ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В странах с большой железнодорожной сетью и большим потоком перемещения поездов, таких как РФ, США, Китай, Индия существует проблема схода составов с рельс, которые могут быть обусловленны различными факторами, их можно классифицировать на:

- внешние: кривизна пути, профиль пути, состояние транспортного пути, проблемы со стрелочным переводом, погодные условия (при экстремальных температурах рельсы могут сильно расширяться или сжиматься);
- внутренние: количество вагонов в составе, загруженность, скорость, невнимательность машиниста, состояние состава.

Некоторые пути могут проходить через национальные парки, национальные заповедники и другие типы особо охраняемых объектов. По этой причине аварии, произошедшие на таких участках могут привести к экологической катастрофе, особенно велика опасность, если поезд был грузовым и перевозил легко воспламеняемые объекты (нефть, газ, метан, уголь, древесина) или высокотоксичные грузы. Следует отметить, что помимо экологической проблемы могут возникнуть и другие проблемы, например, такие как:

- логистическая - если состав сошел с рельс, следующим поездам приходится идти в обход, в некоторых случаях обхода может не быть;
- экономическая - связанна с издержками транспортной компании по решению экологической проблемы, потери части вагонов, локомотива, утрата части груза, временные издержки;
- инфраструктурная - повреждение строения железнодорожного пути, стыков, моста, обрушение тоннеля и др.

В данной работе рассматривается проблема схода состава с рельс, поскольку данная проблема является одной из самых опасных. В зависимости от масштаба происшествия сходы классифицируют на аварии и крушения. Согласно [4] за период с 2013 г. по 2016 г. в Российской Федерации имеется 262 протокола сходов с рельс вагонов как в грузовых поездах, так и в пассажирских поездах, без учета протоколов транспортных происшествий,

классифицированных как крушения. Соответственно, при вычислении среднего числа дней без аварий выходит 4 дня, поэтому проблема представляет интерес для железнодорожных компаний.

В данной работе будет проведен анализ причин схода железнодорожного подвижного состава, а также будут построены предсказательные модели числа сошедших вагонов. Для достижения поставленных задач будут использованы методы теории вероятностей и математической статистики.

ГЛАВА 1

Обзор предметной области

1.1 Тема 1

1.2 Тема 2

1.3 Тема 3

ГЛАВА 2

Предварительный анализ данных

2.1 Структура данных

В данном наборе данных представлена информация о случаях схода составов с рельс по причине излома боковой рамы вагона

Анализ данных будет проведен при помощи языка программирования Python3. Выбор пал на данный язык по нескольким причинам:

- большое количество модулей для анализа данных
- удобство и простота работы с данными в форматах csv, xlsx
- множество встроенных функций и выразительность языка

Определим размеры выборки:

```
print("shape of data frame:", df.shape)
```

```
> shape of data frame: (56, 12)
```

Выведем названия факторов:

```
print(df.columns)
```

```
> Index(['Дата', 'Количество вагонов', 'Макс. число вагонов в сходе',
'Общее количество вагонов', 'Количество сшедших вагонов', 'Скорость',
'Вес', 'Загрузка', 'Стрелочный перевод', 'Кривизна', 'Профиль пути',
'Режим движения'],
dtype='object')
```

Получим первые 5 записей из набора:

№	Дата	Количество вагонов	Макс. число вагонов в сходе	Общее количество вагонов	Количество сшедших вагонов	Скорость	Вес	Загрузка	Стрелочный перевод	Кривизна	Профиль пути	Режим движения
1	2013-01-08	56.0	19.0	58.0	1	57.0	3402.0	0.547101	0	0.000000	0.0007	NaN
2	2013-01-09	60.0	25.0	62.0	1	72.0	4082.0	0.652657	0	0.000000	0.0009	NaN
3	2013-01-10	60.0	4.0	64.0	1	15.0	4420.0	0.734300	0	0.001639	NaN	3.0
4	2013-01-12	66.0	63.0	68.0	21	67.0	5699.0	0.918094	0	0.002326	0.0060	NaN
5	2013-01-19	67.0	34.0	69.0	1	69.0	5854.0	0.932944	0	0.000000	0.0006	2.0

Таблица 2.1 — первые 5 записей в наборе данных

Получим основные статистики по данным с помощью команды `print(df.describe())` (для краткости названия признаков заменены на f1, f2, . . . , f11, признак "Дата"не рассматривается).

