Airlines

Predicting Delays

W261 Section 1 Team 3

Tony Di Sera, Ammara Essa, Andy Hoopengardner, Lee Moore

Question Formation

Data

Airlines Data
~32 million rows
2015-2019 US Commercial Flights

Weather Data
~620 million weather observations
2015-2019

Goal of Analysis

Binary classification task predicting:
“detrimention flight outcomes”

across all USA flights

Question Formation

Outcome Variable

Composite outcome:

- Arrival delays > 15 minutes
- Cancelled flights
- Diverted flights

Evaluation Metrics

Key metric:	Recall
Secondary:	F1-score

Baseline

Crude prediction of “On time” for every flight:

- Accuracy = 80%
- Recall/Precision = 0%

Aspiration:

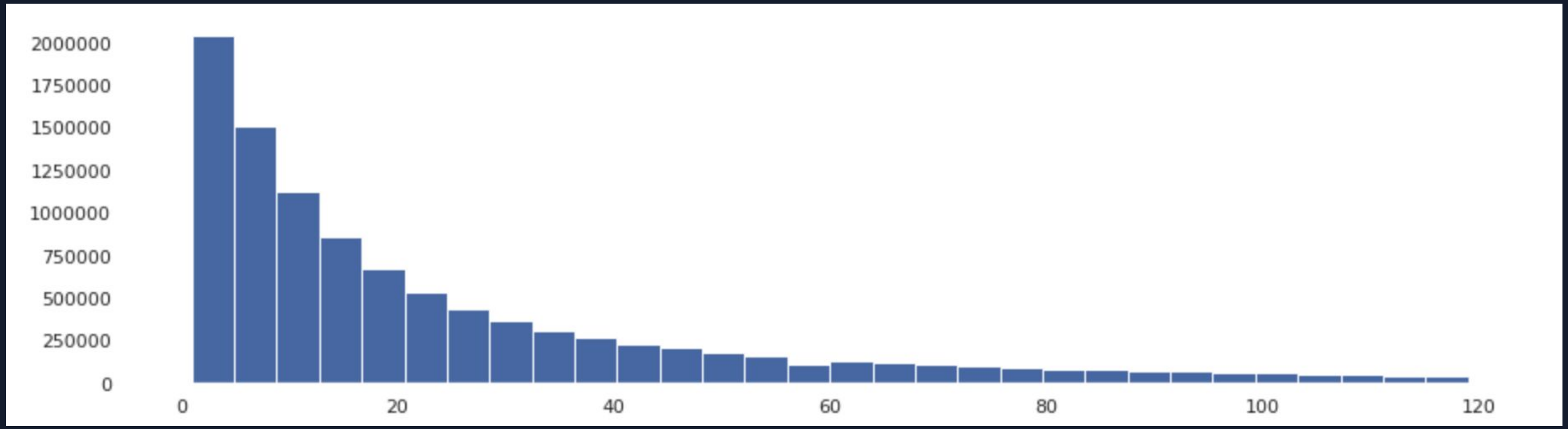
- Recall > 80%

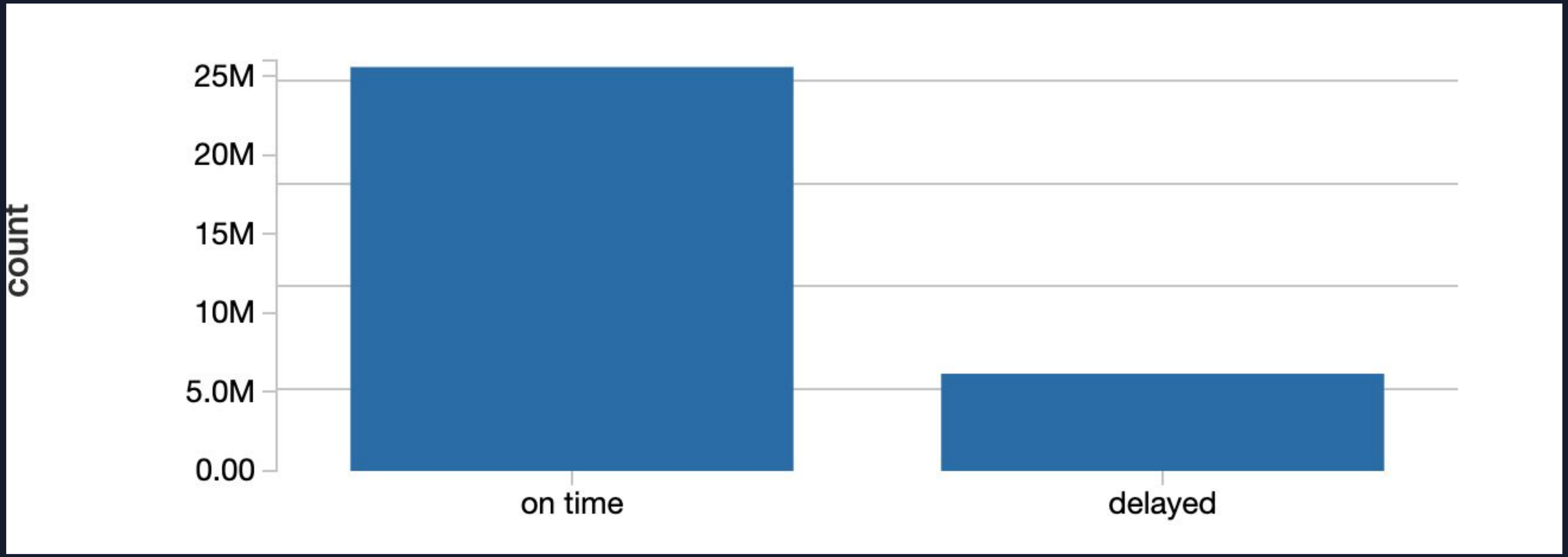
EDA

EDA

1

Is the outcome variable balanced?

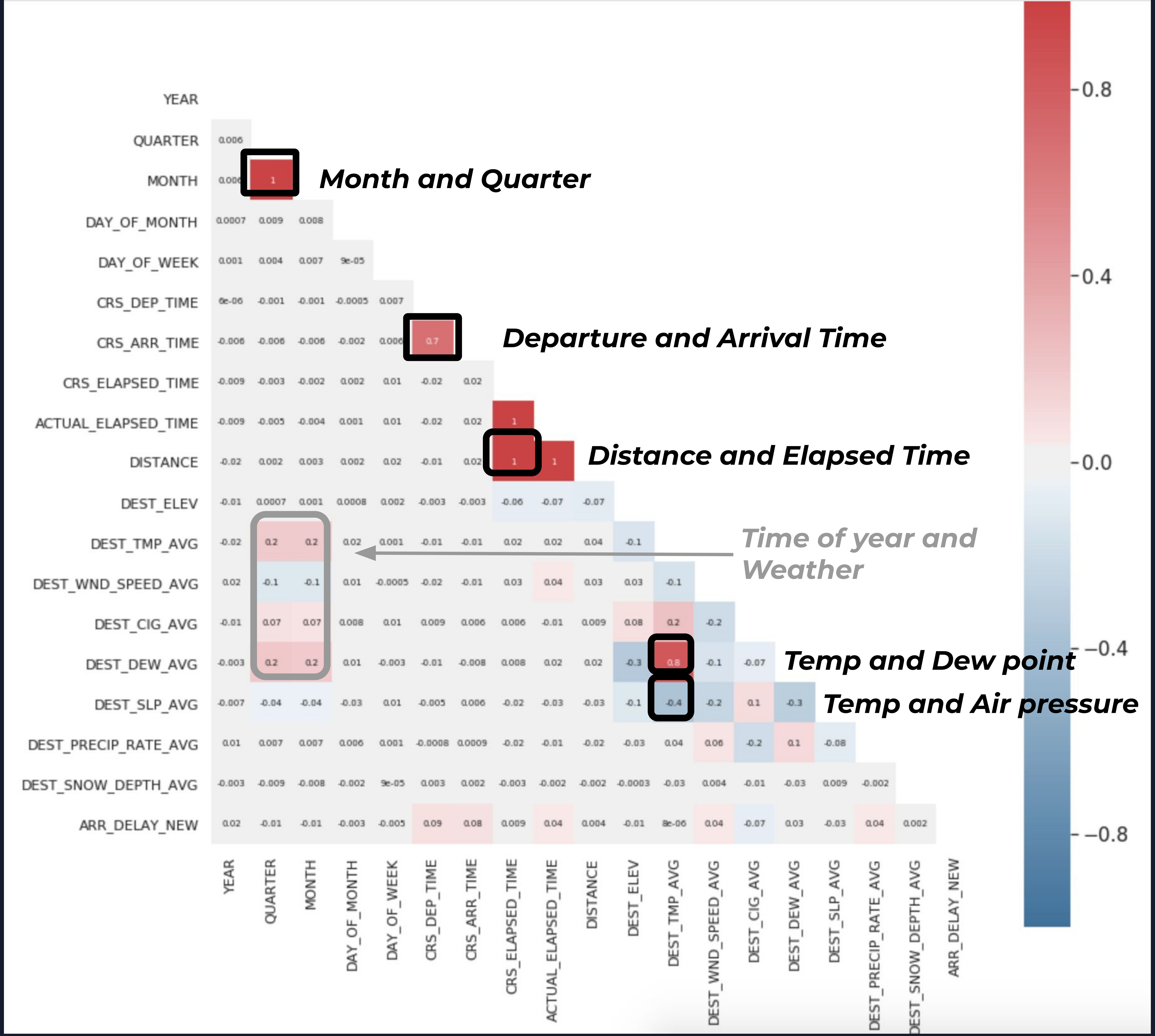




EDA

2

Are continuous variables highly correlated?



