テキストマイニングによる不満から読み取る鉄道に関する問題
ベイズ学習による投稿者の不明な属性の推定

王 帥・三道弘明　関西学院大学

背景・目的

- データマイニング技術がビジネスに活用されつつある
- ベイズ学習の適用可能性

- テキストマイニングの研究結果に基づいて社会に助言
- ベイズ学習による不明な情報を推測できるかどうかの検証

データの分類

トイレ
トイレ ＋ 汚い
女子トイレ＋ 混雑

施設
ゴミ箱 ＋ 臭い
階段 ＋ 長い

人・マナー
荷物 ＋ 邪魔
席 ＋ 譲らない

電車施設
エアコン＋ 寒い
コンセント＋ ない

ダイヤ
電車本数＋ 少ない
終電 ＋ 早い

料金
料金 ＋ 高い
運賃 ＋ 下げる

コーディングルール

- 共通する概念を一つのキーワードに集約

- 施設：駅構内のごみ箱が臭いやホームまでの階段が長いなど駅に対する不満
- ダイヤ：終電が早いや　電車の本数が少ないなどダイヤに対する不満

*トイレ
（ トイレ | 手洗い | 便所 | おしっこ | 小便 ）&（ 汚い | 古い | ゴミ | 臭い | におい | 暗い | 増やす ）| 洗面所

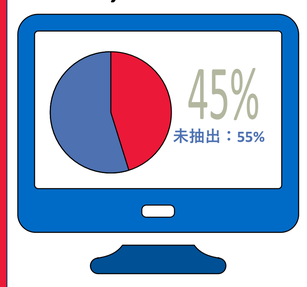
*施設
（ 階段 &（ 長い | 狭い | 混雑 | 多い | 急 | つらい | 辛い | 危ない ））|（ ベビーカー | 車椅子 & 不便 | 専用 ）|（ エレベーター &（ 狭い | 混雑 | 混む | 小さい |（ 広い & 欲しい | ほしい ）| 臭い ））|（ エスカレーター &（ 遅い | 狭い ））|（ 場所 & 座る & ない ）| 困る &（ 駅 & 寒い | 熱い | 冷房 | 暖房 ）| 電波 &（ 駅 | ホーム ）

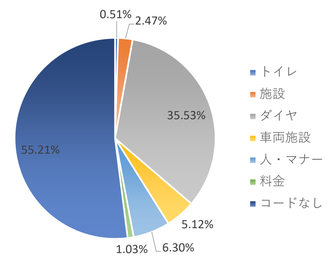
*ダイヤ
（（ 朝 | 帰宅 | 電車 | 通勤 ）& ラッシュ ）| 乗り換え & 時間 |（（ 本数 | 電車 | 車両 ）&（ 少ない | 増やす | 欲しい | ほしい | 混む ））|（（ 特急 | 急行 | 快速 | 列車 ）& 増やす | 少ない ）|（ 終電 & 早い | 遅い | ない ）| 子連れ & 車両 | 電光 | 待ち時間 | 振替 & 輸送 | スピード &（ 遅い | 電車 | 上げる ）| 職員

*車両施設
（ 電車 & 寒い | 熱い | 冷房 | 暖房 | 空調 | クーラー | 臭い | 匂い | におい | ニオイ | 空気 ）| クーラー &（ 効く | 寒い | 暑い ）|（（ 電車 | 車内 ）&（ 狭い | 寒い | 熱い | 広い &（ 欲しい | ほしい ））| 手すり | 吊革 | 足元 &（ 暖房 | ヒーター | 熱い | 暖かい ）| 隙間 &（ ホーム | 広い | 危ない | 怖い ）| 電波 &! （ 駅 | ホーム ）| 揺れ &（ 激しい | 電車 | 大きい | 酔う | 酷い ）| 網棚 &（ 高い | 邪魔 | 狭い | 位置 ）| コンセント &（ 窓側 | 座席 | 席 | 新幹線 | 車両 | 窓際 ）| ガラス &（ 割れる | 曇る | 汚い | 汚れる ）

