Learning and Prediction over Light-Weight Spatio-Temporal Data

BlackSwan Team
chenyitian@jd.com

August 16, 2017

JD.京东.COM

Yang Guo, Yitian Chen, Research engineer@JD.COM

Jie Lin, Master Candidate, Nanjing University of Science and Technology

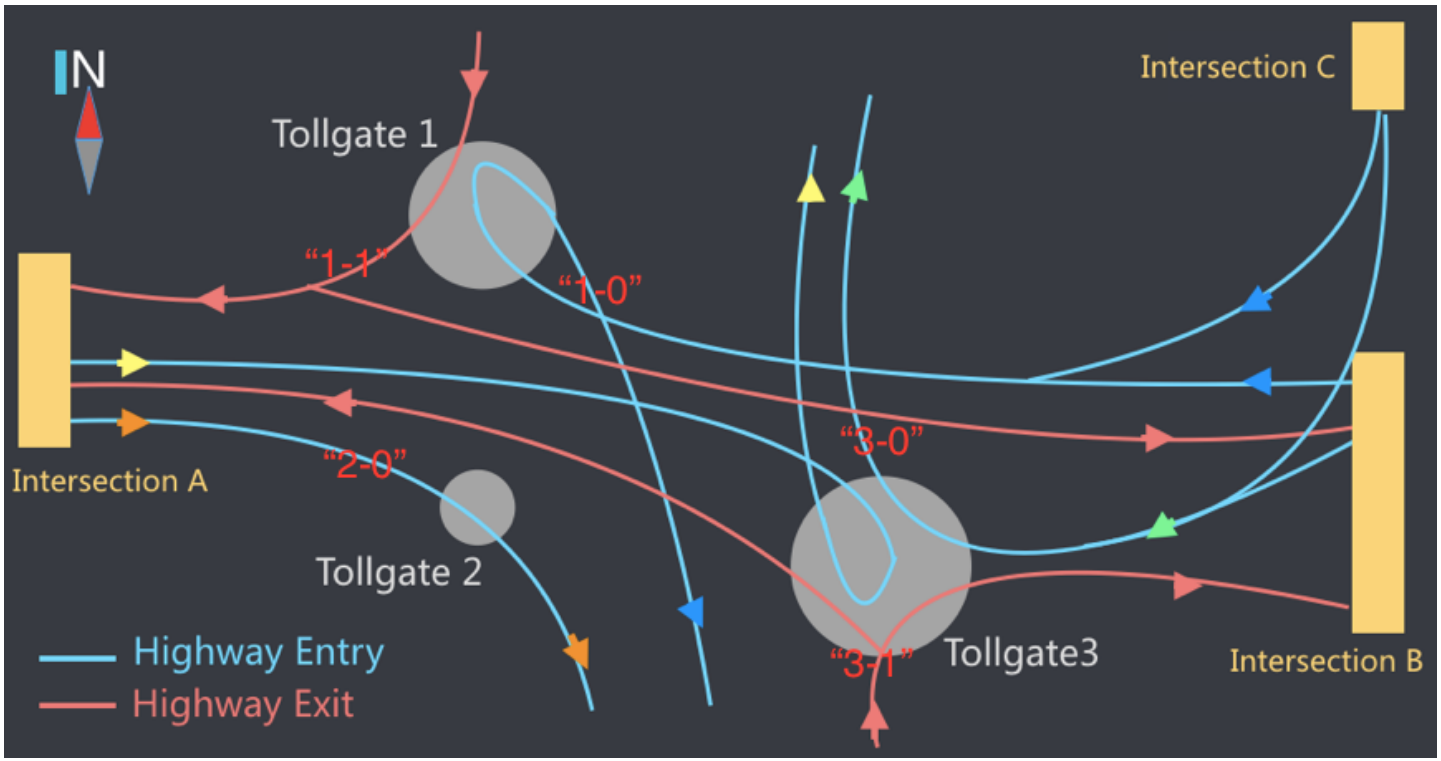
Jie Zhou, PHD Candidate, East China Normal University

Hao Lin, Research engineer@Tencent

JD.京东.COM

Motivation

- Help authorities do better decision making.
- Expedite the toll collection process.
- Streamline future traffic flow and reduce congestion.



Directions: entry:0; exit:1

Task Formulation

Task: Given 5 tollgate-direction pairs and previous two-hour vehicle records, predict the traffic volume of every 20-minute time window for the next 2 hours.

Data

- Testing days: previous 2-hour vehicle records.
- Training days: vehicle records of all days (24 hours).
- Weather: humidity, precipitation, wind....

time	tollgate_id	direction	vehicle_model	has_etc
2016-09-21 05:47:44	1	0	2	0
2016-09-21 05:52:19	1	0	1	0
2016-09-21 05:53:27	1	0	1	0
2016-09-21 05:54:46	1	0	2	0
2016-09-21 05:55:26	1	0	1	0
2016-09-21 10:09:31	1	0	1	0
2016-09-21 10:09:33	1	0	1	1
2016-09-21 10:10:53	1	0	1	0
2016-09-21 10:11:13	1	0	1	0
2016-09-21 10:11:55	1	0	1	0

Directions: entry:0; exit:1

Note: the traffic volume for a given tollgate-direction pair is the total volume of all vehicles that enter/exit the tollgate in that time window. Each 20-minute time window is defined as a right half-open interval, e.g., [2016-09-21 8:00:00, 2016-09-21 8:20:00).

