Кодбуки - Трек # 3

В архиве находятся следующие датасеты:

- `train.csv` — основной датасет: список ДТП на трассах М-4 Дон, М-8 «Холмогоры», М-18 «Кола» по данным ГИБДД и МЧС;
- `test.csv` — тестовый датасет с пропущенным целевым признаком;
- `traffic.csv` — информация об интенсивности движения на трассах М-8 «Холмогоры» и М-18 (Р-21) «Кола» по данным Росавтодора (файл размещен в архиве *.rar);
- `repair.csv` — информация о ремонтных работах на участках трасс М-8 «Холмогоры» и М-18 (Р-21) «Кола» по данным Росавтодора;
- `crash_parts.csv` — информация об аварийно опасных участках трасс М-8 «Холмогоры» и М-18 (Р-21) «Кола» по данным Росавтодора;
- `atmos.csv` – информация Росгидромета об атмосферных явления, собранная на станциях вдоль трасс М-8 и Р-21;
- `meteo.csv` – информация Росгидромета о метеорологической обстановке, собранная на станциях вдоль трасс М-8 и Р-21;
- `tele2_data.csv` – информация Теле2 о количестве вызовов на экстренные номера и количестве населения, передвигающегося по участку дороги;
- `geo_data.csv` - таблица, сопоставляющая километры а/дороги с координатами из Геокодера и ГЛОНАСС.

Большинство данных зависят от времени и покрывают временной период с `2012-01-01` по `2020-12-31`.

Ниже описание полей датасетов.

train.csv — список ДТП на трассах М4, М8 и М18

- `lat` (FLOAT) — широта, градусы.
  - Пример: 39.974849
- `lon` (FLOAT) — долгота, градусы.
  - Пример: 68.974849
- `datetime` (DATETIME) — дата события в формате `yyyy-MM-dd hh-mm-ss`.
  - Пример: 2013-10-09 11:00:00.
- `road_id` (INT) — код автодороги. В датасете атрибут принимает одно из трех значений:

  - 5 — М-4 (Е115) “Дон” Москва - Новороссийск;
  - 9 — М-8 (Е115) “Холмогоры” Москва - Архангельск через Ярославль, Вологду;
  - 14 — М-18 (Е105) “Кола” С.Петербург - Мурманск через Петрозаводск.

- `road_km` (INT) — км участка на котором произошло ДТП.
  - Пример: 16.
- `man_injured_num` (INT) — количество пострадавших.
  - Пример: 1.
- `man_dead_num` (INT) — количество погибших.
  - Пример: 0.
- `car_damaged_num` (FLOAT) — количество машин участников ДТП.
  - Пример: 1.
- `data_source` (TEXT) — источник данных события. В датасете атрибут принимает одно из двух значений:

  - gibdd - ГИБДД;
  - gochs - МЧС.

- `road_name` (TEXT) — Название федеральной автомобильной дороги.
  - Пример: М-4 (Е115) =Дон= Москва - Новороссийск.
- `target` (INT) — целевой признак является ли событие ДТП. В датасете атрибут принимает одно из четырех значений:
  - 0 - событие не является ДТП;

    - 1 - ДТП без пострадавших;
    - 2 - ДТП с пострадавшими;
    - 3 - ЧС (объявляется от 5 погибших и/или 10 пострадавших)* в тестовой выборке события отсутствуют.

test.csv — тестовый датасет с пропущенным целевым признаком

В тестовом датасете находится список событий с 2020-01-01 по 2020-12-31 для которых требуется построить прогноз

- `datetime` (DATETIME) — дата события в формате `yyyy-MM-dd hh-mm-ss`.
  - Пример: 2013-10-09 11:00:00.
- `road_id` (INT) — код автодороги. В датасете атрибут принимает одно из трех значений:

  - 5 — М-4 (Е115) =Дон= Москва - Новороссийск;
  - 9 — М-8 (Е115) =Холмогоры= Москва - Архангельск через Ярославль, Вологду;
  - 14 — М-18 (Е105) =Кола= С.Петербург - Мурманск через Петрозаводск.

- `road_km` (INT) — км участка на котором произошло ДТП.
  - Пример: 16.
- `target` (INT) — целевой признак является ли событие ДТП. В датасете атрибут принимает одно из четырех значений:

- 0 - событие не является ДТП;
- 1 - ДТП без пострадавших;
- 2 - ДТП с пострадавшими;
- 3 - ЧС (объявляется от 5 погибших и/или 10 пострадавших)* в тестовой выборке события отсутствуют.

traffic.csv — датасет с характеристиками загруженности движения

Датасет включает в себя данные наблюдений пунктов учета интенсивности движения (ПУИД) за временной период с `2016-01-01` по `2021-04-14`.

