Mid-Term

Flight Delays

Vamsi & Akhmet

Operational Cost

Predicting flight delays in US Domestic Market

Operational Efficiency

Customer Satisfaction

Plan at hand

DATA COLLECTION

- US Census Reports
- FAA Report Details
- US Weather Events (2016 - 2020) - Kaggle
- Enplanements data
- Traffic

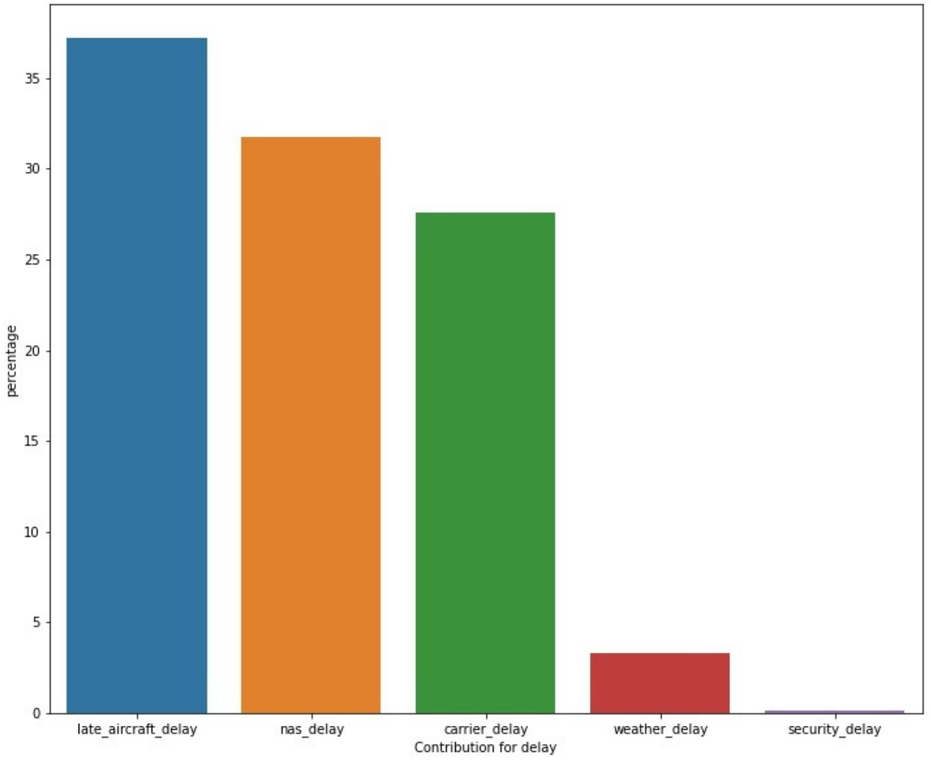
DATA PREPARATION

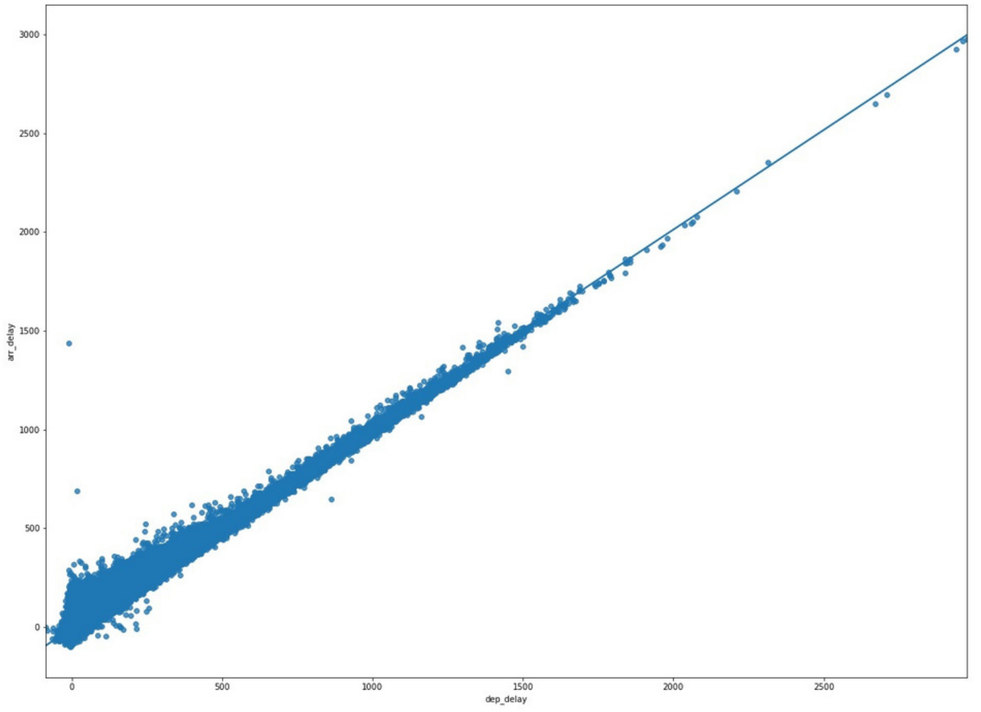
- 2019 Flights Data (7Mil)
- 10% of 2019 (700k+)
- 2% of 2019 (150k+)

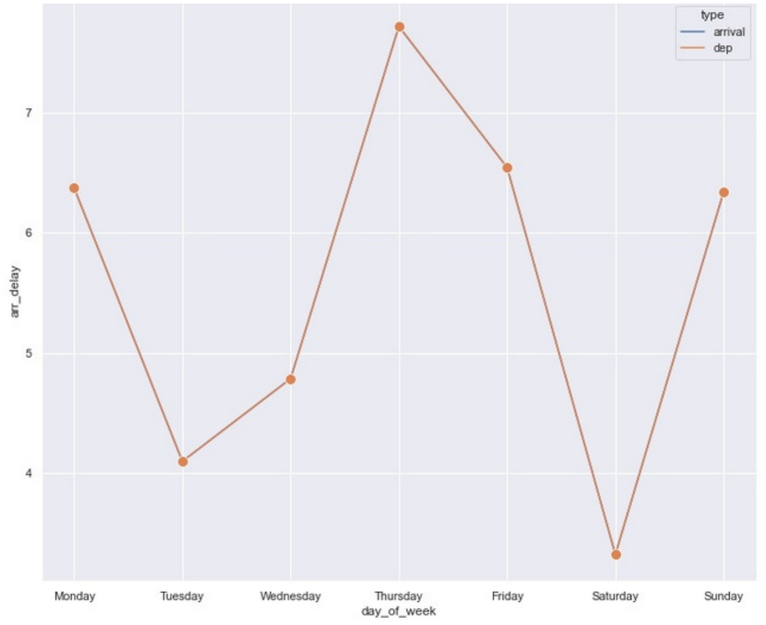
All carriers and airports were equally represented in the sample dataset.

