**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

Институт/факультет: «Академия “Высшая Инженерная Школа”, АВИШ»

Специальность/Направление подготовки: Информатика и вычислительная техника

**КУРСОВАЯ РАБОТА**

**по дисциплине:** Машинное обучение и анализ данных

|  |  |
| --- | --- |
| **на тему:** | «Прогнозирование тяжести ДТП в зависимости от погодных условий: моделирование и анализ с использованием методов машинного обучения". Вариант 9. |

|  |  |
| --- | --- |
| **Студент группы ШАД-212** | **(Гинда Д.А. )** |
| **Проверил** | **(Атамасов Н.В. )** |

**Москва 2023 г.**

*Оглавление*

[Основные термины 3](#_Toc155733003)

[Введение 4](#_Toc155733004)

[Анализ данных 5](#_Toc155733005)

[Подготовка данных к работе 5](#_Toc155733006)

[Разведочный анализ 8](#_Toc155733007)

[Подготовка данных 12](#_Toc155733008)

[Разделение на выборки 18](#_Toc155733009)

[Обучение моделей 19](#_Toc155733010)

[Автоматическое машинное обучение 19](#_Toc155733011)

[Моделирование 20](#_Toc155733012)

[Лучшая модель 23](#_Toc155733013)

[Интерпретация 24](#_Toc155733014)

[Глобальная интерпретация 24](#_Toc155733015)

[Локальная интерпретация 26](#_Toc155733016)

[Выводы 29](#_Toc155733017)

[Приложение 30](#_Toc155733018)

# Основные термины

Машинное обучение - процесс автоматического обучения и совершенствования поведения системы искусственного интеллекта на основе обработки массива обучающих данных без явного программирования.

Гиперпараметр - параметр модели, который задаются специалистом по машинному обучению для улучшения показателей модели.

Автоматическое машинное обучение – способ автоматизации рутинных задач разработки моделей машинного обучения.

Алгоритмы интерпретации – алгоритмы, помогающие определить важность различных признаков для обучения модели.

# Введение

В современном обществе дорожно-транспортные происшествия (ДТП) остаются одной из актуальных и сложных проблем, затрагивающих сферу безопасности на дорогах. Однако, наряду с множеством факторов, воздействующих на данное явление, особое внимание уделяется анализу влияния природных условий на возникновение и тяжесть ДТП. Погодные условия, будучи неотъемлемой частью дорожной среды, выступают важным фактором, оказывающим значительное воздействие на безопасность движения.

В рамках курсовой работы я намерен провести анализ воздействия погодных условий на серьезность дорожно-транспортных происшествий, используя методы машинного обучения.

Цель данного исследования заключается в разработке эффективного алгоритма машинного обучения, предназначенного для решения задачи многоклассовой классификации. Данный алгоритм будет направлен на предсказание тяжести ДТП на основе имеющихся данных о погодных условиях.

# Анализ данных

## Подготовка данных к работе

Изначально набор данных состоит из двух Excel файлов. Pogoda.xlsx содержит данные о погоде, а DataDTPкор.xlsx – данные о ДТП. У обоих датасетов есть соединительный столбец join\_key, который позволяет объединить данные о погоде и ДТП в один датафрейм. Для удобства работы столбцы были переименованы в едином стиле, и для быстроты чтения форматы файлов переведены в csv (Рисунок 1).



Рисунок 1

Далее понадобилось рассчитать столбцы, которые в дальнейшем могут положительно повлиять на качество обучения моделей. Оставлены строки о ДТП в светлое время суток, также рассчитаны переменные: праздник (1\0), перепад температур, был ли переход через 0, сезон, час, день недели, день месяца.

В результате датафрейм для анализа данных состоит из 20749 строк и 33 переменных (Таблица 1).

