TAXI

Trajectory Prediction

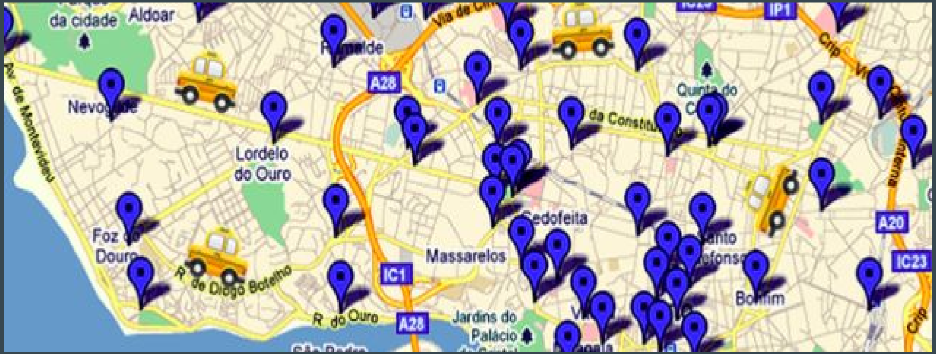
TEAM VWXYJ
Yunsheng Bai
Vivi Chuang
Junheng Hao
Sherly Sun

Table of Contents

- Introduction
- Strategy Analysis
- Implementation
  - Similarity Based
  - Neural Network
- Results

Introduction

- Predict taxi trip destination (Porto, Portugal)
- Based on initial partial trajectories
- Motivation
  - Mobile Dispatch System
  - Easy to see where the taxi has been
  - Hard to know where the taxi is heading to
- Features
  - TRIP_ID
  - ORIGIN_CALL (phone number)
  - ORIGIN_STAND
  - TAXI_ID
  - START_TIMESTAMP
  - DAYTYPE (weekday, holiday, weekend)
  - COORDINATES sequence (every 15s)



Similarity-based approach

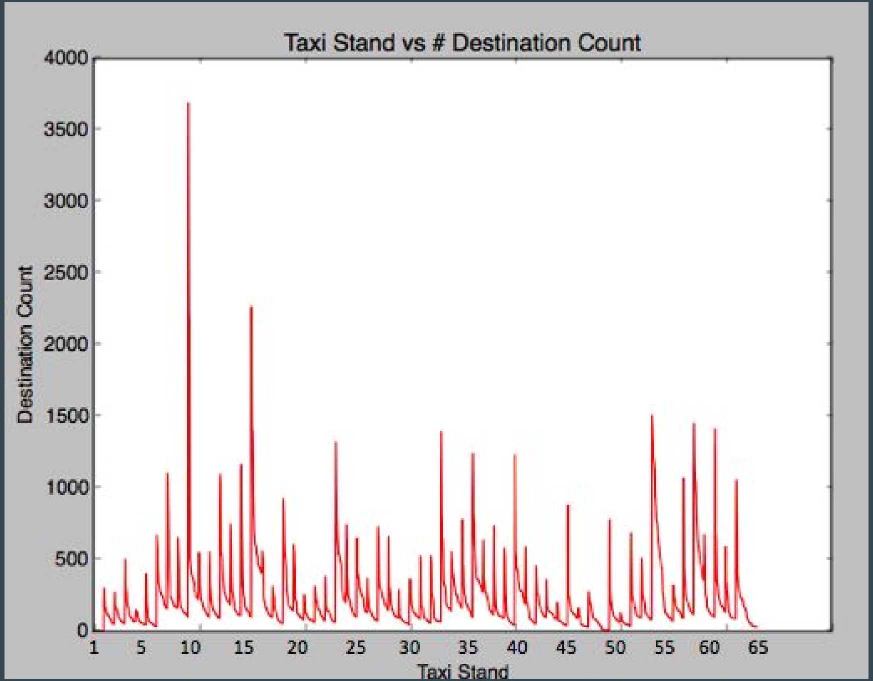
(1) Some trips tend to stop for a long time in middle.

- Count the number of waiting instance
- Statistics: 65.49%: 30s, 15.87%: 45s, 1.13%: 60s

(2) Last 1-2 minutes is more important than the previous part of the trip when predicting.

Similarity-based approach

Insights from data visualization



(3) Nearest taxi stand from a trip's start point also provides hint for its destination.

- Destination counting for each nearest stand

(4) Set out time influences the destination to certain degree.