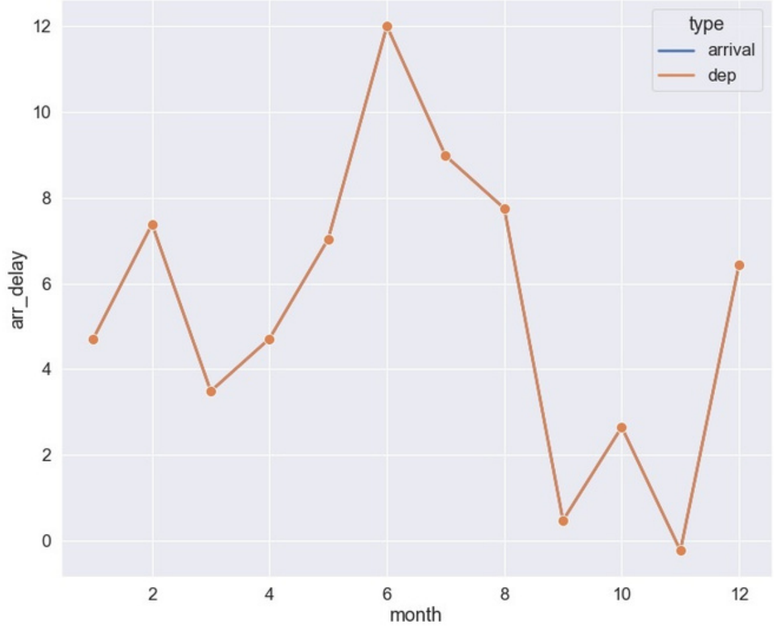
DATA EXPLORATION

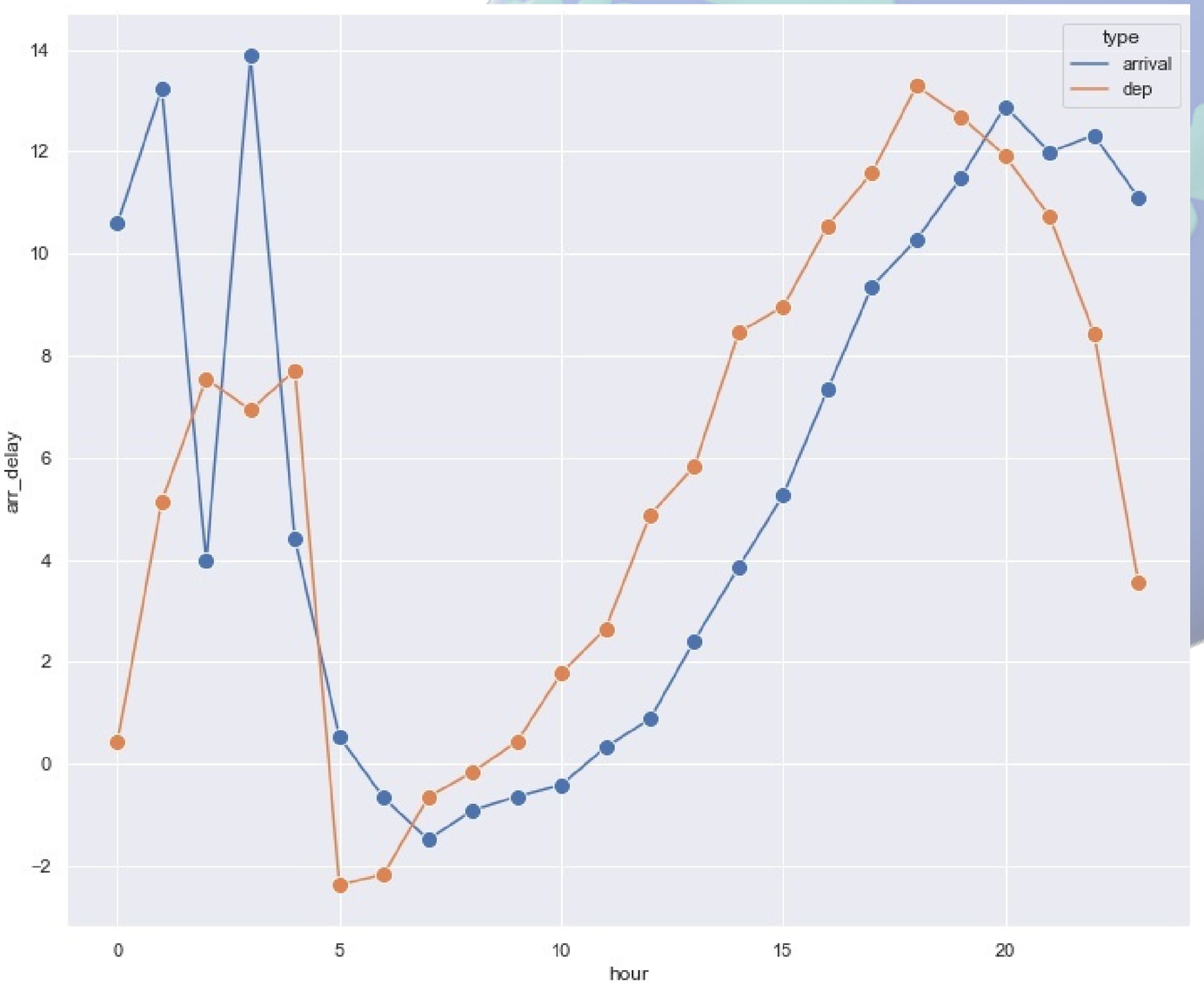
- Main contributor for the delay is Late Arrival
- June & the Rest of the Summer are peak months for delays
- Early mornings have fewer delays
- On average, each carrier arrives on time


