УДК 656.13 **DOI 10.52928/2070-1616-2025-51-1-70-76**

МЕТОДЫ АНАЛИЗА ДАННЫХ И ПРОГНОЗИРОВАНИЯ АВАРИЙНОСТИ НА ПРИМЕРЕ ГОРОДА МИНСКА

М.Р. ЛЕБЕДЕВА, д-р техн. наук, проф. А.О. ЛОБАШОВ, канд. техн. наук С.С. СЕМЧЕНКОВ
(Белорусский национальный технический университет, Минск)

Представлены методы прогнозирования аварийности на дорогах с целью повышения безопасности дорожного движения. Апробирование методов и прогноз аварийности выполнен на примере г. Минска. Оценка количества погибших и раненых проведена с использованием двух моделей и двух методов: модели ARIMA, модели SARIMA, метода линейной регрессии и метода «случайного леса». Каждый метод и каждая модель оцениваются по точности и достоверности прогнозов. Анализ показал: методы линейной регрессии и «случайного леса» наиболее точно предсказывают количество погибших, но для прогнозирования количества раненных требуется дальнейшая доработка моделей; модели ARIMA и SARIMA дают завышенные прогнозы для обеих категорий. В статье также рассматривается возможность использования экзогенных факторов для повышения точности прогноза. Результаты могут быть полезны для разработки эффективных мер по снижению аварийности и улучшению ситуации на дорогах.

Ключевые слова: *дорожно-транспортные происшествия, машинное обучение, прогнозирование, линейная регрессия, модель ARIMA, модель SARIMA, метод «случайного леса», безопасность дорожного движения.*

Введение. Анализ дорожно-транспортных происшествий (ДТП) и прогнозирование их последствий являются важнейшими задачами обеспечения безопасности дорожного движения. Возросший в последние годы поток транспортных средств и увеличившаяся интенсивность дорожного движения в крупных и крупнейших городах Беларуси обострили проблему аварийности, что делает задачу прогнозирования количества пострадавших и погибших в ДТП актуальной. Понимание динамики ДТП позволяет выявить не только основные причины аварий, но и зависимости, что поможет разработать более эффективные меры по предотвращению ДТП на дорогах.

В данной статье рассматриваются современные методы анализа данных, которые могут быть использованы для прогнозирования количества погибших и раненых в ДТП, среди них: модели анализа временных рядов, методы машинного обучения и другие подходы, позволяющие учесть временные тренды, сезонные колебания и нелинейные связи между переменными. На основе данных о ДТП в Минске за последние года проводится сравнительный анализ этих методов с целью выявления наиболее точного и эффективного инструмента для прогнозирования.

Основная *цель работы* – предложить и протестировать различные подходы к прогнозированию последствий ДТП, а также оценить их применимость для анализа аварийности в условиях крупных и крупнейших городов Беларуси. Кроме того, важно отметить, что анализ данных о ДТП может помочь в выявлении факторов, способствующих ДТП, таких как погодные условия, состояние дорожного покрытия, интенсивность движения и поведение водителей. В этом контексте применение кластерного анализа позволит выделить группы аварий с схожими характеристиками для разработки целевых стратегий предотвращения ДТП.

В настоящее время транспортные системы быстро развиваются, и применение современных методов прогнозирования становится необходимым для формирования безопасной среды для участников дорожного движения. В Беларуси применение методов машинного обучения – не редкость. Они активно внедряются в IT-компаниях, банках, финансовых организациях, университетах и пр. Главной отличительной особенностью машинного обучения является адаптивность, что необходимо в области дорожного движения.

Таким образом, *целью статьи* является систематизация существующих методов прогнозирования аварийности и применение их на практике. В виде исходных данных взята статистика количества погибших и раненых в городе Минске, сделан прогноз на 2023 год и проведено сравнение прогноза с фактическими значениями количества погибших и раненых. Результаты исследования могут оказать влияние на разработку мероприятий по повышению безопасности дорожного движения и принять решение о внедрении машинного обучения в существующие практики по анализу статистики ДТП.