Task Formulation

Basic ML approach:Use previous 6 20-minute time-window volume points to predict the next 6 points.

$$X = \begin{array}{c} \begin{array}{cccccc} \textit{Lag1} & \textit{Lag2} & \dots & \textit{Lag6} & \textit{tsDistance} & \textit{otherFeatures} \end{array} \\ \begin{bmatrix} V_{[7:40,8:00]} & V_{[7:20,7:40]} & \cdots & V_{[6:00,6:20]} & 20 & \cdots \\ V_{[7:40,8:00]} & V_{[7:20,7:40]} & \cdots & V_{[6:00,6:20]} & 40 & \cdots \\ V_{[7:40,8:00]} & V_{[7:20,7:40]} & \cdots & V_{[6:00,6:20]} & 60 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ V_{[7:40,8:00]} & V_{[7:20,7:40]} & \cdots & V_{[6:00,6:20]} & 120 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}, \quad Y = \begin{array}{c} \textit{Labels} \\ \begin{bmatrix} V_{[8:00,8:20]} \\ V_{[8:20,8:40]} \\ V_{[8:40,9:00]} \\ \cdots \\ V_{[9:40,10:00]} \\ \cdots \end{bmatrix} \end{array}$$

Objective: minimize the MAPE($\hat{Y}$, Y): $\frac{1}{n}\sum_{t=1}^{n}|\frac{y-\hat{y}}{y}|$. Works to do:

- Feature engineering: transfer raw data into features better represent the problem (Refer GIT for details).
- Modeling: design model framework/policy given the specific problem (**focused in this talk**).

GIT:https://github.com/12190143/Black-Swan

JD.京东.COM

Challenges

Challenge-1 The vehicle volume of a route varies a lot depending on.

- Time of day.
- Day of the week.
- Holidays vs normal days.

Challenge-2 Small dataset: Only 29 (36 for stage 2) days' data of 5 tollgate-direction pairs is provided. And it's very noisy.

Challenge-3 Evaluation metrics (MAPE): Most regression loss functions do not minimize APE(Absolute percentage error) directly.

- MSE (Gaussian distribution): $loss = \frac{1}{2}(y - \hat{y})^2$.
- MAE (Laplace/Quantile distribution): $loss = |y - \hat{y}|$.

JD.京东 .COM

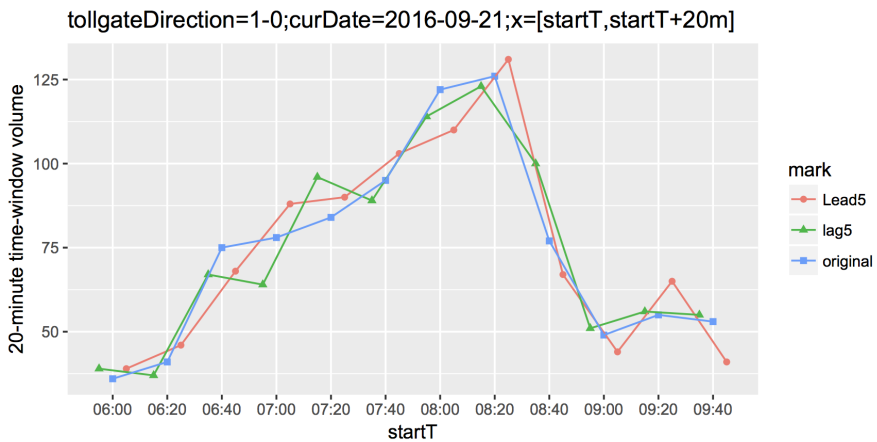
Strategy

Data-Augmentation: Augment the data by sliding time windows from $w(t)$ to $w(t+\pi)$.

	previous-2-hour	to-be-predicted
original	[6:00,6:20], [6:20,6:40],..., [7:40,8:00]	[8:00,8:20], [8:20,8:40],...,[9:40,10:00]
$\pi = 5$	[6:05,6:25], [6:25,6:45],..., [7:45,8:05]	[8:05,8:25], [8:25,8:45],...,[9:45,10:05]
$\pi = -5$	[5:55,6:15], [6:15,6:35],..., [7:35,7:55]	[7:55,8:15], [8:15,8:35],...,[9:35,9:55]
...	...	...

Regression-Objective: We consider two ways of approximating the evaluation objective.

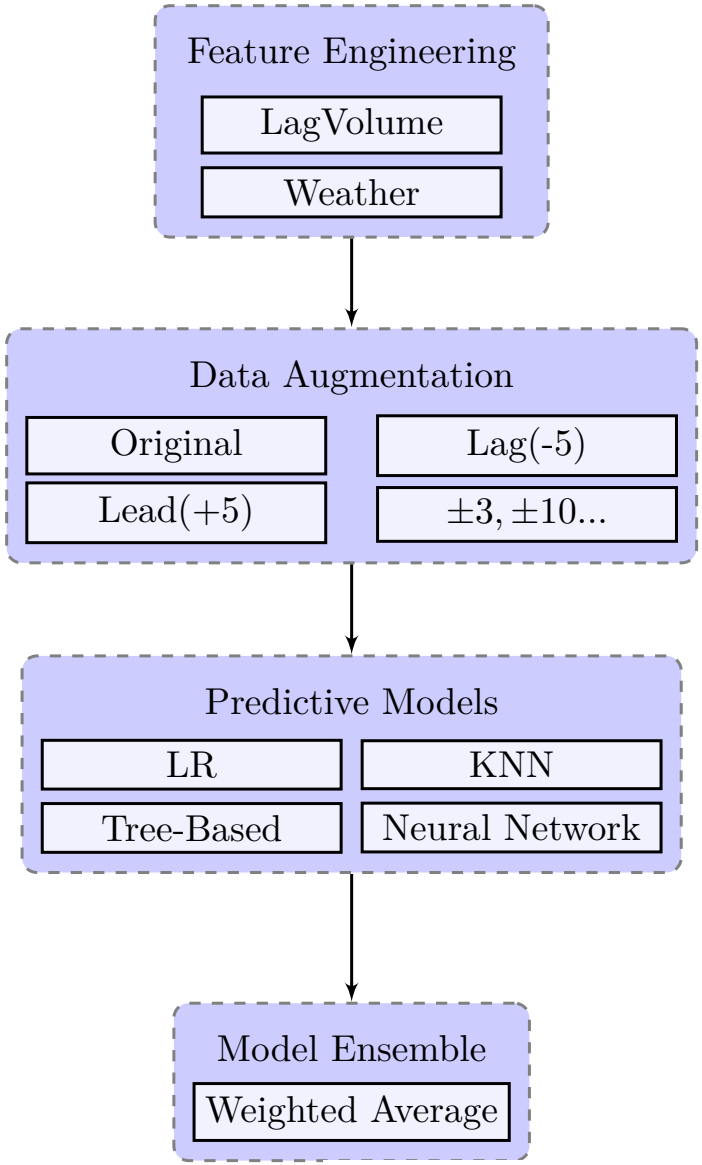
- Logarithm-Transform: $|log\hat{y} - logy| = |log\frac{\hat{y}}{y}| = |log(1 + \frac{\hat{y}-y}{y})| \approx |\frac{\hat{y}-y}{y}|$ (APE).
- Quantile-Regression: Minimize $|y - \hat{y}|$, a quantile point a little smaller than 0.5 (prediction a little small than median).



