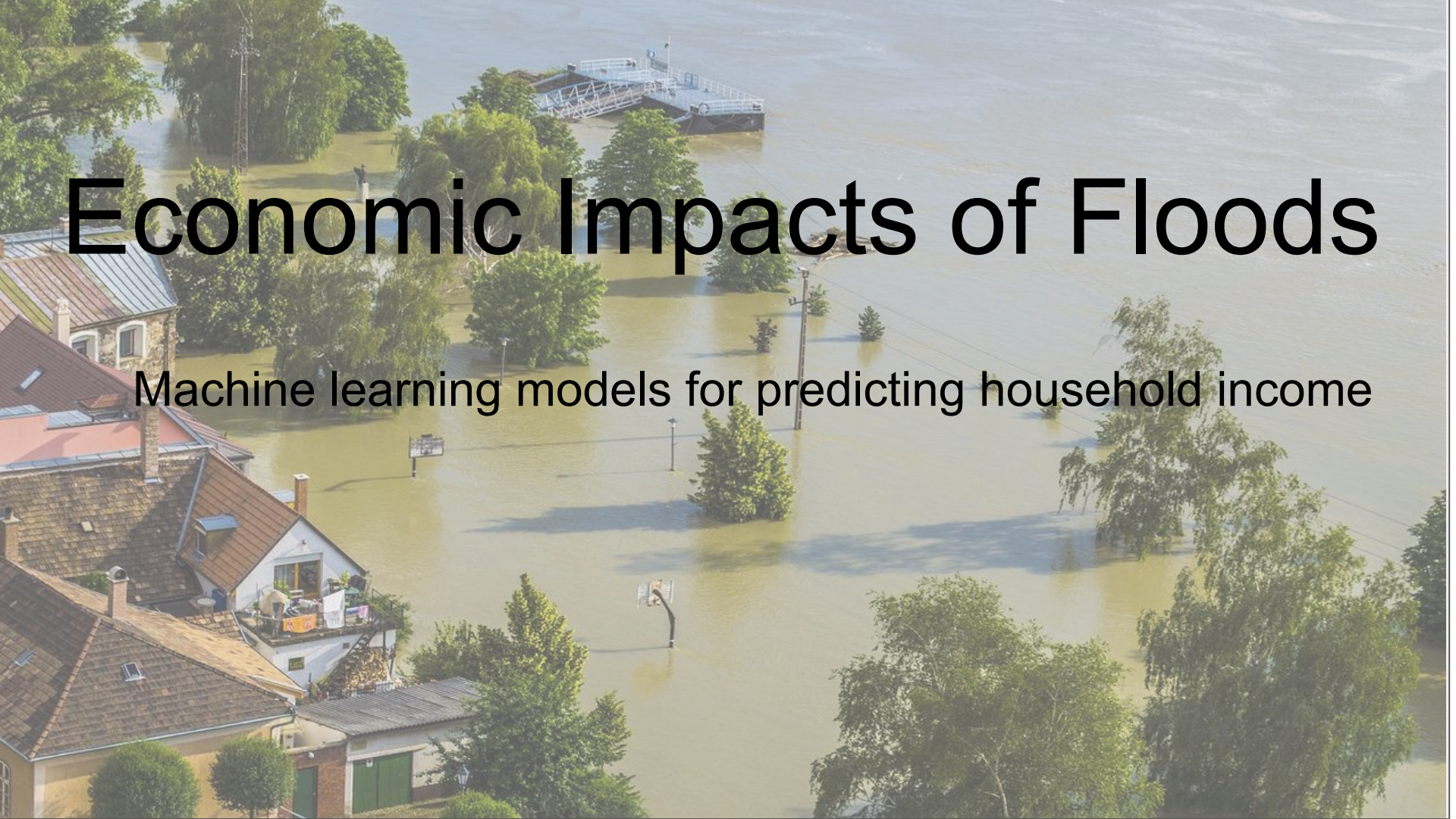


Economic Impacts of Floods

Machine learning models for predicting household income

Economic Impact of Flooding on Vulnerable Communities

Floods are becoming more frequent and more catastrophic with changing climate. Are they hitting our more vulnerable communities harder? Does this reinforce cycles of under-resourced communities?

ML translation: Can we predict median income using flood occurrence?

Data Sources

NOAA storm event database

(https://www.ncdc.noaa.gov/stormevents/ftp.jsp)

Features: flood type, date, location, duration, property damage, human injury/death

US Census Data

Features: median income, public assistance, state

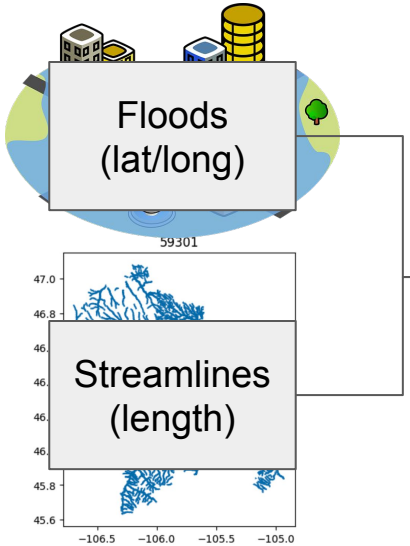
USGS National Hydrography Dataset

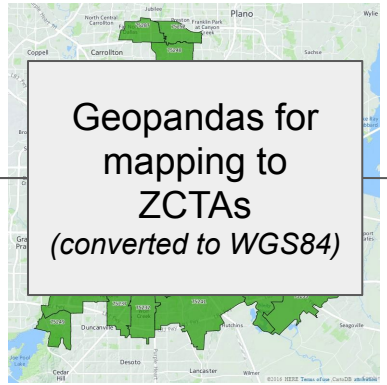
Streamlines → proxy for proximity to flood source

