CalEmber - A Fire Damage Prediction & Insurance Assessment Tool Final Presentation: 12/09/24

Michelle Cheung
Ishika Prashar
Jackie Wang
Hao Xie

Problem - California Wildfires and Insurance

Wildfire severity and risk in California

Lucrative housing market

Fluctuating Insurance Market

High Fiscal Impact

$87,290,000,000 → Total California Wildfire Cost 1980 - 2021

$9,000,000,000 → California property insurance market annual premiums

359,831 → Residential properties sold in CA over past 12 months

Expert Interview - Micah Mumper, PhD

Role: Research Data Specialist at California Department of Insurance

- Insurance premium calculation via regression modeling
- Uninsured homeowners

Our Solution: Interactive CalEmber Website

1. <u>Current Homeowners</u>
   Damage Predictions
2. <u>Prospective Homeowners</u>
   Fire Severity Scores
3. <u>All Users</u>
   Insurance and Damage Dashboards

Cal Fire Damage Inspection (DINS) Data

Structures damaged by wildfires from **2013 - 2021**

Key categories → home metrics, risk features

Damage levels:

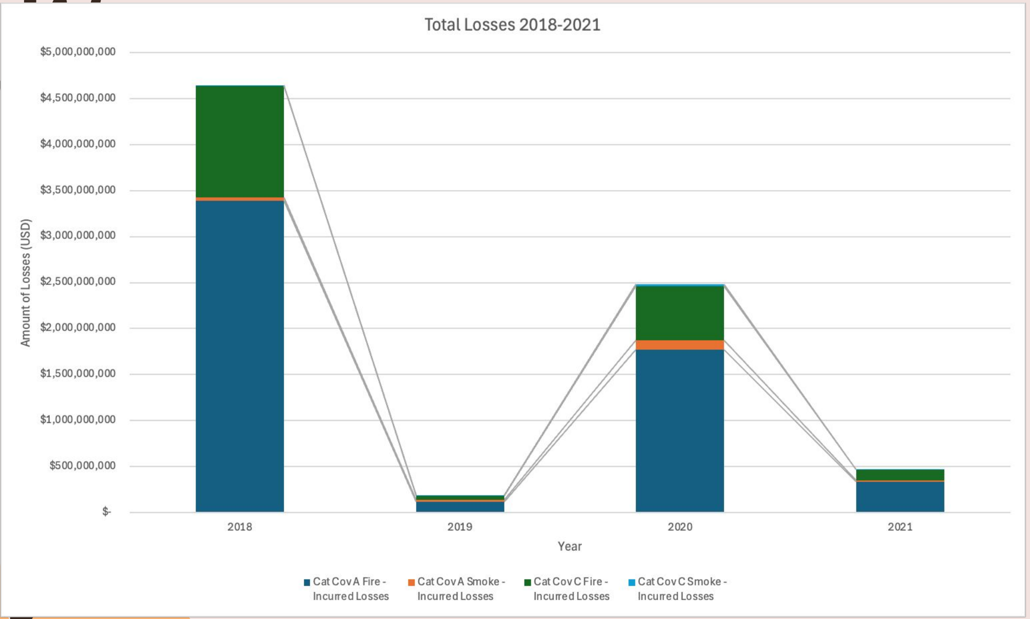
0 %	1 - 10 %	11 - 25 %	26 - 50 %	> 50 %
No damage	Low damage	Minor damage	Major damage	Destroyed

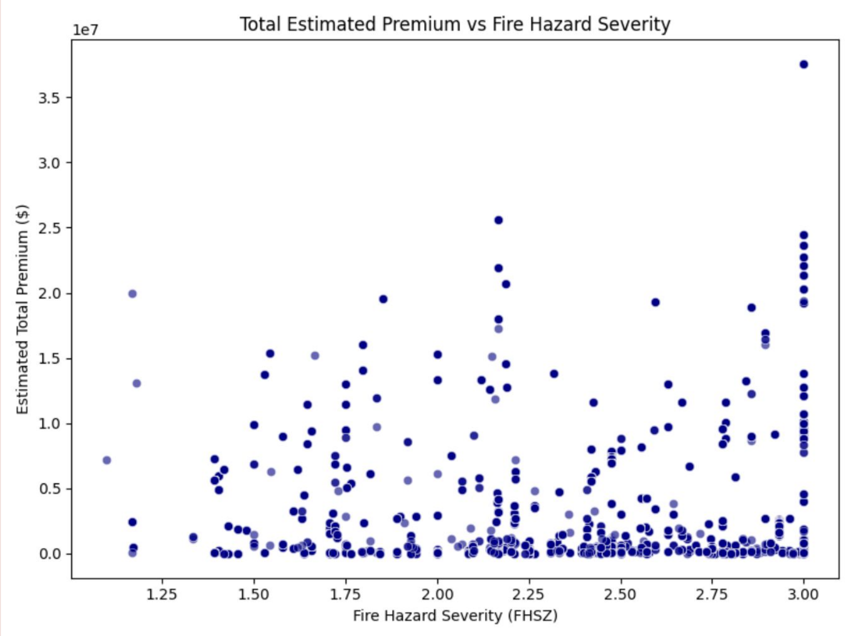
Data Preprocessing: AWS location services, one hot encoding, ordinal scale

Secondary Data Sources

California Department of Insurance Personal Property Experience Data	• 2009-2021 • Fire loss, premiums, insurance metrics • Used for dashboards
Fire Hazard Severity Zones	• Regions in state with moderate, high, and severe fire severity scores • Used for zip code severity look-up tool

Fluctuations in Fire Losses & Increased Premiums in High Fire Hazard Severity Regions





Architecture

AWS S3 Bucket

AWS Sagemaker

Flask

Tableau

Html CSS
JavaScript

Initial Modeling

Model	Accuracy (Test Set)	Comments
Baseline	57.9%	• Always predict **majority class** damage level 4
Linear Regression	39%	• RMSE 1.11 • Key features - single residence, roof, siding
Recurrent Neural Network with LSTM	36%	• **No sequential pattern** between damage and input metrics
K Nearest Neighbors	77%	• Well suited for geographic regions
Support Vector Machine	87%	• These models perform well on the majority classes (0 and 4), but is unable to predict the minority classes (1, 2, and 3) correctly - **class imbalance**
Random Forest	93%	
LightGBM	93.2%	
XGBoost	**93.3%**	

Class Imbalance Concerns

Classification Report:

		precision	recall	f1-score	support
No damage	0	0.92	0.97	0.95	2954
low damage	1	0.63	0.42	0.50	317
minor damage	2	0.26	0.12	0.17	73
major damage	3	0.19	0.11	0.14	27
destroyed	4	0.96	0.96	0.96	4625
	accuracy			0.93	7996
	macro avg	0.59	0.52	0.54	7996
	weighted avg	0.92	0.93	0.93	7996

Model Optimization - Part 1

→ SMOTE (Synthetic Minority Oversampling Technique) oversampling minority classes in attempt to resolve class imbalance
→ Random gridsearch for optimized hyperparameters
→ Combine multiple models

Ensemble - Voting Classifier

	precision	recall	f1-score	support
0	0.93	0.98	0.95	2954
1	0.56	0.45	0.50	317
2	0.25	0.23	0.24	73
3	0.14	0.11	0.12	27
4	0.97	0.95	0.96	4625
accuracy			0.93	7996
macro avg	0.57	0.54	0.55	7996
weighted avg	0.93	0.93	0.93	7996

XGBoost only

	precision	recall	f1-score	support
0	0.92	0.97	0.95	2954
1	0.54	0.46	0.50	317
2	0.29	0.22	0.25	73
3	0.13	0.11	0.12	27
4	0.96	0.94	0.95	4625
accuracy			0.92	7996
macro avg	0.57	0.54	0.55	7996
weighted avg	0.92	0.92	0.92	7996

Reframing the Problem

Refined approach → what matters most to homeowners?

