УДК 330.4

СОВРЕМЕННЫЕ ВЫЗОВЫ И МЕТОДЫ АНАЛИЗА ДАННЫХ

В. О. Сувалов

Аспирант экономического факультета
Белорусского государственного университета, г. Минск

Научный руководитель: **И. А. Карачун**

Кандидат экономических наук, доцент, заведующая кафедры цифровой экономики
экономического факультета Белорусского государственного университета, г. Минск

Работа посвящена современным вызовам науки об анализе данных, возникающих в ходе применения машинного обучения для целей прогнозирования. В частности, обсуждаются подходы к разделению массива данных на тренировочную, контрольную и проверочную выборки. Кроме того поднимается вопрос о возможности применения новых методов прогнозирования, которые позволяет использовать метод машинного обучения. Одним из таких методов является комбинирование прогнозов. Рассматриваются причины появления данного метода, даются рекомендации по его использованию.

Ключевые слова: машинное обучение; большие данные; анализ данных; тренировочная выборка; контрольная выборка; проверочная выборка; метод комбинированного прогноза.

MODERN DATA ANALYSIS` CHALLENGES AND METHODS

V. Suvalov

PhD Student of the Faculty of Economics of the Belarusian State University, Minsk

Supervisor: **I. Karachun**

PhD in Economics, Associate Professor, Head of Digital Economics Department
at the Faculty of Economics of the Belarusian State University, Minsk

This article is devoted to the modern challenges of the science of data analysis arising from the use of machine learning for forecasting purposes. In fairness, approaches to dividing the data set into training, control and test samples are discussed. In addition, question the about the possibility of applying new forecasting methods that can be used due to the machine learning method is raised. Combining forecast is one of these methods. The reasons for the appearance of this method are considered, recommendations for its use are given.

Keywords: machine learning; big data; data analysis; training sample; control sample; validation sample; combined forecasting method.

В условиях продолжающегося технического прогресса происходит закономерное углубление цифровой трансформации. Достижения последних лет в сборе, обработке и хранении данных привели к возможности накапливать и обрабатывать огромные массивов данных. Данный феномен получил название «большие данные» («big data»). Больше данные ставят перед исследователями новые вызовы. К примеру, поиск закономерностей в больших объемах информации вручную становится слишком трудоёмкой задачей. В этой связи все больше исследователей приходят к выводам о необходимости применения методов программной обработки и анализа данных, в том числе с применением методов машинного обучения.

Методы машинного обучения позволяют увеличить скорость обработки и анализа информации, но в тоже время их применение связанно с необходимостью нахождения ответов на новые вопросы. Следует понимать, что классический процесс построения модели, проиллюстрированный на рисунке 1, сохраняется и при применении методов машинного анализа данных. В то же время следует отметить, что в рамках машинного обучения также возможен отход от представленной схемы.

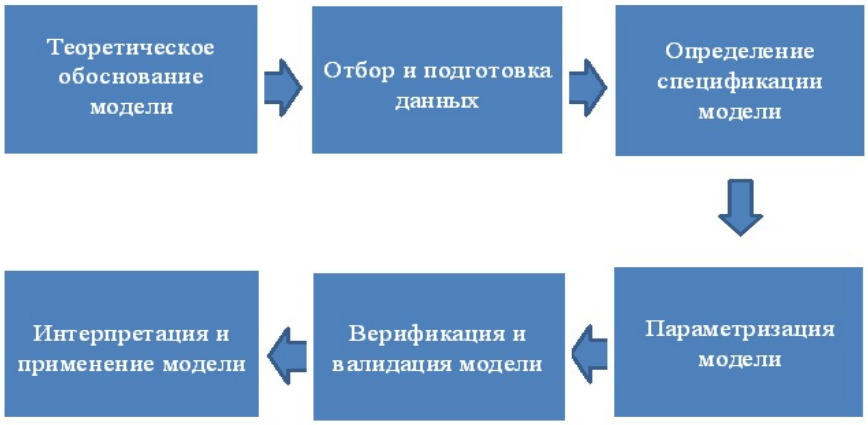

Рисунок 1 – **Порядок построения модели**

Примечание – Источник: составлено автором.

