Using Telematics and Insurance Data to Predict Accident Risk: Evidence from Beijing

Kanyao Han

Research Question

How can we improve accident risk prediction in this digital era?

Telematics Device

What?

How?

Driving behavioral data

Traditional method in Insurance Industry

Customer Records → Logistic regression (or other regression)

Demographic features: Gender, Age, Job…

Vehicle features: Price, Type, Equipment…

Self-report driving mileage

Previous claim record

$$log\left(\frac{P(y_i = 1|d_{ij}, v_{ik}, m_i, c_i)}{1 - P(y_i = 1|d_{ij}, v_{ik}, m_i, c_i)}\right) = \alpha + \sum_{j=1}^{J} \beta_j d_{ij} + \sum_{k=1}^{K} \gamma_k v_{ik} + \delta m_i + \eta c_i + \epsilon_i$$

Traditional Method: Drawbacks

- Actual user ≠ customer in insurance record (like family car).
- Demographic features are not usually good indicators (Jin, Deng & Jiang, 2018) .
- Self-reported records, such as annual driving mileage(White, 1976) , are usually not exactly same as the actual ones.

A solution: combining telematics data and traditional insurance data.

Data! Data! Data!

- A confidential car insurance dataset from insurance company: 150,000 observations

can be merged by vehicle id number

- A confidential telematics dataset from telematics company (10,000 cars over 3 months):

In-vehicle sensor data: acceleration, hard brake, actual mileage…

GPS data: averagely each car contains over 10,000 GPS observation

An example of GPS data structure of a car

For ethical and confidential reason, I don’t display some identifiable variables.

VIN <chr>	lon <chr>	lat <chr>	time <S3: POSIXct>
	116.365120	39.953669	2016-01-01 09:50:06
	116.364989	39.953702	2016-01-01 09:50:39
	116.364955	39.953041	2016-01-01 09:50:51
	116.364960	39.952391	2016-01-01 09:51:03
	116.365124	39.951912	2016-01-01 09:51:16
Confidentiality	116.365120	39.951150	2016-01-01 09:51:26
	116.365201	39.950516	2016-01-01 09:51:35
	116.365211	39.949626	2016-01-01 09:52:10
	116.365295	39.948980	2016-01-01 09:52:19
	116.365168	39.948152	2016-01-01 09:52:40

1–10 of 22,573 rows

Previous 1 2 3 4 5 6 ... 100 Next

Extracting data from GPS (Some examples)

- Actual demographic features

  Where the car owner live → Economic status

- Driving behavior

  Night driving, urban driving, etc.

  Familiarity with the road (how often a driver/drivers driving in one area/road)

- Driving environment

  Various road conditions in which a driver/drivers driving the car (I also have some types of road conditions data).

Method: Spatial data aggregation

Modeling

- Response:

  Self-reported claim in insurance data ⟹ Accident or not

  Claim amount ⟹ Accident loss

- Features:

  Traditional features in insurance data

  Driving behavior in in-vehicle sensor data

  Demography, behavior and environment in GPS data

Model Selection

In data-driven research, there is no golden standard for model selection. Trying different algorithm and parameter.

- Classification (self-reported claim):

  Neural network often performs other algorithm in accident prediction (Paefgen et al. , 2013)

- Regression (claim amount):

  Lasso and elastic net are frequently used when there is dozens of features.

So what?

- For insurance company: improving pricing strategy based on telematics data

Cooperation

- For telematics company: risk scoring service based on both classification and regression

- For academics: most similar research are just based on in-vehicle sensor data. The driving behavior and environment information extracted from GPS data have not been given enough attention. It has large potential for prediction.