- Solution → simplified damage outputs

Old approach:

0 %	1 - 10 %	11 - 25 %	26 - 50 %	> 50 %
No damage	Low damage	Minor damage	Major damage	Destroyed

New approach:

0 %	1-50%	> 50 %
No damage	Moderate Damage	Destroyed

Model Optimization - Part 2

1. Merging minority damage classes

2. Part 1 model optimization Learnings
   - XGBoost, Unscaled, SMOTE

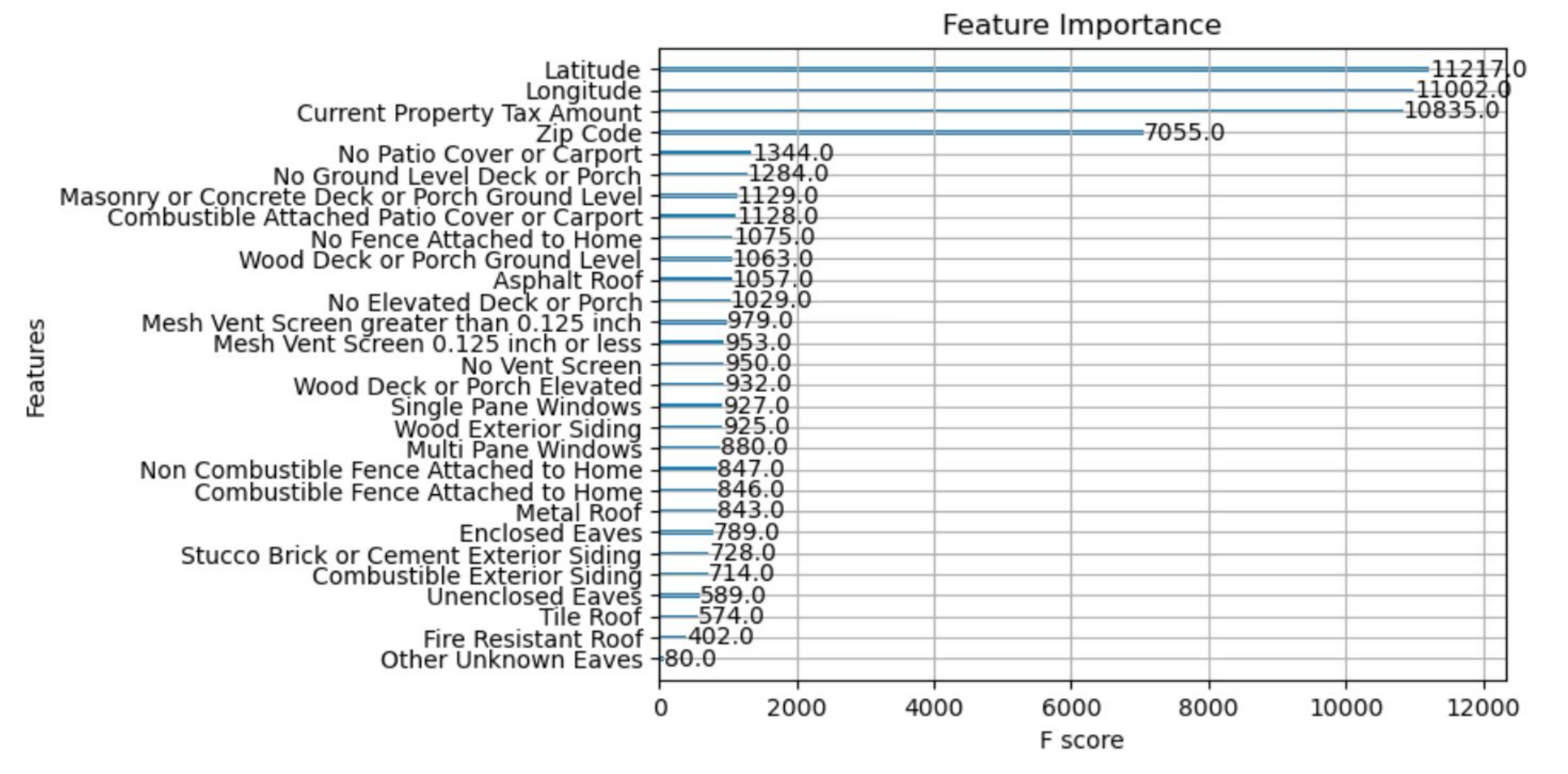
3. Feature Reduction
   - From 60 model features to 30

4. Hyperparameter Tuning
   - L1 & L2 regularization
   - 100 trees (n-estimators) with max depth 10 each tree
   - 60% of features can be used for each tree max

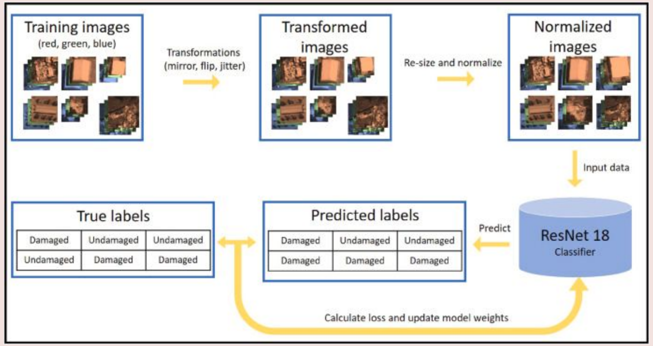
Final Model Results

		precision	recall	f1-score	support
No damage	0	0.94	0.97	0.95	2954
Moderate Damage	1	0.82	0.52	0.64	417
destroyed	2	0.95	0.96	0.96	4625
	accuracy			0.94	7996
	macro avg	0.90	0.82	0.85	7996
	weighted avg	0.94	0.94	0.94	7996

Feature Details



Technical Model Evaluation



Galanis, M., Rao, K., Yao, X., Tsai, Y.-L., Ventura, J., & Fricker, G. A. (2021). DamageMap: A post-wildfire damaged buildings classifier. *International Journal of Disaster Risk Reduction, 65*, 102540.

Test Set	Accuracy	F1
1	0.92	0.96
2	0.98	0.96

CALEMBER

	precision	recall	f1-score	support
0	0.94	0.97	0.95	2954
1	0.82	0.52	0.64	417
2	0.95	0.96	0.96	4625
accuracy			0.94	7996
macro avg	0.90	0.82	0.85	7996
weighted avg	0.94	0.94	0.94	7996

