WILDFIRE DETECTION USING SATELLITE IMAGING

Detect wildfires and empower insurance companies for improved risk management

Team Wildfire Prophet - Liang Li, Iris Lew, Ivy Chan

University of California, Berkeley | School of Information

DATASCI 210 Capstone - Fall 2023

Final Presentation

December 14, 2023

Berkeley
UNIVERSITY OF CALIFORNIA

Team Members

Liang Li

- Machine Learning Engineer
- Data Scientist
- Product Manager

Ivy Chan

- Data Engineer
- Web Developer
- Project Manager

Iris Lew

- Data Scientist
- Data Analyst
- Testing Manager

Berkeley
UNIVERSITY OF CALIFORNIA

AGENDA

1. Problem & Impact
2. SME Interviews Key Takeaways
3. Market Opportunity & Target User
4. Answer User's Questions
5. MVP Demo
6. Data Pipeline & Feature Extraction
7. Models & Generalizability
8. Model Improvements
9. Challenges & Solutions
10. Conclusion

Berkeley
UNIVERSITY OF CALIFORNIA

Problem & Impact

- Increasing the frequency and severity of wildfires as a global threat.
- Better use of high-resolution satellite imaging data for wildfire detection.
- Alarming statistics reveal rising fire-related deaths and extensive property damage.
- Wildfires result in a surge of insurance claims for property damage and losses.
- Insurance companies may experience significant financial losses due to wildfire-related claims.


Rim Fire, California. August, 2013


Camp Fire, California. November, 2018


Paradise Fire, California. November, 2018

Berkeley
UNIVERSITY OF CALIFORNIA

SME Interview Key Takeaways

- Narrow down to the scope
- Many services already use satellite imagery to help inform them of wildfire risk, but cloud cover is always a challenge
- Adding weather is difficult
- Excess vegetation is a contributing factor to wildfires
- Climate change is a driving force of wildfires

Thanks to Meer Wu, Jenna Morabito, and Professor Matthew Potts!

Market Opportunities

Death rate	US fire deaths per million population rose by 18% over a decade, reaching 13.0 in 2021, up 14% from 2020's 11.4.
Land size	Fires impact 4 million sq. km of Earth's land annually, about half the size of the U.S.
Market opportunity	The forest wildfire detection market is set to grow at a 10.4% CAGR, reaching USD 5.1 billion by 2028

Target User - Insurance Companies

Wildfire detection	Accurate wildfire detection and impact assessment.
Underwriting process	Improves underwriting processes and enhances client protection.
Proactive Management	Empowers proactive management and risk reduction.

