Using Foursquare data to predict life expectancies

Your Neighborhood Is Killing You

AARON CHAICLIN MAY 2020

How do neighborhoods affect our life?

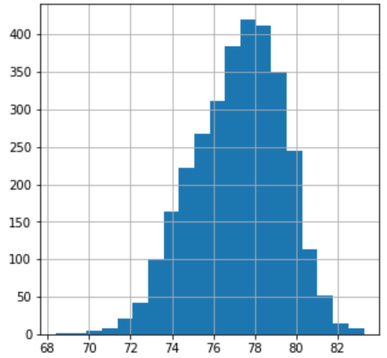
- Research by the University of Washington has pinpointed life expectancies at the county level
- These US life expectancies vary from 68.4 to 83.3 years old
- Why are there regional differences in this health metric?
- In order to improve life expectancies, we need to understand these regional differences
- Can we determine how our neighborhoods may affect these regional variations?

Data collection & cleaning

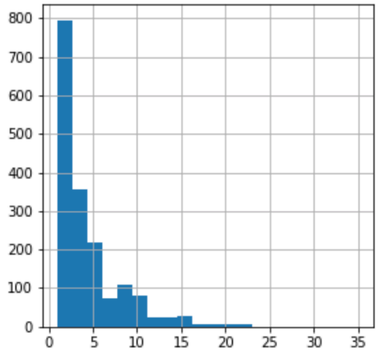
- Life Expectancies by US County from the University of Washington for 3,142
- County shapes & area (GeoJSON) from the US Census for 3,142 counties
- Foursquare data for 199,511 venues throughout 3,126 counties
- Created 1,385 different features

Feature Histograms

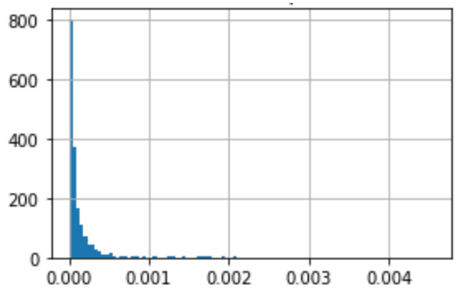
AVERAGE LIFE EXPECTANCY



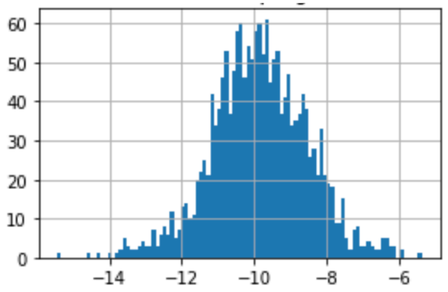
ORIGINAL BAR DISTRIBUTION



BARS PER CAPITA

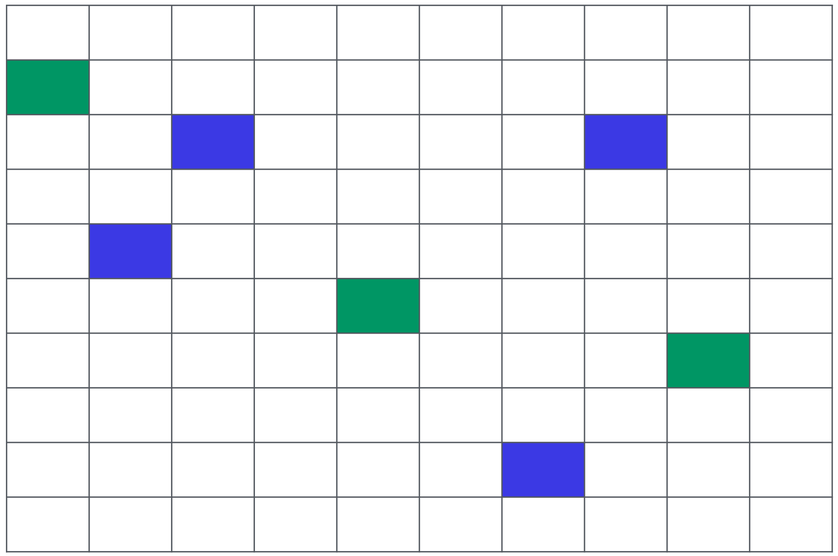


LOG OF BARS PER CAPITA



- Life Expectancy has a skewed normal distribution
- Foursquare data did not match the distribution pattern
- Most features were limited to a value of 1 spread across various counties
- Created per capita and per square mile metrics of Foursquare data
- Created log features to best match the target data

