**К**

УДК 004.8:004.6

Э. В. Конончук, Т. В. Ермоленко, Т. О. Шишунов
Государственное образовательное учреждение высшего профессионального образования
«Донецкий национальный технический университет»
283001, г. Донецк, ул. Университетская, 24

# МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ ВЕРОЯТНОСТИ ПОЯВЛЕНИЯ ДТП И ЕГО СЕРЬЕЗНОСТИ

E. V. Kononchuk, T. V. Ermolenko, T. O. Shishunov
State Educational Institution of Higher Professional Education «Donetsk National Technical University»
283001, Donetsk, University st, 24

# MACHINE LEARNING MODELS FOR ASSESSING THE PROBABILITY OF ACCIDENT AND ITS SEVERITY

Е. В. Конончук, Т. В. Єрмоленко, Т. О. Шишунов
Державна освітня установа вищої професійної освіти
«Донецький національний технічний університет»
283001, м. Донецьк, вул. Університетська, 24

# МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ ОЦІНКИ ЙМОВІРНОСТІ ПОЯВИ ДТП ТА ЙОГО СЕРЙОЗНОСТІ

В статье приведены результаты анализа значимости факторов, влияющих на возникновение ДТП, а также проведен анализ эффективности предиктивных моделей, построенных на основе деревьев решений и нейросетей. Обучение моделей проводилось на наборе данных о ДТП в Америке, взятых с сайта Kaggle.
**Ключевые слова:** разведочный анализ данных, случайный лес, анализ главных компонент, многослойный персептрон.

The article presents the results of the analysis of the significance of factors affecting the occurrence of road accidents, as well as an analysis of the effectiveness of predictive models built on the basis of decision trees and neural networks. The models were trained on a dataset of road traffic accidents in America taken from the Kaggle website.
**Key words:** exploratory data analysis, random forest, principal component analysis, multilayer perceptron.

У статті наведено результати аналізу значущості факторів, що впливають на виникнення ДТП, а також проведено аналіз ефективності передиктивних моделей, побудованих на основі дерев рішень та нейромереж. Навчання моделей проводилося на наборі даних про ДТП в Америці, взятих із сайту Kaggle.
**Ключові слова:** розвідувальний аналіз даних, випадковий ліс, аналіз головних компонентів, багатошаровий персептрон

# Введение

Технологии Data Mining позволяют выявлять скрытые закономерности во множестве данных и связать их с влиянием разных факторов для прогнозирования вероятности появления и развития негативных событий, наиболее массовыми из которых являются дорожно-транспортные происшествия (ДТП). Data Mining используется для выявления скрытых закономерностей в данных о произошедших инцидентах с целью прогноза новых происшествий. Качественно построенная предиктивная модель опасности возникновения ДТП, внедренная в систему мониторинга, на основе предиктивного анализа событий необходима для превентивного реагирования на возможные происшествия в дорожной среде для снижения рисков нештатных и аварийных ситуаций и оказания экстренной помощи.

Предварительное исследование имеющихся данных (разведочный анализ) заключается в определении их основных характеристик, а также взаимосвязей между предикторами с целью сужения набора методов, используемых для создания предиктивной модели.

**Целью данной работы** является сравнительный анализ эффективности различных моделей машинного обучения для оценки вероятности появления ДТП и его серьезности.

Для достижения цели необходимо:

- выполнить разведочный анализ факторов влияния на серьезность ДТП для определения наиболее значимых из них;
- по результатам разведочного анализа построить модели машинного обучения для оценки серьезности ДТП на основе деревьев решений и нейросетевого подхода;
- оценить качество построенных моделей на тестовой выборке.

# Описание набора данных

В данной работе использовался набор данных о ДТП в Америке. Данный набор был взят с сайта Kaggle [1]. Набор содержит 47 полей, это – факторы, описывающие: серьезность аварии; протяженность дороги, пострадавшей от аварии; ее географическое положение в координатах GPS; ряд метеорологических наблюдений; наличие тех или иных знаков, перекрестков, светофоров, остановок, лежачих полицейских; период дня, в который произошла авария.

