УДК 681.3 Н

ПРОГНОЗИРОВАНИЕ СТЕПЕНИ ТЯЖЕСТИ ПОСЛЕДСТВИЙ ДТП С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

© 2019 Д. С. Донченко, Н. П. Садовникова, Д. С. Парыгин

Волгоградский государственный технический университет (г. Волгоград, Россия)

В этой статье рассматривается возможность разработки модели прогнозирования дорожно-транспортных происшествий на основе базы данных ДТП, предоставленной ГИБДД России. Приведен пример использования собранных данных для разработки модели прогнозирования степени тяжести последствий ДТП и анализируются факторы, влияющие на это.

Ключевые слова: машинное обучение, искусственный интеллект, анализ данных.

В данной работе мы использовали открытые данные о ДТП, которые доступны на официальном сайте ГИБДД [1]. Для анализа использовались данные об авариях, произошедших в России с начала 2015 года по апрель 2018 года. Загруженные данные для удобства работы с ними заносятся в базу данных. В качестве меры тяжести ДТП выступает серьезность ранений, полученных их участниками. Мы решили проанализировать ранения, полученные только водителями, потому что по имеющимся данным не всегда получается соотнести ущерб, полученный транспортным средством, с местонахождением пассажира в автомобиле.

В таблице приведены переменные, которые были использованы для построения модели:

Некоторые характеристики, такие как возраст транспортного средства, были нормализованы путем взятия логарифма его значения для повышения эффективности обучения.

Мы применили библиотеку pandas [2] для чтения данных из базы данных и создания библиотеки наборов данных и scikit-learn [3], чтобы разделить ее на обучающий набор и тестовый набор.

Таблица

Переменные модели

Переменная	**Значения**
Время суток [a]	День(10:00-17:30)
	Ночь(19:30-08:00)
	Часы пик (08:00-10:00, 17:30-19:30)
День недели	Пн, Вт и т.д.
Наличие осадков	Без осадков
	Дождь
	Снег
	Метель
	Туман
Использование ремня безопасности	Да/Нет
Наличие дефектов автомобиля	Да/Нет
Наличие значительных дефектов дорожного покрытия	Да/Нет
Алкогольное / наркотическое опьянение	Да/Нет
Наличие повреждений на стороне водителя автомобиля	Да/Нет
Возгорание автомобиля	Да/Нет
Полное повреждение кузова	Да/Нет

Продолжение таблицы

Освещения дорожного покрытия	Да/Нет
Состояние дорожного покрытия	Сухое
	Мокрое
	Гололед

Донченко Дмитрий Сергеевич – Волгоградский государственный технический университет, аспирант кафедры «Системы автоматизированного проектирования и поискового конструирования»,

Садовникова Наталья Петровна – Волгоградский государственный технический университет, доцент, доктор технических наук, профессор кафедры «Системы автоматизированного проектирования и поискового конструирования».

Парыгин Данила Сергеевич – Волгоградский государственный технический университет, кандидат технических наук, доцент кафедры «Системы автоматизированного проектирования и поискового конструирования».

Возраст автомобиля	Логарифм от скорости
Нарушение правил ПДД	Без нарушений
	Проезд на запрещенный сигнал светофора
	Выезд на встречную полосу
	Превышение скорости
	Нарушение правил обгона
	Опасные маневры
	Нарушение правил парковки
	Ослепление фарами
	Другое
Степень тяжести ДТП	Без повреждений / Незначительные повреждения
	Тяжелые повреждения/смерть

Полученный набор данных содержит данные о 599528 водителях, которые участвовали в дорожно-транспортных происшествиях с 2015 по 2018 год. Большинство водителей (84,86 %) не получили повреждения или получили легкие ранения (мы считали ранение незначительным, если пациент получал амбулаторное лечение), тогда как другие 15,14 % умерли или получили тяжелые травмы с необходимостью стационарного лечения. Мы использовали библиотеку imbalanced-learn [4] для исправления дисбаланса в наборе данных, используя технику недостаточной выборки, уменьшая количество примеров наиболее хорошо представленного класса [5].

После применения недостаточной выборки мы получили около 91 тысячи примеров для обоих классов в наборе данных. На следующем шаге мы разбили набор данных на тренировочный набор и тестовый набор, используя соотношение 90/10.