Sparse Matrix



WHITE = NO DATA
BLUE = VALUE OF 1
GREEN = VALUE > 1

- The Foursquare venue API produced a sparse matrix of data points
- 93% of elements in the matrix were blank
- ~4% had a value of 1
- ~3% had a value > 1

Feature Development

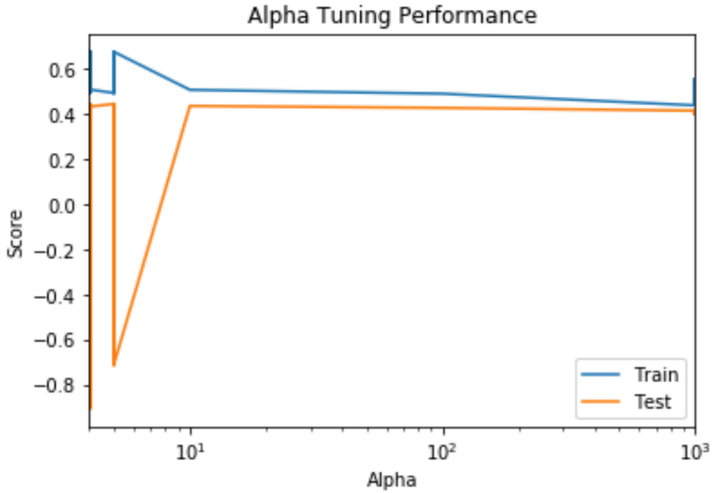
TOP FREQUENCY • CORRELATION FEATURES

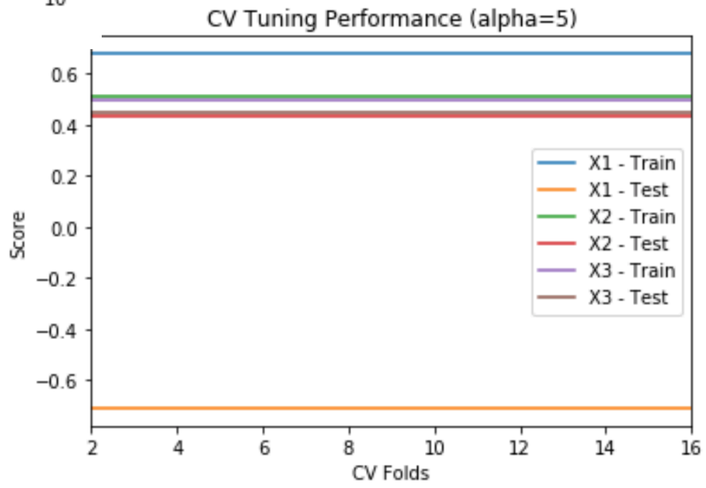
	freq.	corr	corr-abs
fastfoodPerCapLog	0.92	-0.37	0.37
fastfoodPerCap	0.92	-0.27	0.27
Park	0.73	0.32	0.32
doctorsofficePerCapLog	0.79	-0.25	0.25
bbqPerCapLog	0.52	-0.39	0.39
totalPop	1.00	0.20	0.20
medicalPerCapLog	0.76	-0.23	0.23
Trail	0.47	0.36	0.36
churchPerCapLog	0.89	-0.18	0.18
Pizza	0.90	0.18	0.18

- Created a feature subset using recursive feature elimination (RFE)
- Created feature subset of top correlated & relatively high frequency items
- Reduce sparsity and data noise
- Used 3 Feature sets for model training
  - Full
  - RFE
  - FreqXCorr

