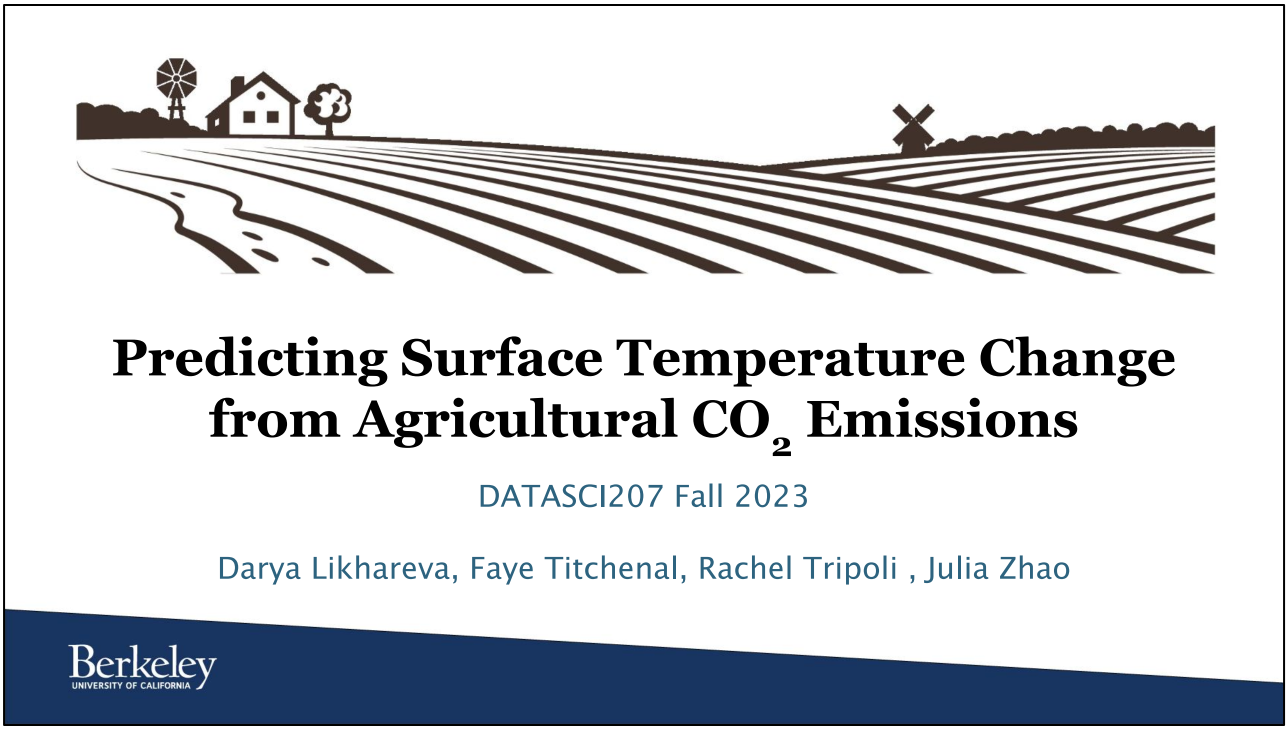
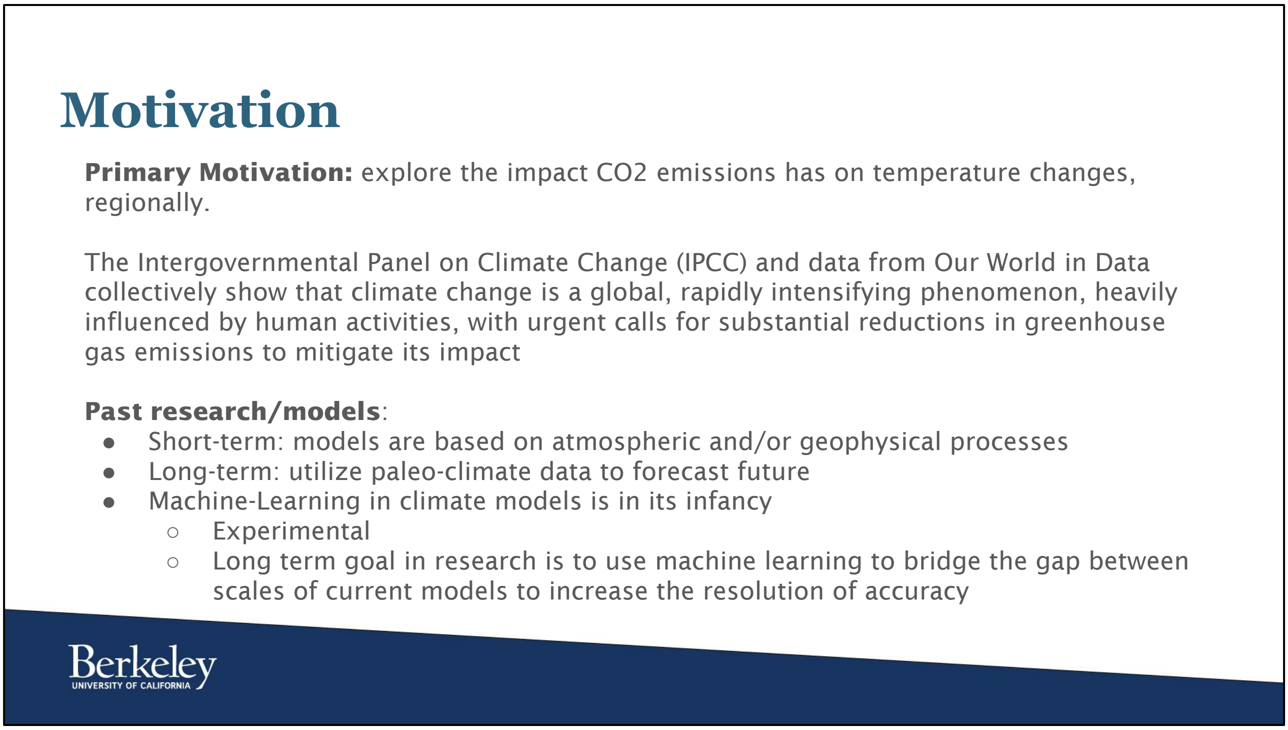


Predicting Surface Temperature Change from Agricultural CO_2 Emissions

DATASCI207 Fall 2023

Darya Likhareva, Faye Titchenal, Rachel Tripoli , Julia Zhao

Good Evening! Tonight, myself, Darya, Faye, and Julia will be giving a presentation on how we used machine learning to create a model that predicts surface temperature changes from agricultural co2 emissions.



Motivation

Primary Motivation: explore the impact CO2 emissions has on temperature changes, regionally.

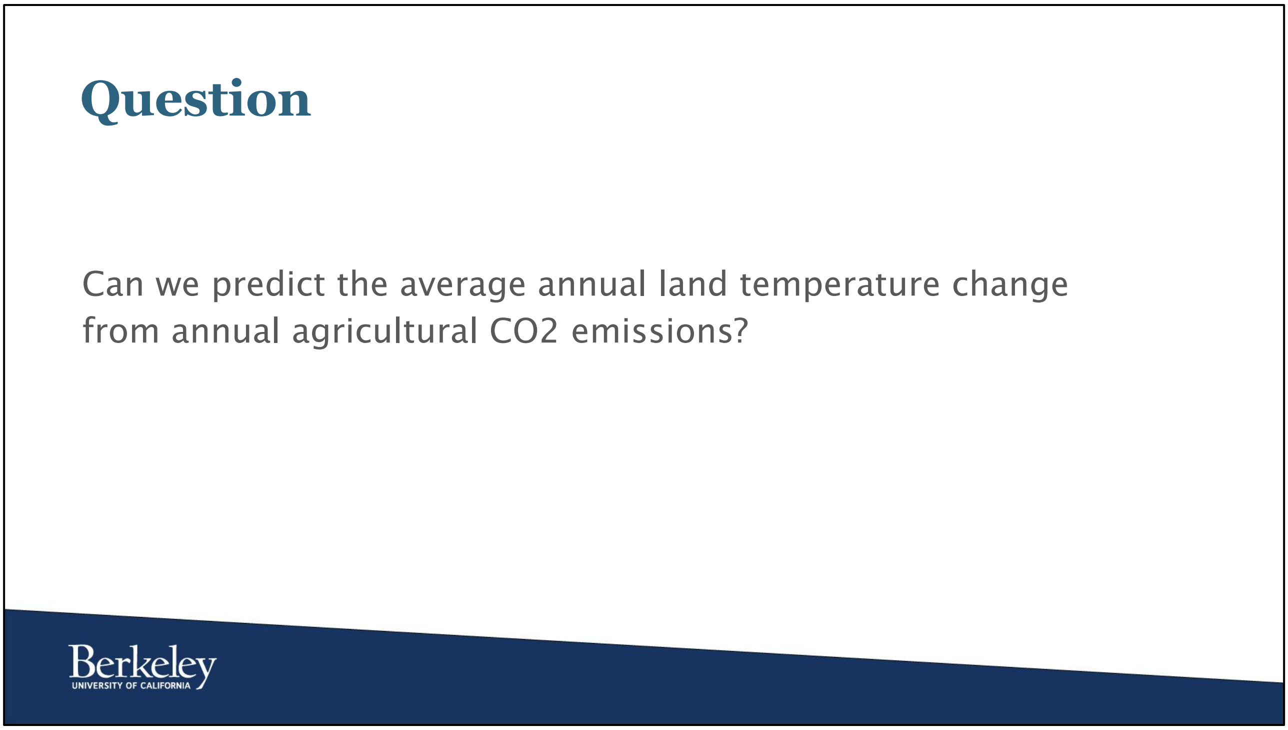
The Intergovernmental Panel on Climate Change (IPCC) and data from Our World in Data collectively show that climate change is a global, rapidly intensifying phenomenon, heavily influenced by human activities, with urgent calls for substantial reductions in greenhouse gas emissions to mitigate its impact

Past research/models:

- Short-term: models are based on atmospheric and/or geophysical processes
- Long-term: utilize paleo-climate data to forecast future
- Machine-Learning in climate models is in its infancy
    - Experimental
    - Long term goal in research is to use machine learning to bridge the gap between scales of current models to increase the resolution of accuracy

Berkeley
UNIVERSITY OF CALIFORNIA

Rising global temperatures and humanity's impact on such was our primary motivation to pick this project. Current models used in predicting how our global climate is shifting in the short term are based on atmospheric processes and, or in combination with, geophysical processes. Long-term climate models utilize paleoclimate data to make large scale forecasts. In regards to machine learning, implementing it in climate modeling is still very much in its infancy. Current research is in the experimental phase, with the NOAA citing a long term goal of using machine learning to bridge the gaps in the scales of current models to increase the resolution of their accuracies.

