NUS National University of Singapore

SWS3017
C2-36

NUS SOC 2019 - TO WIN A DATA COMPETITION

NYC TAXI TRIP DURATION

Overview

- **Background** - nowadays, the intelligent transportation is gradually changing people's way of life.
- **Research Object** - in this project our main aim is to predict the total ride duration of taxi trips in New York City with some primary data provided such as datetime and geo-coordinates.
- **Financial value** - offer better service and gain larger market share.

Data

DATA SOURCE -

kaggle

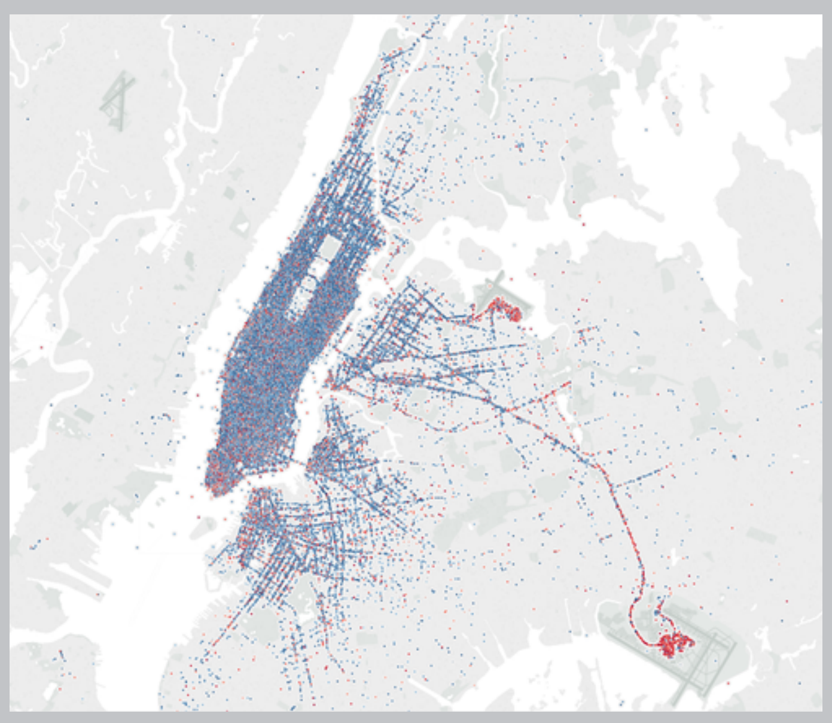
TARGET VARIABLE -

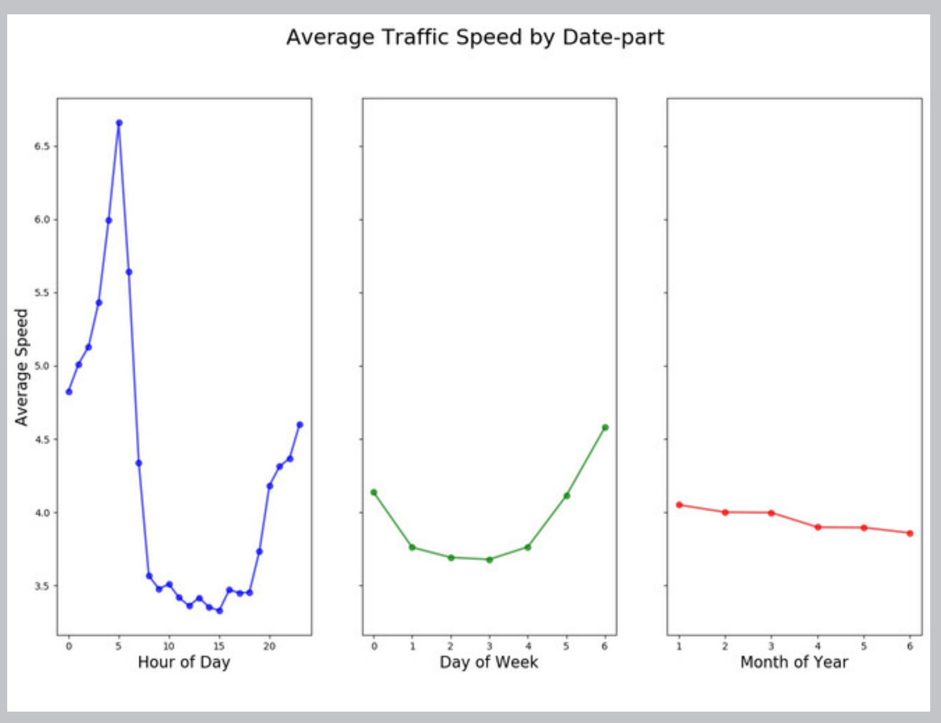
trip duration

FEATURES - *1458644 Samples*

X3 trip id, supplier, whether is stored	X4 longitude and latitude of pickup and dropoff position
X1 passenger numbers	X2 pickup time and dropoff time