Regression Models

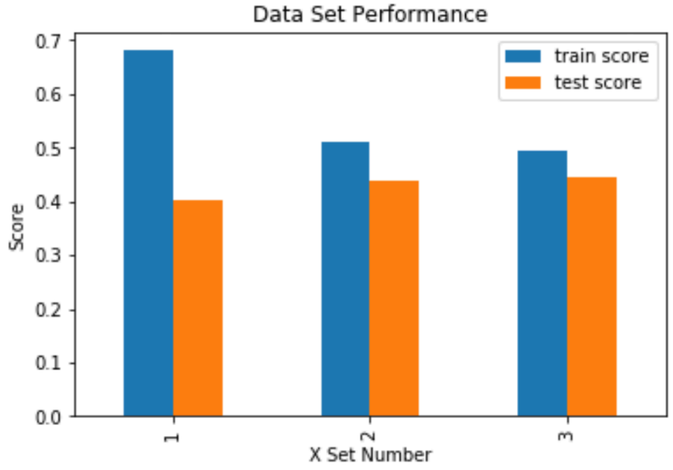
RidgeCV Tuning

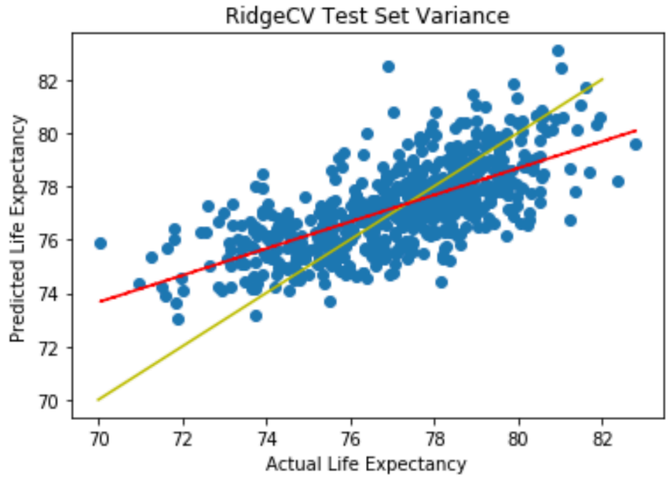




- Iterated to optimize hyperparameters
- Found two optimal alpha values
- CV tuning appeared to make no difference
- Iterating though “solvers” did not make a difference, used default of “auto”

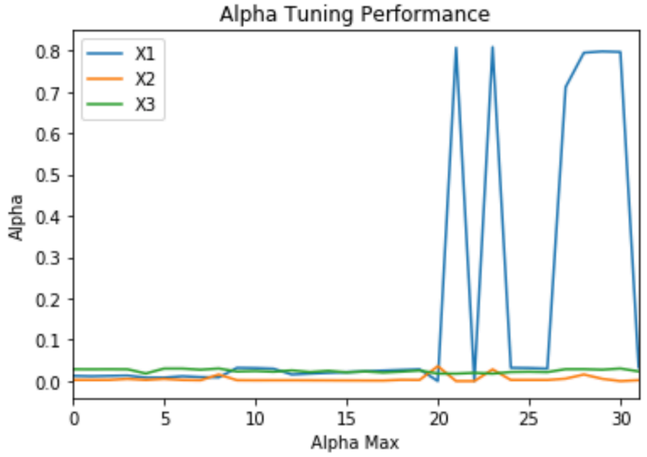
RidgeCV Results

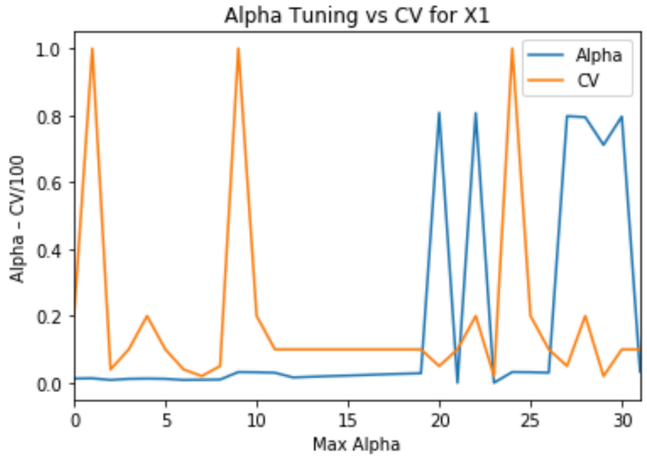




- CorrXFreq data (X3) had the best performance on test sets, but not by much
- RFE outperformed the full feature set
- Top validation score of .45 was not good enough

