



MACHINE LEARNING FINAL

TAXI TRIP TIME PREDICTION (II)

組員：M11115021 蘇峻緯 M11115010 莊歲任 M11152025 陳彥合

OUTLINE

- Theme
- Feature Description
- Dataset Preprocessing
- Feature Engineering
- Method & Model
- Comparison
- Conclusion

THEME

■ Taxi Trip Time Prediction (II)

 Research Prediction Competition

ECML/PKDD 15: Taxi Trip Time Prediction (II)

Predict the total travel time of taxi trips based on their initial partial trajectories

\$250

Prize Money

345 teams · 7 years ago

[Overview](#)

[Data](#)

[Code](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Late Submission](#)

...

FEATURE DESCRIPTION

- TRIP_ID
- CALL_TYPE
- ORIGIN_CALL
- ORIGIN_STAND
- TAXI_ID
- TIME_STAMP
- DAY_TYPE
- MISSING_DATA
- POLYLINE

	TRIP_ID	CALL_TYPE	ORIGIN_CALL	ORIGIN_STAND	TAXI_ID	TIMESTAMP	DAY_TYPE	MISSING_DATA	POLYLINE
0	1372636858620000589	C	NaN	NaN	20000589	1372636858	A	False	[[[-8.618643,41.141412],[[-8.618499,41.141376],[...
1	1372637303620000596	B	NaN	7.0	20000596	1372637303	A	False	[[[-8.639847,41.159826],[[-8.640351,41.159871],[...
2	1372636951620000320	C	NaN	NaN	20000320	1372636951	A	False	[[[-8.612964,41.140359],[[-8.613378,41.14035],[...
3	1372636854620000520	C	NaN	NaN	20000520	1372636854	A	False	[[[-8.574678,41.151951],[[-8.574705,41.151942],[...
4	1372637091620000337	C	NaN	NaN	20000337	1372637091	A	False	[[[-8.645994,41.18049],[[-8.645949,41.180517],[...

FEATURE DESCRIPTION

- POLYLINE of testing data.
 - Not Complete
 - last point is not destination
- But the feature set we used contains:
 - Distance (start point, end point)
 - EndCluster (endpoint)

REVISE METHOD

- 1. Predict the end point of testing data


X: (Call_type, ORIGIN_CALL, ORIGIN_STAND, MISSING_DATA, lon_1st, lat_1st, delta_lon, delta_lat)
y: (end_lon, end_lat)

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [New Notebook](#) [...](#)

Notebooks

[Filters](#)

[All](#) [Your Work](#) [Shared With You](#) [Bookmarks](#) [Best Score](#) [▼](#)