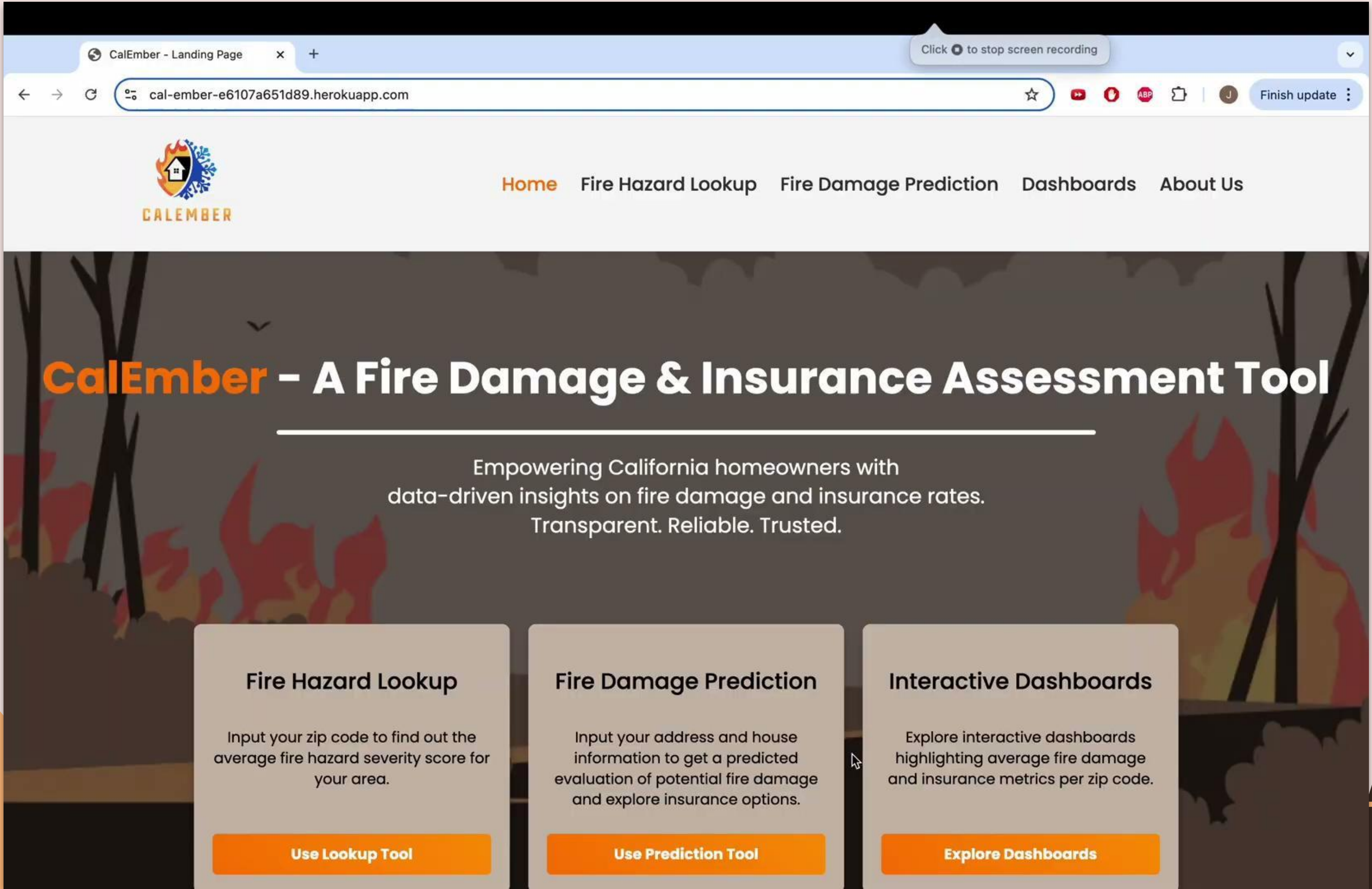
Model Results - Homeowner Perspective

Location Matters → Lat/Long, Zip Code, Property Value are important factors

Protective Measures → Fire Safety measures such as vent screens, non-combustible materials

Save Future Costs → Taking protective measures now may limit future damage levels

Demo: CalEmber User Perspective



User Interviews

Conducted user interviews from 13 people of various ages & professions to get feedback on our MVP!

User 1 (Name and Occupation):	
Walkthrough of website and explanation	Live Demo
Q1: How clear is this tool to use? Scale of 1-5 (1 is unclear, 5 is very clear)	
Q2: Is the tool easy to understand without any further background regarding data science or wildfire knowledge? Scale of 1-5 (1 is not easy, 5 is very easy)	
Q3: How well does this website help answer our research question/objective? Scale of 1-5 (1 is not well, 5 is very well)	
Q4: Any feedback on the visuals/navigation/website appearance?	
Q5: Any feedback on the content of the website that would make it better for consumers?	

Q1: How clear is this tool to use? Scale of 1-5

- Average score: **4.5**

Q2: Is the tool easy to understand without any further background regarding data science or wildfire knowledge? Scale of 1-5

- Average score: **4.23**

Q3: How well does this website help answer our research question or objective? Scale of 1-5

- Average score: **4.73**

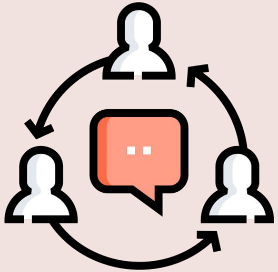
Q4: Any feedback on the visuals/navigation/website appearance?

- Implemented changes in navigation and sectioning of content to make it more user-friendly

Q5: Any feedback on the content of the website that would make it better for consumers?

- Updated information on webpage to make it more informative and helpful from a potential California homeowner's POV

Potential Next Steps



Connect and communicate more with the source of data

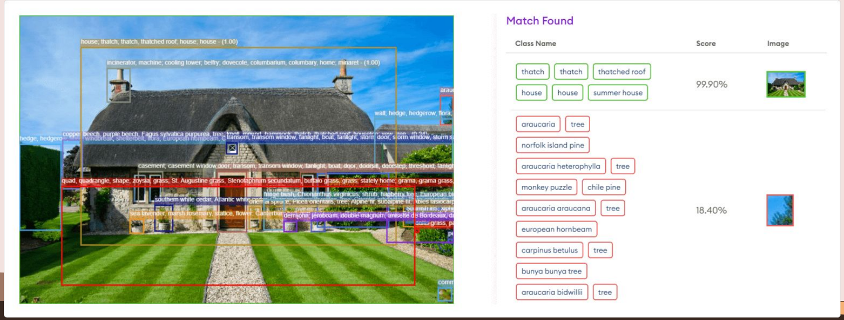
- Limitations of our data

Synthesis between damage and insurance

- Could be interesting to project insurance cost ranges given wildfire risks, projected damage level, and other housing characteristics

Feature recognition through images into the model pipeline

- User can just input pictures of their houses and the features will be picked up



Ethics/Privacy

Unexpected Uses

Use of the tool beyond informative purposes - pricing/bias

Data Limitations

Ultimately, not every insurance company is represented in our data - could cause bias

Re-Identification

Zip Codes with limited properties may make it easy to identify homeowners of the area

Conclusion

End Result: A working prediction tool for users and interactive dashboards showing different insurance metrics and damages across California!

- Intersection of **advanced machine learning** and **impactful decision-making**
- Helping homeowners understand how **specific property features influence risk and damage,** enabling them to make targeted improvements that **save lives, protect properties**, and **reduce insurance costs**
- Providing **transparent insurance insights and fire severity insights** for current and prospective California residents, regardless of homeowner status

CalEmber empowers existing and future California homeowners by providing transparent information on fire damage and insurance rates using data driven insights - results they can trust! Thank you!

Appendix

Acknowledgment:

We extend our deepest gratitude to our advisors, Joyce Shen and Morgan Ames, for their invaluable feedback and steadfast support throughout the semester. We also wish to express our sincere appreciation to Micah Mumper, Research Data Specialist at the California Department of Insurance, for sharing his domain expertise. Finally, we thank our colleagues, friends, and family for their participation in user testing and for providing insightful feedback that greatly contributed to the success of this project.