Таблица 1

|  |  |
| --- | --- |
| **Имя переменной** | **Описание** |
| join\_key | Служебный столбец для объединения датасетов. |
| datetime\_weather | Дата и время замера погоды. |
| year\_weather | Год замера погоды. |
| month\_weather | Месяц замера погоды. |
| region\_weather | Регион замера погоды. |
| temperature | Температура. |
| atmospheric\_pressure | Атмосферное давление. |
| humidity | Влажность. |
| wind\_direction | Направление ветра. |
| wind\_speed | Скорость ветра. |
| cloudiness | Облачность. |
| weather\_conditions | Условия погоды. |
| datetime\_round | Округленная дата и время. |
| id | Служебный столбец, уникальный идентификатор ДТП. |
| light | Освещенность. |
| point | Географические данные о широте и долготе места ДТП. |
| pogoda\_region | Регион замера погоды. |
| region\_dtp | Регион, в котором произошло ДТП. |
| address | Адрес ДТП. |
| category | Категория ДТП. |
| datetime\_dtp | Дата и время ДТП. |
| severity | Тяжесть ДТП (целевая метка). |
| dead\_count | Количество погибших. |
| injured\_count | Количество пострадавших. |
| parent\_region | Город ДТП |
| participants\_count | Количество участников ДТП. |
| is\_holiday | Был ли день государственным праздником (1/0) |
| temp\_diff | Разница температур. |
| zero\_cross | Было ли пересечение отметки 0 градусов (1/0). |
| season | Сезон ДТП. |
| hour | Час ДТП. |
| day\_of\_week | День недели ДТП. |
| day\_of\_month | День месяца ДТП. |

## Разведочный анализ

В данном разделе построены визуализации, помогающие понять структуру данных, чтобы в дальнейшем лучше подготовить данные для обучения моделей.

Всего было построено 5 визуализаций. Сначала стоит посмотреть, как ДТП распределены по тяжести, то есть по целевой метке.

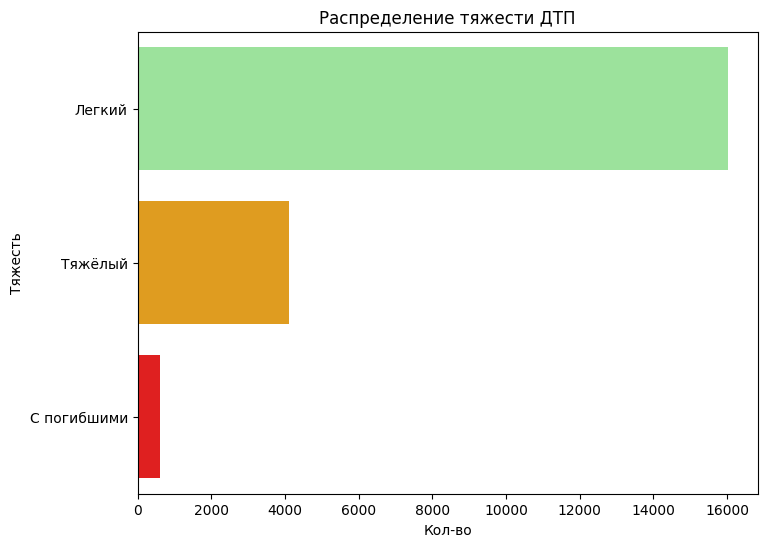


Рисунок 2

Как видно из столбчатой диаграммы (Рисунок 2) наиболее распространены лёгкие ДТП, тяжёлых ДТП в 4 раза меньше, а с погибшими меньше 1000. По этой причине на этапе обработки данных нужно устранить дисбаланс классов.

Ранее для каждой строки была посчитана переменная сезон. Чтобы оценить полезность данной переменной, посмотрим, отличается количество ДТП разной тяжести в разные сезоны, для этого построим столбчатую диаграмму.