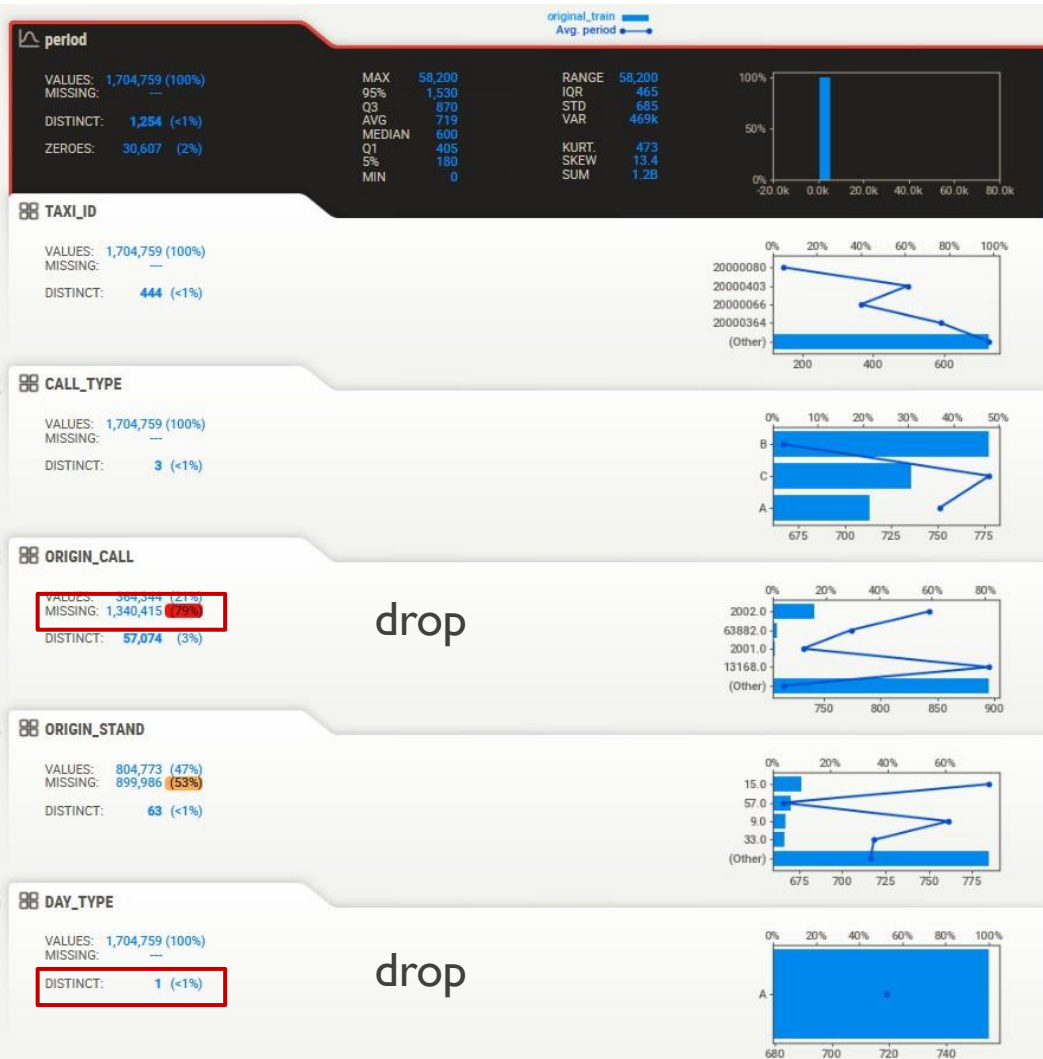
Taxi trajectory Prediction I
Updated 2y ago
Score: 2.70197 · 2 comments · ECML/PKDD 15: Taxi Trajectory Prediction (I)

[▲](#) [11](#) [...](#)

- 2. Predict the EndCluster of testing data

X: (CALL_TYPE_STAND, Hour, WEEKDAY, StartCluster)
y: (EndCluster)

