A Simple But ***Quantifiable*** Approach to

# DYNAMIC PRICE PREDICTION WITH MULTI-SOURCE URBAN DATA

## in Ride-on-Demand Services

Suiming Guo, Chao Chen, Jingyuan Wang, Yaxiao Liu, Ke Xu, Daqing Zhang, Dah Ming Chiu

Suiming Guo, Chao Chen, Jingyuan Wang, Yaxiao Liu, Ke Xu, Daqing Zhang, Dah Ming Chiu, ***A Simple but Quantifiable Approach to Dynamic Price Prediction in Ride-on-demand Services Leveraging Multi-source Urban Data***, Volume 2, Issue 3, 112:1-112:24, IMWUT, 2018

香港中文大學 The Chinese University of Hong Kong
重慶大學 CHONGQING UNIVERSITY
北京航空航天大學 BEIHANG UNIVERSITY
清華大学 Tsinghua University
北京大学 PEKING UNIVERSITY

# Outline

## A Brief Presentation of Our Work

**Background**

*-- Problem, Solution & Methodology*

**Multi-source Urban Datasets**

*-- RoD, Taxi, Public Transportation, POI, Weather*

**Basic Features**

**Composite Features**

**Model & Evaluation**

**Feature Contribution & Discussions**

### 01 RoD Service

Emerging transportation services with dynamic pricing -- Uber, Didi, Lyft, Shenzhou UCar, etc.

### 02 Dynamic Pricing

Trip fare = (dynamic) price multiplier * (fixed) normal price.

### 03 What We Do?

Based on a number of features extracted from fare estimation, calculate (or estimate, or predict) the price multiplier for this estimation.

### 04 What Data?

Multi-source urban datasets: RoD, Taxi, POI, weather, public transportation, etc.

**In Short...**

# Background

## Problem, Solution & Methodology

## PROBLEM

TAXI → Dynamic Pricing → UBER, 神州专车, •••

### Passenger's Worry

- price too high
- do I get a good price now?
- better price later or nearby?
- ...

### Industry & Policy Maker's Worry

- price manupulation?
- killing taxi/buses?
- fairness?
- ...

## SOLUTION

### Price Predicton

- What will be the price multiplier, given features?
- & the price multiplier later?
- & the price multiplier nearby?
- ...

**For Passenger:** SOLVED

**For Industry & Policy Maker:**

- price manipulation -- solved
- killing taxi/buses -- introduce features relevant to taxi and buses
- fairness -- all passengers now have the same guidance

## METHODOLOGY

✗ **Unveil the "secret algorithm"**

✓ **Learn from data, don't care the truth**

↓ **Model selection?**

✗ **Complex & non-linear models:**
**highly accurate, non-interpretable**

✓ **Linear models:**
**interpretable, hard to describe non-linear correlation between features.**

# But, How To Compensate for the Drawbacks of a Linear Model?

1. Have more features -- use multi-source urban datasets
2. Add non-linear terms -- construct composite features

# Multi-source Urban Datasets

| | |
|---|---|
| RoD Service | **Event-log data**: record the time, location, price multiplier, etc. for every fare estimation on the app. |
| Taxi Service | **Taxi GPS trajectories**: used to describe the traffic condition and taxi availability around given locations. |
| Public Transport. | **Distribution of lines/stations**: how many bus/metro lines/stations around given locations. |
| POI | **Number of POIs**: how many POIs of each categories are there around given locations. |
| Weather | **Weather**: the weather condition (temp, wind speed, humidity, pressure, etc.) around given locations. |

**Information about time, price, etc.**

**Information about location.**

# Feature Engineering

## Basic Features

| | |
|---|---|
| **RoD Service** 神州专车 | **Temporal**: hour-of-day, day-of-week, day-of-month, isHoliday, isWeekend, ...<br>**Price**: estimated trip fare, price multiplier, historical price multiplier, ... |
| **Taxi Service** | **Taxi availability**: number of full taxis, full taxi ratio, up/down count, ...<br>**Traffic condition**: average speed, variance of speed, ... |
| **Public Transport.** | **Public transport. availability**: number of bus/metro lines/stations nearby |
| **POI** | **Number of POIs**: the number of POIs of 14 different categories (i.e., restaurant, shopping, business, etc.) |
| **Weather** | **Weather**: temperature, wind speed, humidity, pressure, visibility, weather condition (e.g., 'rainy', 'sunny', 'sprinkler', ...), etc. |