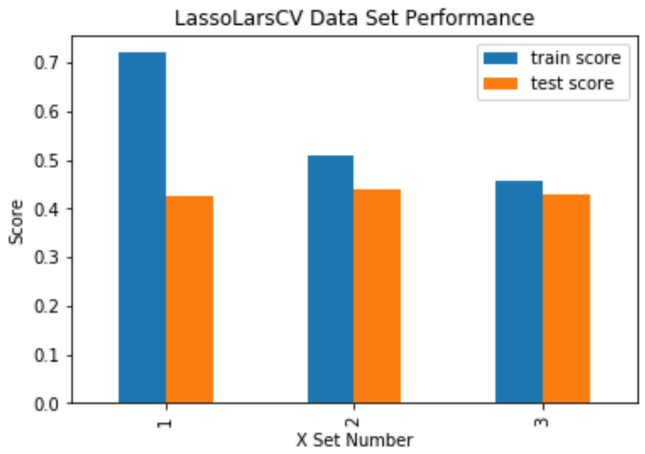
LassoLarsCV

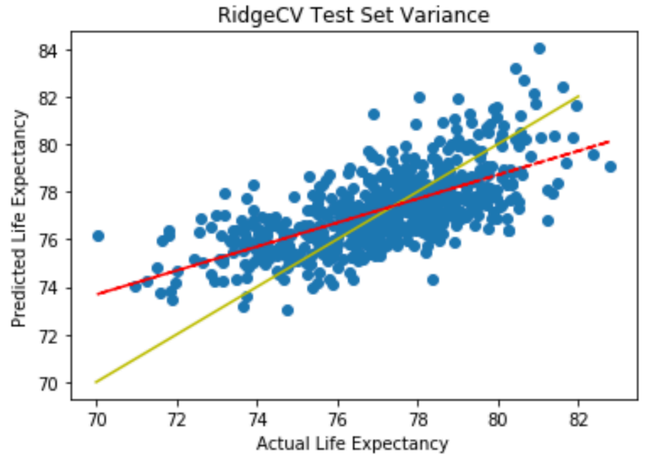




- Tuning was not straightforward
- Could not find correlation between Max Alpha setting and actual alpha
- Alpha seemed more somewhat dependent on a combination of the Max Alpha and the number of cross validation folds

LassoLarsCV

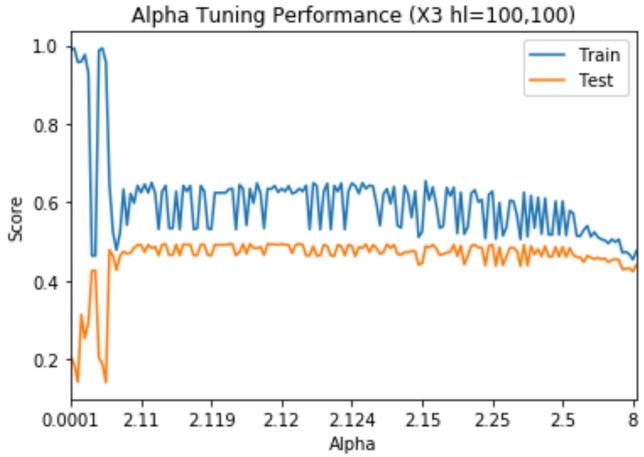


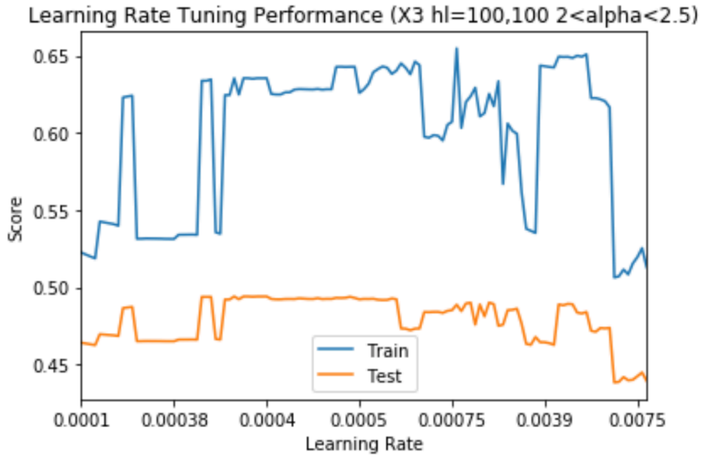


Yellow line shows perfect prediction.
Red line shows trend of our prediction.