	f1	f2	f3	f4	f4	f6	f7	f8	f9	f10	f11
count	54.000000	51.000000	54.000000	56.000000	53.000000	54.000000	54.000000	56.000000	46.000000	44.000000	33.000000
mean	63.870370	37.137255	66.407407	3.875000	49.150943	5126.629630	0.817678	0.107143	0.000806	-0.000384	1.666667
std	9.790342	21.543463	10.053665	6.081455	18.450971	1438.743887	0.243936	0.312094	0.001171	0.005689	0.777282
min	24.000000	2.000000	26.000000	1.000000	9.000000	998.000000	0.179710	0.000000	0.000000	-0.011500	1.000000
25%	60.000000	17.500000	62.500000	1.000000	35.000000	4155.500000	0.690451	0.000000	0.000000	-0.004750	1.000000
50%	66.000000	43.000000	68.000000	1.000000	51.000000	5722.000000	0.925519	0.000000	0.000000	0.000000	1.000000
75%	68.000000	56.500000	71.750000	2.250000	64.000000	6010.250000	0.995586	0.000000	0.001479	0.001875	2.000000
max	96.000000	72.000000	100.000000	26.000000	78.000000	8806.000000	1.076087	1.000000	0.005000	0.010900	3.000000

Таблица 2.2 — основные статистики

Заметим, что в данных есть пропуски, так признак "Режим движения"(f11) содержит только 33 записи. Также много пропусков у признаков "Кривизна"(f9) и "Профиль пути"(f10).

Построим матрицу корреляции признаков:

```
corrmat = df.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, square=True)
```

Из матрицы видно, что признаки "Количество вогонов"и "Общее число вагонов"имеют сильную корреляцию. Также "Вес"и "Загрузка"сильно коррелируют. Менее сильная корреляция наблюдается у признаков "Вес"и "Общее число вагонов". Также заметим, что у "Профиль пути"и "Режим движения"наблюдается сильная обратная корреляция. Многие зависимости можно нетрудно объяснить: чем больше вагонов в составе, тем больше вес, чем больший вес, тем, как правило, большая загруженность. Таким образом, можно прийти к выводу, что в данные в наборе избыточны, поскольку несколько признаков несут одинаковое количество информации. Поэтому эти зависимости приводят к проблеме мультиколлинеарности, что приведет к эффекту переобучения в линейных моделях. Для решения данной проблемы нужно исключить коррелирующие признаки, и, возможно, добавить

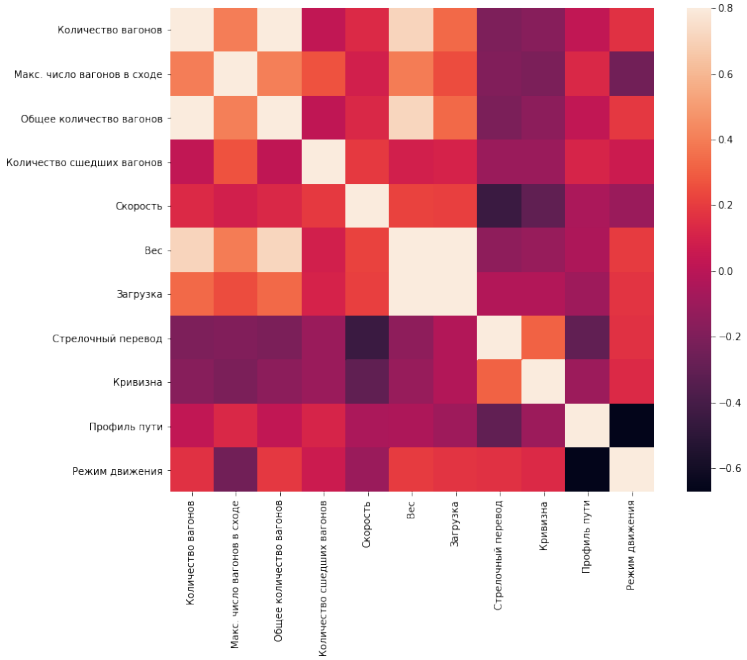

Рисунок 2.1 — Корреляция признаков

новые. Решение проблемы мультиколлинеарности смотри в главе "Линейная регрессия".

2.2 Пропуски в данных

Из таблицы 2.2 видно, что в последних четырех признаках присутствуют пропуски в данных.

Существует методы по решению проблемы с пропусками в данных:

- удалить все записи в которых есть хотя бы одно пустое поле. При использовании этого метода для данного набора данных существует риск того, что оставшегося множества записей не хватит для получения приемлемого качества построенной модели;
- заменить пропуски на средние значение по признаку;
- заменить пропуски на медианные значение по признаку. В отличие от среднего значения замена на медианное позволяет избежать сильного влияния выбросов на итоговое значение.

При решении задачи будут поочередно использованы все 3 метода борьбы с

пропусками, предпочтение будет отдаваться тем моделям, у которых будут более лучшие показатели метрик качества.

2.3 Экстремальные значения

Для поиска выбросов построим графики, изображающие отношения между парами признаков.