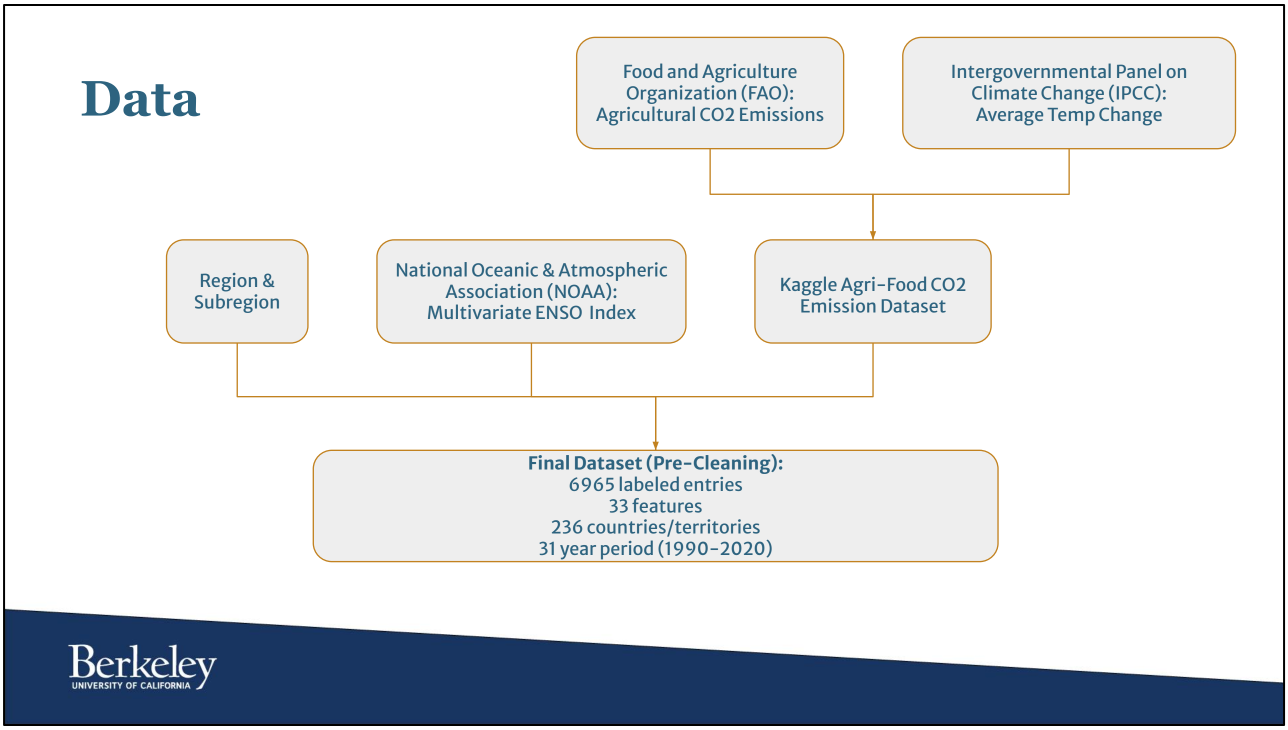


Question

Can we predict the average annual land temperature change from annual agricultural CO2 emissions?

Berkeley
UNIVERSITY OF CALIFORNIA

Our main research question is can we predict the average annual land temperature change from annual agricultural CO2 emissions?



Data

Food and Agriculture Organization (FAO): Agricultural CO2 Emissions

Intergovernmental Panel on Climate Change (IPCC): Average Temp Change

Region & Subregion

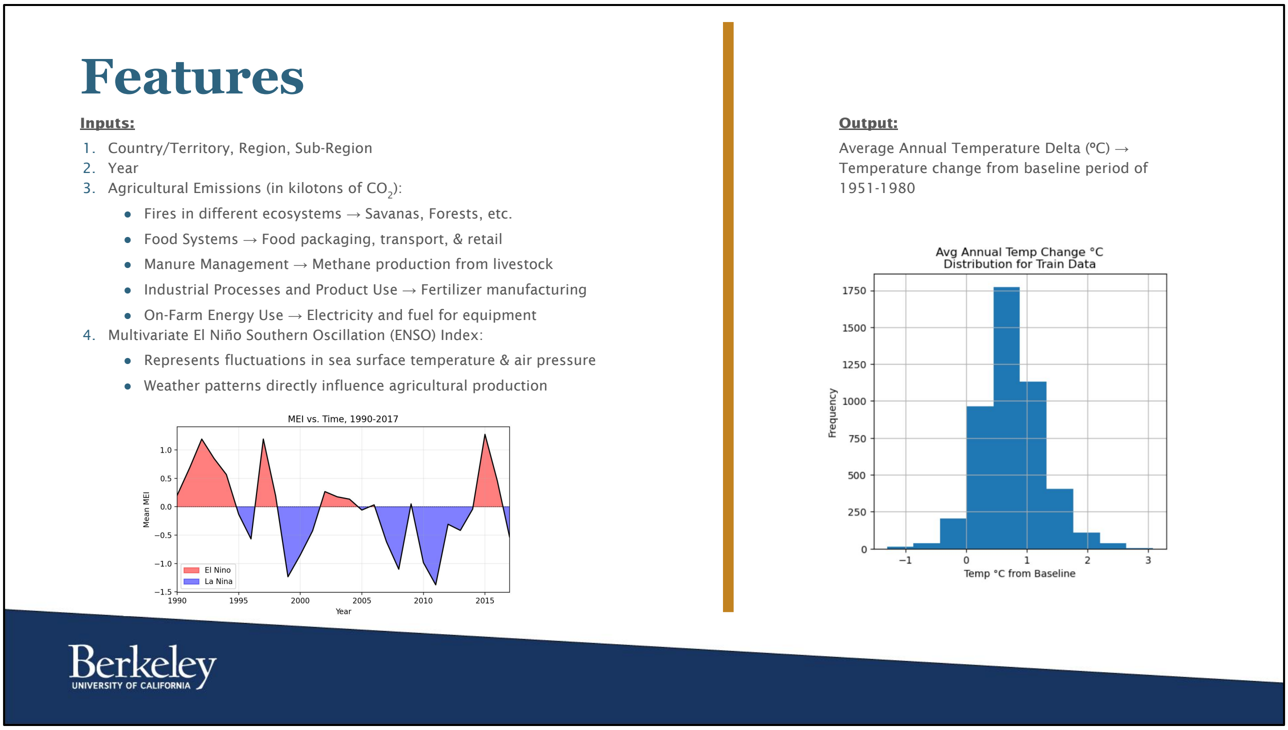
National Oceanic & Atmospheric Association (NOAA): Multivariate ENSO Index

Kaggle Agri-Food CO2 Emission Dataset

Final Dataset (Pre-Cleaning):
6965 labeled entries
33 features
236 countries/territories
31 year period (1990-2020)

Berkeley
UNIVERSITY OF CALIFORNIA

We sourced our raw data from Kaggle. This data is a compilation of agricultural CO2 emissions, from the FAO, and Average Temperature Change, from the IPCC. We added in outside data to segment the countries by region, more on that later, and for the Multivariate ENSO Index, or MEI. Pre-cleaning, our final dataset had just under 7000 labeled entries covering 236 countries over a 31 year period.



Features

Inputs:

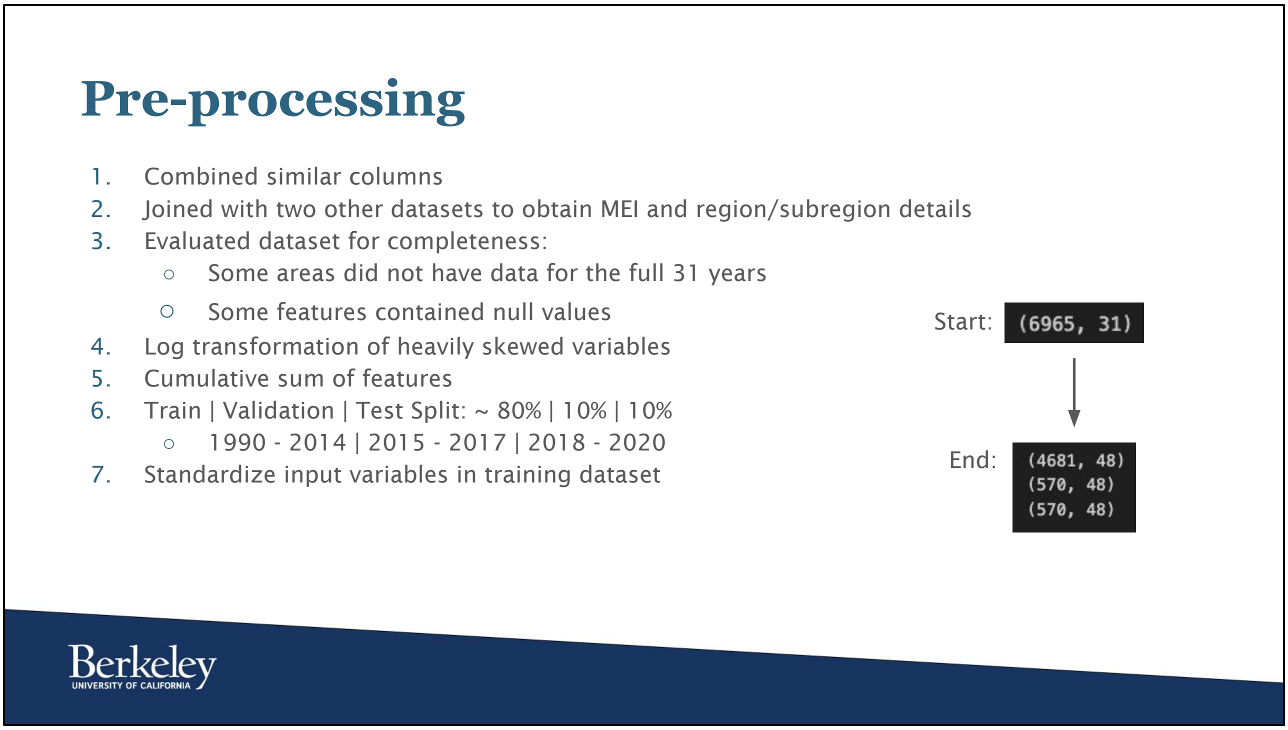
1. Country/Territory, Region, Sub-Region
2. Year
3. Agricultural Emissions (in kilotons of CO_2):
   - Fires in different ecosystems → Savanas, Forests, etc.
   - Food Systems → Food packaging, transport, & retail
   - Manure Management → Methane production from livestock
   - Industrial Processes and Product Use → Fertilizer manufacturing
   - On-Farm Energy Use → Electricity and fuel for equipment
4. Multivariate El Niño Southern Oscillation (ENSO) Index:
   - Represents fluctuations in sea surface temperature & air pressure
   - Weather patterns directly influence agricultural production

Output:

Average Annual Temperature Delta (ºC) → Temperature change from baseline period of 1951-1980

There were initially 33 features in our dataset. We added in region and sub-region features in order to dwindle down a potentially high number of one-hot-encoded variables with the countries. We then did a few merges of features, including merging all types of fires into one feature, and all aspects of food systems into one feature, manure management into another, industrial processes and product use into their own feature, and on-farm energy its own feature. Finally, as I previously mentioned, we added in this MEI, or multivariate El Nino Southern Oscillation index into our dataframe. MEI is a measure of the strength and variability of the El Nino phenomenon that occurs in the equatorial Pacific Ocean. This phenomenon has global effects on both weather and marine life, and both El Nino - represented as red in the plot - and it's counterpart La Nina - represented as blue - can have strong impacts on agricultural practices across the globe due to its correlation with the jet stream and temperature changes. As such, we elected to include it in our dataframe to ensure we were capturing a clearer picture of agricultural output and temperature changes as a whole. Our output variable is the average annual temperature delta from a baseline period of 1951 - 1980.

And now I'm going to pass it off Darya who will further our pre-processing discussion.



Pre-processing

1. Combined similar columns
2. Joined with two other datasets to obtain MEI and region/subregion details
3. Evaluated dataset for completeness:
    - Some areas did not have data for the full 31 years
    - Some features contained null values
4. Log transformation of heavily skewed variables
5. Cumulative sum of features
6. Train | Validation | Test Split: ~ 80% | 10% | 10%
    - 1990 - 2014 | 2015 - 2017 | 2018 - 2020
7. Standardize input variables in training dataset

Start:

```
(6965, 31)
```

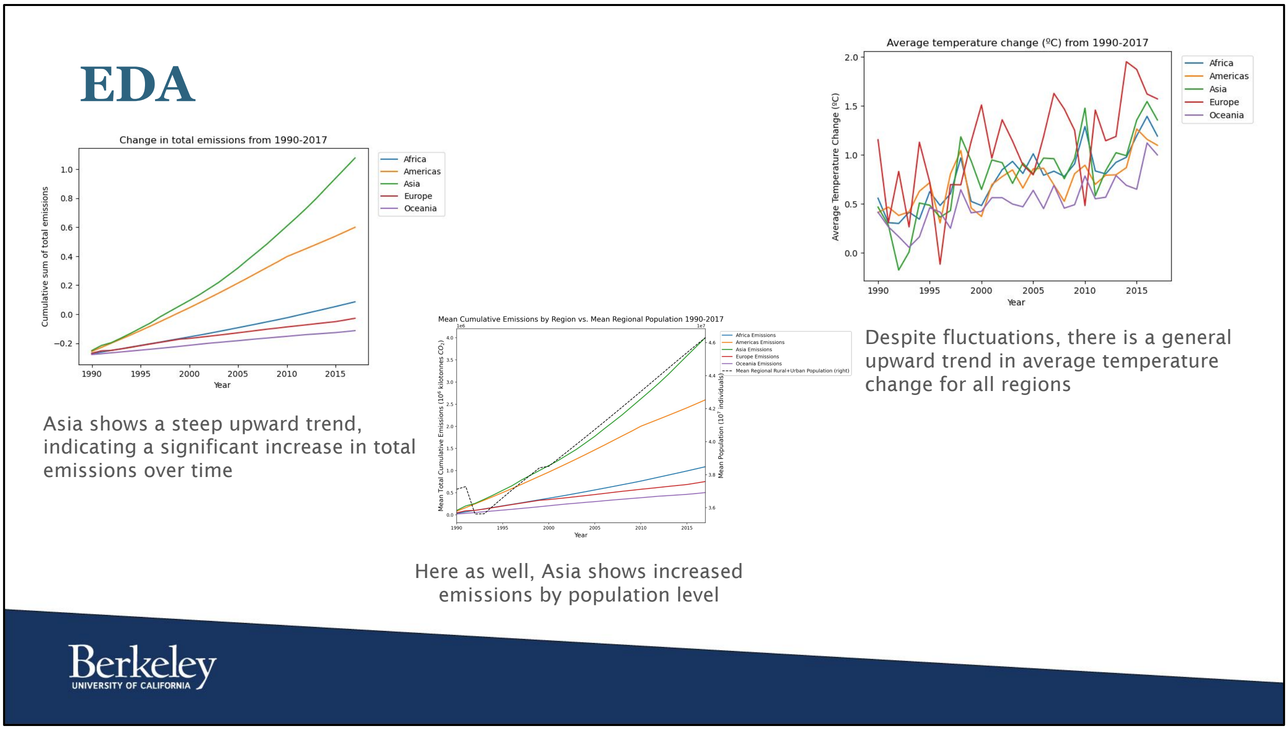
End:

```
(4681, 48)
(570, 48)
(570, 48)
```

Now I'll talk about how we prepared our data. First, we looked at and combine similar columns (like anything relating to fires or food systems) in our dataset to reduce redundancy. Then, we wanted to incorporate 2 additl. datasets - a MEI dataset and the Region/Subregion dataset. Since this is a data set that spans 31 years, we dropped some countries that don't exist anymore. Another key point was to make sure that our data set was as complete as possible, so a couple of issues that we ran into was that some areas did not have data for the entire 31 years and smaller countries contained only null values.We ended up addressing the missing data through a data imputation approach- through an average by region.

Next, we applied a log transformation to non-normally distributed variables. We also calculated the cumulative sums of some features over the years.

As expected, we split our data into three sets to support model development. 80% of our data is allocated to the training set (1990-2014), with 10% each for validation (2015-2017) and test (2018-2020) sets. Finally, we standardized input variables within the training dataset to ensure consistent scaling.



EDA

Asia shows a steep upward trend, indicating a significant increase in total emissions over time

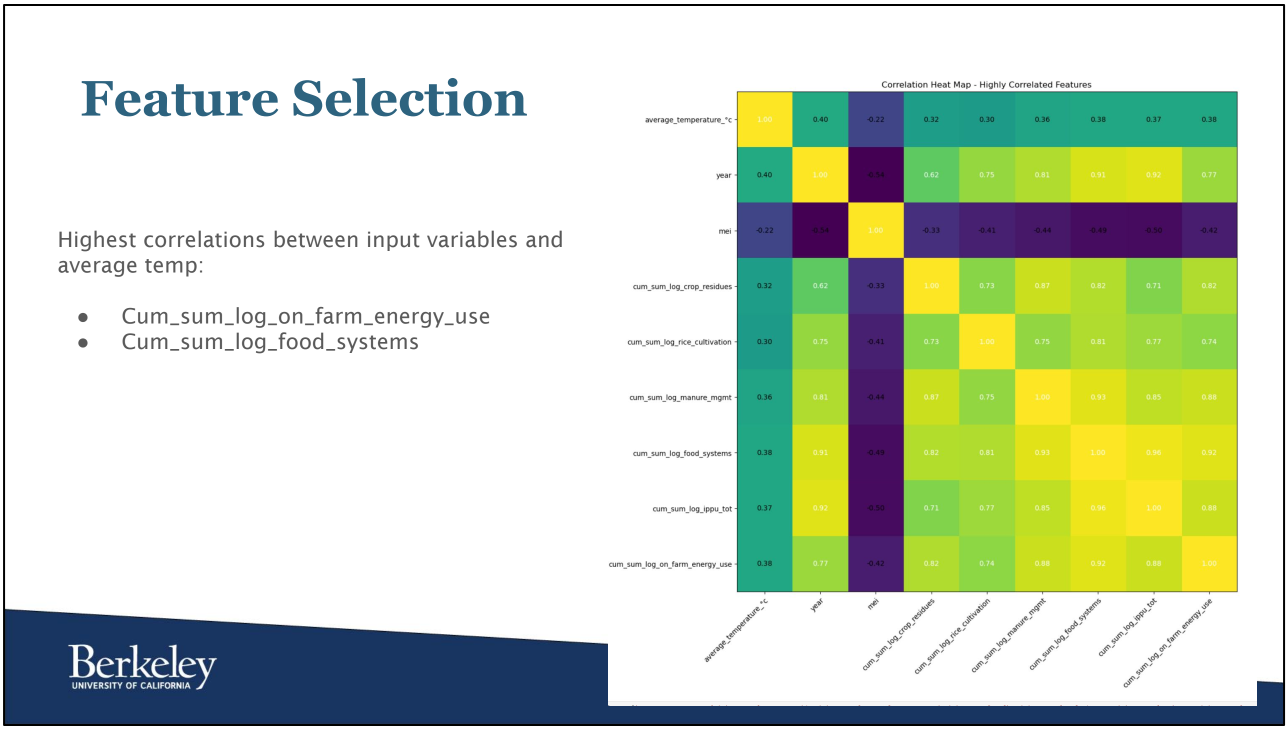
Here as well, Asia shows increased emissions by population level

Despite fluctuations, there is a general upward trend in average temperature change for all regions

Berkeley
UNIVERSITY OF CALIFORNIA

Here are some examples of some EDA - we plotted all of the different regions here and took a look at both the changes in emissions and temperature change, plotted over the years.