Описание переменных:

- `datetime` (DATETIME) - Данные о дате измерения в формате гггг-мм-дд чч:мм:сс.
  - Пример: 2021-04-14 07:59:59
- `road_id` (INT) - ID автомобильной дороги. В датасете переменная принимает 2 значения:
  - 9 — М-8 (Е115) "Холмогоры" Москва - Архангельск через Ярославль, Вологду";
  - 14 — Р-21 (Е105) "Кола" Санкт-Петербург - Мурманск через Петрозаводск.
- `road_km` (INT) - километр автомобильной дороги в соответствии с разметкой километража по направлению движения.
- `name` (TEXT) - расположение данного участка по километрам рассматриваемой трассы
  - Пример: "км 180+700", где 180 - номер километра, а 700 - метры от последнего пройденного километра.
- `data_id` (INT) - представляет идентификационный номер (ID) конкретного измерения
  - Пример: 9505362.
- `station_id` (INT) - ID пункта учета интенсивности движения (ПУИДа), с использованием которого и было получено измерение.
- `direction` (TEXT) - направление движения транспортного средства по данному наблюдению. Представлена 2 значениями:
  - forward - при движении от нулевого километра;
  - backward - при движении к нулевому километру.
- `lane_count` (INT) - общее количество полос на участке автомобильной дороги по отношению к расположению ПУИДа. Как и переменная lane принимает целые значения от 1 до 6.
- `lane` (INT) - Номер полосы, по которой осуществляется движение.

    - В датасете представлена целыми значениями, от 1 до 6.
    - Значение для данной переменной определяется для наблюдения по полосе и направлению.

- `volume` (FLOAT) - значение для общей интенсивности на дороге, определяемого как общее количество транспортных средств, пересекших конкретное сечение автодороги за единицу времени по соответствующему направлению движения.
    - Пример: 240.
- `occupancy` (FLOAT) - значение загрузки полосы в процентах.

    - В датасете принимает значения от 0 до 100.

- `speed` (FLOAT) - средняя скорость на данном участке дороге, в км/ч за время наблюдения.
- `latitude` (FLOAT) - координата широты для данного километра дороги в градусах.
    - Пример: 60.000762.
- `longitude` (FLOAT) - координата долготы для данного километра дороги в градусах.
    - Пример: 41.072852.

crash_parts.csv — таблица с характеристиками из Формы 7а по аварийно опасным участкам

Данные таблицы представляют сведения из Формы 7а по существующим аварийно опасным участкам автомобильной дороги. Таблица покрывает период с `2015` по `2020` год включительно.

`datetime` (DATETIME) - Данные о дате измерения в формате гггг-мм-дд.

    - Пример: 2017-01-01.

`road_id` (INT) - идентификационный номер автомобильной дороги. Принимает два целочисленных значения:

- 9 — М-8 (Е115) "Холмогоры" Москва - Архангельск через Ярославль, Вологду";
- 14 — Р-21 (Е105) "Кола" Санкт-Петербург - Мурманск через Петрозаводск”.

`road_km` (INT) - километр автомобильной дороги в соответствии с разметкой километража по направлению движения.

    - Пример: 311.

`avuch_start` (FLOAT) - обозначение точки начала аварийного участка с точностью до метра.

- Пример: 25.18.

`avuch_end` (FLOAT) - обозначение точки окончания аварийного участка с точностью до метра.

- Пример: 1180.32.

`length` (INT) - протяженность аварийно опасного участка в метрах. Тут же представлены значения для идентификации особенных случаев, в частности:

- 1 - транспортная развязка;
- 2 - пересечение;
- 3 - примыкание.

`avuch_loc` (INT) - местоположение аварийного участка. Принимает два целочисленных значения:

- 1 - в пределах населенного пункта;
- 2 - вне пределов населенного пункта.

`stabchar_type` (INT) - характеристика стабильности местоположения участка концентрации ДТП. Принимает следующие 3 значения:

- 1 - Стабильный;
- 2 - Мигрирующий (подразумевается варьирующий характер частоты ДТП на данном участке при сравнении по годам);
- 3 - Вновь возникший (подразумевает возникновение нескольких аварийно опасных ситуаций в рамках одного года).