Набор данных содержит 1 516 064 записей. Число пустых значений в столбцах составляет 2 427 878. Число записей после удаления строк с пустыми значениями составляет 334 821. Факторы, которые вошли в предиктивную модель, приведены в табл. 1.

Таблица 1 – Описание используемых переменных

| № | Обозначение | Описание |
|---|---|---|
| 1. | Severity | серьезность аварии, число от 1 до 4, где 1 означает наименьшее воздействие на движение (т.е. короткую задержку в результате аварии), а 4 означает значительное влияние на движение |
| 2. | Temperature | температура в Фаренгейтах |
| 3. | Humidity | влажность воздуха в процентах |
| 4. | Pressure | давление воздуха в дюймах |
| 5. | Visibility | показывает видимость в милях |
| 6. | Wind_Speed | скорость ветра в милях в час |
| 7. | Precipitation | количество осадков в дюймах, если они есть |
| 8. | Amenity | наличие удобств в близлежащем месте |
| 9. | Bump | наличие лежачих полицейских или горных дорог в близлежащем месте |
| 10. | Crossing | наличие перекрестка в близлежащем месте |

Продолжение таблицы 1

| | | |
|---|---|---|
| 11. | Give_way | наличие знака «уступи дорогу» в близлежащем месте |
| 12. | Junction | наличие перекрестка в соседнем месте |
| 13. | No_Exit | отсутствие выхода в близлежащем месте |
| 14. | Railway | наличие железной дороги в близлежащем месте |
| 15. | Roundabout | наличие кругового перекрестка в близлежащем месте |
| 16. | Station | наличие остановки в ближайшем месте |
| 17. | Stop | наличие знака стоп в ближайшем месте |
| 18. | Traffic_Calming | наличие зон спокойного движения в ближайшем месте |
| 19. | Traffic_Signal | наличие светофора в ближайшем месте |
| 20. | Turning_Loop | наличие поворотной петли |
| 21. | Hour | время происшествия |

Переменная Severity является зависимой (целевой), остальные – предикторы.

## Разведочный анализ данных

Первый этап разведочного анализа – проверка данных на выбросы. Для определения аномальных значений использовалось 2 метода: метод средних отклонений [2] и метод боксплотов [3]. Оба метода показали наличие выбросов в данных. Для построения качественной модели записи, содержащие выбросы, были удалены. В итоге после удаления записей с выбросами и пустыми значениями объем выборки (обучающей и тестовой) уменьшился до 307 565.

Следующий этап – проверка переменных на коллинеарность. Оценить коллинеарность переменных можно из значений элементов корреляционной матрицы, изображенной на рис. 1.

