2018年度　卒業論文

ランダムフォレストを用いた IMFにおける先進国決定要因の分析

慶應義塾大学　経済学部
尾髙智洋

要旨

本稿では、IMFが国家を先進国と分類する際どのような要素を重視しているのか、分析を行った。分析には、IMF、世界銀行、エコノミスト・インテリジェンス・ユニットが発表した、2017年の各国の経済、健康、インフラ、科学技術、教育データ、2016年の政治データを用い、ランダムフォレストによってモデル化し、どのような要素を重視しているのか分析し、可視化した。

1. はじめに

昨今、世界では度々AIによってどんな職業が置き換えられるのか、また、AIは人間を超えるのかといった議論が見られる。これからの将来どうなっていくかは分からないが、少なくともAIによって我々の生活や仕事が大きく変わっていくことは間違いないだろう。しかし、AIとは明確に定義されていない。これに対し、松尾(2014、P45)では、AIとは「人工的に作られた人間のような知能、ないしはそれを作る技術」としている。この定義もとても幅広く、「ルール自体を自動で考えて作業を行ってくれるもの」など様々なものがAIに分類される。

本稿では、このAIの初歩的な手法である、機械学習の一種のランダムフォレストを用いた分析の実証を行う。具体的には、ランダムフォレストを用いて、IMFが国家を先進国と分類する際の、要素の重要性を可視化する。先進国には明確な定義がなく曖昧となっている。その結果、OECDやIMFなど様々な国際機関で先進国と指定される国が異なっている。この原因として、何処の機関もどのような要素を重視して先進国と定めているのか公表していないためだ。よって、IMFがどのような要素を重視し国家を先進国として定めているのか分析を行った。

2. データ

具体的な分析について記述する前に、本稿で用いるデータについて記述する。以下で記述するデータは全て2017年のデータである。本稿で用いるデータは、以下の7項目である。先進国か否か、一人当たりGDP、平均寿命、電気普及率、発表論文数、人間開発指数、民主主義指数。

まず、IMFより経済の指標として、一人当たりのGDP。また、公開されている先進国リストを得た。次に、世界銀行より、健康の指標として、平均寿命。また、インフラの指標として電気普及率。また、科学力の指標として、発表論文数。また、教育の指数として、人間開発指数を得た。ただ、発表論文数に関しては、香港のデータが欠損している。これは、中国の発表数の中に香港のデータも含まれているためである。そのため、中国のデータを、中国と香港の人口比で分け、欠損値を補完した。最後に、政治の指標として、エコノミスト・インテリジェンス・ユニットが発表している民主主義指数を得た。これは、選挙の過程と多様性、

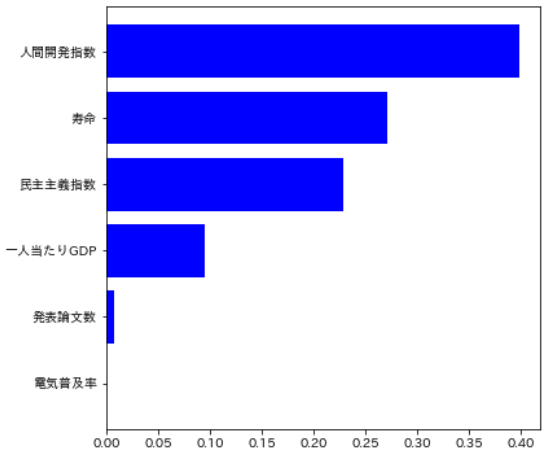
政府機能、政治参加、政治文化、市民自由度のそれぞれを10.00満点で評価し、平均を取った指標である。

分析対象は以上の7項目で欠損値のなかった161か国である。先進国で欠損値があったのは、台湾である。台湾は国連に加盟できていないため、世界銀行の調査対象から外されている。この161か国の約7割である、112か国を学習データにし、約3割である49か国をテストデータとした。分け方は、完全なランダムである。付録1 に今回の分析で用いた国々の表を載せておく。

3. 結果と考察

まず、分類の精度はテストデータにおいて100%であった。このことから、このモデルによってすべての国を先進国と途上国にほぼ間違えることなく分けることが出来る。次に、以下の図1に変数重要度の可視化結果を示す。

図1. 変数重要度



この図の横軸が変数の重要度を示しており、全ての変数の重要度の和は1となる。

この結果を見ると、科学技術力とインフラはあまり先進国の決定に寄与していないことがわかる。特に最も重要度が高いのが教育であり、二番手が平均寿命である。そして、経済の指標である一人当たりGDPは四番手となっている。これは、教育や健康といった指標のほうが経済よりも人々の生活をより具体的に表し、生活の豊かさを示しているためであると考えられる。また、教育や健康の指標にはある程度、経済の情報も含まれているだろうことも要因であると考えられる。それは、お金に余裕がなければ、健康も教育も維持できないと考えられるためである。また、三番手の重要度となっているのが民主主義指数である。この指標は政治について強く示されている。例えば、中国のように共産党による一党独裁の場合や、サウジアラビアのように絶対君主制の場合選挙の過程と多様性が0点となっている。また、ロシアといった一応選挙が開かれるものの、過程などに不可解な点がある場合も2点台が設定されている。そして、総じてそういう国家は市民自由度が低い。そういった国家は、市民の意見などを抑圧しているため、人権部分に大きな懸念があるため先進国に選定されていないのだろう。実際に経済の指標である1人当たりGDPが高いものの、カタール、クウェート、アラブ首長国連邦、サウジアラビアなどの国は先進国に選ばれていない。これらの国は表1