	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	CRS_DEP_TIME	CRS_ARR_TIME	CRS_ELAPSED_TIME	ACTUAL_ELAPSED_TIME	DISTANCE	DEST_ELEV	DEST_TMP_AVG	DEST_WND_SPEED_AVG	DEST_CIG_AVG	DEST_DEW_AVG	DEST_SLP_AVG	DEST_PRECIP_RATE_AVG	DEST_SNOW_DEPTH_AVG	ARR_DELAY_NEW
YEAR																			
QUARTER	0.006																		
MONTH	0.006	1																	
DAY_OF_MONTH	0.0007	0.009	0.008																
DAY_OF_WEEK	0.001	0.004	0.007	9e-05															
CRS_DEP_TIME	6e-06	-0.001	-0.001	-0.0005	0.007														
CRS_ARR_TIME	-0.006	-0.006	-0.006	-0.002	0.006	0.7													
CRS_ELAPSED_TIME	-0.009	-0.003	-0.002	0.002	0.01	-0.02	0.02												
ACTUAL_ELAPSED_TIME	-0.009	-0.005	-0.004	0.001	0.01	-0.02	0.02	1											
DISTANCE	-0.02	0.002	0.003	0.002	0.02	-0.01	0.02	1	1										
DEST_ELEV	-0.01	0.0007	0.001	0.0008	0.002	-0.003	-0.003	-0.06	-0.07	-0.07									
DEST_TMP_AVG	-0.02	0.2	0.2	0.02	0.001	-0.01	-0.01	0.02	0.02	0.04	-0.1								
DEST_WND_SPEED_AVG	0.02	-0.1	-0.1	0.01	-0.0005	-0.02	-0.01	0.03	0.04	0.03	0.03	-0.1							
DEST_CIG_AVG	-0.01	0.07	0.07	0.008	0.01	0.009	0.006	0.006	-0.01	0.009	0.08	0.2	-0.2						
DEST_DEW_AVG	-0.003	0.2	0.2	0.01	-0.003	-0.01	-0.008	0.008	0.02	0.02	-0.3	0.8	-0.1	-0.07					
DEST_SLP_AVG	-0.007	-0.04	-0.04	-0.03	0.01	-0.005	0.006	-0.02	-0.03	-0.03	-0.1	-0.4	-0.2	0.1	-0.3				
DEST_PRECIP_RATE_AVG	0.01	0.007	0.007	0.006	0.001	-0.0008	0.0009	-0.02	-0.01	-0.02	-0.03	0.04	0.06	-0.2	0.1	-0.08			
DEST_SNOW_DEPTH_AVG	-0.003	-0.009	-0.008	-0.002	9e-05	0.003	0.002	-0.003	-0.002	-0.002	-0.0003	-0.03	0.004	-0.01	-0.03	0.009	-0.002		
ARR_DELAY_NEW	0.02	-0.01	-0.01	-0.003	-0.005	0.09	0.08	0.009	0.04	0.004	-0.01	8e-06	0.04	-0.07	0.03	-0.03	0.04	0.002	