Geospatial Transformation: Zip Code Tabulation Area

Zip Code: list of addresses, not an area

US Census Bureau create Zip Code Tabulation Areas (ZCTAs)
(https://www2.census.gov/geo/tiger/TIGER2020/ZCTA520/)





Final Dataset by ZCTA
- # floods (total, decadal, season
- Median household income
- State
- Stream length in ZCTA
- Median flood duration

Data Cleaning

Missing Data

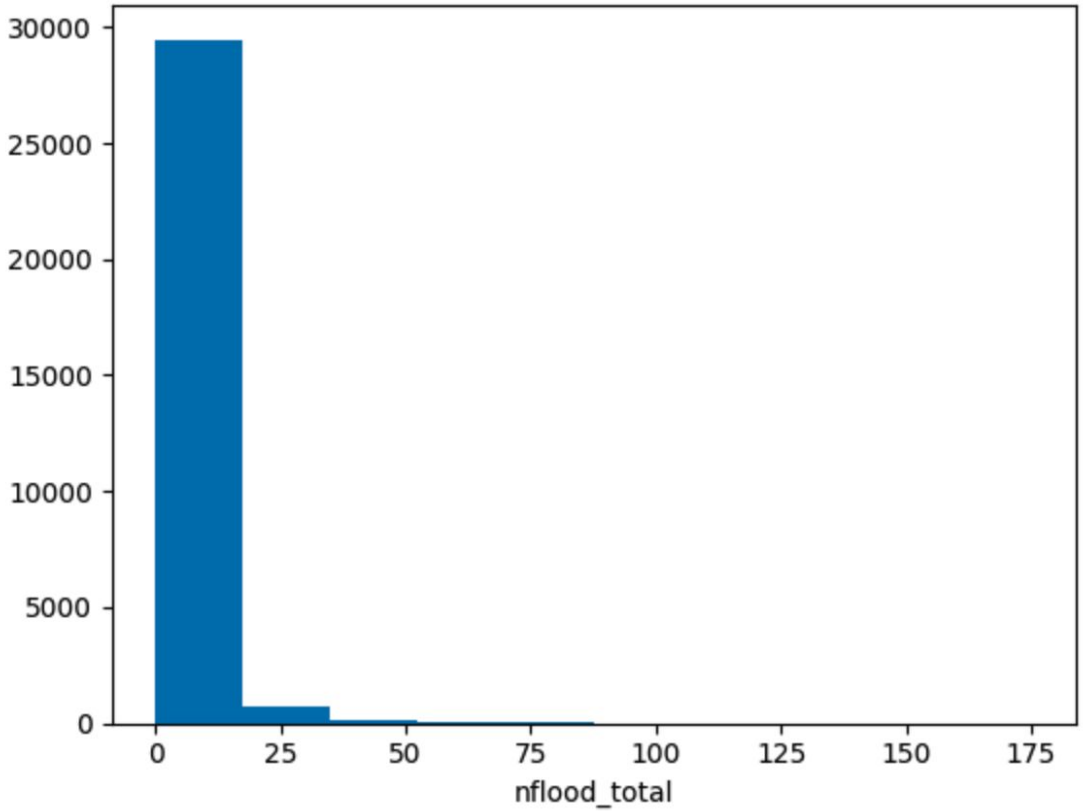
- Flood location not consistently recorded until 2006 → dropped
- Some floods in rural areas outside of mapped ZCTA → dropped
- Analysis for lower 48 states only

Data suggests:

- Many floods occur first of the month; might also mean unknown → not used
- Injury/death data mostly zero; zero might also be unknown → not used