Основная часть. Существуют разнообразные техники анализа данных, применение которых направлено на выявление «полезной» информации из имеющейся базы данных. В работе со статистикой ДТП в первую очередь следует выделить классический статистический подход – описательную статистику. Данный метод используют на начальных этапах исследования, он позволяет выявить общие закономерности и тренды. Вторым по популярности методом является регрессионный анализ. Его применяют для моделирования зависимостей между переменными, что позволяет отследить, как изменение одной переменной влияет на изменения второй. И для наиболее сложных анализов применяются методы машинного обучения.

Возвращаясь к методам анализа ДТП, отметим три классических метода: количественный, качественный и топографический. Выбор метода определяется видом поставленной задачи. Количественный метод позволяет оценить частоту ДТП, тяжесть последствий ДТП, сравнить статистику по различным временным периодам и пр. Это говорит о том, что метод фокусируется на сборе и анализе числовых данных. Качественный метод позволяет определить причины и факторы, способствовавшие возникновению ДТП. Данный метод учитывает поведение водителей, состояние дороги и причины ДТП. Топографический метод подразумевает использование картографических данных, которые позволяют выявить опасные участки дороги и проанализировать влияние дорожных условий (повороты, пересечения, изменение количества полос) на создание аварийной ситуации.

С развитием информационных технологий и внедрением машинного обучения началось применение новых инструментов анализа – это моделирование аварийных ситуаций, использование искусственного интеллекта для прогнозирования и анализа факторов риска, а также применение программ, способных анализировать большое количество данных.

Таким образом, методы анализа данных можно представить в виде схемы (рисунок 1).

