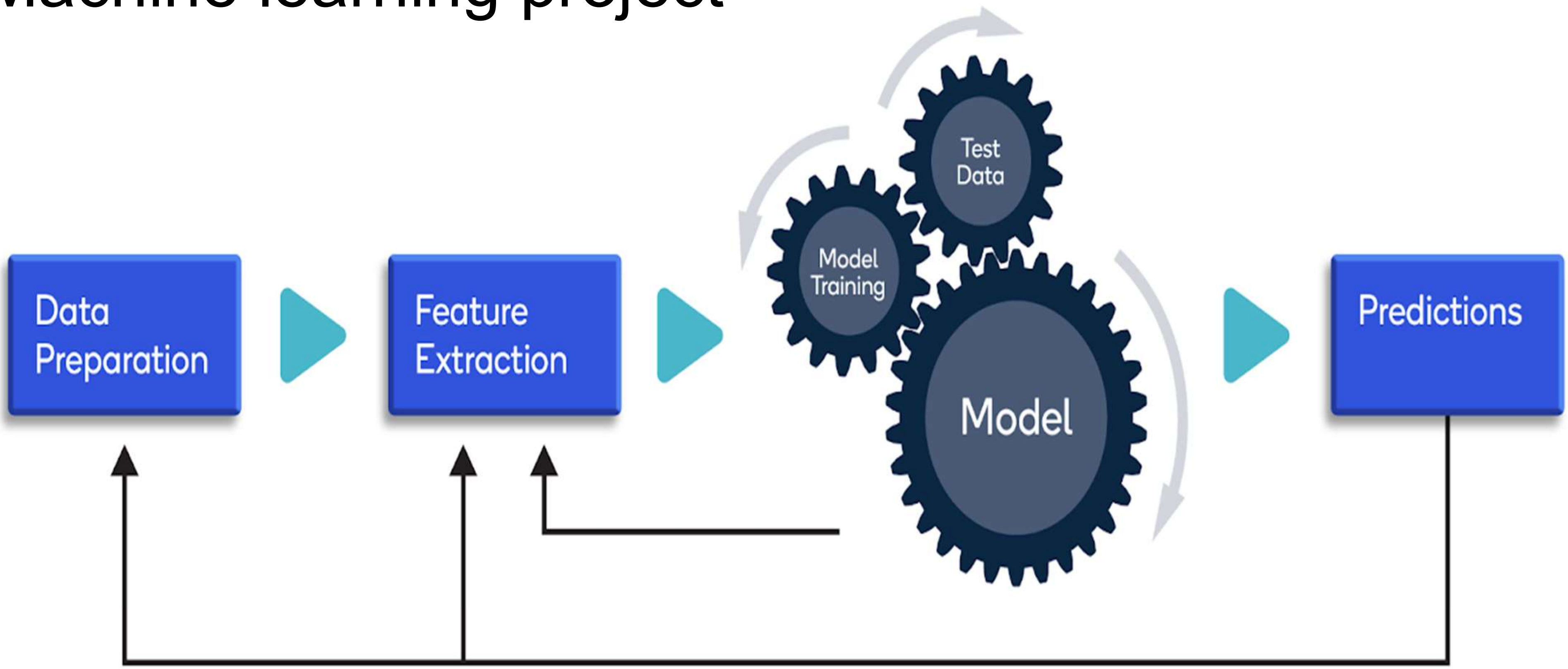
Machine learning project



Asmaa Chraibi & Anderson Chris

Presentation

I - Insights on the dataset

II - Features engineering

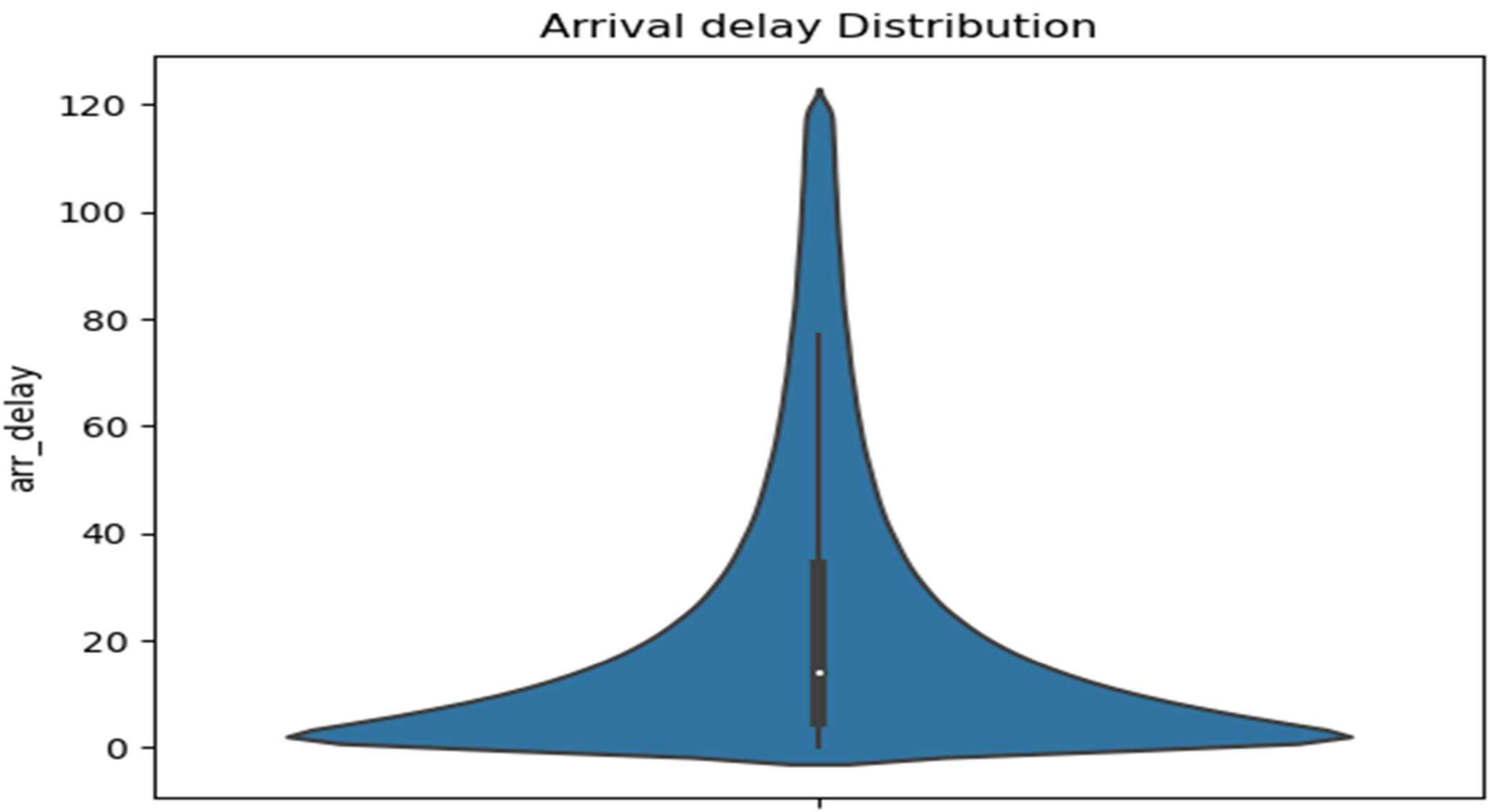
III - Model presentation

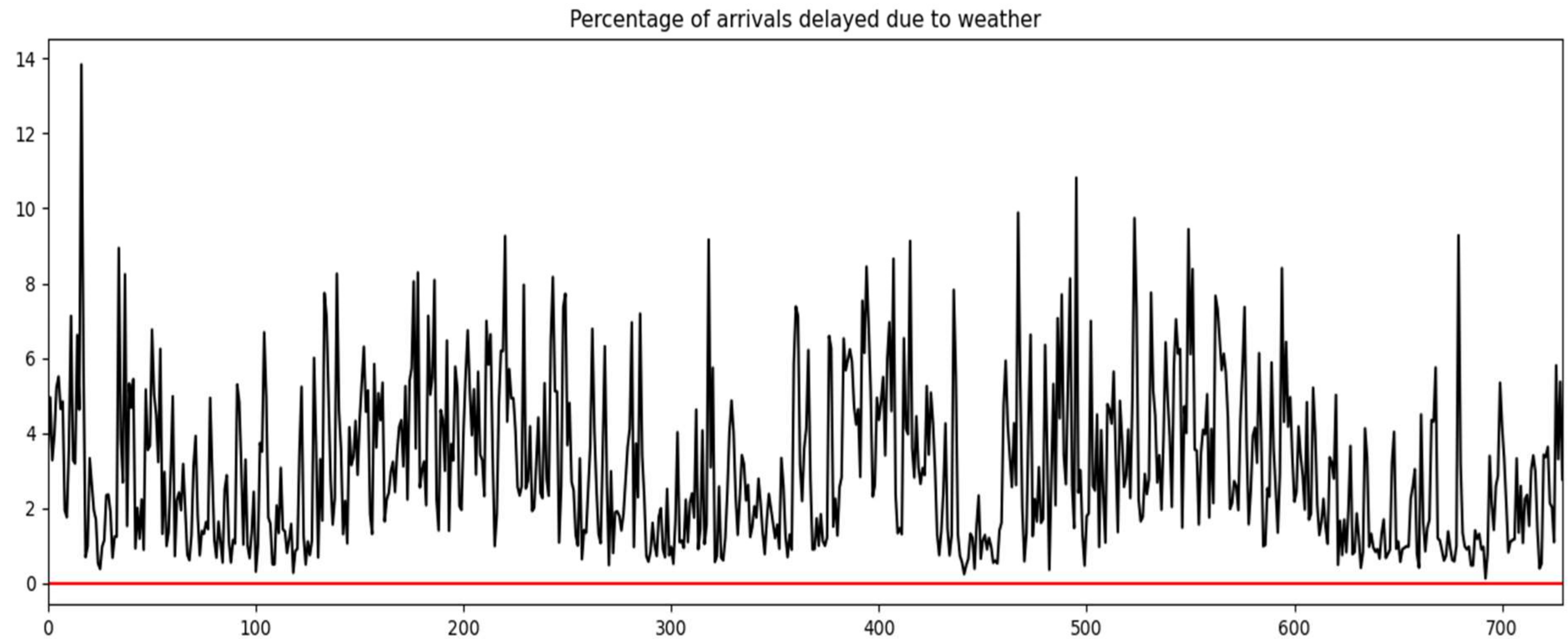
IV - Challenges

Insights on the dataset

❖ Weather and cancellation

	cancelled	weather_delay
mkt_unique_carrier		
AA	0.024995	3.782428
UA	0.019588	4.388874
VX	0.018786	0.530233
WN	0.018213	1.405329
F9	0.017311	0.877153
B6	0.014498	1.914842
NK	0.013025	2.497221
AS	0.012302	1.398075
HA	0.007112	1.633382
DL	0.006930	6.296495
G4	0.006626	4.888876


Arrival delay Distribution
arr_delay
120
100
80
60
40
20
0


Percentage of arrivals delayed due to weather
14
12
10
8
6
4
2
0
0
100
200
300
400
500
600
700