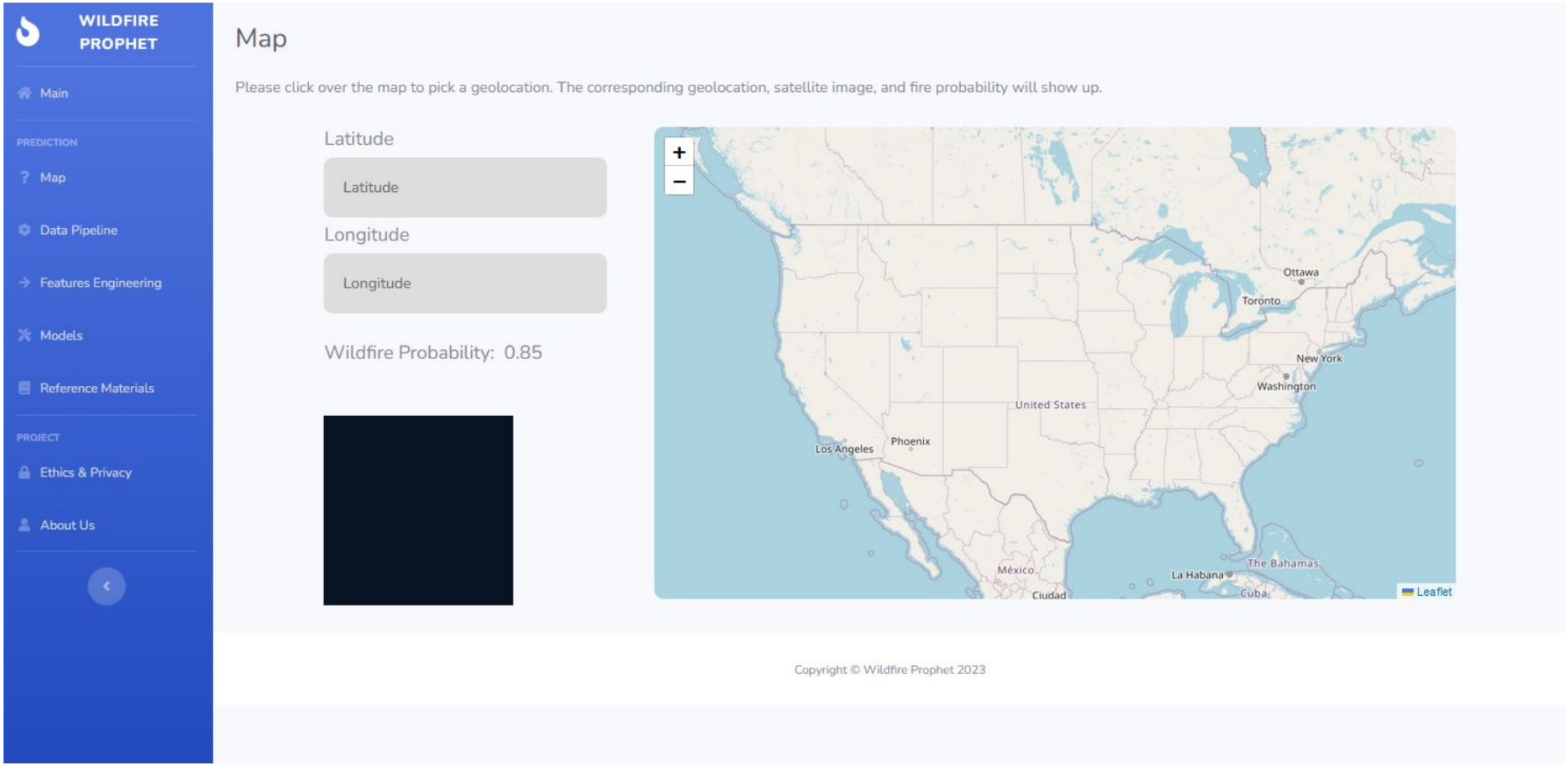
Answer User's Questions

1. What is the wildfire probability given the location?

2. Should the business expand to this location/region/county given the wildfire probability?

MVP Demo

MVP Demo



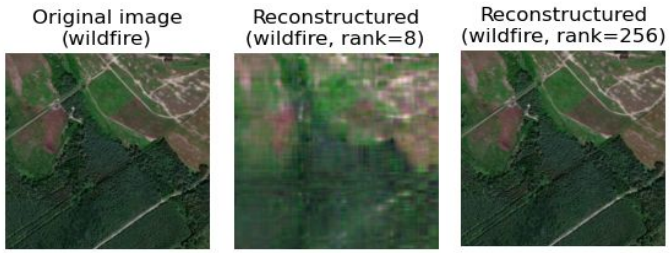
Please click the image to open the website.

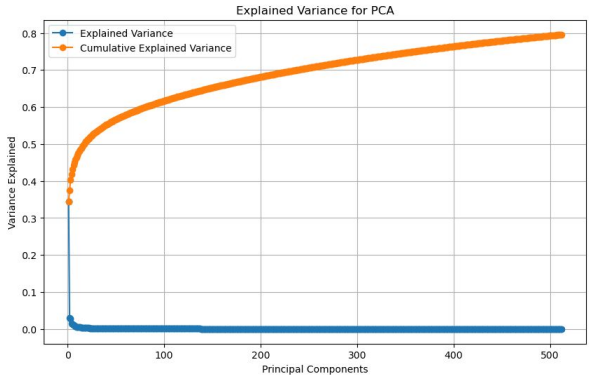
Berkeley
UNIVERSITY OF CALIFORNIA

Data Pipeline & Feature Extraction

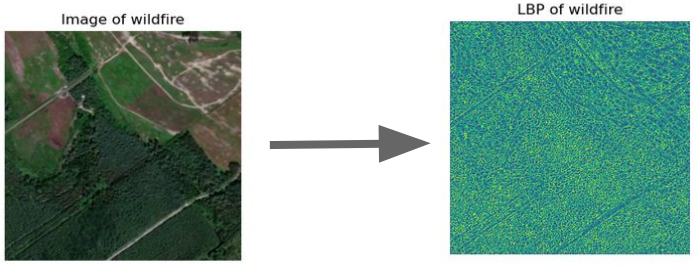
Data Pipeline (Canada)

