Prediction of Domain Name Renewal Rate by Machine Learning Focusing on Registered Domain Name String

4 Nov 2019

ICANN66 TechDay

Yoshiro Yoneya, Kentaro Mori, Osamu Inomoto

<yoshiro.yoneya@jprs.co.jp> <kentaro@jprs.co.jp> <inomoto@jprs.co.jp>

Table of Contents

1. Introduction (1/3)

- To maintain DNS operational stability is essential to maintain stability of the Internet
  - Capital of critical DNS servers' operation is depending on TLD registries' revenue
- JPRS is making next year’s revenue plan with predicting renewal rate† of newly registered (.jp) domain name in high precision
  - Prediction is performed heuristically by extremely experienced person in charge
  - This is too individualistic and difficult to transfer to others
  - Thus, we have formed small team (three members including the experienced person) to evaluate prediction with machine learning, and achieved meaningful results
- We expect our proposed method will be applicable to other TLDs

† Renewal period of .jp Domain Names is one year, and its expiration date is the end of the month of one year later. Registration renewal process runs monthly.

1. Introduction (2/3)

- Heuristic method by person
  - Perform prediction of domain name renewal rate of targeted month by statistical analysis (least square method)
    - With compensation of experienced person
    - i.e. lowering the rate if the amount of domain name labels with random string or embedding words such as “test” is relatively high
- Our proposed method
  - Perform prediction from the result of decision tree analysis with supervised machine learning
    - Used 49 features that can be obtained from just label string (explained later)

1. Introduction (3/3)

- Examples of ‘Heuristically’ checked out domain names
  - Random string registered simultaneously