`planactiv_type` (INT) - код планируемого/планируемых работ на данном участке. Принимает следующие 4 значения:

- 1 - реконструкция;
- 2 - капитальный ремонт;
- 3 - ремонт;
- 4 - содержание.

`planactiv_descr` (TEXT) - подробное описание планируемого/планируемых работ на данном участке.

- Пример: "Дублирование дорожных знаков 3.20 «Обгон запрещен» на щитах с флуорисцентной пленкой желто-зеленого цвета. Щит аварийно опасный участок."

`planactiv_year` (DATETIME) - планируемая дата окончания работ на данном участке.

- Пример: 2018-07-01.

repair.csv - сведения о проводимых ремонтных работах

Данные таблицы представляют сведения о проводимых в рассматриваемый период ремонтных работах на участках автомобильной дороги. Таблица покрывает период с `2015` по `2020` год включительно.

`datetime` (DATETIME) - Данные о дате измерения в формате гггг-мм-дд чч:мм:сс.

- Пример: 2016-01-01.

`road_id` (INT) - идентификационный номер автомобильной дороги. Принимает два целочисленных значения:

- 9 — М-8 (Е115) "Холмогоры" Москва - Архангельск через Ярославль, Вологду";
- 14 — Р-21 (Е105) "Кола" Санкт-Петербург - Мурманск через Петрозаводск".

`road_km` (INT) - километр автомобильной дороги в соответствии с разметкой километража по направлению движения.

- Пример: 151.

`repair_id` (INT) - идентификационный номер проведенных ремонтных работ.

- Пример: 45447320152017.

`repair_description` (TEXT) - наименование ремонтных работ согласно отчетным данным.

- Пример: "Капитальный ремонт автомобильной дороги М-8 "Холмогоры" Москва-Ярославль-Вологда-Архангельск на участке км 640+000 - км 670+000 в Вологодской области".

`repair_period` (TEXT) - период, в который проходили ремонтные работы.

- Пример: "2015 — 2016".

`length` (FLOAT) - протяженность отремонтированного участка в километрах.

- Пример: 11.982.

`price` (FLOAT) - стоимость ремонтных работ, тыс. руб.

- Пример: 1349782,841.

atmos.csv — датасет с характеристиками атмосферных явлений

> **Важно:** поскольку данные Росгидромета (atmos.csv и meteo.csv) сгруппированы по метеостанциям, а не километрам трассы, то для их использования к наблюдениям обучающей и тренировочной выборок нужно добавить название ближайшей метеостанции. Скрипт как это можно сделать (с примером) лежит в ноутбуке match_station.ipynb.

Датасет включает в себя данные наблюдений с `1 января 2015` по `31 декабря 2020.`

Описание переменных:

`road_id` (INT) – ID автомобильной дороги. В датасете переменная принимает 2 значения:

- 9 — М-8 (Е115) "Холмогоры" Москва - Архангельск через Ярославль, Вологду";
- 14 — Р-21 (Е105) "Кола" Санкт-Петербург - Мурманск через Петрозаводск.
    - Пример: “Р-21“.

`station` (TEXT) – Буквенное обозначение станции.

- Пример: “АРАТИТ“.

`lat` (FLOAT) – Географическая широта местонахождения станции.

- Параметр определён не для всех записей.
    - Пример: 56.4.

`lon` (FLOAT) – Географическая долгота местонахождения станции.

- Параметр определён не для всех записей.
    - Пример: 41.646667.

`phenomenon` (TEXT) – Название атмосферного явления.

- Параметр определён не для всех записей.
    - Пример: “туман поземный“.

`intensity` (TEXT) – Обозначение интенсивности или силы атмосферного явления.

- Параметр определён не для всех записей.

- Пример: “слабая интенсивность“.

`start_date` (DATETIME) – Дата начала атмосферного явления.

- Пример: “2017-11-17“.

`start_ts` (TEXT) – Время начала атмосферного явления.

- Параметр определён не для всех записей.
    - Пример: “12:00“.

`end_date` (DATETIME) – Время окончания атмосферного явления.

- Пример: “2018-06-08“.

`start_ts` (TEXT) – Время окончания атмосферного явления.

- Параметр определён не для всех записей.
    - Пример: “19:12“.

Кроме того, в датасете присутствуют поля `phenomenon_q`, `intensity_q`, `start_q` и `end_q`, характеризующие достоверность соответствующих данных.