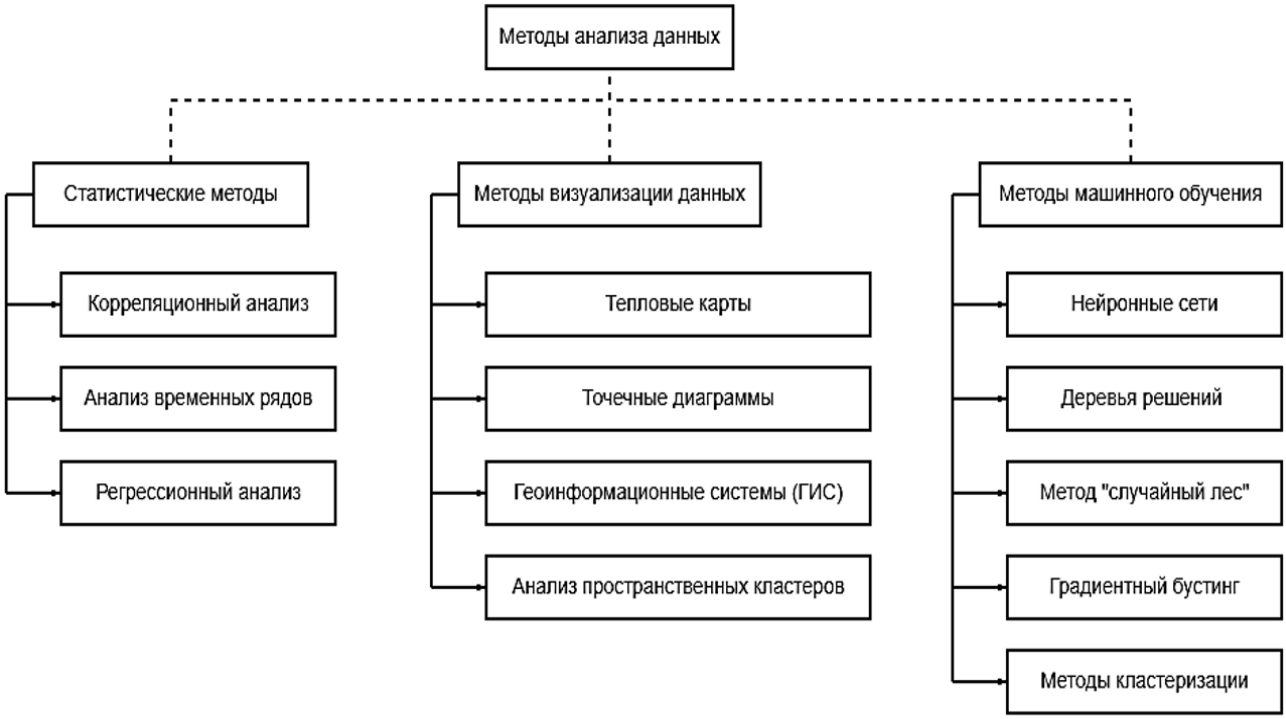

Рисунок 1. – Методы анализа данных

При выборе метода прогнозирования необходимо учесть несколько факторов: какова цель анализа, доступные ресурсы, какими характеристиками обладают имеющиеся данные и необходимая точность прогнозирования.

Прогнозирование – обоснованное предположение, которое базируется на анализе данных предыдущих происшествий, которые могут включать в себя множество характеристик, таких как: время суток, дорожные условия, погодные условия, характеристики транспортного средства и пр. Его необходимость обусловлена тем, что будущее состояние объекта существенно влияет на принимаемые в настоящем решения. Будущая ситуация неизбежно связана с неопределенностью, которую невозможно полностью устранить. В условиях неопределенности основная функция лица, принимающего решения, заключается в выборе наилучшего варианта из предложенных. Прогнозирование является одним из основных инструментов поиска такого решения на основе эмпирического и научно-обоснованного анализа проблемы.

Для выбора подходящего метода необходимо изучить данные, обратить внимание на сезонные колебания, корреляции между переменными и другие особенности, которые могут помочь определить подходящий метод прогнозирования. Также требуется верно поставить цель анализа и представить, какой конечный результат необходим. Требования, выставляемые к анализу, напрямую будут влиять на выбор метода.

Для удобства определения метода был создан алгоритм, представленный на рисунке 2.

Заключительным шагом будет оценка полученной модели и результатов анализа. Оценка достоверности модели – это процесс проверки, насколько хорошо модель отражает фактические данные и насколько эффективно она может решать задачи, для которых была разработана. Это важный этап в анализе дан-

ных и машинном обучении, т.к. он позволяет удостовериться, что полученная модель будет давать точные и надежные результаты применительно к новым данным.