Month and Quarter

Departure and Arrival Time

Distance and Elapsed Time

Time of year and Weather

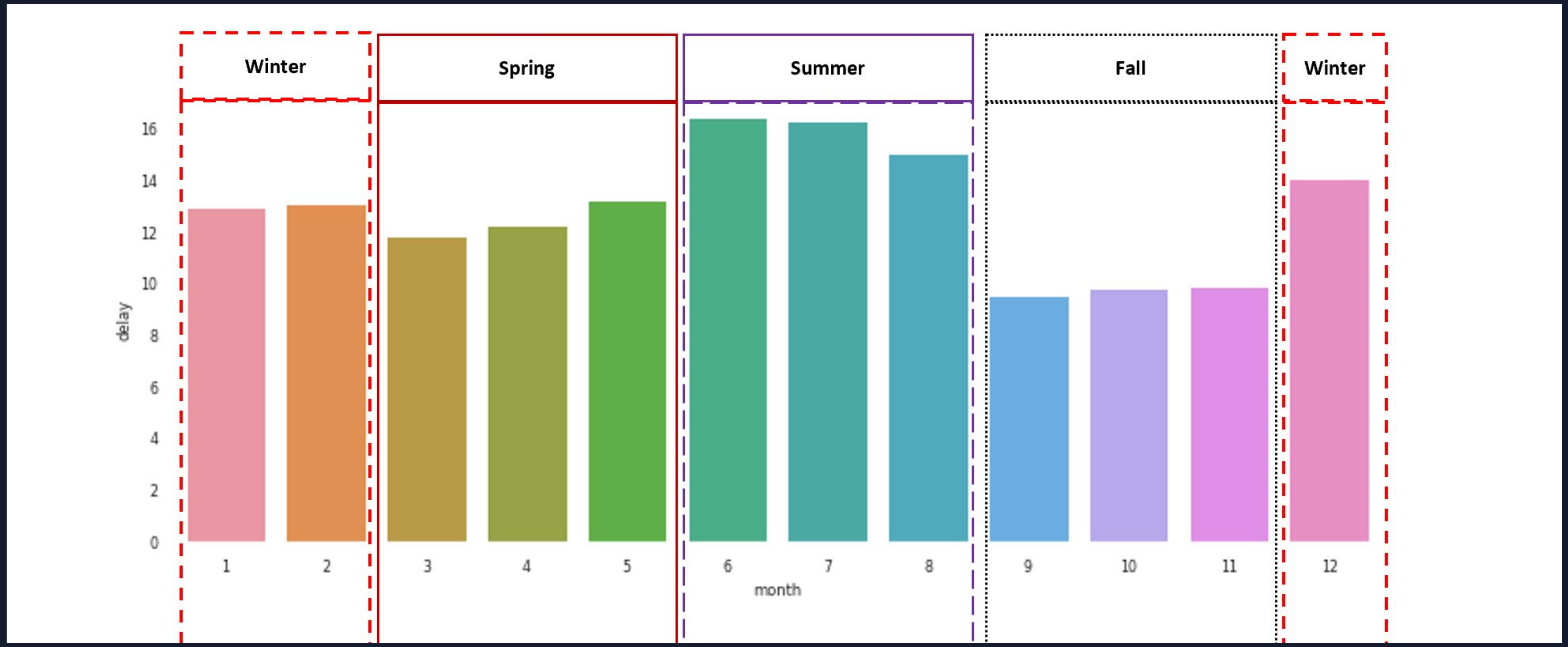
Temp and Dew point

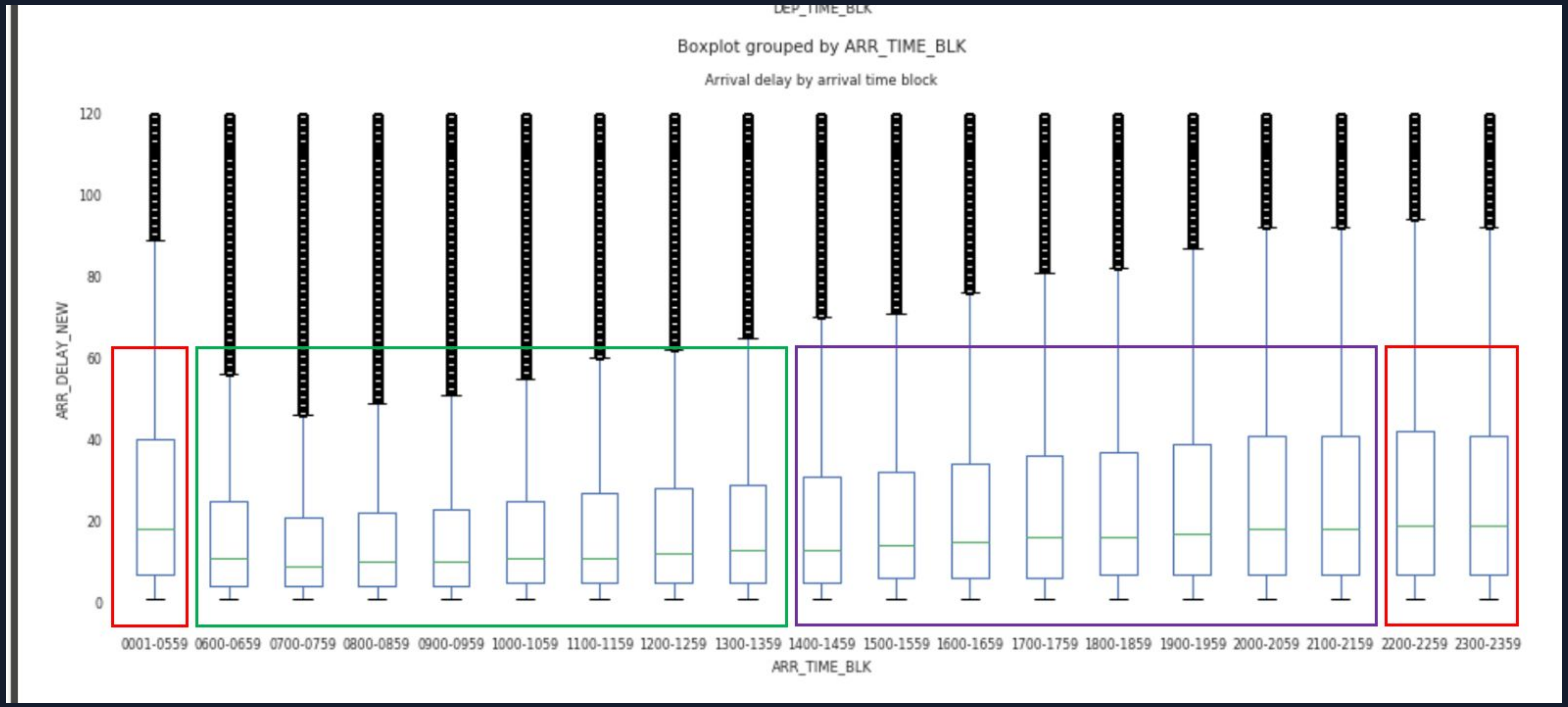
Temp and Air pressure

EDA

3

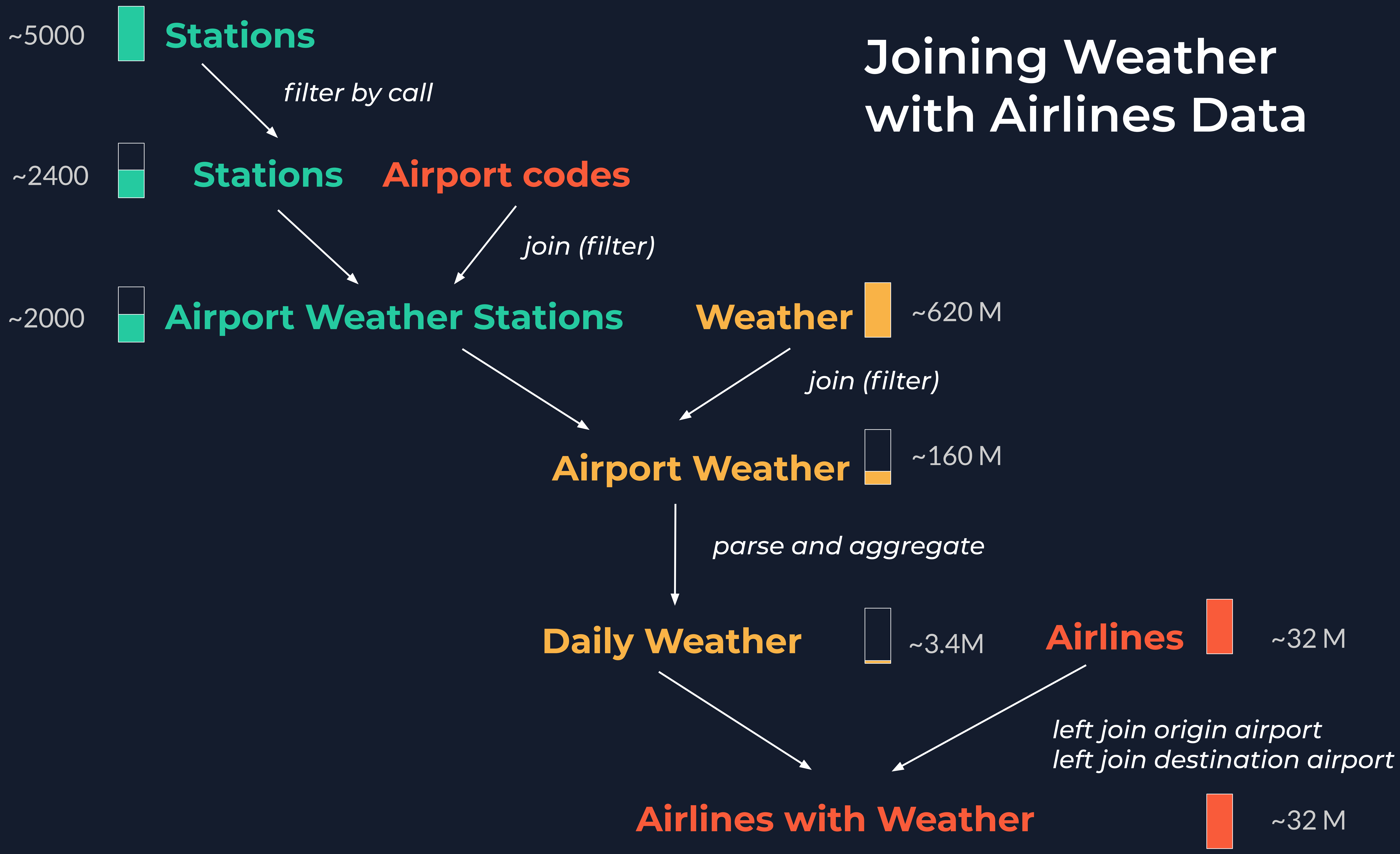
Are there seasonal and time-of-day trends?





Feature Engineering

Joining Weather with Airlines Data



~5000 Stations

filter by call

~2400 Stations

Airport codes

join (filter)

~2000 Airport Weather Stations

Weather ~620 M

join (filter)

Airport Weather ~160 M

parse and aggregate

Daily Weather ~3.4M

Airlines ~32 M

left join origin airport
left join destination airport

Airlines with Weather ~32 M

Feature Engineering

- Exclude features not available at prediction time
- Exclude features that represent outcome
- Graph features

Logistic Regression

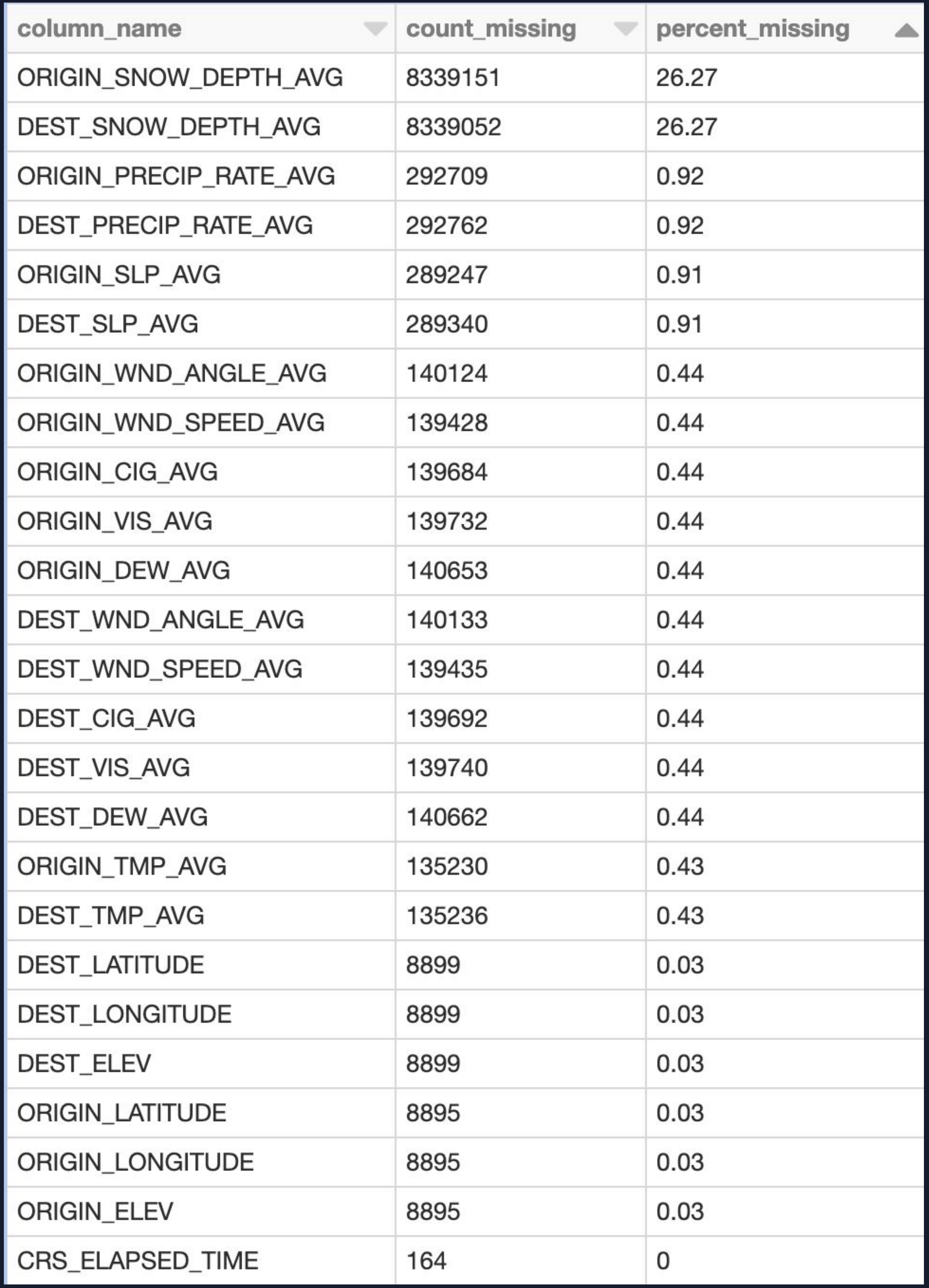
- One-hot encoding
- Standardization
- Exclude highly correlated features