Принимают значения: `Значение элемента достоверно` или `Значение элемента забраковано на станции`.

meteo.csv — датасет с характеристиками метеорологических явлений

Датасет включает в себя данные наблюдений с `1 января 2015` по `31 декабря 2020.`

`road_id` (INT) - ID автомобильной дороги. В датасете переменная принимает 2 значения:

- 9 — М-8 (Е115) "Холмогоры" Москва - Архангельск через Ярославль, Вологду";
- 14 — Р-21 (Е105) "Кола" Санкт-Петербург - Мурманск через Петрозаводск.
    - Пример: “Р-21“.

`station` (TEXT) - Буквенное обозначение станции.

- Пример: “АРАТИТ“.

`lat` (FLOAT) – Географическая широта местонахождения станции.

- Параметр определён не для всех записей.
    - Пример: 56.4.

`lon` (FLOAT) – Географическая долгота местонахождения станции.

- Параметр определён не для всех записей.
    - Пример: 41.646667.

`measure_dt` (DATETIME) – Дата и время фиксации параметров атмосферных явлений.

- Пример: “2017-11-17“.

`vsp_1` (INT) - Высота снежного покрова в сантиметрах на снегомерной рейке №1.

- Пример: 42.

`vsp_2` (INT) - Высота снежного покрова в сантиметрах на снегомерной рейке №2.

- Пример: 42.

`vsp_3` (INT) - Высота снежного покрова в сантиметрах на снегомерной рейке №3.

- Пример: 42.

`visib` (INT) – Горизонтальная дальность видимости. Это то наибольшее расстояние, с которого в светлое время суток перестает быть видимым абсолютно черный объект размером более 15′, проектирующийся на фон неба у горизонта. Дальность видимости является показателем оптического состояния атмосферы. На метеорологических станциях измерение МДВ производится с помощью приборов, а в их отсутствие – визуально с помощью специально выбранных ориентиров. Горизонтальная дальность видимости приводится в цифрах кода. При инструментальном способе измерения используются цифры от 00 до 89, за исключением 51-55, а при визуальном – от 90 до 99.

- Коды обозначают следующее:
    - 00 – менее 0,1 км;
    - 01-50 – указывают видимость в десятых долях км, т.е от 0,1 км до 5,0 км;
    - Например, 25 = 2,5 км
    - 51-55 – не используются;
    - 56-80 –видимость от 6 до 30 км с шагом в 1км. Видимость в целых км может быть определена вычитанием 50 из кода, т.е. цифра кода 65 означает горизонтальную видимость в 15км;

- 81-88 – видимость от 35 до 70 км с шагом в 5км;
- 89 – видимость более 70 км;
- 90 – видимость менее 0,05 км;
- 91 – видимость 0,05 км;
- 92 – 0,2 км;
- 93 – 0,5 км;
- 94 – 1 км;
- 95 – 2 км;
- 96 – 4 км;
- 97 – 10 км;
- 98 – 20 км;
- 99 – более 50 км.

`clouds` (INT) – Общее количество облачности и количество облаков нижнего яруса.Оценивается визуально как степень покрытия небосвода облаками по 13-бальной шкале. Кодируется в баллах от 0 до 13.

- 0 означает полное отсутствие облаков или покрытие облаками менее 1/10 небосвода, а значение 10 означает, что небосвод полностью покрыт облаками;
- 11 обозначает наличие следов облаков;
- 12 – 10 баллов с просветами;
- 13 – облака невозможно определить.

`weather_range` (TEXT) – Погода в течение трёх часов, предшествующих сроку наблюдения.

  - Пример: “Ливневые осадки”.

`weather_on_measure` (TEXT) – Погода в срок наблюдения или в течение последнего часа перед сроком наблюдения.

  - Пример: “Небо без изменений”.

`wind_dir` (INT) – Направление ветра в градусах.

- Штиль кодируется одной цифрой 0, а переменное направление – 999.

  - Пример: 316.

`avg_wind` (INT) – Средняя скорость ветра (м/сек).

  - Пример: 12.

`max_wind` (INT) – Максимальная скорость ветра (в м/сек) за 3 часа, включая порывы.

  - Пример: “APATIT“.

precip (FLOAT) – Сумма осадков за период между сроками, когда измеряются осадки, в мм с точностью до десятых долей.

- Пример: 1.7.

temp_on_measure (FLOAT) – Температура воздуха во время замера в градусах Цельсия.

