BIG DATA CHALLENGE

SHAASTRA 2019

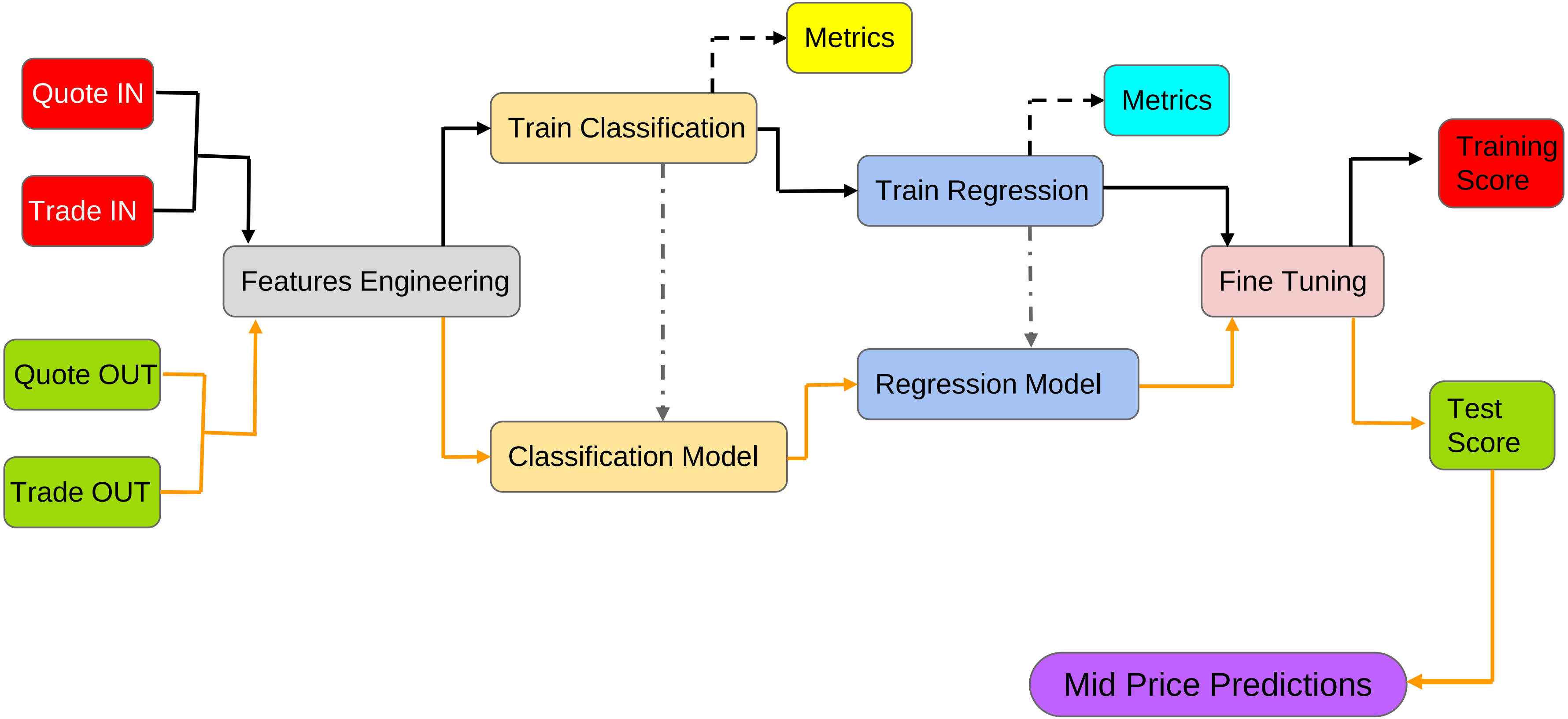
TEAM - DATA DRIVERS

E. Naveen

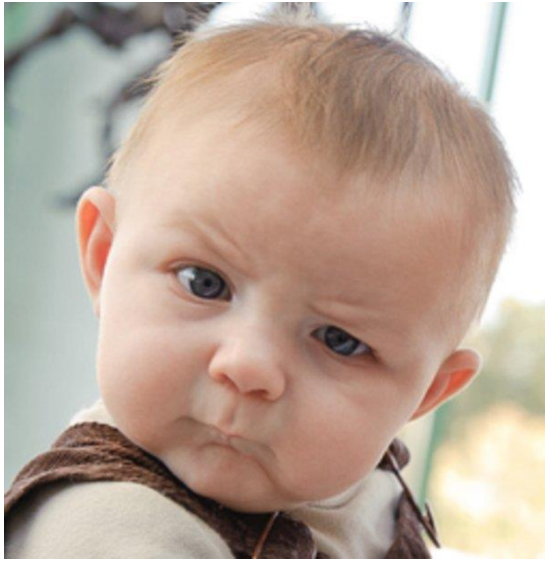
I. Hariharan

MID PRICE PREDICTION

FINAL PIPELINE



R2 SCORE 0.98 ??



