

Разработка программного решения для прогноза площади и длительности пожаров в России

Батыршин К.

Сухинин В.

ПРО-ТРСИИ-102М

Анализ проблемы

Россия занимает первое место в мире по площади сгоревшего леса.

Также для снижения рисков при пожаре полезно знать возможные масштабы его распространения а также время, которое может потребоваться для ликвидации.



Содержательная постановка задачи

Имеем: набор данных из 1 459 464 значений, предоставляющий информацию об отдельных очагах пожаров:

- точке регистрации (приведены ее координаты и положение относительно ближайшего населенного пункта, а также сопутствующие территориальные данные),
- датах первого и последнего наблюдения,
- площади, уже пройденной огнём на момент регистрации пожара.

Данные доступны за период с 2000 года по 12 сентября 2024 года.

Требуется: разработать предсказательную систему, предоставляющую следующие возможности:

- формирование предсказаний возможной площади пожара,
- формирование предсказаний возможной длительности пожара.

Состав программных средств разработки и методов обучения

Предобработка, анализ данных, обучение модели:

- Python, Jupyter Notebook.

Библиотеки:

- pandas
- numpy
- matplotlib
- sklearn
- catboost
- lightgbm

Разработка интерфейса:

- C#, Visual Studio 2022.
Windows Forms

Рассмотренные алгоритмы и методы обучения:

1. Линейная регрессия sklearn - довольно простая модель линейной регрессии;
2. Логистическая регрессия sklearn - довольно простая модель логистической регрессии;
3. Алгоритм Cat-Boost - алгоритм, способный принимать категориальные параметры без их предварительной категоризации. Устойчив к переобучению. В основе - градиентный бустинг;
4. Алгоритм LightGBM - алгоритм, показывающий очень высокую производительность. Основан на градиентном бустинге деревьев решений.
5. Нейросеть MLPRegressor из sklearn.neural_network

Результаты тренировки модели регрессии разными методами:

Рассмотренные алгоритмы обучения:

1. Линейная регрессия sklearn: **нехватка памяти**
2. Логистическая регрессия sklearn: **нехватка памяти**
3. Алгоритм Cat-Boost: долго и неточно. Результат обучения на выборке..
..из 100 000:

RMSE: 1.2997146629864997
R²: 0.5322129448935901
4. Алгоритм LightGBM: лучший вариант:

RMSE: 0.955064606123774
R²: 0.7498517197477828
5. Нейросеть MLPRegressor: долго и неточно:

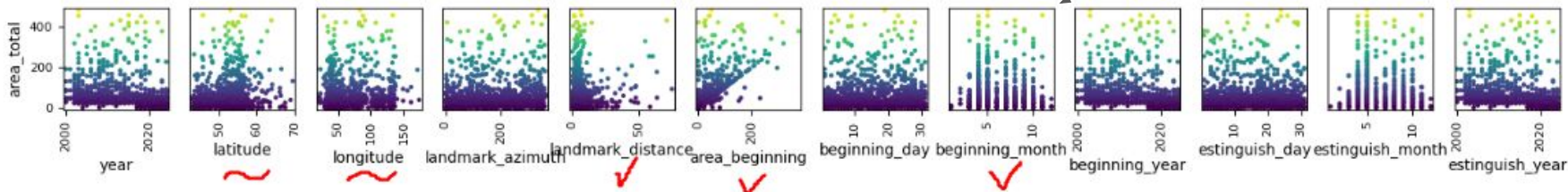
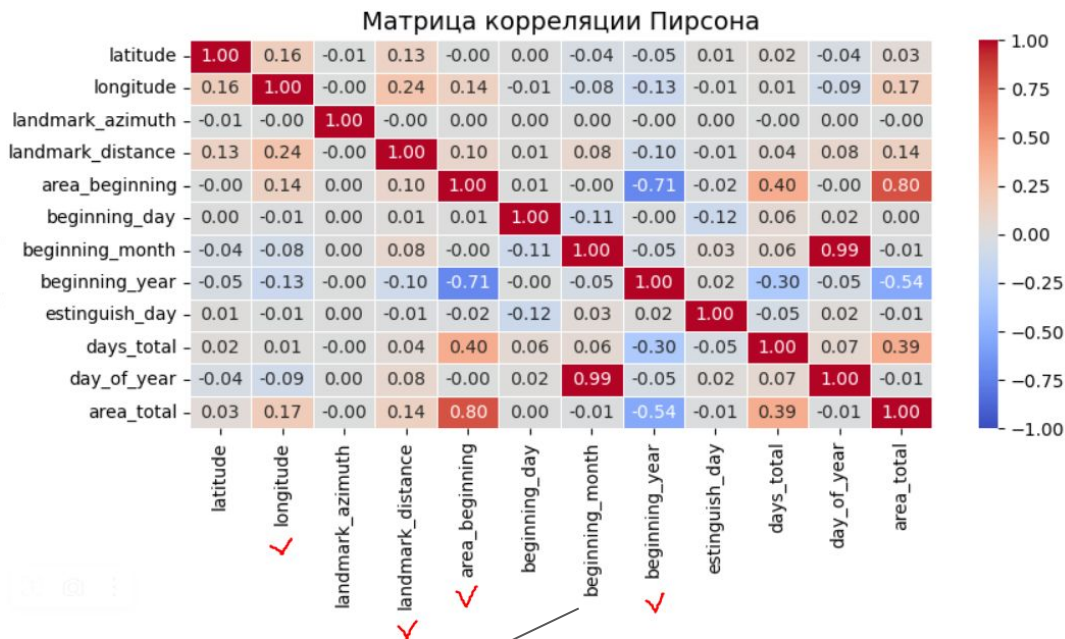
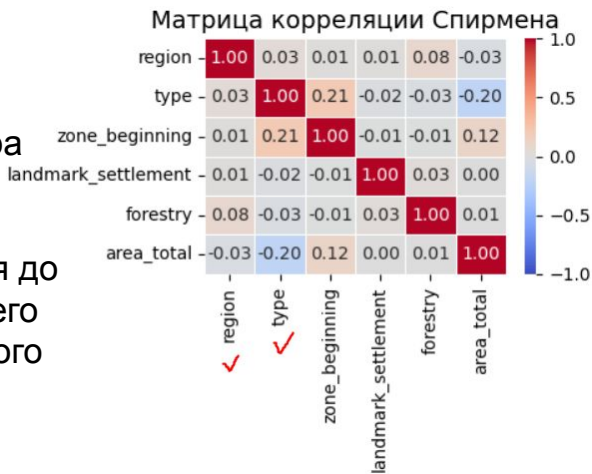
RMSE: 1.4421147036479285
R²: 0.5695891157846288

- Выбранным для работы алгоритмом стал LightGBM ввиду его быстродействия, высокой точности и способности вообще обработать более миллиона значений.

Выбор входных данных для обучения

Выбранные признаки:

- регион
- тип пожара
- широта
- долгота
- дистанция до ближайшего населённого пункта
- месяц
- год