From the EDA, we saw that Asia shows a steep upward trend, indicating a significant increase in total emissions over time and also when plotted by population. As expected, there is a general upward trend and average temperature change for all regions.



Feature Selection

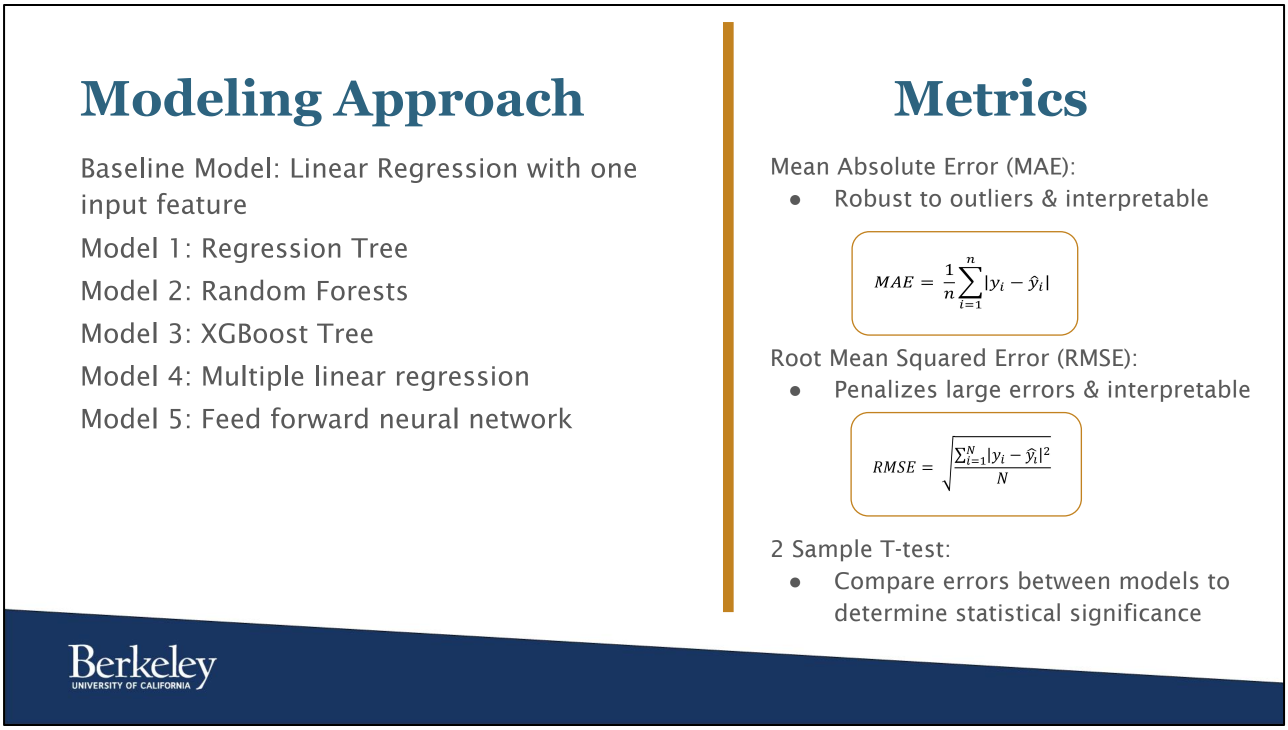
Highest correlations between input variables and average temp:

- Cum_sum_log_on_farm_energy_use
- Cum_sum_log_food_systems

Correlation Heat Map - Highly Correlated Features

	average_temperature_°c	year	mei	cum_sum_log_crop_residues	cum_sum_log_rice_cultivation	cum_sum_log_manure_mgmt	cum_sum_log_food_systems	cum_sum_log_ippu_tot	cum_sum_log_on_farm_energy_use
average_temperature_°c	1.00	0.40	-0.22	0.32	0.30	0.36	0.38	0.37	0.38
year	0.40	1.00	-0.54	0.62	0.75	0.81	0.91	0.92	0.77
mei	-0.22	-0.54	1.00	-0.33	-0.41	-0.44	-0.49	-0.50	-0.42
cum_sum_log_crop_residues	0.32	0.62	-0.33	1.00	0.73	0.87	0.82	0.71	0.82
cum_sum_log_rice_cultivation	0.30	0.75	-0.41	0.73	1.00	0.75	0.81	0.77	0.74
cum_sum_log_manure_mgmt	0.36	0.81	-0.44	0.87	0.75	1.00	0.93	0.85	0.88
cum_sum_log_food_systems	0.38	0.91	-0.49	0.82	0.81	0.93	1.00	0.96	0.92
cum_sum_log_ippu_tot	0.37	0.92	-0.50	0.71	0.77	0.85	0.96	1.00	0.88
cum_sum_log_on_farm_energy_use	0.38	0.77	-0.42	0.82	0.74	0.88	0.92	0.88	1.00

Taking a look at a selection of the highest correlations between the input and average temp, the farm energy use and food systems are the highest correlated, and the Industrial Processes and Product Use (IPPU) is a close follow up. These are included in the final linear model.



Modeling Approach

Baseline Model: Linear Regression with one input feature

Model 1: Regression Tree

Model 2: Random Forests

Model 3: XGBoost Tree

Model 4: Multiple linear regression

Model 5: Feed forward neural network

Metrics

Mean Absolute Error (MAE):

- Robust to outliers & interpretable

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE):

- Penalizes large errors & interpretable

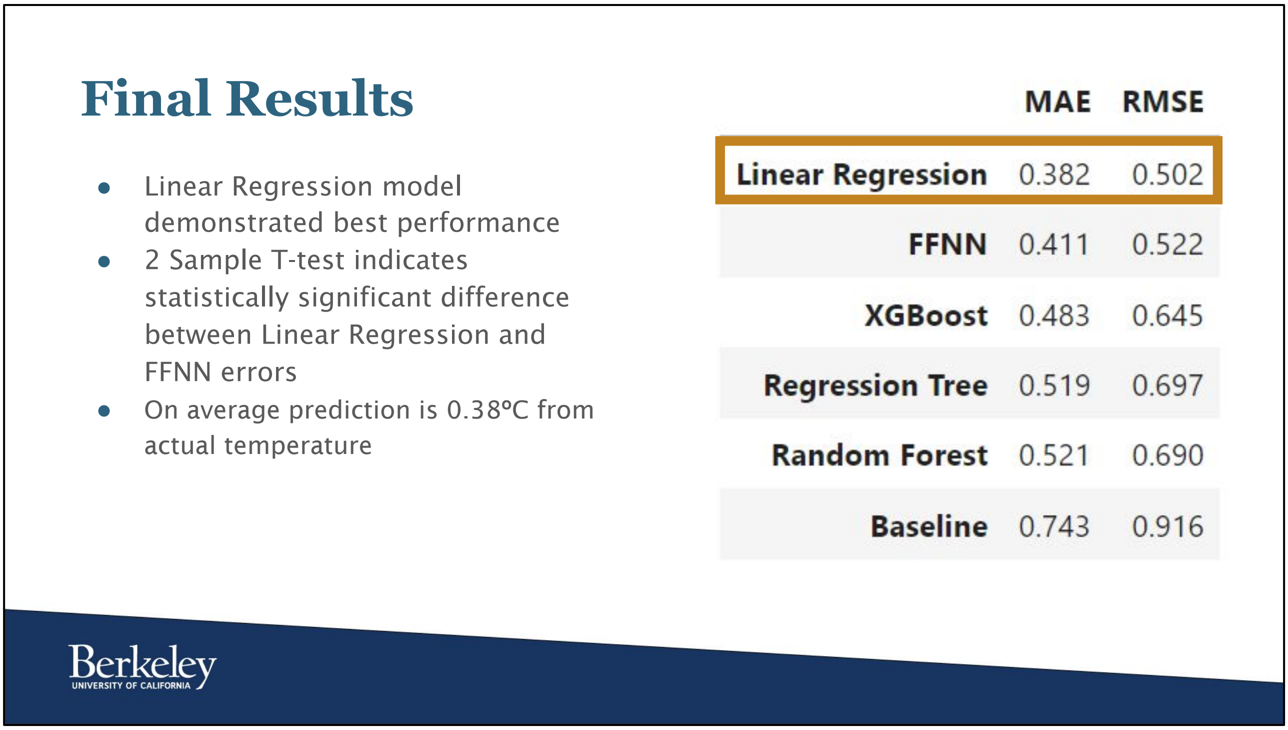
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|^2}{N}}$$

2 Sample T-test:

- Compare errors between models to determine statistical significance

Berkeley UNIVERSITY OF CALIFORNIA

After establishing a baseline model for comparison we modeled our data using multiple regression tree variants, a multiple linear regression model, and finally, a feed forward neural network. Because our output variable is continuous data, we chose Mean Absolute Error as the main metric to evaluate performance of these models. This metric is robust to outliers and is interpretable as it is reported in the same units as our output variable. As a secondary metric, we used the Root Mean Squared Error as this metric is slightly more conservative than the MAE. In order to compare models that produced similar MAE values, we chose to run 2 sample t-tests to compare the error values between models and determine any statistically significant differences between the error populations.