FEATURES

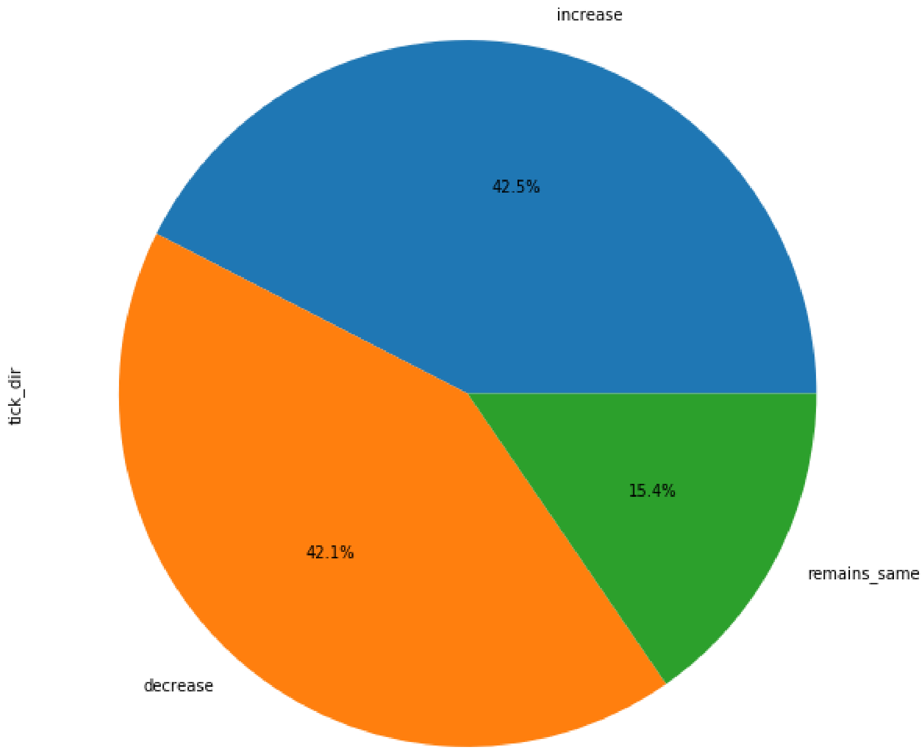
DATETIME	→	date and time of record
SYM	→	system used to record data
BSIZE	→	bid size of quote
BID	→	bid price of quote
ASIZE	→	ask size of quote
ASK	→	ask price of quote
MID	→	average of ASK and BID
MID CHANGE GROUP	→	every change in mid price is specified by a group number

FEATURES

FUT MID	→	future mid price to predict
TICK	→	FUT MIID - MID
TICK SIZE	→	absolute value of TICK
TICK DIR	→	direction of MID PRICE movement
IMBALANCE	→	(BSIZE - ASIZE) / (BSIZE + ASIZE)
SHARE IMBALANCE	→	BSIZE/(BSIZE + ASIZE)