にまとめてある。IMFが政治関連について点数付けなどを行っているのかは不明瞭であるが、経済よりも重視しているのはデータを見てみると、納得できる。

表1. 日本と一人当たりGDPが高いが先進国でない国との比較

国名	一人当たりGDP(U.S.dollars)	民主主義指数
日本	38448.57	7.99
カタール	61024.77	3.18
アラブ首長国連邦	37732,66	2.75
クウェート	27393.91	3.85
サウジアラビア	21096.44	1.93

最後に、経済の指標である一人当たりGDPが四番手となっている。今回の分析を行う前は最も先進国の決定に作用すると考えていたが、今はあくまで最低限しか見られていないのだと考えている。

今回、学習データで精度が100%であったのはもちろんであるが、テストデータにおいても精度が100%であった。このことから、過学習を起こしていないであろうと考えられるため、学習データの分類も正常に成されている。つまり、今回の変数で先進国を分類できることを示しているが、それと同時に矛盾したデータが存在しないということも示している。昔から先進国にである国々は、昔から先進国であるからという理由ではなく、しっかりとした理由があって今現在においても先進国に分類されているのだということがわかる。

4. まとめ

今回の分析から、IMFがその国家を先進国として分類する際に、教育、健康、政治、経済を重視していることが分かった。これと同様の分析をOECDなどほかの機関でも行うことで、それぞれの機関が先進国と分類する際にどのような要素を重視しているのか、また、重要度はどのくらい違うのかの比較なども行うことが出来るだろう。また、今回の分析はランダムフォレストによってどのような分析ができるのかについての一つの具体例として意義がある。

付録 1　分析対象国の分類

以下の表2に分析対象国およびそれらが先進国か否かの分類を載せておく。国名はIMFの表記である。

表2. 分析対象国家

	国名	先進国か否か		国名	先進国か否か
1	Afghanistan	0	9	Azerbaijan	0
2	Albania	0	10	Bahrain	0
3	Algeria	0	11	Bangladesh	0
4	Angola	0	12	Belarus	0
5	Argentina	0	13	Belgium	1
6	Armenia	0	14	Benin	0
7	Australia	1	15	Bhutan	0
8	Austria	1	16	Bolivia	0

(続)表1. 分析対象国

17	Bosnia and Herzegovina	0	57	Greece	1
18	Botswana	0	58	Guatemala	0
19	Brazil	0	59	Guinea	0
20	Bulgaria	0	60	Guinea-Bissau	0
21	Burkina Faso	0	61	Guyana	0
22	Burundi	0	62	Haiti	0
23	Cabo Verde	0	63	Honduras	0
24	Cambodia	0	64	Hong Kong SAR	1
25	Cameroon	0	65	Hungary	0
26	Canada	1	66	Iceland	1
27	Central African Republic	0	67	India	0
28	Chad	0	68	Indonesia	0
29	Chile	0	69	Iraq	0
30	China	0	70	Ireland	1
31	Colombia	0	71	Islamic Republic of Iran	0
32	Comoros	0	72	Israel	1
33	Costa Rica	0	73	Italy	1
34	Cote d'Ivoire	0	74	Jamaica	0
35	Croatia	0	75	Japan	1
36	Cyprus	1	76	Jordan	0
37	Czech Republic	0	77	Kazakhstan	0
38	Democratic Republic of the Congo	0	78	Kenya	0
39	Denmark	1	79	Korea	1
40	Djibouti	0	80	Kuwait	0
41	Dominican Republic	0	81	Kyrgyz Republic	0
42	Ecuador	0	82	Lao P.D.R.	0
43	Egypt	0	83	Latvia	0
44	El Salvador	0	84	Lebanon	0
45	Equatorial Guinea	0	85	Lesotho	0
46	Eritrea	0	86	Liberia	0
47	Estonia	0	87	Libya	0
48	Ethiopia	0	88	Luxembourg	1
49	Fiji	0	89	Madagascar	0
50	Finland	1	90	Malawi	0
51	FYR Macedonia	0	91	Malaysia	0
52	France	1	92	Mali	0
53	Gabon	0	93	Malta	1
54	Georgia	0	94	Mauritania	0
55	Germany	1	95	Mauritius	0
56	Ghana	0	96	Mexico	0