Final Results

- Linear Regression model demonstrated best performance
- 2 Sample T-test indicates statistically significant difference between Linear Regression and FFNN errors
- On average prediction is 0.38°C from actual temperature

	MAE	RMSE
Linear Regression	0.382	0.502
FFNN	0.411	0.522
XGBoost	0.483	0.645
Regression Tree	0.519	0.697
Random Forest	0.521	0.690
Baseline	0.743	0.916

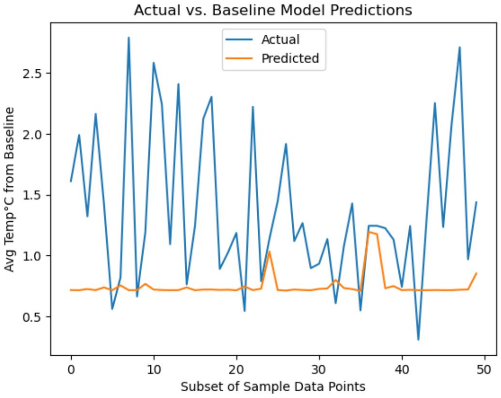
Berkeley
UNIVERSITY OF CALIFORNIA

Let's look at the final results. Our expanded linear regression model outperformed all other models. The results of a two-sample T-test indicate a statistically significant difference in prediction errors between the Linear Regression model and another more complex model we tried, the Feedforward Neural Network. The regression model's predictions were only off by 0.38 degrees Celsius from the actual temperature.

Experiments

Baseline Model

- Linear regression with single input feature (Cumulative Sum Total CO_2 Emissions)
- Established baseline loss for comparison on future models
- MAE = 0.743ºC



Berkeley
UNIVERSITY OF CALIFORNIA

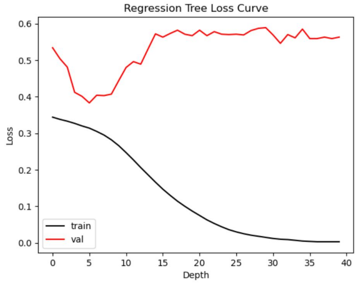
Faye

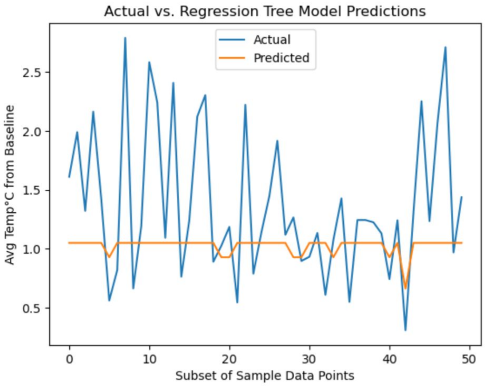
For our baseline model, we chose to run a univariate linear regression model. As our input variable we chose the cumulative summation of total CO2 emissions, which is simply the total agricultural CO2 emissions for each year, summed year over year.

The predictions from this model were on average, 0.743 degrees celsius from the actual temperature values in our test data set. As shown in the graph on the right, the input to this model was not that accurate in predicting actual temperature changes. However, this model establishes a baseline comparison of agricultural CO2 emissions in general against our variable of interest and acts as a good foundation for comparison with future, more complex models.

Regression Tree

- Potential explainability benefit
- Moderate improvement over baseline model → MAE = 0.519ºC
- Optimal Hyperparameter: Max Depth = 5

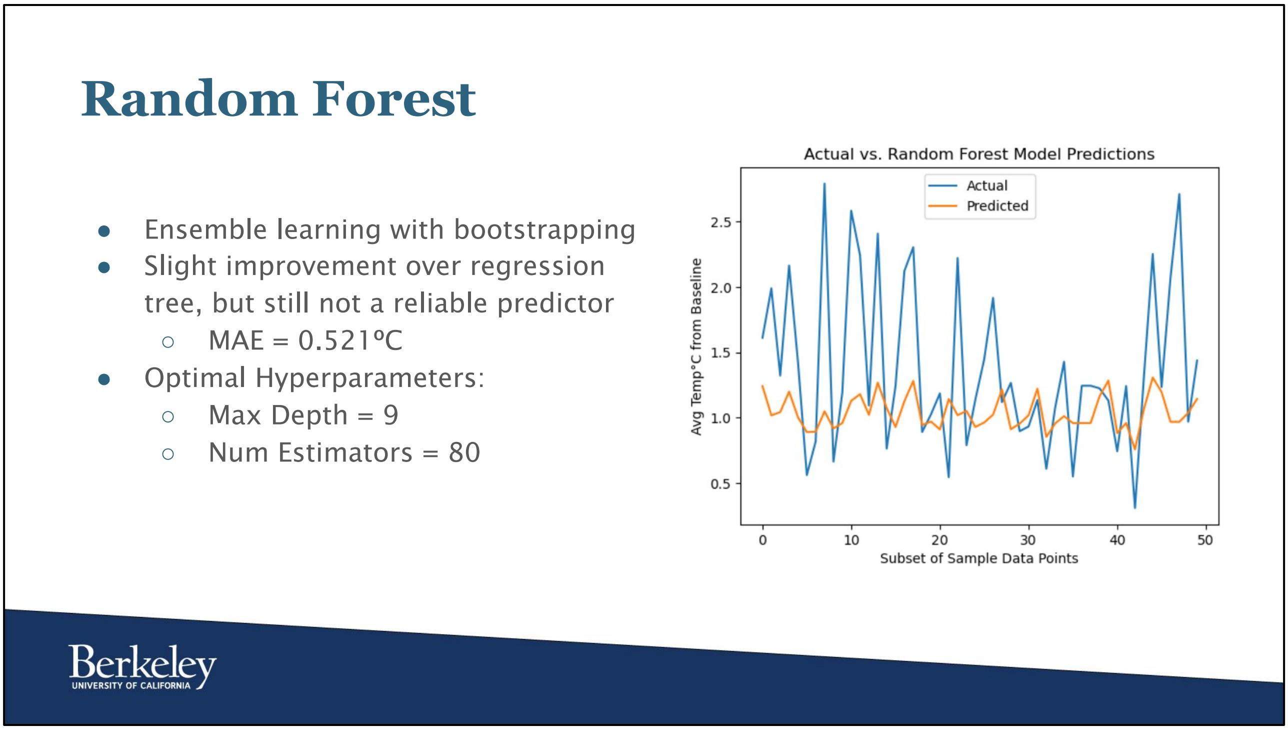




Berkeley
UNIVERSITY OF CALIFORNIA

Faye
Next, we experimented with a few decision tree model variants. As the predictions made by decision trees follow intuitive logic that can be easily visualized, we thought that these models would be a good first approach to attempting to answer our research question.

Because our output variable is continuous, we used a regression tree architecture. We limited our input variables to those with the highest correlation to the temperature output, which included CO2 emissions from manure management, food systems, industrial processes, and on farm energy use. While we saw a moderate improvement compared with the baseline model, the predictions were still an average of a half a degree celsius away from the actual temperature values. As shown in the graph on the left, our analysis indicated that at higher depths, the model quickly began to overfit to the training data. To limit this overfitting, we chose a maximum depth of 5 to optimize this model.



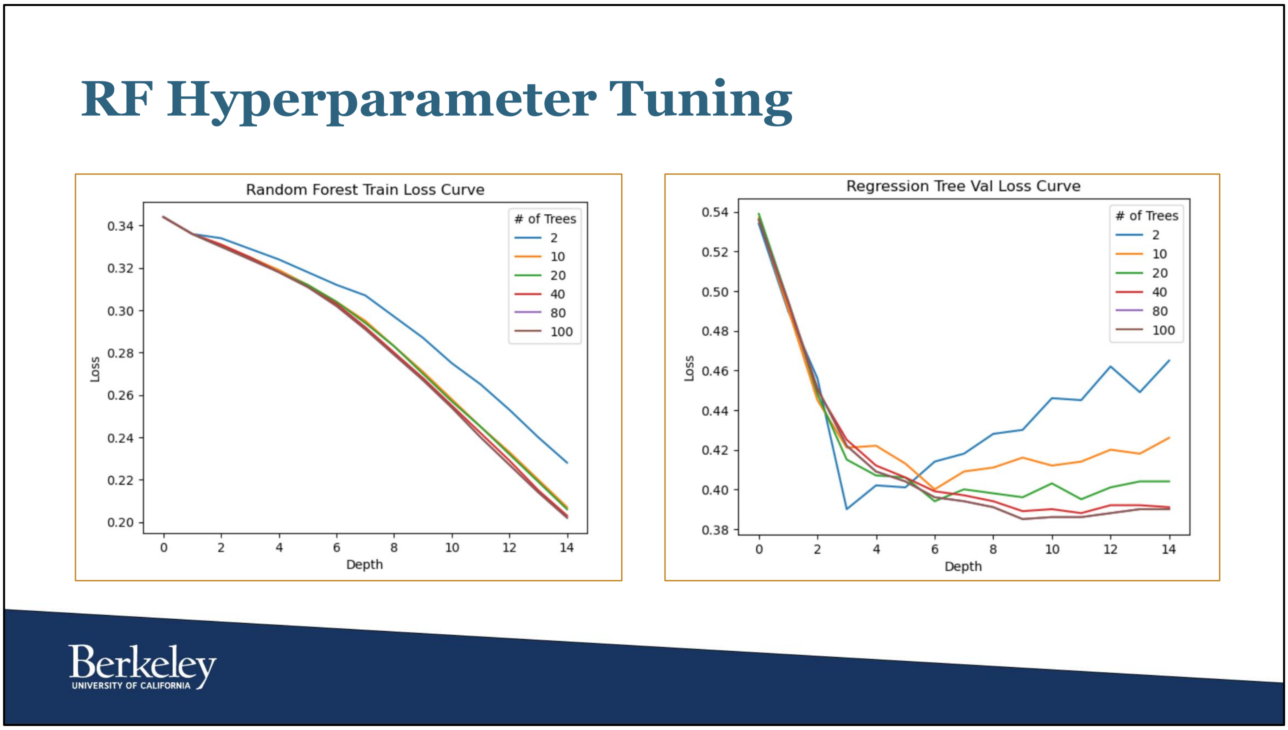
Random Forest

- Ensemble learning with bootstrapping
- Slight improvement over regression tree, but still not a reliable predictor
  - MAE = 0.521ºC
- Optimal Hyperparameters:
  - Max Depth = 9
  - Num Estimators = 80