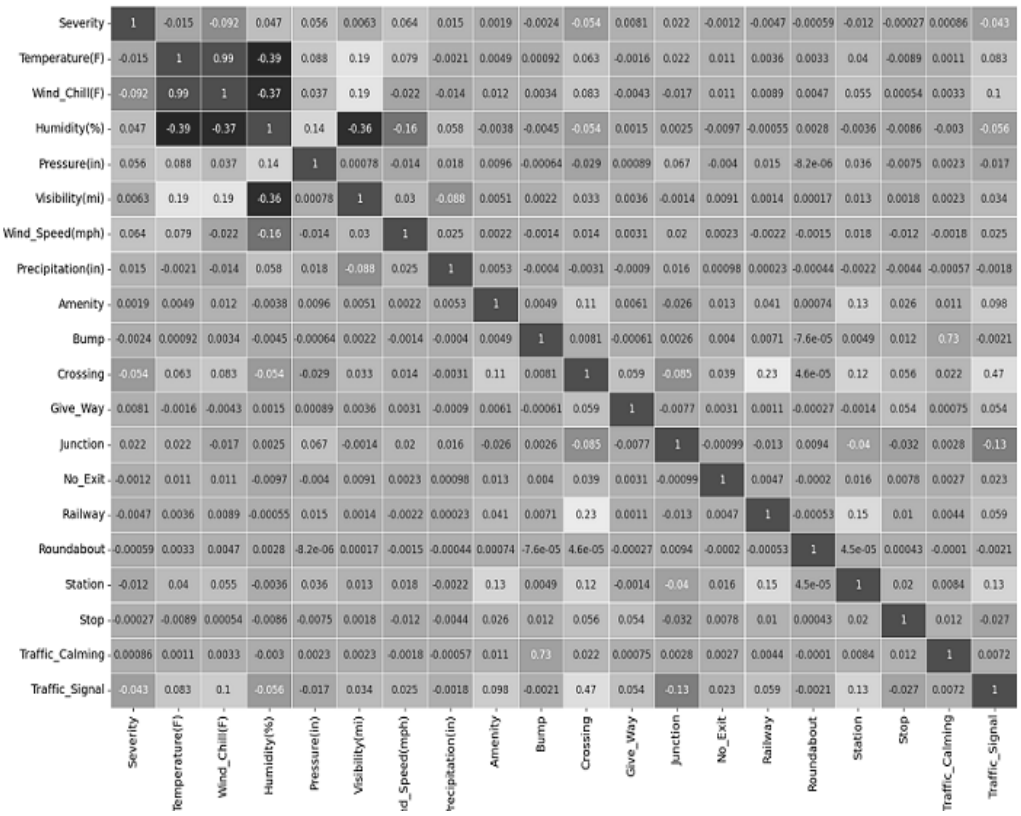


Рисунок 1 – Корреляционная матрица признаков в виде тепловой карты

Видна сильная корреляция между переменными Temperature и Wind_Chill, а также корреляция послабее между Bump и Traffic_Calming. Таким образом, в модели из первой пары коррелированных признаков лучше оставить только один. Для устранения корреляции между признаками и сокращения размерности перспективным представляется использовать анализ главных компонент [4].

Первая строка корреляционной матрицы указывает на силу линейной связи между зависимой переменной Severity и остальными предикторами и свидетельствует в нашем случае об отсутствии таковой. Следовательно, методы линейного регрессионного анализа для построения модели неприменимы.

По данным за 2016 – 2020 гг. проведено численное исследование влияния времени суток (Hour) на количество ДТП, результаты которого отображены на рис. 2.