(続)表1.　分析対象国

97	Moldova	0	130	Singapore	1
98	Mongolia	0	131	Slovak Republic	0
99	Montenegro	0	132	Slovenia	1
100	Morocco	0	133	South Africa	0
101	Mozambique	0	134	Spain	1
102	Myanmar	0	135	Sri Lanka	0
013	Namibia	0	136	Sudan	0
104	Nepal	0	137	Suriname	0
105	Netherlands	1	138	Sweden	1
106	New Zealand	1	139	Switzerland	1
107	Nicaragua	0	140	Tajikistan	0
108	Niger	0	141	Tanzania	0
109	Nigeria	0	142	Thailand	0
110	Norway	1	143	The Gambia	0
111	Oman	0	144	Timor-Leste	0
112	Pakistan	0	145	Togo	0
113	Panama	0	146	Trinidad and Tobago	0
114	Papua New Guinea	0	147	Tunisia	0
115	Paraguay	0	148	Turkey	0
116	Peru	0	149	Turkmenistan	0
117	Philippines	0	150	Uganda	0
118	Poland	0	151	Ukraine	0
119	Portugal	1	152	United Arab Emirates	0
120	Qatar	0	153	United Kingdom	1
121	Republic of Congo	0	154	United States	1
122	Romania	0	155	Uruguay	0
123	Russia	0	156	Uzbekistan	0
124	Rwanda	0	157	Venezuela	0
125	Lithuania	0	158	Vietnam	0
126	Saudi Arabia	0	159	Yemen	0
127	Senegal	0	160	Zambia	0
128	Serbia	0	161	Zimbabwe	0
129	Sierra Leone	0			

付録2 ランダムフォレストについて

ランダムフォレストとは、無数の決定木を作成し、それぞれの決定木で予測を行い、多数決によって値を予測する手法である。決定木とは、データを任意の変数の任意の値で分割することを繰り返し、似たデータ群に分ける手法である。

無数の決定木を作成する際には、学習データをすべて用いるのではなくその中からランダムにデータと変数を選択する。そのデータを利用して決定木を作成することで、それぞれの

決定木で使用するデータや変数が異なるようになっているため、多様性のある決定木の作成を可能にしている。この多様な決定木で多数決して予測することで、一つの決定木で予測するより精度の良い予測を目指している。これは、一人の人に中国は先進国かを尋ねるより、1000人にアンケートを取り多数決したほうが真の値に近い結果が得られる可能性が高いことと同じことである。

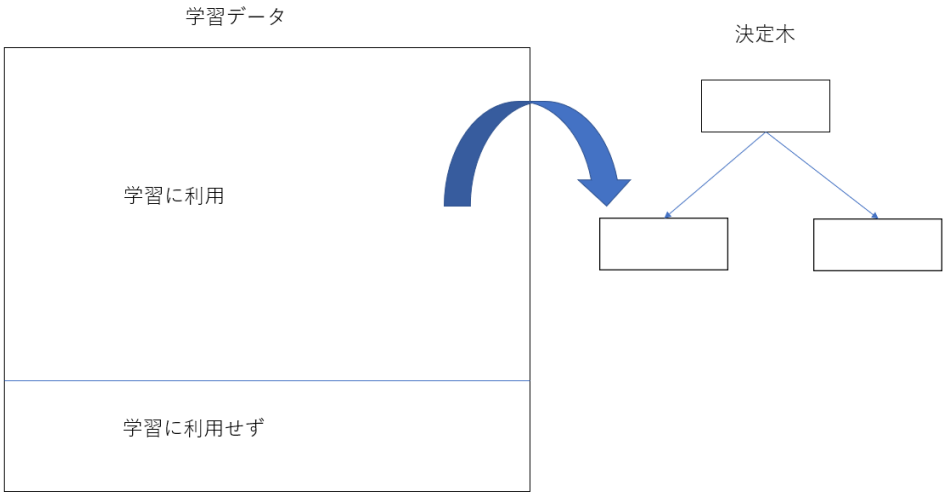


図2.　ランダムフォレスト

ランダムフォレストにおいて、変数の重要度を算出する際には、図2において、学習に利用しなかったデータを活用する。Aという変数の重要度を算出する場合、学習に利用しなかったデータを決定木に入れ、予測を行う。その際に予測を誤ったデータの割合を誤り率αとして算出する。次に学習に利用しなかったデータのA変数の値をランダムにシャッフルし、決定木に入れ、予測を行う。その際に予測を誤ったデータの割合を誤り率βとして算出する。

図3. 変数重要度の算出

学習に利用しなかったデータ

A	B	C
1	2	3
4	5	6
7	8	9

誤り率α

変数Aの値を
ランダムにシャッフル

決定木に当てはめ、予測

A	B	C
7	2	3
1	5	6
4	8	9

誤り率β

重要度は誤り率α-誤り率βで算出する。これを、全ての決定木で行い、重要度を平均する。これがその変数の重要度で、全ての変数で重要度を算出し、全ての変数の重要度の和が1になるように調整して算出される。
ランダムフォレストについてより詳しくは、Leo Breiman(2001)『Random Forests』を参照のこと。

参考文献

松尾豊(2014)『人工知能は人間を超えるのか』 KADOKAWA.
Leo Breiman(2001) 『Random Forests』 (Machine Learning, 45, P5-P32, 2001, Kluwer Academic Publishers.)

WEB
エコノミスト・インテリジェンス・ユニット『Democracy Index 2016』
URL: http://felipesahagun.es/wp-content/uploads/2017/01/Democracy-Index-2016.pdf