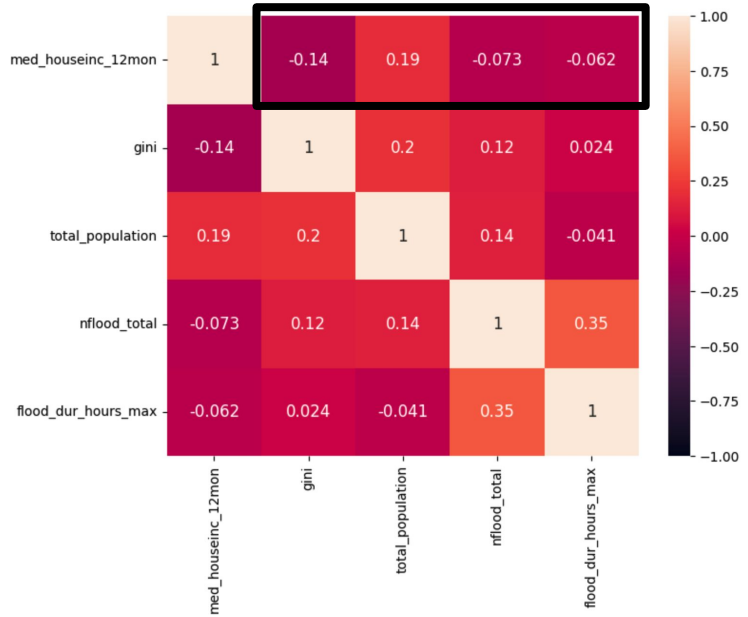
Data Summary

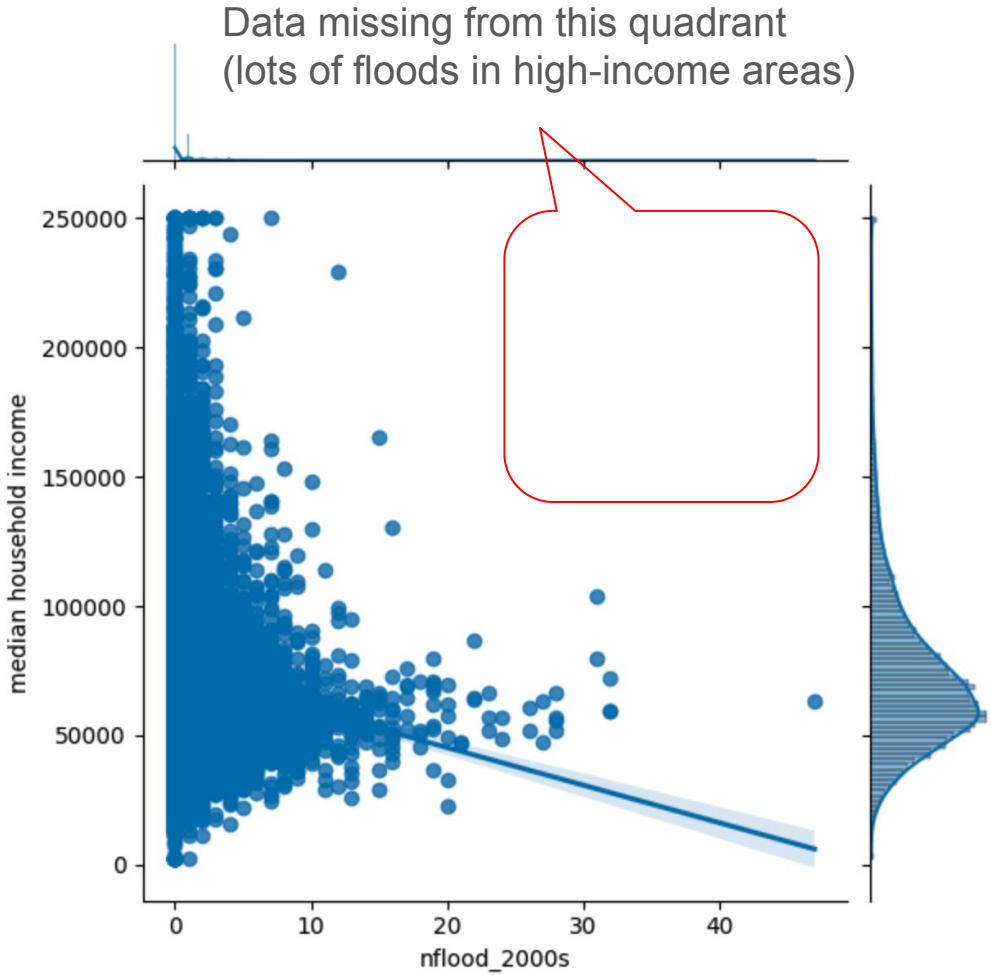
Number of floods heavily skewed; most ZCTAs have only a few floods, a few ZCTAs have lots of floods



Data Summary

Skewed flood data means low correlation with other individual features





Feature Engineering

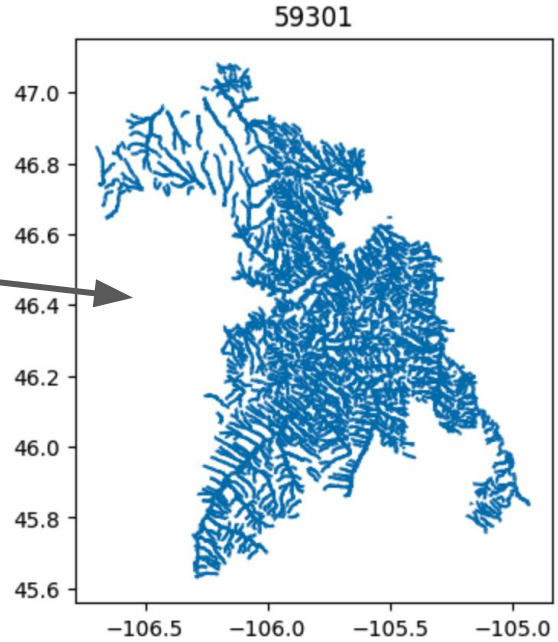
Number of floods in each season (fall, winter, spring) and decade (2000 - 2010, 2010 - 2020)

Mean, median, max flood duration

Sum stream channel length in ZCTA

Power transform features to address heavy tails

Note: Census Gini index is a measure of how far the Lorenz curve is from line of equality, ranges from 0 to 1 and higher values indicate higher income inequality.