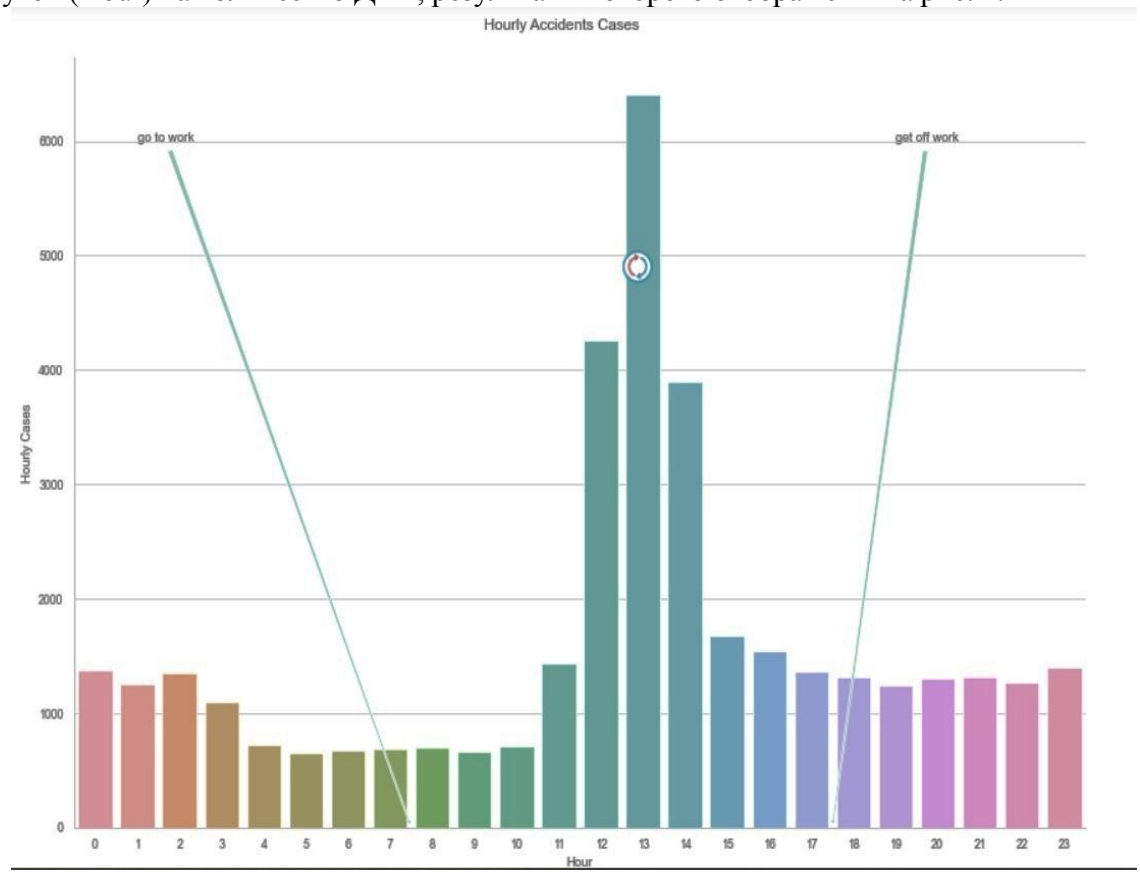


Рисунок 2 – Зависимость количества ДТП от времени суток

Как видно из рис. 2, ощутимый рост количества ДТП приходится на период с 12 до 14 часов. Этот факт должен учитываться моделью.

На основе визуального анализа графиков зависимостей целевой переменной от других предикторов сделать предположение о форме зависимостей сделать невозможно. В этом случае для построения предиктивной модели наиболее логичным представляется использовать деревья и нейросети. Деревья могут отражать сложные структуры взаимодействия, скрытые в данных и, если их деревья достаточно глубокие, имеют относительно небольшое смещение [5]. Скрытые слои нейронной сети, в свою очередь, извлекают из неструктурированных входных данных внутренние представления, которые затем преобразуются в такое расположение точек в гиперпространстве, чтобы выходной линейный слой смог их легко разделить на классы [6].

Кроме того, среди предикторов, отражающих факторы, влияющие на серьезность ДТП, выявлена мультиколлинеарность, для борьбы с которой используют факторный анализ, позволяющий решать две важные задачи: устранить зависимость факторов и сократить их число. Для выявления наиболее значимых факторов в данной работе использовался метод главных компонент, который с помощью ортогональных вращений в пространстве предикторов формирует новые факторы – главные компоненты, определяет наиболее информативные, а остальные исключает из анализа [4].

## Анализ значимости предикторов

Для имеющегося набора данных методом главных компонент линейными многообразиями аппроксимированы данные, в результате построен ортогональный базис и найден вклад в общий разброс данных для каждого из компонент. При помощи представления данных в таком базисе, можно исключить из рассмотрения составляющие с меньшим вкладом в суммарную дисперсию, применив, таким образом, метод главных компонент как фильтр.

На рис. 3 показан суммарный вклад в дисперсию первыми $n$ компонентами ($n$ = =0, …, 16).

