

ランダムフォレストを用いたインドにおける 政府からの支援を受けられる家庭の特徴推定

Estimating the characteristics of households

that can receive government support in India using Random Forest

新領域創成科学研究科 国際協力学専攻
47-196757 新里 智樹
指導教員：鈴木 綾 教授

キーワード：Random Forest, インド, COVID-19, 経済的支援, R, 機械学習法

1. はじめに

COVID-19 パンデミックによって、世界各国の政府は、コロナ感染者の拡大を防止するため都市封鎖などの政策をとっており、同時に経済的なダメージを緩和させるための政策を迅速に打ち出す必要に迫られている。今回、私は全土封鎖や直接現金給付金の支給などの政策を行っているインドに注目をした。

インドの財務相は2020年3月26日、緊急経済対策を発表し、対策の1つとして、国民ID「アダール」(日本のマイナンバーに相当)などのデジタル産業やデジタル公共インフラの強みを生かした直接現金給付を実施すると発表した。これはインド全土の貧しい農家を対象として、国民IDに紐づいた銀行口座に現金給付をするものである[1]。インドでは、仲介人の介在や人々が銀行口座を持たないことなどにより、肝心の貧困層に補助金が届かない問題や、大規模な不正受給が長年の課題だった。しかし、直接現金給付が導入されたことにより、執行コストが縮小され、スピーディーに膨大な数の国民に現金給付することが可能となった。しかし、それでも政府から給付金を受け取れていない人がいる。

そこで本研究では、インドの往生する出稼ぎ労働者や学生などの地元への帰還などが多い農村部である6つの州(Jharkhand, Rajasthan, Uttar Pradesh, Andhra Pradesh, Bihar, and Madhya Pradesh)のデータを使用して、世帯の誰かが6月に銀行口座のなどで政府からお金を受け取ったことがある家庭の特徴を分析し、経済的支援家庭の基準と今回の計算結果を比較する。そうすることで、インド政府が目指していた支援対象と現実の支援受給者の特徴の違いがあるかを調べる。違いがあれば、どのようなことがボトルネックになって支援が当初の目的通りに行き渡らなかったのかを考察する。

また、COVID-19のような緊急事態の場合に行政が支援対象先を決めるためには、既存のデータを用いて決める必要があるが、その手法として機械学習が有効かどうかを検証する。もし有効ならば、今後の政策にも反映できると考える。

2. インド政府の支援

インド政府の財務相による直接現金給付をする緊急経済対策の主な内容の一部が以下の通りである[1]。

- ・貧困層の女性2億人に1人当たり1,500ルピー支給
- ・全国農村雇用保証法対象者(MNREGA)の日給を182ルピーから202ルピーに引き上げ
- ・*MNREGAとは、The Mahatma Gandhi National Rural Employment Guarantee Actの略称で、地方部の貧困層をインフラ整備事業で雇用して職と収入を提供する貧困削減策
- ・貧困層の年配者、未亡人、障害者向け補助金増額
- ・建設関係労働者が福祉基金を利用できるよう各州政府に要請
- ・女性の自助組織(Self-help groups)への無担保融資上限額を100万ルピーから200万
- ・公立病院および医療センターでCOVID-19と戦う医療従事者のための保険制度

本経済的支援は、貧困層(特に女性)の家庭や医療従事者などの重要な職業の方々を対象にしている傾向がみられる。

3. 研究手法

3.1. 使用するデータセット

世界銀行によるコンピュータ支援電話インタビュー(CATI)によってインド農村部の6つの州で収集されたデータを使用する。これらのデータのサンプリングは、世界銀行、インド農村開発省、などが以前に実施した調査と影響評価から抽出された。全部で5004個のデータの中からrel_transfer_rec_r2のアンケート項目にYesかNoで回答しているデータ、計3867個を使用する。変数名:rel_transfer_rec_r2は、「6月にあなたまたはあなたの家族の誰かが、銀行口座、現金、小切手、または郵便局の口座のいずれかで政府からお金を受け取りましたか?」という設問である。[2]を参考に、学習に使用するトレーニングデータを9割の3480個、性能の評価に用いるテストデータを1割の387個に設定した。本研究で使用するデータの概要は、以下の通りである。

- 1.基本データ：州、年齢、宗教、屋根材、世帯主の性別、教育レベル、世帯数など
- 2.収入と消費：賃金率、雇用期間、消費支出、必需品の価格、食料安全保障の状況などの変化
- 3.移住：移住率、移民の収入と雇用状況、帰国移民計画など
- 4.救済へのアクセス：現物での現金および労働救済へのアクセス、受け取った救済の量、および救済へのアクセスの制約
- 5.健康：医療施設へのアクセスと過去の医療の割合、COVID-19関連の症状と保護行動に関する知識