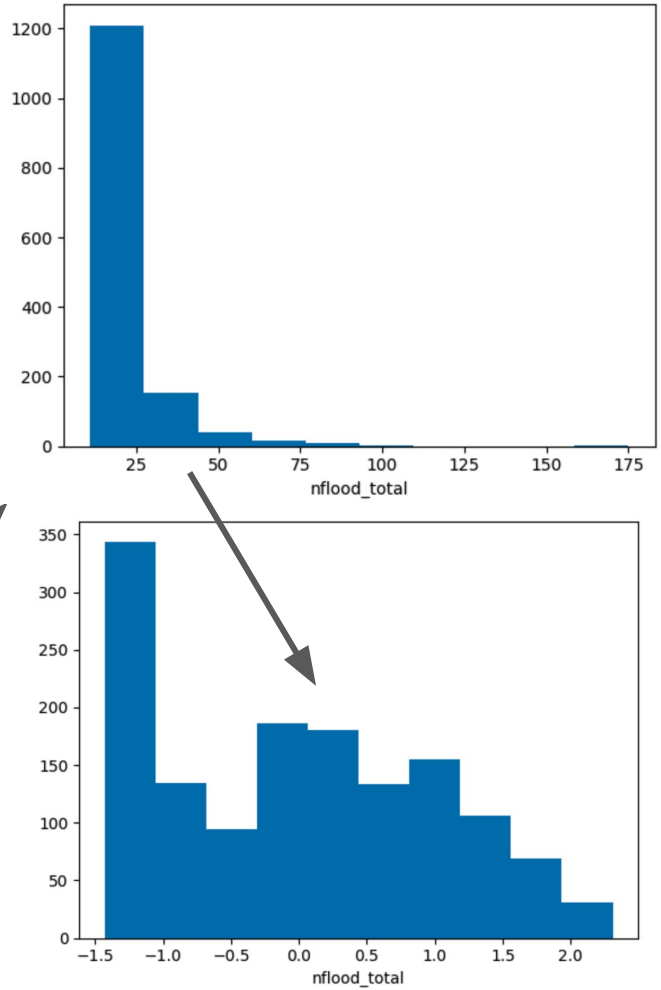
Feature Engineering

Number of floods in each season (fall, winter, spring) and decade (2000 - 2010, 2010 - 2020)

Mean, median, max flood duration

Sum stream channel length in ZCTA

Power transform features to address heavy tails

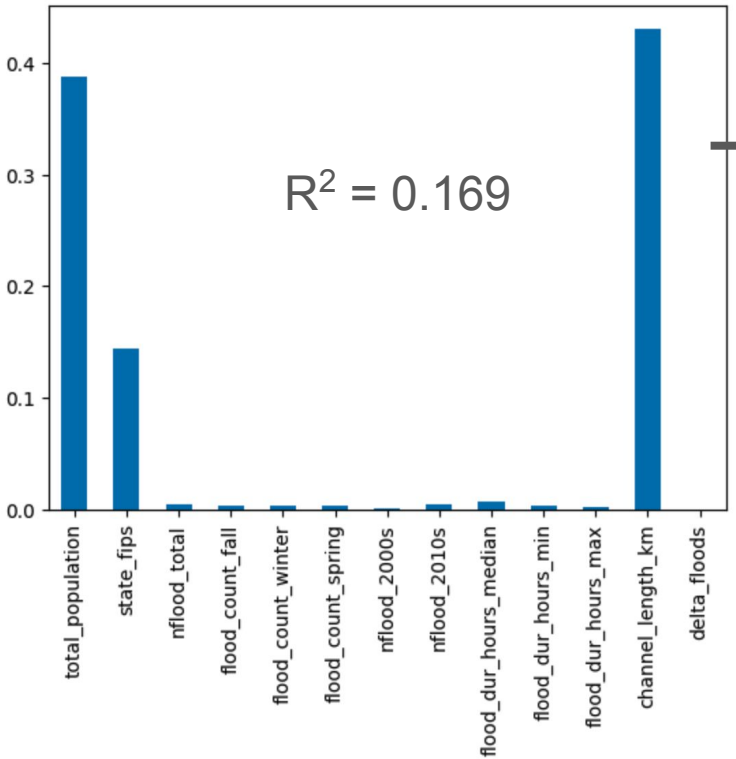


Note: Census Gini index is a measure of how far the Lorenz curve is from line of equality, ranges from 0 to 1 and higher values indicate higher income inequality.

