EXECUTIVE SUMMARY

Who's Being Priced Out of Protection? Predicting FAIR Plan Enrollment in California

Leo Barleta, Abhay Chaudhary, Allison Lucas, Tiana Townsend
https://github.com/Abhay-Chaudhary/Climate-Insurance-Redlining

Introduction

California's residential insurance market faces unprecedented disruption as major insurers retreat due to escalating wildfire risk, construction cost increases, and volatile reinsurance markets. State Farm, the state's largest residential insurer, suspended new policy issuance in 2023 and did not renew about 72,000 existing policies. Insurers are now passing higher premiums to consumers and pushing homeowners toward the California FAIR Plan—a state-backed insurance program that provides basic property coverage when traditional insurers refuse to write policies. In Sacramento County, FAIR Plan usage more than doubled in one year. With FAIR Plan resources limited (~$377 M liquidity, as of Jan 2025), regulators and advocates warn of an "uninsurable future" for wildfire-exposed homeowners.

This analysis examines whether FAIR Plan enrollment patterns and financial liability can be predicted using disaster risk exposure, socioeconomic indicators, and insurance cost metrics. Developing a reliable forecasting model would enable policymakers, community organization, and homeowners to anticipate coverage gaps, allocate resources proactively, and identify communities vulnerable to insurance market disruption.

Data Collection

We compiled a comprehensive dataset combining FAIR Plan data for 2022 with publicly available information about residential insurance policies, housing prices, socioeconomic conditions, and climate-related disasters. These data was retrieved from the following sources (a longer description of the data sources is available on the data_prep.ipynb notebook):

- American Community Survey - reported income, race, and housing conditions.
- Zillow Housing Value Index
- Governor-proclaimed disasters from 1991 to present.
- California Department of Insurance residential policy data, based on insurance company bi-annual reports.

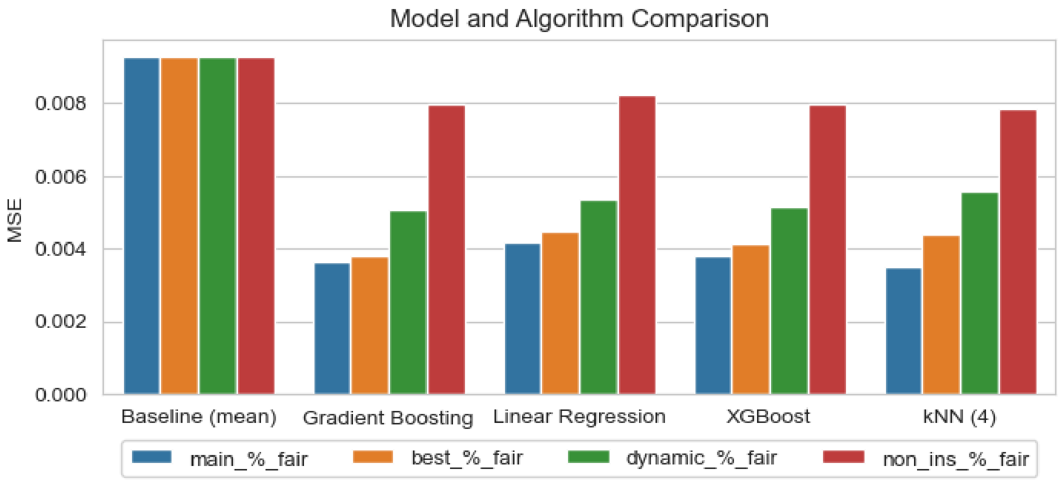
The primary geographic unit of analysis is the ZIP Code Tabulation Area (ZCTA), which aligns with the resolution of most available datasets. We predict the usage of FAIR Plan policies in a given ZIP Code, measured in percentage of residential units. As a secondary target, we also estimate the financial impact of the program's expansion, based on total exposure covered in FAIR Plan policies. This variable measures the maximum potential loss California could face in case of climate disaster.

We focus on the 2018–2021 period, where data from multiple sources are available and comparable. Given that detailed FAIR Plan data is only available for 2022, we computed the rate of growth of a given indicator (e.g., growth in house prices) for the four previous years for some key variables.

