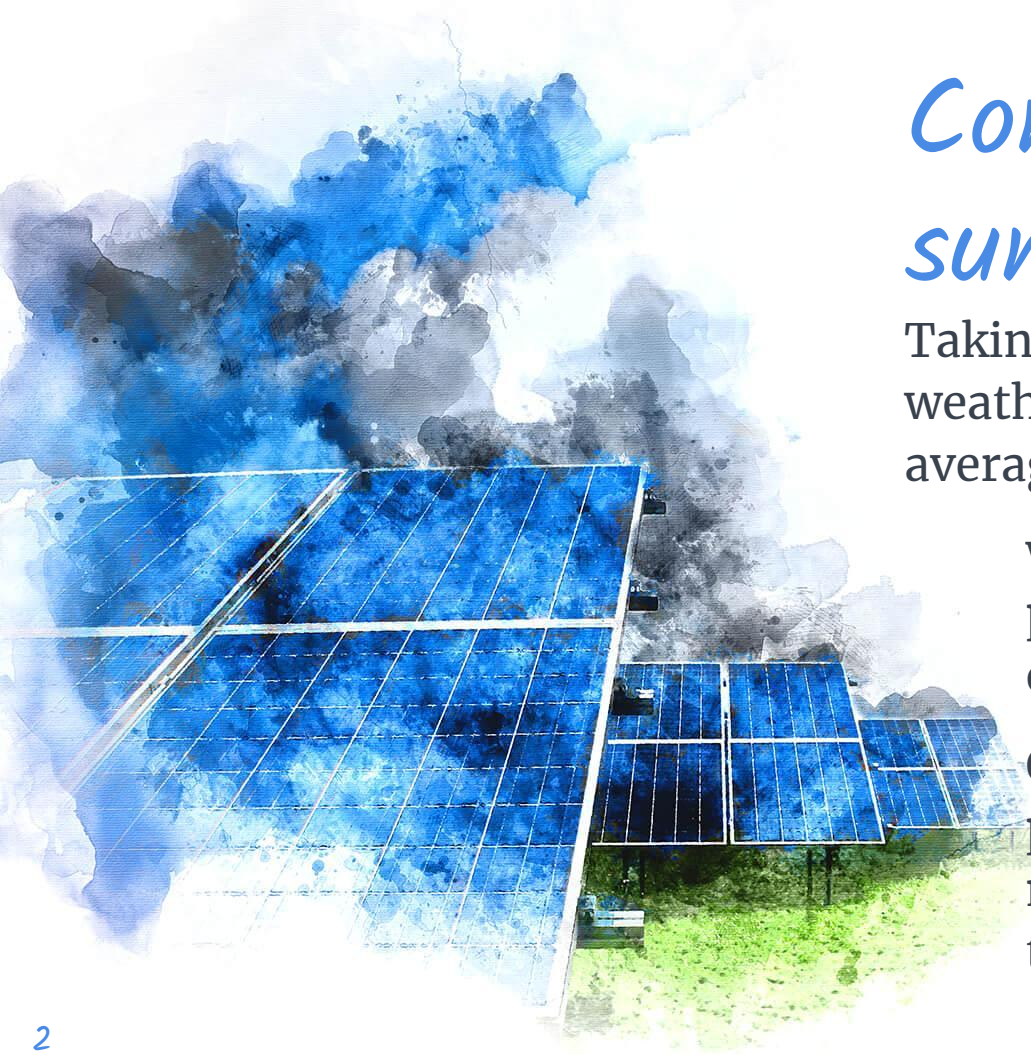
Solar Spark

By: Sid Surapaneni, Nikki Aaron,
Kevin Hoffman, and Ashley Scurlock



Component Executive summary

Taking into account location factors, weather, solar radiation levels, and average amount of energy used.

we want to build a model that can predict the yearly dollars saved on energy after installing 30 solar panels.

Our supervised model is accurate at predicting this (3% avg error) despite missing variables and not knowing all the equations that govern the result.

Data journey

To make our dataset for this project we combined both government and private data from ***seven*** different sources.

Average Annual Energy Consumption By State
Source: US Energy Information Administration

US Elevation by County
Source: US Geological Survey

Average Monthly Temperatures (℃) By Coordinate
Source: Climatic Research Unit

Cost of Energy by State ($/kWH)
Source: US Department of Energy

Average Monthly Solar Energy Output (W/m²) By Zip
Source: National Renewable Energy Laboratory

Average Monthly Cloudy Days by Coordinate
Source: Climatic Research Unit

Continental US Zip Codes (with latitude and longitude)
Source: SimpleMaps

Final Dataset

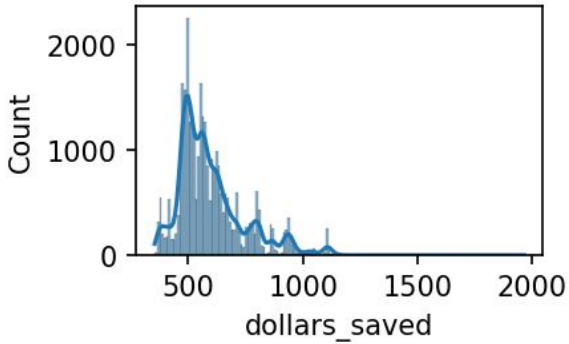
Data summary

Preprocessing

- Combined datasets and scaled all of the features except the response.