В случае применения методов машинного обучения теоретическое обоснование и выбор применяемого перечня моделей остаётся на исследователе. Следует понимать, что данные, применяемые при анализе, могут не содержать в себе причинно-следственные связи, но при этом обладать определенной степенью корреляции. Так, временные ряды данных могут быть связаны через промежуточный показатель, или коррелировать ввиду наличия зависимостей от другого показателя. Определить наличие теоретической взаимосвязи между данными и отсеять ложные корреляции – задачи исследователя. Кроме того, перед исследователями стоит задача определения целей построения моделей. К примеру, наличие корреляции между показателями – основание для построения сигнальных моделей, а для моделей принятия решений требуется наличие не просто корреляции временных рядов, но причинно-следственных связей.

Следующий этап – отбор и подготовка данных к анализу может быть так или иначе автоматизирован программными методами. Данный этап безусловно важен, но не интересен с теоретической точки зрения.

Этапы моделирования и оценки качества полученных моделей являются основными из подвергающихся программной автоматизации, и именно они ставят перед исследователями новые вызовы. Среди таковых следует указать проблему выбора оптимального разделения всей имеющейся совокупности данных на тренировочную, контрольную и проверочную выборки. Разделение всей совокупности данных необходимо для получения независимых выборок для разных этапов построения моделей. В противном случае оценки качества построенных моделей и оценка выбранной оптимальной модели становятся смещенными, что не позволит использовать их в дальнейшем для целей исследователя. На текущий момент не существует строгого подхода для определения оптимального соотношения данного разделения, но существуют практические рекомендации и несколько исследований в этом направлении.

Так в исследовании Kevin K. Dobbin и Richard M. Simon [1], производится анализ влияния пропорции разделения данных на среднеквадратичную ошибку (MSE) оценки

точности прогноза. Авторы утверждают, что для набора данных с более чем 100 наблюдениями, оптимальным является отделение 2/3 общей выборки в качестве тренировочной. Кроме того исследователями было подтверждено, что оптимальная пропорция разделения коррелирует с размером полного набора данных и требуемой точностью модели. Так, чем меньше общая совокупность данных и выше требование к точности прогноза, тем больше данных необходимо включить в тренировочную выборку. В тоже время следует отметить, что исследователи не ставили целью нахождение оптимального соотношения из всех возможных. В рамках данной работы рассматривались лишь два варианта стратегий разделения: в качестве тренировочной выборки использовали или 1/2, или 2/3 все имеющейся совокупности. В тоже время отметим, что вопрос определения контрольной выборки в данной работе не рассматривался.

В альтернативном исследовании Georgios Afendras и Marianthi Markatou [2] приводят теоретические и эмпирические обоснования оптимального разбиения выборки данных с применением метода перекрестной проверки (cross-validation). Авторы, решая задачу оптимизации для определения оптимального размера выборки, приходят к выводу, что для широкого класса функций оптимальный размер обучающей выборки равен половине общего размера выборки, независимо от распределения данных. Таким образом, стремясь установить правила, позволяющие оптимально выбирать размер тренировочной выборки для фиксированного набора данных размера n, авторы подтверждают мнение своих коллег-практиков о необходимости отделения лишь половины данных для целей обучения. В тоже время, как и в исследовании Dobbin и Simon не рассматривается вопрос определения контрольной выборки.

В собственных исследованиях было определено, что оптимальным является отделение 80 % всей имеющейся совокупности данных для целей обучения и валидации модели, что несколько больше рекомендаций предшествовавших исследователей. Дополнительно было обнаружено, что для некоторых временных рядов отделение для этих целей не более 60 % совокупной выборки данных позволяет достичь минимального уровня ошибки, а другие ряды не демонстрируют значительного увеличения точности прогноза при увеличении обучающих выборок данных уже после отделения от совокупной выборки 25 %.

Другой важной особенностью применения машинного обучения следует указать возможность применения новых методов анализа и прогнозирования данных. Среди прочих следует указать метод комбинированного прогноза и метод построения так называемых нейронных сетей. Остановимся на первом более подробно.

Разнообразие множества вариантов прогнозирования данных привело к появлению идеи об объединении результатов нескольких методов для получения комбинированного прогноза. Практика показывает, что данный подход позволяет улучшить точность прогноза. Ряд авторов утверждает: это связано с тем фактом, что комбинированная модель является более сложной и гибкой по отношению к моделям, из которых она состоит [3]. Следует также указать, что простые, базовые модели не могут учитывать всех факторов, взаимодействующих в реальных экономических системах.