3.2. 研究手法

今回使用するデータは、全ての家庭が全ての質問に対して答えることがほとんどなかった。つまり、欠損値が多い。また、多種多様な質問があるため、説明変数が136個と多い。多数の説明変数から重要な変数、つまり、経済的支援を受け取ることができた人の特徴を知りたい。そこで分析方法として、機械学習法のRandom Forestに注目した。Random Forestは、学習用のデータをランダムにサンプリングして多数の決定木を作成し、作成した決定木をもとに多数決で結果を決める方法で、精度、汎用性が高く扱いやすい[3]。さらに、以下のような特長がある。

- ・説明変数が数百、数千でも効率的に作動
- ・目的変数に対する説明変数の重要度を推定
- ・欠損値を持つデータでも有効に動作
- ・個体数がアンバランスでもエラーバランスが保たれる

以上の点から、Random Forestを使用することとする。実際に本研究のようなケースにおいて、Random Forestが特徴分析や予測する能力に優れた機械学習法であると研究されている[4]。

3.3. 計算手順

ルピーへ引き上げ計算手順は、以下の通りである。

1. Random Forestのチューニングをする。その際に、構築する決定木を500個に指定する。

そして、選択する説明変数の個数の初期値がいくつの時にOut-of-Bag 誤差(OOB error)が最も少なくなるかを割り当てる。

2. 1 で求めた説明変数の数で Random Forest を実行する。
3. 重要度(importance)を測る。重要度とは「その特徴量の分割がターゲットの分類にどれくらい寄与しているかを測る指標」である。
4. テストデータを使って、トレーニングデータで学習した Random Forest で予測し、実測値と比べる

4. 結果・考察

4.1. 結果

1. Random Forest のチューニングをした結果、図 1 の通り、説明変数が 6 個の時に Out-of-Bag 誤差(OOB error)が 38.02%と、最も少なくなった。つまり、正判率は 61.98%である。説明変数が 11 個の時、OOB error は 38.56%、22 個の時、OOB error は 39.63% と

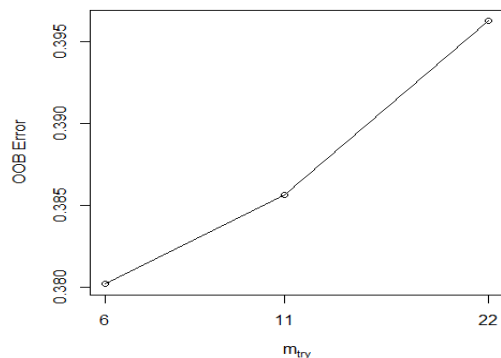


図 1：Out-of-Bag 誤差(OOB error)と説明変数(mtry)の数の相関図

2. 1 で求めた説明変数の数である 6 個で Random Forest を実行した。全体の OOB error : 38.82%というのが今回作成した予測モデルから予測した値が答え合わせの結果、間違っていた割合を示している。逆に言えば、今回の予測モデルの精度は 61.18%ということになる。つまり、新規のデータが手に入った場合、経済的支援を受けられるかどうかを 61.18%の精度で予測できるということになる。
3. 重要度(importance)を測った結果を表 1 に示す(説明変数が多いため、表 1 は Mean Decrease Gini が高い上位 6 個の説明変数のみ示す。Mean Decrease Gini(ジニ不純度)の値が大きいほど、影響度が大きいということを意味する。今回の例では demo_age が 61.179 で最も影響が大きい要素であることが分かる。

表 1：各説明変数の Mean Decrease Gini (ジニ不純度)

	Mean Decrease Gini	説明変数名
1	61.18	demo_age
2	51.21	con_lckdwn_r2
3	51.16	con_wk_r2
4	46.80	demo_hh_size
5	43.23	rel_mnrega_avg_r2
6	36.85	rel_mnrega_r2

4. テストデータを使って、トレーニングデータで学習した Random Forest で予測し、実測値と比べた結果、rel_transfer_rec_r2 を Yes と予測した時の判別率は 81.6%と高水準なのにに対し、No と予測した時の判別率は 41.4%と低水準だと分かった。