Decision Trees

- Briemans Method

Feature Engineering

Missing Value treatment

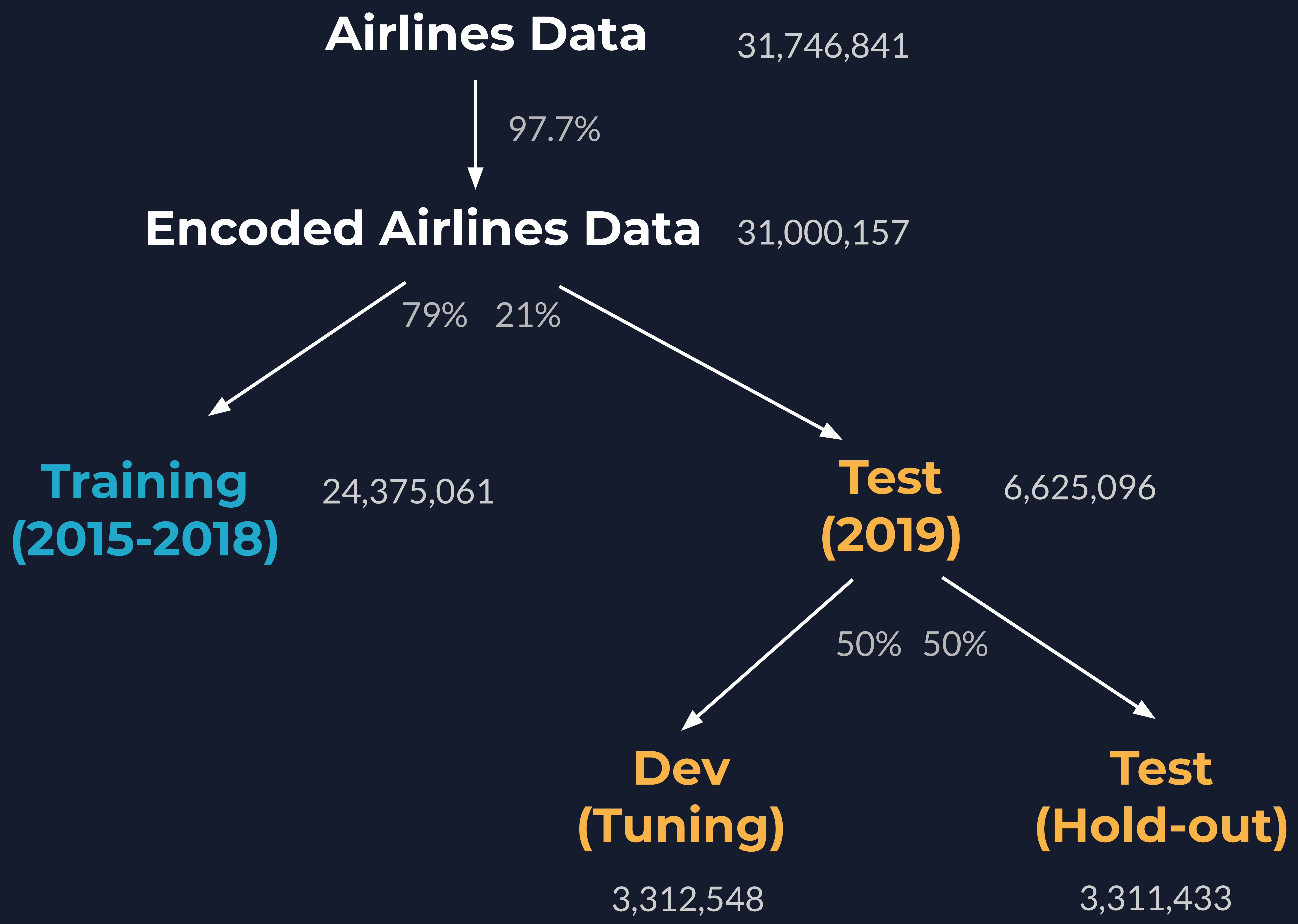


column_name	count_missing	percent_missing
ORIGIN_SNOW_DEPTH_AVG	8339151	26.27
DEST_SNOW_DEPTH_AVG	8339052	26.27
ORIGIN_PRECIP_RATE_AVG	292709	0.92
DEST_PRECIP_RATE_AVG	292762	0.92
ORIGIN_SLP_AVG	289247	0.91
DEST_SLP_AVG	289340	0.91
ORIGIN_WND_ANGLE_AVG	140124	0.44
ORIGIN_WND_SPEED_AVG	139428	0.44
ORIGIN_CIG_AVG	139684	0.44
ORIGIN_VIS_AVG	139732	0.44
ORIGIN_DEW_AVG	140653	0.44
DEST_WND_ANGLE_AVG	140133	0.44
DEST_WND_SPEED_AVG	139435	0.44
DEST_CIG_AVG	139692	0.44
DEST_VIS_AVG	139740	0.44
DEST_DEW_AVG	140662	0.44
ORIGIN_TMP_AVG	135230	0.43
DEST_TMP_AVG	135236	0.43
DEST_LATITUDE	8899	0.03
DEST_LONGITUDE	8899	0.03
DEST_ELEV	8899	0.03
ORIGIN_LATITUDE	8895	0.03
ORIGIN_LONGITUDE	8895	0.03
ORIGIN_ELEV	8895	0.03
CRS_ELAPSED_TIME	164	0