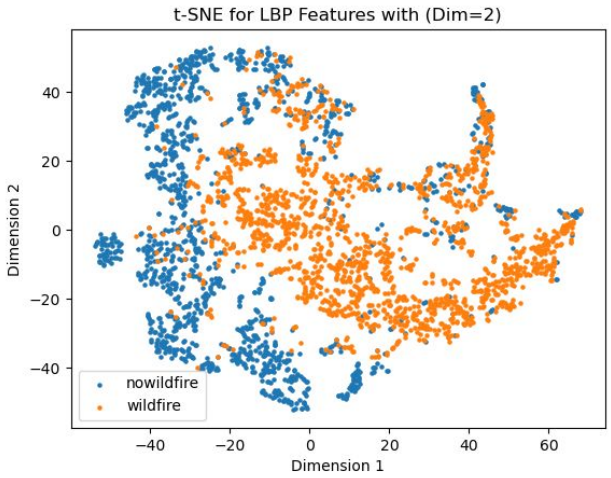
- Download 42,850 rows from Kaggle*
- PCA for raw pixel feature
  - Need 512+ principal components to capture 80%+ explained variance
  - Compression rate ~1.02, i.e., almost no compression





- Feature extraction
  - HOG feature
  - Edge feature
  - LBP feature (recommended, 367,500 dims ⇒ 9 dims)
    - radius = 1 (radius of the LBP circle)
    - n_points = 8 (#points to sample on the LBP circle)
    - method = "uniform"