original vs new

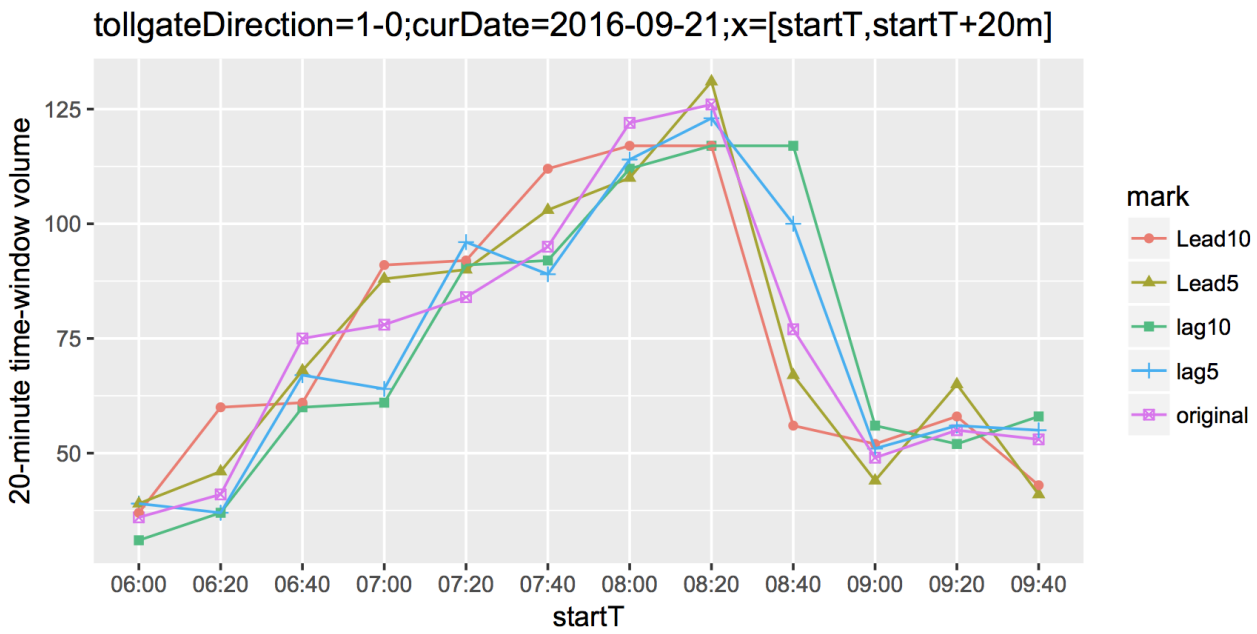
Solution Framework

- **Original data**: 573140 vehicle records of 5 tollgate-direction pairs from Sep 19 to Oct 24.
- **Data augmentation**: Sliding time-window with different timestamps.
- **Feature engineering**: Vehicle records aggregation, weather data preprocessing;reformulate the data for ML training.
- **Model training**: Train different models with absolute-loss or log-transform.
- **Model ensemble**: Weighted average of multiple model results.



JD.京东 JD.COM

Experiment: Data Augmentation



π(Lead)	mean	sd	aveCor	π(Lag)	mean	sd	aveCor
...	...	...	...	...	...	...	...
3	0.157	5.53	0.954	-3	-0.199	5.61	0.954
5	0.224	7.39	0.915	-5	-0.327	7.41	0.919
10	0.358	11.18	0.812	-10	-0.693	11.51	0.838
15	0.562	14.67	0.717	-15	-1.079	14.45	0.737
20	0.714	17.89	0.621	-20	-1.47	17.42	0.638
...	...	...	...	...	...	...	...

Experiment

Experiment:Take last 7 days (Oct 18 to Oct 24) as test data (leaderboard), the remains as training data.

Model	Data	Approach	MAPE
LR	original	Log-Trans	0.146
KNN	original	Abs-Dist	0.140
GBDT	original	Gaussian	0.147
LightGBM	original	Log-Trans	0.133
LightGBM	original, ±1, ±5, ±10	Log-Trans	0.1222
NN	original, ±3, ±5	Quantile	0.1219
Ensemble		Weighted Average	0.1150

Experiment

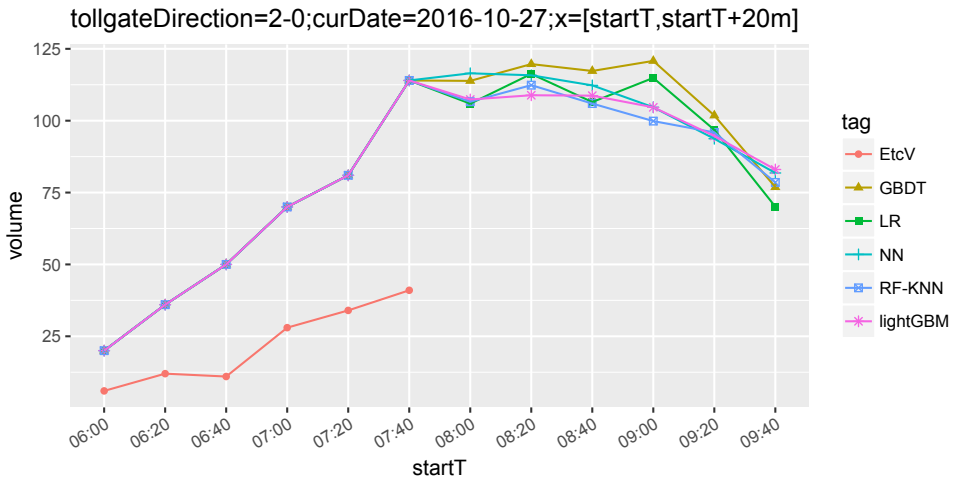
Performance of Neural Network:

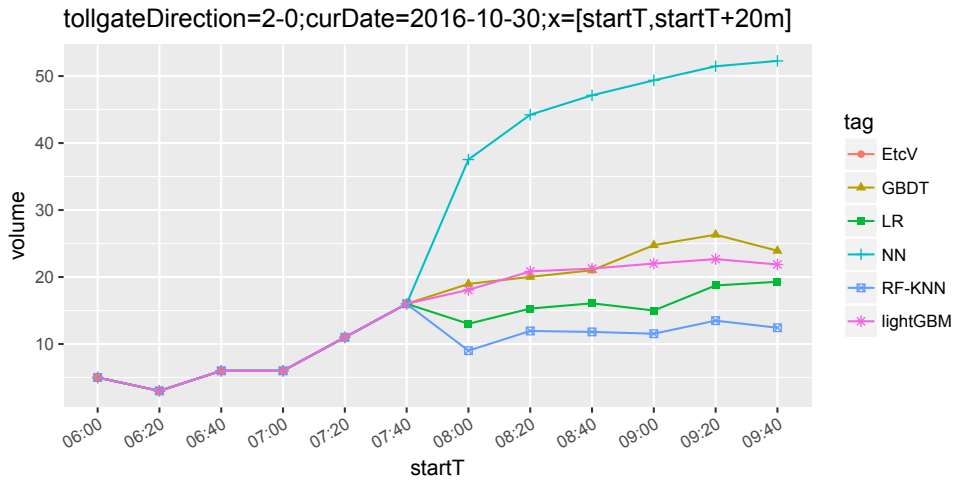
- Laplace/Quantile regression for absolute loss $|\hat{y} - y|$.
- Two NN Model results, eg, "with-dropout" A and "without-dropout" B.
- $\alpha \cdot min(A, B) + (1 - \alpha) \cdot max(A, B)$ $(\alpha \in [0.5, 0.7])$.
- Apply cross-validation to find the best α $(\alpha \in seq(0.5, 0.7, by = 0.05))$.
- Train with "adadelta" for adaptive learning rate.

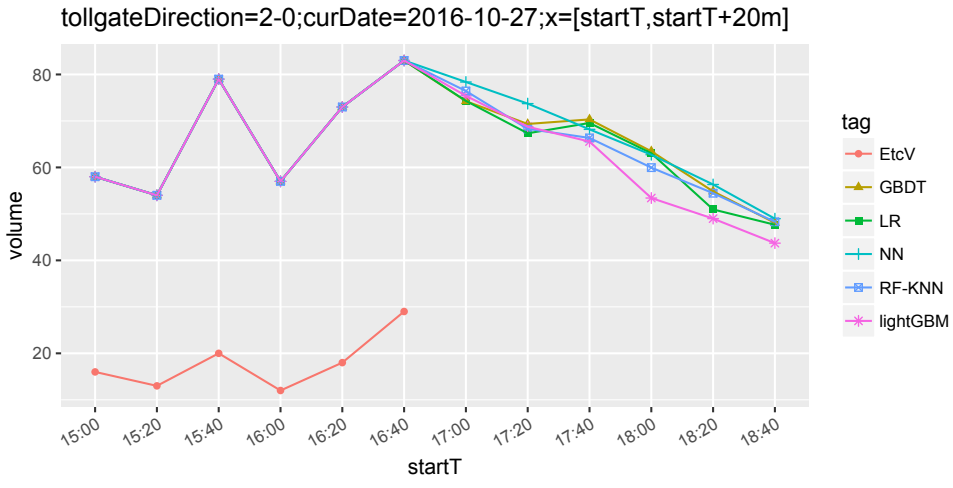
Model	Data	Approach	MAPE
3-Layers	original	gaussian	0.144
3-Layers	original	new	0.138
3-Layers	original, ± 3, ± 5	...	0.133
2-Layers	original, ± 3, ± 5	...	0.1277
1-Layer	original, ± 3, ± 5	...	0.1219

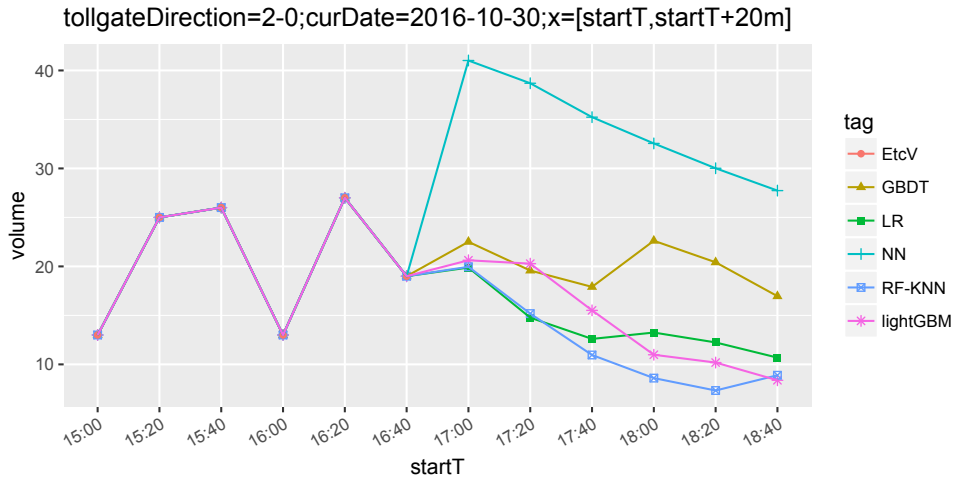
JD.京东.COM

Decision Making: which models can we trust?