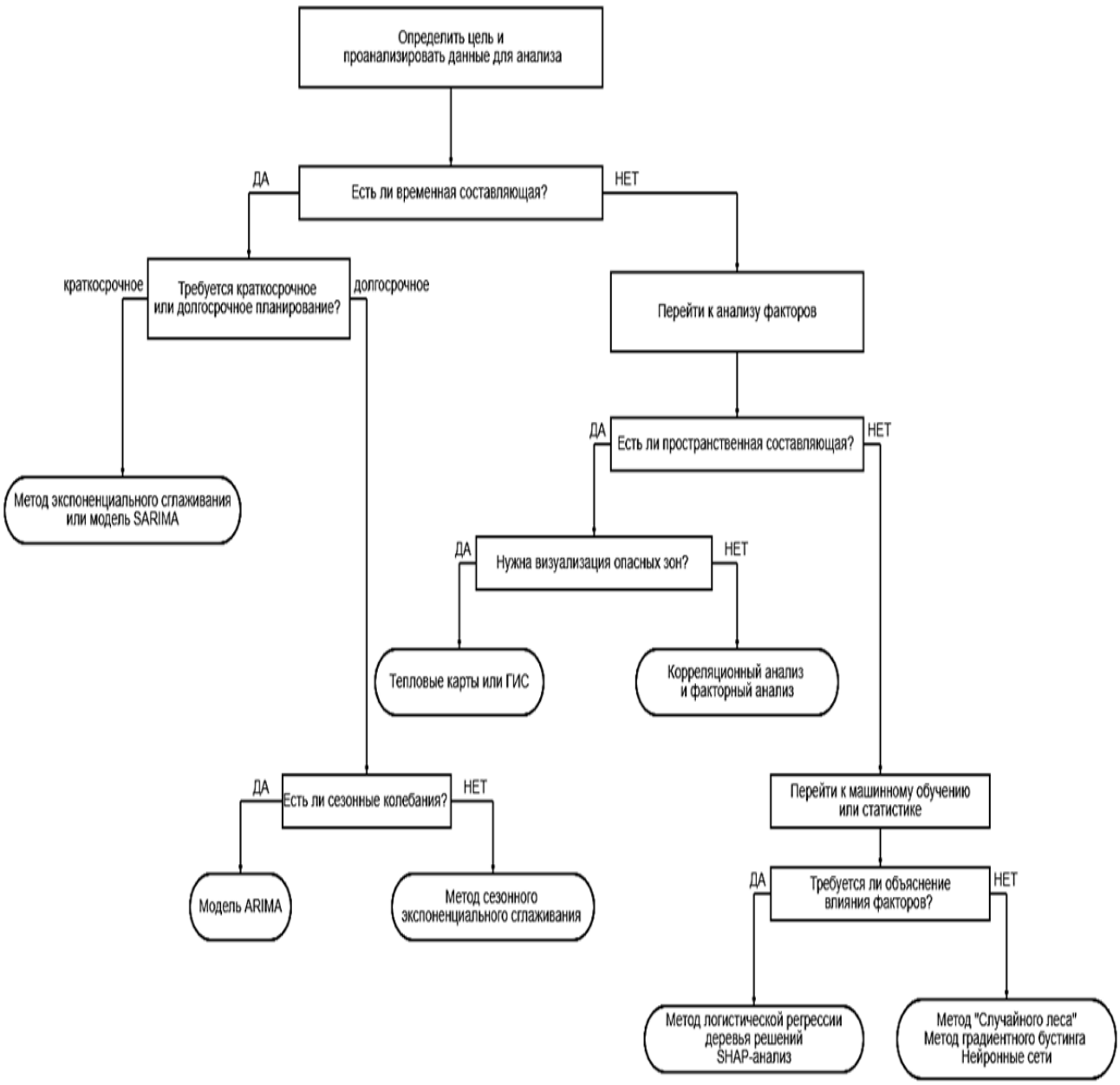

Рисунок 2. – Алгоритм выбора метода анализа и прогнозирования

Далее с использованием программного комплекса Python был проведен анализ методами линейной регрессии и «случайного леса», с помощью моделей ARIMA и SARIMA.

Линейная регрессия относится к статистическому методу, который применим для моделирования связи между зависимой переменной и одной или несколькими независимыми переменными, что предполагает линейную связь [1–5]. Данный метод является самым простым и понятным, его легко применить и интерпретировать результаты. Целью линейной регрессии является подбор к анализируемым данным линейной функции, которая лучше всего отобразит взаимосвязь переменных. Код, используемый для прогноза методом линейной регрессии, представлен на рисунке 3.

Согласно данным, полученным методом линейной регрессии, прогноз количества погибших в ДТП на 2023 г. равен 26, прогноз количества раненых – 658,72. Для оценки достоверности прогноза можно использовать различные метрики качества модели. В случае линейной регрессии одной из таких метрик является коэффициент детерминации R2, который показывает, насколько хорошо модель соответствует данным, и принимает значения от 0 до 1, где 1 – идеальное соответствие, 0 – отсутствие соответствия. Коэффициент детерминации для прогноза количества погибших составил 0,6, а для прогноза количества раненых – 0,29.

```python
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, r2_score

# Подготовка данных (например, X - признаки, y - целевая переменная)
# X_train, y_train, X_test, y_test - это обучающие и тестовые данные

# Инициализация модели
model = LinearRegression()

# Обучение модели
model.fit(X_train, y_train)

# Прогнозирование
y_pred = model.predict(X_test)

# Оценка точности
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Средняя абсолютная ошибка: {mae}")
print(f"Коэффициент детерминации: {r2}")
```