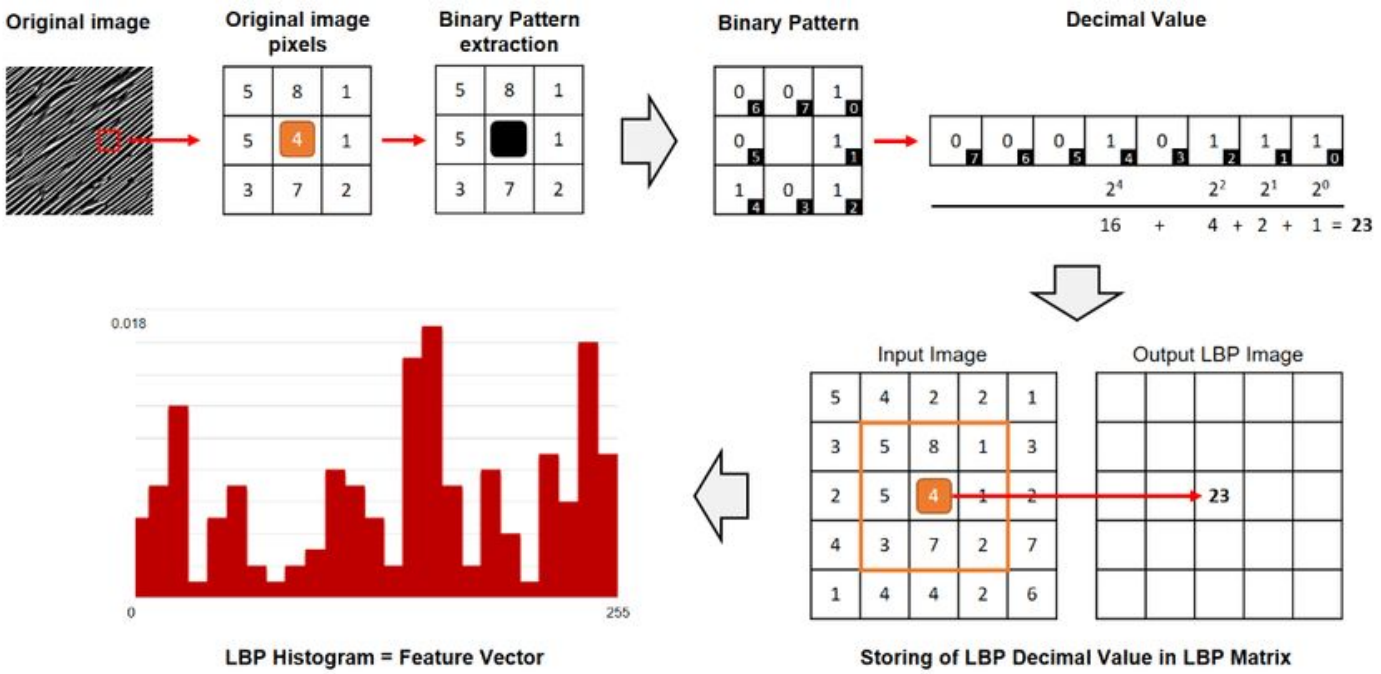




*https://www.kaggle.com/datasets/abdelghaniaaba/wildfire-prediction-dataset

LBP Feature

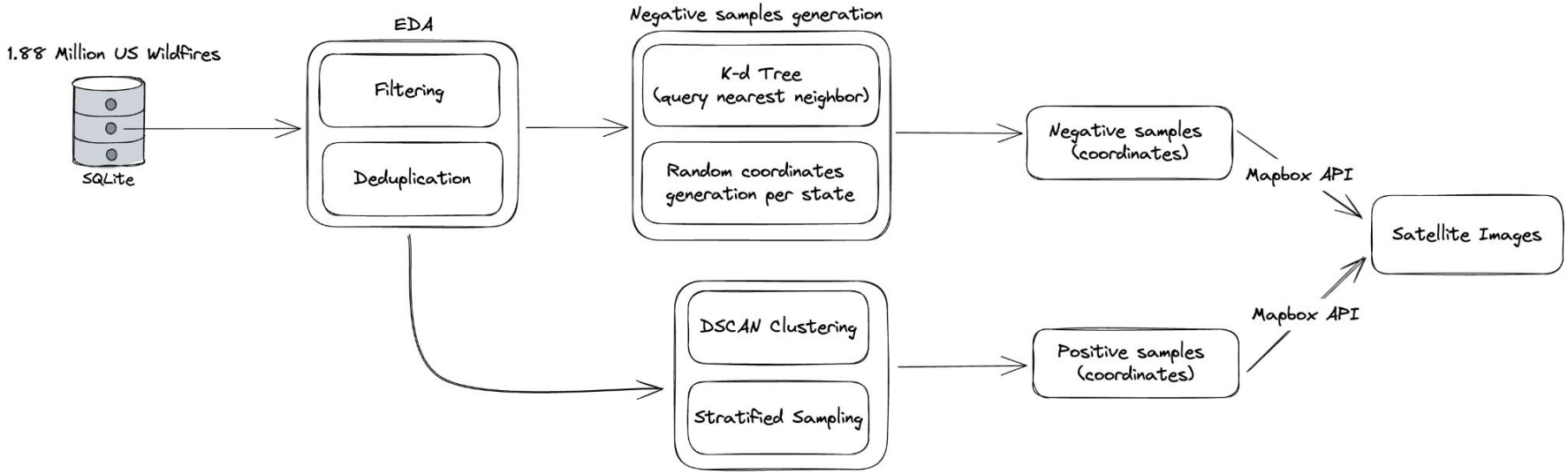
- LBP (Local Binary Pattern) feature extraction flow



Source: Müller, Martin & Britz, Dominik & Ulrich, Laura & Staudt, Thorsten & Mücklich, Frank. (2020). Classification of Bainitic Structures Using Textural Parameters and Machine Learning Techniques. Metals. 10. 630. 10.3390/met10050630.

Data Pipeline (US)

- Download 1.88M US wildfire data from Kaggle*
- EDA for positive samples
  - filtering, deduplication, DBSCAN clustering, stratified sampling
- Data generation for negative samples
  - k-d tree, random coordinates generation
- Download satellite images through Mapbox API