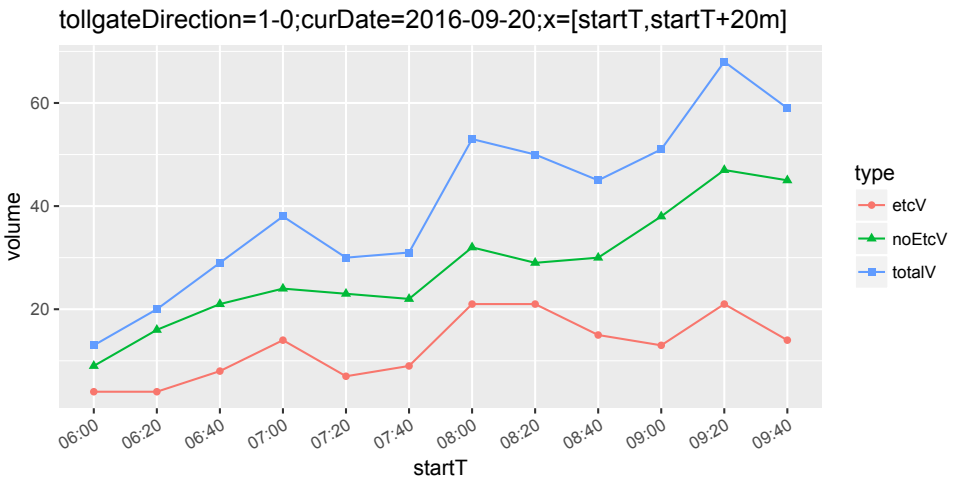


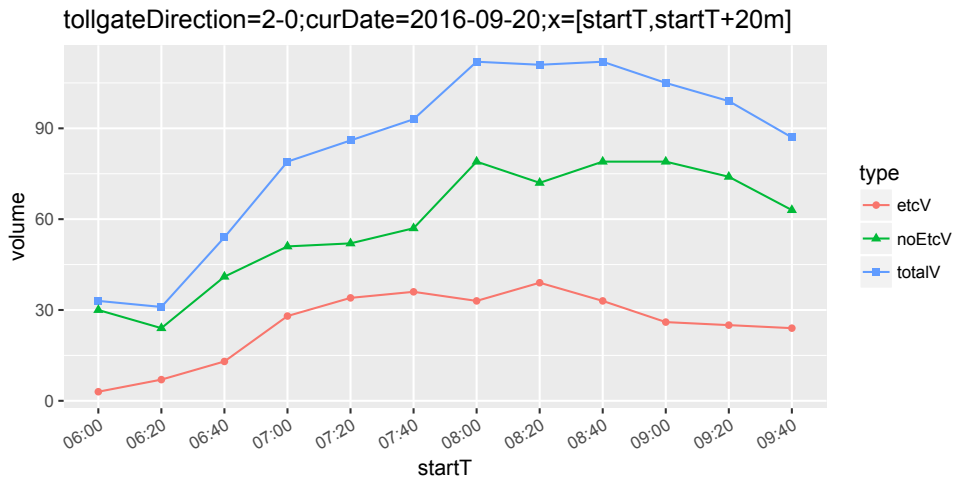


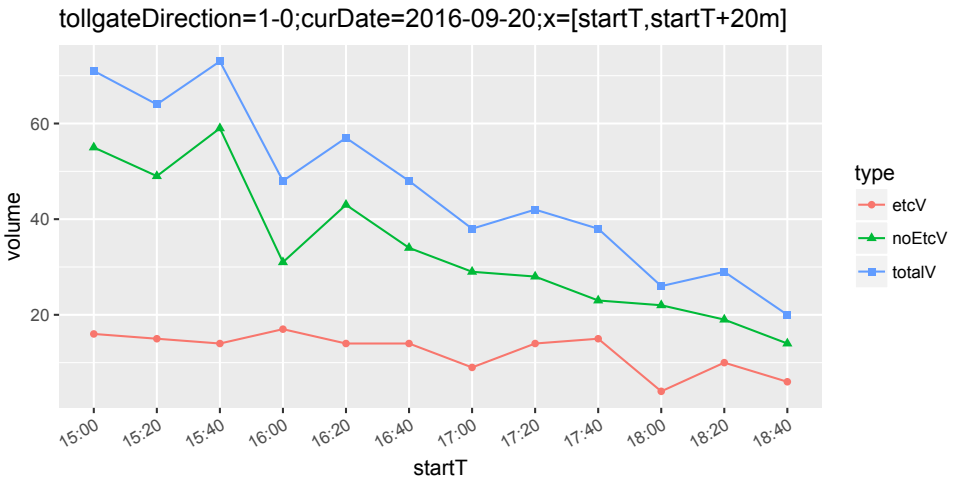
Normal model results.

