Высокая ошибка уравнения (4) в 2016 г. обусловлена наличием структурных изменений во В.р. В целом характеристики уравнений являются приемлемыми. Средняя относительная ошибка прогнозирования не превысила 5,07%, поэтому уравнения (1)–(6) и могут быть использованы при разработке прогнозов.

Литература

1. Савчишина, К.Е. Прогнозирование показателей налогово-бюджетной сферы в рамках квартальной макроэкономической модели QUMMIR / К.Е. Савчишина // Научные труды Ин-та народнохозяйственного прогнозирования РАН. – 2008. – Т. 6. – С. 225–241.

2. Борейко, Н.Н. Прогнозирование ВВП Республики Беларусь производственным методом на основе системы эконометрических моделей / Н.Н. Борейко, А.Ю. Селицкая, А.А. Никонович // Экономика, моделирование, прогнозирование: сб. науч. тр. – Минск: НИЭИ М-ва экономики Респ. Беларусь. – 2018. – Вып. 12. – С. 248–259.

ПРИМЕНЕНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ И АЛГОРИТМА СЛУЧАЙНОГО ЛЕСА ДЛЯ КЛАССИФИКАЦИИ АДМИНИСТРАТИВНО-ТЕРРИТОРИАЛЬНЫХ ЕДИНИЦ ПО ПРИЗНАКУ ОТНЕСЕНИЯ К ОТСТАЮЩИМ

Драгун Н.П.,
кандидат экономических наук, доцент,
Гнедько Н.Ю.,
НИЭИ Министерства экономики Республики Беларусь, г. Минск

В контексте низкоразмерных данных (когда число признаков по сравнению с размером выборки невелико) логистическая регрессия считается классическим инструментом бинарной классификации. Область применения этого метода в экономике достаточно широка: от научных работ, где использование логистической регрессии носит самостоятельный характер [1], до исследований, включающих метод в качестве вспомогательной модели [2]. Распространенность данного подхода среди эконометристов объясняется интуитивной схожестью с моделью множественной регрессии, а также наличием встроенных инструментов для построения логистической регрессии в основных статистических программных пакетах.

Альтернативным методом классификации низкоразмерных данных считаются методы машинного обучения, в частности, алгоритм случайного леса. С момента своего изобретения данный метод [3], в котором основное внимание уделяется прогнозированию, а не объяснению, приобретает все большую популярность и все чаще становится общепринятым стандартным инструментом, который также используется учеными, не имеющими достаточного опыта в статистике или машинном обучении. В настоящее время экономическое сообщество достаточно скептически относится к методам машинного обучения по причине отсутствия привычных критериев оценки полученных моделей и четкого понимания работы алгоритмов – зачастую методы машинного обучения представляют из себя «черный ящик». Однако появление новых эконометрических исследований в этой области [4] говорит о том, что случайные леса могут оказаться полезным инструментом в задачах прогнозирования и классификации.

В целях сравнения логистической регрессии и случайного леса была решена задача по классификации административно-территориальных единиц (районов и городов областного подчинения) базового уровня Республики Беларусь (всего 128 единиц) по признаку отнесения административно-территориальной единицы (АТЕ) к отстающим районам в соответствии с постановлением Совета Министров Республики Беларусь от 2 февраля 2019 г. № 74 «Об утверждении порядка отнесения административно-территориальных единиц к территориям, отстающим по уровню социально-экономического развития» [5].

Полученные результаты применения двух методов – логистической регрессии и случайного леса – для решения задачи классификации АТЕ базового уровня позволяют сделать следующие выводы.

1. Качество полученных на тестовой выборке прогнозов лучше у логистической модели (92,3% против 82,1% у случайного леса). С одной стороны, можно говорить о том, что имеет место переобучение случайного леса. С другой стороны, следует отметить, что случайный лес, как и любые методы машинного обучения, гораздо лучше работает при больших выборках, значительно превышающих количество АТЕ. Однако в целом точность прогнозирования достаточно высокая у обеих моделей. При этом у алгоритма случайного леса имеет место несколько большая устойчивость получаемых результатов при изменении состава обучающей и тестовой выборок, что подтверждают теоретические ожидания работы алгоритма.

2. Результаты применения двух методов с точки зрения определения наиболее значимых факторов отнесения АТЕ к отстающим по уровню социально-экономического развития в целом совпадают. Так, к факторам высокой значимости отнесены: объем производства продукции (работ, услуг) в расчете на душу населения в среднегодовом исчислении; объем производства продукции (работ, услуг) в сопоставимых ценах в расчете на одного списочного работника в среднем за год. При этом случайный лес предоставляет исследователю большие возможности по классификации факторов с точки зрения их значимости для прогнозирования, а также оценке ее относительной (других факторов) величины.

3. Интерпретируемость случайного леса позволяет визуализировать и исследовать характер влияния факторов на результаты классификации, что особенно важно тогда, когда это влияние носит сложный нелинейный характер (для рассматриваемой модели – это, например, влияние площади территории, расстояния от центра АТЕ до областного центра). Это дает возможность установить границы значений факторов, в рамках которых результирующая переменная относится к определенной категории. Полученные границы количественных значений факторов могут использоваться для решения широкого круга задач управленческого характера.

Проведенная апробация позволила установить следующие достоинства и недостатки исследуемых методов для решения достаточно часто встречающихся в управленческой практике задач, которые характеризуются небольшим размером выборочной совокупности, сочетанием количественных и категориальных переменных:

- логистическая регрессия обладает высокой точностью классификации, однако результат сильно зависит от выборки. Выбор признаков для построения логистической регрессии остается за исследователем, что, с одной стороны, позволяет сделать вывод о незначимости некоторых факторов, с другой – усложняет процесс подбора модели;
- алгоритм случайного леса обладает высокой устойчивостью по отношению к вариативности выборки, что обеспечивает хороший результат при проведении тестовых симуляций. Кроме задач классификации, случайный лес позволяет получить интерпретацию результатов благодаря мерам важности и графикам частичной зависимости. Данные инструменты полезны при определении наиболее важных предикторов, а также помогают установить границы значений того или иного признака, при которых объект будет классифицирован к определенной группе.

Литература

1. Матраева, Л.В. Использование логистической регрессии при выявлении приоритетов региональной инвестиционной политики в отношении иностранных инвесторов в регионы РФ / Матраева Л.В. // Статистика и Экономика. – 2013. – № 6. – С. 170–174.

2. Безбородова, А. Кредитный импульс и восстановление экономики: опыт Республики Беларусь / Безбородова А. // Банкаўскі веснік. – 2018. – № 4/657. – С. 10–19.

3. Breiman, L. Random forests / Breiman L. // MachineLearning. – 2001. – Vol. 45, № 1. – P. 5–32.

4. Денисов, Д.В. Применение метода случайных лесов для оценки резерва произошедших, но еще не заявленных убытков страховой компании / Денисов Д.В., Смирнова Д.К. // International Journal of Open Information Technologies. – 2016. – Vol. 4, № 7. – P. 45–51.

5. Об утверждении порядка отнесения административно-территориальных единиц к территориям, отстающим по уровню социально-экономического развития [Электронный ресурс]: постановление Совета Министров Респ. Беларусь, 2 февр. 2019 г., № 74 // Совет Министров Республики Беларусь. – Режим доступа: http://government.by/upload/docs/file3e1b0a660c08b501.PDF.