Appendix

Datasets:

- Fire Hazard Severity Zones from California Open Data Portal [link]
- Residential Property Insurance from CA Department of Insurance [link]
- Property Market Share Data by California Department of Insurance [link]
- Zip Code Level Premium & Exposure data from CA Department of Insurance [link]
- Data and Analysis and Wildfires and Insurance from CA Department of Insurance [link]
- California Fire Incidents from CalFire [link]

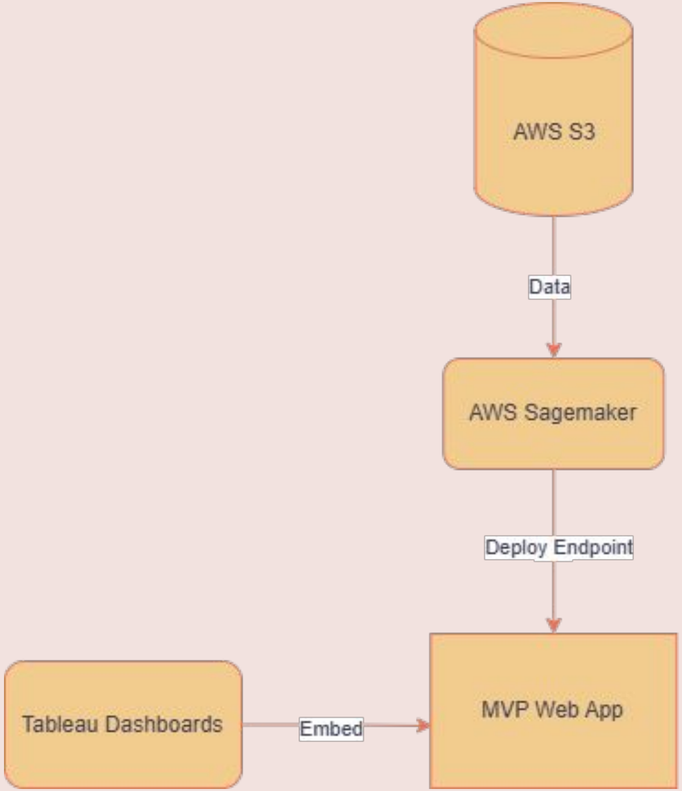
References:

- CNN [link]
- Redfin [link]
- SF Chronicle [link]
- LA Times [link]
- ATTOM [link]

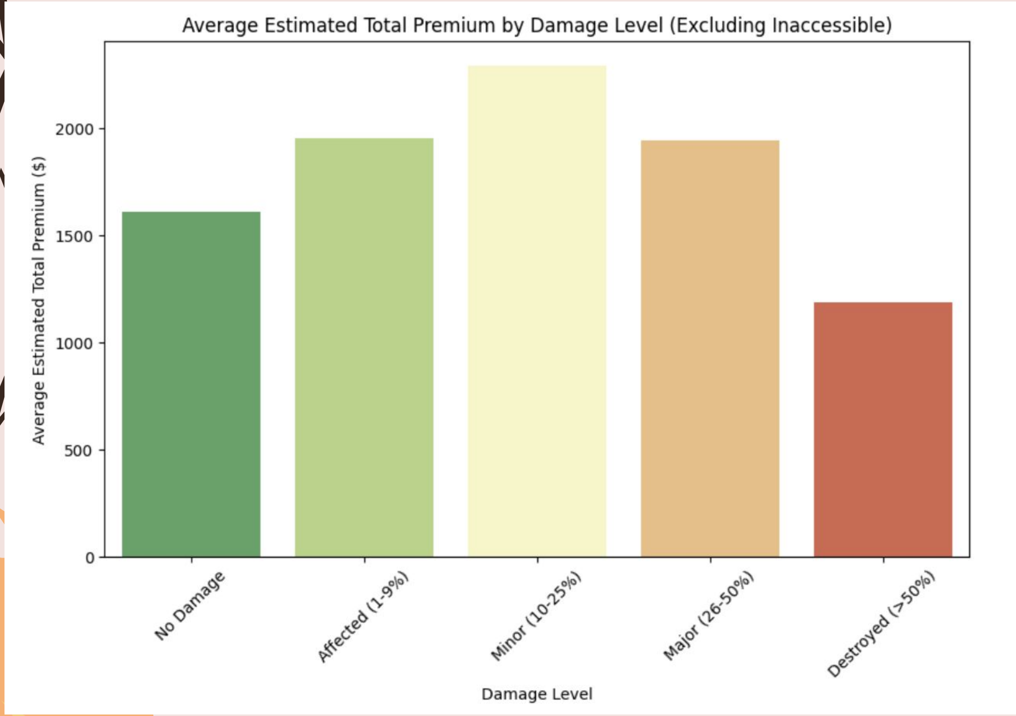
Appendix: Key Data Processing Steps

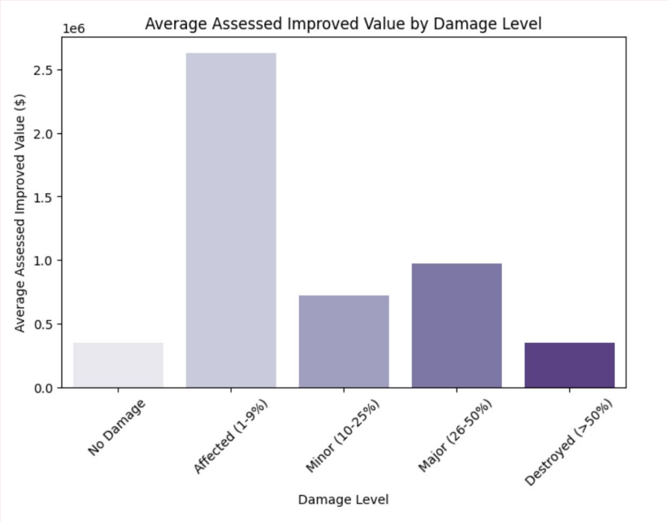
Damage Data	• AWS location services to transform lat/long into zip code • One hot encode categorical variables • Transform damage to ordinal scale
Insurance Data	• Estimate missing years
Fire Severity Zones	• GeoPandas transformations • Conversion to Zip Code Regions

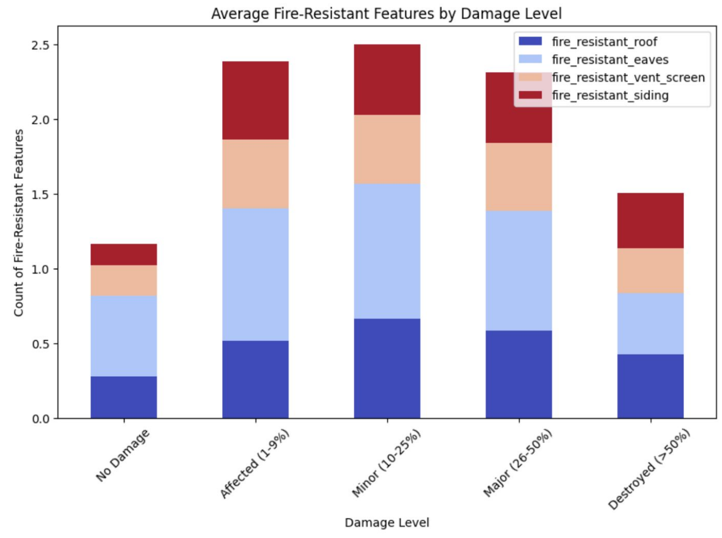
Appendix: Architecture → Data to Final Solution