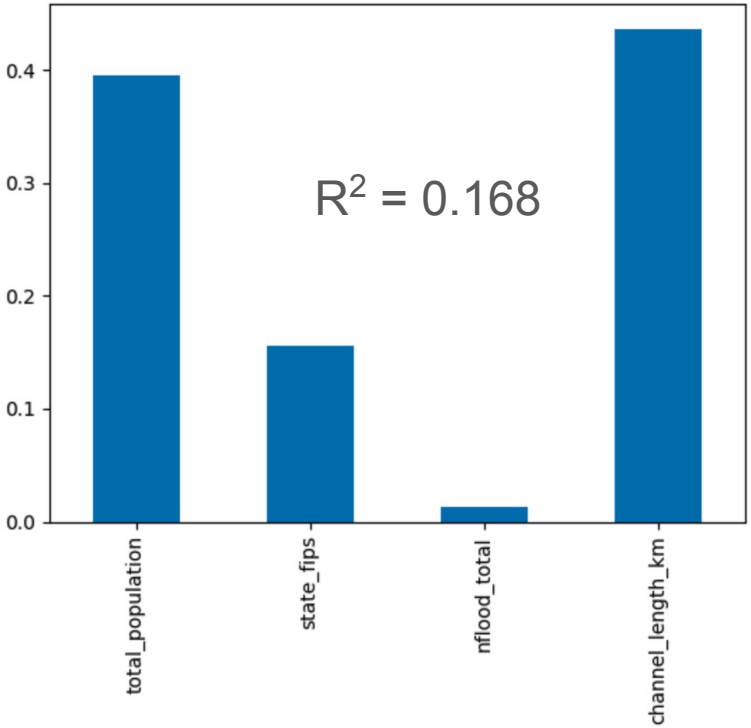
Model: Random Forest

Step 1: Remove correlated features

Target: Median household income
Test/Train Split: 20/80

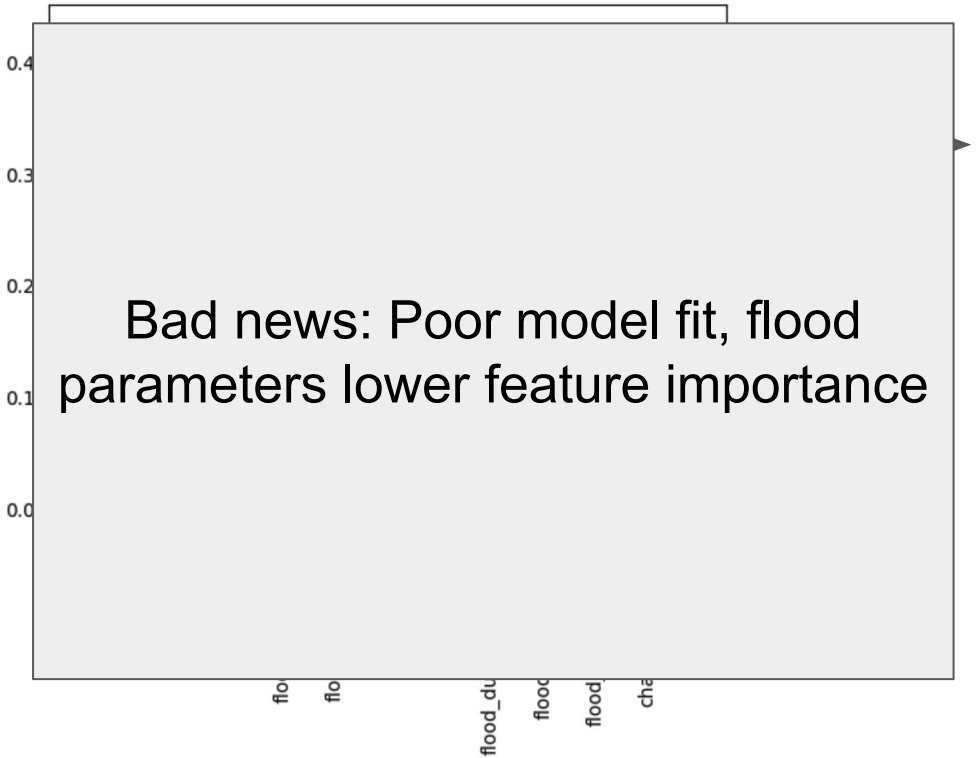


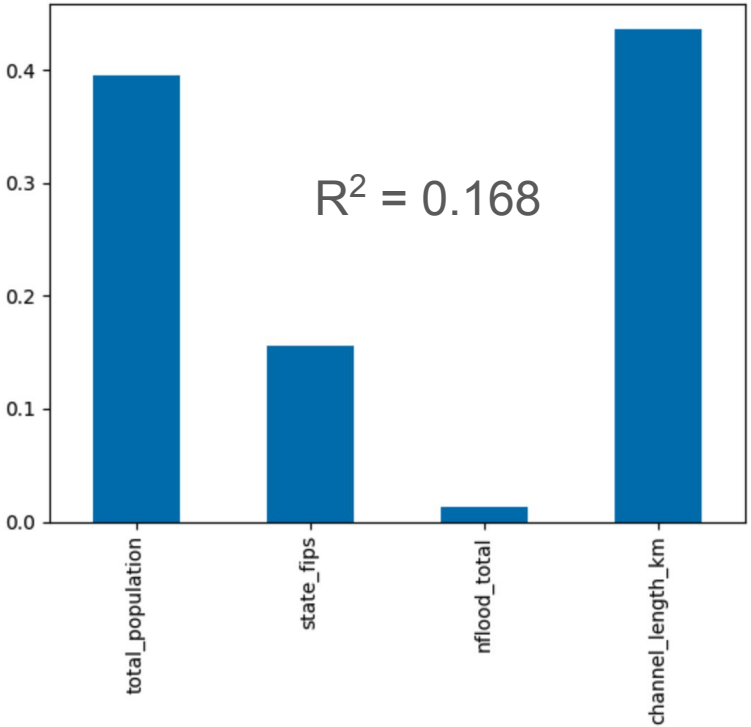
Remove features one at a time until R^2 drops