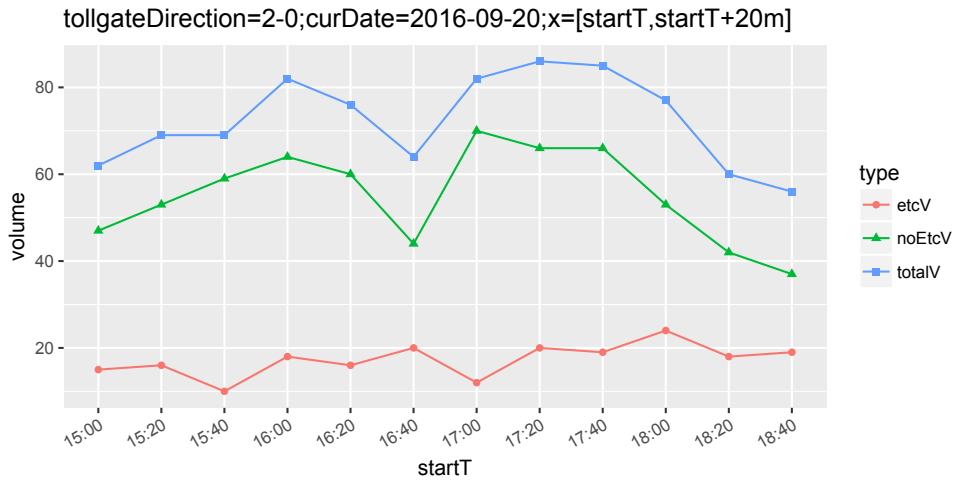
Huge differences among model results

Observation: Normal days







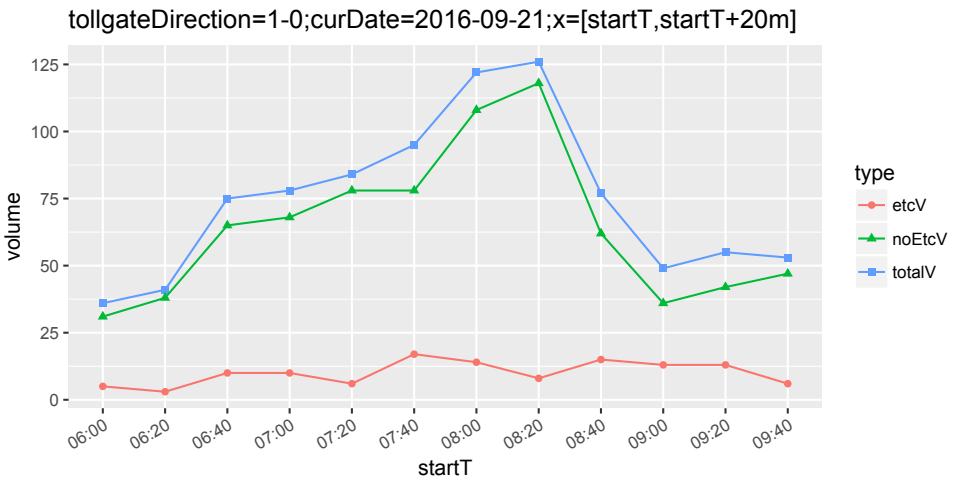


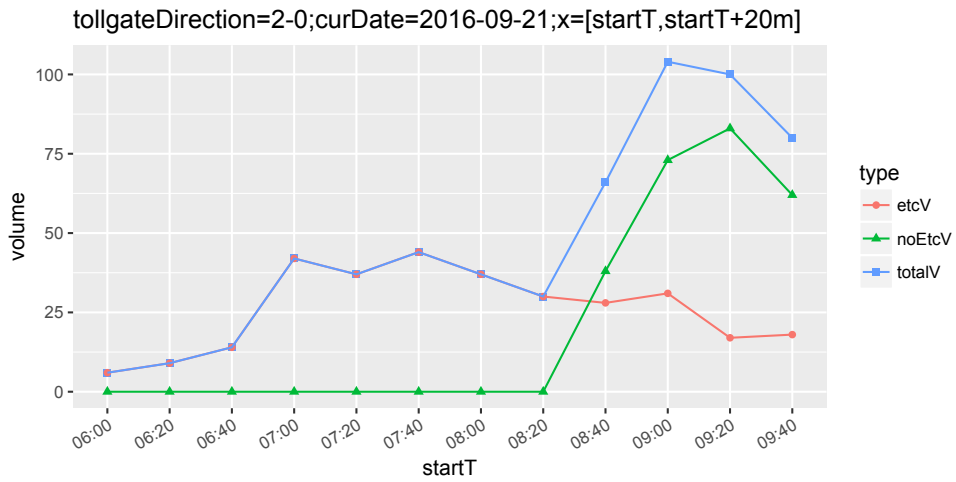
ETC introduction guide

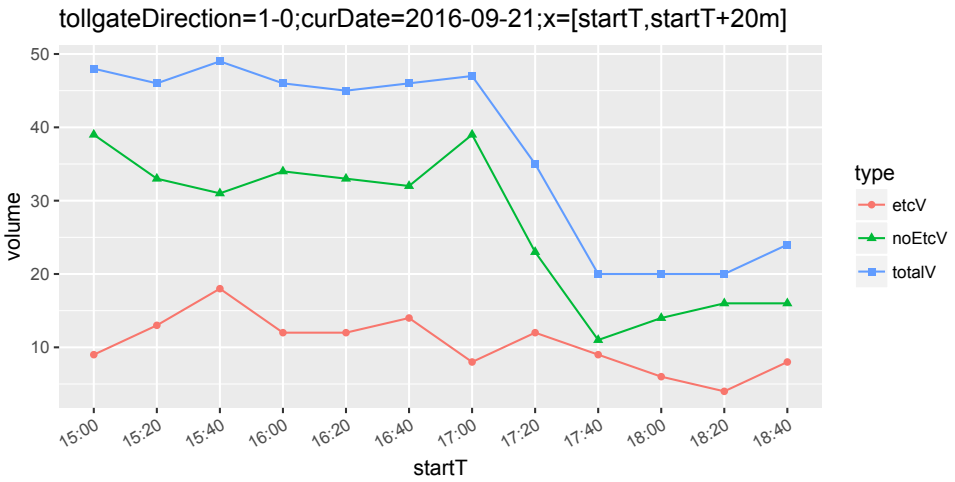
The ETC in-vehicle device

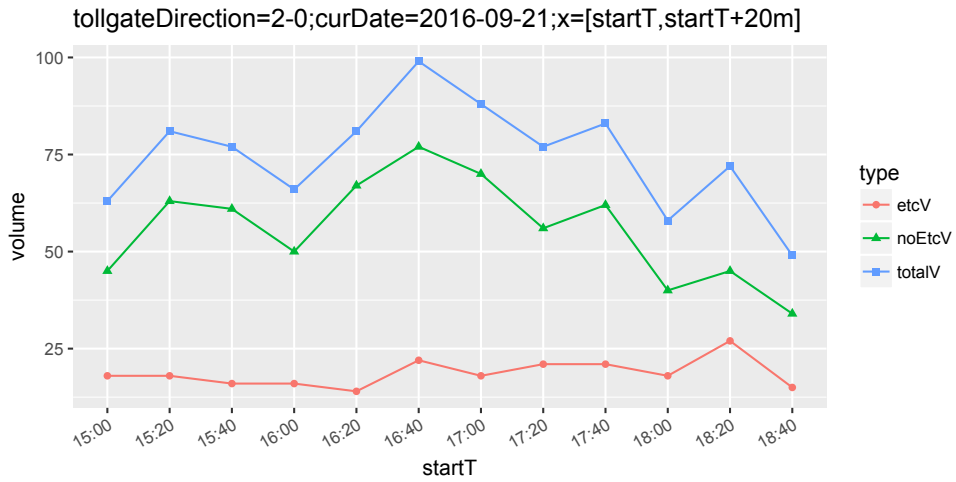
ETC systems can be used simply by inserting the ETC card into the in-vehicle device. The ETC in-vehicle device is equipped with a function which wirelessly communicates with the antenna set up at the toll booths to send and receive vehicle information necessary for paying the appropriate toll fare.