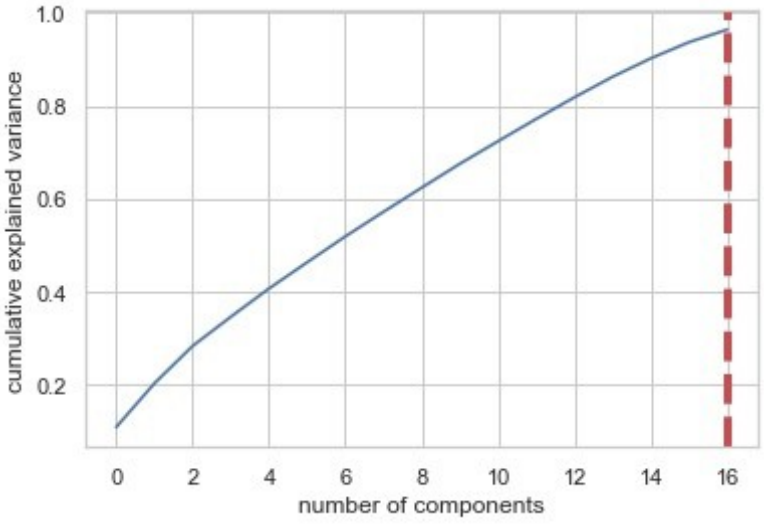


Рисунок 3 – Вклад в дисперсию главных компонент

Как видно из рис. 4, первые 13 главных компонент ($n$ = 0, …, 13) объясняют более 80% дисперсии данных, следовательно, остальные компоненты можно исключить из анализа.

Для первых 8 главных компонент в табл. 2 приведены веса исходных предикторов в базисе главных компонент.

Таблица 2 – Веса исходных предикторов в базисе главных компонент

| Предикторы | Номера главных компонент | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Temperature | 0.4387 | -0.0426 | -0.10825 | 0.2905 | 0.2097 | -0.0176 | 0.0549 | -0.1229 |
| Humidity | -0.494 | 0.0407 | 0.28117 | -0.0115 | 0.1866 | -0.0277 | -0.035 | -0.044 |
| Pressure | -0.064 | -0.001 | 0.14158 | 0.11791 | 0.7152 | -0.0606 | 0.0711 | -0.2859 |
| Visibility | 0.3591 | -0.0352 | -0.2937 | -0.3544 | 0.1898 | -0.0161 | -0.04 | -0.0095 |
| Wind_Speed | 0.2633 | -0.0348 | -0.077 | 0.476 | -0.063 | -0.0273 | 0.0508 | 0.01748 |
| Precipitation | -0.216 | 0.0280 | 0.263145 | 0.581 | -0.207 | 0.0032 | 0.0718 | -0.0001 |

Продолжение таблицы 2

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Amenity | 0.1041 | 0.0144 | 0.2589 | -0.037 | 0.2109 | -0.0828 | 0.0974 | 0.50692 |
| Bump | 0.0527 | 0.7036 | -0.0308 | 0.0124 | 0.0034 | 0.00422 | -0.024 | -0.0112 |
| Crossing | 0.3213 | 0.0068 | 0.5141 | -0.1559 | -0.173 | 0.07905 | -0.001 | -0.1228 |
| Give_way | 0.0100 | 0.0046 | 0.0525 | -0.0557 | -0.106 | -0.0366 | 0.558 | -0.1786 |
| Junction | -0.022 | -0.0026 | -0.04 | 0.0558 | 0.1241 | 0.6842 | 0.066 | 0.04031 |
| No_Exit | 0.0502 | 0.0048 | 0.0546 | -0.0478 | -0.108 | 0.0077 | 0.126 | 0.58242 |
| Railway | 0.1221 | -0.003 | 0.36591 | -0.219 | -0.07 | 0.1244 | 0.027 | -0.3957 |
| Roundabout | -0.006 | -0.0003 | -0.0095 | 0.0181 | 0.0809 | 0.6903 | 0.066 | 0.10840 |
| Station | 0.1575 | 0.00203 | 0.26493 | 0.0785 | 0.4127 | -0.1186 | 0.048 | 0.2729 |
| Stop | -0.018 | 0.03909 | 0.00017 | -0.14 | -0.047 | -0.0682 | 0.755 | 0.0072 |
| Traffic_Calming | 0.0563 | 0.70358 | -0.022 | 0.0102 | 0.0068 | 0.0037 | -0.023 | -0.0121 |
| Traffic_Signal | 0.3132 | -0.0240 | 0.4083 | -0.0075 | -0.160 | 0.0237 | -0.215 | 0.01912 |
| Turning_Loop | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hour | 0.2315 | -0.0266 | -0.1284 | 0.320 | -0.061 | 0.0038 | 0.117 | -0.1115 |

Из табл. 2 можно сделать вывод, что наиболее значимыми предикторами являются: Humidity, Temperature, Visibility, Crossing, Traffic_Signal и Hour, т.е. факторы, связанные с погодными условиями (Humidity, Temperature, Visibility), временем суток (Hour), а также с наличием перекрёстка (Crossing) и светофора (Traffic_Signal).