    2ymc4hyv　b23fk9hl
  - Including event name and/or year number

    joker-movie　election2018
  - Embedding somewhat ‘test’ flavored string

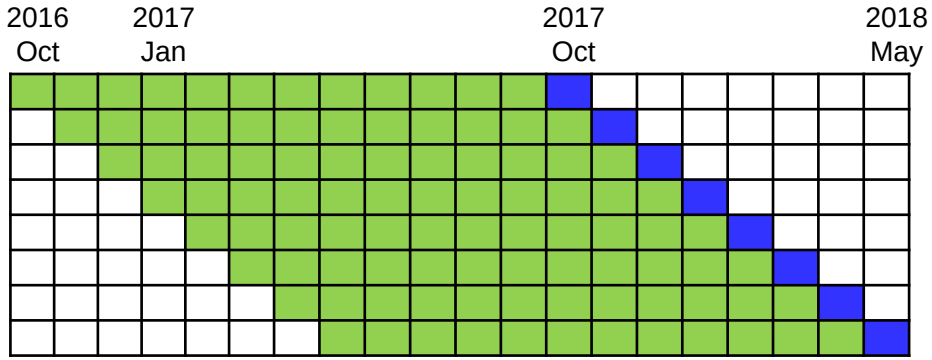
    unit-test-1　unit-test-2

    unit-test-3　unit-test-4
  - Including online game name

    rankup-gta　cheating-pubg

2. Methods / Dataset (1/9)

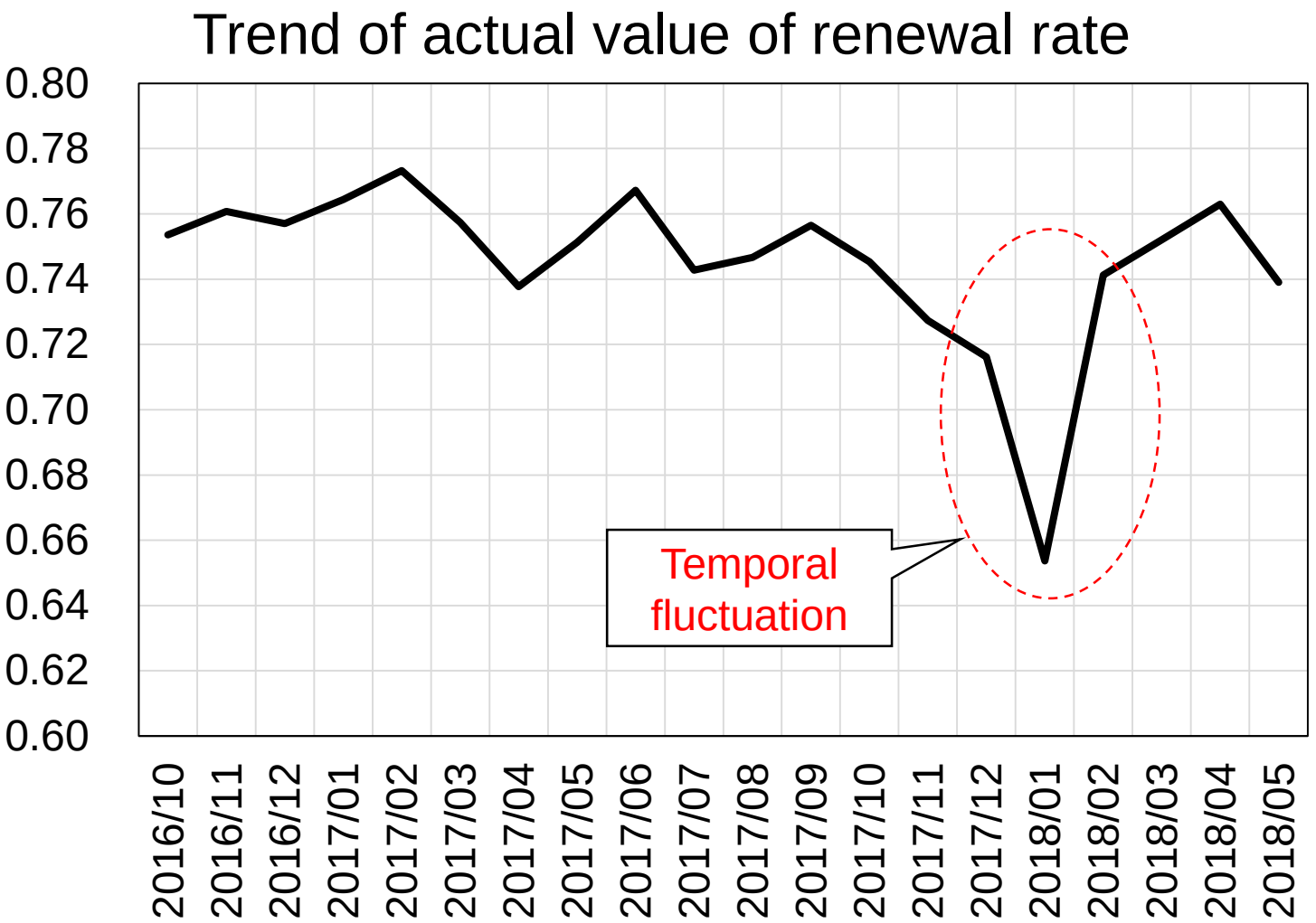
- Test data (for prediction)
  - Newly registered .jp domain names† of each month, from Oct 2017 to May 2018 (8 months == 8 test sets)
  - They all have ‘renewal status’ (*i.e.* renewed or not) at the end of the first registered year
  - So we can validate the ‘prediction’ results comparing to the ‘fact’.
- Training data (for generating decision tree)
  - Newly registered .jp domain names† over the last 12 months preceding to the targeted month (test data)



Green: Training data
Blue: Test data

†: ASCII SLD (General-Use ASCII JP Domain Name) only

2. Methods / Dataset (2/9)



Actual value of renewal rate per month

YYYY/MM	Renewal Rate	YYYY/MM	Renewal Rate
2016/10	0.7536	2017/10	0.7453
2016/11	0.7608	2017/11	0.7273
2016/12	0.7570	2017/12	0.7162
2017/01	0.7644	2018/01	0.6537
2017/02	0.7733	2018/02	0.7413
2017/03	0.7574	2018/03	0.7521
2017/04	0.7377	2018/04	0.7630
2017/05	0.7513	2018/05	0.7390
2017/06	0.7672		
2017/07	0.7428		
2017/08	0.7467		
2017/09	0.7565		