Мы попытались применить следующие классификаторы, используя библиотеку scikit-learn, чтобы определить, какой из них показывает наилучшую точность в наборе данных:

1. Стохастический градиентный спуск (Stochastic gradient descent classifier, SGD).
2. Метод K-ближайших соседей (K-Nearest Neighbors Classifier, KNN).
3. Логистическая регрессия.
4. Деревья решений.
5. Однослойный персептрон.
6. Random Forest Classifier.
7. Gaussian Naïve Bayes.
8. Градиентный бустинг (Gradient Boosting Classifier).

Наилучшую точность получили дерево решений (66 %), Random Forest (69 %) и классификаторы на основе метода градиентного бустинга (67 %). Производительность других классификаторов была значительно ниже, при этом точность не превышала 55 %, поэтому мы решили сосредоточиться на повышении производительности классификаторов Random Forest и градиентного бустинга. Для следующего шага мы решили применить реализацию классификатора Gradient Boosting с использованием библиотеки XGBoost [6].

Мы применили метод поиска по сетке (Grid Search) с применением кросс-валидации на обучающем наборе для поиска оптимальных значений гиперпараметров как для классификатора Random Forest, так и для классификатора XGBoost.

После оптимизации гиперпараметров мы обучили классификаторы Random Forest и XGBoost и измерили характеристики полученных моделей на тестовом наборе. Точность модели на основе классификатора Random Forest была несколько выше, чем точность модели на основе классификатора XGBoost (73,2 % для модели случайного леса и 71,07 % для модели XGBoost).

На рисунке 1 представлена визуализация характеристик моделей с использованием матриц ошибок (confusion matrix), где каждая строка матрицы представляет экземпляры в прогнозируемом классе, а каждый столбец представляет экземпляры в реальном классе:

Еще одной важной метрикой, которую можно рассчитать на основе матриц ошибок, являются точность precision (которая показывает долю положительных идентификаторов, которые были действительно правильными) и полнота recall (который показывает, какая доля фактических положительных результатов была правильно определена). Эти метрики важны для наших наборов данных о дорожно-транспортных происшествиях, поскольку они показывают, была ли проблема дисбаланса классов решена верно.

В итоге, средний показатель precision-recall для модели случайного леса (AP = 0,67) получился немного лучше, чем для модели XGBoost (AP = 0,65).

Наконец, мы сравнили значения важности переменных для моделей с Random

Forest и XGBoost. Важность каждого признака определяется как изменчивость между деревьями этого признака. Он показывает, сколько атрибутов используется для принятия ключевых решений с деревьями.

Время суток оказалось наиболее важной переменной как для моделей с Random Forest, так и для моделей XGBoost. . Степень тяжести дорожно-транспортных происшествий увеличивается в ночное время из-за сочетания нескольких факторов, таких как плохие условия освещения на дорогах, более длительное время реагирования и более высокие показатели вождения в состоянии опьянения. Другие важные особенности, которые являются общими для обеих моделей, включают такие признаки, как использование ремня безопасности, день недели, дефекты транспортного средства, наличия осадков и дорожного дефекта. Однако некоторые функции («Наличие дефектов автомобиля» «Возгорание автомобиля»), которые имеют большое значение в модели Random Forest, не оказали большого влияния на производительность модели XGBoost. Одна из возможных причин заключается в том, что эти функции могут коррелировать друг с другом и с функцией «Наличие дефектов автомобиля», которая в случае классификатора XGBoost будет означать, что будет выбрана только одна из них, в то время как для Random Forest все признаки будут выбраны случайным образом.