Actual vs. Random Forest Model Predictions
Actual
Predicted
Avg Temp°C from Baseline
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
50
Subset of Sample Data Points
Berkeley
UNIVERSITY OF CALIFORNIA

Faye
To see if we could improve further on our regression tree model, we moved on to an ensemble learning architecture using random forest regression. We leverage a bootstrapping method for this model, meaning that each individual tree within the random forest was built using a random subset of the training data and then the results of each tree within the forest were aggregated to determine the optimal predictions.

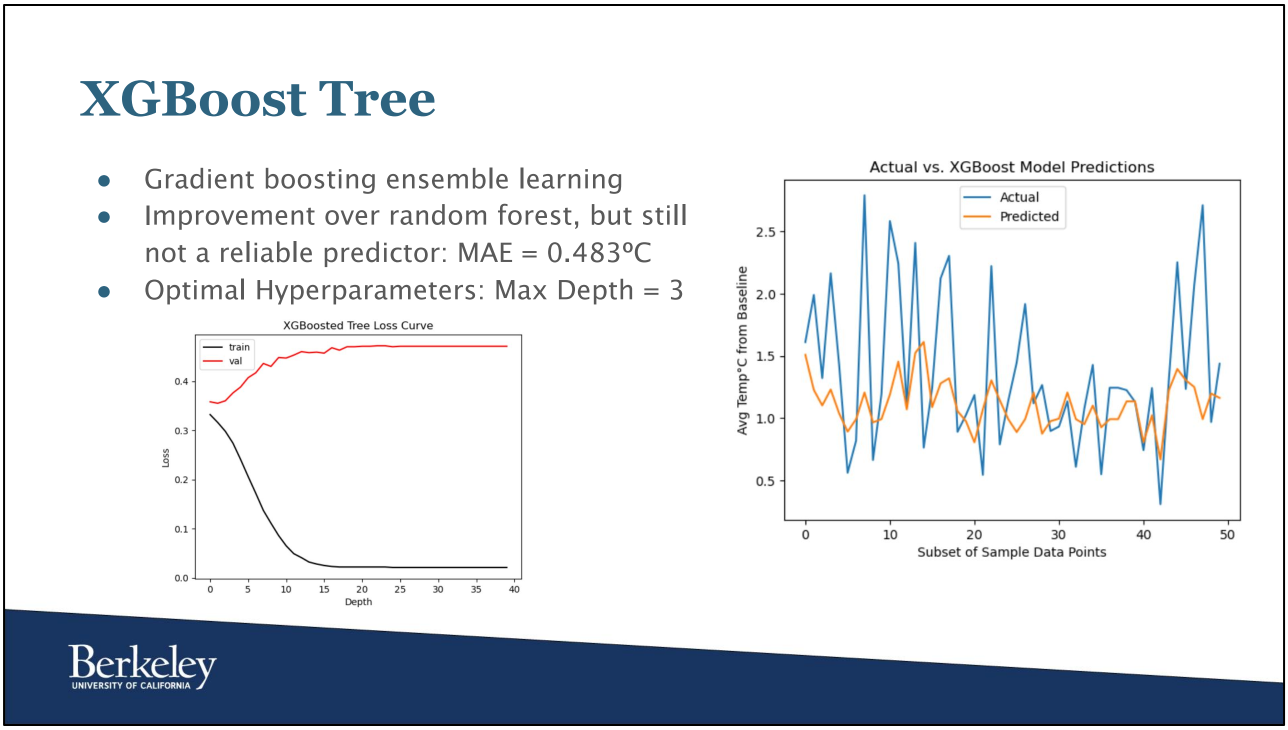
We saw a minor improvement over the single regression tree model, but still had a fairly significant error when running on the test data set.



RF Hyperparameter Tuning

Faye
As shown in the previous slide, we optimized this model by using 80 trees at a maximum depth of 9. To arrive at these optimal hyperparameters, we plotted the training and validation loss curves with a varying number of trees at varying depths.

The graph on the left shows the loss curves for the training data set. From this graph, there was a large drop in loss between 2 and 10 trees, but little difference in the loss with an increasing number of trees. While we could have run models of greater depth to decrease the training loss further, the loss curves of our validation data began to increase at greater depths. Based on our analysis, the validation loss was minimized with a forest made up of 80 trees at a depth of 9.

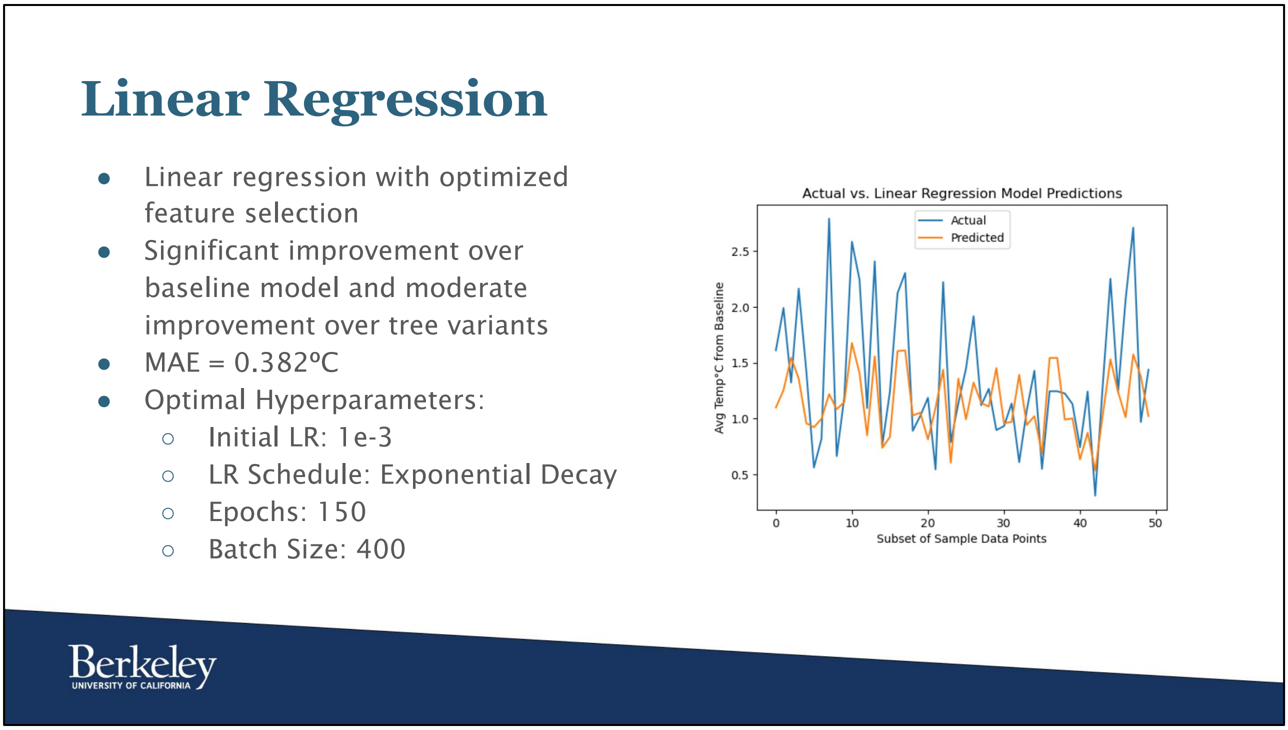


XGBoost Tree

- Gradient boosting ensemble learning
- Improvement over random forest, but still not a reliable predictor: MAE = 0.483ºC
- Optimal Hyperparameters: Max Depth = 3

Faye
As a final regression tree variant, we chose to build an xgboosted tree model. The XGboosted model combines ensemble learning with boosting which fits each tree in sequence, using the predictions from the preceding tree as input to the subsequent tree in the sequence. In this way, each tree learns from the weaknesses of the preceding tree, which theoretically, should help to minimize the final loss of the model as a whole. Similar to the previous two models, we had to limit the depth of the tree in order to prevent overfitting to the training data. This model was able to improve the loss to approximately 0.48 degrees celsius.

While we were able to demonstrate improvement over the baseline model using these various tree architectures, there is still a significant error in the temperature predictions. Therefore, we switched gears to evaluate linear regression and neural network models. I'll pass it over to Julia to discuss these in more detail.