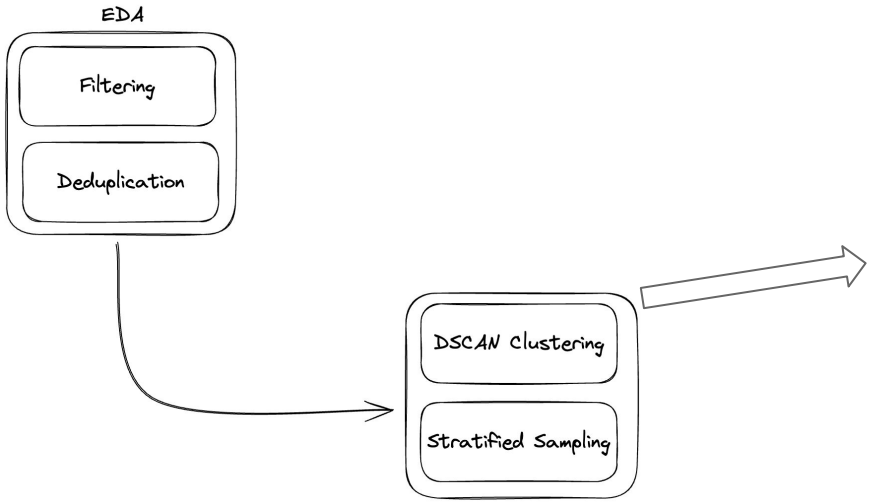


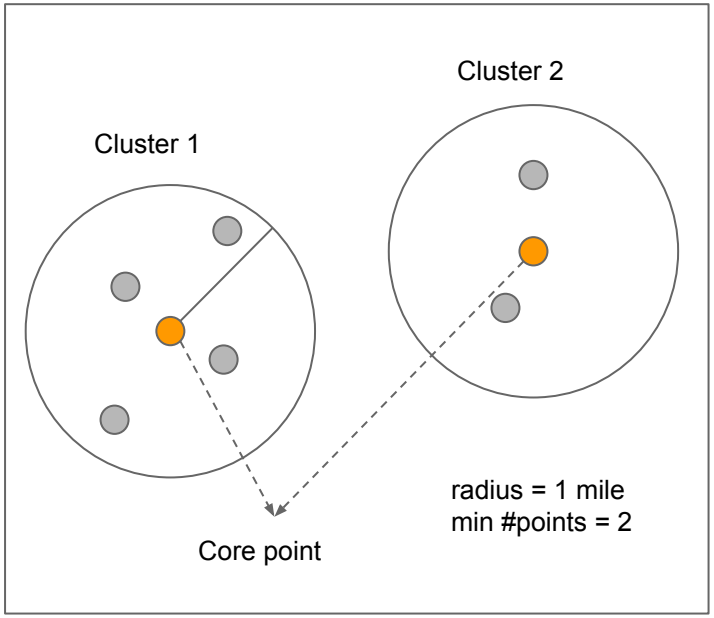
Berkeley
UNIVERSITY OF CALIFORNIA

*https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires

Data Pipeline (US, Clustering)

- Motivation
  - Optimize satellite image acquisition and processing for adjacent fire events
- Clustering
  - Cluster closely located coordinates within a 1-mile radius
  - Employ DBSCAN for spatial clustering
  - Assign a weight of N based on the cluster size

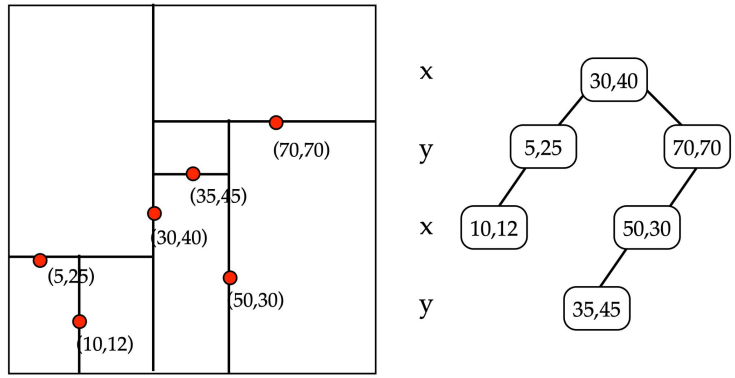




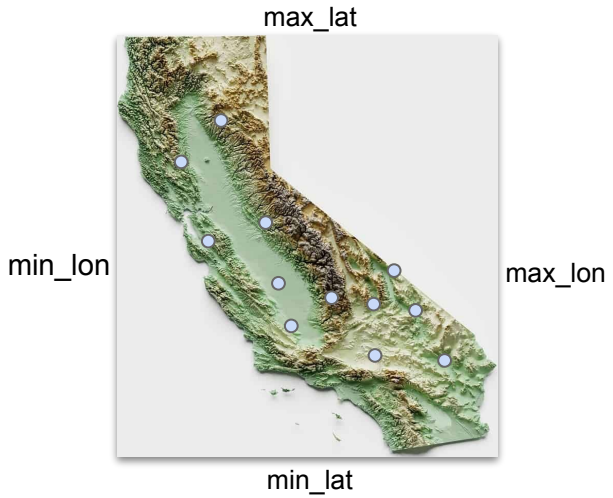
DBSCAN Clustering

Negative Samples Generation

- Goal and challenge
  - Goal: generate coordinates without wildfires as negative samples
  - Challenge: no collision/overlap with existing 1.88M US wildfire coordinates
- Proposed solution
  - Build k-d tree for existing 1.88M wildfire coordinates
  - Query the nearest neighbor for new coordinate (in O(logn) time)
- New coordinates generation
  - Step 1: for each existing coordinate, randomly generate a coordinate in the same state
  - Step 2: compute the distance between new coordinate and its nearest neighbor
  - Step 3: break if the distance is within 2~50 miles range, otherwise repeat step 1-3