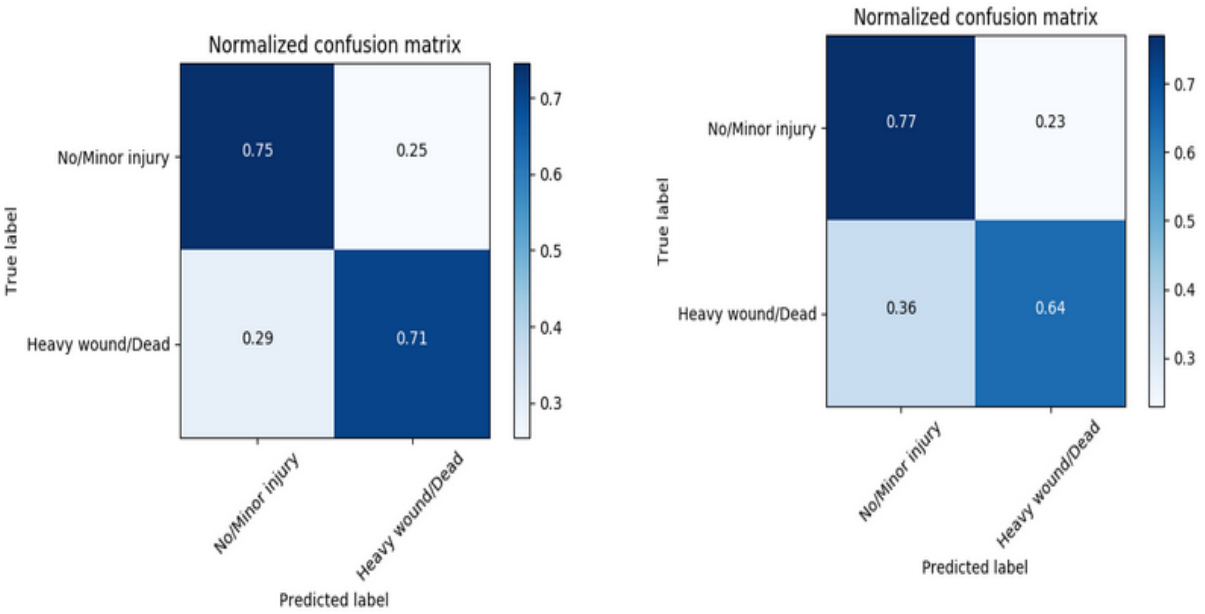

Рисунок 1. Матрицы ошибок Random Forest и XGBoost.

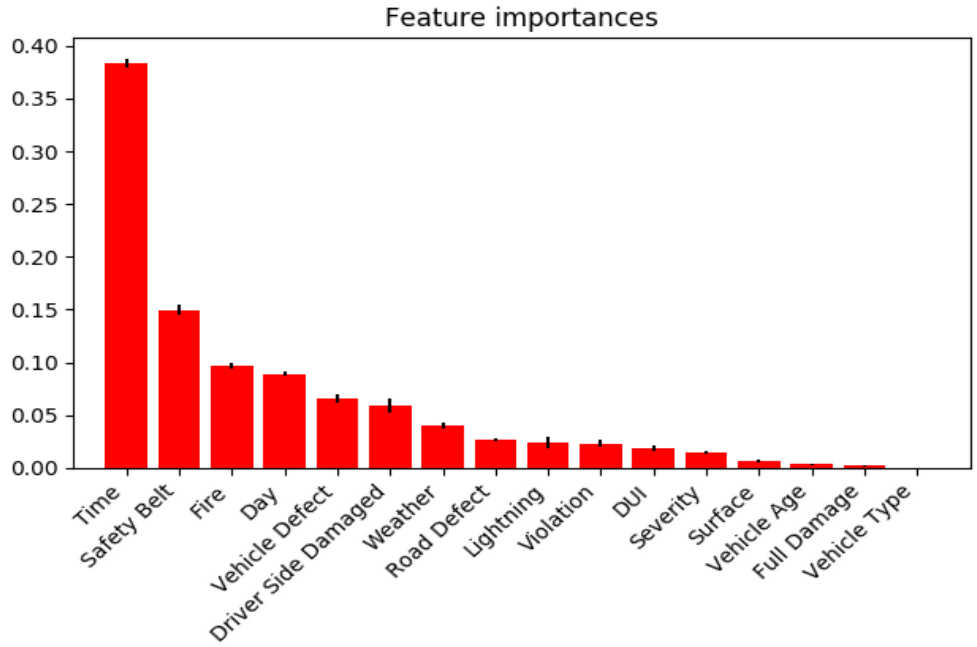

Рисунок 2. Важность признаков в модели Random Forest.