percentage
0
5
10
15
20
25
30
35
late_aircraft_delay
nas_delay
carrier_delay
weather_delay
security_delay
Contribution for delay

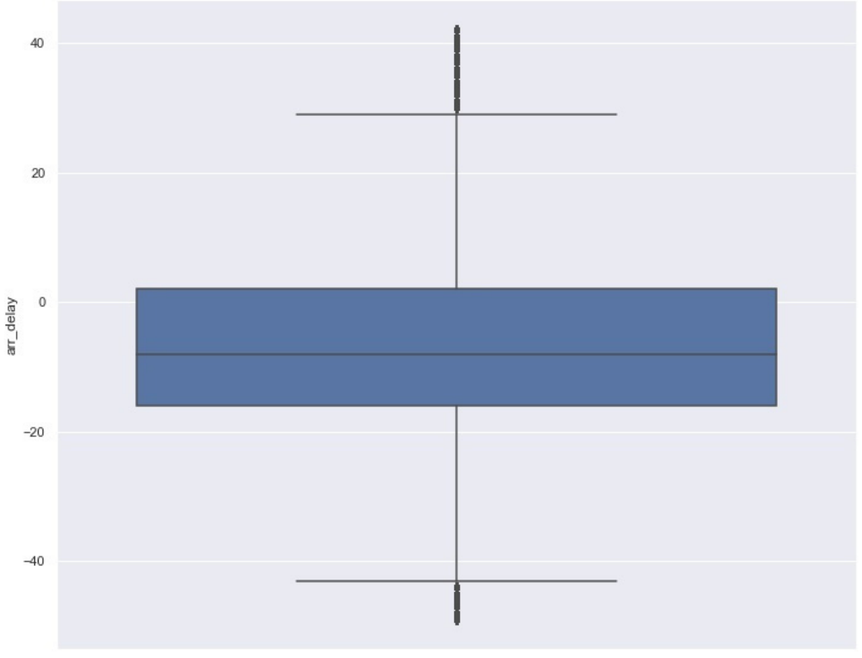

arr_delay
dep_delay
0
500
1000
1500
2000
2500
3000

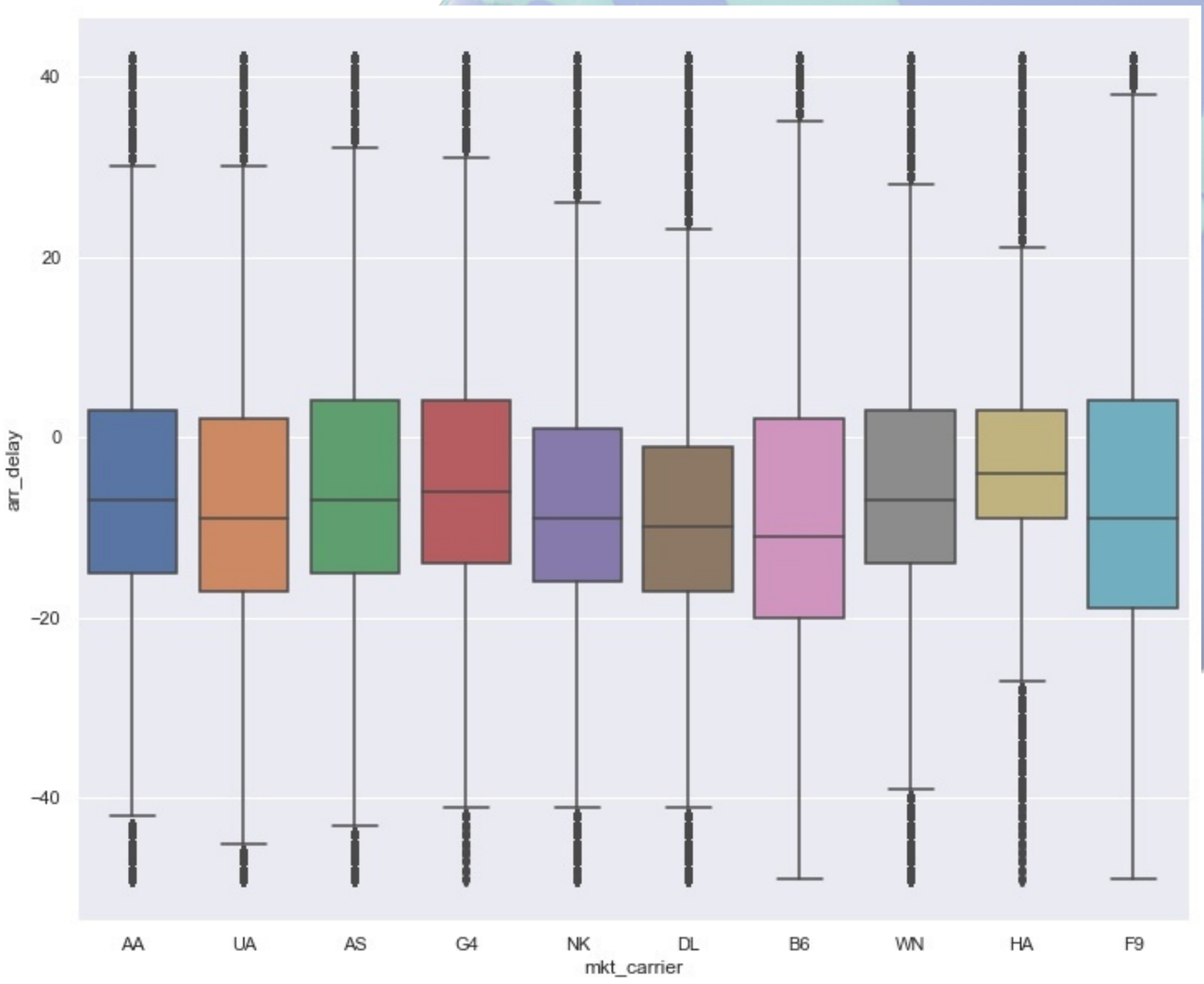

type
arrival
dep
arr_delay
7
6
5
4
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
day_of_week


type
arrival
dep
arr_delay
12
10
8
6
4
2
0
2
4
6
8
10
12
month

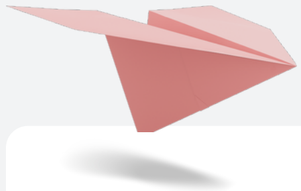

type
arrival
dep
arr_delay
14
12
10
8
6
4
2
0
−2
0
5
10
15
20
hour


arr_delay
3000
2500
2000
1500
1000
500
0


arr_delay
40
20
0
−20
−40

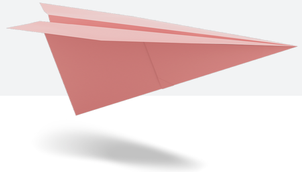

arr_delay
40
20
0
−20
−40
AA
UA
AS
G4
NK
DL
B6
WN
HA
F9
mkt_carrier

Plan at hand II



MODELING

- 'Arrival Delay' Regression
- Multinomial 'Delay' Classification
- Binary 'Cancelled' Classification



MODEL REFINEMENT

- Hyper Parameter Tuning
- Feature Selection
- Feature Extraction
- Feature Transformation
- Resampling