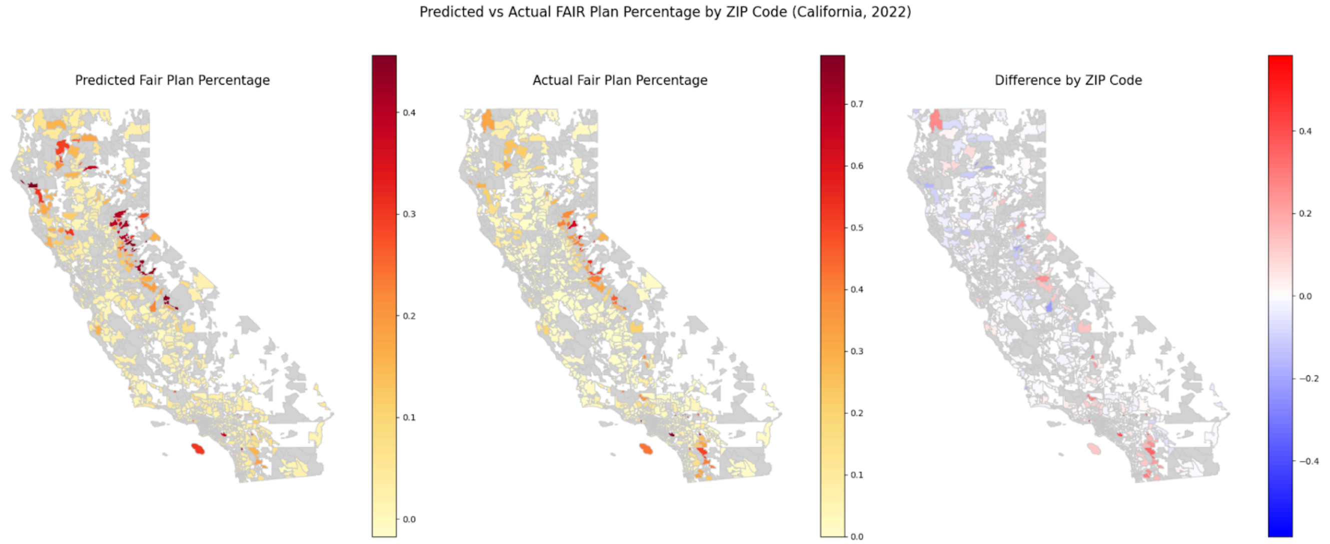
Modeling Approach

EDA and Feature Engineering: We performed common EDA tasks, such as determining p-values and feature importance via Random Forest modeling. We engineered features from the cleaned dataset to transform data into more informative and suitable features. Two examples of these features are Renewal Resilience, the proportion of policies successfully renewed, and Housing Value to Median Income Ratio, both created by combining other features. The final collection of all features was then sorted by p-value or feature importance and redundancies were removed with lower p-values and higher feature importance taking precedence. Additionally, visual elements such as a correlation matrix and plots of the main features overlaid on a map of California were helpful in preparing our most indicative features for model selection.

Model Selection: we developed different features sets based on exploratory data analysis and use cases that we envision that our project could provide insight, as well as compare predicting techniques to produce the optimal regressor. We compared linear regression, kNN regression, Gradient Boosting, XGBoost, and logistic regression (in this case, using the prediction probabilities) to a baseline model (mean) using cross-validation. Our metrics were MSE, RMSE, MAE and, for exposure, the total monetary amount underwritten in FAIR Plan policies.



The most performant model significantly improved baseline predictions for FAIR Plan usage, as measured by MSE. It uses Gradient Boosting on the “main_features” set, which includes important variables detected in EDA. However, models only using the top 2 predictors ('Renewal Resilience' and 'Change in Earned Premiums') performed very similarly. The sets “dynamic_features” and “non_ins_%_fair”–respectively, features that capture changes over time and features not using data reported by insurance companies–had the worst performance. A visual inspection of our predictions, as presented in the maps below, suggests that our models have been able to capture the spatial patterns in the data.


Predicted vs Actual FAIR Plan Percentage by ZIP Code (California, 2022)
Predicted Fair Plan Percentage
0.0
0.1
0.2
0.3
0.4
Actual Fair Plan Percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Difference by ZIP Code
−0.4
−0.2
0.0
0.2
0.4

The secondary model predicting total exposure also improved baseline estimates, nearly halving the difference between the actual values underwritten in FAIR Plan policies and their predictions. In absolute values, our predictions were $19 billion dollars (compared to $35 billions of baseline estimates) off, from a total of $164 bi in exposure covered in California-backed insurance policies. However, this specific model is not very stable and would require additional tuning for a more realistic assessment.