BEFORE PREPROCESSING: EDA



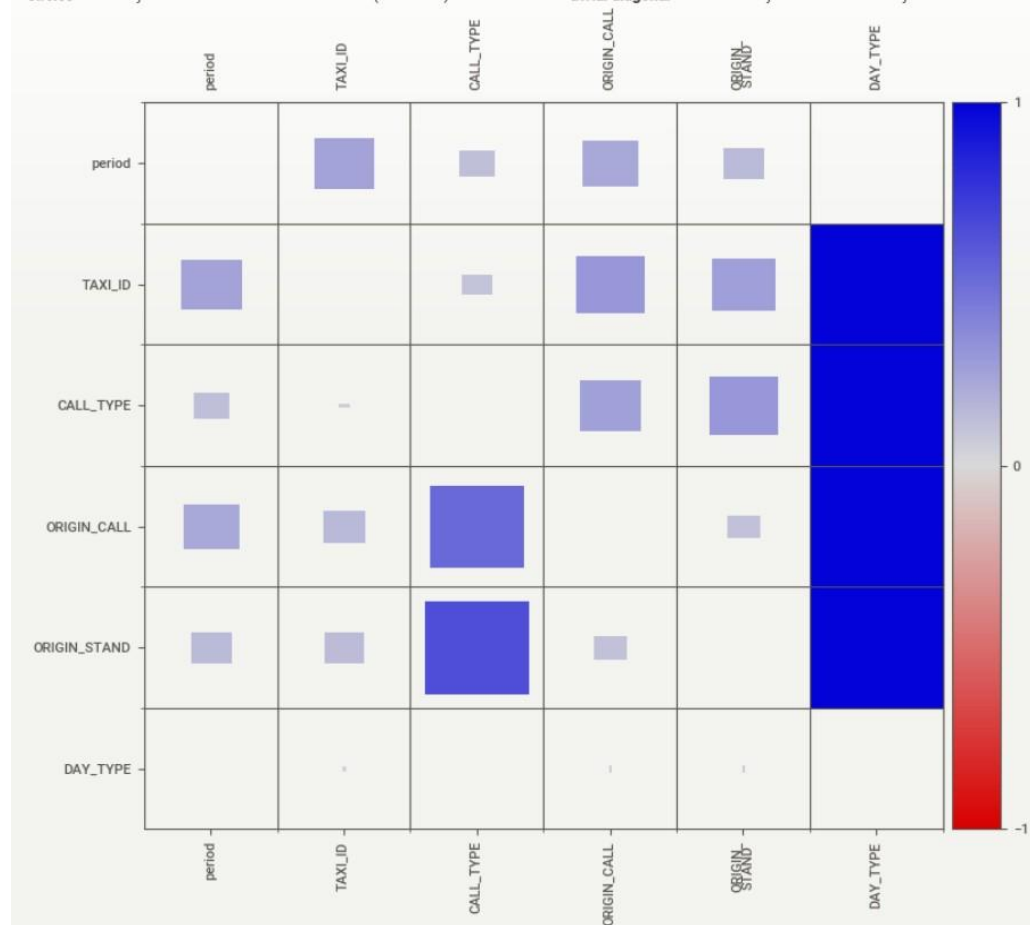
Info
about
start

Associations

[Only including dataset "original_train"]

■ Squares are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **assymetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• Circles are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.



DATA PREPROCESSING

- **CALL_TYPE_STAND** \leftarrow CALL_TYPE + CALL_STAND

e.g. 'B7' \leftarrow 'B' + 7

- **YEAR, MONTH, DAY, HOUR, MIN** \leftarrow TIMESTAMP

1372636858 \rightarrow 2013, 7, 1, 0, 0, 0 (monday)

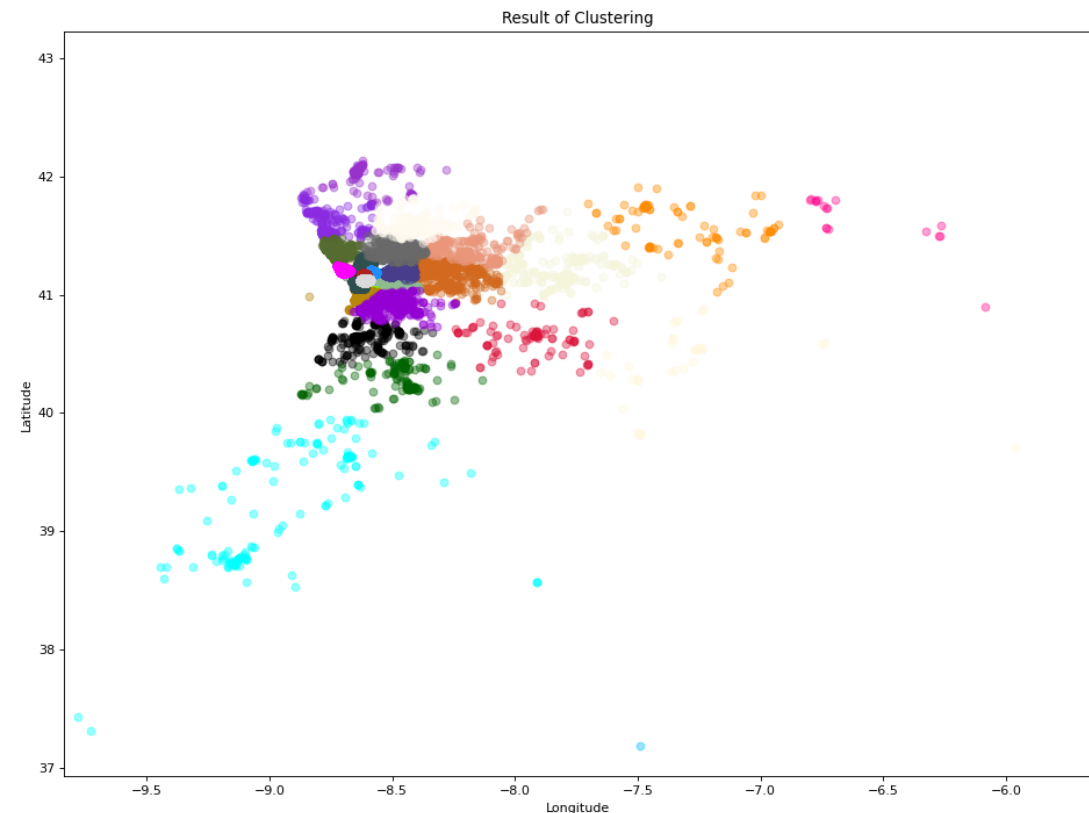
- **START_LON, START_LAT, END_LON, END_LAT** \leftarrow POLYLINE