Помимо анализа главных компонент для анализа значимости предикторов использовалась модель случайного леса. Случайный лес (random forest) – модель, представляющая собой композицию классификаторов, которая создает большую коллекцию декоррелированных деревьев, а затем усредняет их [7]. При каждом разделении в каждом дереве показателем значимости переменной является улучшение критерия расщепления, этот показатель приписывается переменной разделения и накапливается по всем деревьям, входящим в лес, отдельно для каждой переменной. Гистограмма значимости переменных в рамках проведенного исследования приведена на рис. 4.

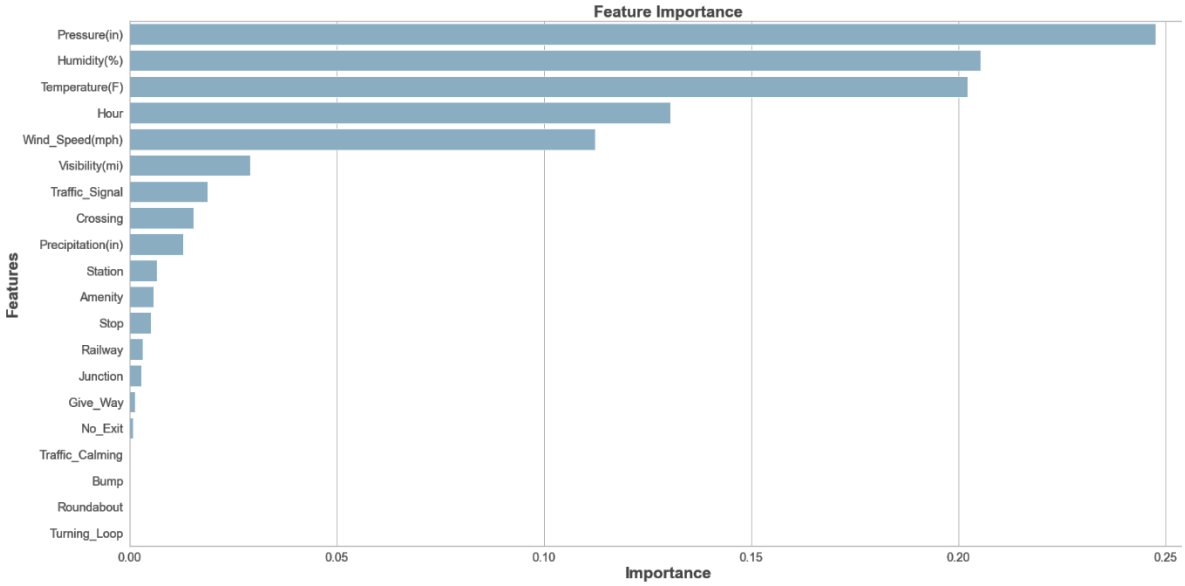


Рисунок 4 – Результат ранжирования исходных предикторов по значимости

Определение значимости предикторов с помощью модели случайного леса дало похожие результаты, что и анализ главных компонент: наиболее значимыми являются Pressure, Humidity, Temperature, Hour и Wind_Speed, т.е. факторы, связанные с погодой и временем суток.

## Построение предиктивных моделей

В работе в качестве предиктивных моделей использовались случайный лес и многослойный персептрон.

Модели обучены на выборке объемом 200 тысяч записей, объем тестовой выборки – 107 тысяч записей.

Реализованы две модели случайного леса: с исходными предикторами, число которых равно 20, а также с 13 предикторами, полученными в результате анализа главных компонент. На тестовой выборке точность первой модели составила 90,01%, второй – 89,81%.

Помимо моделей на основе случайного леса реализована нейросетевая предиктивная модель, в частности, многослойный персептрон со следующими параметрами:

- размер bath – 100;
- количество эпох – 15;
- размер рецепторного слоя – 20 нейронов;
- размер выходного слоя – 5 нейронов;
- количество скрытых слоев – 2;
- количество нейронов в скрытых слоях – 625;
- функция активации – ReLu, Softmax на выходе;
- функция потерь – кросс-энтропия (перекрестная энтропия);
- оптимизация градиентного спуска – Adam;
- Dropout – 0.2.