О причинах, побуждающих к построению комбинированных прогнозных моделей, говорили в своей работе «Комбинация прогнозов» такие авторы как J. M. Bates и C. W. J. Granger ещё в 1969 году [4]. В работе «Комбинирование прогнозов» 1989 года R. T. Clemen проводится обзор теории и практики применения комбинированных прогнозов и сфер их применения на основе анализа исследования множества авторов [5]. В последующие годы авторы занимались вопросами улучшения качества комбинированных прогнозов, в частности вопросами определения оптимального соотношения результатов прогнозирования входящих в объединенный прогноз [6]. Из данной работы следует вывод, что в оптимальных условиях комбинированный прогноз может даже превосходить по точности прогноз, полученный от наиболее точной модели, входящей в комбинированный прогноз.

Данный метод сохраняет свою актуальность благодаря сравнительной простоте его реализации. Кроме того, возможность получения более точного прогноза всегда будет оставаться одной из главных причин применения новых методов прогнозирования. В собственном исследовании [7] был получен вывод о том, что комбинированный прогноз может приводит к улучшению общего качества прогноза за счёт взаимного нивелирования разнонаправленных ошибок, полученных из других методов. При этом стоит отказаться от применения такого подхода, если все используемые в комбинированном подходе методы имеют одинаковый знак ошибки, или в ситуации, когда направление ошибки применяемых методов непостоянно.

Таким образом, можно заключить, что цифровая трансформация ставит перед исследователями новые вызовы, но в то же время позволяет развивать новые методы анализа и прогнозирования данных, позволяющие получать более качественные результаты.

Библиографические ссылки

1. Dobbin K. K., Simon R. M. Optimally splitting cases for training and testing high dimensional classifiers // BMC Med Genomics. 2011. Vol. 4 (31). P. 31–38.

2. Afendras G., Markatou M. Optimality of training/test size and resampling effectiveness in cross-validation // Journal of Statistical Planning and Inference. 2019. Vol. 199. P. 286–301.

3. Stock J. H., Watson M. W. Combination Forecasts of Output Growth in a Seven-Country Data Set // Journal of Forecasting. 2004. Vol. 23. P. 405–430.

4. Bates J. M., Granger C. W. J. The Combination of Forecasts // Operational Research Society. 1969. Vol. 20. № 4. P. 451–468.

5. Clemen R. T. Combining forecasts: A review and annotated bibliography // International Journal of Forecasting. 1989. Vol. 5. P. 559–583.

6. Armstrong J. S. Combining forecasts // Principles of forecasting: a handbook for researchers and practitioners. 2001. P. 417–439.

7. Сувалов В. О. Свойства комбинированного прогноза и особенности его применения // Банкаўскі веснік. 2020. № 4 (681). С. 18–21.

УДК 353:330.322

УПРАВЛЕНИЕ ИНВЕСТИЦИОННОЙ ДЕЯТЕЛЬНОСТЬЮ В УСЛОВИЯХ ЦИФРОВОЙ ТРАНСФОРМАЦИИ НА РЕГИОНАЛЬНОМ УРОВНЕ

Ю. В. Сухина[1)], Н. Г. Яблонская[2)]

[1)] *Кандидат наук по государственному управлению, доцент кафедры государственной, муниципальной службы и менеджмента Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации, Липецкий филиал, г. Липецк, Россия*

[2)] *Старший преподаватель кафедры менеджмента Донецкой академии управления и государственной службы при Главе ДНР, г. Донецк*

В статье рассмотрены концептуальные основы взаимосвязи региональной инвестиционной политики в условиях цифровой трансформации и уровня развития региона в условиях цифровой трансформации, особенности управления инвестиционной деятельностью в условиях цифровой трансформации на уровне региона. Определены основные подходы к формированию эффективного управления инвестиционной деятельностью в условиях цифровой трансформации на региональном уровне.

Ключевые слова: управление; инвестиционная деятельность региона; инвестиционная политика региона; цифровая трансформация; инвестиционная привлекательность региона; инвестиции.