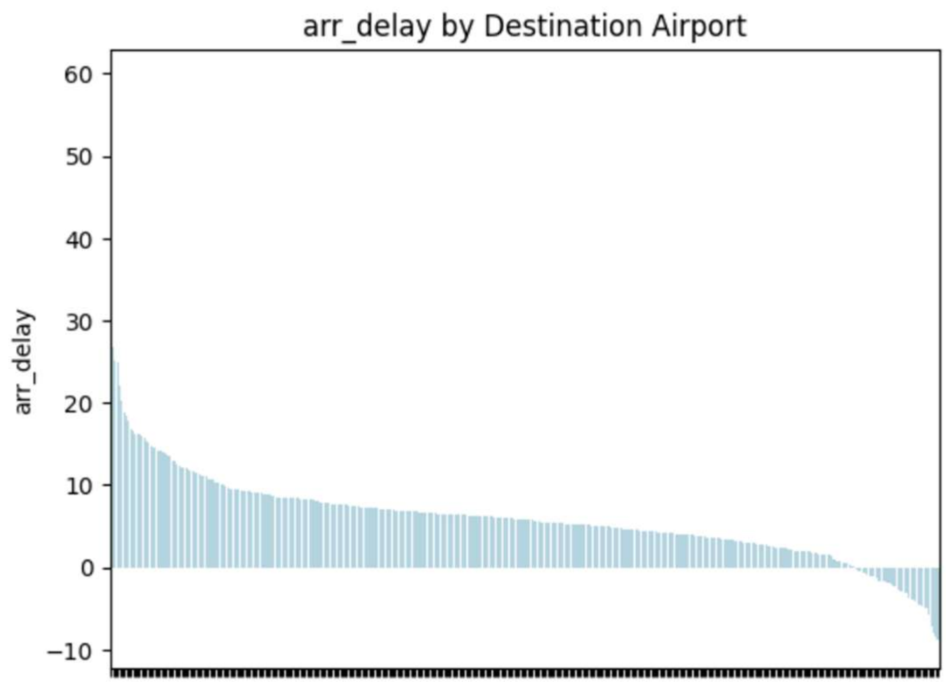
Insights on the dataset

- Flight with late departure have higher airtime speed statistically significant
  - percentage of arrival delays caused by departure delays: 69.39
  - percentage of arrival delays caused by departure delays and compensated during flight: 60.53
  - average percentage of departure delays time compensated:
- Carrier

	mkt_unique_carrier	arr_delay
2	B6	11.328906
4	F9	11.294149
8	UA	9.095866
5	G4	8.948751
0	AA	6.407416
7	NK	5.135043
10	WN	3.549976
3	DL	2.511255
9	VX	1.727978
6	HA	1.245525
1	AS	0.746585

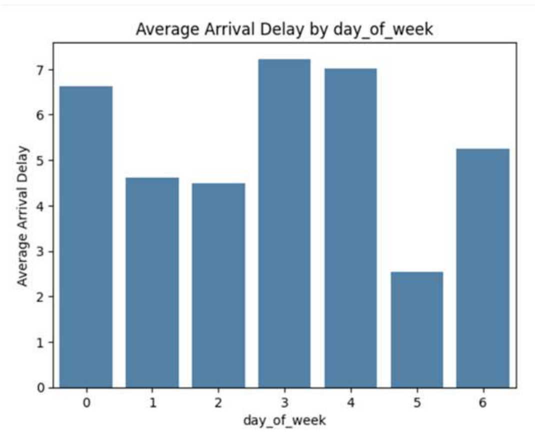
Insights on the dataset

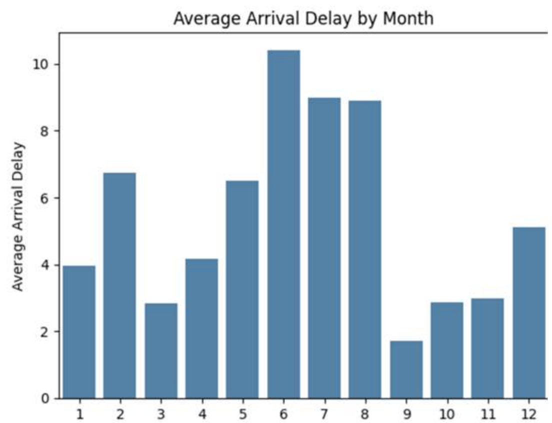
- Airport:

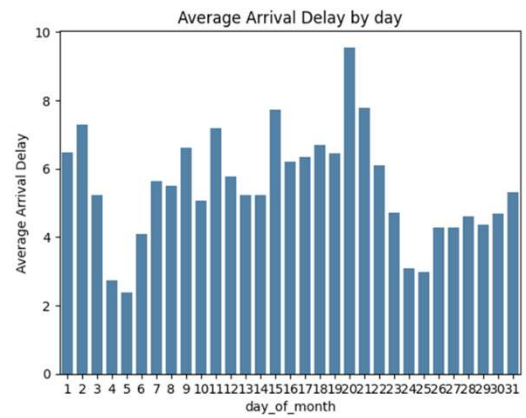


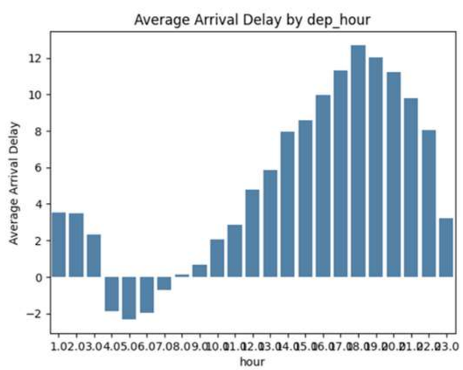
	dest	arr_delay
374	YNG	59.500000
104	DUT	26.706150
281	PPG	25.037344
282	PQI	24.971257
97	DIK	22.056373
77	CMX	20.231227
119	EWR	18.860065
326	SHD	18.470943
260	OTH	17.846262
71	CKB	16.870748

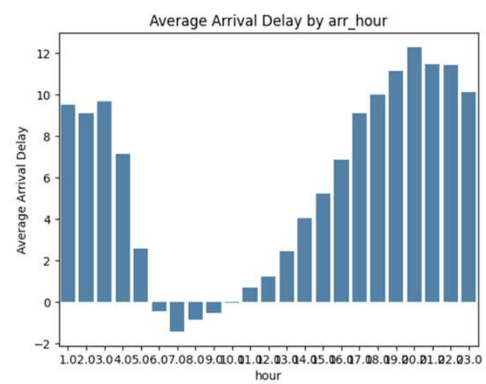
Insights on the dataset:time related











Forecasting arrivals delay magnitude at airports in US

Features considered

	arr_delay	month_day	week_day	dep_hour	arr_hour	origin-dest-mean	month_carrier_mean
0	-19.0	12	3	13.0	17.0	26.84	-0.50
1	-17.0	30	0	15.0	18.0	-1.84	-0.86
2	7.0	14	4	15.0	17.0	6.81	2.55
3	-21.0	2	5	15.0	4.0	-1.48	10.71
4	-10.0	28	4	11.0	12.0	0.81	-1.39

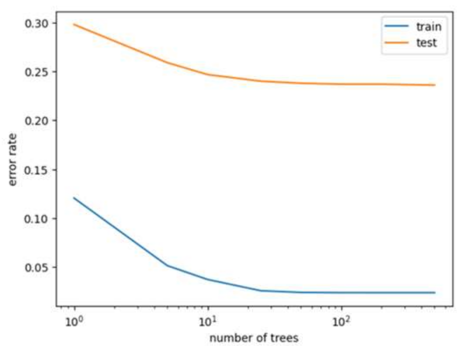
Forecasting arrivals delay magnitude at airports in US

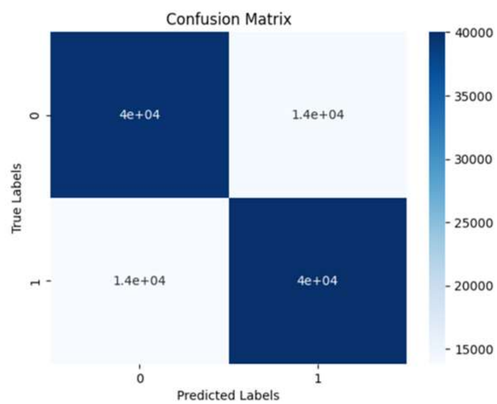
```
ElasticNet regression: R_squared of: 0.031512609451930705
Random Forest Regression: Negatif score
XGBoost Regressor
Train R^2:    0.08507206011548818
Test R^2:    0.04028756208307038

The best Performing Model was a the lineaire regression with:
Train R^2:    0.040355745822241995
Test R^2:    0.039288631080021674
```