На тестовой выборке точность нейросетевой модели составила 94,12%, что свидетельствует о превосходстве нейросетевого подхода над деревьями решений в задачах прогнозирования возникновения ДТП и оценке его серьезности.

## Выводы

В результате разведочного анализа данных, отражающих факторы, влияющие на серьезность ДТП, можно заключить, что среди них наиболее значимыми являются факторы, связанные с погодными условиями и временем суток. Анализ главных компонент показал, что наличие перекрестка и светофора также влияет на возникновение ДТП.

Численные исследования эффективности предиктивных моделей на основе деревьев и нейросетей показали преимущество многослойного персептрона, точность этой модели достигла более 94%, что на 4% больше точности случайного леса.

## Список литературы

1. US Accidents: A Countrywide Traffic Accident Dataset (2016 - 2020) [Электронный ресурс]. – URL: https://www.kaggle.com/sobhanmoosavi/us-accidents (дата обращения 16.12.2021).
2. Senthamarai Kannan K.. Labeling Methods for Identifying Outliers [Электронный ресурс] / K. Senthamarai Kannan, K. Manoj, S. Arumugam. – URL: https://www.researchgate.net/publication/283755180_Labeling_Methods_for_Identifying_Outliers (дата обращения 15.12.2021).
3. Андреа Клитон. Двумерный боксплот на основе высокоэффективных робастных оценок масштаба и корреляции [Электронный ресурс] / Андреа Клитон, Смирнов Павел Олегович, Шевляков Георгий Леонидович // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2013. – № 1 (22). – URL: https://cyberleninka.ru/article/n/dvumernyy-boksplot-na-osnove-vysokoeffektivnyh-robastnyh-otsenok-masshtaba-i-korrelyatsii (дата обращения 15.12.2021).

4. Метод Главных Компонент (PCA) [Электронный ресурс]. – URL: https://rcs.chemometrics.ru/old/Tutorials/pca.htm (дата обращения 16.12.2021).
5. Паклин Н. Б. Глава 9 [Текст] / Н. Б. Паклин, В. И. Орешков // Бизнес-аналитика: от данных к знаниям: Учебное пособие. 2-е изд. – СПб. : Питер, 2013. – С. 428–472.
6. Zhang Y. Extracting deep neural network bottleneck features using low-rank matrix factorization / Y. Zhang, E. Chuangsuwanich, J. Glass // IEEE international conference on acoustics, speech and signal processing (ICASSP). – 2014. – С. 185–189.
7. Глава 15 – Случайные леса [Электронный ресурс]. – URL: http://www.williamspublishing.com/PDF/978-5-907144-42-2/part.pdf (дата обращения 16.12.2021).

# References

1. US Accidents: A Countrywide Traffic Accident Dataset (2016 - 2020) [Elektronnyj resurs]. – URL: https://www.kaggle.com/sobhanmoosavi/us-accidents (data obrashcheniya 16.12.2021).
2. K. Senthamarai Kannan, K. Manoj, S. Arumugam. Labeling Methods for Identifying Outliers [Elektronnyj resurs]. URL: https://www.researchgate.net/publication/283755180_Labeling_Methods_for_Identifying_Outliers (data obrashcheniya 15.12.2021).
3. Andrea Kliton, Smirnov Pavel Olegovich, SHevlyakov Georgij Leonidovich Dvumernyj boksplot na osnove vysokoeffektivnyh robastnyh ocenok masshtaba i korrelyacii [Two-dimensional boxplot based on highly efficient robust scale and correlation estimates]. *Vestn. Tom. gos. un-ta*.[ Vestn. Volume. state university Management, computer technology and informatics] Upravlenie, vychislitel'naya tekhnika i informatika. 2013. №1 (22). [Elektronnyj resurs]. URL: https://cyberleninka.ru/article/n/dvumernyy-boksplot-na-osnove-vysokoeffektivnyh-robastnyh-otsenok-masshtaba-i-korrelyatsii (data obrashcheniya 15.12.2021).
4. *Metod Glavnyh Komponent* (PCA) [Method of Principal Components] [Elektronnyj resurs]. URL: https://rcs.chemometrics.ru/old/Tutorials/pca.htm (data obrashcheniya 16.12.2021).
5. Paklin N.B., Oreshkov V.I. Glava 9 [Chapter 9] *Biznes-analitika: ot dannyh k znaniyam: Uchebnoe posobie*. 2-e izd. [Business Analytics: From Data to Knowledge: Textbook. 2nd ed.], SPb.: Piter, 2013, S. 428-472.
6. Zhang Y., Chuangsuwanich E., Glass J. Extracting deep neural network bottleneck features using low-rank matrix factorization / // IEEE international conference on acoustics, speech and signal processing (ICASSP). – 2014. – С. 185-189.
7. Glava 15 – Sluchajnye lesa [Chapter 15 - Random Forests] [Elektronnyj resurs]. URL: http://www.williamspublishing.com/PDF/978-5-907144-42-2/part.pdf (data obrashcheniya 16.12.2021).