Model: Random Forest

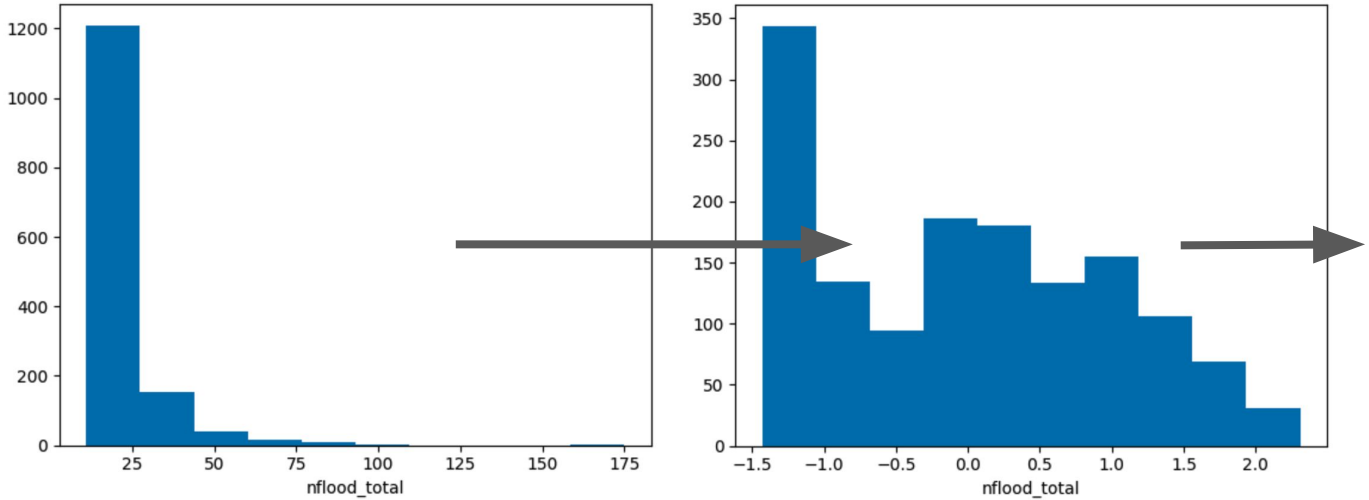
Step 1: Remove correlated features





Model: Random Forest

Step 2: Transform heavy-tailed data

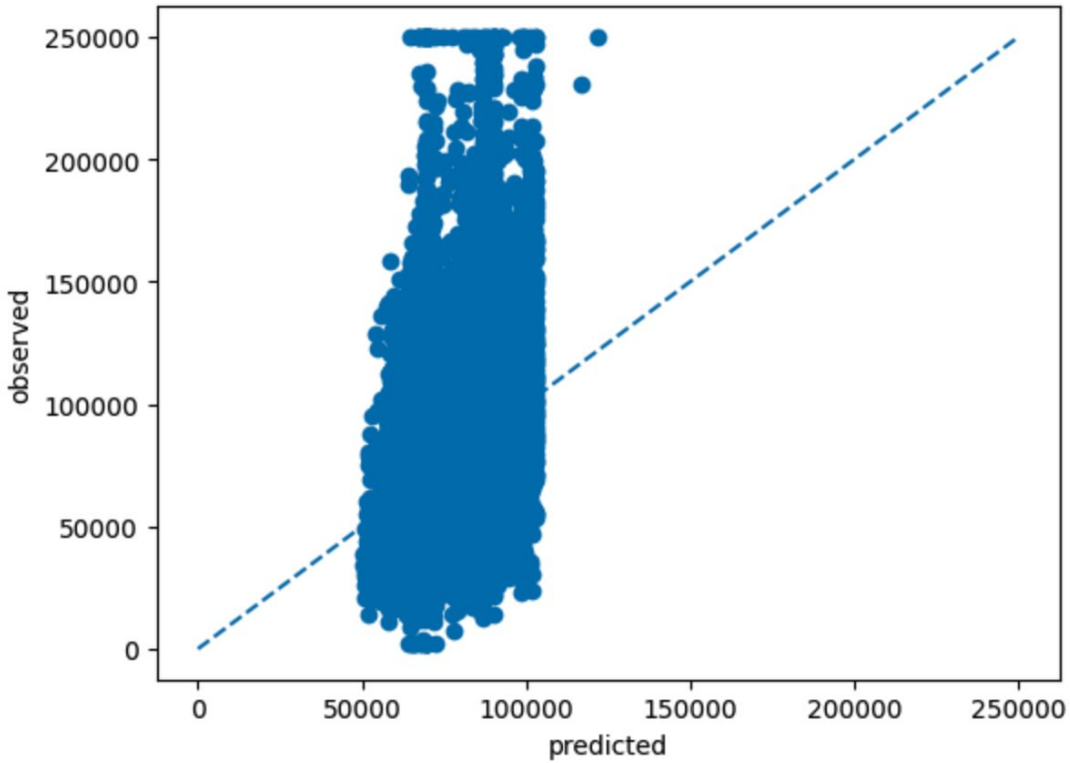


Same results: RFR insensitive to transform

$R^2 = 0.169$

Model: Random Forest

Step 3: Where are the model misses?