**Features Extracted from Single Dataset**

# Feature Engineering

## Composite Features

**Lack of non-linear terms**: how to describe correlations between features?

**Idea**: Adding product-form terms.

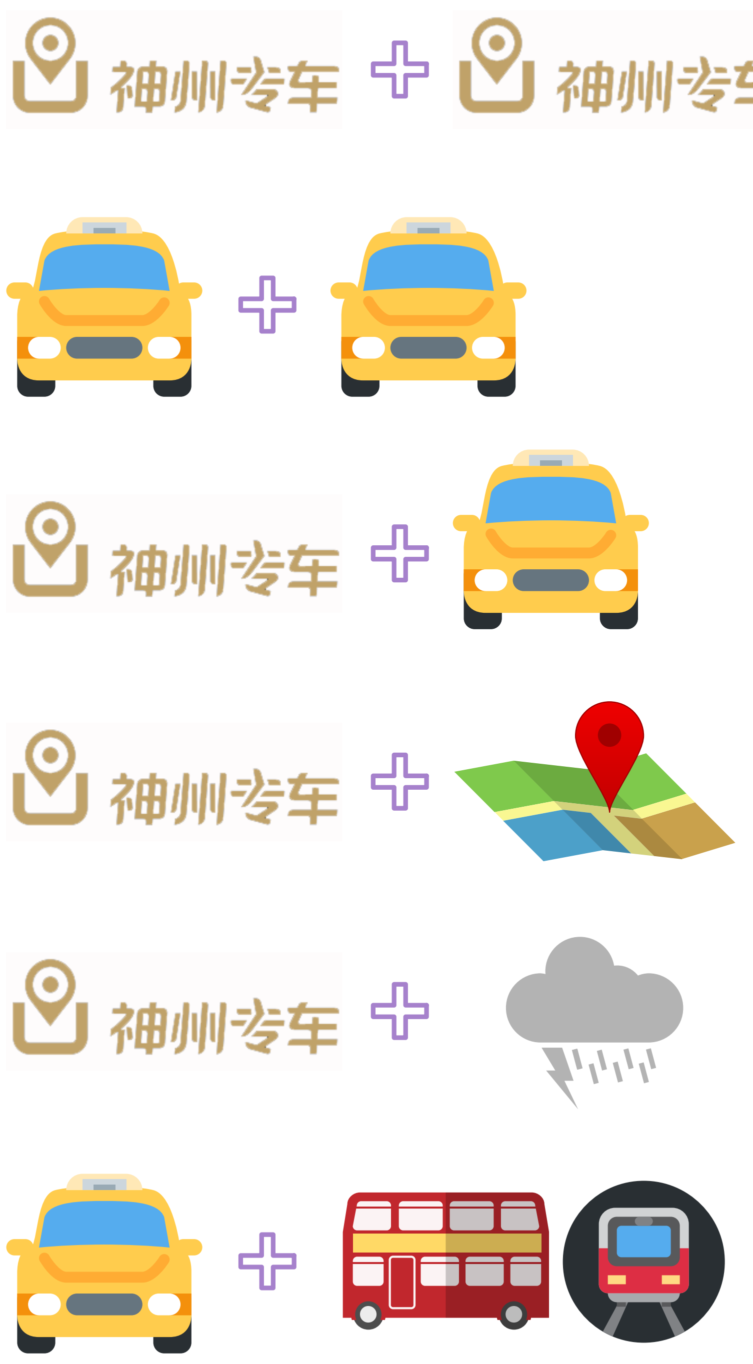
$$y = \omega_a x_a + \omega_b x_b + b$$

$$y = \omega_a x_a + \omega_b x_b + \omega_c x_a x_b + b$$

The linear model is transformed into a non-linear one, but still interpretable.

**Composite features**: combining basic features in product-form

**Examples of Combining Basic Features**




(day-of-week, hour-of-day)
(isWeekend, hour-of-day)...

(full taxi count, full taxi ratio)
(average speed, up count)...

(hour-of-day, full taxi ratio)
(historic price multiplier, full taxi ratio)...

(day-of-week, POI counts)
(hour-of-day, POI counts)...

(historic price multiplier, temperature)
(day-of-week, weather condition)...

(taxi count, bus station count)
(full taxi count, metro line count)...



Details of *normalization*, *generating product forms*, and *discussions* are omitted -- please refer to the paper.

# MODEL AND EVALUATION

## A Linear Regression Model

### What to Model?

- Given features describing a fare estimation, output the estimated or predicted price multiplier.
- As long as one can generate these features, the price multiplier could be estimated.

### How to Model?

- Simple linear regression model.
- Squared-error loss function: the difference between the predicted price multiplier and the ground truth;
- L1/L2 regularization: controls spasity and over-fitting;
- Spatio-temporal regularization: makes sure that the price multiplier does not change abruptly;
- Dimension of features: about 4000. Not very large as we don't combine more than 2 basic features together.

### Overall Results

- Use sMAPE as the evaluation metric;
- Have a very good prediction (with difference of price multiplier <= 0.1) in 81.17% cases;
- Using **composite features** reduces the sMAPE by 35.86%;
- Using **multi-source urban datasets** reduces the sMAPE by 25.17%;
- Tentatively, using a four-layer neural network could further reduce the sMAPE by 6.50%, and has a very good prediction in 85.68% cases. But neural network is not interpretable and quantifiable.

# FEATURE CONTRIBUTION

## Top-20 Features According to Absolute Weights

**Top datasets**

1. RoD Service
2. Weather
3. Taxi Service

**Top features (examples)**

- **historic price multipliers** (in the last 1, 2, 3 hours):
- --> with weights more than twice than others;
- --> basic and composite features;
- --> shows that prices are predictable.

- **weather**:
- --> higher temperature;
- --> when there is rain;
- --> quantifiable.

- **Competition between Taxi and RoD**:
- --> there is, but not that fierce;
- --> fewer taxis, higher prices;
- --> only obvious during evening rush;

**Public Transport.?**

- no strong influence on price multiplier: with weights far smaller than the top features.
- possible reasons: different customer bases --> needs further verification by other data.

**Influence of POI Features?**

- Again, no strong influence, with weights even smaller than public transportation.
- possible reasons: "POI counts" is not an accurate description; rather, the importance of POI should be addressed.

- **LBS checkin data:**
- --> more checkins, more important
- --> indeed increases weights by 20 times

- **TF-IDF of POI counts:**
- --> common POI categories get weights diminished.
- --> increases weights by at most 17.5 times

# THANK YOU FOR YOUR INTEREST!


## I'm from

The Chinese University of Hong Kong,
and will soon join Jinan University, Guangzhou, China.

## Email Address

guosuiming@email.jnu.edu.cn

## Wechat