Set null to 0

Skip

Feature Engineering - Observations



Airlines Data 31,746,841

97.7%

Encoded Airlines Data 31,000,157

79% 21%

Training (2015-2018) 24,375,061

Test (2019) 6,625,096

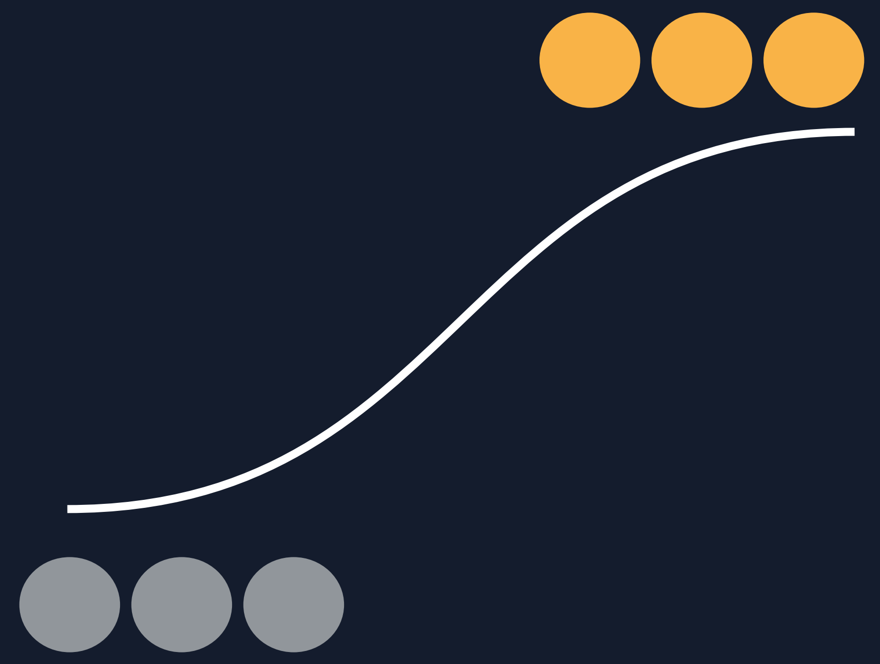
50% 50%

Dev (Tuning) 3,312,548

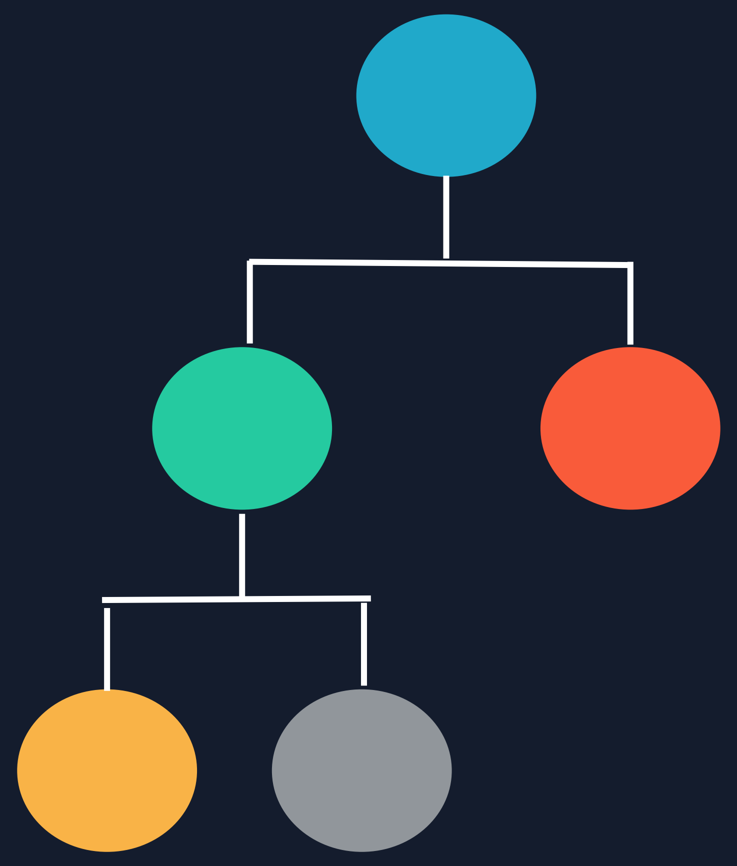
Test (Hold-out) 3,311,433

Algorithmic Exploration

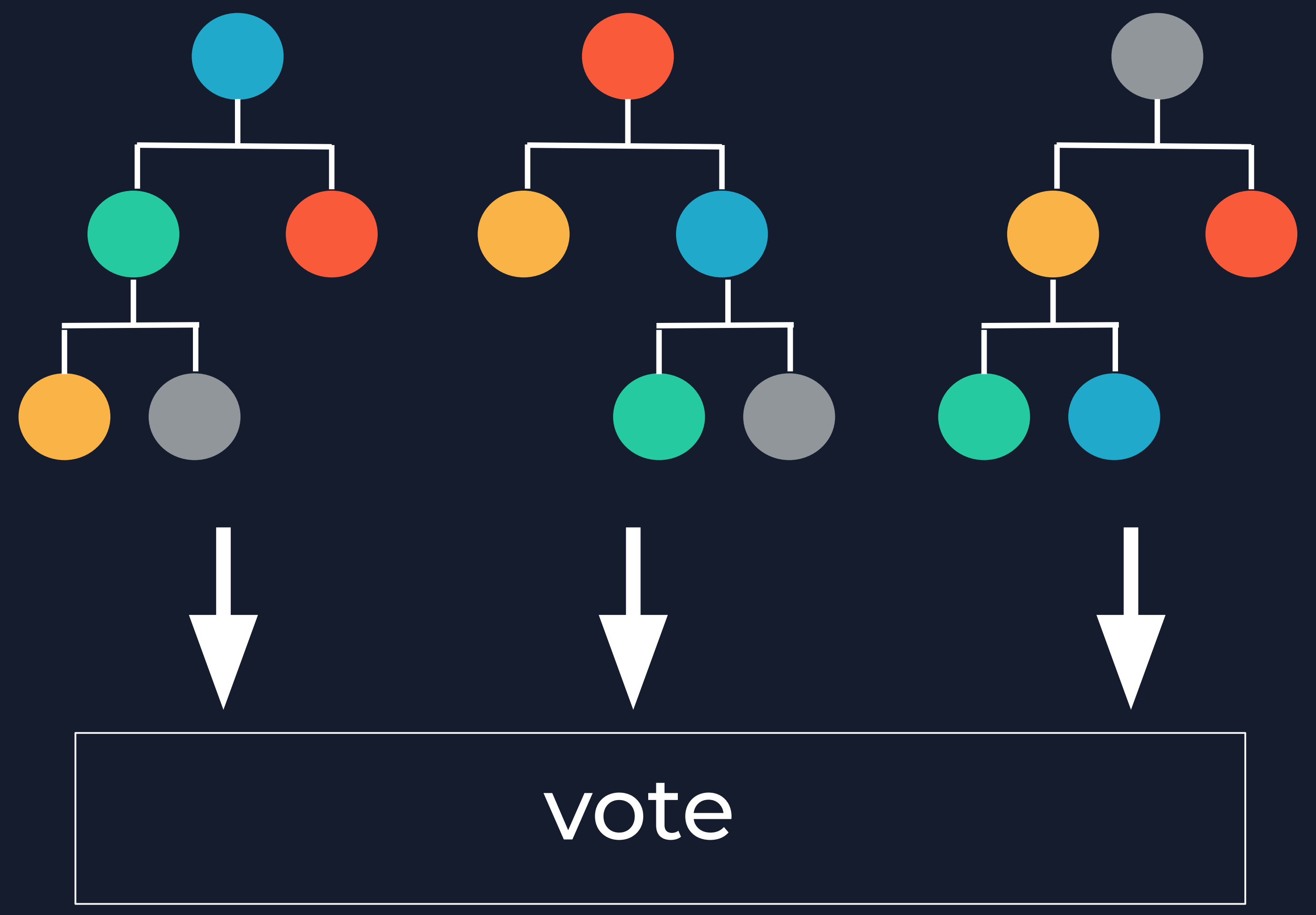
Algorithmic Exploration



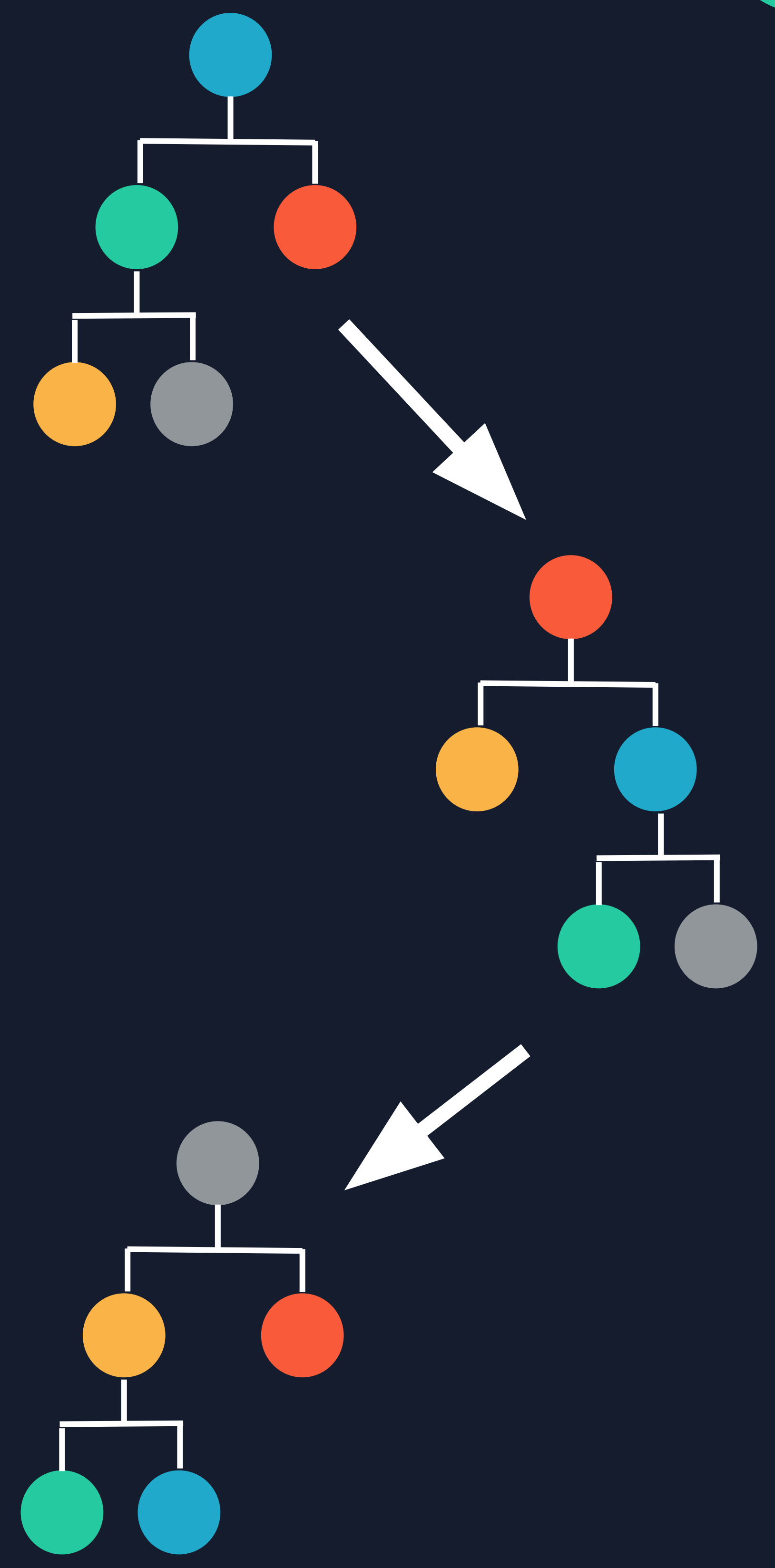
Logistic Regression



Decision Trees



Random Forest



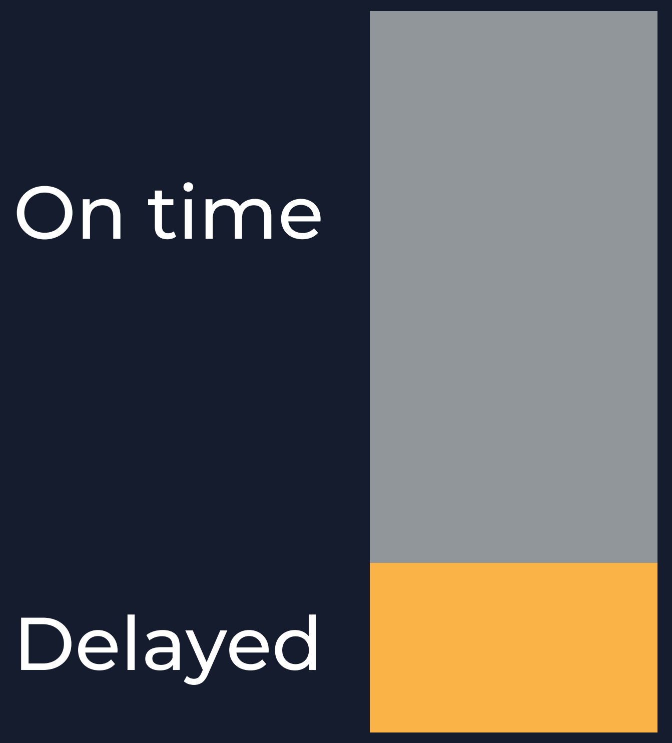
Gradient Boosted Tree

- *Gridsearch and crossfold validation (CV)*
- *Human-based gridsearch*
- *Regularization*

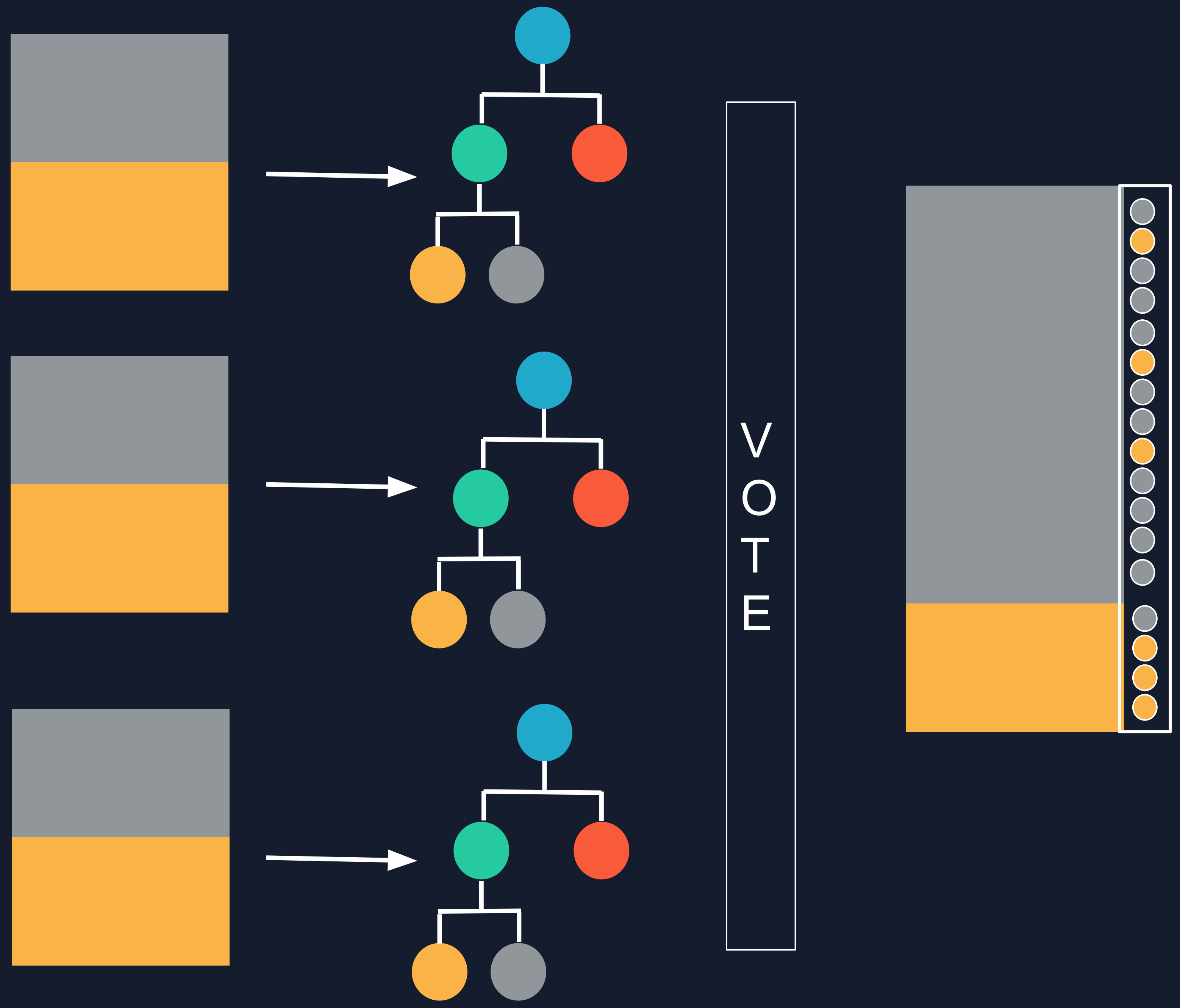
- *Tree Depth*
- *Trees in Forest*
- *Iterations for GBT*