Similarity-based approach

Data Preprocessing:

(1) Divide the area covered by the training set into grids to bin different locations. (Grid resolution: 100m)

(2) In one training example, if the taxi does not move more than 1 meter within 45 seconds, we ignore the previous trajectory.

(3) Calculate the nearest taxi stand from starting point of the journey

(4) For testing set, cut the journey and only keep the last 90 seconds features.

Similarity-based approach

General Model:

For every test case, use geometric median of top K related trips as final prediction.

Related trips mean:

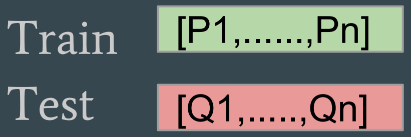
(1) Trips start from same taxi stand

(2) Trips set out from within -t ~ +t hour (time within day) (hyperparameter)

Similarity-based approach

General Model:

Ranking function to calculate distance between journey:



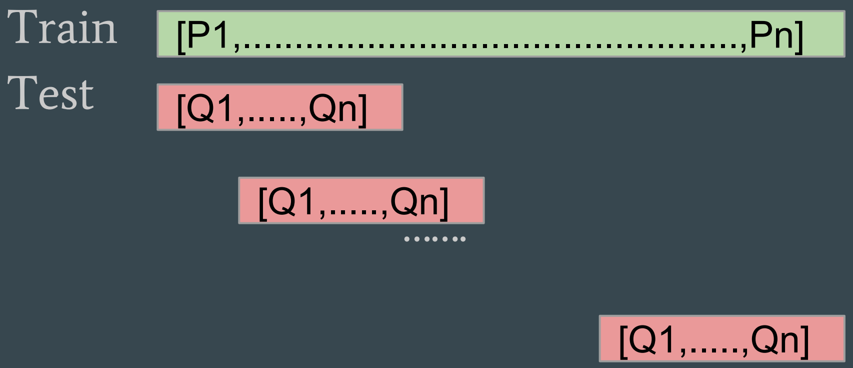
Train

[P1,......,Pn]

Test

[Q1,.....,Qn]

$$dis(train, test) = dis(P_1, Q_1) + \lambda dis(P_2, Q_2) + ... + \lambda^{n-1} dis(P_n, Q_n)$$



Train

[P1,..,Pn]

Test

[Q1,.....,Qn]

[Q1,.....,Qn]

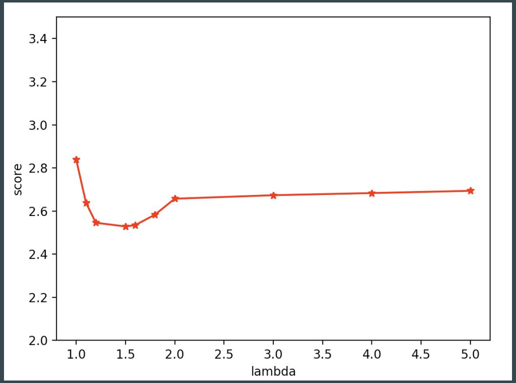
....

[Q1,.....,Qn]

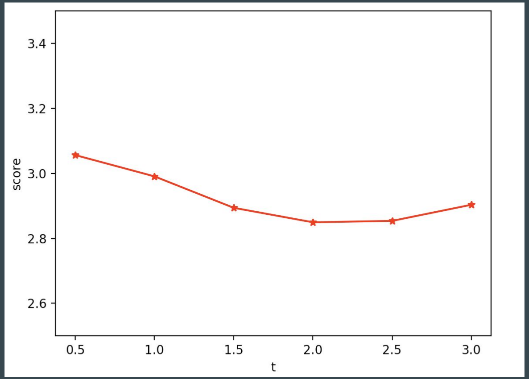
Slide the test data window, and use the smallest one as final distance between train and test.

Similarity-based approach

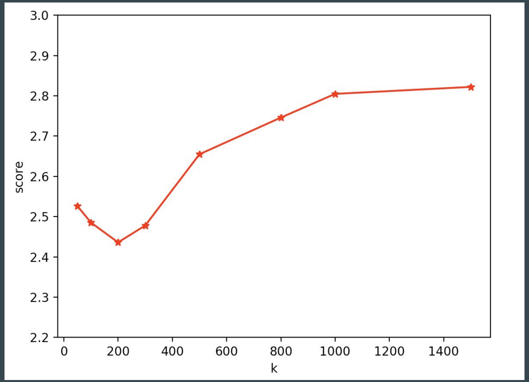
Parameter tuning result on validation set:



Lambda variation with fix t and k



t variation with fix lambda and k



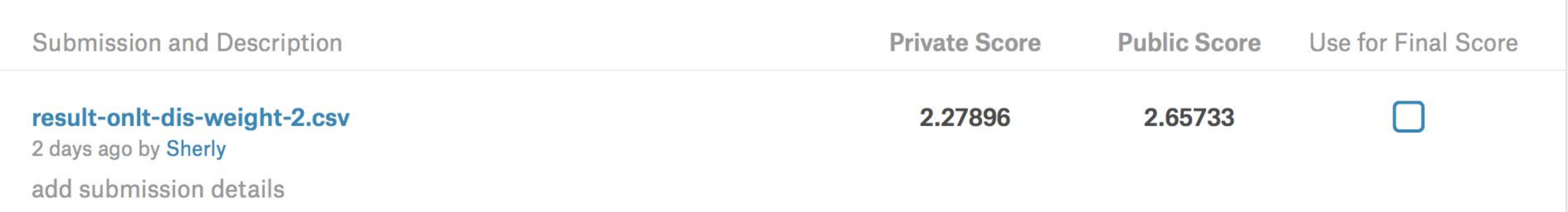
k variation with fix lambda and t

Similarity-based approach

Final parameter setting

k = 200, t = 2, lambda = 1.4

Final performance



Submission and Description	Private Score	Public Score	Use for Final Score
result-onlt-dis-weight-2.csv 2 days ago by Sherly add submission details	**2.27896**	**2.65733**	□

Neural Network Approach

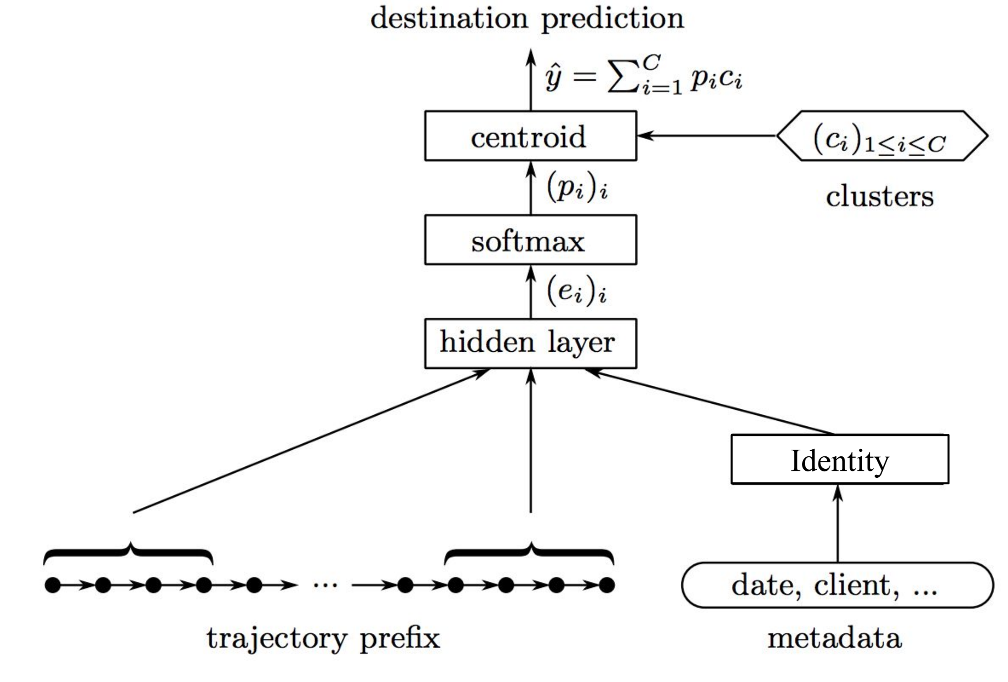


Fig. 1: Architecture of the winning model.

De Brébisson, Alexandre, et al. "Artificial neural networks applied to taxi destination prediction." arXiv preprint arXiv:1508.00021 (2015).