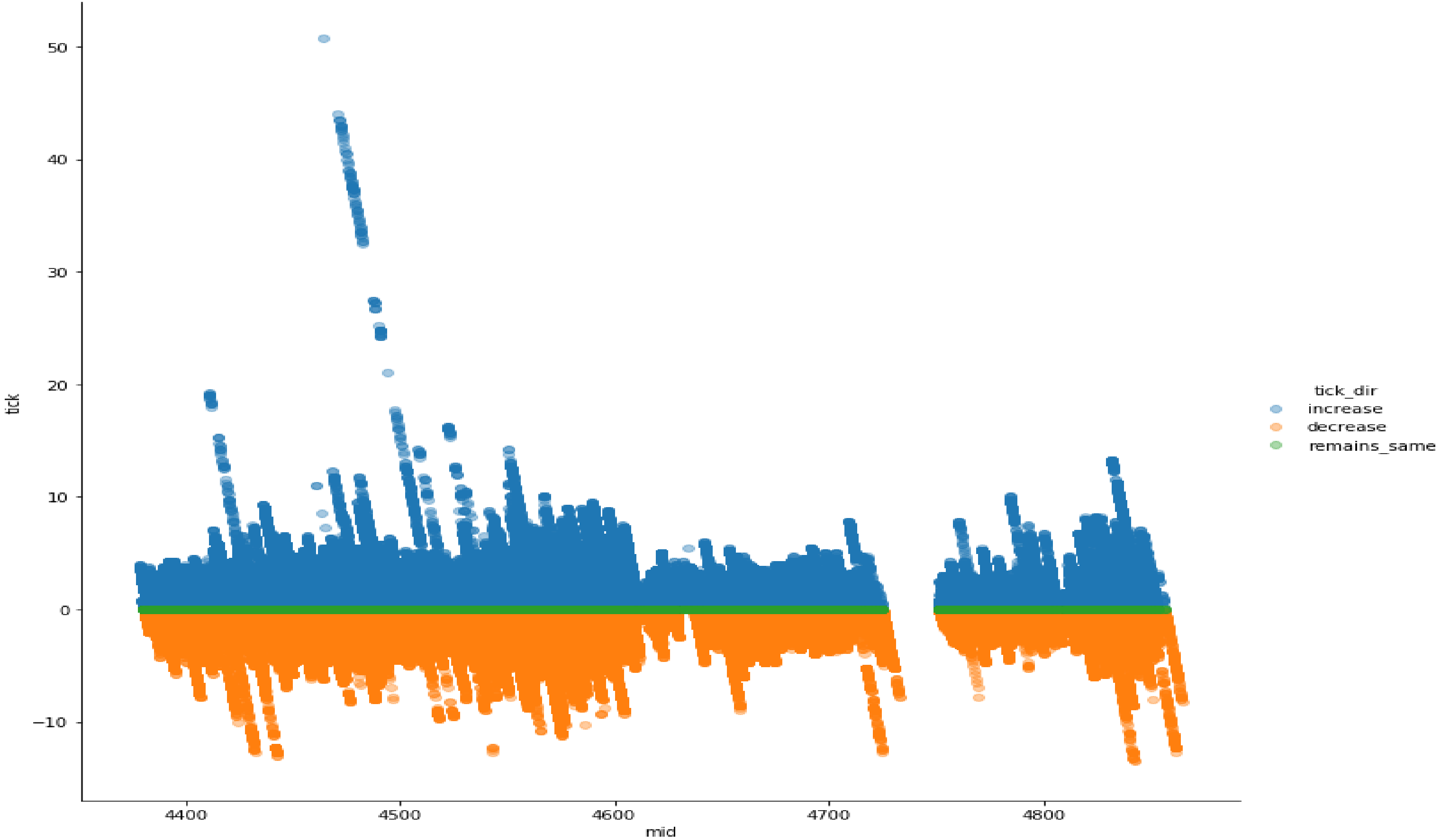
EXPLORATORY DATA ANALYSIS

PROPORTION OF LABELS



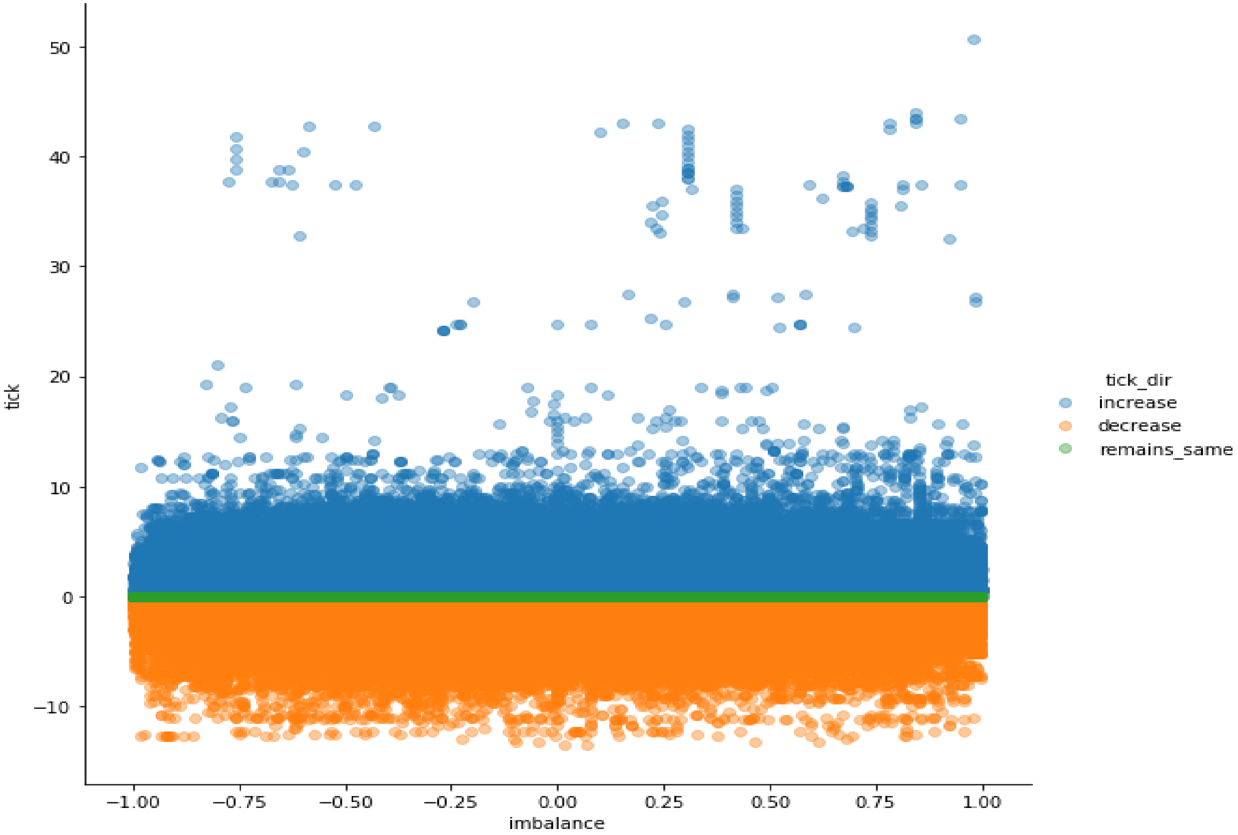
EXPLORATORY DATA ANALYSIS

TICK VS MID PRICE

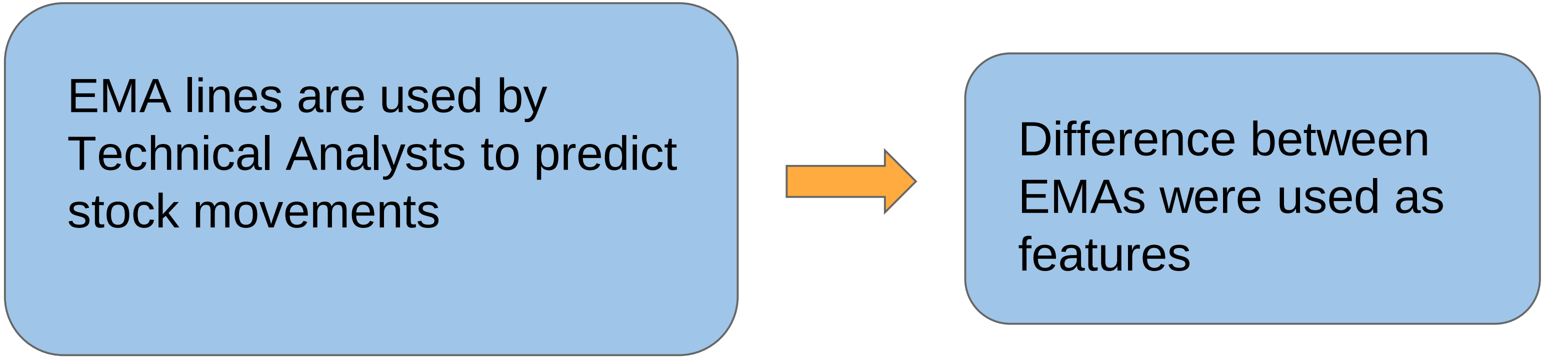


EXPLORATORY DATA ANALYSIS

TICK VS IMBALANCE