Algorithmic Exploration

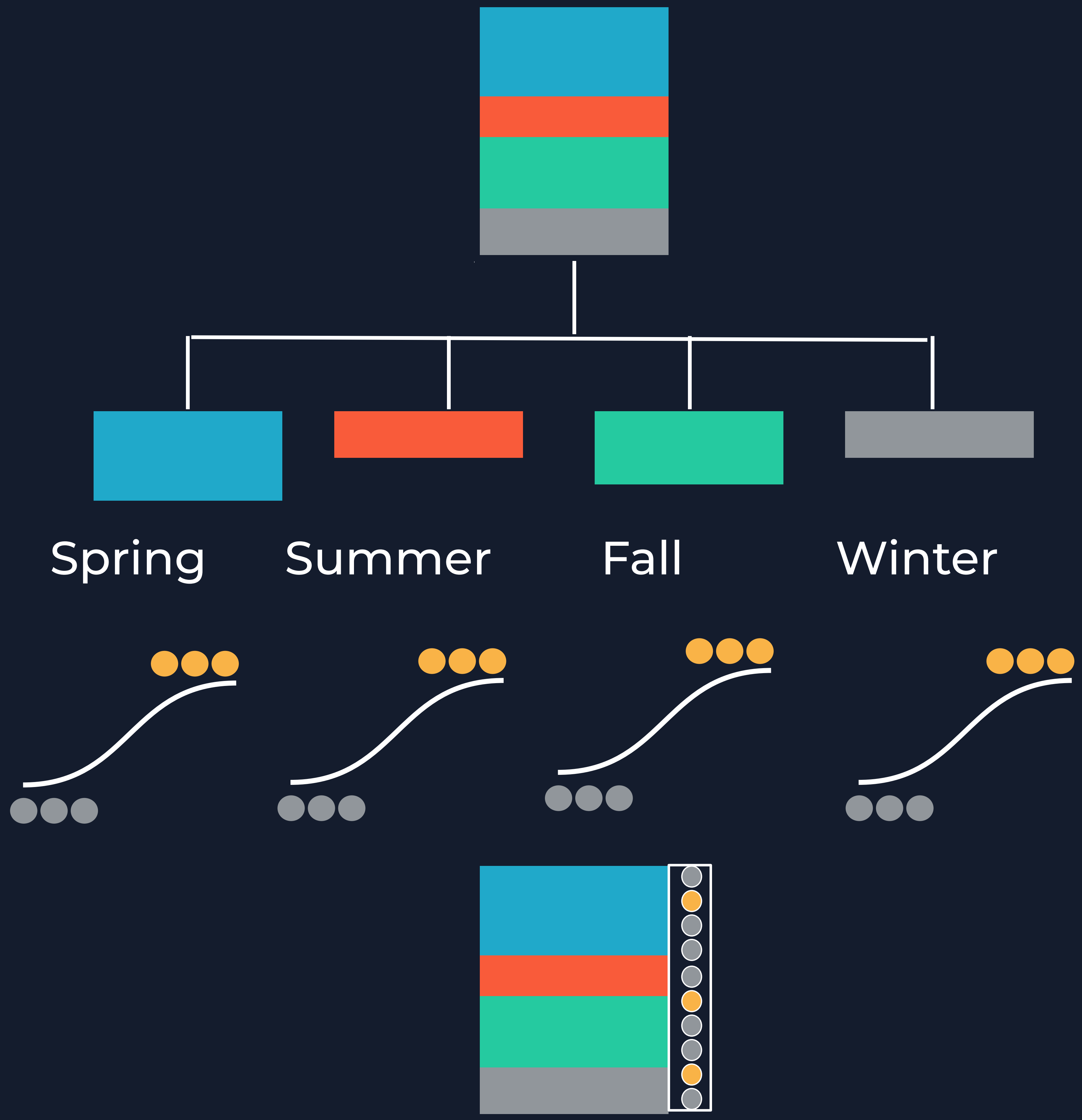
Balance



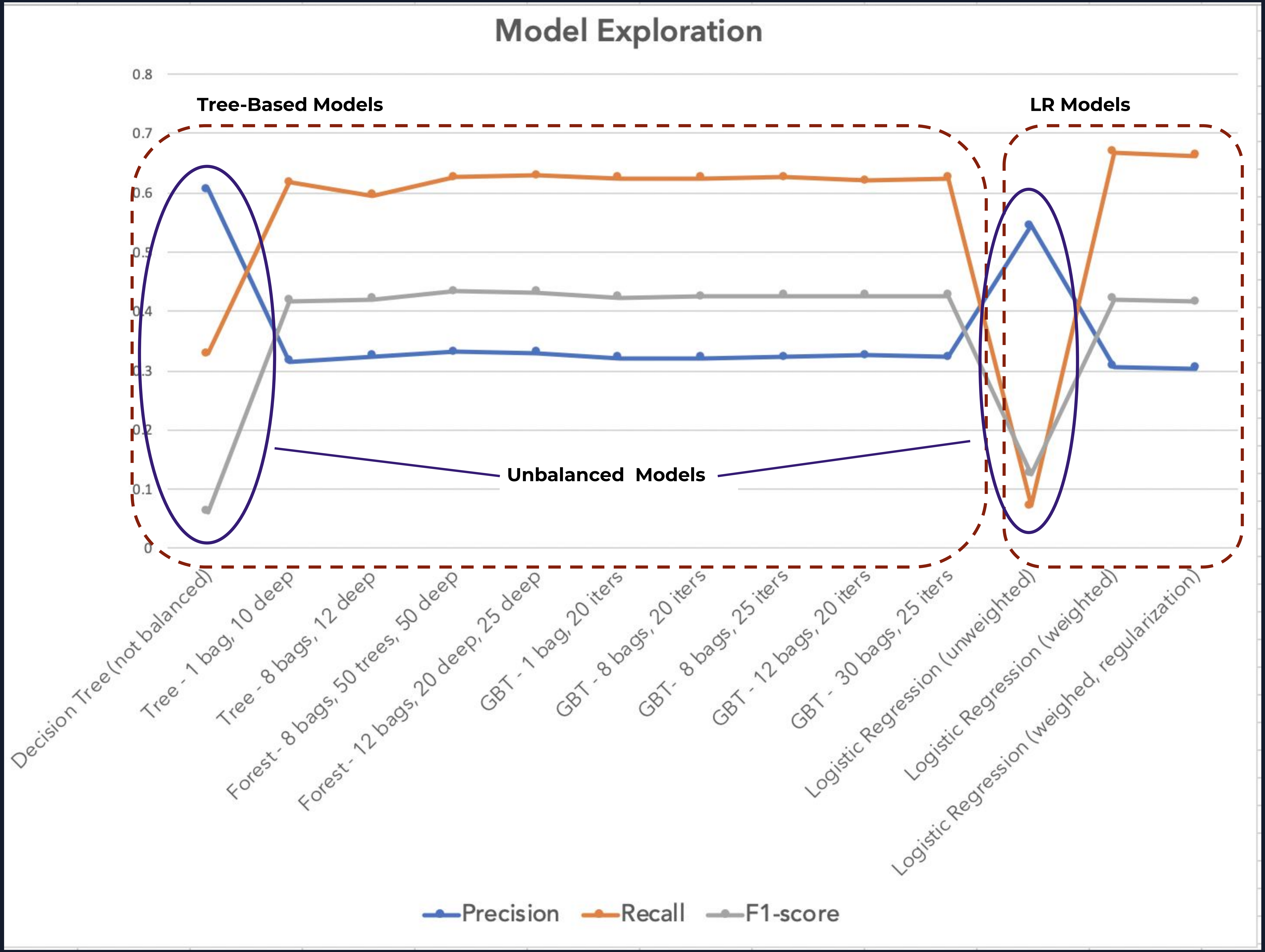
Bootstrap Aggregating (Bagging)



