Chevron Challenge

Zheran Li, Benson Chi, Zida Wang
01.29.2023

Overview

Expected Due Date

January 29th, 2023

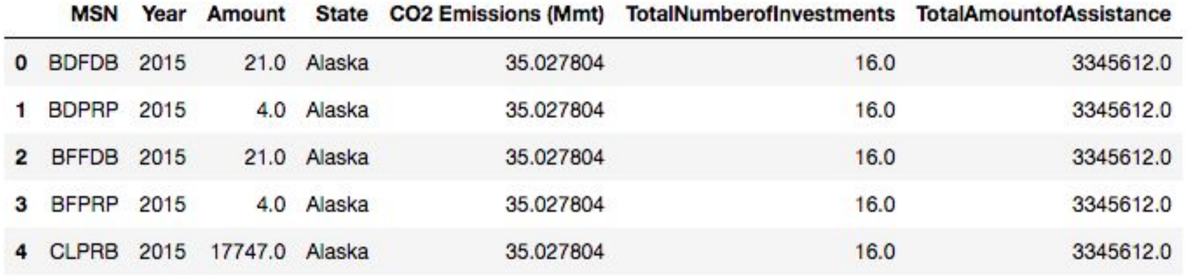
Main Goal

Construct ML model to predict on the total amount of green energy investment in 2020 given the 2020 energy related data