## RESUME

*E. V. Kononchuk, T. V. Yermolenko, T. O. Shishunov*
*Machine Learning Models for Assessing the Probability of Accident and its Severity*

Data Mining technologies make it possible to identify hidden patterns in a variety of data and link them with the influence of various factors to predict the likelihood of the occurrence and development of negative events, the most widespread of which are road traffic accidents. A qualitatively constructed predictive model of the risk of an accident, implemented in the monitoring system, based on predictive analysis of events is necessary for a preventive response to possible accidents in the road environment to reduce the risks of emergency and emergency situations and provide emergency assistance.

The models were trained on a dataset of road traffic accidents in America taken from the Kaggle website. To identify significant factors, the principal component analysis method was used, and to predict the severity of the accident, a random forest algorithm was used, a multi-layer perceptron neural network was trained.

Among the advantages of the random forest algorithm: work with a large number of features and classes, high scalability, there are methods for assessing the significance of individual components of the model, equally well handles discrete and continuous features.

One of the disadvantages is the large size of the model, and quite a long training with a large number of trees in the forest. On the other hand, the implementation of a multilayer perceptron has advantages: the ability to study nonlinear models, the ability to study models in real time. Among the disadvantages: it requires setting up a number of hyperparameters, such as the number of hidden neurons, layers and iterations, it is sensitive to scaling functions.

Numerical studies of the effectiveness of predictive models based on trees and neural networks have shown the advantage of a multilayer perceptron, the accuracy of this model has reached more than 94%, which is 4% more than the accuracy of a random forest.

## РЕЗЮМЕ

*Э. В. Конончук, Т. В. Ермоленко, Т. О. Шишунов*
*Модели машинного обучения для оценки вероятности появления ДТП и его серьезности*

В статье приведены результаты анализа значимости факторов, влияющих на возникновение ДТП, а также проведен анализ эффективности предиктивных моделей, построенных на основе деревьев решений и нейросетей.

Обучение моделей проводилось на наборе данных о ДТП в Америке, взятых с сайта Kaggle. На базе этих данных были построены модели: многослойного персептрона, случайного леса, а также случайного леса с использованием метода главных компонент для устранения коррелированности предикторов и уменьшения пространства признаков.

В результате разведочного анализа данных с помощью метода главных компонент и случайного леса были выявлены наиболее значимые предикторы, влияющие на серьезность дорожно-транспортного происшествия. Наиболее значимыми являются факторы, связанные с погодными условиями, временем суток и сложностью перекрестка.

Для решения задачи прогнозирования ДТП и его серьезности лучшим показал себя многослойный персептрон. На тестовой выборке точность этой модели достигла более 94%, что на 4% больше точности случайного леса.

Реализация предложенных моделей позволит найти потенциально опасные участки возникновения автокатастроф, определить их тяжесть. Благодаря этому появляется возможность предотвратить или снизить серьезность аварии.

Статья поступила в редакцию 05.12.2021.