集計結果

- 駅・電車サブカテゴリ
  186,595レコードから
- 83,576レコードが抽出され
- 83,576レコードが6つに帰属





ベイズによる投稿者不明な属性の推定

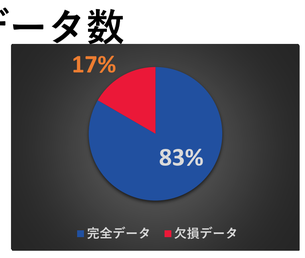
不明データサンプル

G	H	I	J	K	L	M	N	O
product	text	crated_at	gender	age	birth	job	prefecture	arital_stat
トイレ	ない駅のトイレ	42291.69	女	不明	不明	学生	大阪府	不明
。カビなのか、しばらく体		42291.92	女	不明	不明	上員（事務系	神奈川県	不明
記慮に加えて開く行為は公共		42292.14	不明	不明	不明	不明	不明	不明
微妙な遅延困る!乗換うまく		42292.07	不明	不明	不明	上員（事務系	大阪府	不明
自動改札機を持ってて左手		42292.03	男	不明	不明	上員（その他	不明	不明
みどりの窓口、どういうこと		42292.59	男	不明	不明	上員（技術系	兵庫県	不明
電車	とかつらすぎ。系	42292.04	不明	不明	不明	不明	不明	不明

データの整理

- 不明な属性（外部変数）
  性別・結婚歴・年収・年齢・職業・地域..etc..

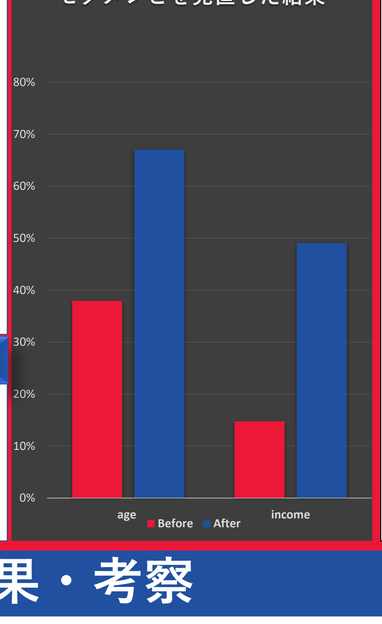
- 不明な属性のない　データ数
- 91399→76201
- 約2割弱が「欠損データ」



セグメントの見直し

	年齢	年収	
青少年	10	100	低
	20	200	
中年	30	300	
	40	400	中-高
中高年	50	500	
	60	700	
	70	800	
		900	高-超高
		1000	
		1200	
		1500	
		2000	