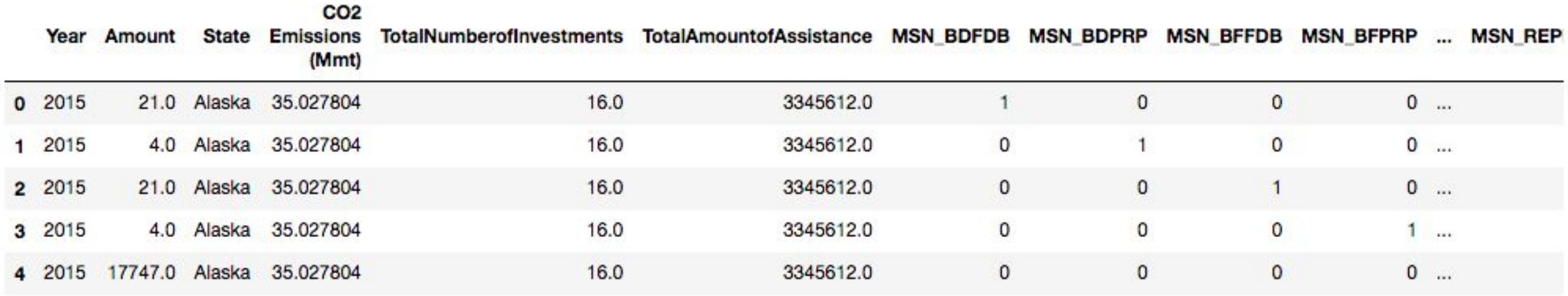
Also Important

Find out which state is the most promising in regards to renewable investments

Preprocessing

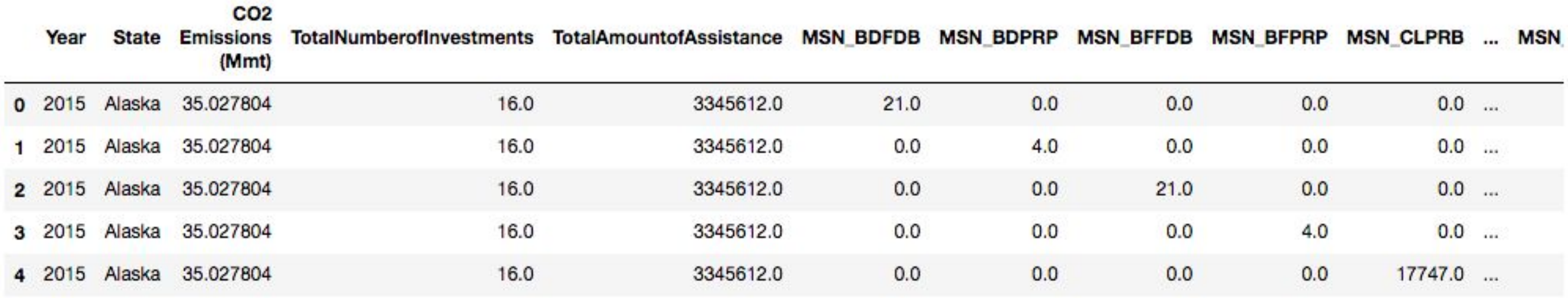


	MSN	Year	Amount	State	CO2 Emissions (Mmt)	TotalNumberofInvestments	TotalAmountofAssistance
0	BDFDB	2015	21.0	Alaska	35.027804	16.0	3345612.0
1	BDPRP	2015	4.0	Alaska	35.027804	16.0	3345612.0
2	BFFDB	2015	21.0	Alaska	35.027804	16.0	3345612.0
3	BFPRP	2015	4.0	Alaska	35.027804	16.0	3345612.0
4	CLPRB	2015	17747.0	Alaska	35.027804	16.0	3345612.0

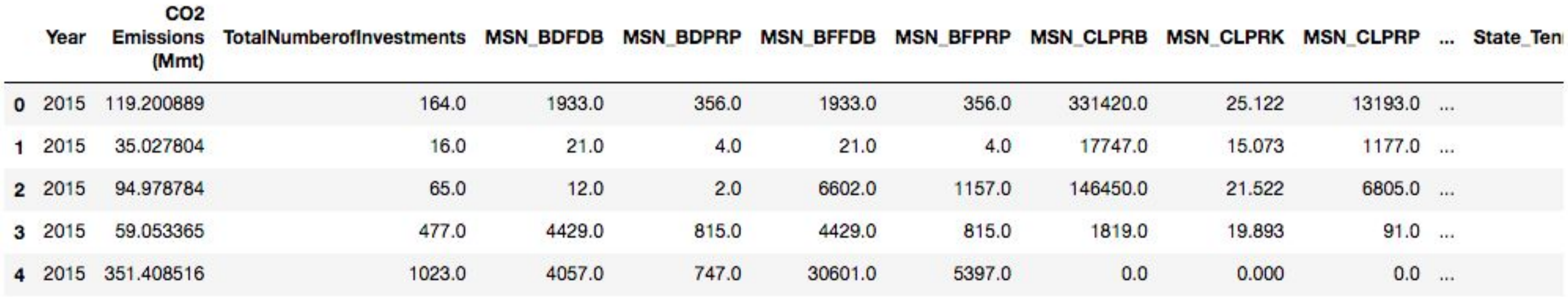


	Year	Amount	State	CO2 Emissions (Mmt)	TotalNumberofInvestments	TotalAmountofAssistance	MSN_BDFDB	MSN_BDPRP	MSN_BFFDB	MSN_BFPRP	...	MSN_REP
0	2015	21.0	Alaska	35.027804	16.0	3345612.0	1	0	0	0	...	
1	2015	4.0	Alaska	35.027804	16.0	3345612.0	0	1	0	0	...	
2	2015	21.0	Alaska	35.027804	16.0	3345612.0	0	0	1	0	...	
3	2015	4.0	Alaska	35.027804	16.0	3345612.0	0	0	0	1	...	
4	2015	17747.0	Alaska	35.027804	16.0	3345612.0	0	0	0	0	...	

Preprocessing



	Year	State	CO2 Emissions (Mmt)	TotalNumberofInvestments	TotalAmountofAssistance	MSN_BDFDB	MSN_BDPRP	MSN_BFFDB	MSN_BFPRP	MSN_CLPRB	...	MSN_
0	2015	Alaska	35.027804	16.0	3345612.0	21.0	0.0	0.0	0.0	0.0	...	
1	2015	Alaska	35.027804	16.0	3345612.0	0.0	4.0	0.0	0.0	0.0	...	
2	2015	Alaska	35.027804	16.0	3345612.0	0.0	0.0	21.0	0.0	0.0	...	
3	2015	Alaska	35.027804	16.0	3345612.0	0.0	0.0	0.0	4.0	0.0	...	
4	2015	Alaska	35.027804	16.0	3345612.0	0.0	0.0	0.0	0.0	17747.0	...	



