Analyzing Socioeconomic and Housing Factors to Predict Arrest Trends in New York State

AUTHOR
Joyce Yan (qy249), Lydia Lin (dl2253), Victoria Xiao (sx287)

PUBLISHED
December 16, 2024

```
<Training/Testing/Total>
<440/112/552>
```

```
#  5-fold cross-validation using stratification
# A tibble: 5 × 2
  splits           id
  <list>           <chr>
1 <split [352/88]> Fold1
2 <split [352/88]> Fold2
3 <split [352/88]> Fold3
4 <split [352/88]> Fold4
5 <split [352/88]> Fold5
```

```
# A tibble: 3 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard     459.
2 rsq     standard       0.999
3 mae     standard     224.
```

```
# A tibble: 3 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard     771.
2 rsq     standard       0.998
3 mae     standard     351.
```

A model card provides brief, transparent, responsible reporting for a trained machine learning model.

Model details

-Developed by TEAM Skillful-Wombat.

-Model Description: A Poisson regression model predicting Total arrests using year and socioeconomic factors.

-The model was trained on county-level data from 2015 to 2022, with 80% of the data used for training and 20% for testing. The dataset includes 21 features such as unemployment rates, poverty levels, housing characteristics, and year.

-Feature Engineering and Preprocessing: Detailed feature transformations were applied to enhance predictive power, such as:Interactions between poverty levels and unemployment rates. Logarithmic transformations of housing related features. One-hot encoding of categorical predictors, including Year.The dataset was standardized and stratified during resampling to ensure balanced splits.

-Version: 20241216T193757Z-4c1b8

-Publication Date: 12/16/2024

-Citation:If you use this model, please cite:“Poisson Regression Model for Arrest Predictions in New York State, developed by Skillful-Wombat. Published on December 16, 2024.”

-If you have questions about this model, please contact sx287@cornell.edu.

Intended use

-Primary Intended Uses: Predict total adult arrests at the county-year level. Support policymakers in resource allocation and crime prevention strategies. Analyze the relationship between socioeconomic factors and arrest trends.

-Primary Intended Users: Local government agencies. Policymakers and resource planners. Researchers in sociology, criminology, and economics.

-Out-of-Scope Uses: Individual level arrest predictions. Real time crime detection or intervention. Legal decisions, such as sentencing or bail determination.

Important aspects/factors

-Relevant Aspects:

Demographic factors: Population density, language barriers, and education levels. Economic conditions: Unemployment rates, poverty levels, and household income. Housing characteristics: Housing affordability and ownership metrics.

-Model Evaluation Focus:

Calibration of predictions at the county level. Observed vs. predicted trends across socioeconomic features.

Metrics

-Metrics Used:

RMSE (Root Mean Square Error): Measures prediction error magnitude. R^2 (Coefficient of Determination): Evaluates model fit to the data. MAE (Mean Absolute Error): Provides average error magnitude.

-How Metrics Were Computed:

Metrics were evaluated using the yardstick package in the tidymodels framework. 5-fold cross-validation was used to validate the model on training data. Final metrics were computed on an 80/20 train-test split.

-Why These Metrics:

RMSE is sensitive to large errors, making it a useful diagnostic tool. R² explains how well the model captures the variability in arrest counts. MAE offers a more interpretable average error magnitude.

Training data & evaluation data

-The model was trained on socioeconomic and arrest data for 62 New York State counties from 2015 to 2022.

-The dataset was stratified by Total arrests to ensure balanced train test splits.

```
Rows: 0
Columns: 21
$ Mortgage_affordability_index     <dbl>
$ Log_housing_units                <dbl>
$ Log_mortgage_ratio               <dbl>
$ Language_education_interaction   <dbl>
$ Unemployment_poverty_interaction <dbl>
$ Unemployment_rate                <dbl>
$ Median_household_income          <dbl>
$ Below_poverty_level              <dbl>
$ No_health_insurance              <dbl>
$ `Felony Total`                   <dbl>
$ `Drug Felony`                    <dbl>
$ `Violent Felony`                 <dbl>
$ `DWI Felony`                     <dbl>
$ `Other Felony`                   <dbl>
$ `Misdemeanor Total`              <dbl>
$ `Drug Misdemeanor`               <dbl>
$ `DWI Misdemeanor`                <dbl>
$ `Property Misdemeanor`           <dbl>
$ `Other Misdemeanor`              <dbl>
$ County                           <chr>
$ Year                             <dbl>
```

-The test dataset consists of 20% of the data, stratified by the target variable.

-We chose an **80/20 training-test split** stratified by the target variable, Total (total arrest counts), to ensure both subsets reflect the distribution of arrest counts.