2. Methods / Dataset (3/9)

- Statistical methods (to compare with proposed method) -

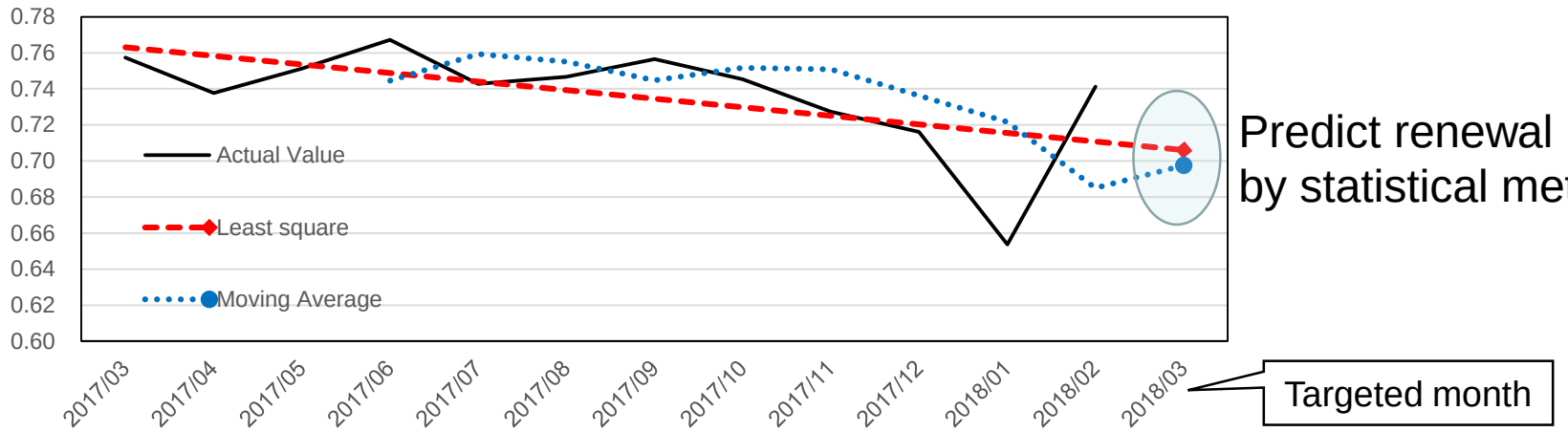
Calculate renewal rate of targeted month by following methods

Method 1: least square (linear approximation)

– With over the last 12 months actual value of renewal rate

Method 2: moving average

– With over the last 2 months actual value of targeted month

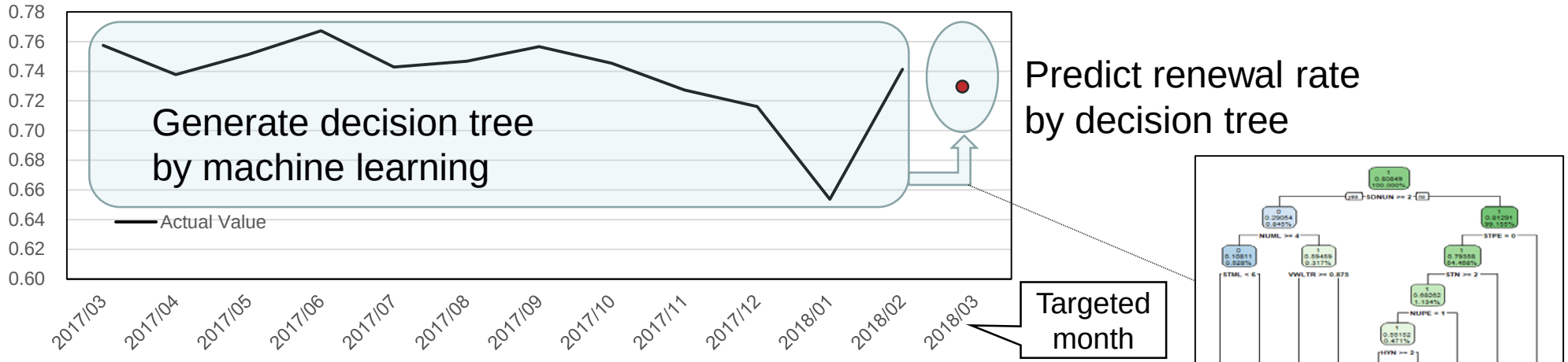


Result of prediction using least square and moving average (example of Mar 2018)

2. Methods / Dataset (4/9)

- Proposed method -

- Generate decision tree from over the last 12 months actual value of targeted month (same period with statistical methods)
- Then, apply condition of each node in the tree to domain names registered on targeted month and obtain predicted renewal rate for each domain name
- Predicted renewal rate of targeted month is calculated as mean value of the domain names



Result of prediction using decision tree (example of Mar 2018)

2. Methods / Dataset (5/9)

- Description of proposed method -

- Definition of features
  - Focused on features based on past experience
    Ex. Factors indicating randomness or temporariness:
    - Including numbers such as year number or sequential number
    - Random string such as continuous consonants or existence of single (independent) digit

➢ Length of label string

➢ Disposition of digits ("0" - "9"): Where? Continuous? etc.