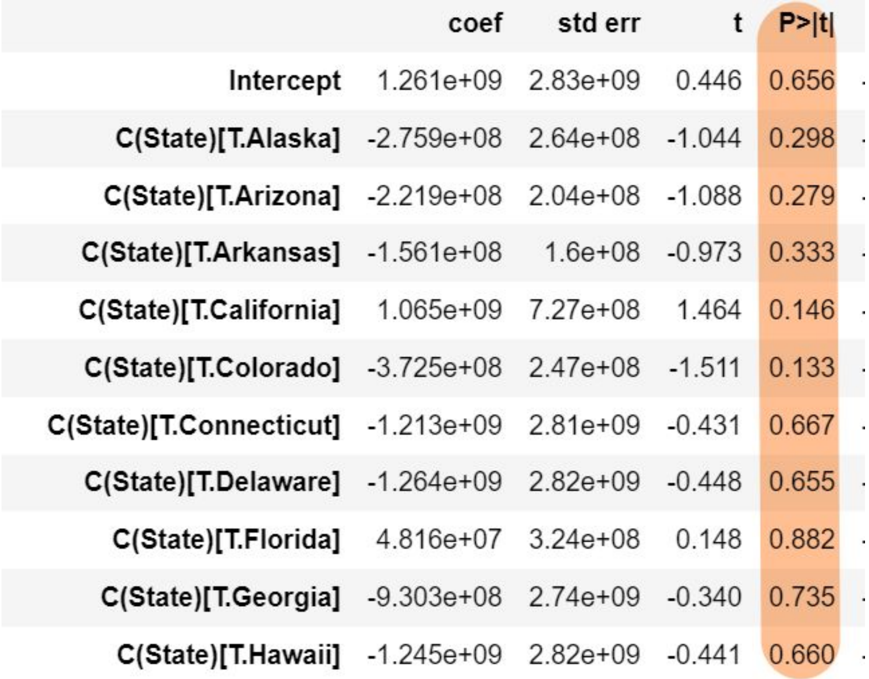
	Year	CO2 Emissions (Mmt)	TotalNumberofInvestments	MSN_BDFDB	MSN_BDPRP	MSN_BFFDB	MSN_BFPRP	MSN_CLPRB	MSN_CLPRK	MSN_CLPRP	...	State_Ten
0	2015	119.200889	164.0	1933.0	356.0	1933.0	356.0	331420.0	25.122	13193.0	...	
1	2015	35.027804	16.0	21.0	4.0	21.0	4.0	17747.0	15.073	1177.0	...	
2	2015	94.978784	65.0	12.0	2.0	6602.0	1157.0	146450.0	21.522	6805.0	...	
3	2015	59.053365	477.0	4429.0	815.0	4429.0	815.0	1819.0	19.893	91.0	...	
4	2015	351.408516	1023.0	4057.0	747.0	30601.0	5397.0	0.0	0.000	0.0	...	

Model Comparison

Linear Model

RMSE = 153,928,095,447.2145

P_value > 10%



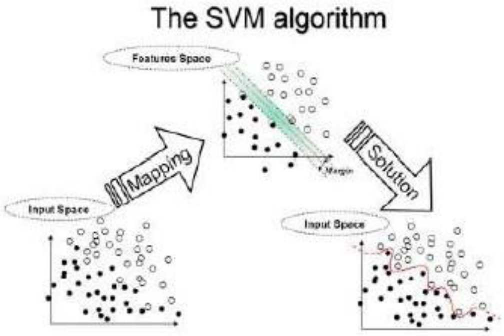
	coef	std err	t	P>|t|
Intercept	1.261e+09	2.83e+09	0.446	0.656
C(State)[T.Alaska]	-2.759e+08	2.64e+08	-1.044	0.298
C(State)[T.Arizona]	-2.219e+08	2.04e+08	-1.088	0.279
C(State)[T.Arkansas]	-1.561e+08	1.6e+08	-0.973	0.333
C(State)[T.California]	1.065e+09	7.27e+08	1.464	0.146
C(State)[T.Colorado]	-3.725e+08	2.47e+08	-1.511	0.133
C(State)[T.Connecticut]	-1.213e+09	2.81e+09	-0.431	0.667
C(State)[T.Delaware]	-1.264e+09	2.82e+09	-0.448	0.655
C(State)[T.Florida]	4.816e+07	3.24e+08	0.148	0.882
C(State)[T.Georgia]	-9.303e+08	2.74e+09	-0.340	0.735
C(State)[T.Hawaii]	-1.245e+09	2.82e+09	-0.441	0.660