Observation: Abnormal days







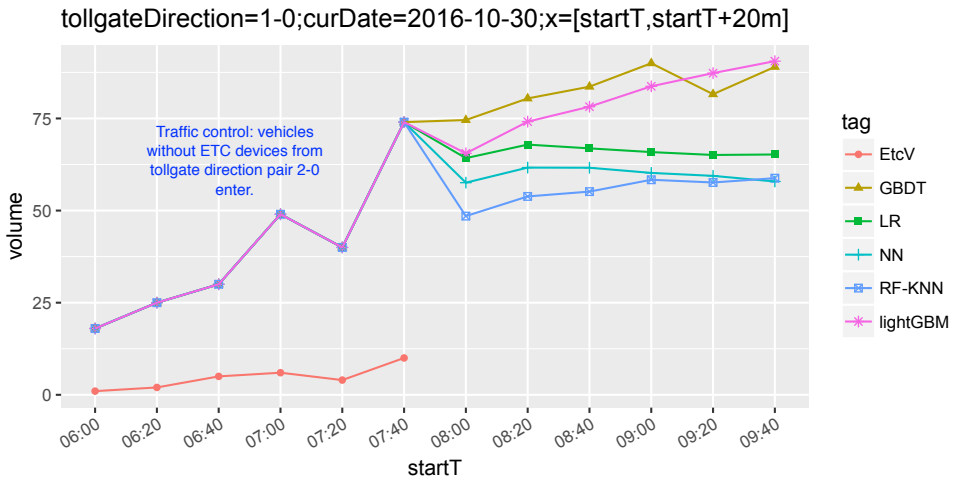


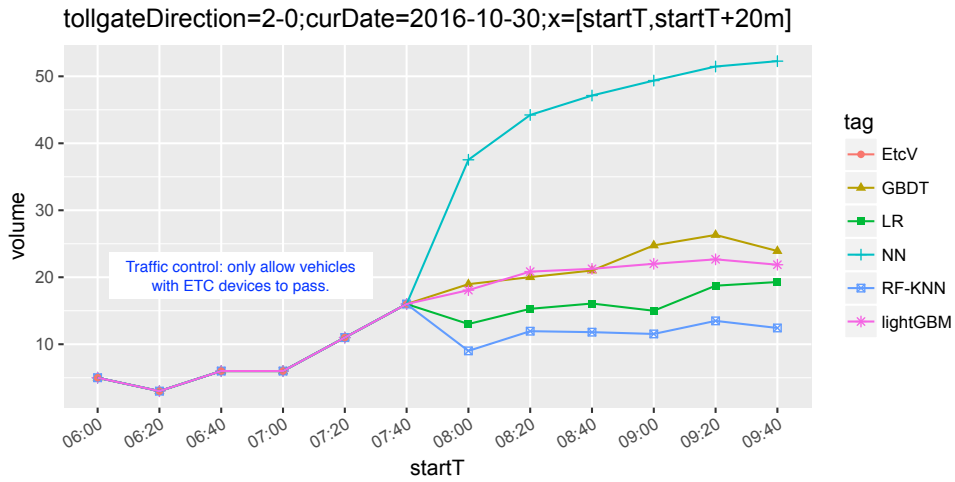
ETC introduction guide

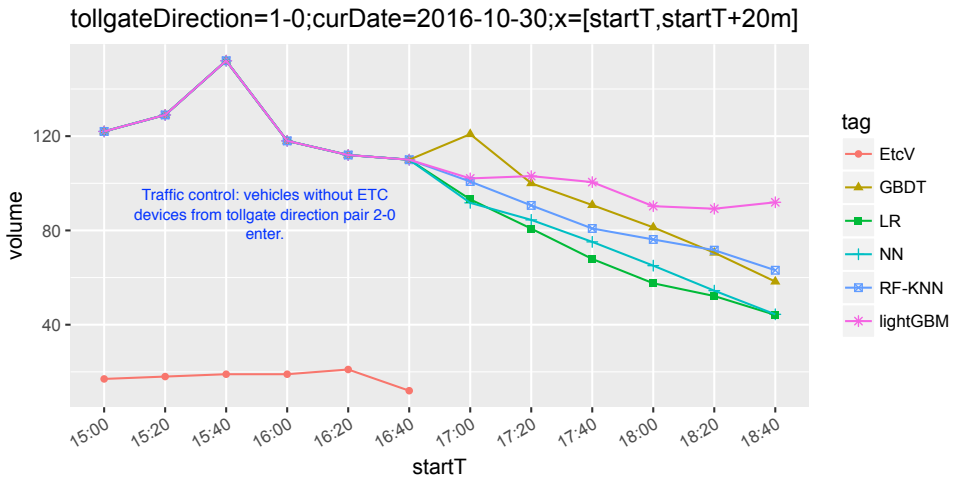
The ETC in-vehicle device

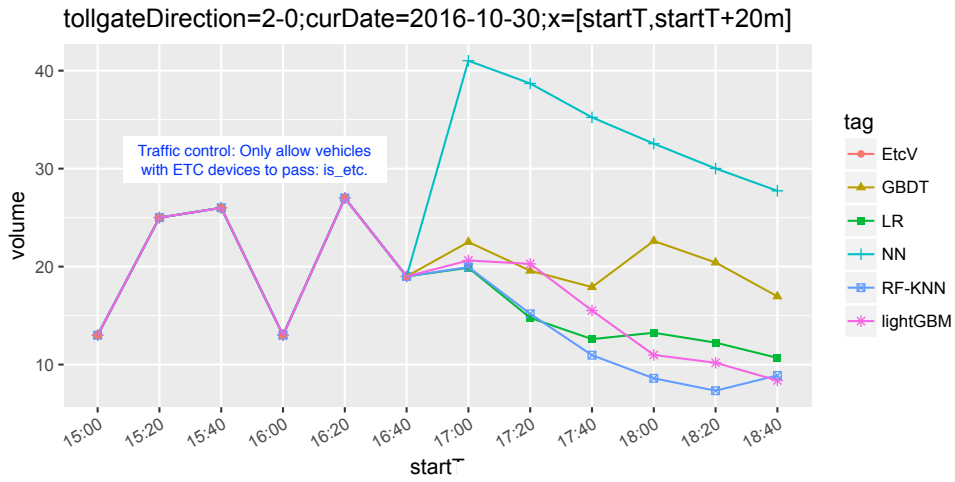
ETC systems can be used simply by inserting the ETC card into the in-vehicle device. The ETC in-vehicle device is equipped with a function which wirelessly communicates with the antenna set up at the toll booths to send and receive vehicle information necessary for paying the appropriate toll fare.

Analysis and Inference









Decision Making Continued

Business Assumption

- Total traffic volumes entering tollgate 1,2 are stable (1-0, 2-0).
- When tollgate 2 only allow is_etc vehicles to enter, those without ETC devices will turn to tollgate 1 (1-0).

Model Selection

- Build new baseline models with data:
  - 2-0: is_etc volume data.
  - 1-0: total volume + no_etc volume of 2-0.
- Bagging of selected model result(s) close to the baseline model results.

JD.COM & Y-Business-Units

About JD.COM

- China's largetst retailer, online or offline–236.5 million shoppers.
- World's third largest internet company by revenue – $37.5bn in 2016.
- Renowned for our zero-fakes policy and amazingly fast delivery.