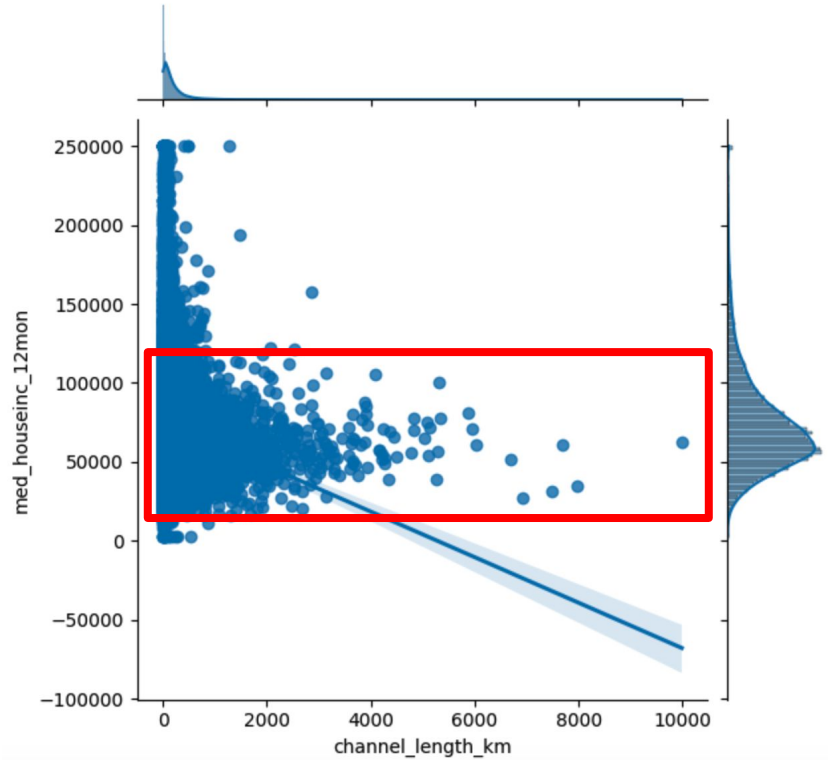


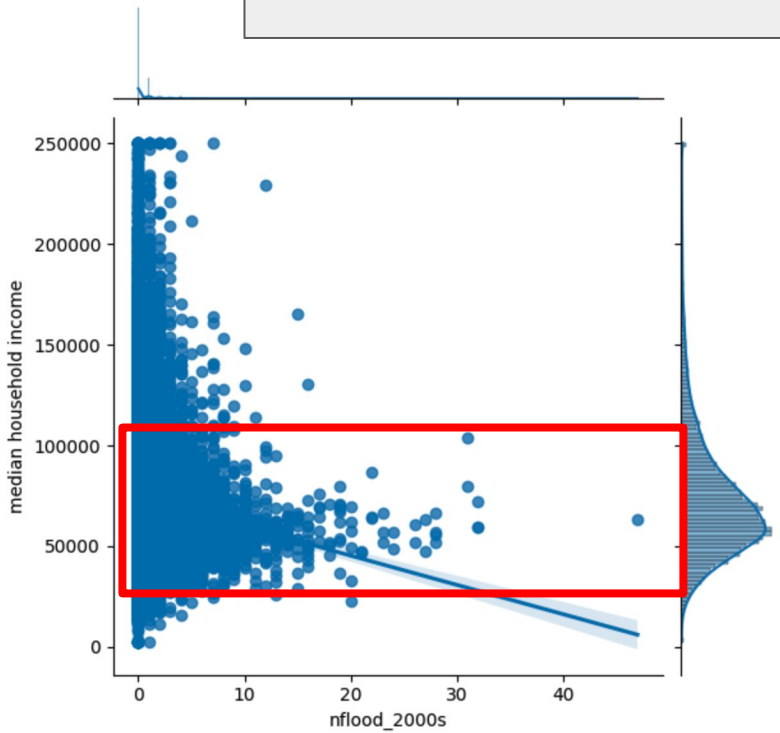
Highest error for lowest/highest incomes

Model: Random Forest

Step 3: Where are the model misses?