Ensemble of LR Models



Algorithmic Selection and Tuning



Model Exploration

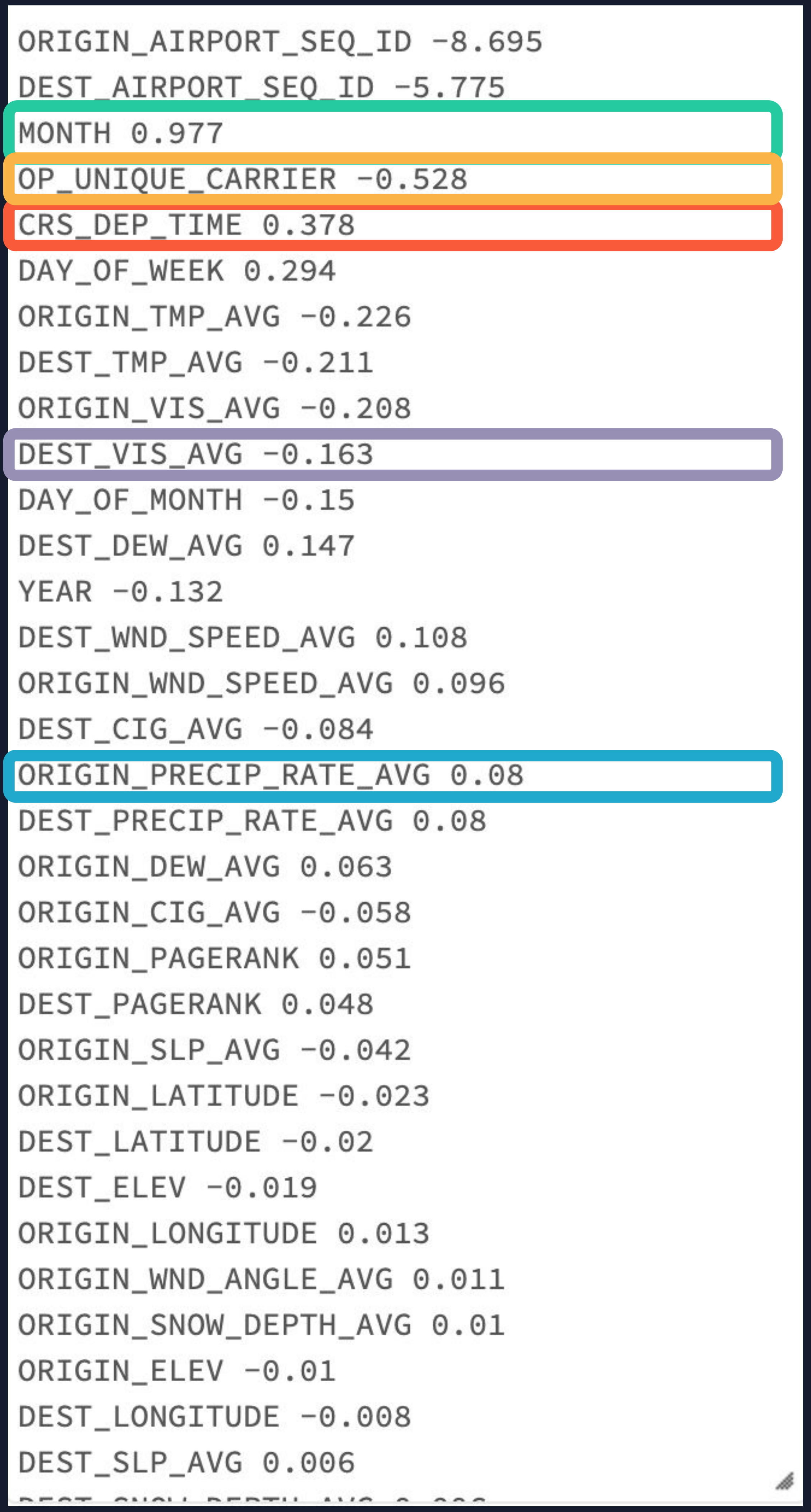
Tree-Based Models

LR Models

Unbalanced Models

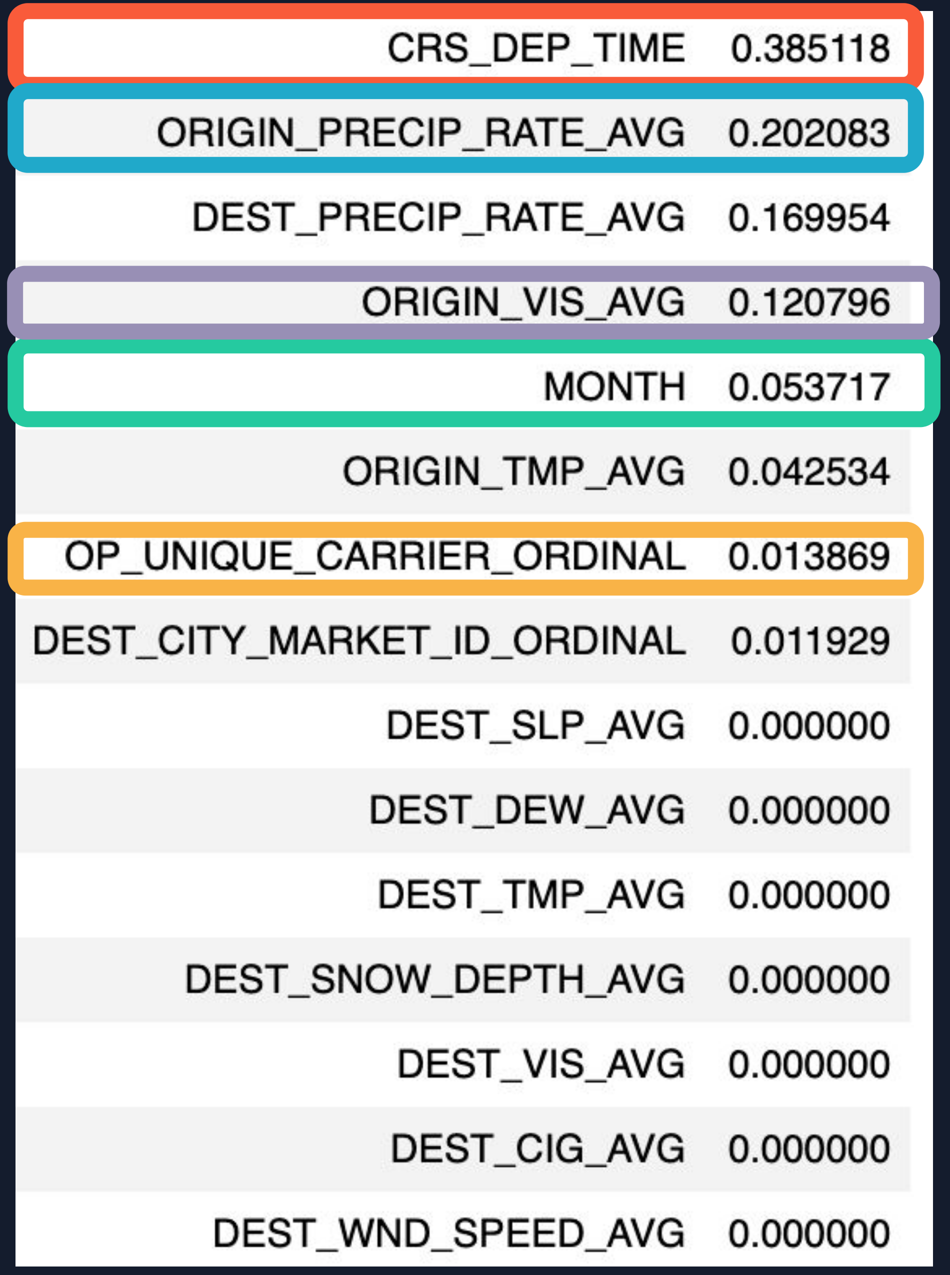
Feature Importance

Logistic Regression



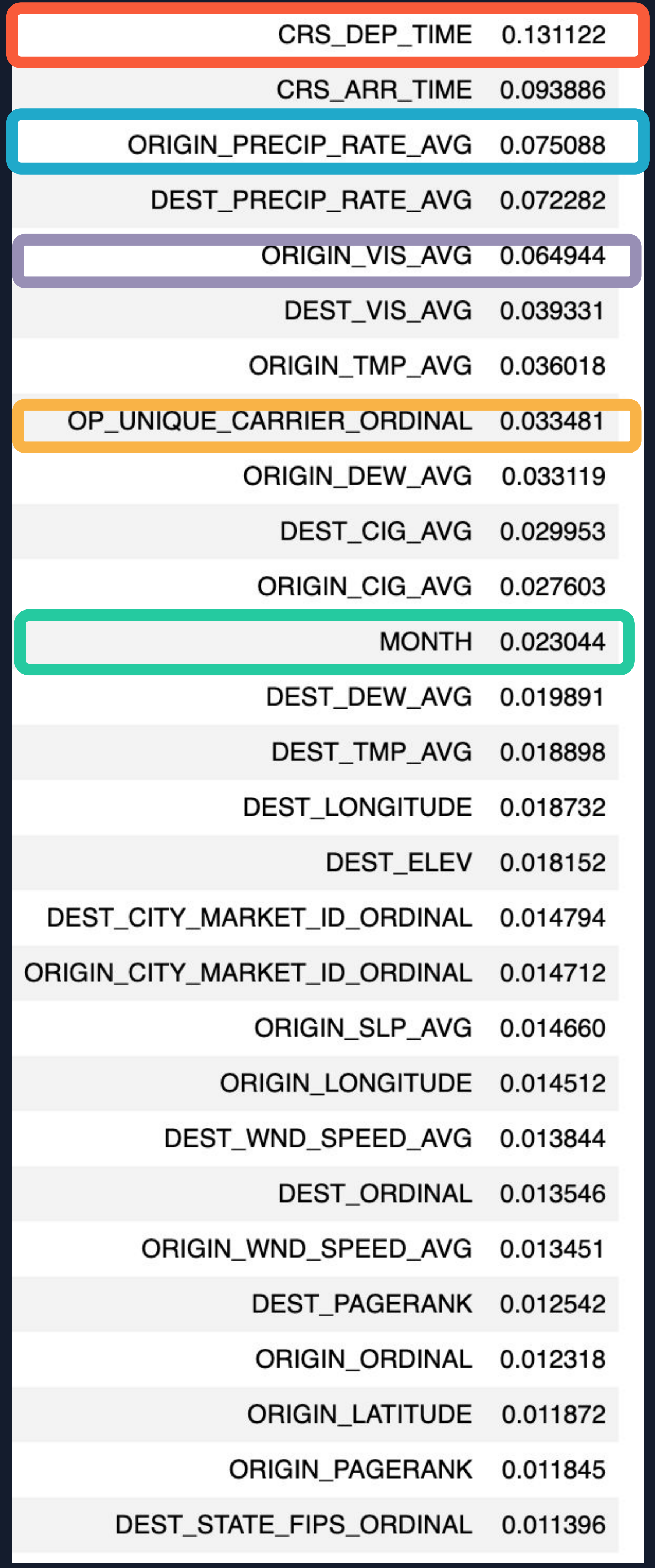
```
ORIGIN_AIRPORT_SEQ_ID -8.695
DEST_AIRPORT_SEQ_ID -5.775
MONTH 0.977
OP_UNIQUE_CARRIER -0.528
CRS_DEP_TIME 0.378
DAY_OF_WEEK 0.294
ORIGIN_TMP_AVG -0.226
DEST_TMP_AVG -0.211
ORIGIN_VIS_AVG -0.208
DEST_VIS_AVG -0.163
DAY_OF_MONTH -0.15
DEST_DEW_AVG 0.147
YEAR -0.132
DEST_WND_SPEED_AVG 0.108
ORIGIN_WND_SPEED_AVG 0.096
DEST_CIG_AVG -0.084
ORIGIN_PRECIP_RATE_AVG 0.08
DEST_PRECIP_RATE_AVG 0.08
ORIGIN_DEW_AVG 0.063
ORIGIN_CIG_AVG -0.058
ORIGIN_PAGERANK 0.051
DEST_PAGERANK 0.048
ORIGIN_SLP_AVG -0.042
ORIGIN_LATITUDE -0.023
DEST_LATITUDE -0.02
DEST_ELEV -0.019
ORIGIN_LONGITUDE 0.013
ORIGIN_WND_ANGLE_AVG 0.011
ORIGIN_SNOW_DEPTH_AVG 0.01
ORIGIN_ELEV -0.01
DEST_LONGITUDE -0.008
DEST_SLP_AVG 0.006
DEST_SNOW_DEPTH_AVG 0.006
```

Decision Tree



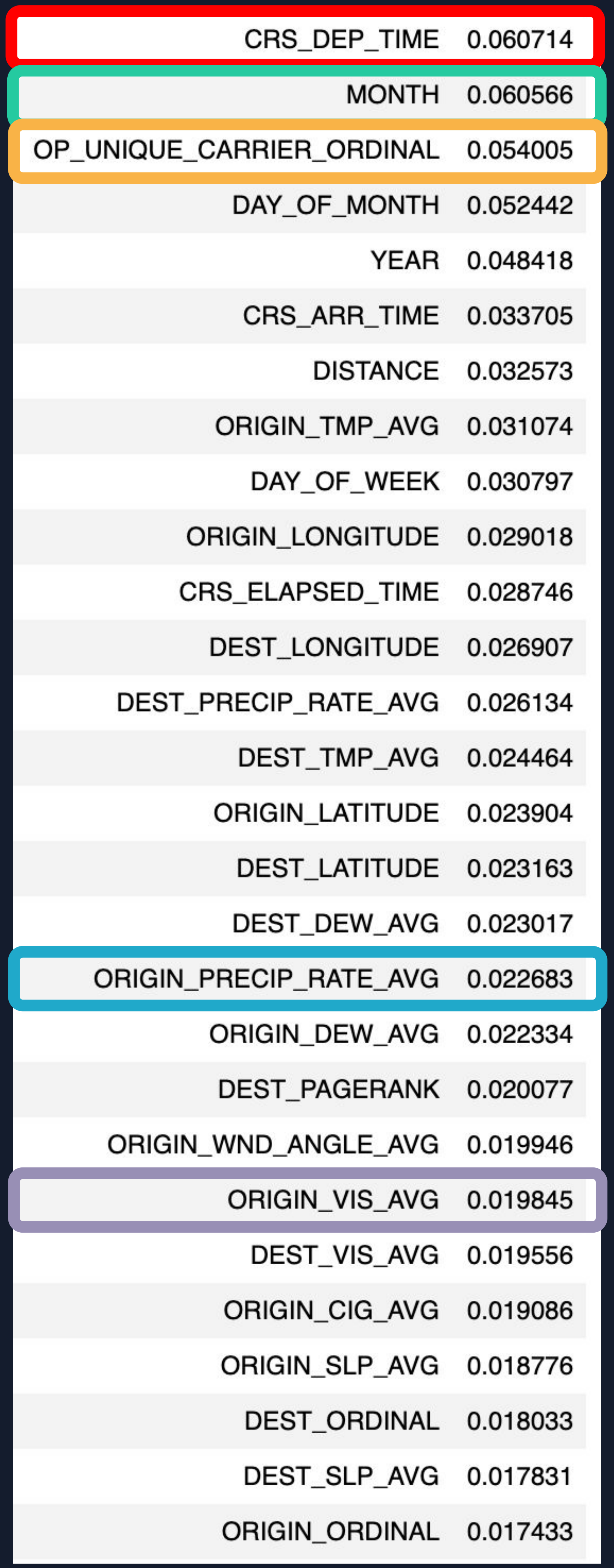
CRS_DEP_TIME	0.385118
ORIGIN_PRECIP_RATE_AVG	0.202083
DEST_PRECIP_RATE_AVG	0.169954
ORIGIN_VIS_AVG	0.120796
MONTH	0.053717
ORIGIN_TMP_AVG	0.042534
OP_UNIQUE_CARRIER_ORDINAL	0.013869
DEST_CITY_MARKET_ID_ORDINAL	0.011929
DEST_SLP_AVG	0.000000
DEST_DEW_AVG	0.000000
DEST_TMP_AVG	0.000000
DEST_SNOW_DEPTH_AVG	0.000000
DEST_VIS_AVG	0.000000
DEST_CIG_AVG	0.000000
DEST_WND_SPEED_AVG	0.000000