Рисунок 3. – Метод линейной регрессии в Python

Модель ARIMA (Autoregressive Integrated Moving Average) –статистическая модель, используемая для анализа и прогнозирования временных рядов. Другими словами, ARIMA помогает спрогнозировать будущее значение на основе ранее наблюдавшихся данных [6; 7].

Прогноз количества погибших и количества раненых моделью ARIMA составил 36 и 701,59, соответственно. Для оценки достоверности модели используется метод средней абсолютной ошибки, которая составила 12 для количества погибших и 109,59 для количества раненых.

Модель SARIMA – это расширенная модель ARIMA, но с сезонной корректировкой. Данная модель учитывает сезонные закономерности и периодические колебания [8; 9]. Прогноз количества погибших и раненых моделью SARIMA совпал с прогнозом моделью ARIMA, Это значит, что при использовании модели в анализ не был включен сезонный фактор. В таких случаях данные становятся равны. Для изменения результатов при прогнозировании моделью SARIMA необходимо включить погодные условия или сезон года.

Следующий метод – метод «случайного леса» – один из алгоритмов машинного обучения, используемый в прогнозировании данных. Он создает множество деревьев решений [10]. Каждое отдельно взятое дерево в модели работает с использованием случайно выбранных подмножествах данных, что исключает переобучение и делает модель стабильной и устойчивой. Таким образом, каждое дерево в процессе прогноза предлагает собственное решение, а в итоге выбирается тот результат прогноза, за который проголосовало большинство.

Метод дал результат 25,35 для количества погибших и 716,08 для количества раненых. Коэффициент детерминации для погибших составил 0,772 и 0,17 для раненых. Результат коэффициента детерминации для раненых близок к 1, что говорит о том, что прогноз близок к фактическому значению количества погибших, а его уровень можно считать высоким. Результат коэффициента детерминации для прогноза количества раненых низок, значит, прогноз далек от фактического значения.

В таблице 1 приведены результаты прогнозирования и значения коэффициента детерминации или средней абсолютной ошибки в зависимости от метода или модели.

Таблица 1. – Значения прогноза

Методы и модели	Показатель	Количество погибших	Количество раненых
Линейная регрессия	Значения прогноза	26,00	658,72
	Коэффициент детерминации	0,60	0,29
Модель ARIMA	Значения прогноза	36,00	701,59
	Средняя абсолютная ошибка	12,00	109,59
Модель SARIMA	Значения прогноза	36,00	701,59
	Средняя абсолютная ошибка	12,00	109,59
Метод «случайного леса»	Значения прогноза	25,35	716,08
	Коэффициент детерминации	0,77	0,17
Фактические значения		24	592

На рисунках 4 и 5 представлено визуальное сравнение результатов прогнозирования с фактическими значениями погибших и раненых за 2023 г, соответственно.