2-d tree (binary search tree with 2 dim for each node)



Negative samples generation using California as example

Berkeley
UNIVERSITY OF CALIFORNIA

Models & Generalizability

Machine Learning Models (Canada)

Model	Precision		Recall		F1 Score		Accuracy	Notes
	wildfire	nowildfire	wildfire	nowildfire	wildfire	nowildfire		
LR + raw pixel	0.82	0.82	0.86	0.76	0.84	0.79	81.75%	Baseline
LR + LBP	0.81	0.80	0.85	0.75	0.83	0.77	80.32%	
LR + HOG	0.81	0.80	0.85	0.76	0.83	0.78	80.79%	
SVM + LBP	0.92	0.92	0.94	0.90	0.93	0.91	92.06%	
SVM + ResNet	0.91	0.92	0.94	0.89	0.93	0.91	91.75%	
NN + LBP	0.92	0.92	0.94	0.90	0.93	0.91	92.38%	Best

Generalizability

- A hybrid CV model trained with both Canada and US data
  - Overall accuracy 83%
    - ‘nowildfire’ class: precision 86%, recall 77%, f1-score 81%
    - ‘wildfire’ class: precision 80%, recall 88%, f1-score 84%
  - Deployed to production
- The quality of Canada dataset is higher than US dataset

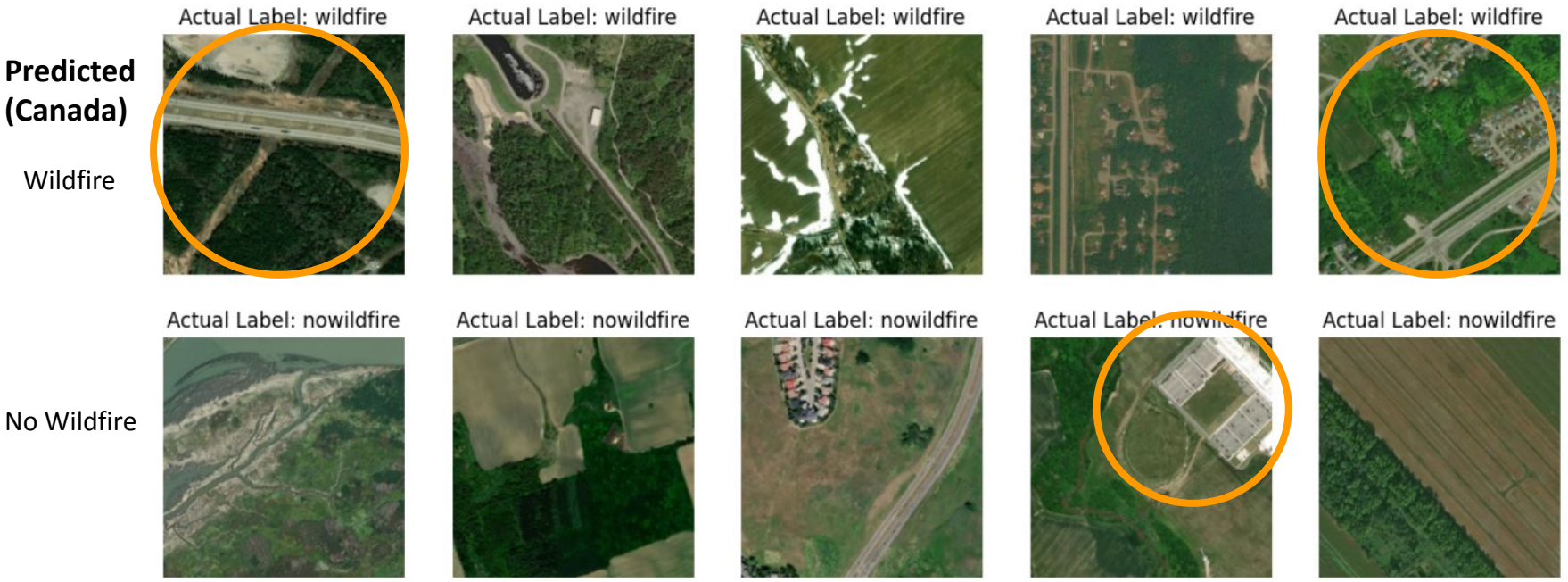
ID	Training Data	Samples	Vectorized	Model	Accuracy		
					Canada	US	Overall
1	Canada	30K	Yes	NN+LBP	93%	63% (65% N.E. US)	67%
2	US	20K	Yes	NN+LBP	48%	74%	66%
3	Canada + US	30K+20K	No	NN+LBP	92%	67%	83%
4	Canada + US	30K+30K	No	NN+LBP	N/A	N/A	79%