Neural Network Approach

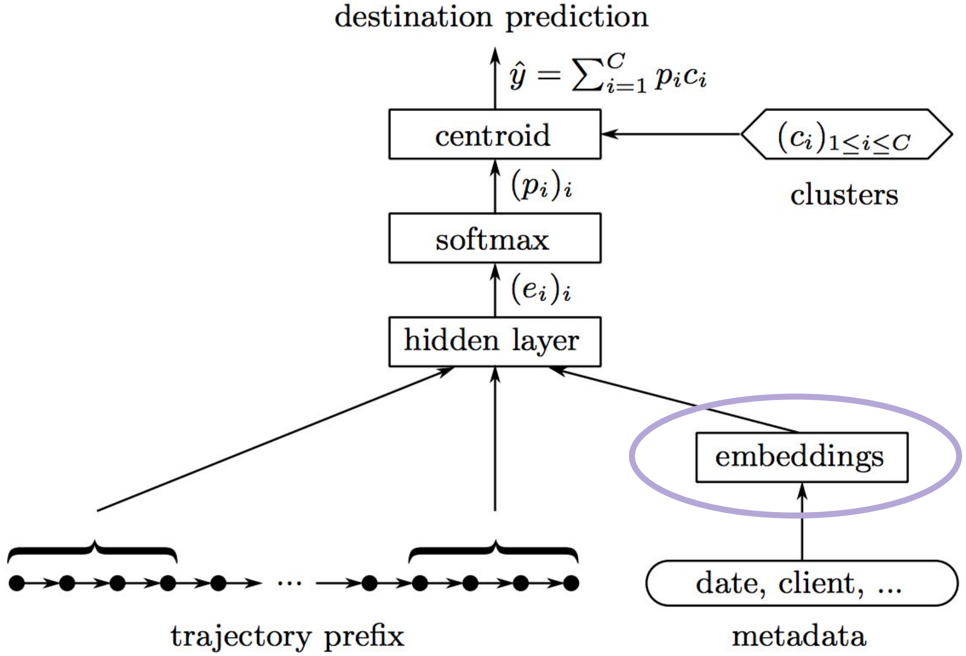


Fig. 1: Architecture of the winning model.

De Brébisson, Alexandre, et al. "Artificial neural networks applied to taxi destination prediction." arXiv preprint arXiv:1508.00021 (2015).

Current results (updated on Dec 4)

Model	Cost on Kaggle Private
Similarity-based	2.27896
NN-no embeddings	2.40713
NN (our best results)	1.84468
Winning method	1.87

References

- [1] De Brébisson, Alexandre, et al. "Artificial neural networks applied to taxi destination prediction." arXiv preprint arXiv:1508.00021(2015)
- [2] Hoang Thanh Lam, et al. “(Blue) Taxi Destination and Trip Time Prediction from Partial Trajectories” arXiv preprint arXiv:1509.05257(2015) https://arxiv.org/pdf/1509.05257.pdf
- [3] ECML/PKDD 15 Taxi Trip Time Prediction (II): 1st Place Solution write-up
- [4] Kaggle discussion board: https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/discussion