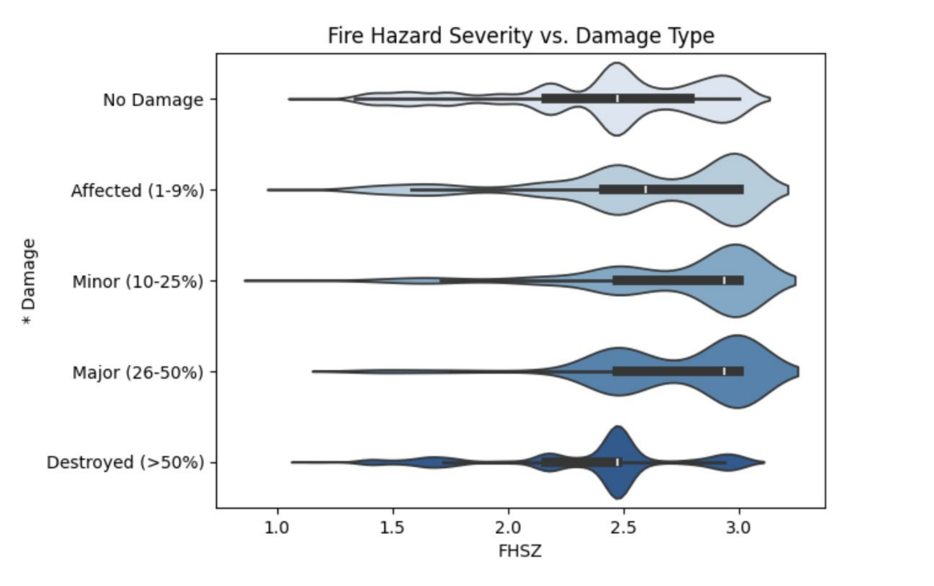
Appendix: EDA

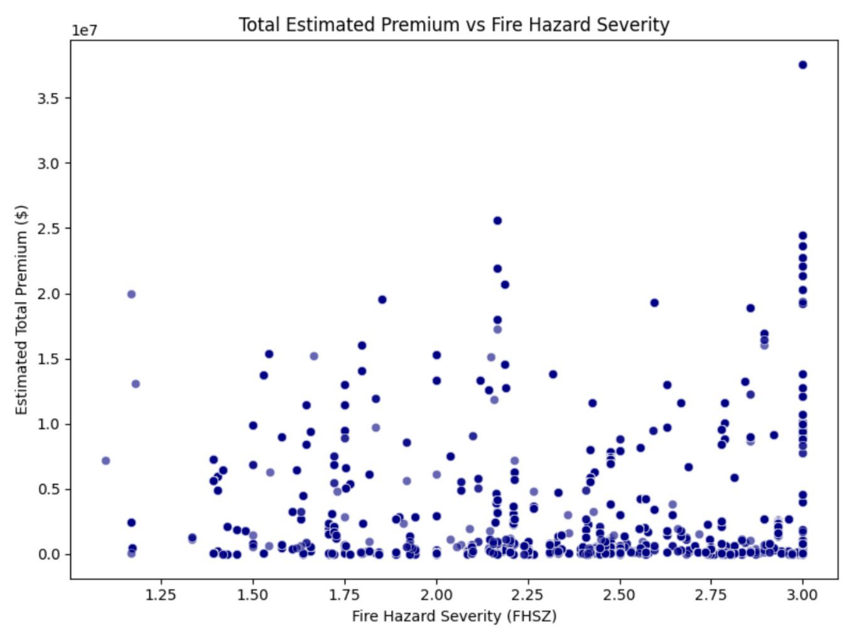




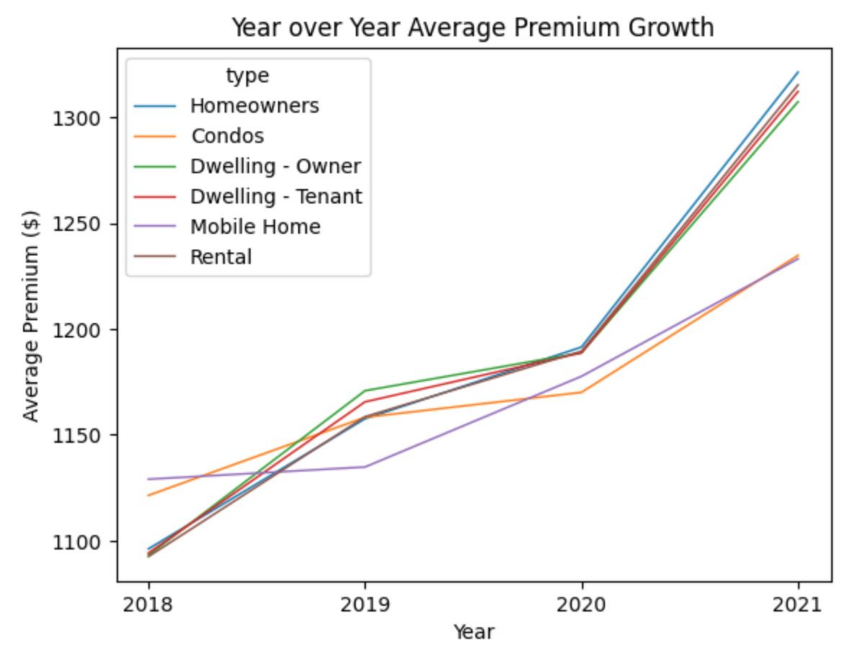


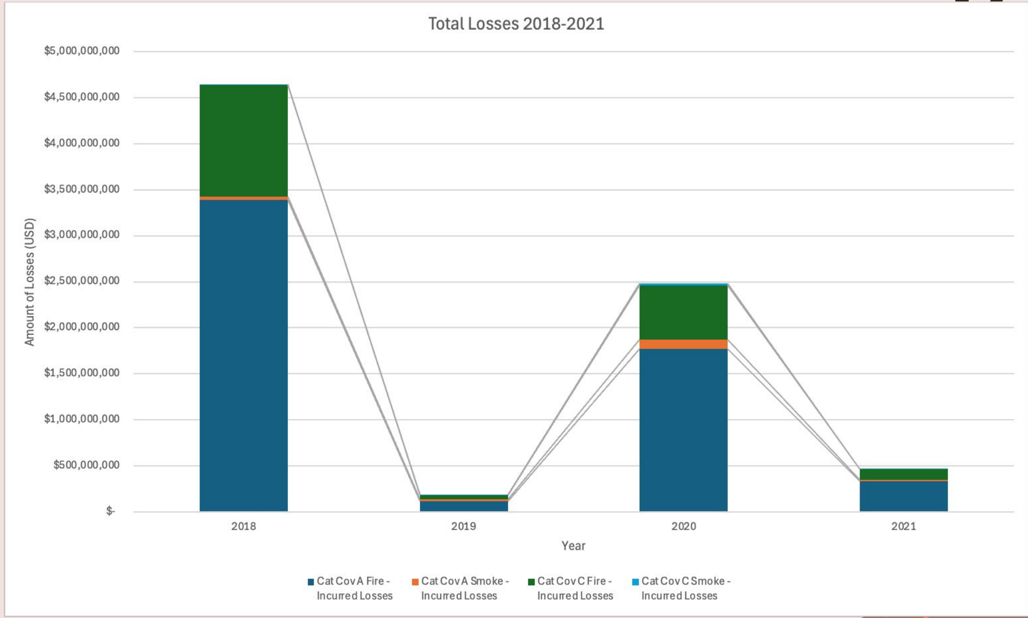
Appendix: EDA



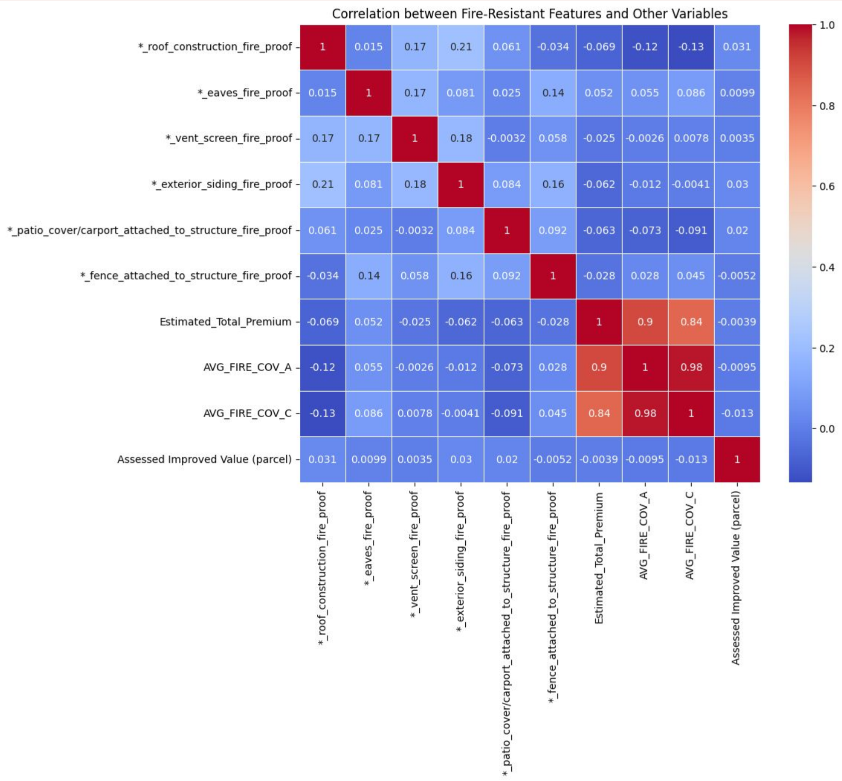


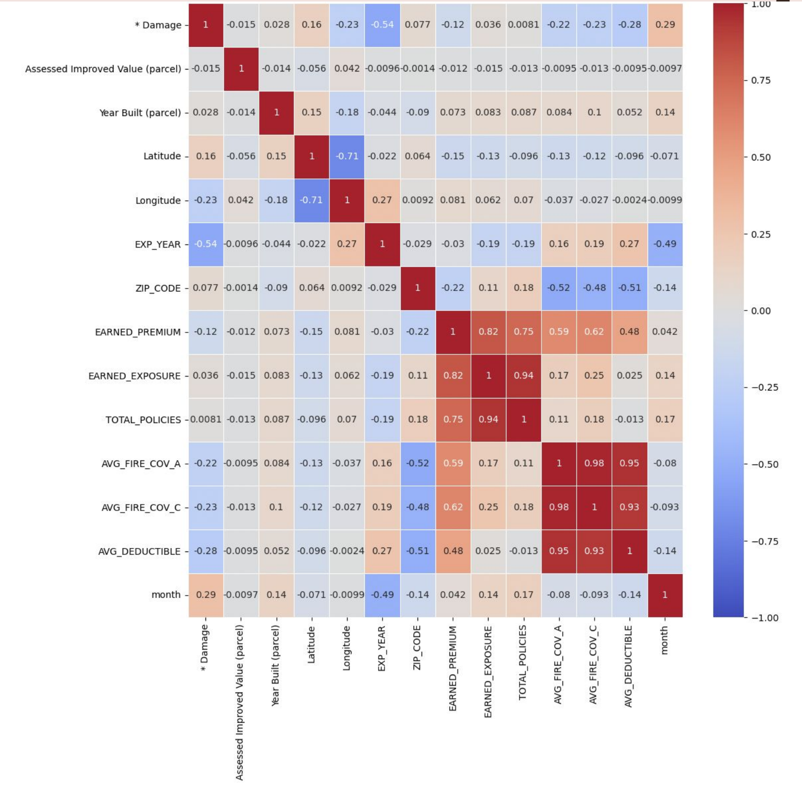
Appendix: EDA





Appendix: Key EDA



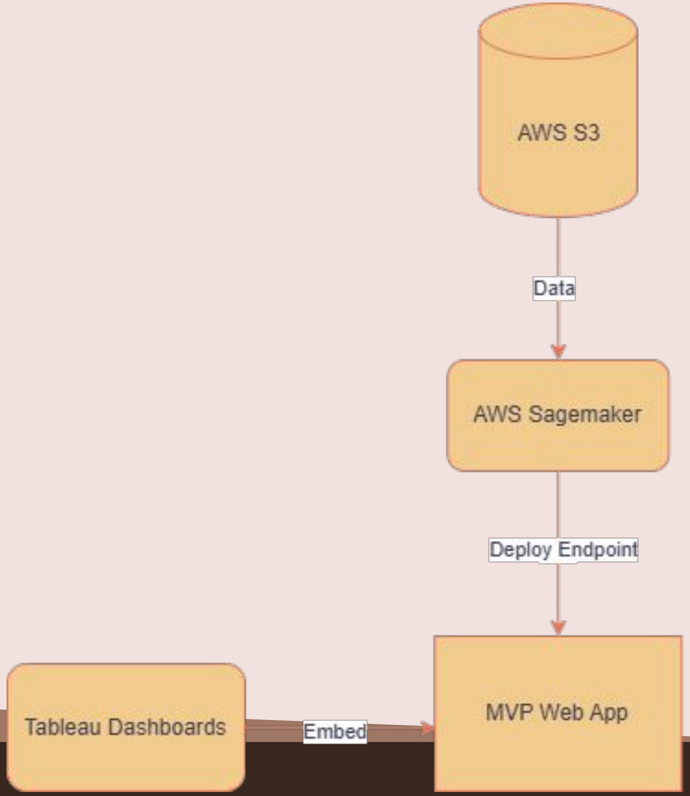


State	Wildfire Cost (Billions)	Percentage of U.S. Total	Annual Cost per Capita (Dollars)	Annual Cost per Square Km (Dollars)
California	$87.29	73%	$53	$4,902
Colorado	$5.27	4%	$22	$465
Oregon	$4.97	4%	$28	$464
Montana	$2.91	2%	$65	$182
Texas	$2.85	2%	$2	$97
Idaho	$2.85	2%	$39	$313
Washington	$2.51	2%	$8	$324
Alaska	$2.03	2%	$66	$28
Tennessee	$1.64	1%	$6	$357
New Mexico	$1.42	1%	$16	$108
Utah	$1.25	1%	$9	$135
Arizona	$1.17	1%	$4	$95
Nevada	$1.11	1%	$9	$92
Wyoming	$0.98	1%	$40	$92
Alabama	$0.66	1%	$3	$116
Oklahoma	$0.31	0%	$2	$41
Florida	$0.28	0%	$0	$39
Georgia	$0.27	0%	$1	$41
South Dakota	$0.10	0%	$3	$11
Minnesota	$0.09	0%	$0	$10
North Carolina	$0.08	0%	$0	$14
Nebraska	$0.05	0%	$1	$6
Mississippi	$0.04	0%	$0	$7
North Dakota	$0.01	0%	$0	$2
United States	$120.13	100%	$9	$21

Appendix: Updates 9/30/24

- Continued literary review
  - Research paper: “Catastrophe Models for Wildfire Mitigation: Quantifying Credits and Benefits to Homeowners and Communities” [link]
- Able to find more supporting evidence for the impact of our project
- Hone in on our modeling approach

Appendix: Overall Architecture



Appendix: Modeling

Initial models tested and considerations:

- Linear Regression: helps identify basic patterns and relationships between the data, find highly statistically significant correlations between damage level and other features
- Random Forest Classifier, LightGBM, & XGBoost: perform well for data that includes a variety of features that are in mixed categories such as ordinal, categorical, and numerical
- Recurrent Neural Network with LSTM: handle sequential data (multiple years in our dataset)
- K-Nearest Neighbors: naturally suited for geographic data, easy to explain and interpret for users
- Support Vector Machine: robust and handles noise well, flexible and takes in "mixed" types of data