Berkeley
UNIVERSITY OF CALIFORNIA

Model Improvements?

Improving on the Baseline Model?

Would eliminating urban coordinates lead to an improved model?



Removing Urban Points Improve Accuracy?

“A fire that is burning strongly and out of control on an area of grass or bushes in the countryside” - Cambridge dictionary’s definition of a wildfire.

There seems to be a lot of buildings when we sample the satellite images from when the model does do a good job at predicting.

Dataset	Correlation Coefficient between Urban and Wildfire column	Correlation Coefficient between Urban and Predicted column
Canada Train	r=0.81	
Canada Validation	r=0.80	r=0.72
US Train	r=-0.09	
US Validation	r=-0.06	r=-0.18

Identify which points are in urban areas* and eliminate those points.

Berkeley
UNIVERSITY OF CALIFORNIA

*Canada: https://www.oecd.org/cfe/regionaldevelopment/functionalurbanareasbycountry.htm
*US: https://www.census.gov/geographies/mapping-files/2018/geo/carto-boundary-file.html

Canada Dataset Predicts Urbanism

Training Data (NN+LBP model)	**Accuracy (on wildfire column)**			
	Canada (w/ Urban): 6,300 rows, 55.2% wildfire	**Canada** (w/o Urban): 3,192 rows 16.1% wildfire	**US** (w/ Urban): 3,806 rows 50.0% wildfire	**US** (w/o Urban): 3,642 rows 49.3% wildfire
Canada (w/ Urban): 30,250 rows, 52.1% wildfire	**93.0%**	93.2%		
Canada (w/o Urban): 14,099 rows 95.1% wildfire	**57.1%**	**16.9%**		
US (w/ Urban): 12,412 rows 38.7% wildfire			71.5%	71.9%
US (w/o Urban) 11,880 rows 37.7% wildfire			71.9%	72.4%

Berkeley
UNIVERSITY OF CALIFORNIA

Struggles With the Canada Dataset

Training Data (NN+LBP model)	Accuracy (on wildfire column)	
	Canada (w/ Urban): 6,300 rows, 55.2% wildfire	**Canada** (w/o Urban): 3,192 rows 16.1% wildfire
Canada (w/ Urban): 30,250 rows, 52.1% wildfire	93.0%	93.2%
Canada (w/o Urban): 14,099 rows 95.1% wildfire	57.1%	16.9%

Accurately classified
- 2,666 no wildfire
- 3,196 wildfire

Inaccurately classified
- 154 no wildfire images
- 284 wildfire images

Accurately classified
- 124 no wildfire
- 3,474 wildfire

Inaccurately classified
- 2,696 no wildfire images
- 6 wildfire images

Accurately classified
- 27 no wildfire
- 514 wildfire

Inaccurately classified
- 2,651 no wildfire images

Berkeley
UNIVERSITY OF CALIFORNIA

US Dataset Not Predicting Urbanism

Training Data (NN+LBP model)	Accuracy (on wildfire column)	
	US (w/ Urban): 3,806 rows 50.0% wildfire	**US** (w/o Urban): 3,642 rows 49.3% wildfire
US (w/ Urban): 12,412 rows 38.7% wildfire	**71.5%**	**71.9%**
US (w/o Urban) 11,880 rows 37.7% wildfire	**71.9%**	**72.4%**