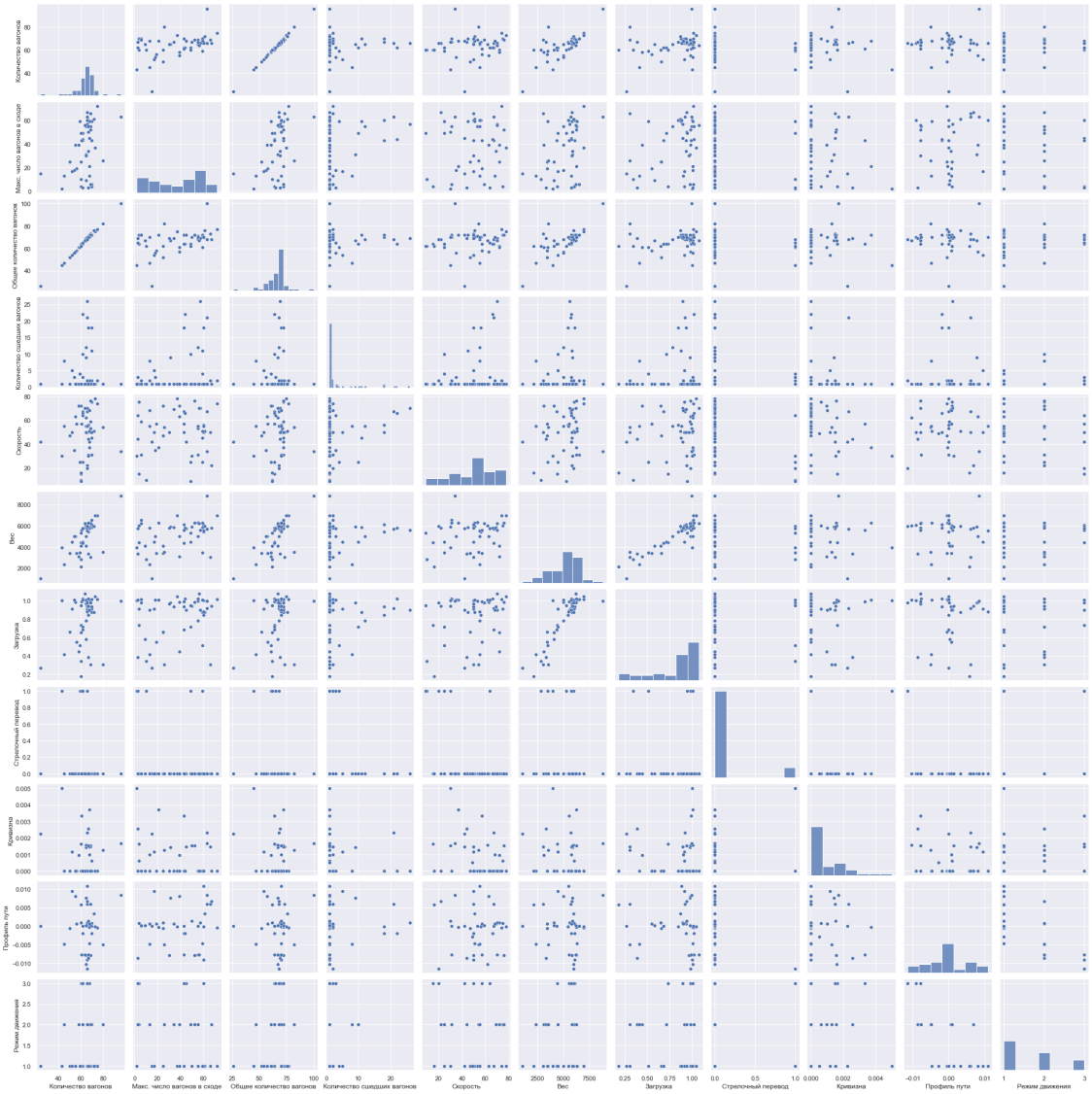

Рисунок 2.2 — Пары признаков

Изучив таблицу 2.2, а также при детальном рассмотрении графиков 2.2 выбросов в данных не обнаружено.

2.4 Кластеризация данных

ГЛАВА 3

Методы решения задачи

3.1 Линейная регрессия

3.1.1 Достоинства и недостатки ЛР для данной задачи

3.2 Пуассоновская регрессия

3.3 Геометрическая регрессия

ГЛАВА 4

Сравнительный анализ

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Andrew Ng., Machine Learning от Stanford University. https://www.coursera.org/learn/machine-learning
2. Воронцов К.В., Введение в машинное обучение от НИУ ВШЭ & Yandex School of Data Analysis. https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie
3. Пуассоновская регрессия. https://en.wikipedia.org/wiki/Poisson_regression
4. Замышляев А.М., Игнатов А.Н., Кибзун А.И., Новожилов Е.О. Функциональная зависимость между количеством вагонов в сходе из-за неисправностей вагонов или пути и факторами движения // Надежность. 2018. Т. 18, № 1. С. . . .. DOI: 10.21683/1729-2646-2018-18-1. . .

Список иллюстраций

Список таблиц

Список программных листингов