In-depth evaluation of these models in following slides

Appendix: Modeling - Part 1

Baseline	Majority Class ● 4 - greater than 50% damage ● 57.9% accuracy
Linear Regression	Linear: ● RMSE: 1.11, 60+ features → feature selection during optimization phase ● 39% accuracy ● Outcome variable ordinal from 0-4, larger values representing higher % of damage ● Notable features include single residence, roof construction, and fencing
Recurrent Neural Network with LSTM	Train Accuracy: 36% Test Accuracy: 36% Poor performance: Does not perform well, implies that there may not be a strong sequential (time based) relationship between the damage prediction and features perhaps because the level of damage a wildfire can cause may vary
K-Nearest Neighbors	Train Accuracy: 81% Test Accuracy: 77%

Appendix: Model Improvements

Linear Regression

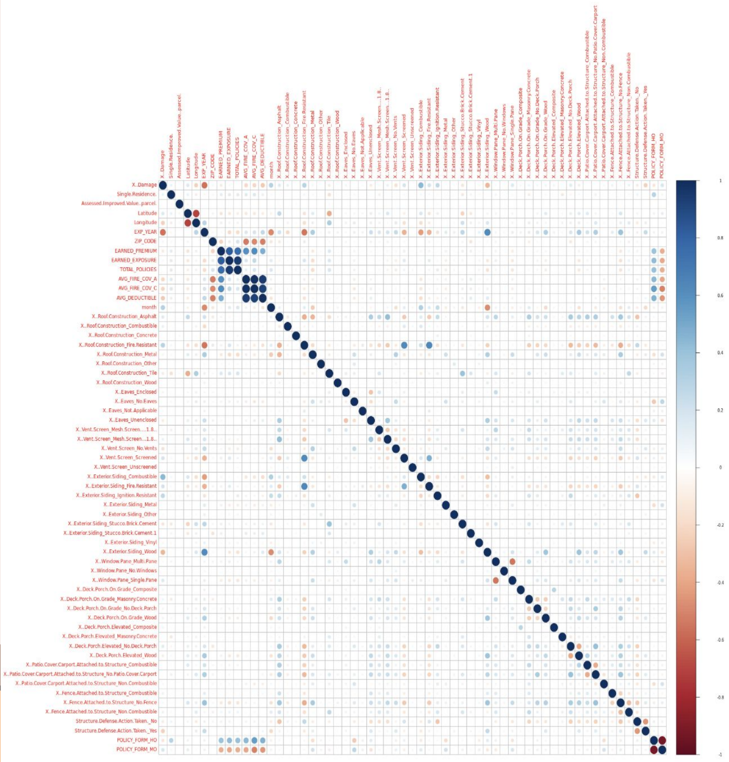
- Process:
  - dropping statistically insignificant features at alpha=0.05 level
  - dropping features that have low correlation (less than |0.1|)
  - iterative process completed 4 times
  - final product: dropped 40+ features and ended up with 18 final features
- Plotted Feature Correlation Matrix and Hierarchical Feature Heatmap
- Best accuracy throughout process: 39%
- Lowest RMSE: 1.11
- Not a good final model

Linear Regression Results

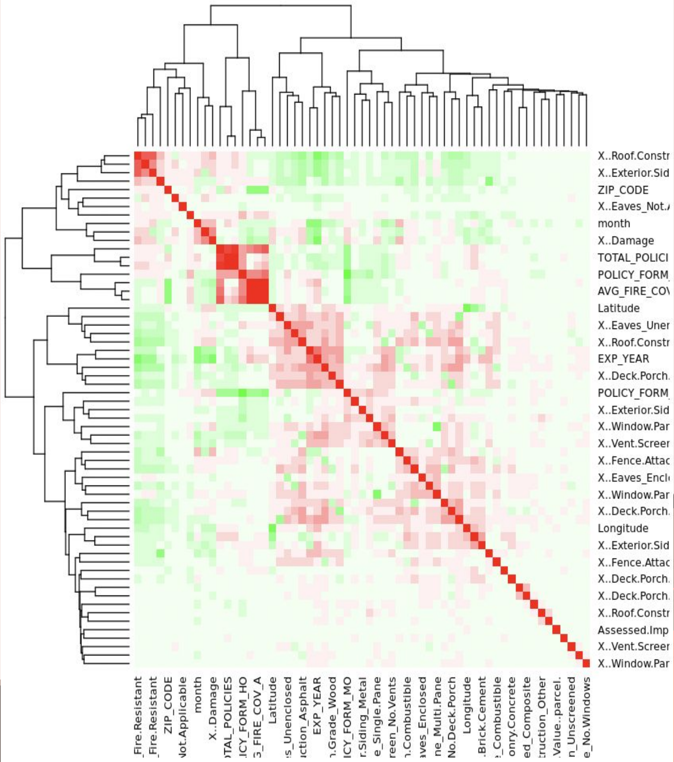
	Dependent variable:
	X..Damage
Longitude	-0.056*** (0.006)
EXP_YEAR	-0.357*** (0.008)
EARNED_PREMIUM	-0.00000*** (0.000)
month	-0.149*** (0.007)
X..Roof.Construction_Asphalt	0.450*** (0.021)
X..Roof.Construction_Fire.Resistant	0.227*** (0.037)
X..Roof.Construction_Tile	-0.313*** (0.043)
X..Eaves_Unenclosed	-0.408*** (0.018)
X..Vent.Screen_No.Vents	-0.311*** (0.027)
X..Vent.Screen_Screened	-0.138*** (0.035)
X..Exterior.Siding_Combustible	1.661*** (0.030)
X..Exterior.Siding_Fire.Resistant	1.600*** (0.044)
X..Exterior.Siding_Ignition.Resistant	1.708*** (0.034)
X..Exterior.Siding_Stucco.Brick.Cement	0.575*** (0.033)
X..Window.Pane_Single.Pane	0.613*** (0.018)
X..Deck.Porch.Elevated_No.Deck.Porch	0.410*** (0.019)
Structure.Defense.Action.Taken._No	-0.365*** (0.023)
Structure.Defense.Action.Taken._Yes	-1.108*** (0.039)
Constant	716.119*** (16.125)
Observations	28,070
R2	0.551
Adjusted R2	0.551
Residual Std. Error	1.289 (df = 28051)
F Statistic	1,911.551*** (df = 18; 28051)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix: Model Improvements

Feature Correlation Matrix



Hierarchical Feature Heatmap



Appendix: Modeling - Part 2

Support Vector Machine	Train Accuracy: 88% Test Accuracy: 87% Missing minority classes: The model is performing well on the majority classes (0 and 4), but is unable to predict the minority classes (1, 2, and 3) correctly
Random Forest	Train Accuracy: 99% Test Accuracy: 93%
LightGBM	Train accuracy: 97.2% Test accuracy: 93.2% Imbalance Problem: The model performs well on the majority classes (0 and 4) but struggles with the minority classes (like 1, 2, 3)
XGBoost	Train Accuracy: 97% Test Accuracy: 93.3% Imbalance Problem: The model performs well on the majority classes (0 and 4) but struggles with the minority classes (like 1, 2, 3)

Appendix: Modeling Problem - Class Imbalance

Best performing model:

XGBoost

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.97	0.95	2954
1	0.63	0.42	0.50	317
2	0.26	0.12	0.17	73
3	0.19	0.11	0.14	27
4	0.96	0.96	0.96	4625
accuracy			0.93	7996
macro avg	0.59	0.52	0.54	7996
weighted avg	0.92	0.93	0.93	7996

Appendix: Modeling - Optimization Part 1

Final model choice - XGBoost: best model performance thus far

- SMOTE undersampling minority classes and oversampling majority classes in attempt to resolve class imbalance
- Random gridsearch for optimized hyperparameters
- Scaled vs unscaled data