Name	test_data
Number of rows	112
Number of columns	22

________________________	
Column type frequency:	
character	1
numeric	21
________________________	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
County	0	1	5	12	0	53	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25
Mortgage_affordability_index	0	1	1.37	0.41	0.65	1.03
Log_housing_units	0	1	11.46	1.72	8.98	10.26
Log_mortgage_ratio	0	1	0.00	0.00	0.00	0.00
Language_education_interaction	0	1	17999948.89	91300258.02	197.40	7158.90
Unemployment_poverty_interaction	0	1	80.14	44.39	18.88	52.73
Unemployment_rate	0	1	6.28	1.63	3.20	5.30
Median_household_income	0	1	64579.25	16024.13	34299.00	52018.75
Below_poverty_level	0	1	12.28	3.56	5.30	10.28
No_health_insurance	0	1	5.83	2.07	2.30	4.20
Felony Total	0	1	3605.36	5834.84	8.00	282.25
Drug Felony	0	1	524.62	919.62	1.00	49.00
Violent Felony	0	1	1021.54	1811.98	1.00	48.75
DWI Felony	0	1	81.90	97.35	0.00	27.75
Other Felony	0	1	1977.29	3236.24	5.00	156.50
Misdemeanor Total	0	1	6943.37	11759.18	26.00	714.00
Drug Misdemeanor	0	1	1314.79	2592.38	1.00	80.50
DWI Misdemeanor	0	1	516.16	571.49	5.00	146.25
Property Misdemeanor	0	1	2551.28	4986.29	2.00	175.75
Other Misdemeanor	0	1	2561.14	4169.22	12.00	246.00
Year	0	1	2018.45	2.33	2015.00	2016.00
Total	0	1	10548.72	17331.43	34.00	1045.50

Quantitative analyses

```
# A tibble: 3 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard     459.
2 rsq     standard       0.999
3 mae     standard     224.
```

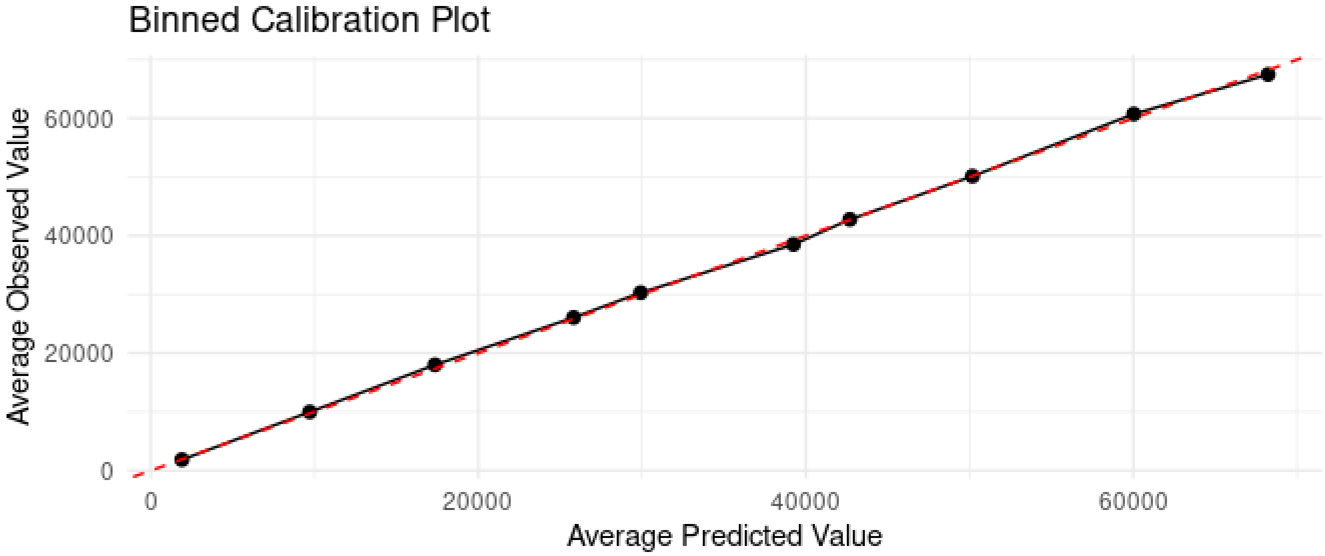
Overall model performance

```
# A tibble: 3 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard     459.
2 rsq     standard       0.999
3 mae     standard     224.
```