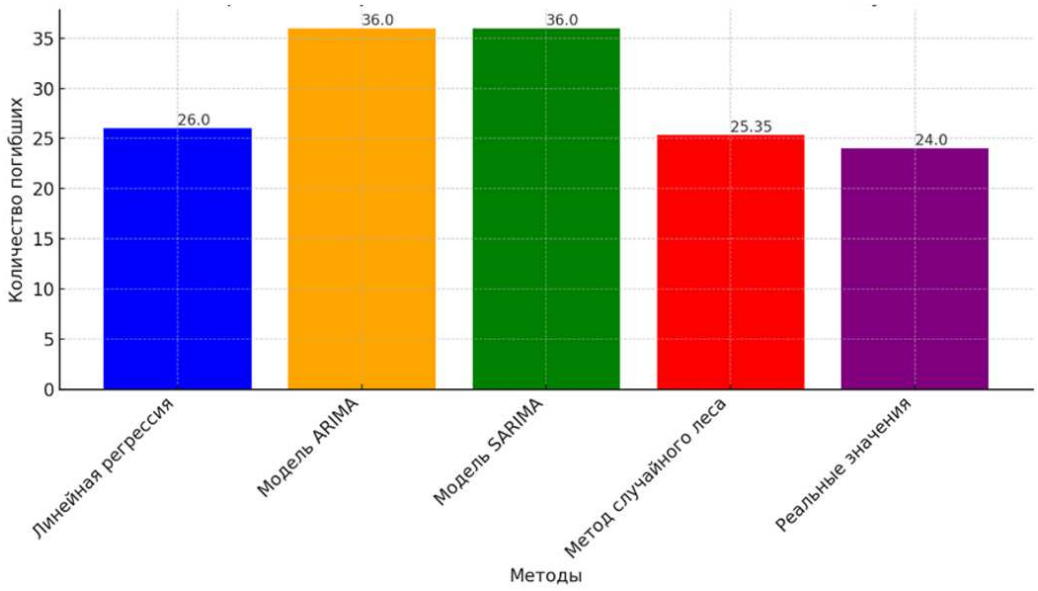

Рисунок 4. – Сравнение результатов прогнозирования с фактическими значениями погибших в ДТП

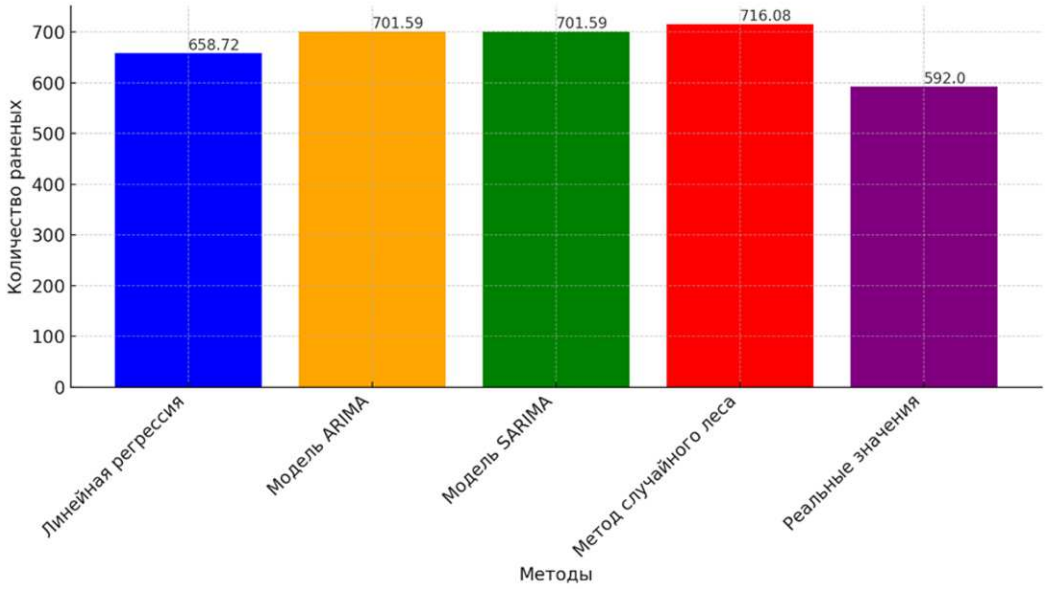

Рисунок 5. – Сравнение результатов прогнозирования с фактическими значениями раненых в ДТП

Заключение. Линейная регрессия прогнозировала 26 погибших – близко к фактическому значению 24. Коэффициент детерминации 0,6, что указывает на высокую достоверность модели для прогнозирования количества погибших.