- **DISTANCE, period(y)** \leftarrow START_LON, START_LAT, END_LON, END_LAT

DATA PREPROCESSING

- **StartCluster, EndCluster** \leftarrow START_LON, START_LAT, END_LON, END_LAT
- Using KMEANS algorithm to cluster all the start and end points

	lon	lat	Cluster
0	-8.630838	41.154489	21
1	-8.665740	41.170671	1
2	-8.607996	41.142915	42
3	-8.687268	41.178087	9
4	-8.578224	41.160717	22
...
2917441	-8.606385	41.144742	42
2917442	-8.612469	41.146020	28
2917443	-8.610138	41.140845	28
2917444	-8.630712	41.154885	21
2917445	-8.615538	41.140629	3
2917446 rows × 3 columns			



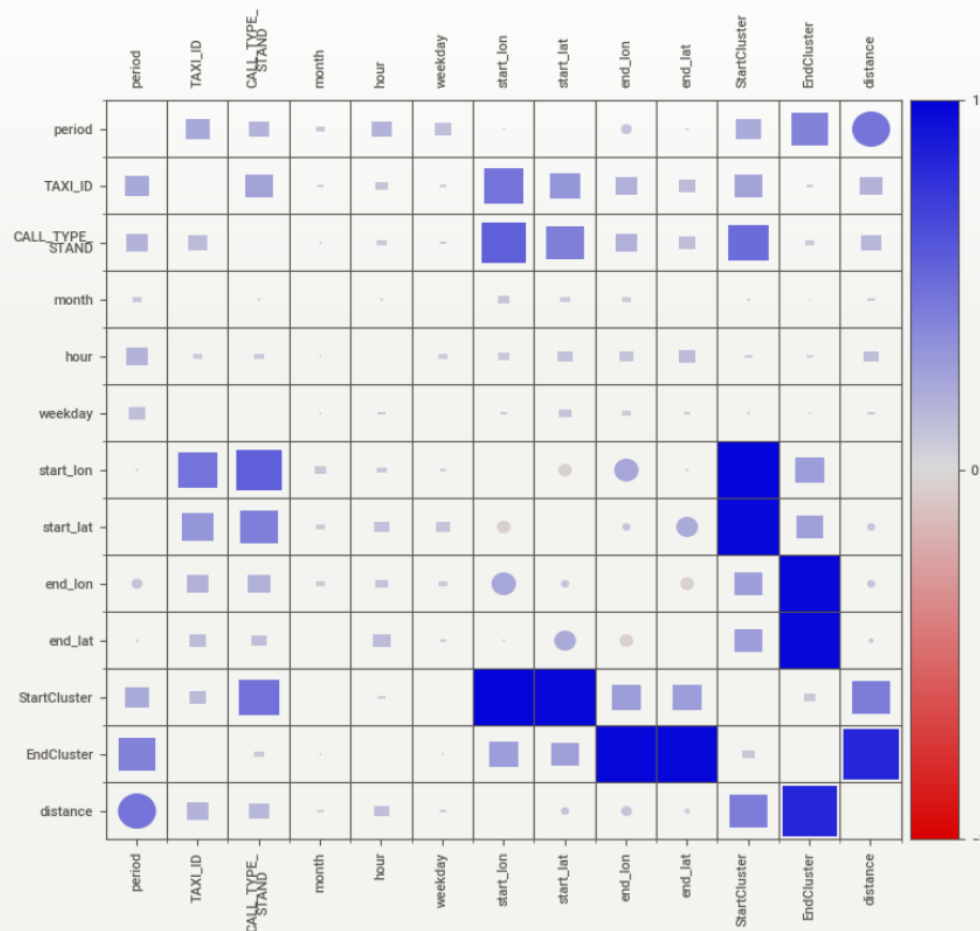
DATA PREPROCESSING

Associations

[Only including dataset "train_EDA"]

■ Squares are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **asymmetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• Circles are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.