- Could not beat RidgeCV
- Variance Score topped out at 0.44
- LassoLars performed best on the RFE feature set (X2)

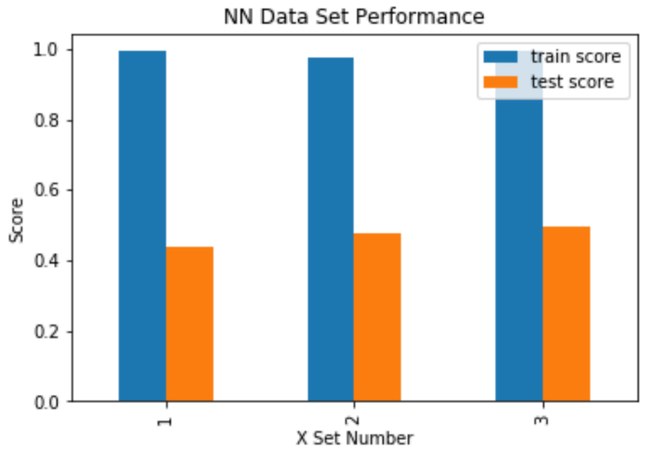
Neural Network

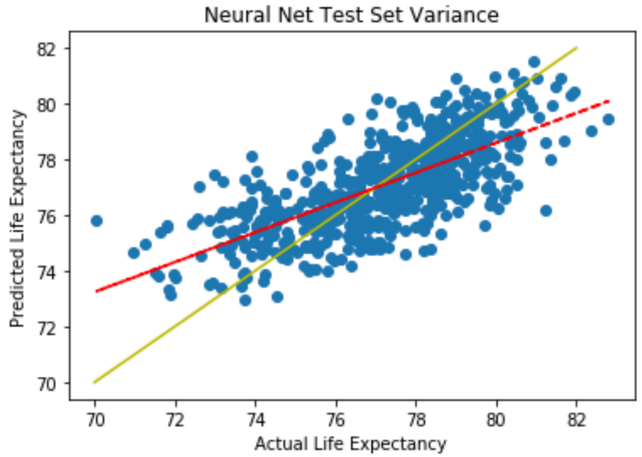




- Optimal hidden layers landed two hidden layers with 100 nodes in each
- Tuning seemed a little chaotic
- Alpha parameter had a “good range” of 2.11 - 2.13
- Initial Learning Rate was tuned around 0.0004