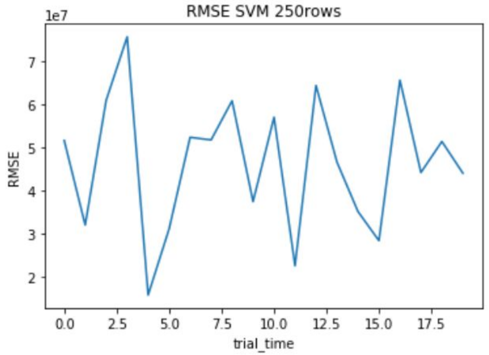
GridSearch

SVM

RMSE = 46,478,462.69

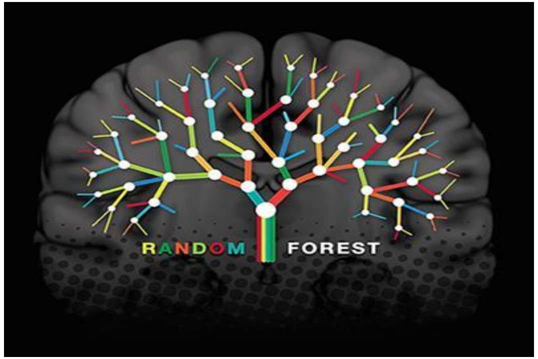
Fine tuned: C

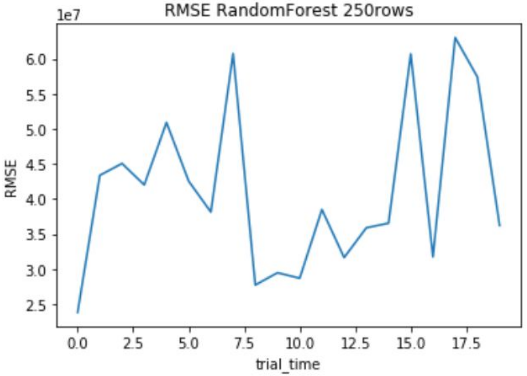




RandomForest

RMSE = 41,203,279.39
Fine tuned: n_estimators,   max_depth



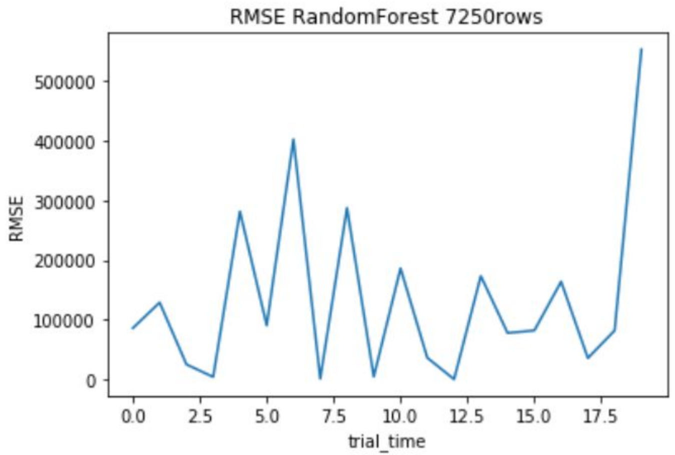


Final Model: RandomForest

- Go back to 7250 rows:

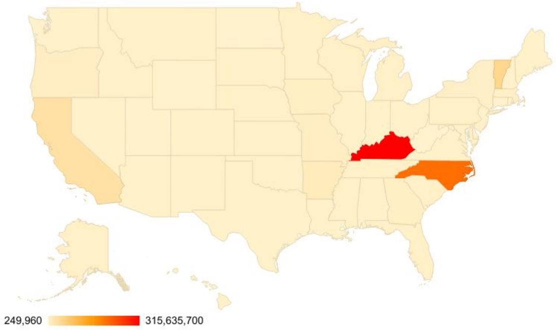
  250 rows vs 83 features

  7250 rows vs 83 features
- Use GridSearch
- Hyperparameter tuning
- Recalculate RMSE = 134,932.1668