➢ Letters ("A" - "Z") and Hyphen ("-") are as well

2. Methods / Dataset (6/9)

- Features related to label itself and digits -

- Definition of features obtained from domain name label itself

Target of features	No
Label	**1**
Digit	**13**
Letter	15
Hyphen	20
Total	49

Features of **label** (1)

Name	Explanation	Val
LEN	Length	int

Features of **Digit** (13)

Name	Explanation	Val
DIN	Number of digits	int
SDN	Number of single digit	int
NUN	Number of numerics	int
NUML	Max length of numerics	int
SDNUN	Sum of SDN and NUN	int
DIMI	Max distance between digits	int
NUPW	Whole label is digits?	bool
SDPB	Head of label is SD?	bool
SDPM	Mid of label is SD?	bool

Name	Explanation	Val
SDPE	End of label is SD?	bool
NUPB	Head of label is NU?	bool
NUPM	Mid of label is NU?	bool
NUPE	End of label is NU?	bool

SD: Single Digit
NU: Numeric

2. Methods / Dataset (7/9)

- Features related to letters -

- Definition of features obtained from domain name label itself

Target of features	No
Label	1
Digit	13
Letter	**15**
Hyphen	20
Total	49

Features of **Letter** (15)

Name	Explanation	Val
LTN	Number of letters	int
SLN	Number of single letter	int
STN	Number of strings	int
STML	Max length of strings	int
VWN	Number of vowels	int
CSN	Number of consonants	int
VWMI	Max distance between vowels	int
VWLTR	Ratio between vowels and letters	real

Name	Explanation	Val
STPW	Whole label is letters?	bool
SLPB	Head of label is SL?	bool
SLPM	Mid of label is SL?	bool
SLPE	End of label is SL?	bool
STPB	Head of label is ST?	bool
STPM	Mid of label is ST?	bool
STPE	Mid of label is ST?	bool

SL: Single Letter
ST: String

2. Methods / Dataset (8/9)

- Features related to hyphen -

SH: Single Hyphen
SD: Single Digit
SL: Single Letter
NU: Numeric
ST: String
HS: Hyphen sequence

- Definition of features obtained from domain name label itself

Target of features	No
Label	1
Digit	13
Letter	15
Hyphen	**20**
Total	49

Features of **Hyphen** (20)