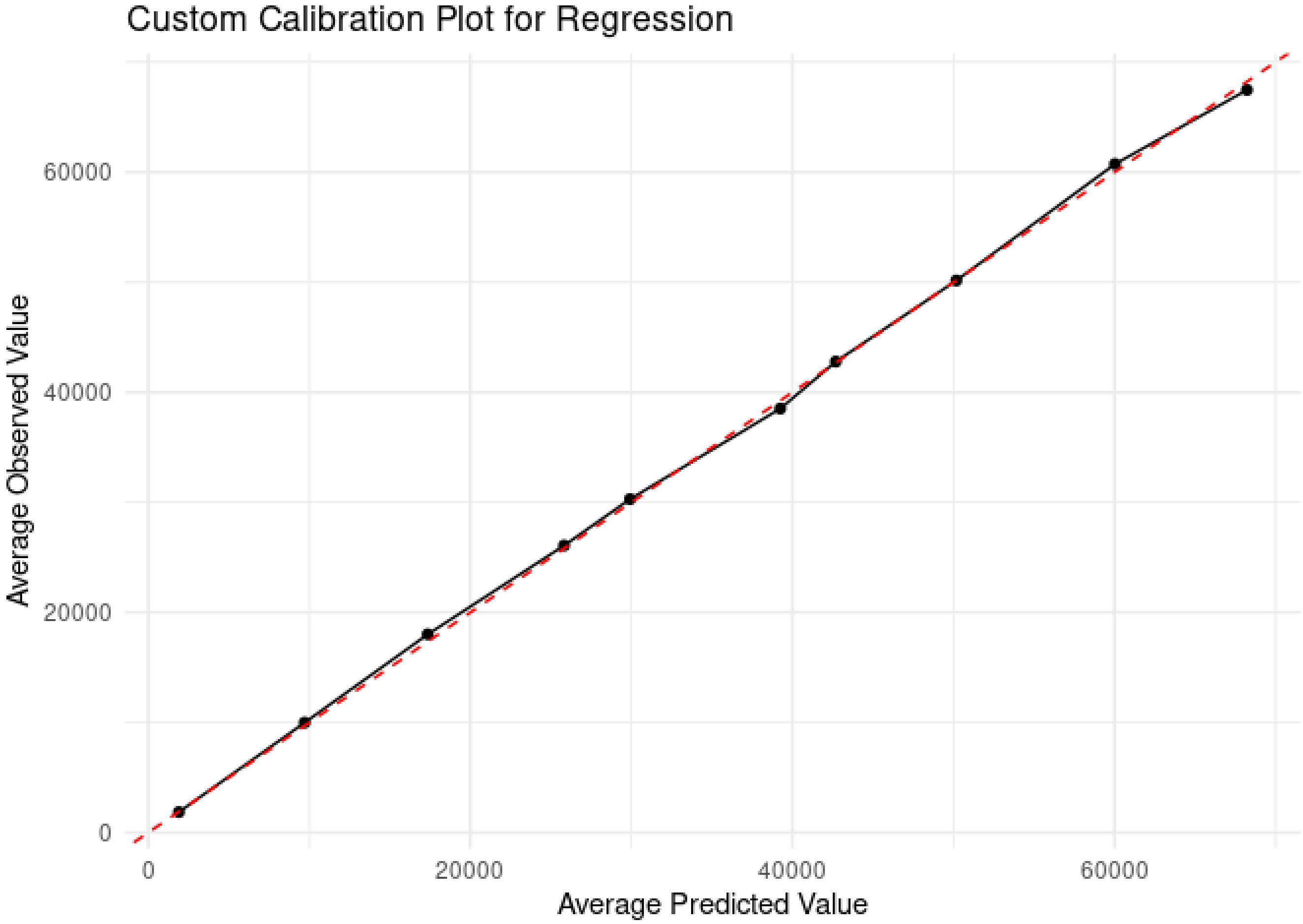
Disaggregated model performance

```
# A tibble: 159 × 4
   County      .metric .estimator .estimate
   <chr>       <chr>   <chr>          <dbl>
 1 Albany      rmse    standard       184.
 2 Allegany    rmse    standard        20.5
 3 Bronx       rmse    standard      1674.
 4 Broome      rmse    standard       286.
 5 Cattaraugus rmse    standard       157.
 6 Cayuga      rmse    standard        21.6
 7 Chautauqua  rmse    standard       136.
 8 Chenango    rmse    standard        93.3
 9 Clinton     rmse    standard       312.
10 Columbia    rmse    standard       121.
# ℹ 149 more rows
```

Visualize model performance



Make a custom plot



Ethical considerations

-Bias in Data: Arrest records may reflect systemic biases (e.g.,over-policing in disadvantaged areas) rather than actual crime levels.

-Resource Allocation: Predictions must be used cautiously to avoid reinforcing inequities in law enforcement resource distribution.

-Transparency:Clear explanations should accompany model predictions toensure ethical use and avoid misinterpretation.

Caveats & recommendations

-What the Model Does: Predict total adult arrests based on socioeconomic conditions at the county level. Identify key drivers of arrest trends, such as unemployment and poverty rates.

-What the Model Does Not Do: Predict individual-level arrests. Establish causal relationships between features and arrests.

-Recommendations: Combine predictions with qualitative insights to guide decisions. Regularly update the model with new data to ensure relevance and accuracy. Use results as part of broader policy discussions, not as definitive outcomes.