Y Business Units: Focus on retailing service and smart supply chain, working out demand forecasting, inventory optimization, dynamic pricing with artificial intelligence and operation research technologies.

JD.COM 京东

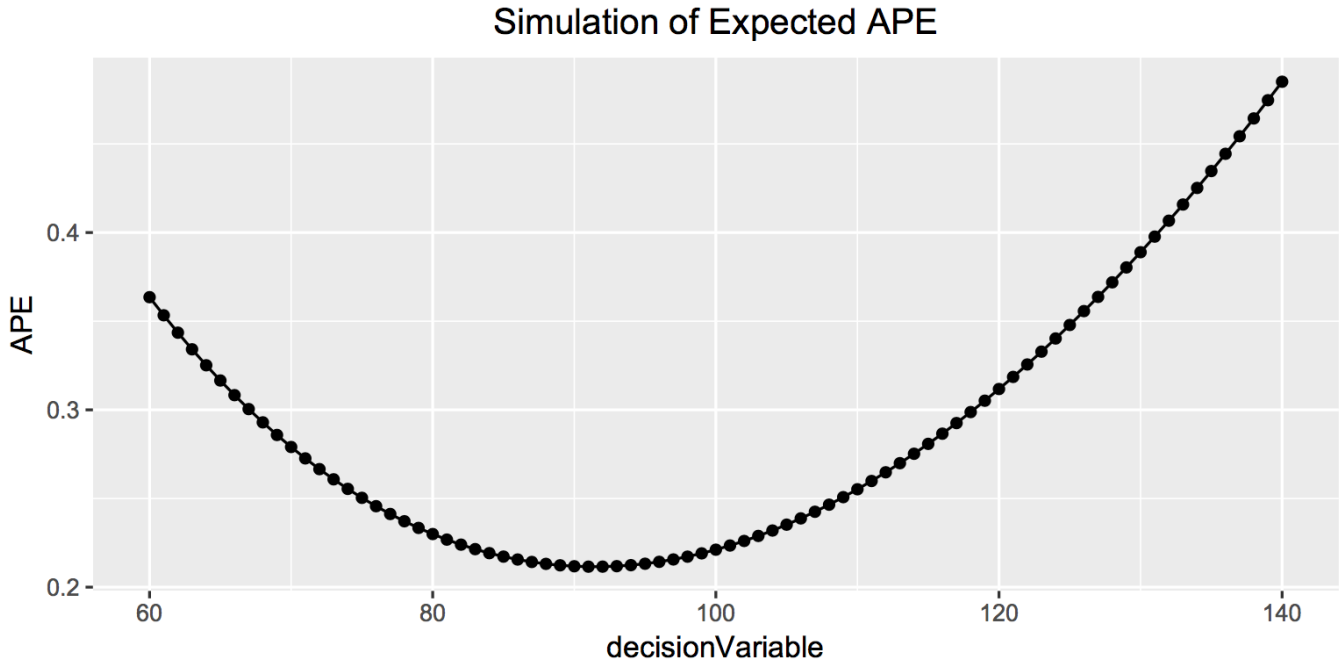
Thank you!

JD.京东
JD.COM

Appendix: Quantile regression approach

- We are trying to minimize the APE (absolute percentage error) $|\frac{\hat{y}-y}{y}|$ for every training sample.
- Laplace/quantile regression(with quantle_alpha=0.5) minimize the absolute error $|\hat{y}-y|$.
- Assume the true value y randomly distributed around the predicted value $\hat{y}$.
- We want to do optimal decision $\hat{y}_2$ that minimize the expected APE.
- A decision variable of $\hat{y}_2$ which is a little smaller than $\hat{y}$ achieve better(smaller) absolute percentage error.

Example: prediction value $\hat{y} = 100$, true value follows discrete uniform distribution in the interval $[60, 140]$.



Appendix: Neural network training details

- Apply randomized grid search with parameters list below.
- Bagging of blend of top-K results.

parameters	list
activation	"Rectifier","Tanh"
l1,l2 regularization	c(0, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7)
input_dropout_opt	c(0,0.05,0.1,0.2)
max_w2_opt	10,20,30,40,50
adadelta-epsilon	c(1e-4,1e-6,1e-8,1e-10)
adadelta-rho	c(0.9,0.95,0.99,0.999)

grid search parameters list