Neural Network



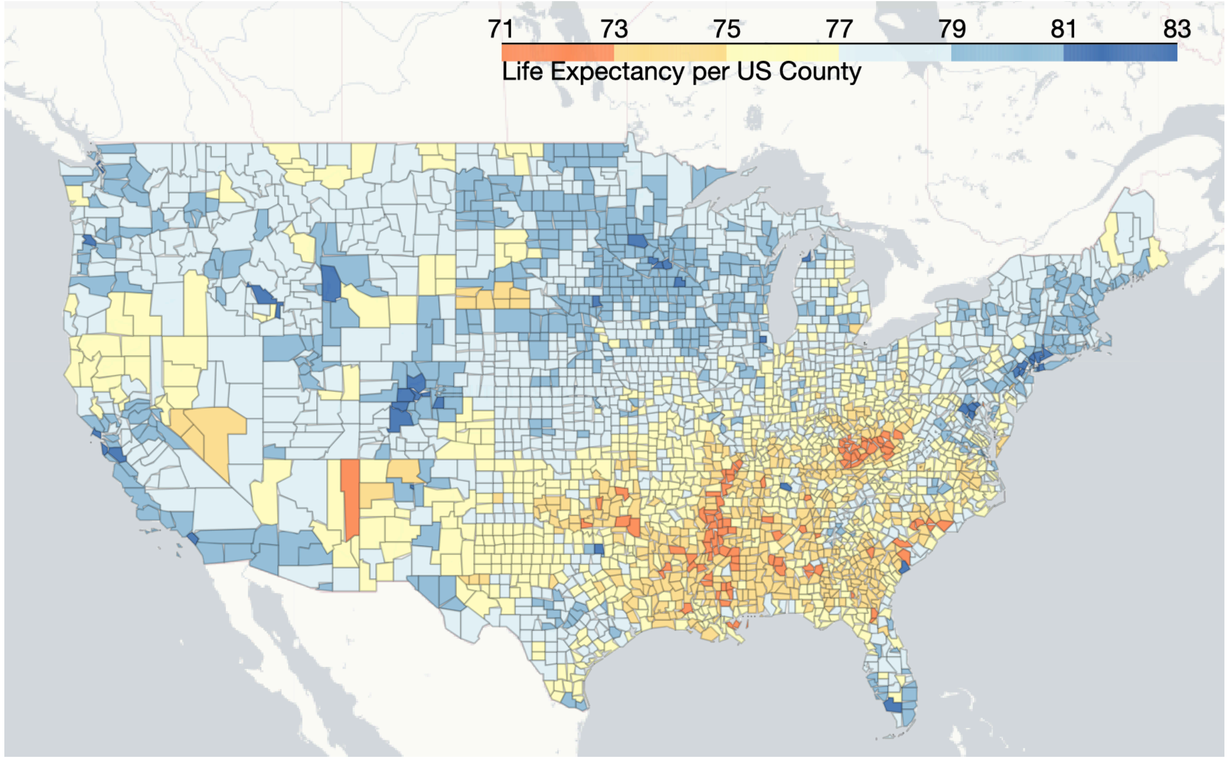


Yellow line shows perfect prediction.
Red line shows trend of our prediction.

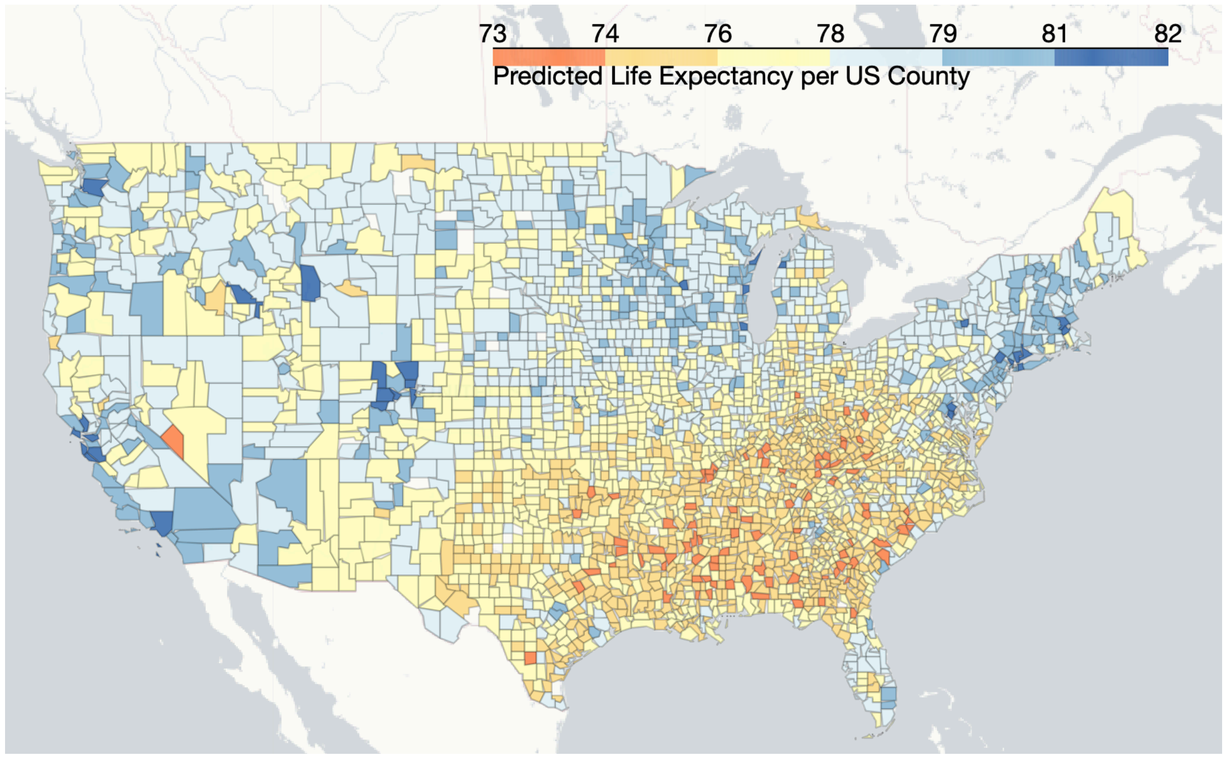
- Performed best on the CorrXFreq (X3) dataset
- Neural Network has the best score of 0.49 variance

Actual vs Prediction Map

Actual



Prediction



Conclusion

- Did not hit target threshold for a successful prediction
- Given API & time constraints model performed admirably
- Additional data and processing power likely to achieve a better result