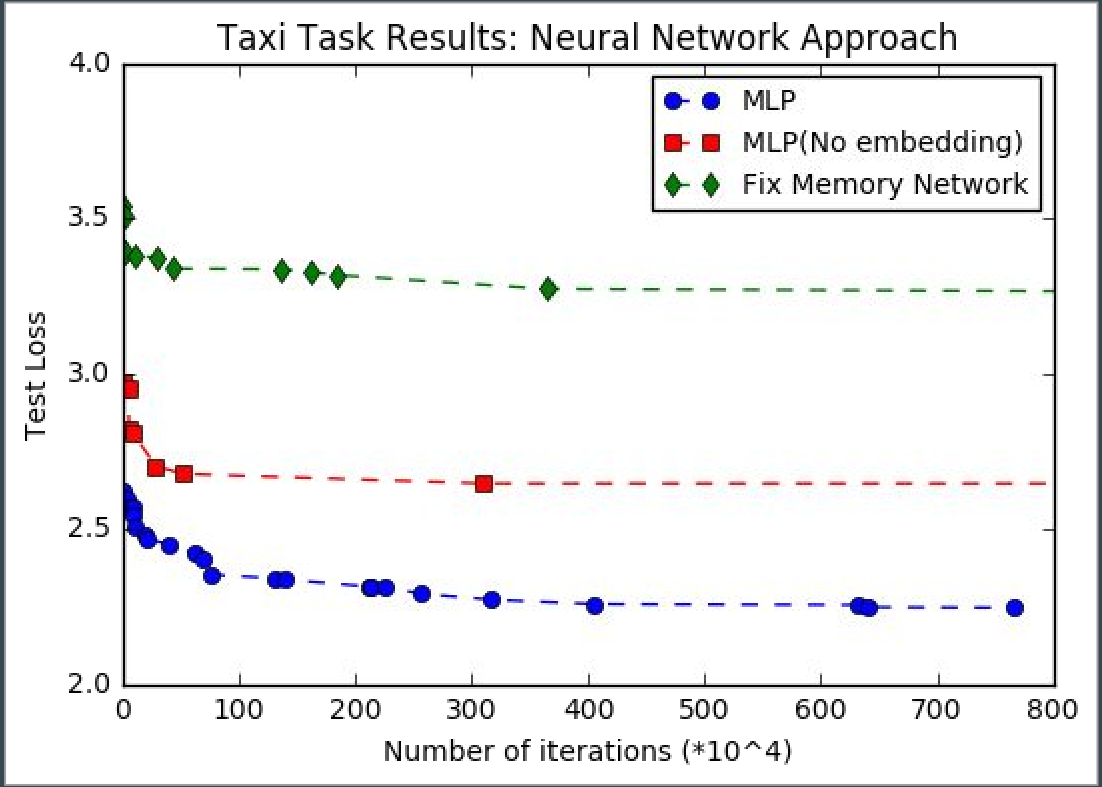
The End

Thank you for listening :)

Q&A

Backup slides: Training process

Model design mostly counts for final results.