RESULTS

- Best Results achieved with XGB in both Classification and Regregression problems
- R2 Regression = 0.0700
- F1 Delay (5class) = 0.6744
- F1 Cancelled = 0.9906

Regression 'Arrival Delays'

VERSION 1

- 10% 2019 - 700k - FAA (Total~
- Linear = 0.0312
- XGB = 0.0403

VERSION 2

- 80k - Weather - FAA (Total~ 35)
- Linear = 0.0412
- XGB = 0.0683

VERSION 3

- 2% 2019 - 132k - Weather - FAA Enplanements (Total~89-48)
- Linear = 0.0134
- XGB = 0.0284

VERSION_FINAL

- 80k - Weather - FAA - Enplanements - hand picked(Total~26)

Arr_Delay_Regression	R2_SCORE	MAE	MSE
LinearRegression	0.0611	12.37	246.82
RandomForestRegressor	0.0307	12.54	254.82
XGBRegressor	0.0700	12.26	244.49

Classification 'Delays'

- weather_delay
- nas_delay
- security_delay
- late_aircraft_delay
- carrier_delay

SMOTE("all") 13k -> 26k

Delays_Multinomial	ACCURACY	F1_SCORE	PRECISION SCORE	RECAL SCORE
LogisticRegression	0.5340	0.5138	0.5075	0.5340
KNeighborsClassifier	0.6531	0.6335	0.6292	0.6531
GaussianNB	0.3734	0.2690	0.3792	0.3734
RandomForestClassifier	0.5836	0.5519	0.5593	0.5836
XGBClassifier	0.6799	0.6744	0.6754	0.6799

Classification 'Cancelled'

SMOTE("minority") 150k -> 295k

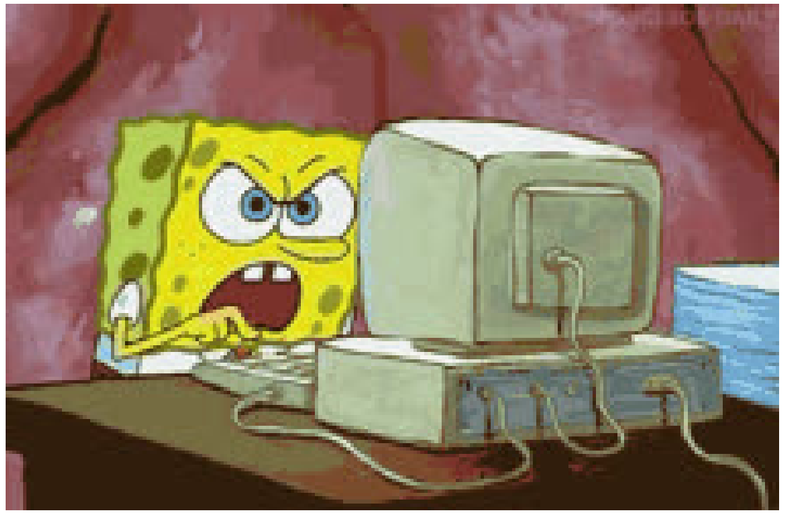
Cancelled_Binary	ACCURACY	F1_SCORE	PRECISION SCORE	RECAL SCORE
LogisticRegression	0.6977	0.6968	0.6999	0.6977
KNeighborsClassifier	0.9499	0.9498	0.9544	0.9499
GaussianNB	0.6098	0.5599	0.6993	0.6098
RandomForestClassifier	0.8109	0.8101	0.8164	0.8109
XGBClassifier	0.9906	0.9906	0.9907	0.9906

Conclusion

- Data Matters (ex. Kaggle Weather set matched with 13 %)
- Finding the Right Features is the hardest part
- More Data is not always better (700k - 120k - 80k)
- Be Aware of Computational Power - Utilize time in between.
- Be Organized - Things Get Dirty
- Leave GridSearch for the end - Basic Models - Priorities

GitHub


HAVE TO WAIT A BIT
FOR GRIDSEARCHCV() TO FINISH
2 MONTHS LATER




"MORE DATA THE BETTER"
EVERYTHING ABOVE 500K