Accurately classified
- 1,506 no wildfire
- 1,216 wildfire

Inaccurately classified
- 397 no wildfire images
- 687 wildfire images

Accurately classified
- 1,435 no wildfire
- 1,162 wildfire

Inaccurately classified
- 371 no wildfire images
- 634 wildfire images

Berkeley
UNIVERSITY OF CALIFORNIA

Predicting on the Urban Column Instead

Training Data (NN+LBP model)	**Accuracy (on urban column)**	
	Canada (w/ Urban): 6,300 rows, 49.3% urban	**US** (w/ Urban): 3,806 rows 4.3% urban
Canada (w/ Urban): 30,250 rows, 53.4% urban	**86.6%**	
US (w/ Urban): 12,412 rows 4.3% urban		**96.5%**

Correct Predictions

Instead of predicting urbanism, it may be predicting whether there's empty images

Canada

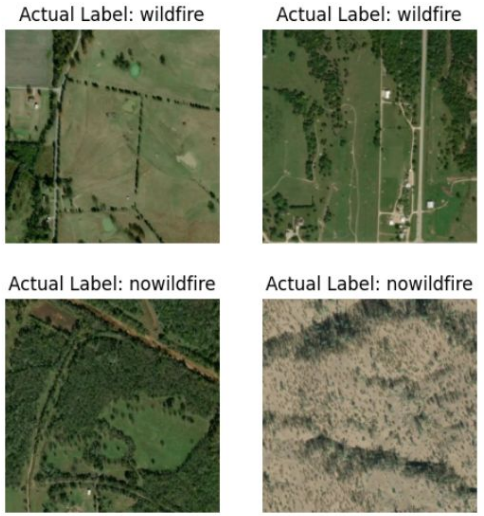
US

Incorrect Predictions

There's still some "urban" images?

Canada

US



Challenges & Solutions

Challenges and Solutions

Cloud Cover and Time	• Used MapBox instead of LandSat satellites to get images without clouds • Gave up time stamp
Generalizability	• Combine Canada and US datasets (leading to memory challenges and changing code)
Precision of GPS Coordinates	• Used DBSCAN to eliminate coordinates within a 1 mile radius • MapBox images are far enough away to not be able to make out distinct humans
Data Quality	• US had better data (but can evaluate further) • Need to re-evaluate Canada data

Mission

Detect wildfires and empower insurance companies for improved risk management

Berkeley
UNIVERSITY OF CALIFORNIA

Q & A

Acknowledgements

- Edwin Figueroa (DATASCI 231 Ethics and Privacy Expert)
- Meer Wu (Wildfire Rx)
- Jenna Morabito (Wildfire Rx)
- Professor Matthew Potts
- Joyce Shen
- Kira Wetzel
- Prahbu Narsina

Deliverables

- Project Website
- Project Website Demo
- Final Presentation Slide Deck
- ISchool Project Web Page

References

Links:

- https://www.insurance.ca.gov/01-consumers/200-wrr/DataAnalysisOnWildfiresAndInsurance.cfm
- https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires
- https://www.kaggle.com/datasets/abdelghaniaaba/wildfire-prediction-dataset
- https://www.mapbox.com/pricing/#static-images-api
- https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/kdtrees.pdf
- https://www.oecd.org/cfe/regionaldevelopment/functionalurbanareasbycountry.htm
- https://www.census.gov/geographies/mapping-files/2018/geo/carto-boundary-file.html
- https://calmatters.org/environment/2019/10/paradise-california-camp-fire-anniversary-cleanup/
- https://atlantic.ctvnews.ca/halifax-area-wildfires-caused-more-than-165-million-in-insured-damage-1.6467917

Images:

- Multiple images from Google Images
- https://earthobservatory.nasa.gov/images/81919/rim-fire-california
- https://www.theatlantic.com/photo/2018/11/camp-fire-ravages-paradise-california/575461/
- https://en.wikipedia.org/wiki/List_of_California_wildfires#/media/File:AERONET_La_Jolla.2007295.terra.250m.jpg
- https://www.creativefabrica.com/product/photo-image-icon/
- https://careersnews.ie/using-google-maps-in-education/
- https://upload.wikimedia.org/wikipedia/commons/b/b1/Camp_Fire_oli_2018312_Landsat.jpg
- https://a-z-animals.com/blog/discover-how-california-got-its-unique-shape/
- https://cliparting.com/wp-content/uploads/2016/06/Cloud-clip-art-images-free-clipart-images.png
- https://clipground.com/images/green-plant-clipart-8.jpg