	Min	Mean	Std	Max
vendor_id	1	1.53495	0.498777	2
passenger_count	0	1.66453	1.314242	9
pickup_longitude	-121.933	-73.9735	0.070902	-61.3355
pickup_latitude	34.3597	40.75092	0.032881	51.88108
dropoff_longitude	-121.933	-73.9734	0.070643	-61.3355
dropoff_latitude	32.18114	40.7518	0.035891	43.92103
trip_duration	1	959.4923	5237.432	3526282





point: pick-up points in New York City
color: the gradient color from red to blue indicates the duration of trips from long to short

Baseline & Advanced Model

Method	RMSLE	Square of R
Mean	0.6961	0.00201
DT	0.5556	0.6327
RF	0.3612	0.8092
NN	0.7949	0.0144
KNN	0.4717	0.8092
XGB	0.3966	0.7866
LGBM	0.3469	0.7992

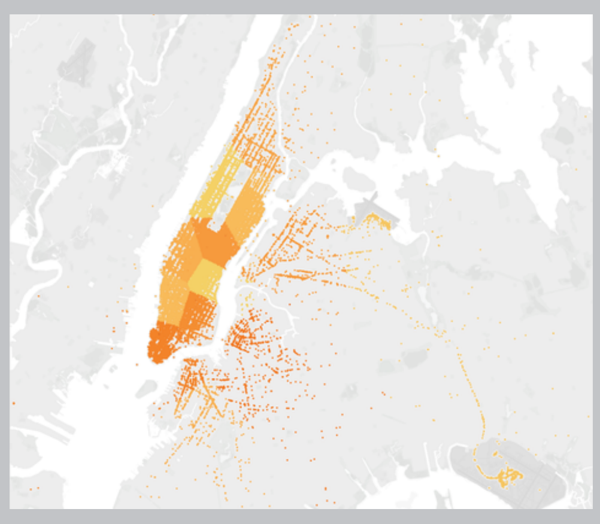
Feature Engineering

EXTERNAL DATA -

OSRM
total distance
total travel time
number of steps

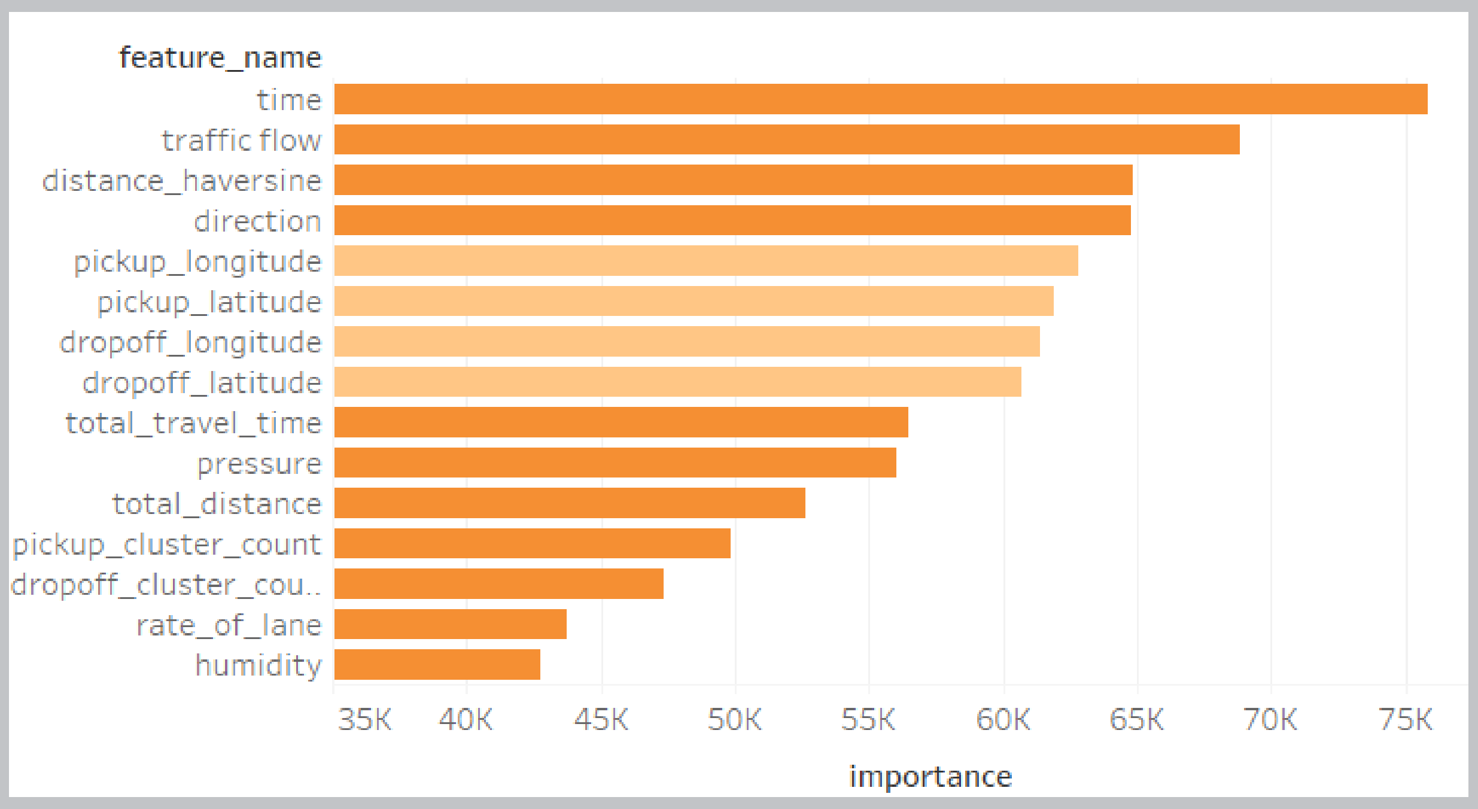
Weather
atmosphere pressure
humidity
dewpoint