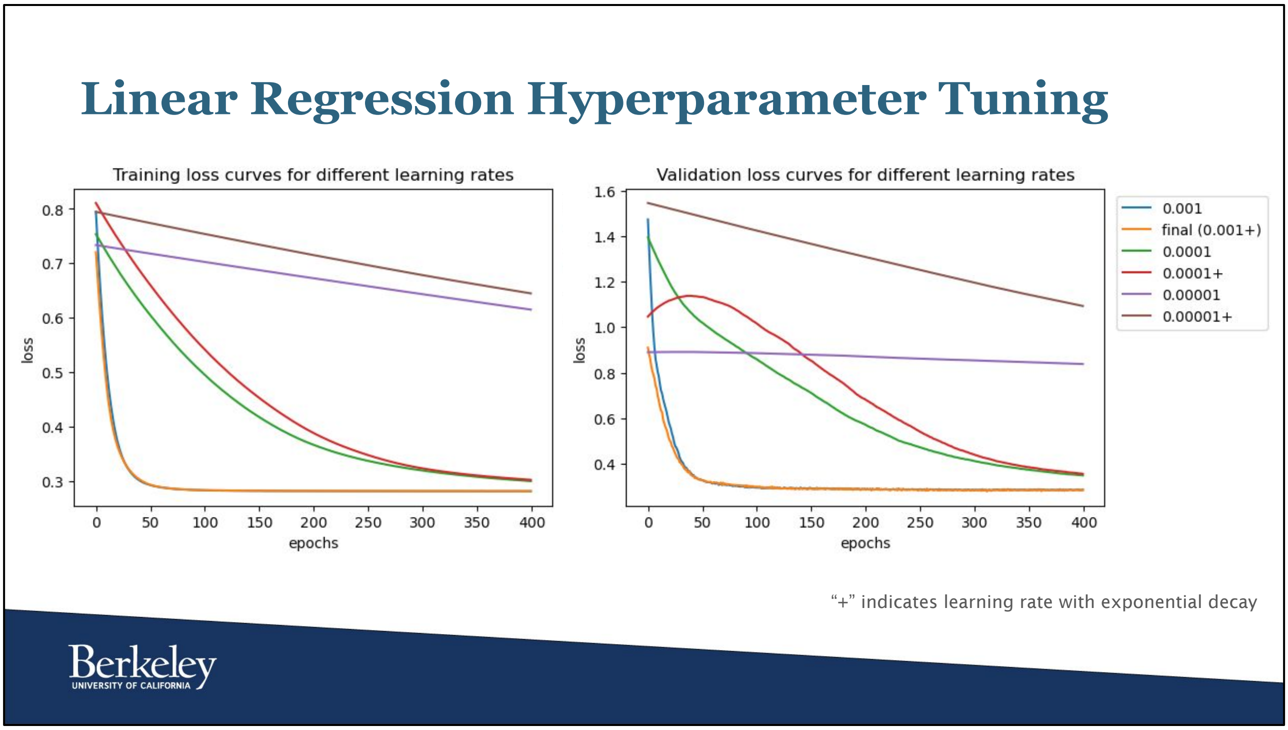


Linear Regression

- Linear regression with optimized feature selection
- Significant improvement over baseline model and moderate improvement over tree variants
- MAE = 0.382ºC
- Optimal Hyperparameters:
  - ○ Initial LR: 1e-3
  - ○ LR Schedule: Exponential Decay
  - ○ Epochs: 150
  - ○ Batch Size: 400

- Inputs - MEI, Urban Pop, Rice, Manure, Food Systems, IPPU, On Farm Energy Use, Sub-Region, Area

Going back to linear regressions, we expanded upon the baseline model by including more inputs. Our final model included a selection of the more correlated inputs. These were MEI, Urban Pop and emissions from rice cultivation, manure management, food systems, industrial processes, and on-farm energy use. We also used one-hot encodings for sub-region and area. The optimal parameters were a learning rate of 0.001 which we implemented with exponential decay. We used a batch size of 400 and 150 epochs. The linear regression produced the lowest mean absolute error out of all of our models of 0.38ºC.



Linear Regression Hyperparameter Tuning

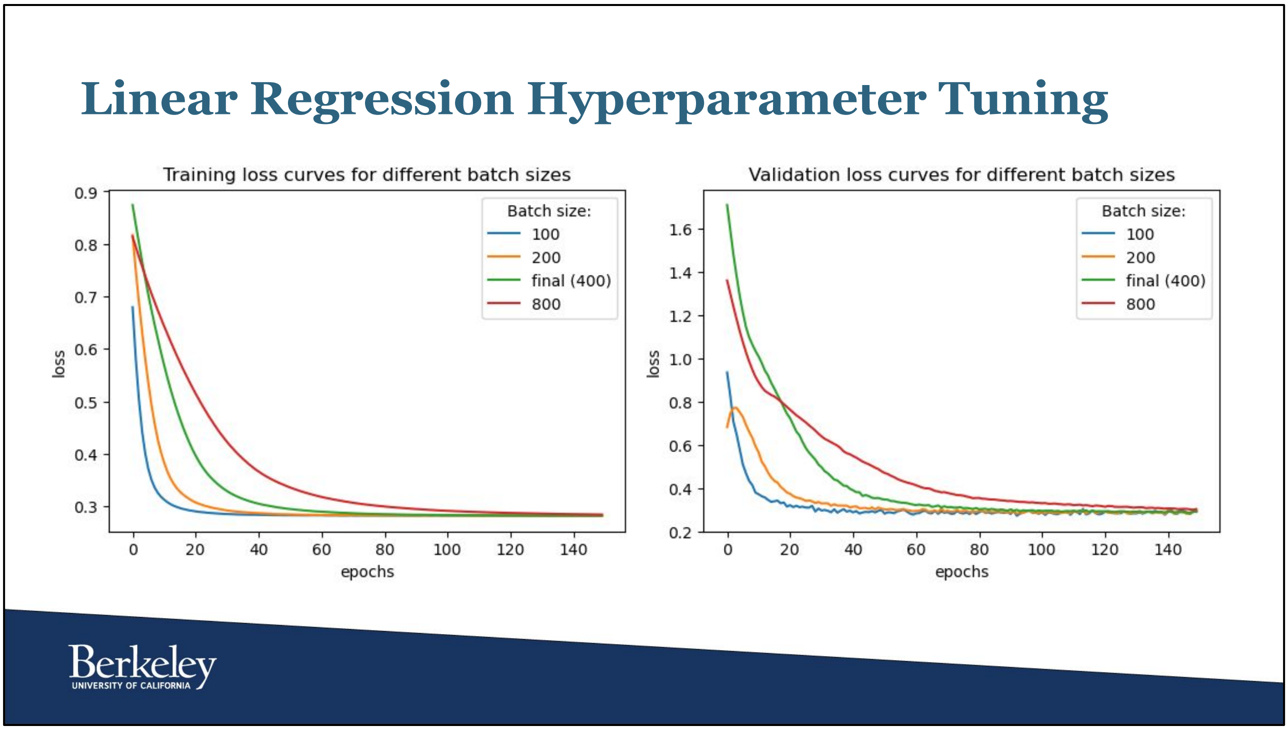
Training loss curves for different learning rates

Validation loss curves for different learning rates

“+” indicates learning rate with exponential decay

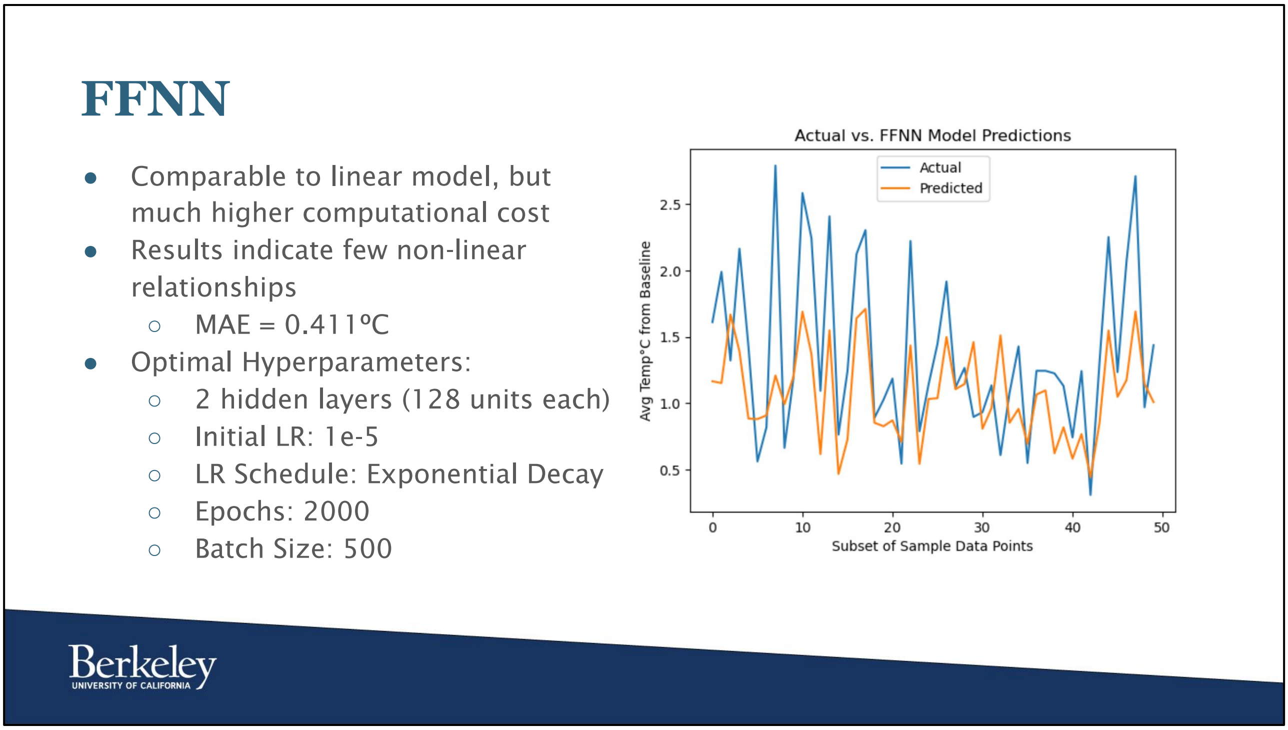
- showing more epochs than actually was used for the final model, because smaller learning rates require more epochs to train

Aside from experimenting with including different features, we also tried different learning rates and batch sizes. For illustration purposes, these graphs are showing more epochs than what we ran on the final model to show the impact of smaller learning rates. For the final linear regression model, we were able to use a much higher learning rate than what was used in the feed forward neural network.