4.2. 考察

表 1 の結果から、基本データの中では年齢(demo_age)や世帯数(demo_hh_size)、州(state)が、政府の支援受給を予測する上で重要な説明変数であることが分かった。そして、1.1.研究背景で示している通り、今回の経済的支援は、貧困層の年配者、未亡人、障害者向けに補助金が増額されているので、インドの高齢者の定義

である 60 歳以上の年齢の方は受けとりやすいと考えられる。年齢を政府からの資金受給の有無で分けてみたところ、経済的支援を受け取っていない家庭の方が、高齢者の割合が 0.5%と僅かだが高いことが分かった。さらに、demo_age の年齢は、アンケートに答えた方の年齢なので、60 歳未満の方が回答していても、60 歳以上の方が家庭にいる可能性がある。

また、「封鎖中、あなたの世帯は特定の月にいくらのお金を使いましたか?(con_lckdwn_r2)」、「あなたの世帯は過去 7 日間にいくらのお金を使いましたか?(con_wk_r2)」が影響の出る説明変数と判断されたことから、支出が多い・少ないが経済的支援を受けられることに影響していると考えられる。

さらに、「6 月のあなたの地域の MNREGA の 1 日あたりの平均賃金はいくらでしたか?(rel_mnrega_avg_r2)」、「6 月に MNREGA の作業に関してあなたに当てはまるのは次のうちどれですか?(rel_mnrega_r2)」という説明変数も Random Forest を使用することで重要だと分かった。「6 月に、MNREGA の作業に関してあなたに当てはまるのは次のうちどれですか?(rel_mnrega_r2)」の問いに対する回答者の答えを、政府からの資金受給の有無で分けてみた。その結果、経済的支援を受けている家庭は「試しましたが、どの日も仕事に就けませんでした」、経済的支援を受けてない家庭は「(MNREGA による事業に参加するために必要な) ジョブカードを持っていません」の回答率が最も高いことが分かった。ジョブカードを持っていない方の中には、サンプリング方法の性質上、MNREGA の仕事に就く必要がない位、収入に余裕がある方もいると考えられる。しかし、MNREGA は希望者に年 100 日の仕事を提供するとしているが、インフラ整備事業の遅滞から仕事が入らず、100 日の仕事を得ることができたのは参加世帯の 4 割にとどまっているので、収入に余裕がない方がジョブカードを持っていないケースが多いと考えられる。

そして、既存のデータで学習した Random Forest で経済的支援を受けられるかどうかをテストデータから予測し、実測値と比べた結果、「経済的支援を受けられる」と予測した時の判別率は 81.6%と高水準だったので、その時のインド政府の政策方針によるが、Random Forest で支援対象先をある程度決めることはできると判断する。

5. まとめ

本研究では、インドの往生する出稼ぎ労働者や学生などの地元への帰還などが多い農村部である 6 つの州(Jharkhand, Rajasthan, Uttar Pradesh, Andhra Pradesh, Bihar, and Madhya Pradesh)のデータを使用して、世帯の誰かが 6 月に銀行口座などで政府からお金を受け取ったことがある家庭の特徴を Random Forest を使って分析した。インド政府が示した経済的支援を受け取れる家庭の基準と今回の計算結果を比較した結果、今回の経済的支援は貧困層の年配者の補助金が増額されるにも関わらず、経済的支援を受け取っていない家庭の方が、高齢者の割合が 0.5%と僅かだが高いことが分かった。また、経済的支援を受け取っていない家庭は、MNREGA にも参加する資格がない可能性が高いことが分かった。また、既存のデータ(直近のセンサスなど)で学習した Random Forest で経済的支援を受けられるかどうかを予測し、実測値と比べた結果、「経済的支援を受けられる」と予測した時の判別率は 81.6%と高水準だったので、その時のインド政府の政策方針によるが、Random Forest で支援対象先をある程度決めることはできると判断する。

5. 参考文献

- [1] PIB Delhi, "Finance Minister announces Rs 1.70 Lakh Crore relief package under Pradhan Mantri Garib Kalyan Yojana for the poor to help them fight the battle against Corona Virus", Press Information Bureau Government of India, 2020/3/26
- [2] Taro Sawaki, Takuya Tanaka, and Ryosuke Kasahara, "Credit Scoring for SMEs Using Machine Learning Techniques", 人工知能学会研究会資料, SIG-FIN-019
- [3] LEO BREIMAN 『Random Forests』 Machine Learning, 45(1), 2001, 5-32
- [4] Kazusa Yoshimura, Nobuo Yoshida, "Machine Learning for monitoring the twin goals – Put falls and Solutions", World Bank, 2019/6/19