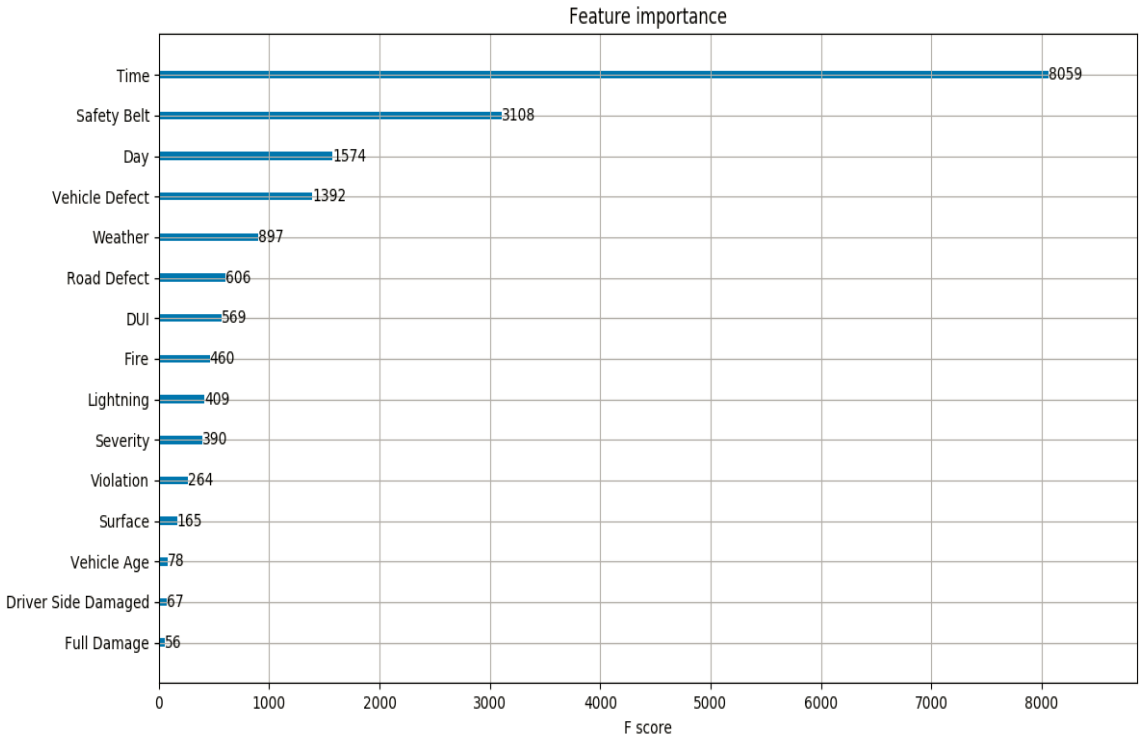

Рисунок 3. Важность признаков в модели XGBoost.

Вывод. Методы ансамблевого машинного обучения (деревья решений, Random forest, градиентный бустинг) превзошли другие классификаторы (такие как логистическая регрессия и Naive Bayes) в построении модели прогнозирования серьезности дорожно-транспортных происшествий. Наилучшая производительность была получена с моделями Random Forest и XGBoost. Однако полученные нами показатели precision (73 %) и recall (0,67) говорят о том, что у нас есть проблемы с данными, используемыми для обучения моделей машинного обучения. Одним из возможных способов повышения точности прогноза является добавление некоторых предположительно важных характеристик, таких как рейтинг безопасности транспортных средств, плотность движения, количество полос движения и тип движения (в одну или две полосы) дороги и т. д. Другой способ улучшить модель - попробовать передискретизацию или различные веса для

классов, чтобы решить проблему несбалансированных классов. Использование этих методов позволит использовать больше данных для обучения по сравнению с техникой недостаточной выборки. Также, возможно, стоит попытаться применить глубокие нейронные сети для построения моделей прогнозирования тяжести дорожно-транспортных происшествий.

ЛИТЕРАТУРА

1. Показатели состояния безопасности дорожного движения (Электронный ресурс - http://stat.gibdd.ru/).

2. Python Data Analysis Library (Электронный ресурс – https://pandas.pydata.org/).

3. scikit-learn (Электронный ресурс – https://scikit-learn.org/stable/).

4. imbalanced-learn (Электронный ресурс – https://imbalanced-learn.readthedocs.io/en/stable/).

5. Haibo He, Edwardo A. Garcia “Learning from Imbalanced Data”, IEEE Transactions on Knowledge and Data Engineering (Volume: 21 , Issue: 9 , Sept. 2009) (Электронный ресурс – https://ieeexplore.ieee.org/document/5128907/).

6. XGBoost documentation (Электронный ресурс – https://xgboost.readthedocs.io/en/latest/).

FORECASTING THE SEVERITY OF THE CONSEQUENCES OF ACCIDENTS WITH THE USE OF METHODS OF MACHINE LEARNING

© 2019 D. S. Donchenko, N. P. Sadovnikova, D. S. Parygin

Volgograd State Technical University (Volgograd, Russia)

This article discusses the possibility of developing a model for predicting traffic accidents on the basis of a crash database provided by the State Traffic Safety Inspectorate of Russia. An example of using the collected data to develop a model for predicting the severity of the consequences of an accident is given, and factors influencing this are analyzed.

Keywords: machine learning, artificial intelligence, data analysis.