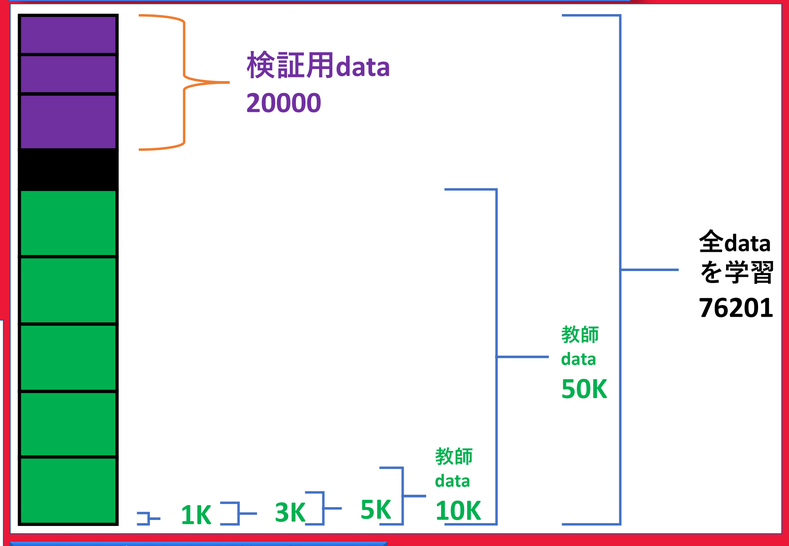
セグメンとを見直した結果



全データの推測

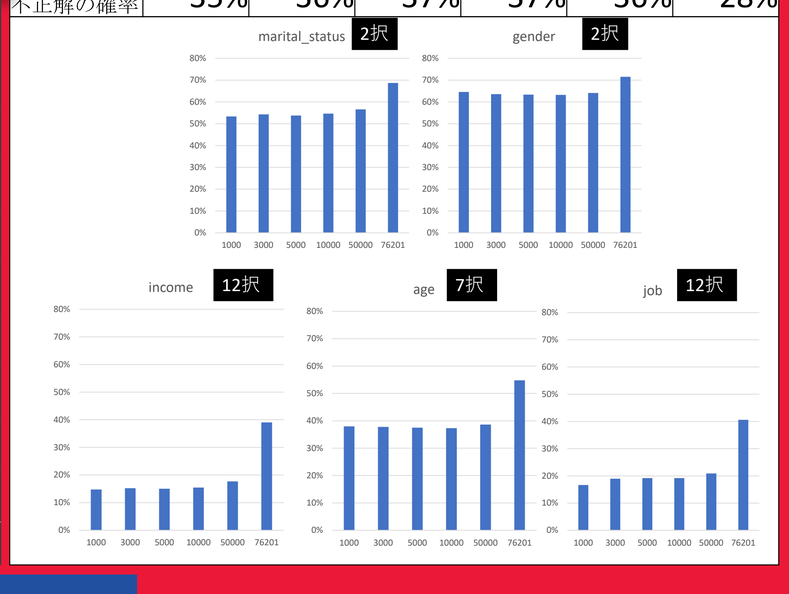
- 1,000件だけ学習
- 91,399の不明を含むdata不明の部分を推測
- クロス集計,対応分析

教師データと検証データの用意



データの集計

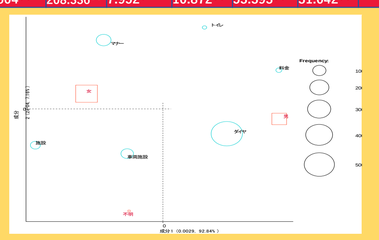
回答×正解 / 件数	1000	3000	5000	10000	50000	76201
男×男	454	924	1200	1551	1970	3160
女×女	12464	11794	11467	11091	10849	11149
女×男	6391	5921	5645	5294	4875	3685
男×女	691	1361	1688	2064	2306	2006
正解の確率	65%	64%	63%	63%	64%	72%
不正解の確率	35%	36%	37%	37%	36%	28%



外部変数による不明を推測する前後の変化

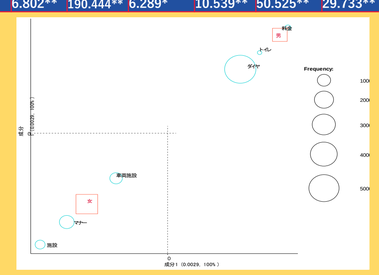
Before,gender

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
女	666 (0.50%)	4985 (3.73%)	48673 (36.38%)	7116 (5.32%)	9013 (6.74%)	1297 (0.97%)	133796
男	307 (0.59%)	1265 (2.45%)	18456 (35.76%)	2552 (4.94%)	3002 (5.82%)	646 (1.25%)	51616
不明	2 (0.17%)	70 (5.92%)	450 (38.04%)	66 (5.58%)	67 (5.66%)	7 (0.59%)	1183
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12082 (6.47%)	1950 (1.05%)	186595
カイ2乗値	9.604**	208.336**	7.952*	10.872**	53.393**	31.042**	