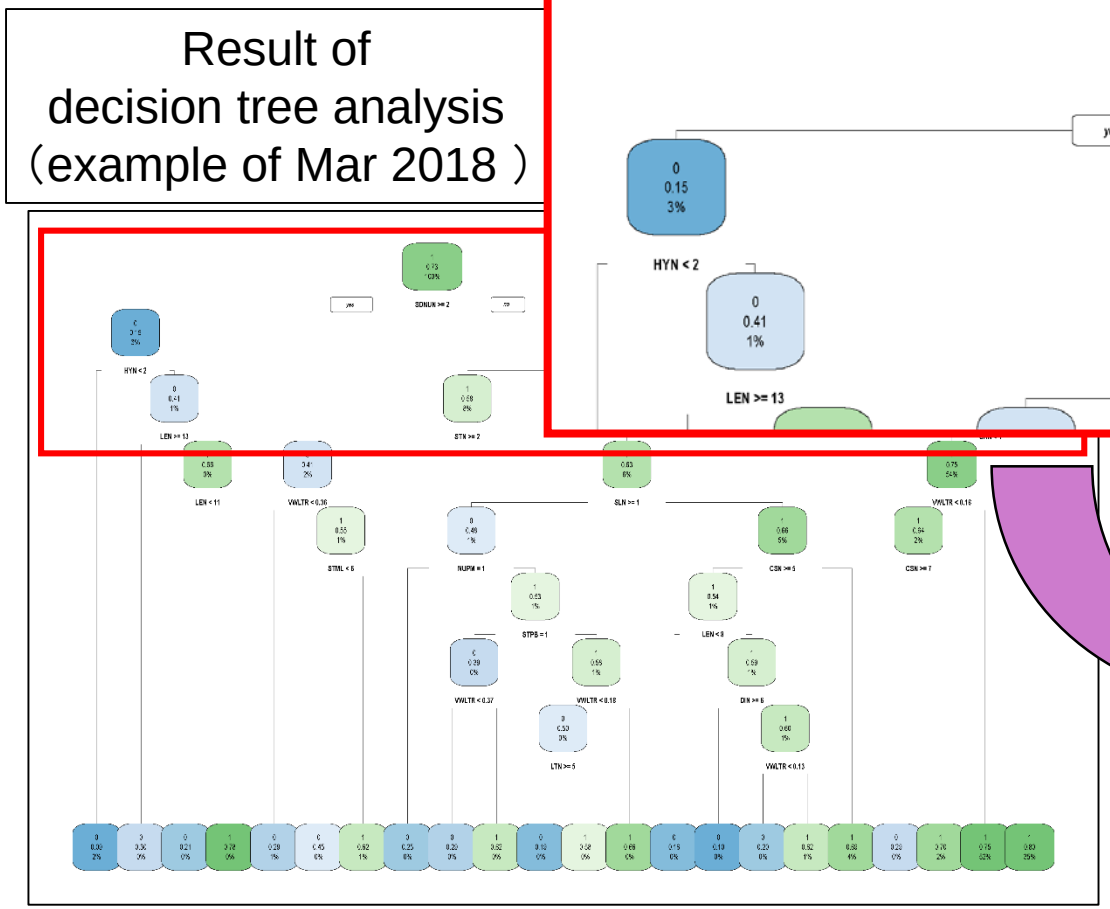
Name	Explanation	Val
HYN	Number of hyphens	int
SHN	Number of single hyphen	int
HSN	Number of hyphen sequences	int
HSML	Max length of hyphen sequences	int
SHPASD	SH after SD?	bool
SHPBSD	SH before SD?	bool
SHPASL	SH after SL?	bool
SHPBSL	SH after SL?	bool
SHPANU	SH after NU?	bool
SHPBNU	SH before NU?	bool

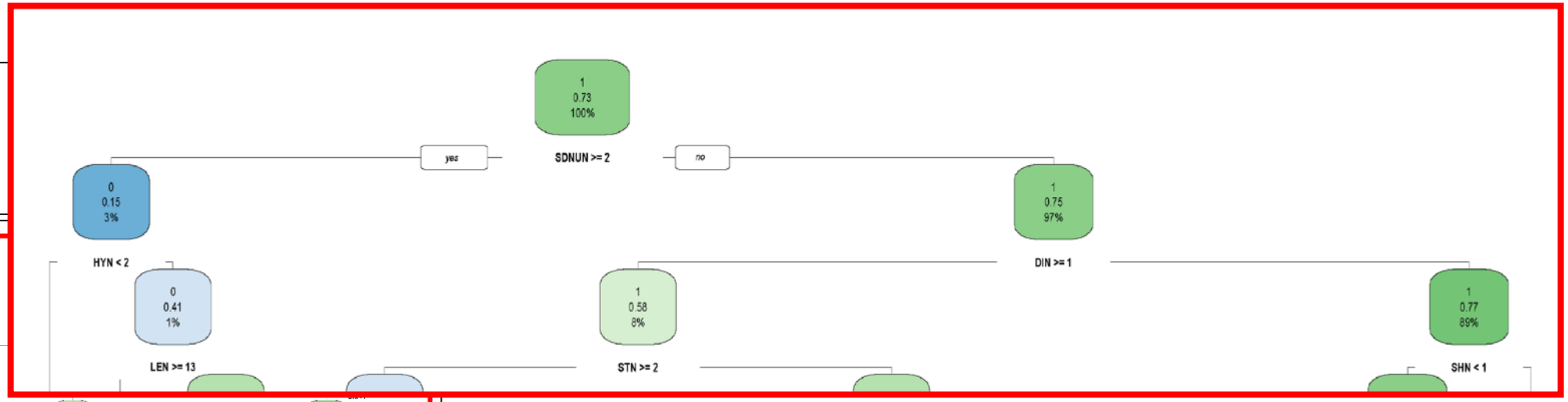
Name	Explanation	Val
SHPAST	SH after ST?	bool
SHPBST	SH before ST?	bool
HSPASD	HS after SD?	bool
HSPBSD	HS before SD?	bool
HSPASL	HS after SL?	bool
HSPBSL	HS before SL?	bool
HSPANU	HS after NU?	bool
HSPBNU	HS before NU?	bool
HSPAST	HS after ST?	bool
HSPBST	HS before ST?	bool