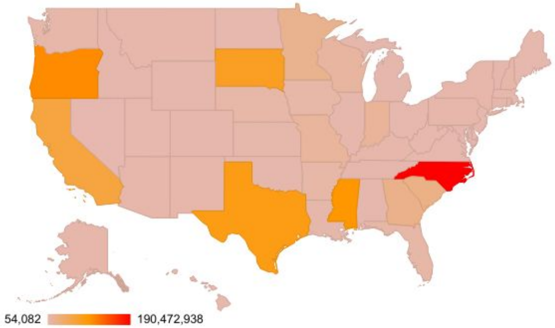


Total Amount of Assistance from 2015 - 2019

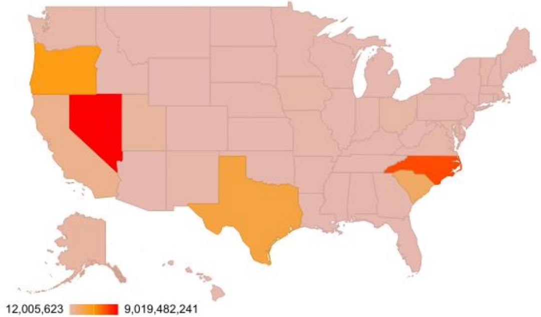
2015



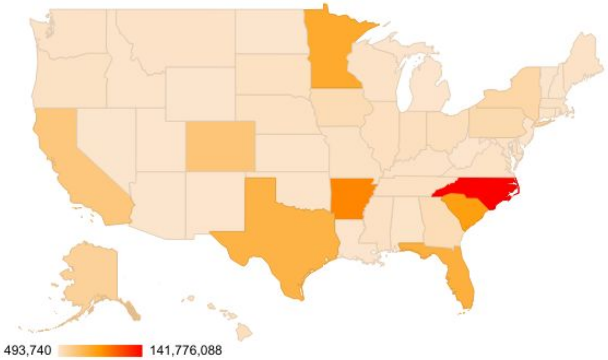
2016



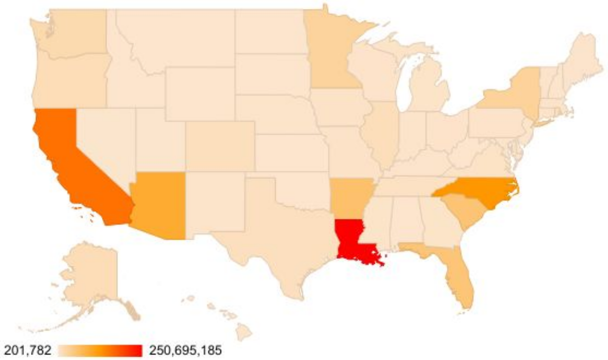
2017



2018

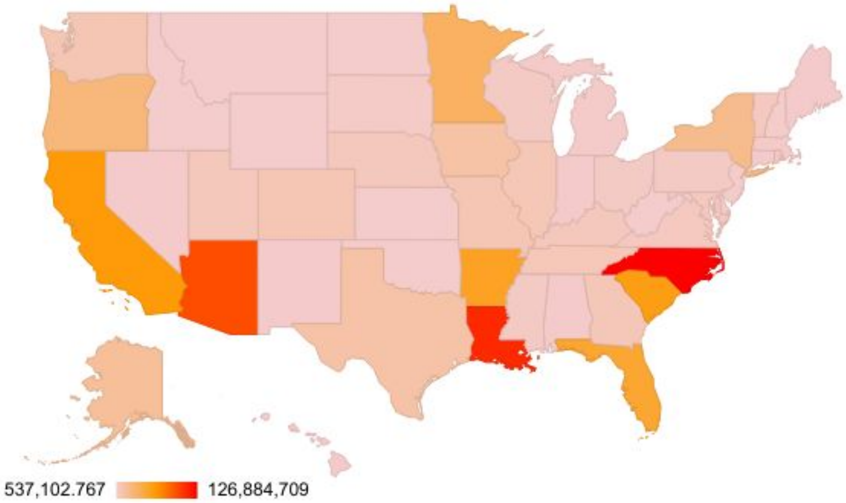


2019