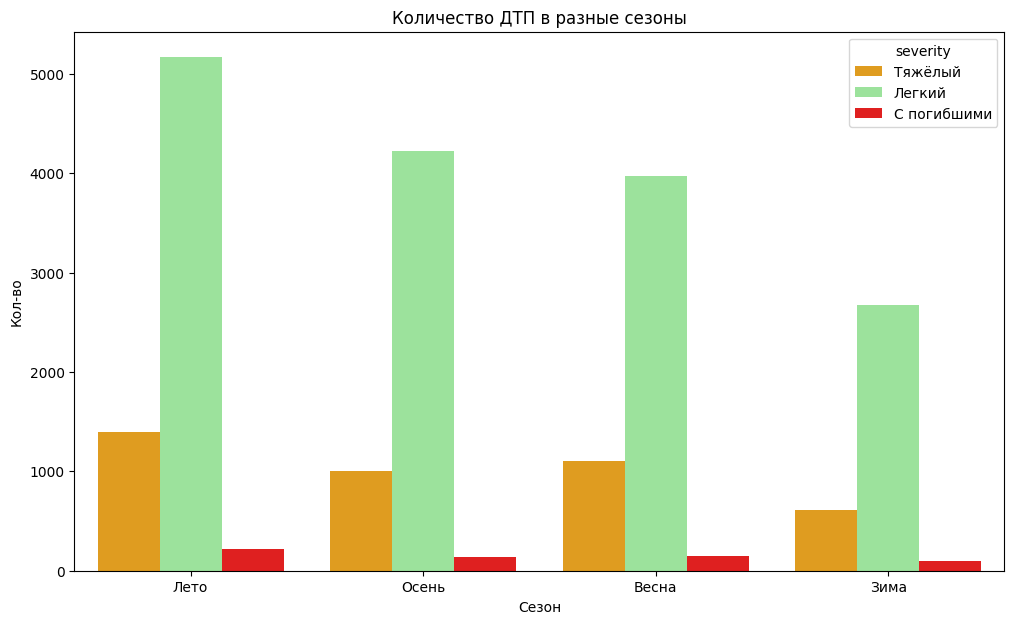


Рисунок 3

Как видно из рисунка 3, летом происходило больше всего ДТП разной тяжести. Также весной больше тяжёлых ДТП, чем осенью и зимой, что может говорить о влиянии погоды на тяжесть ДТП.

Далее посмотрим переменную, которая может лучше всего характеризовать погодные условия, weather\_conditions.

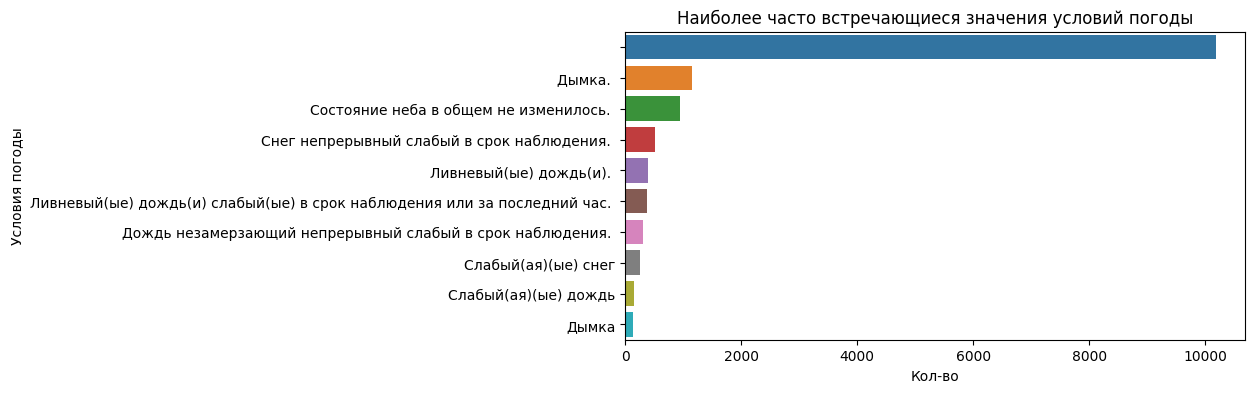


Рисунок 4

Рисунок 4 демонстрирует, что в переменной weather\_conditions большинство значений бесполезны. Также есть повторяющиеся значения, такие как «Дымка.» и «Дымка».

Далее рассмотрим распределение ДТП разной тяжести по часам в сутках.

Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание

Рисунок 5

По диаграмме (Рисунок 5) видно, что большинство ДТП разной тяжести происходит днём, в промежутке с 9 часов до 16 часов. Также можно сказать, что распределения близки к нормальному, однако данные всё равно нужно стандартизировать.

Далее построим бокс плоты, чтобы посмотреть, насколько изменение температур в разных регионах влияет на тяжесть ДТП.

Изображение выглядит как текст, диаграмма, Параллельный, линия

Автоматически созданное описание

Рисунок 6

По бокс плотам (Рисунок 6) можно сделать вывод, что с увеличением тяжести ДТП увеличивается разброс температур. Однако регион незначительно влияет на разницу температур. Во всех регионах разброс изменения температуры находится в пределах от -2 градусов, до 5 градусов.