Scaled

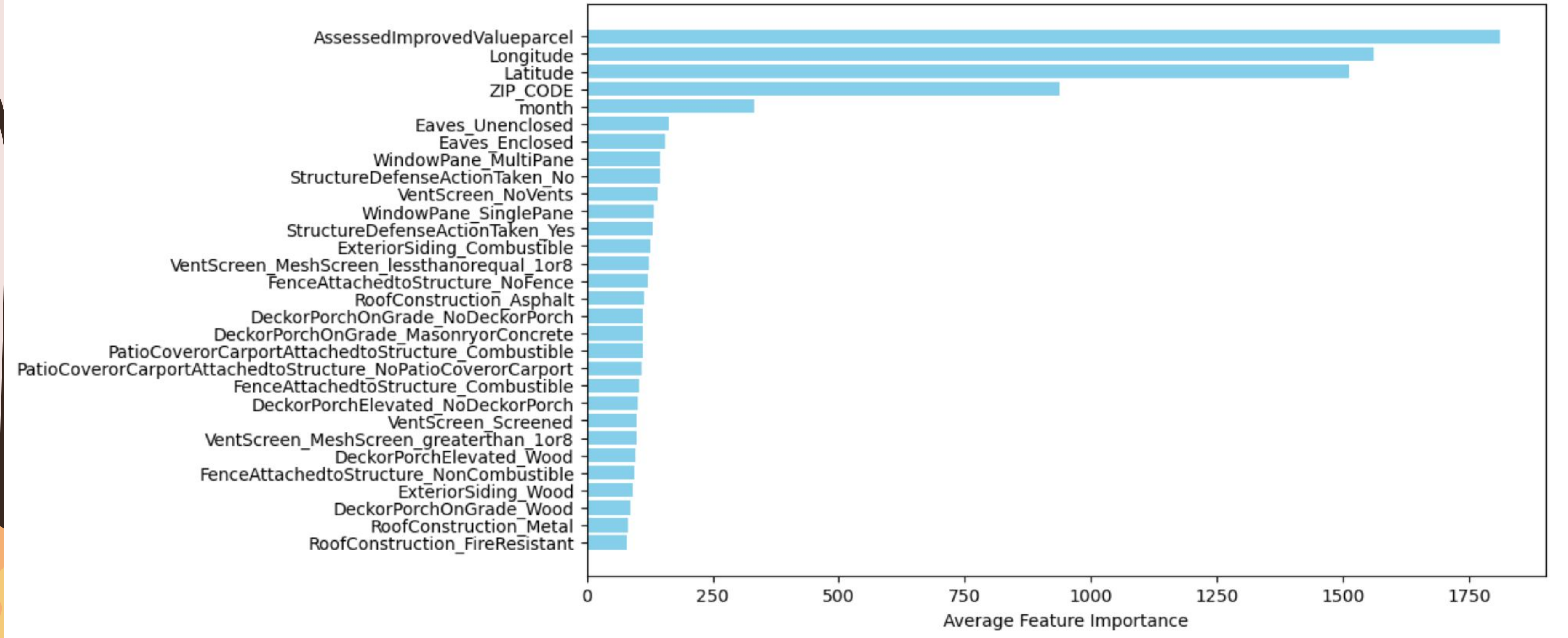
	precision	recall	f1-score	support
0	0.93	0.98	0.95	2954
1	0.56	0.45	0.50	317
2	0.25	0.23	0.24	73
3	0.14	0.11	0.12	27
4	0.97	0.95	0.96	4625
accuracy			0.93	7996
macro avg	0.57	0.54	0.55	7996
weighted avg	0.93	0.93	0.93	7996

Non-Scaled

	precision	recall	f1-score	support
0	0.92	0.97	0.95	2954
1	0.54	0.46	0.50	317
2	0.29	0.22	0.25	73
3	0.13	0.11	0.12	27
4	0.96	0.94	0.95	4625
accuracy			0.92	7996
macro avg	0.57	0.54	0.55	7996
weighted avg	0.92	0.92	0.92	7996

Appendix: Feature Importance

> 60 features currently in model (due to categories being one hot encoded), top features plotted



Appendix: Modeling - Optimization Part 2

- Previously noted that class imbalance issue remained with SMOTE, based on capstone instructor feedback we decided to collapse minority classes into one category
- This made sense as users most likely do not care about a 10% versus 20% difference in damage but rather would prefer a simplified output which we can achieve
- New classes are 0 for no damage, 1 for mild damage (greater than 0% to less than or equal to 50% damage) and 2 for moderate damage (greater than 50% damage)
- Reduction of features by removing few at a time and monitoring performance - 30 features kept led to optimal performance with minimal accuracy decreases

Final choices

- Non-scaled data fits better as majority of features one hot encoded, don't necessarily fit classic normalization and may impact overall data pipeline and interpretability of results
- Reduced feature classes
- 30 features total chosen for XGBoost Model
- Key parameters: L1 & L2 regularization to help reduce overfit,, 100 trees (n-estimators) with max depth 10 each tree, 60% of features can be used for each tree max

Appendix: Final Model Evaluation

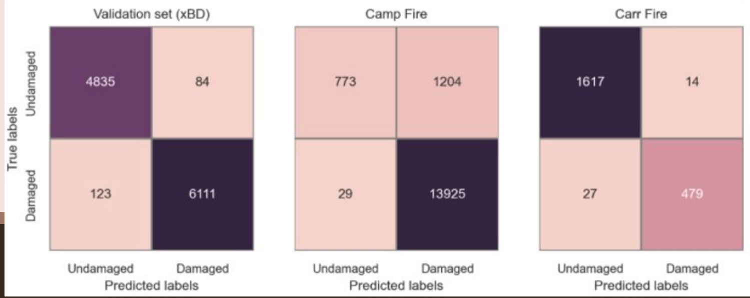
Compare evaluation metrics

- Accuracy, and F1 score of each class, which accounts for both recall and precision
- Similar study: 92% and 98% accuracy on two test sets with a F1 score of 0.96

	precision	recall	f1-score	support
0	0.94	0.97	0.95	2954
1	0.83	0.57	0.68	417
2	0.96	0.97	0.96	4625
accuracy			0.94	7996
macro avg	0.91	0.83	0.86	7996
weighted avg	0.94	0.94	0.94	7996

Table 1. Evaluation metrics of the model on each dataset. Refer to sec. 3.1 for definitions of metrics.

Dataset	Accuracy	Precision	Recall	F1 Score
xBD wildfires (validation set)	0.98	0.99	0.98	0.98
Camp Fire (Test set 1)	0.92	0.92	0.99	0.96
Carr Fire (Test set 2)	0.98	0.97	0.95	0.96



Appendix: Conclusion

Main Takeaways: Our efforts on this data science project have proven fruitful!

- **Great team dynamics** and apportioning of semester-long project tasks and checkpoints alongside learning how to use new tools
- Extensive data searching, merging, cleaning, etc. to create a **unique new dataset**
- Implemented **multiple ML models** with repeated efforts in **optimization**
- Familiarity with **AWS, Sagemaker, and S3 bucket** functionalities
- **Website creation** and hosting our solution

End Result: A working prediction tool for users and interactive dashboards showing different insurance metrics and damages across California!

- Intersection of **advanced machine learning** and **impactful decision-making**
- Helping homeowners understand how **specific property features influence risk and damage,** enabling them to make targeted improvements that **save lives, protect properties**, and **reduce insurance costs**
- Providing **transparent insurance insights and fire severity insights** for current and prospective California residents, regardless of homeowner status

Appendix: Note

At the end of the final class, via email, please share with instructors final presentation slides, web deliverable link, and any supplemental materials within a day of the live session.