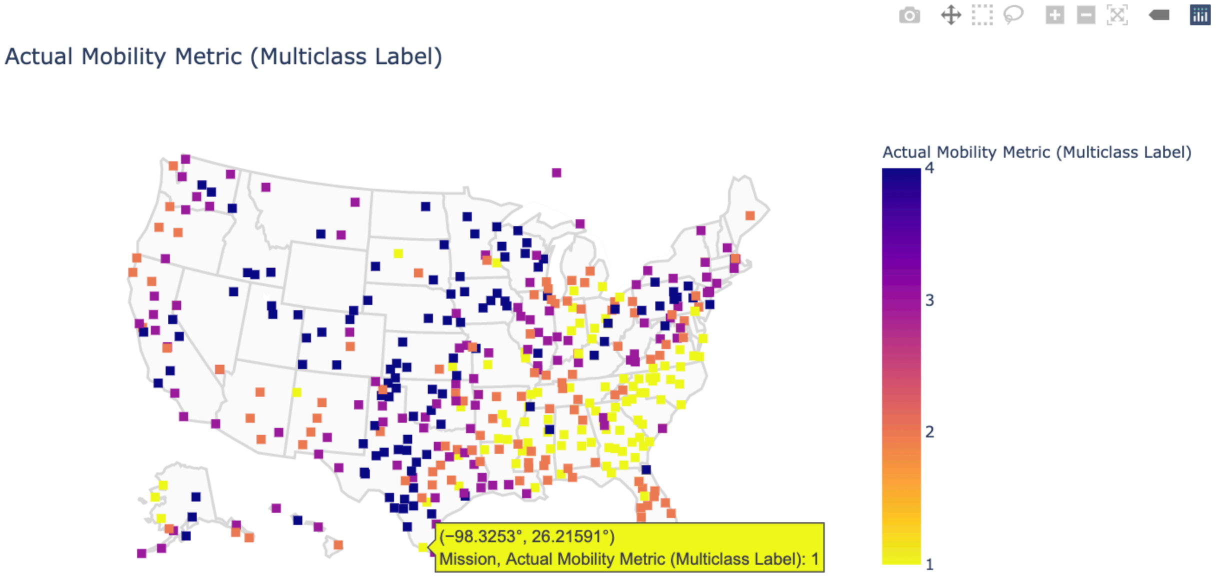
Utilizing Plotly for interactive graphical visualizations, displayed the results for both Binary and Multi-label Classification. *<Link to Plotly visualizations here>.*

Hover over any city to view its Actual vs. Predicted mobility label. Note the higher accuracy of the model, the more identical the map colors will be.


Predicted Mobility (Multiclass Label)
Predicted Mobility Metric (Multiclass Label)
4
3
2
1
(−98.3253°, 26.21591°)
Mission, Predicted Mobility Metric (Multiclass Label): 1


Actual Mobility Metric (Binary Label)
Actual Mobility Metric (Binary Label)
1
0
(−134.4197°, 58.30195°)
Juneau, Actual Mobility Metric (Binary Label): 1


Predicted Mobility (Binary Label)
Predicted Mobility Metric (Binary Label)
1
0
(−134.4197°, 58.30195°)
Juneau, Predicted Mobility Metric (Binary Label): 1


Actual Mobility Metric (Multiclass Label)
Actual Mobility Metric (Multiclass Label)
4
3
2
1
(−98.3253°, 26.21591°)
Mission, Actual Mobility Metric (Multiclass Label): 1

Conclusions

Our hypothesis is confirmed that using data on community-level attributes, we can predict the level of future socioeconomic mobility of children who grow up in that community. This shows not only that there are geographical variations in the likelihood of the success of children, but also the community-level features which are most important in determining this.

We can identify who is disadvantaged or advantaged, why, and hopefully how to help more children rise up. Once able to identify the factors helping or preventing children's success in rising out of poverty, we can start to use this information to inform social policy, community activism, education reform, and targeted solutions for communities across the country.