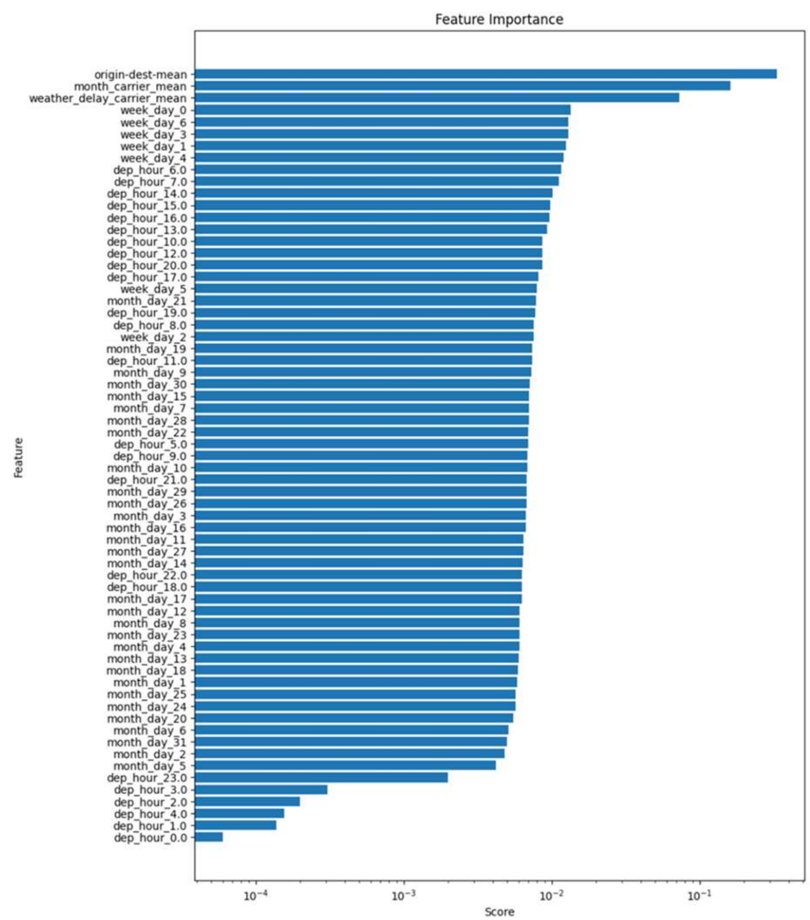
Final cancellation Model:
Features Selection and Balancing the classes

	cancelled	month_day	week_day	dep_hour	weather_delay_carrier_mean	origin-dest-mean	month_carrier_mean
0	1	24	1	10.0	0.75	0.70	0.36
1	1	13	6	10.0	0.20	0.39	0.55
2	1	13	1	10.0	0.34	0.67	0.64
3	1	6	1	14.0	0.34	0.77	0.64
4	1	13	2	7.0	0.11	0.67	0.67

Final cancellation Model: Random Forest Classifier








```
Training Accuracy: 0.9492422202204619
Testing Accuracy: 0.7402492477991159
```

```
Classification Report (Default Threshold):
              precision    recall  f1-score   support

           0       0.74      0.74      0.74     53825
           1       0.74      0.74      0.74     53859

    accuracy                           0.74    107684
   macro avg       0.74      0.74      0.74    107684
weighted avg       0.74      0.74      0.74    107684
```

Challenges & future steps

- Size of Data and Memory Ram limitation
- More Features engineering
- Data on Administrative constraints by airports
- Data on Carrier limitations (Aircraft, Staff..)
- Learn to use google collab
- Try Ensemble models
- Try predicting delays at airport level