- Пример: 31.

temp_min (FLOAT) – Минимальная температура воздуха между замерами в градусах Цельсия.

- Пример: 24.

temp_max (FLOAT) – Максимальная температура воздуха между замерами в градусах Цельсия.

- Пример: 42.

humidity (INT) – Относительная влажность воздуха во время замера в процентах.

- Пример: 85.

pressure (FLOAT) – Атмосферное давление во время замера на уровне станции в гПа.

- Пример: 990.5.

Кроме того, в датасете присутствуют поля vsp_1_q, vsp_2_q, vsp_3_q, visib_q, clouds_q, weather_range_q, weather_on_measure_q, wind_dir_q, avg_wind_q, max_wind_q, precip_q, temp_on_measure_q, temp_min_q, temp_max_q, humidity_q и pressure_q, характеризующие достоверность соответствующих данных.

Принимают значения: Значение элемента достоверно, Значение элемента достоверно и восстановлено автоматически, Значение элемента достоверно и восстановлено вручную, Значение элемента забраковано на станции или Значение элемента отсутствует.

tele2_data.csv — датасет с данными теле2

Информация Теле2 о количестве вызовов на экстренные номер и количестве населения, передвигающегося по участку дороги.

`datetime` (DATETIME) — дата события в формате `yyyy-MM-dd hh-mm-ss`. Пример: 2013-10-09 11:00:00.

`road_id` (INT) — код автодороги. В датасете атрибут принимает одно из трех значений:

- 9 — М-8 (Е115) «Холмогоры» Москва - Архангельск через Ярославль, Вологду;
- 14 — М-18 (Е105) «Кола» С.Петербург - Мурманск через Петрозаводск.

`road_km` (INT) — км участка на котором произошло ДТП.

  - Пример: 16.

`cnt_subs` (INT) — количество населения, передвигавшегося по сегменту дороги по временному интервалу.

geo_data.csv - таблица, сопоставляющая километры а/дороги с координатами

Датасет включает информацию о координатах для каждого километра трасс М-8 Холмогоры и М-18 (Р-21) Кола. Информация была получена с использованием двух источников: Геокодера Яндекса и данных ГЛОНАСС. Источники дифференцированы по полноте представления координат километров трасс и предлагаются к использованию как альтернативные варианты.

Значения координат представлены для абсолютного большинства километров (столбов) трасс.

ОГРАНИЧЕНИЯ ДАННЫХ:

1. Имеют место пропуски данных в отношении координат ряда километров. Основная причина наличия пропусков - непоследовательное обозначение километров, что может объяснятся:

- ренумерацией километровых столбов (например, после 520-ого километра а/дороги Р-21 на картах отображается 565-ый километр);

- отсутствием обозначений километров трасс на участках, проходящих по территории городов и поселений (например, 457-ой километр а/дороги переходит в Окружное шоссе при городе Вологде; следующее значение по столбам - 472 километр, на выезде с территории города).

2. 112 км а/дороги М-8 и 260 км а/дороги Р-21 имеют по две координаты:

- в первом случае (112 М-8) это объясняется параллельным расположением столбов на смежных дорогах;

- во втором случае (260 Р-21) это объясняется расположением километра на границе субъектов (Ленинградская область и Республика Карелия), имеющих свои обозначения 260-го километра трассы Р-21.

ПЕРЕМЕННЫЕ:

`road_id`(INT) — код автодороги. Атрибут принимает одно из 2 значений:

- 9 — М-8 (Е115) «Холмогоры» Москва - Архангельск через Ярославль, Вологду;
- 14 — М-18 (Е105) «Кола» С.Петербург - Мурманск через Петрозаводск.

`road_km`(INT) — километр а/дороги.

  - Пример: 127.

`km_name`(TEXT) - наименование присвоенное километровому столбу а/дороги.

  - Пример: Р-21 Кола, 1482-й километр.

`lat_geoc`(FLOAT) - значение координат широты, полученные с использованием Геокодера.

  - Пример: 69.622973.

`lon_geoc`(FLOAT) - значение координат долготы, полученные с использованием Геокодера.

  - Пример: 30.231478.

`lat_glonass`(FLOAT) - значение координат широты, полученные из данных ГЛОНАСС.

  - Пример: 56.546578.

`lon_glonass`(FLOAT) - значение координат долготы, полученные из данных ГЛОНАСС.

  - Пример: 38.594389.