# Подготовка данных

Начнём этап подготовки данных с удаления ненужных столбцов, многие из которых показали свою ненужность на этапе EDA.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 7

Далее проработаем пропуски и дубликаты, наличие которых снизит качество обучения моделей. Так как пропусков оказалось очень мало, можно просто удалить строки с ними.

Изображение выглядит как текст, снимок экрана, число, программное обеспечение

Автоматически созданное описание

Рисунок 8

После обработки пропусков и дубликатов, поменяем тип данных целевой переменной со строкового на числовой.

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Автоматически созданное описание

Рисунок 9

Выбросы и аномалии могут значительно повлиять на предсказания моделей, поэтому им тоже следует уделить внимание. Далее посмотрим наличие выбросов в разных столбцах с помощью бокс плотов (Рисунок 10) и оставим только данные, входящие в квантиль 0,99.

Изображение выглядит как текст, чек, снимок экрана, число

Автоматически созданное описание

Рисунок 10

После удаления ненужных столбцов в датафрейме осталась одна категориальная переменная, которую нужно закодировать для дальнейшего обучения моделей. В данном случае лучше всего подойдет OneHotEncoder.

Закодировав категориальную переменную, можно построить корреляционную матрицу, чтобы оценить линейную зависимость целевой метки от других столбцов.

Изображение выглядит как снимок экрана, текст, диаграмма, линия

Автоматически созданное описание

Рисунок 11

Как можно заметить по корреляционной матрице (Рисунок 11) данные имеют слабую линейную зависимость.

На этапе EDA был выявлен дисбаланс классов в целевой переменной. Так как количество данных меньшего класса очень малы, меньше тысячи, увеличим их количество одним из методов передискретизации SMOTE (Рисунок 12).

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 12

Разведочный анализ данных также дал понять, что данные нужно нормировать, для этого воспользуемся методом стандартизации StandardScaler.

После обработки данных датафрейм состоит из 19681 строк и 16 столбцов: month\_weather, temperature, atmospheric\_pressure, humidity, wind\_speed, cloudiness, severity, injured\_count, participants\_count, is\_holiday, temp\_diff, zero\_cross, hour, day\_of\_week, day\_of\_month, x0\_Весна, x0\_Зима, x0\_Лето, x0\_Осень.

# Разделение на выборки

Разделение на выборки было произведено с помощью простого holdout метода с сохранением соотношения классов. Однако при подборе гиперпараметров, метрики будут считаться с учетом кросс-валидации.

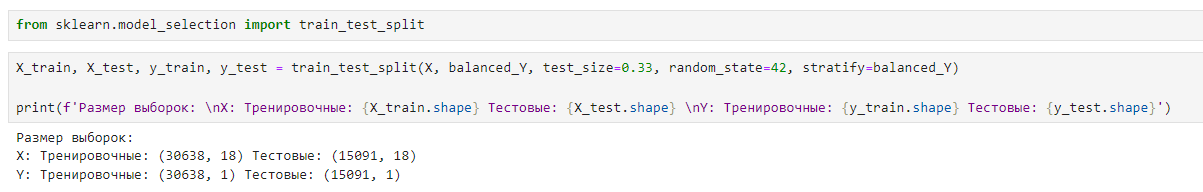


Рисунок 13

# Обучение моделей

## Автоматическое машинное обучение

Для автоматического машинного обучения была использована библиотека mljar, которая не требует долгого подбора гиперпараметров и имеет возможность использовать библиотеку optuna в качестве инструмента подбора гиперпараметров.

Сначала была построена простая baseline модель.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 14

Далее построена модель автоматического машинного обучения с использованием более сложных алгоритмов таких как случайный лес и различные реализации градиентного бустинга. К каждому алгоритму по 300 секунд подбирались гиперпараметры с помощью optuna. В результате лучшим алгоритмом оказался Xgboost с метрикой logloss равной 0,237.

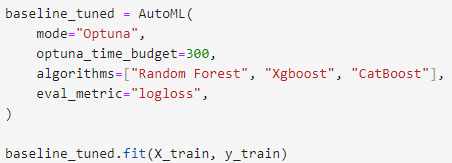


Рисунок 15

## Моделирование

На этапе моделирования были построены модели, решающие задачи многоклассовой классификации. Гиперпараметры подбирались с помощью библиотеки optuna методами grid search и random search, в зависимости от количества параметров. У каждой модели замерялись метрики accuracy, f1 score macro и logloss, также были построены матрицы ошибок.