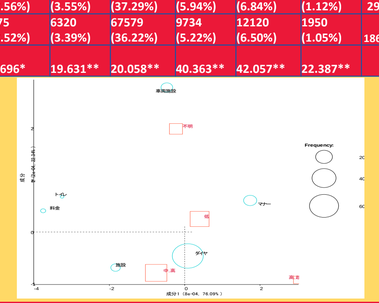
After,gender

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
女	668 (0.50%)	5052 (3.75%)	49088 (36.39%)	7177 (5.32%)	9101 (6.75%)	1302 (0.97%)	134894
男	307 (0.59%)	1268 (2.45%)	18491 (35.77%)	2557 (4.95%)	3019 (5.84%)	648 (1.25%)	51701
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12120 (6.50%)	1950 (1.05%)	186595
カイ2乗値	6.802**	190.444**	6.289*	10.539**	50.525**	29.733**	



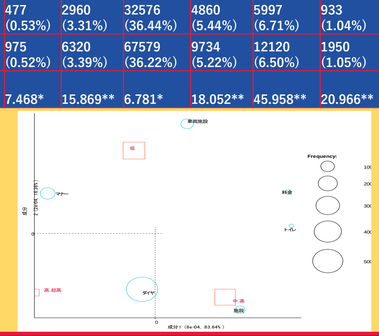
Before, marital_status

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
既婚	429 (0.49%)	3662 (4.16%)	31509 (35.78%)	4039 (4.59%)	5340 (6.06%)	928 (1.05%)	88061
不明	43 (0.37%)	528 (4.60%)	4262 (37.11%)	737 (6.42%)	809 (7.04%)	114 (0.99%)	11485
未婚	503 (0.58%)	2130 (2.45%)	31808 (36.54%)	4958 (5.70%)	5933 (6.82%)	908 (1.04%)	87049
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12082 (6.47%)	1950 (1.05%)	186595
カイ2乗値	12.090**	446.687**	15.149**	144.563**	47.392**	0.374	



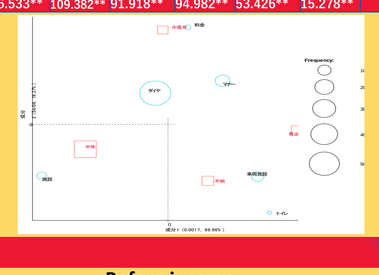
After, marital_status

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
既婚	446 (0.46%)	4162 (4.30%)	34846 (36.02%)	4467 (4.62%)	5987 (6.19%)	1014 (1.05%)	96741
未婚	529 (0.59%)	2158 (2.40%)	32733 (36.43%)	5267 (5.86%)	6133 (6.83%)	936 (1.04%)	89854
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12120 (6.50%)	1950 (1.05%)	18659 5
カイ2乗値	14.372**	513.637**	3.359	145.609**	31.002**	0.013	



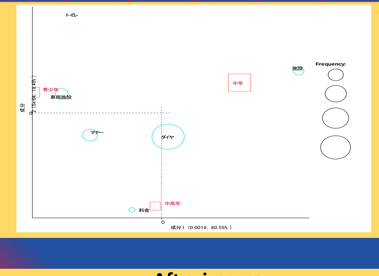
Before,age

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
青少年	217 (0.73%)	806 (2.70%)	11043 (37.02%)	1722 (5.77%)	2163 (7.25%)	345 (1.16%)	29826
中高年	73 (0.32%)	619 (2.68%)	7791 (33.73%)	1043 (4.51%)	1384 (5.99%)	274 (1.19%)	23101
中年	521 (0.49%)	3878 (3.68%)	38141 (36.18%)	5254 (4.98%)	6599 (6.26%)	1019 (0.97%)	105420
不明	164 (0.58%)	1017 (3.60%)	10604 (37.54%)	1715 (6.07%)	1936 (6.85%)	312 (1.10%)	28248
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12082 (6.47%)	1950 (1.05%)	186595
カイ2乗値	46.533**	109.382**	91.918**	94.982**	53.426**	15.278**	