2. Methods / Dataset (9/9)

- Generation of decision tree by training data using ‘R’†, then applying the tree to test data -

Result of decision tree analysis （example of Mar 2018 ）





Zoom in

†: R is a free software environment for statistical computing and graphics <https://www.r-project.org>

3. Results (1/3)
- .jp domain names -

- We calculated predicted value of renewal rate for each method from Oct 2017 to May 2018 (8 months), and its deviation (absolute value) from actual value
- ML is the best for 5 times out of 8 times

Year	Month	Actual value	Predicted value of renewal rate		
			Least Sq.	Mov. Ave.	ML
2017	10	0.7453	0.7497 (0.0044)	0.7516 (0.0063)	0.7542 (0.0089)
	11	0.7273	0.7458 (0.0185)	0.7509 (0.0236)	0.7436 (0.0163)
	12	0.7162	0.7377 (0.0215)	0.7363 (0.0201)	0.7399 (0.0237)
2018	1	0.6537	0.7267 (0.0730)	0.7217 (0.0680)	0.7045 (0.0508)
	2	0.7413	0.6983 (0.0430)	0.6850 (0.0563)	0.7316 (0.0097)
	3	0.7521	0.7060 (0.0461)	0.6975 (0.046)	0.7317 (0.0204)
	4	0.7630	0.7160 (0.0470)	0.7467 (0.0163)	0.7443 (0.0187)
	5	0.7390	0.7242 (0.0148)	0.7576 (0.0186)	0.7402 (0.0012)

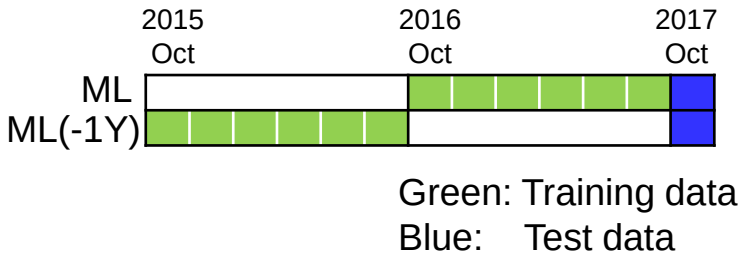
Up : Prediction
Dn : Deviation

Yellow cell indicates the nearest prediction value from actual value

3. Results (2/3)

- .jp domain names -

- We also calculated the value of renewal rate for ML by using further one year earlier training data (‘-1Y’: there is one year gap between training and test data, see below)
- Still, ML(-1Y) is the best for 4 times out of 8 times



Green: Training data
Blue: Test data

Year	Month	Actual value	Predicted value of renewal rate		
			Least Sq.	Mov. Ave.	ML(-1Y)
2017	10	0.7453	0.7497 (0.0044)	0.7516 (0.0063)	0.7532 (0.0079)
	11	0.7273	0.7458 (0.0185)	0.7509 (0.0236)	0.7486 (0.0213)
	12	0.7162	0.7377 (0.0215)	0.7363 (0.0201)	0.7476 (0.0314)
2018	1	0.6537	0.7267 (0.0730)	0.7217 (0.0680)	0.7286 (0.0749)
	2	0.7413	0.6983 (0.0430)	0.6850 (0.0563)	0.7488 (0.0075)
	3	0.7521	0.7060 (0.0461)	0.6975 (0.046)	0.7459 (0.0062)
	4	0.7630	0.7160 (0.0470)	0.7467 (0.0163)	0.7543 (0.0087)
	5	0.7390	0.7242 (0.0148)	0.7576 (0.0186)	0.7500 (0.011)