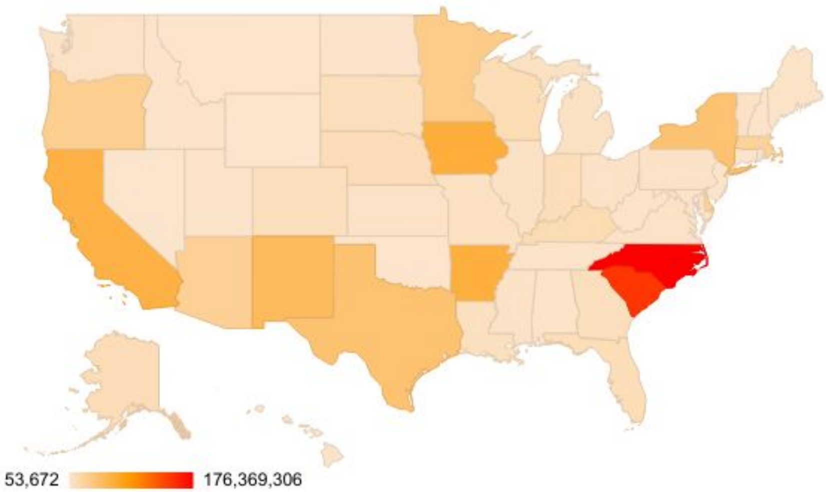


Our Prediction on 2020 vs 2020 in Reality

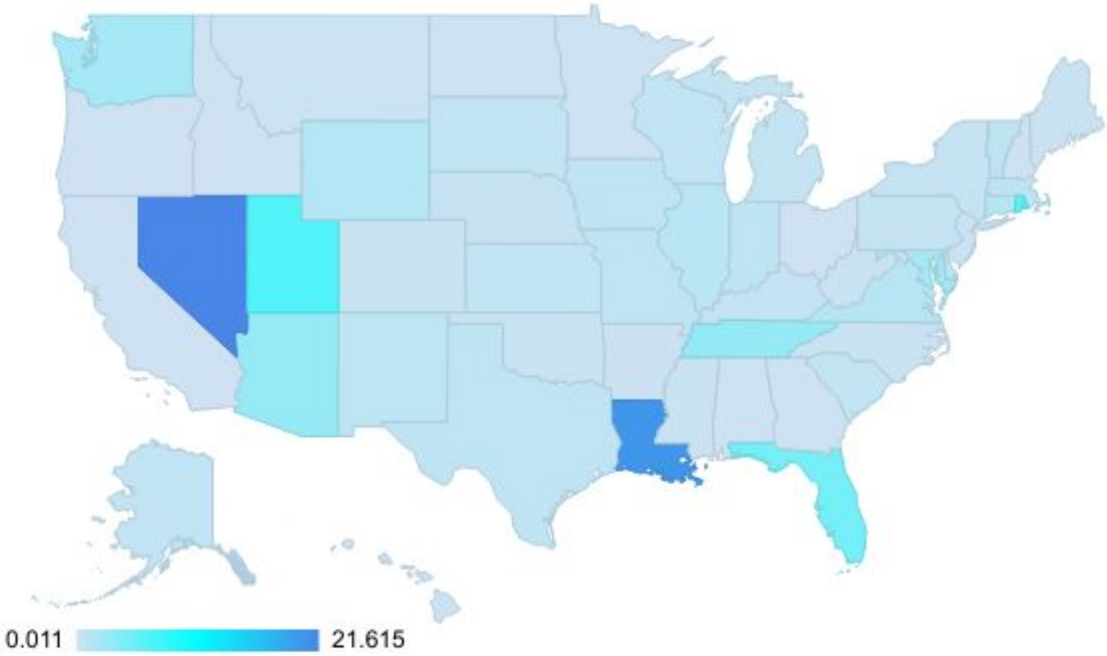
2020 Prediction



2020 in Reality



Prediction’s percentage difference from Reality



Something we also considered

1. More data?
2. Give up categorical for more accuracy?
3. Try more models?

Thanks