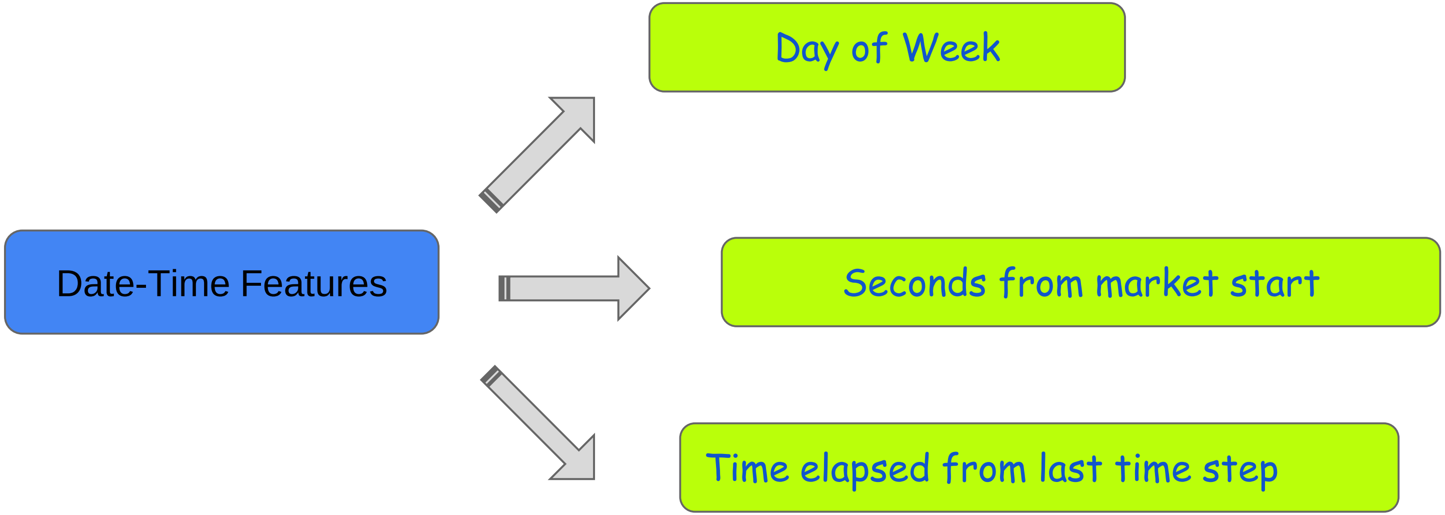
FEATURE ENGINEERING - EMA



EMA lines are used by Technical Analysts to predict stock movements

Difference between EMAs were used as features

FEATURE ENGINEERING - TIME BASED



Date-Time Features

Day of Week

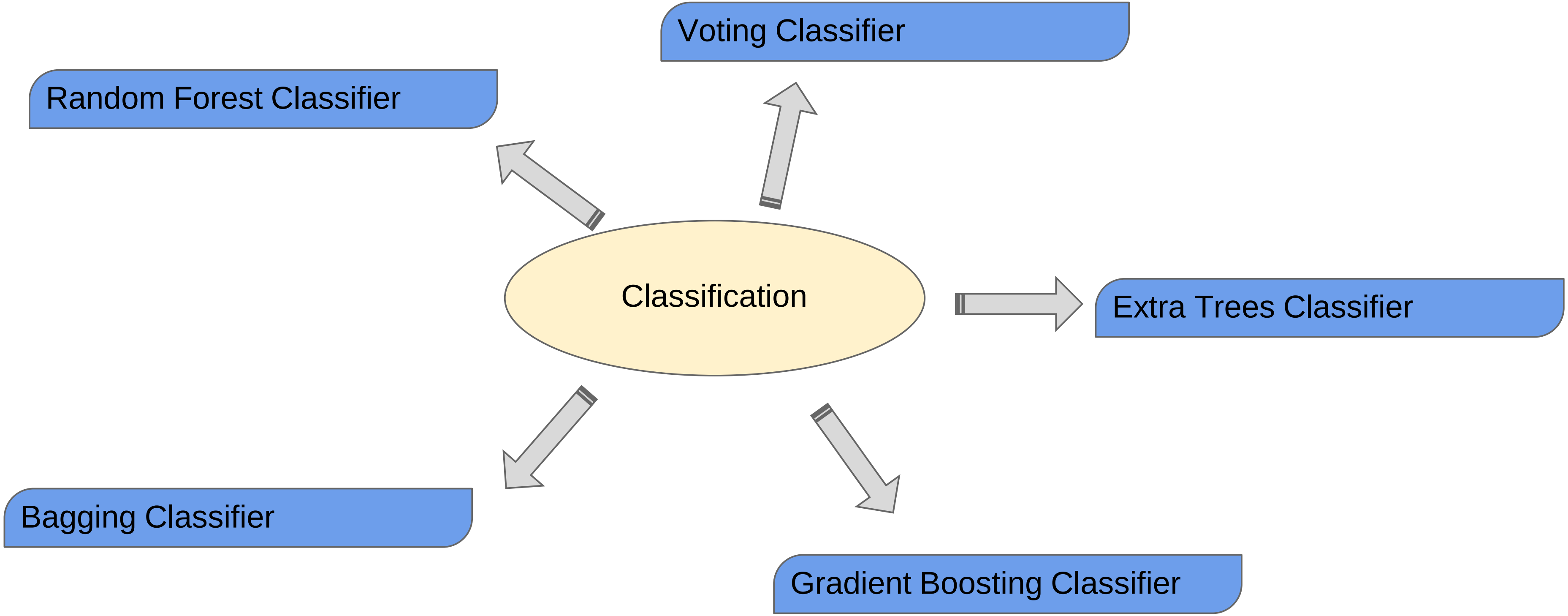