Response Variable:

- ***Dollars saved*** is derived from the estimated energy output using 30 solar panels multiplied by the cost of electricity in the area.



Variation Std.Dev. / Mean		**Correlations** with Dollars Saved	
Dollars saved	24%	Energy used	-0.33
Elevation	125%	Cloudy	-0.23
Temp	37%	DNI	0.25
Energy used	18%	Temp	-0.18
DNI	16%	Elevation	0.14
Cloudy	12%	Lat	0.15
Solar output	5%	Lng	-0.11

Models constructed

1. Linear Regression
   - Baseline model for benchmarking

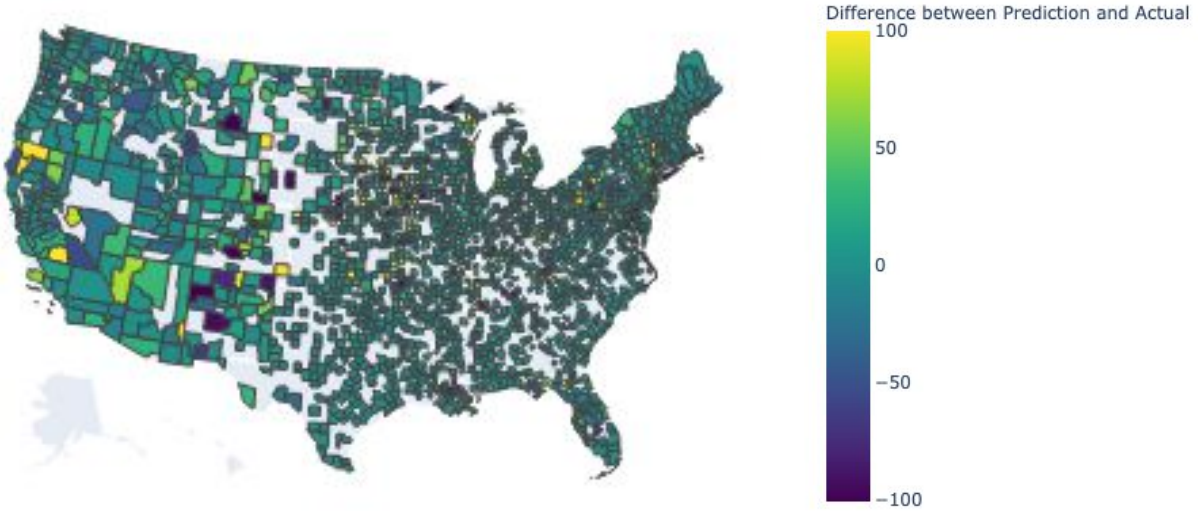
2. Random Forest Regression
   - Ensemble of decision trees
   - Great for non-linear relationships

3. Gradient Boosted Trees
   - Best model
   - One drawback is long training times

Model performance

Model Name	RMSE	Hyperparameters
Linear Regression	$109.23	MaxIterations: 100 regParam: 0 elasticNetParam: 0
Random Forest Regression	$54.95	MaxDepth: 8 numTrees: 250 minInstancesPerNode: 1
Gradient Boosted Trees	$39.59	MaxIterations: 225 MaxDepth: 7 MinInstancesPerNode: 10

Prediction Choropleth



*RMSE gives a high weight to large errors

Conclusions

]:

	feature	score
6	annual_kwh_used	0.195357
1	lng	0.127142
0	lat	0.051254
17	temp_Oct	0.048217
26	pct_cloudy_days_Mar	0.048105
21	pct_cloudy_days_Dec	0.039097
4	elevation	0.036069
22	pct_cloudy_days_Feb	0.032837
25	pct_cloudy_days_Jun	0.028367
15	temp_May	0.027375

- The RMSE was sensitive to high errors. MAE was 18.11.
- Among the models tried out, the gradient boosted model had the most effective performance at predicting the dollars saved
- Feature Importances of Gradient Boosted Trees Model indicated that factors related to consumption and location(longitude, latitude) were important predictors for dollars saved
- Longitude and Annual_kwh_used also had highest feature importances in the Random Forest Model
- Future work can involve using more updated solar panel data as well as utilizing data from outside of the United States, as well as maintenance cost data of solar panels

Thanks!