Модели ARIMA и SARIMA дали одинаковый прогноз – 36 погибших. Значение существенно выше фактического, средняя абсолютная ошибка равна 12. Это означает, что среднее абсолютное отклонение между прогнозами модели и фактическим количеством погибших на 2023 г. составляет около 12 человек, что является высоким отклонением.

Метод «случайного леса» прогнозировал 25,35 погибших, что близко к фактическому значению. Коэффициент детерминации также достаточно высок, что свидетельствует о хорошем качестве модели.

Линейная регрессия прогнозировала 658,72 раненых, что выше фактического значение – модель могла бы быть точнее.

ARIMA и SARIMA – обе модели дали одинаковый прогноз в 701,59 раненых. Это значительно выше фактического значения 592, что указывает на возможные ошибки в модели или несоответствия в данных.

Метод «случайного леса» прогнозировал 716,08 раненых, что выше фактического значения. Коэффициент детерминации для этой модели был отрицательным, что свидетельствует о плохом качестве модели для прогнозирования количества раненых.

Таким образом, для прогнозирования количества погибших можно использовать методы линейной регрессии и «случайного леса». Оба метода оказались наиболее точными для прогнозирования количества погибших и дали прогнозы, близкие к фактическим значениям.

Для прогнозирования количества погибших метод «случайного леса» показал наихудший результат среди всех методов и моделей, что говорит о его неспособности точно прогнозировать этот показатель.

Модели ARIMA и SARIMA для обоих показателей дали одинаковые прогнозы, оказавшиеся завышенными по сравнению с фактическими значениями. Это указывает на возможные проблемы с этими моделями в контексте данных.

Исходя из вышесказанного, для прогнозирования количества раненых стоит рассмотреть улучшение моделей или использование других методов, т.к. текущие модели дают значительные отклонения от фактических значений. Для повышения точности прогнозов стоит рассмотреть комбинирование моделей временных рядов с методами машинного обучения. Такой подход поможет учесть как временные зависимости, так и нелинейные взаимодействия между переменными.

ЛИТЕРАТУРА

1. Amin M., Sadia S. Traffic Accident Prediction Using a Machine-Learning-Enabled Data Analysis // International Journal of Advanced Computer Science and Applications. – 2021. – Vol. 12, no. 1. – P. 104–111.
2. Assessing Driver Fatigue During Urban Traffic Congestion Using ECG Method / N.Gyulyev, A. Galkin, T. Schlosser et al. // Dynamics in Logistics. – 2022. – May. – P. 449–461. DOI:10.1007/978-3-031-05359-7_36
3. The driver's visual perception research to analyze pedestrian safety at twilight / B. González-Hernández, D.S. Usami, O. Prasolenko et al. // Transportation Research Procedia. – 2020. – Vol. 45. – P. 827–834. DOI: 10.1016/J.TRPRO.2020.02.087
4. Chowdhury M.S., Khondoker M.R. Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis // Journal of Transportation Technologies. – 2022. – Vol. 12. – P. 221–235.
5. Choudhary D., Gupta S. Traffic Accident Forecasting using ARIMA and SARIMA Models. International Journal of Engineering and Technology. – 2023. – Vol. 12, no. 4. – P. 505–511.
6. Jha A.K., Prasad P. Comparison of Time-Series Methods for Accident Forecasting // Journal of Transportation Safety & Security. – 2021. – Vol. 13, no. 5. – P. 682–696.
7. Shafique U., Farooq U. Predicting Traffic Accidents Using Statistical and Machine Learning Methods // Journal of Traffic and Transportation Engineering. – 2022. – Vol. 10, no. 2. – P. 132–141.
8. Almeida A., Silva P., Rodrigues P. Seasonal ARIMA Model for Traffic Accident Forecasting // Journal of Transportation Engineering. – 2022. – Vol. 148, no. 3. – P. 04022012.
9. Sangare S., Sene M. Predicting Road Traffic Accidents Using Analytical Measures and Hybrid Machine Learning // International Journal of Advanced Computer Science and Applications. – 2021. – Vol. 12, no. 5. – P. 123–130.
10. Almeida A., Silva P., Rodrigues P. Seasonal Auto Regressive Integrated Moving Average (SARIMA) Model for Traffic Flow Forecasting // Journal of Transportation Engineering. – 2021. – Vol. 147, no. 10. – P. 04021047.