Seconds from market start

Time elapsed from last time step

THE TEAM

MOVING AVERAGE
MID PRICE
TRADE SIZE
ASK SIZE
TRADE PRICE
IMBALANCE
BID PRICE
BID SIZE
DAY OF WEEK
SECONDS
ASK PRICE

CLASSICAL MODELS USED



Classification

Random Forest Classifier

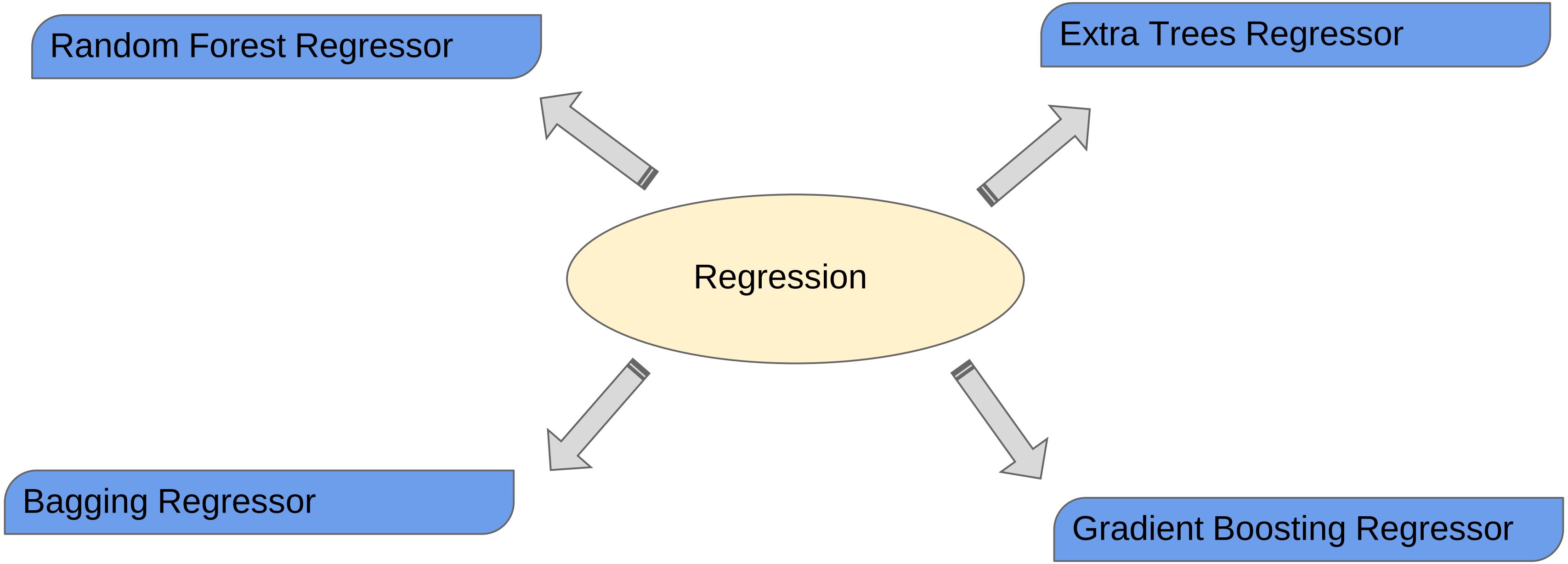
Voting Classifier

Extra Trees Classifier

Bagging Classifier

Gradient Boosting Classifier

CLASSICAL MODELS USED



Regression

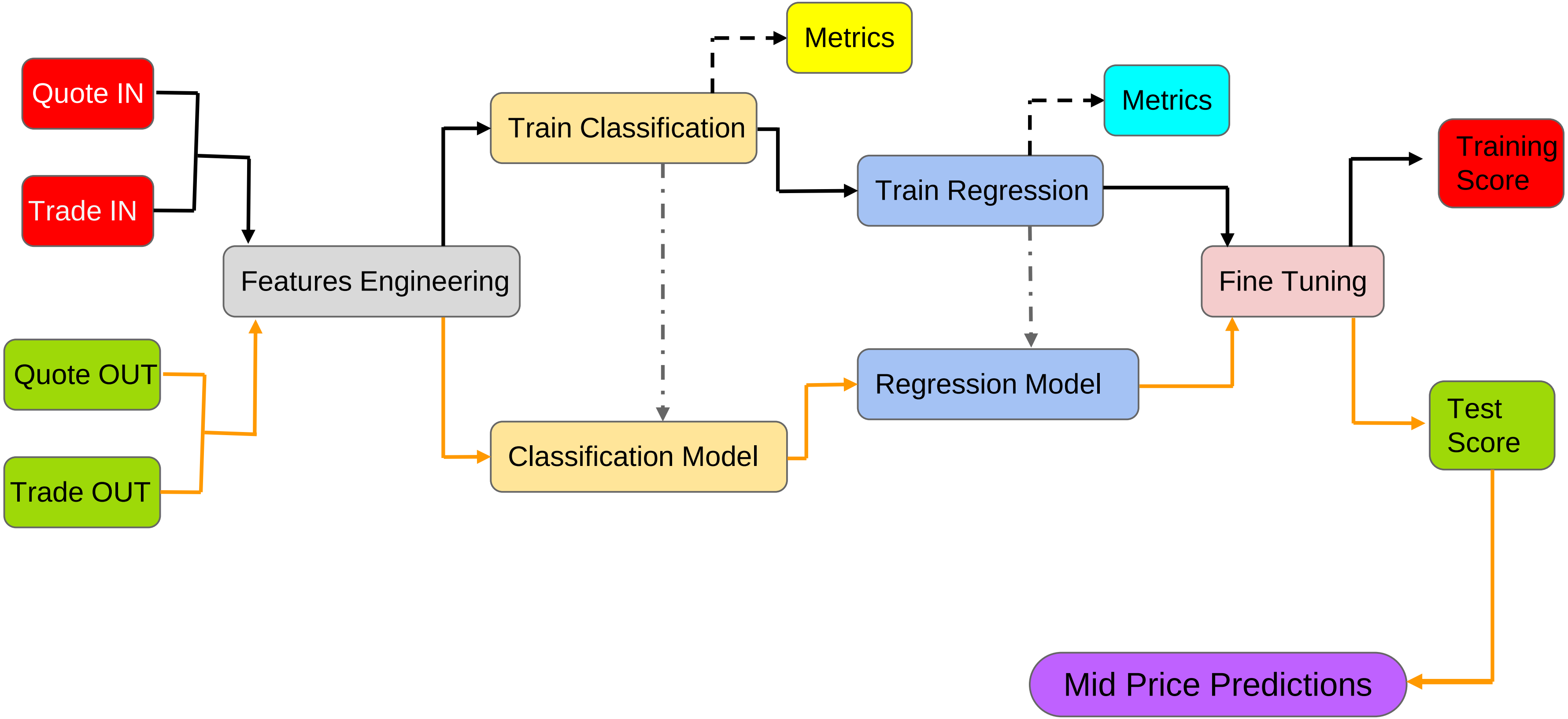
Random Forest Regressor

Extra Trees Regressor

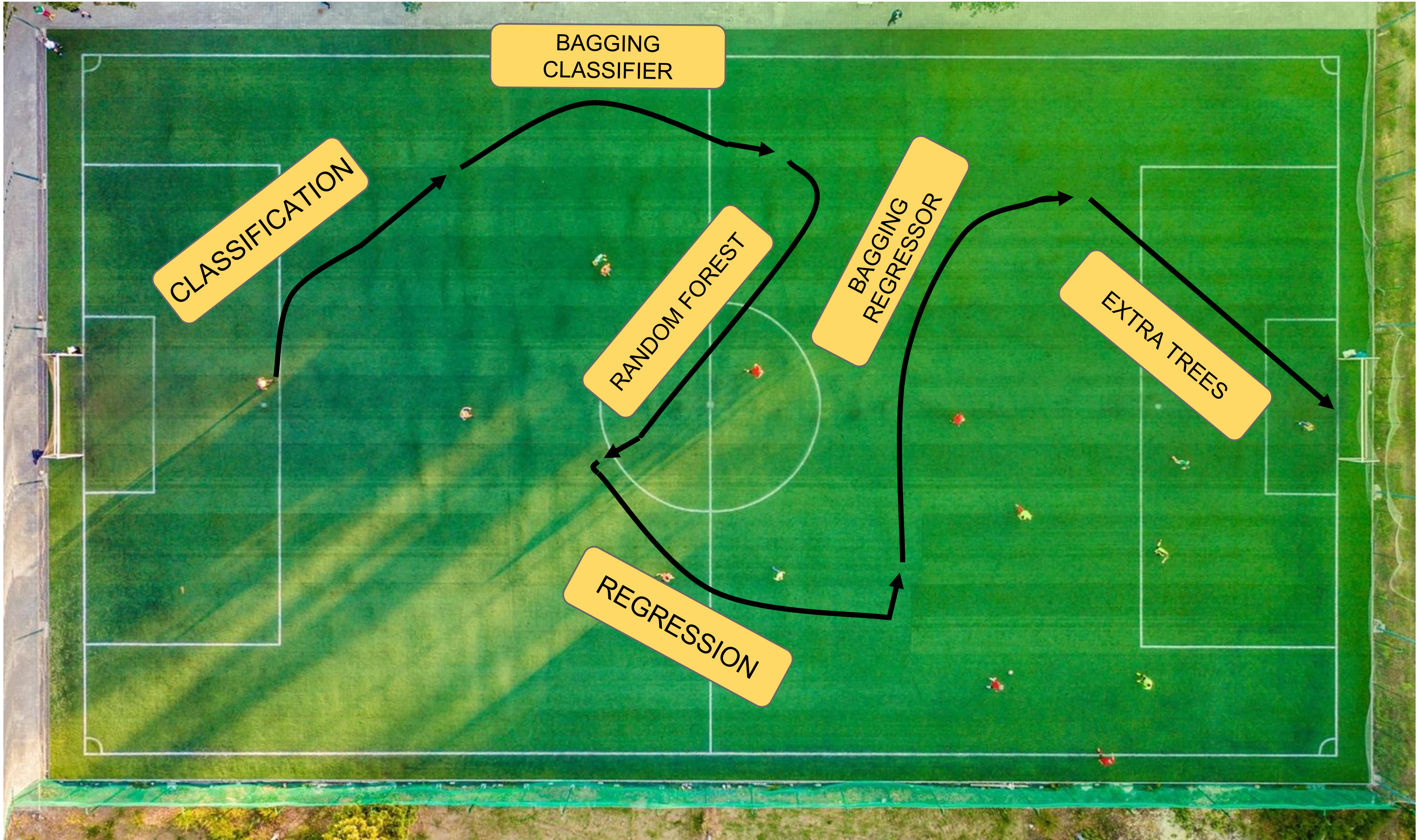