After,age

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
青少年	218 (0.70%)	830 (2.67%)	11405 (36.68%)	1770 (5.69%)	2225 (7.16%)	375 (1.21%)	31097
中高年	74 (0.33%)	603 (2.68%)	7579 (33.74%)	1036 (4.61%)	1347 (6.00%)	261 (1.16%)	22465
中年	683 (0.51%)	4887 (3.67%)	48595 (36.53%)	6928 (5.21%)	8548 (6.43%)	1314 (0.99%)	133033
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12120 (6.50%)	1950 (1.05%)	186595
カイ2乗値	35.396**	116.269**	68.236**	30.856**	32.575**	14.969**	



Before,income

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
高,超高	52 (0.36%)	422 (2.96%)	5237 (36.67%)	688 (4.82%)	1055 (7.39%)	98 (0.69%)	14280
中,高	440 (0.55%)	2841 (3.52%)	28993 (35.94%)	4068 (5.04%)	4953 (6.14%)	891 (1.10%)	80666
低	319 (0.51%)	2018 (3.23%)	22449 (35.97%)	3242 (5.19%)	4113 (6.59%)	635 (1.02%)	62415
不明	164 (0.56%)	1039 (3.55%)	10900 (37.29%)	1736 (5.94%)	1999 (6.84%)	326 (1.12%)	29234
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12120 (6.50%)	1950 (1.05%)	186595
カイ2乗値	8.696*	19.631**	20.058**	40.363**	42.057**	22.387**	

After,income

	*トイレ	*施設	*ダイヤ	*車両施設	*マナー	*料金	ケース数
高,超高	52 (0.36%)	422 (2.96%)	5235 (36.67%)	688 (4.82%)	1056 (7.40%)	98 (0.69%)	14277
中,高	446 (0.54%)	2938 (3.54%)	29768 (35.90%)	4186 (5.05%)	5067 (6.11%)	919 (1.11%)	82920
低	477 (0.53%)	2960 (3.31%)	32576 (36.44%)	4860 (5.44%)	5997 (6.71%)	933 (1.04%)	89398
合計	975 (0.52%)	6320 (3.39%)	67579 (36.22%)	9734 (5.22%)	12120 (6.50%)	1950 (1.05%)	186595
カイ2乗値	7.468*	15.869**	6.781*	18.052**	45.958**	20.966**	

結果・考察

- ダイヤに対する不満が圧倒的に多い
  女＞男,青少年＞中年＞中高年,
- 本研究においては、1000件のテキストデータと50000件のデータの予測結果は大した差がない
  1000件のテキストデータで外部変数の傾向がわかる
- 適切なセグメンテーションにより学習効果が向上
- ベイズ学習により推測後、有意でなくなるデータがある
  結婚歴×ダイヤ
- 欠損データを推定することによって、
  テキストマイニングの結果はより簡潔,明確になった

参考文献

- 小木 しのぶ,テキストマイニングの技術と動向,計算機統計学 ,28(1), 31-40, 2015
- 樋口 耕一,テキスト型データの計量的分析:―2つのアプローチの峻別と統合―,理論と方法, 19(1), 101-115, 2004.
- 奥村 学 ,ソーシャルメディアを対象としたテキストマイニング,電子情報通信学会 基礎・境界ソサイエティ, Fundamentals Review 6(4), 285-293, 2013.
- 打田 裕樹, Webユーザレビューにおける評価情報の時系列変化の可視化,知能と情報：日本知能情報ファジィ学会誌：journal of Japan Society for Fuzzy Theory and Intelligent Informatics 22(3), 377-389, 2010-09-30.
- 高野 敦子,因果関係に着目した口コミＷｅｂサイトからの評価表現抽出,人工知能学会論文誌 24(3), 322-332, 2009.
- 樋口 耕一,社会調査のための計量テキスト分析―内容分析の継承と発展を目指して,ナカニシヤ出版,2014.