Далее представлены метрики для всех моделей.

Логистическая регрессия:

Лучшие параметры: {'C': 0.1}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 16

Дерево решений:

Лучшие параметры: {'max\_depth': 8, 'min\_samples\_split': 10, 'min\_samples\_leaf': 15, 'max\_leaf\_nodes': 13, 'criterion': 'gini'}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 17

Случайный лес:

Лучшие параметры: {'n\_estimators': 73, 'max\_depth': 24, 'min\_samples\_split': 3, 'min\_samples\_leaf': 3}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 18

Градиентный бустинг sklearn:

Лучшие параметры: {'learning\_rate': 0.03807947176588889, 'n\_estimators': 98, 'max\_depth': 15, 'min\_samples\_split': 7, 'min\_samples\_leaf': 4, 'max\_features': 'sqrt'}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 19

Градиентный бустинг CatBoost:

Лучшие параметры: {'learning\_rate': 0.05512431765498469, 'n\_estimators': 59, 'depth': 16, 'min\_data\_in\_leaf': 16, 'max\_bin': 10, 'grow\_policy': 'Lossguide'}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 20

Градиентный бустинг Xgboost:

Лучшие параметры: {'max\_depth': 6, 'learning\_rate': 0.0420167205437253, 'n\_estimators': 284, 'min\_child\_weight': 6, 'gamma': 3.0118659882617117e-07, 'subsample': 0.9698887814869129, 'colsample\_bytree': 0.7773814951275034, 'reg\_alpha': 0.32808889626606236, 'reg\_lambda': 0.14408501080722544}

Метрики:

Изображение выглядит как текст, Шрифт, белый, алгебра

Автоматически созданное описание

Рисунок 21

## Лучшая модель

В результате лучшей моделью, согласно метрикам (Рисунок 19), оказалась модель градиентного бустинга библиотеки sklearn.

Матрица ошибок для данной модели:

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Рисунок 22

Как можно заметить по матрице ошибок (Рисунок 22) модель достаточно хорошо определяет ДТП с погибшими и лёгкие ДТП.

# Интерпретация

## Глобальная интерпретация

Для лучшей модели были построены графики интерпретаций с помощью partial dependence plot, который отображает, насколько различные переменные влияют на предсказание.

Изображение выглядит как текст, диаграмма, План, Параллельный

Автоматически созданное описание

Рисунок 23

Изображение выглядит как текст, диаграмма, План, Параллельный

Автоматически созданное описание

Рисунок 24

Изображение выглядит как текст, диаграмма, План, Параллельный

Автоматически созданное описание

Рисунок 25

По графикам интерпретаций видно, что почти все переменные влияют на результат предсказания, однако самыми важными переменными являются wind\_speed, hour, cloudiness и injured\_count.

## Локальная интерпретация

Далее с помощью метода lime была проведена локальная интерпретация. Были проверены строки, где скорость ветра близка к 0 (Рисунок 26), значение часа близко к 2 (Рисунок 27) и количество пострадавших близко к 0.5 (Рисунок 28). Выбраны именно эти строки, так как на графиках глобальных интерпретаций в значениях данных переменных происходили интересные изменения на графиках.

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Рисунок 26

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Рисунок 27

Изображение выглядит как текст, снимок экрана, диаграмма, программное обеспечение

Автоматически созданное описание

Рисунок 28

Графики локальной интерпретации подтвердили важность переменных wind\_speed, hour, cloudiness и injured\_count. Также можно заметить, что сильное влияние оказывает переменная zero\_cross.

# Выводы

В ходе работы было построено множество моделей, решающих задачу многокассовой классификации, однако лучшей оказалась модель градиентного бустнига в реализации sklearn.

По метрикам моделей видно, что можно достаточно точно предсказывать тяжесть ДТП, основываясь на данных о погоде, так как минимальная точность моделей выше 0,7, а максимальная больше 0,9.

Кроме того, выяснились основные факторы погоды, влияющие на тяжесть ДТП, к ним относятся: скорость ветра, облачность и пересечение отметки 0 градусов.

# Приложение

GitHub с файлом кода: <https://github.com/Ho1ms/ml-course-work>