FEATURE ENGINEERING

Feature	Encoding Method
CALL_TYPE_STAND	One-hot Encoding
MONTH	One-hot Encoding
HOURL	One-hot Encoding / Group+One-hot Encoding
WEEKDAY	One-hot Encoding / Group+One-hot Encoding
StartCluster / EndCluster	One-hot Encoding / Frequency Encoding
Distance	Numerical

METHOD & MODEL

Machine Learning

MACHINE LEARNING(1)

- Based Method + Revise Method(1) :
 - Data_dimension = 112
 - Score :

MODEL	Training time(s)	Kaggle private	Kaggle public
LR	0.5885	0.6829	0.7600
XGB	17.1914	0.7192	0.7564

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	frequency	frequency	numerical

MACHINE LEARNING(2)

- Based Method + Revise Method(1) :
 - Data_dimension = 207
 - Score :

MODEL	Training time(s)	Kaggle private	Kaggle public
LR	1.3191	0.6937	0.76968
XGB	24.730	0.6910	0.75931

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	numerical

MACHINE LEARNING(3)

- Based Method + Revise Method(2) :
 - Data_dimension = 111
 - Score :

MODEL	Training time(s)	Kaggle private	Kaggle public
LR	0.5513	0.7146	0.8185
XGB	12.6156	0.6898	0.7954

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER
one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	frequency	frequency

MACHINE LEARNING(4)

- Based Method + Revise Method(2) :
 - Data_dimension = 206
 - Score :

MODEL	Training time(s)	Kaggle private	Kaggle public
LR	1.1102	0.7043	0.8268
XGB	20.747	0.7022	0.8081

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER
one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding	one-hot encoding

MACHINE LEARNING

Add group method to time feature

MACHINE LEARNING(5)

- Method of Using grouping result :
 - Data_dimension = 88 / 183
 - Score :

MODEL	Preprocessing	Training time(s)	Kaggle private	Kaggle public
LR	Frequency	0.4211	0.6856	0.7619
LR	One-hot encoding	1.5100	0.6956	0.7720

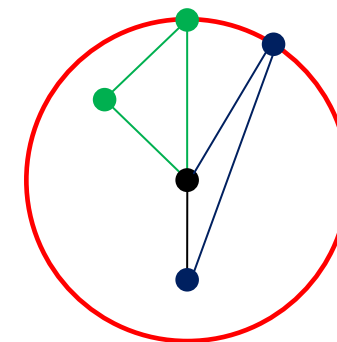
CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding (group)	one-hot encoding (group)	frequency/ one-hot encoding	frequency/ one-hot encoding	numerical

MACHINE LEARNING

Add angle filter

MACHINE LEARNING(6)

- Method of Using angle filter :
 - Data_dimension = 88/183
 - Score :



MODEL	Preprocessing	Training time(s)	Kaggle private	Kaggle public
LR	Frequency	0.4513	0.6940	0.7683
LR	One-hot encoding	15.449	0.7094	0.7823

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding (group)	one-hot encoding (group)	frequency/ one-hot encoding	frequency/ one-hot encoding	numerical

MACHINE LEARNING

Try different number of clustering

MACHINE LEARNING(7)

- Method of Using difference clustering number :
 - Data_dimension = 88
 - Score :

MODEL	Clustering size	Training time(s)	Kaggle private	Kaggle public
LR(poly = 2)	2500	70.473	0.6858	0.7417
LR(poly = 2)	50	69.694	0.6810	0.7587

CALL_TYPE_ STAND	MONTH	HOURL	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding group	one-hot encoding group	frequency	frequency	numerical

MACHINE LEARNING(7)

- Method of Using cluster = 2500 :
 - Data_dimension = 112/88
 - Score :

MODEL	Using group	Training time(s)	Kaggle private	Kaggle public
LR	No	0.6163	0.6819	0.7601
LR	Yes	0.3216	0.6844	0.7620

CALL_TYPE_ STAND	MONTH	HOUR	WEEKDAY	START_CLUSTER	END_CLUSTER	DISTANCE
one-hot encoding	one-hot encoding	one-hot encoding / group	one-hot encoding / group	frequency	frequency	numerical

COMPARISON

- Compare to other approach:

Leader Board Rank	Kaggle private	Kaggle public	Approach
1	0.52528	0.50390	Not provided.
96	0.59623	0.59949	$15 * (\text{len}(\text{polyline}) - 1) + 495$.
Ours	0.6810	0.7587	Cluster-Frequency Encoding + Polynomial LR
Sample Benchmark	0.83835	0.89277	All 660.
Formula	1.30529	1.48417	$15 * (\text{len}(\text{polyline}) - 1)$.

CONCLUSION

- Some general guidelines when novices begin to analyze the dataset
 - Exploratory Data Analysis
 - \hat{y} generation
 - Timestamp -> datetime
 - Start/End info
- Some factors in real world would make this prediction much difficult



THANK FOR
LISTENING