REFERENCES

1. Amin, M. & Sadia, S. (2021). Traffic Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *International Journal of Advanced Computer Science and Applications*, *12*(1), 104–111.
2. Gyulyev, N., Galkin, A., Schlosser, T., Capayova, S. & Lobashov, O. (2022). Assessing Driver Fatigue During Urban Traffic Congestion Using ECG Method. *Dynamics in Logistics*, (May), 449–461. DOI:10.1007/978-3-031-05359-7_36
3. González-Hernández, B., Usami, D.S., Prasolenko, O., Burko, D., Galkin, A., Lobashov, O. & Persia, L. (2020). The driver's visual perception research to analyze pedestrian safety at twilight. *Transportation Research Procedia*, (45), 827–834. DOI: 10.1016/J.TRPRO.2020.02.087
4. Chowdhury, M.S. & Khondoker, M.R. (2022). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Journal of Transportation Technologies*, (12), 221–235.
5. Choudhary, D. & Gupta, S. (2023). Traffic Accident Forecasting using ARIMA and SARIMA Models. *International Journal of Engineering and Technology*, *12*(4), 505–511.
6. Jha, A.K. & Prasad, P. (2021). Comparison of Time-Series Methods for Accident Forecasting. *Journal of Transportation Safety & Security*, *13*(5), 682–696.
7. Shafique, U. & Farooq, U. (2022). Predicting Traffic Accidents Using Statistical and Machine Learning Methods. *Journal of Traffic and Transportation Engineering*, *10*(2), 132–141.
8. Almeida, A., Silva, P. & Rodrigues, P. (2022). Seasonal ARIMA Model for Traffic Accident Forecasting. *Journal of Transportation Engineering*, *148*(3), 04022012.
9. Sangare, S. & Sene, M. (2021). Predicting Road Traffic Accidents Using Analytical Measures and Hybrid Machine Learning. *International Journal of Advanced Computer Science and Applications*, *12*(5), 123–130.

10. Almeida, A., Silva, P. & Rodrigues, P. (2021). Seasonal Auto Regressive Integrated Moving Average (SARIMA) Model for Traffic Flow Forecasting. *Journal of Transportation Engineering*, *147*(10), 04021047.

Поступила 18.11.2025

METHODS OF DATA ANALYSIS AND ACCIDENT RATE PREDICTION ON THE EXAMPLE OF THE CITY OF MINSK

M. LEBEDEVA, A. LOBASHOV, S. SEMCHENKOV
(Belarusian National Technical University, Minsk)

The article discusses methods for predicting accidents on the roads in order to improve road safety. The methods were tested and the accident rate forecast was carried out using the example of Minsk, the estimate of the number of dead and injured was carried out using two models and two methods: the ARIMA model, the SARIMA model, the linear regression metric and the "Random Forest" method. Each method and each model is evaluated according to the accuracy and reliability of the forecasts. The analysis showed that linear regression and "Random Forest" methods most accurately predict the number of deaths, while the ARIMA and SARIMA models provide overestimated forecasts for both categories, and further refinement of the models is required to predict the number of injured. The article also discusses the possibility of using exogenous factors to improve the accuracy of the forecast. The results can be useful for developing effective measures to reduce accidents and improve the situation on the roads.

Keywords: *traffic accidents, machine learning, forecasting, linear regression, ARIMA model, SARIMA model, "Random forest" method, road safety.*