Random Forest



CRS_DEP_TIME	0.131122
CRS_ARR_TIME	0.093886
ORIGIN_PRECIP_RATE_AVG	0.075088
DEST_PRECIP_RATE_AVG	0.072282
ORIGIN_VIS_AVG	0.064944
DEST_VIS_AVG	0.039331
ORIGIN_TMP_AVG	0.036018
OP_UNIQUE_CARRIER_ORDINAL	0.033481
ORIGIN_DEW_AVG	0.033119
DEST_CIG_AVG	0.029953
ORIGIN_CIG_AVG	0.027603
MONTH	0.023044
DEST_DEW_AVG	0.019891
DEST_TMP_AVG	0.018898
DEST_LONGITUDE	0.018732
DEST_ELEV	0.018152
DEST_CITY_MARKET_ID_ORDINAL	0.014794
ORIGIN_CITY_MARKET_ID_ORDINAL	0.014712
ORIGIN_SLP_AVG	0.014660
ORIGIN_LONGITUDE	0.014512
DEST_WND_SPEED_AVG	0.013844
DEST_ORDINAL	0.013546
ORIGIN_WND_SPEED_AVG	0.013451
DEST_PAGERANK	0.012542
ORIGIN_ORDINAL	0.012318
ORIGIN_LATITUDE	0.011872
ORIGIN_PAGERANK	0.011845
DEST_STATE_FIPS_ORDINAL	0.011396

Gradient Boosted Tree

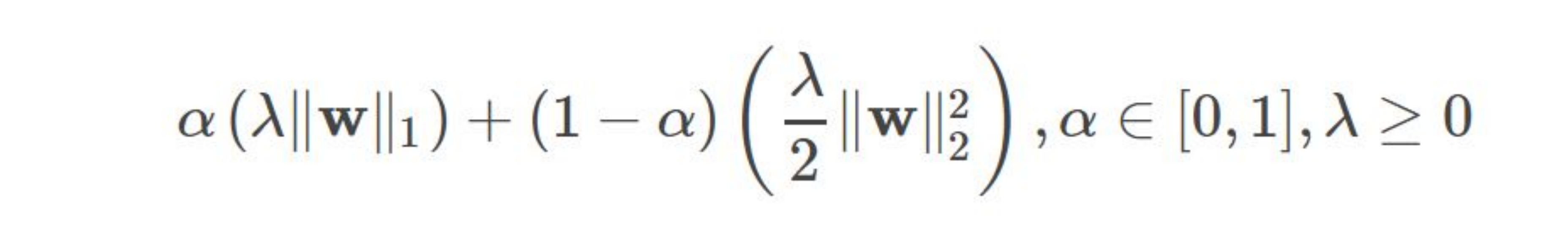


CRS_DEP_TIME	0.060714
MONTH	0.060566
OP_UNIQUE_CARRIER_ORDINAL	0.054005
DAY_OF_MONTH	0.052442
YEAR	0.048418
CRS_ARR_TIME	0.033705
DISTANCE	0.032573
ORIGIN_TMP_AVG	0.031074
DAY_OF_WEEK	0.030797
ORIGIN_LONGITUDE	0.029018
CRS_ELAPSED_TIME	0.028746
DEST_LONGITUDE	0.026907
DEST_PRECIP_RATE_AVG	0.026134
DEST_TMP_AVG	0.024464
ORIGIN_LATITUDE	0.023904
DEST_LATITUDE	0.023163
DEST_DEW_AVG	0.023017
ORIGIN_PRECIP_RATE_AVG	0.022683
ORIGIN_DEW_AVG	0.022334
DEST_PAGERANK	0.020077
ORIGIN_WND_ANGLE_AVG	0.019946
ORIGIN_VIS_AVG	0.019845
DEST_VIS_AVG	0.019556
ORIGIN_CIG_AVG	0.019086
ORIGIN_SLP_AVG	0.018776
DEST_ORDINAL	0.018033
DEST_SLP_AVG	0.017831
ORIGIN_ORDINAL	0.017433

Results

ML Lib Logistic Regression Inputs

- Regularization via Elastic net
  - Convex combination of the L1 and the L2 regularization terms
  - LR Hyper parameters:
    - elasticNetParam (α)
    - regParam (λ)



$$\alpha\left(\lambda\|\mathbf{w}\|_1\right) + (1-\alpha)\left(\frac{\lambda}{2}\|\mathbf{w}\|_2^2\right), \alpha \in [0,1], \lambda \geq 0$$

- maxIter

- Class Weighing (weightCol)

Identified by gridsearch and crossfold validation

Logistic Regression Models

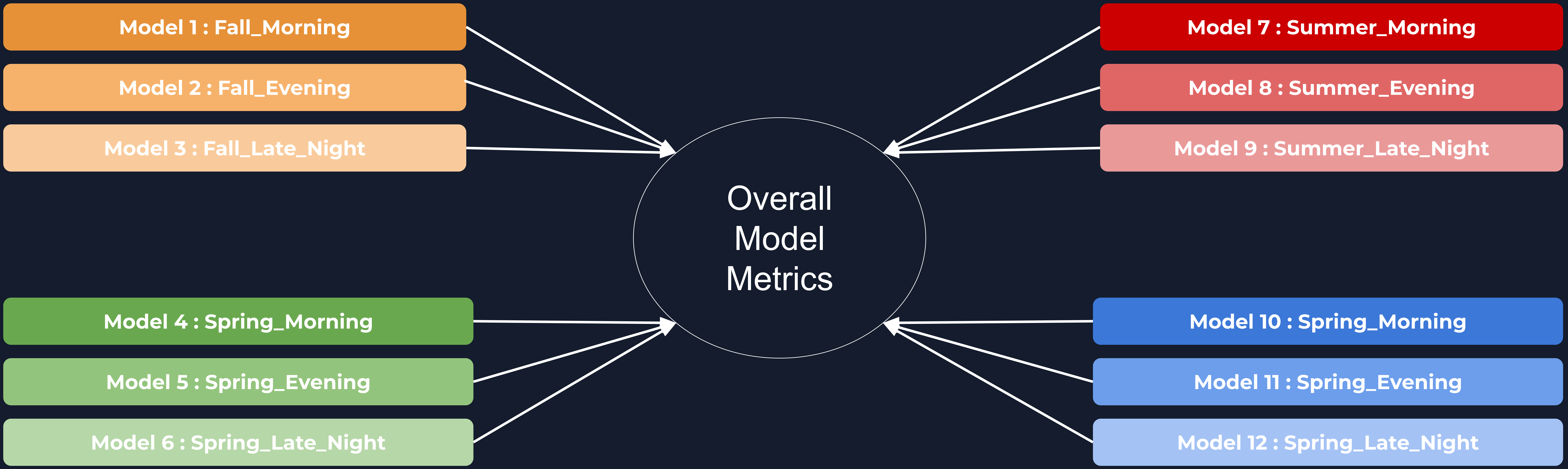
1 Single Model (no Class Weights)

2 Single Model (with Class Weights)

Logistic Regression Models

3

Multi Model Strategy :
Temporal Model (with Class Weights)



Model 1 : Fall_Morning

Model 2 : Fall_Evening

Model 3 : Fall_Late_Night

Model 4 : Spring_Morning

Model 5 : Spring_Evening

Model 6 : Spring_Late_Night

Overall Model Metrics

Model 7 : Summer_Morning

Model 8 : Summer_Evening

Model 9 : Summer_Late_Night

Model 10 : Spring_Morning

Model 11 : Spring_Evening

Model 12 : Spring_Late_Night

Logistic Regression Models - Results

Model	Number of models	Model Type	ClassWeighted	maxIter	reg parm	elastic net	Accuracy	precision	recall	f1
LR_Unweighted_DefaultParms	1	No temporal models	No	100	0	0	0.8009	0.5599	0.0580	0.1051
LR_Weighted_DefaultParms	1	No temporal models	Yes	100	0	0	0.6544	0.3181	0.6247	0.4216
LR_Weighted_OptimalParms	1	No temporal models	Yes	20	0.001	0.25	0.6434	0.3117	0.6363	0.4184
LR_Weighted_OptimalParms_MultiModel	12	Season + Time	Yes	20	0.001	0.25	0.6351	0.3111	0.6670	0.4243
Test_Data	12	Season + Time	Yes	20	0.001	0.25	0.6351	0.3105	0.6665	0.4236

Challenges & Limitations

1 **Outcome not sufficiently customer focused**

2 **Complexity of weather dataset resulted in limited utilisation**

3 **Ensembled trees did not meet expectations**

Proposed Improvements

1.

Utilize Spark
Parallelize
Multi-Model

2.

Real-time
dynamic
graph-based
modelling
approach

Thank you.

Any questions?