Backup slides: Metavalue Embedding

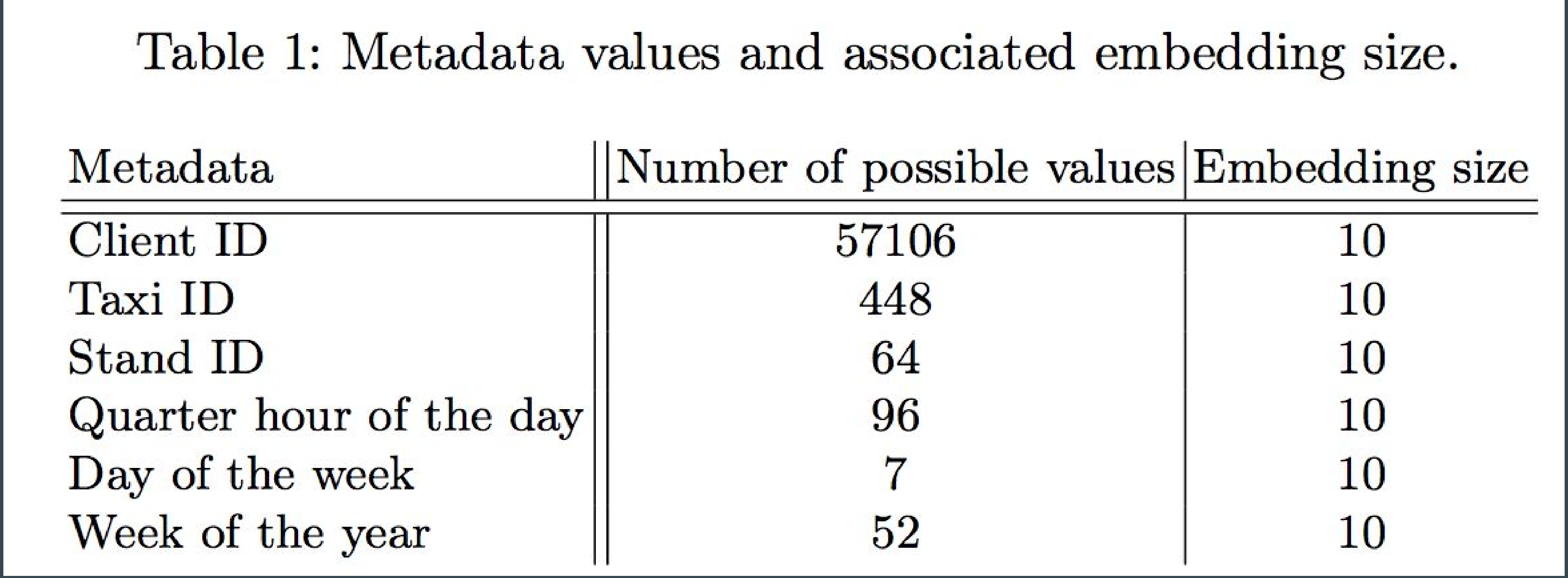
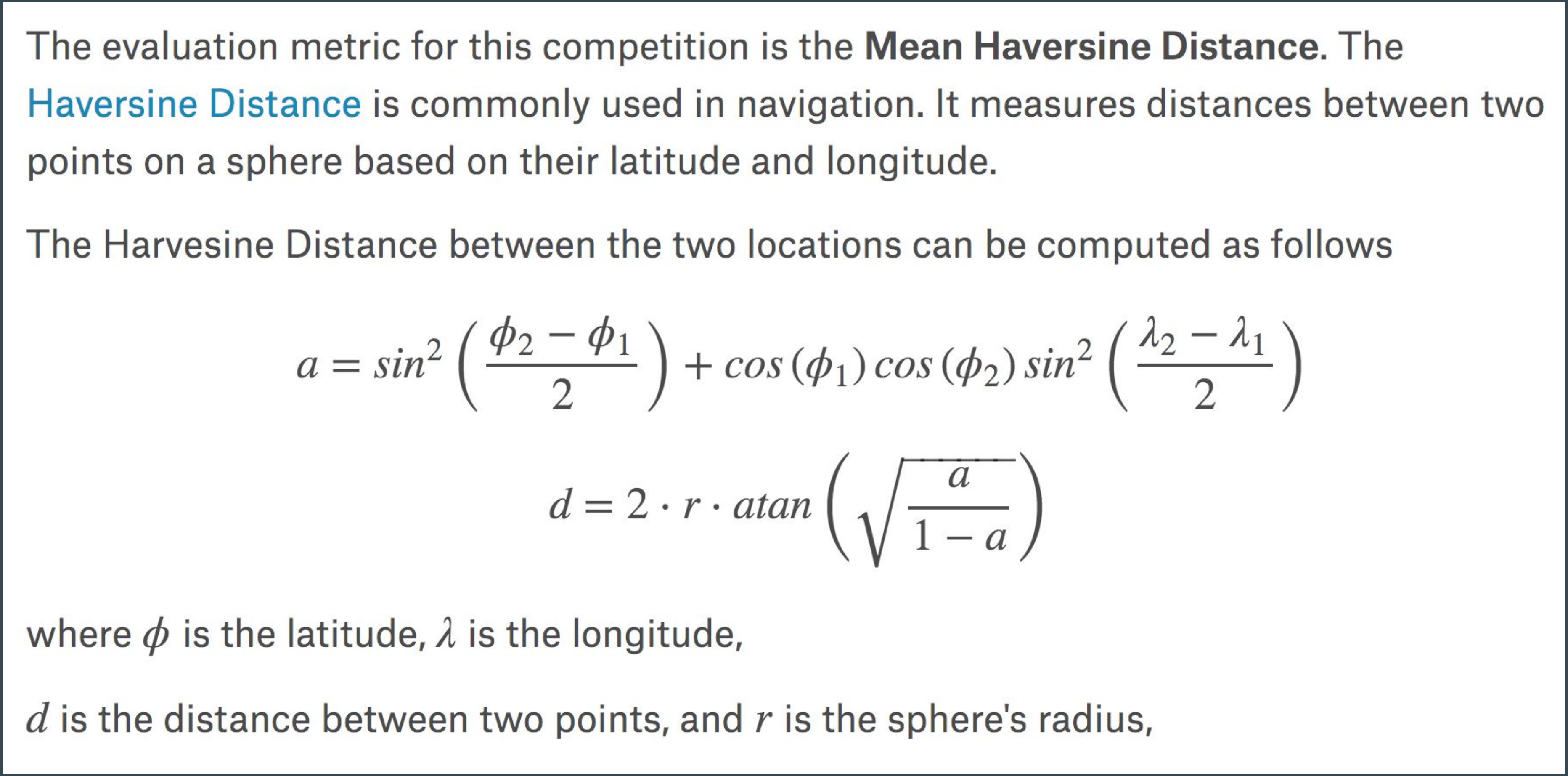


Table 1: Metadata values and associated embedding size.

Metadata	Number of possible values	Embedding size
Client ID	57106	10
Taxi ID	448	10
Stand ID	64	10
Quarter hour of the day	96	10
Day of the week	7	10
Week of the year	52	10



The evaluation metric for this competition is the **Mean Haversine Distance.** The Haversine Distance is commonly used in navigation. It measures distances between two points on a sphere based on their latitude and longitude.

The Harvesine Distance between the two locations can be computed as follows

$$a = sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos\left(\phi_1\right) cos\left(\phi_2\right) sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)$$

$$d = 2 \cdot r \cdot atan\left(\sqrt{\frac{a}{1 - a}}\right)$$

where ϕ is the latitude, λ is the longitude,

d is the distance between two points, and r is the sphere's radius,