INTERESTING FEATURES -



Cluster: use k-means to cluster the pick-up and drop-off position
Airport: if the pick-up or drop-off is within 2km from airports
Direction: the direction of the trip
Holiday: if the trip is on the holiday
Speed of the road: the average speed from one cluster to another
Trip count in hour: the number of records during the near hour

TOP 15 IMPORTANT FEATURES -



Final Result

RERUN MODELS -

	RMSLE		Square of R	
Method	Before	After	Before	After
DT	0.5556	0.4885	0.6327	0.6454
RF	0.3612	0.3394	0.8092	0.8185
XGB	0.3966	0.3420	0.7866	0.8069
LGBM	0.3469	0.3201	0.7992	0.8464

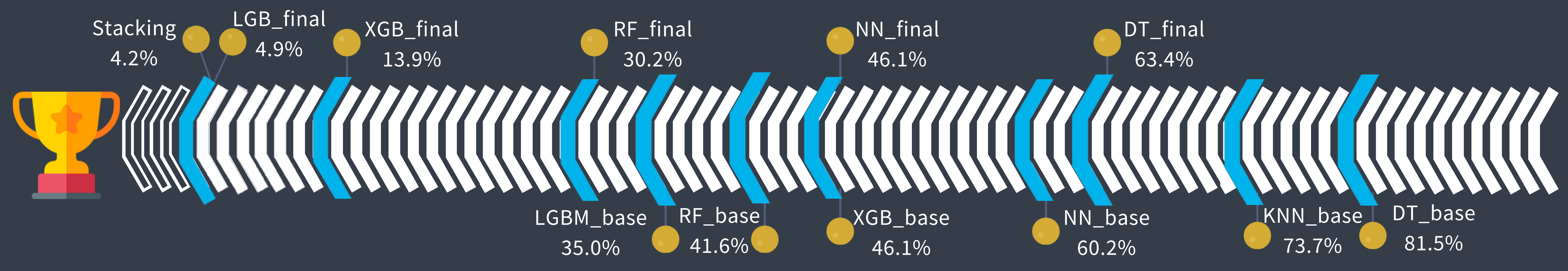
BEST PERFORMACE MODEL ---- STACKING

- level-1 training data set is the predictions of LightGBM and XGBoost, meta-regressor is linear regression
- it is likely to get a better result when combining best single model with others
- when several models both get a good result, using stacking can add marginal improvement
- the chosen models should have different edges

Conclusion

- due to the size of data, over fitting is not that easy
- a feature too simple is also not that good
  - eg.holiday--only a few samples make a contribution
- trust the algorithm, but not so much
  - pressure, temperature, hum .etc can be learnt

POSITION IN LEADERBOARD



HUANG LINLING SHEN XINDI XU XINYUAN XU ZHITONG ZHANG HANYUAN