Linear Regression Hyperparameter Tuning

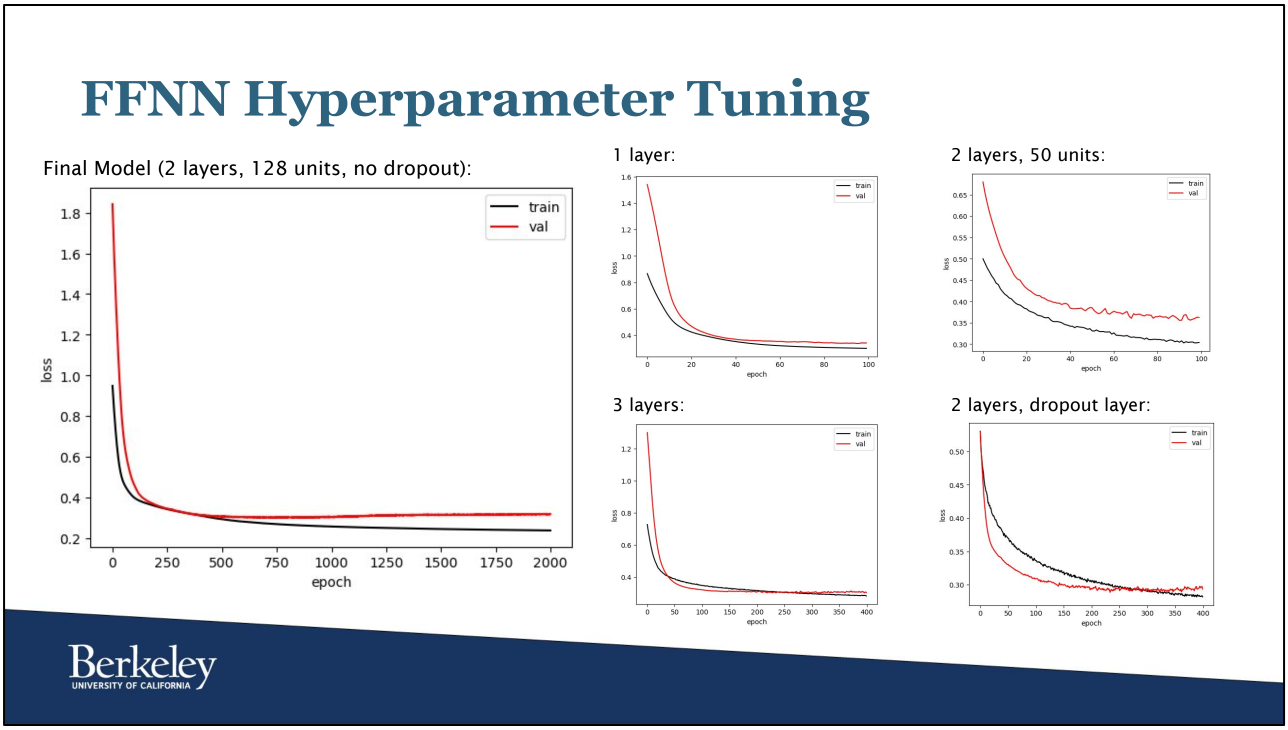
Here are comparisons of different batch sizes all using a learning rate of 0.001 with exponential decay. The smaller batch sizes produced more noise in the validation loss and so the final batch size that we used was 400.



FFNN

- Comparable to linear model, but much higher computational cost
- Results indicate few non-linear relationships
  - MAE = 0.411ºC
- Optimal Hyperparameters:
  - 2 hidden layers (128 units each)
  - Initial LR: 1e-5
  - LR Schedule: Exponential Decay
  - Epochs: 2000
  - Batch Size: 500

Lastly, moving on to feed forward neural networks. In our final model, we included all of the CO2 emission features as well as MEI, urban population and we had embeddings for the area and sub-region inputs. Compared to a linear model, the computational cost is much higher. The final FNN has almost 20,000 trainable parameters whereas the linear model only had 216 trainable parameters. The mean absolute error was slightly higher than that of the linear model at 0.41ºC so this model also performed better than the baseline and tree variants, but slightly worse than the linear model.



FFNN Hyperparameter Tuning

Final Model (2 layers, 128 units, no dropout):

1 layer:

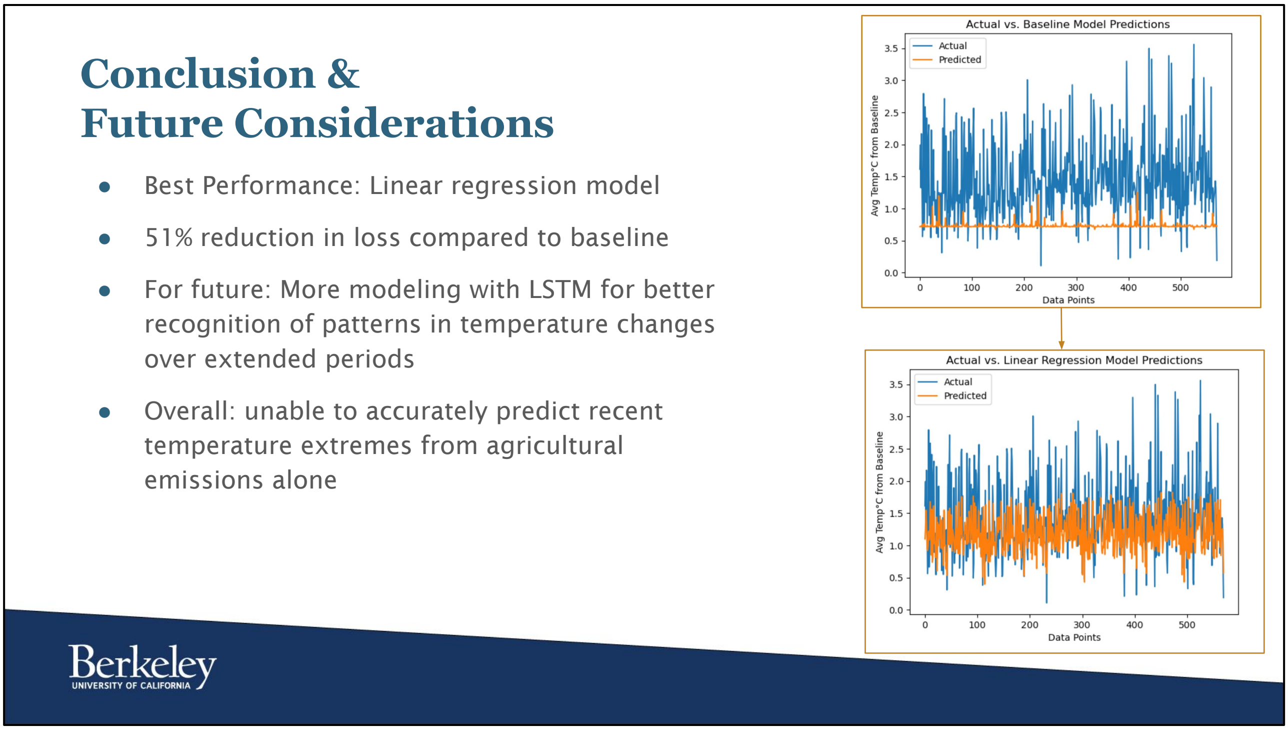
2 layers, 50 units:

3 layers:

2 layers, dropout layer:

We experimented with a lot of different hyperparameters, here are just a small selection. Some of our observations were that here was much more improvement going from 1 layer to 2 layers than from 2 layers to 3 layers. We also found that having 100 or more units in a layer seemed necessary to produce a complex enough model. We ended up not having a dropout layer because it resulted in more underfitting for the train curve and did not improve the validation loss much, but rather introduced more noise.

- Generally used smaller learning rates for FFNN than linear regression and larger batch sizes



Conclusion & Future Considerations

- Best Performance: Linear regression model
- 51% reduction in loss compared to baseline
- For future: More modeling with LSTM for better recognition of patterns in temperature changes over extended periods
- Overall: unable to accurately predict recent temperature extremes from agricultural emissions alone

To wrap things up, after working through all of our previous models, our final results indicate that our best performing model was the linear regression model. It demonstrated a 51% loss reduction compared to our baseline model. If we were to continue working on this problem, we would foray into experiments with LSTM modeling, or long short term memory modeling, for better recognition of patterns in temperature changes over extended periods of time. With all of that being said, through this process, we came to realize that climate modeling in this capacity is extremely complicated, and we do not feel like we can confidently predict temperature changes from agricultural practices that focus solely on CO2 emissions alone. Beyond LSTM modeling in the future, we would like to incorporate more features outside of agriculture to create a more accurate model.

Questions?

Thank you! We can now take questions.

Contributions

	Domain Research	Pre-processing/ Feat. Eng	EDA	Decision Tree & Variants	Linear Regression	FFNN	Slides
Rachel	X		X	X			X
Darya	X				X	X	X
Julia		X			X	X	X
Faye		X	X	X	X	X	X

Github Repository

https://github.com/rachtripoli/DATASCI207_finalproject_Likhareva_Titchenal_Tripoli_Zhao/tree/main

Data Sources/References

1. https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml/data
2. https://psl.noaa.gov/enso/mei/
3. https://ourworldindata.org/greenhouse-gas-emissions-food#:~:text=The%20specific%20number%20that%20answers,we%20include%20all%20agricultural%20products.
4. https://www.ipcc.ch/2021/08/09/ar6-wg1-20210809-pr/
5. https://www.gfdl.noaa.gov/news/noaa-scientists-harness-machine-learning-to-advance-climate-models/

Berkeley
UNIVERSITY OF CALIFORNIA