Bagging Regressor

Gradient Boosting Regressor

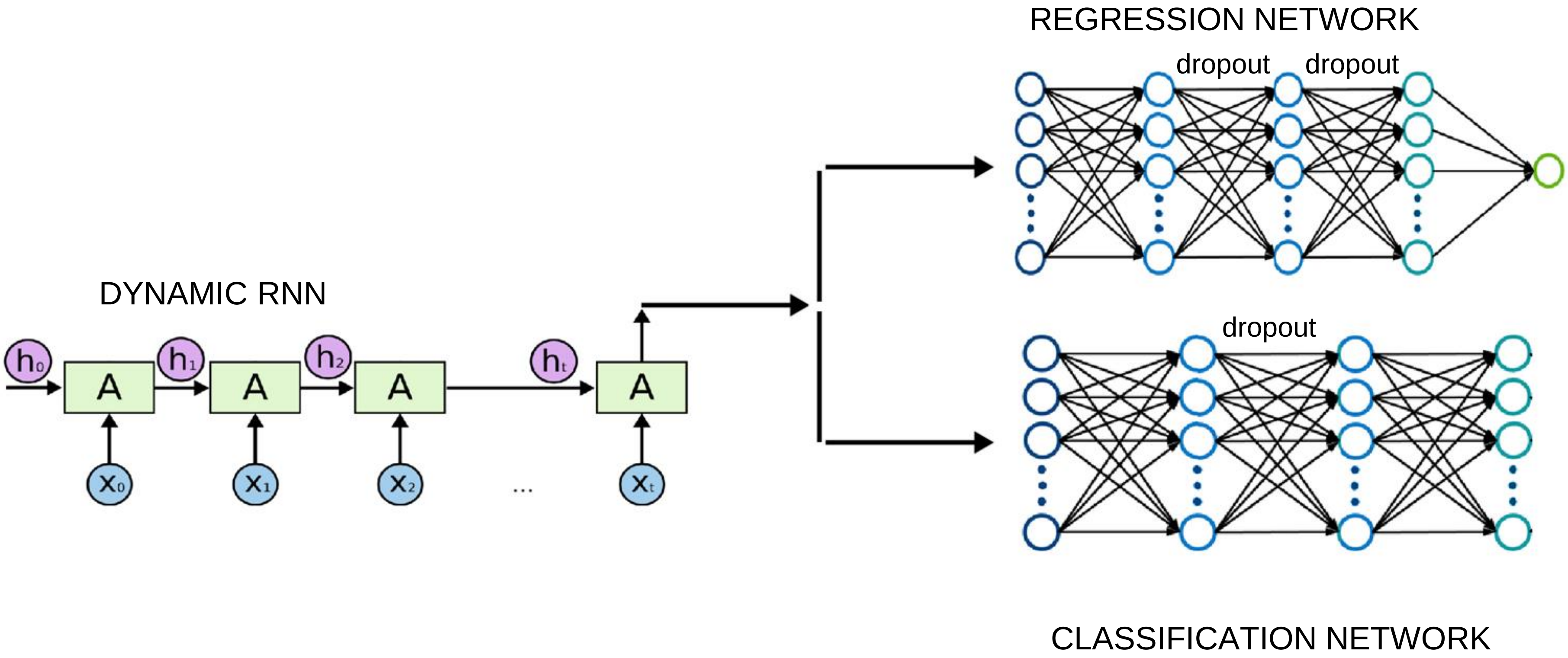
FINAL PIPELINE



STRATEGY


CLASSIFICATION
BAGGING CLASSIFIER
RANDOM FOREST
REGRESSION
BAGGING REGRESSOR
EXTRA TREES

RECURRENT NEURAL NETWORK



REGRESSION NETWORK