Insights and Implications

Data-driven Forecasting: Our models achieved strong predictive performance for the percentage of FAIR Plan-covered units, suggesting that key drivers in our dataset reliably signal where traditional homeowners insurance is becoming harder to get.

Policy Utility: These forecasting tools could help different stakeholders prepare for shifting insurance markets and target interventions more effectively, including:

- Regulators and policymakers: FAIR Plan models into regulatory oversight to ensure coverage availability in vulnerable communities

- Community organizations and homeowners: leverage localized predictive insight to build targeted advocacy campaigns, negotiate collective mitigation funding, and challenge unjust non-renewals; identify communities likely facing increasing reliance on FAIR Plan due to climate pressure.
- Insurance companies and Real Estate Speculators: make informed decisions about market entry, exit, or reinsurance strategy, or forecast future property value volatility. ZIP codes flagged for insurance instability may be seen as high-risk investments—or, conversely, as opportunities for speculative acquisition ahead of state intervention or market recovery.

Reliance on Insurer-Provided Data: The accuracy of our models relied heavily on features derived from data provided by insurance companies, such as policy renewals and total premiums. Although not surprising, this finding raises concerns about potential conflict of interest between private companies, regulators, and affected communities and reinforces the need for consistent, fair, and transparent policies regarding data reporting.

Broader Legislative Lessons
California's ability to analyze these trends stems in part from 2018's Senate Bill 824, which requires insurers with more than $10 million in written premiums to submit detailed biennial reports. Other states with FAIR Plan programs — or those anticipating similar insurance market pressures — could benefit from adopting comparable reporting requirements to track market shifts and plan early interventions.

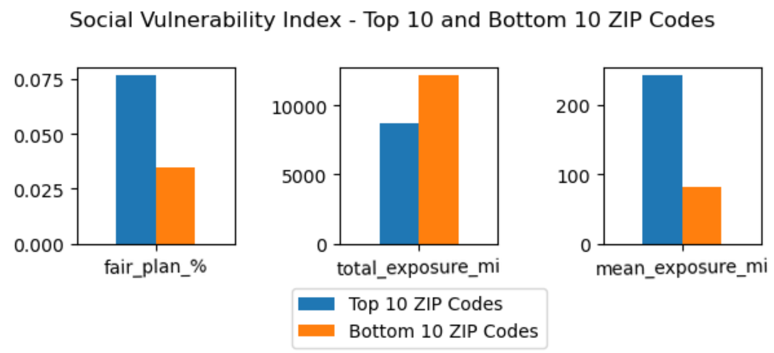
Future Directions

Dataset Enhancements. While we assembled a multi-source dataset and engineered features that proved to be key to our models, our experience suggests that further work in data collection and preprocessing could enhance the quality of our predictions. This includes:

- Improve transformations with datasets in different resolutions, since we have to perform transformations from county-level data to ZIP Code area, by incorporating new data and using geospatial techniques. This issue was particularly acute in climate disaster and racial composition data, which weren't strong predictors.
- Examine potential proxy data from census and state agencies that could serve as an independent predictor or proxy to insurer-provided data. I'll mitigate the absence of data reporting mandates and allow other states to use our models.

Temporal Modeling: Current data availability restricted our ability to perform time-series analysis in this particular dataset, but future data improvements may allow us to do it. That will enable us to evaluate lag effects of disasters and socioeconomic changes and forecast changes in insurance markets.

Socioeconomic Impact and Biases: further our investigation of whether certain groups have been systematically "priced out" or "left out" by premium increases or end of coverage in certain areas, considering underlying effects of this phenomenon (e.g., one's capacity to move to a less disaster-prone area.) As the preliminary analysis below indicates, the least vulnerable ZIP Codes (based on Social Vulnerability Index) have much higher enrollment rates in the FAIR Plan program.



Indexing: Develop a single-value "Insurance-Climate Stress Index" to increase communicability of our analysis to decision makers and replicability in different data contexts.