Up : Prediction
Dn : Deviation

Yellow cell indicates the nearest prediction value from actual value

3. Results (3/3)
- gTLDs -

- We applied one year gap (‘-1Y’) ML method to gTLD (.com, .net, .info, .biz, *etc*.) domain names† registered through JPRS (as registrar)
- ML(-1Y) is the best for 6 times out of 8 times

†: ASCII Domain Name only

Year	Month	Actual value	Predicted value of renewal rate		
			Least Sq.	Mov. Ave.	ML(-1Y)
2017	10	0.8024	0.7739 (0.0285)	0.7903 (0.0121)	0.7916 (0.0108)
	11	0.7857	0.7816 (0.0041)	0.7895 (0.0038)	0.7918 (0.0061)
	12	0.8033	0.7862 (0.0171)	0.7941 (0.0092)	0.8023 (0.0010)
2018	1	0.7849	0.7954 (0.0105)	0.7945 (0.0096)	0.7994 (0.0145)
	2	0.8117	0.7965 (0.0152)	0.7941 (0.0176)	0.8124 (0.0007)
	3	0.8137	0.8037 (0.0100)	0.7983 (0.0154)	0.8079 (0.0058)
	4	0.7973	0.8142 (0.0169)	0.8127 (0.0154)	0.8061 (0.0088)
	5	0.7911	0.8146 (0.0235)	0.8055 (0.0144)	0.8029 (0.0118)

Up: Prediction
Dn: Deviation

Yellow cell indicates the nearest prediction value from actual value

4. Findings (1/2)

- Features contributing to renewal rate -

- We found existence of ‘Digit’ in a label significantly affects to down side (expectedly)
- We found existence of ‘Hyphen’ in a label affects to up side (unexpectedly)
- We found outlying value of vowel and consonants ratio in a label affects to down side (expectedly)

4. Findings (2/2)

- Advantages of prediction by machine learning -

- We confirmed that our proposed method improves precision of prediction compared to statistical methods
- Our proposed method was not affected when temporal fluctuation of renewal rate happened
  - Statistical method keeps deviation of predicted renewal rate value from actual value for a certain period when temporal fluctuation happened
- Our proposed method predicts renewal rate for each domain name individually
  - Which statistical method cannot
- Our proposed method may have ability to predict next year's renewal rate immediately when a domain name registered
  - Statistical method is not appropriate to predict renewal rate with one year gapped data

5. Discussions (1/3)

- Ongoing and future works for improvement -

(1) Prediction in much earlier training data ← Ongoing (P16-17)

– In this work, we predict targeted month just before its renewal period, but we'd like to predict it much earlier with keeping preciseness

• By improving selection of training data period, algorithm of analysis, etc.

(2) Correspondence to the trend changes (fluctuation)

– Look for another features to detect trend of string style which brings fluctuation of renewal rate

• Preliminary trying N-Gram (N=2, 3) analysis

(3) Applicability to others ← Ongoing (P17)

– Candidates are domain registration industries (registries, registrars) who have comprehensive administrating domain names, and researchers

5. Discussions (2/3)

- Ongoing and future works for improvement -

(4) Impacts from information other than domain name string

- Candidates are information that can obtain immediately after registration (ex. from which registrar), information that changes time-to-time (ex. DNS host information, Web site existence)
- Define new features derived from above and consider impact to the preciseness of prediction

(5) Prediction of renewal rate of 2nd year and later

- Confirm the applicability of proposed method to the prediction of renewal rate of 2nd year and later
  - Empirically, it is known that renewal rate of renewed domain names increases according to its renewal times, so renewal times is also new features' candidate

5. Discussions (3/3)

- Ongoing and future works for improvemet -

(6) Vocabulary analysis of domain name string

- Perform analysis of words and terms in string that we did not do in this work
  - Using English dictionaries, Japanese (Romanized) dictionaries, etc.

(7) Corresponding to IDNs

- Consider corresponding to Internationalized Domain Names (IDNs) which enables using non-ASCII characters
  - IDNs are deploying world widely
  - In IDNs, characteristics based on languages appears much more

Q&A

Any ideas to improve our method?
All of your questions, comments and suggestions are highly appreciated!
We will continue this work according to this discussions and will publish evaluation tool (contact me for alpha)