Most data variability in mid-range incomes



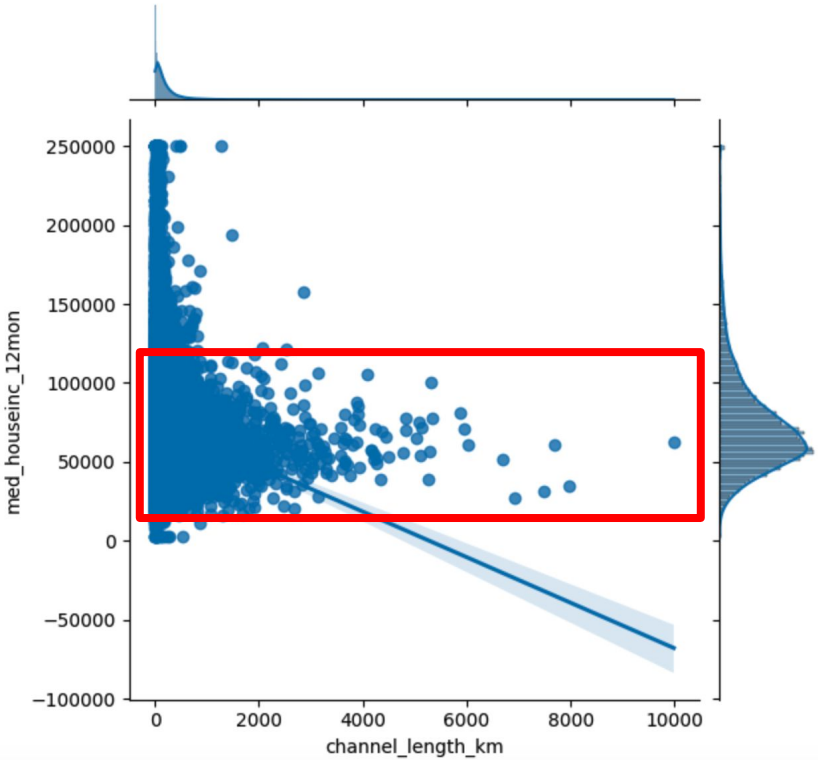


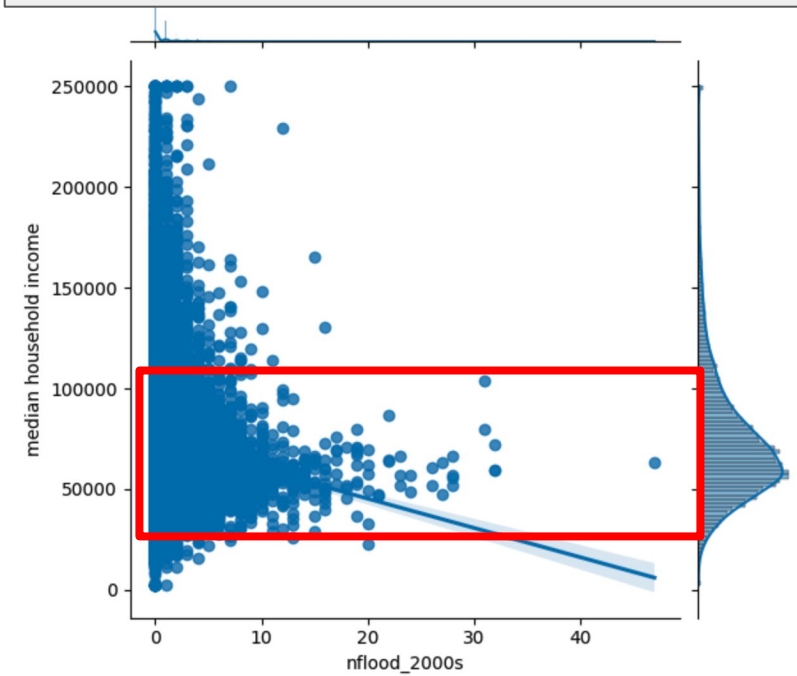
Model: Random Forest

Step 3: Remove outliers

Things to try:

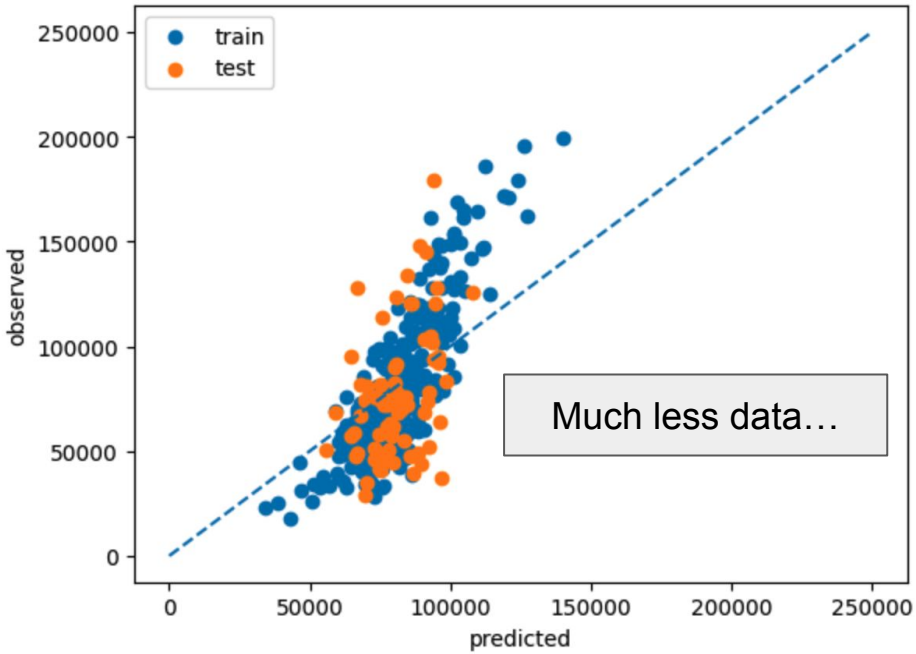
- Limit to ZCTA with more floods (>10)
- 250k max income, limit max income (<200k)
- Low channel lengths have all incomes; perhaps not applicable to study of where flood impacts income (<100km)





Model: Random Forest

Step 3: Remove outliers



Train R^2: 0.52
Test R^2: 0.14

Model overfitting, not enough data

Model: Random Forest

Step 4: Model score without flood features?

First model Train R^2: 0.17

Optimized hyperparameter model R^2: 0.19
(grid search)

Without flood features and optimized hyperparameters R^2: 0.16

So flood features explain 1% of income variability… not great

Summary

- At ZCTA level, flood occurrence does not seem to affect median household income

  (also tried Gradient Boost → model overfit)

  (similar results with Neural Network Model)

Next Steps

- Try at Census Tract level