dropout dropout

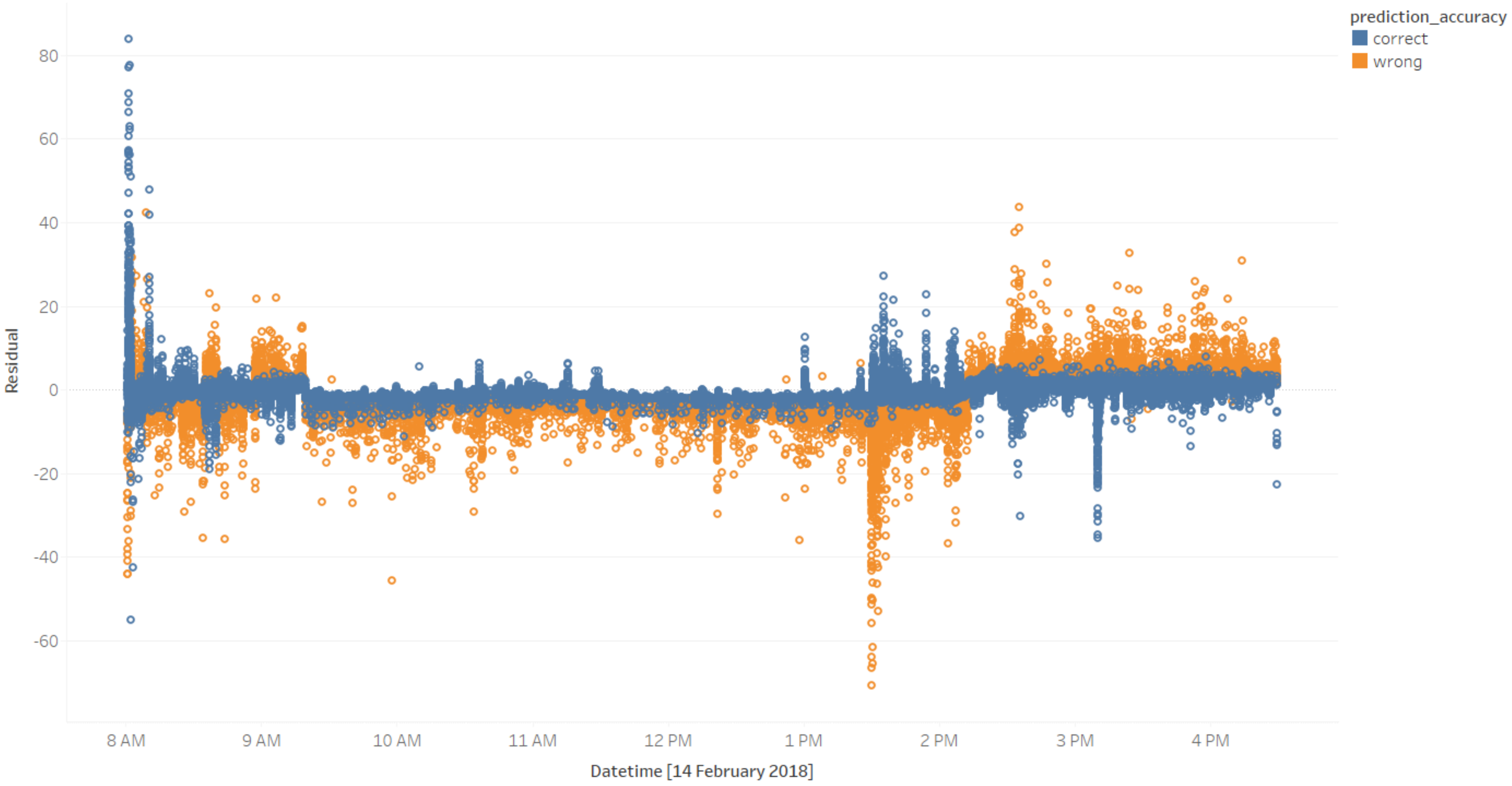
DYNAMIC RNN

dropout

CLASSIFICATION NETWORK

RESULT AND MANUAL TWEAKING

ResidualPlot



The plot of sum of Residual for Datetime. Color shows details about prediction_accuracy.

SCORE

TRAIN

RMS : 1.202

TEST

RMS : 1.106