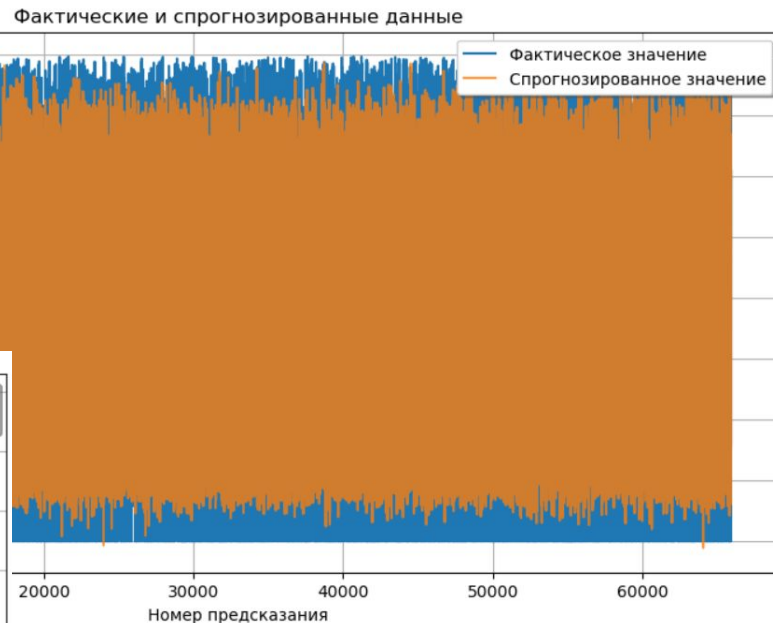
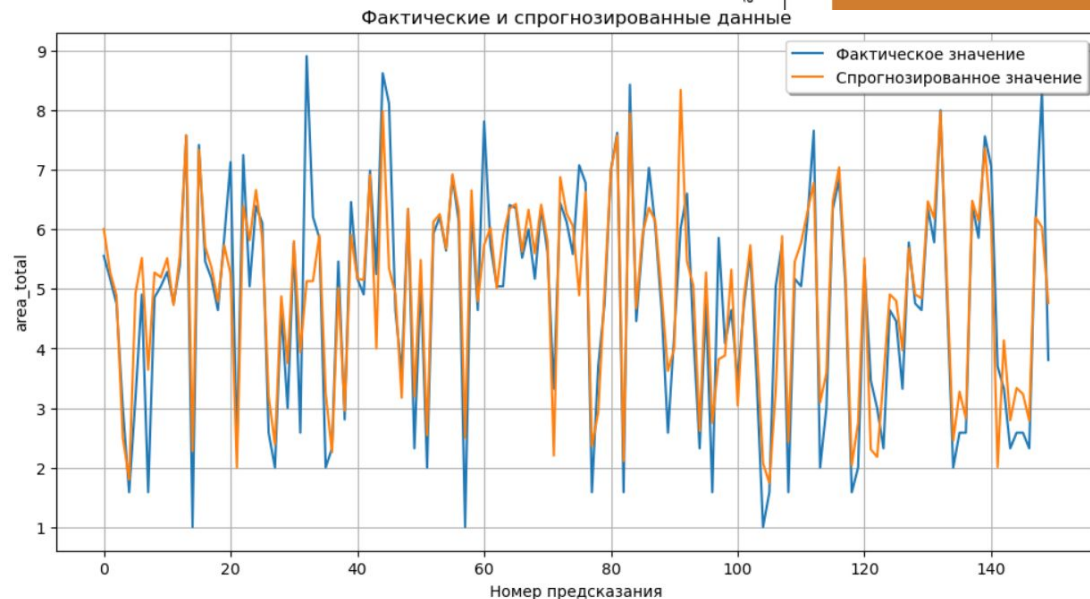


Результаты обучения модели

Прогноз площади

RMSE: 0.955064606123774

R^2 : 0.7498517197477828



все тестовые примеры

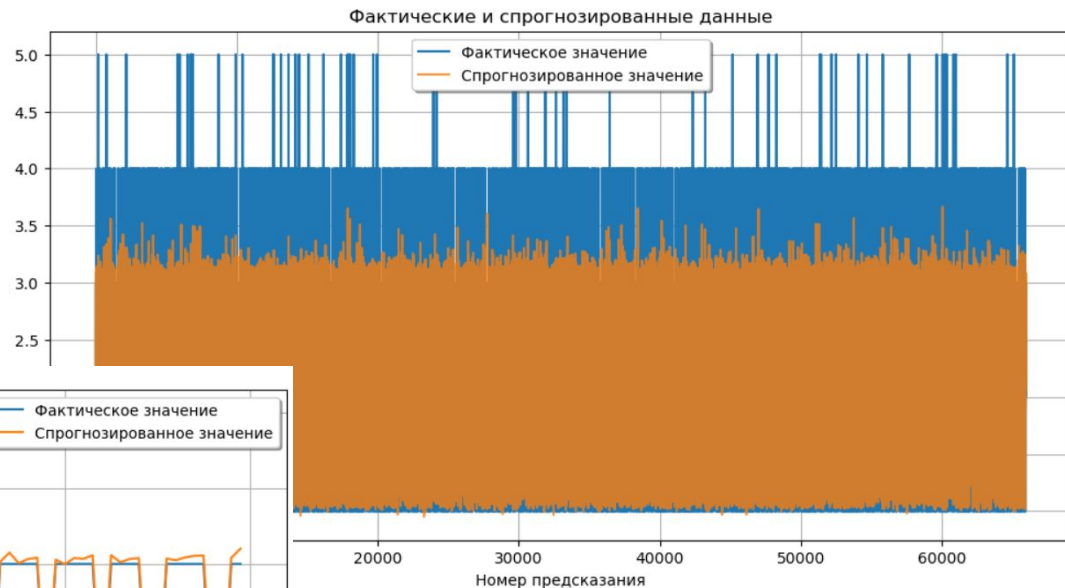
первые 150 примеров

Результаты обучения модели

Прогноз длительности

RMSE: 0.2381135299225969

R^2 : 0.8490690915545629



все тестовые примеры

первые 100 примеров

Анализ результатов обучения

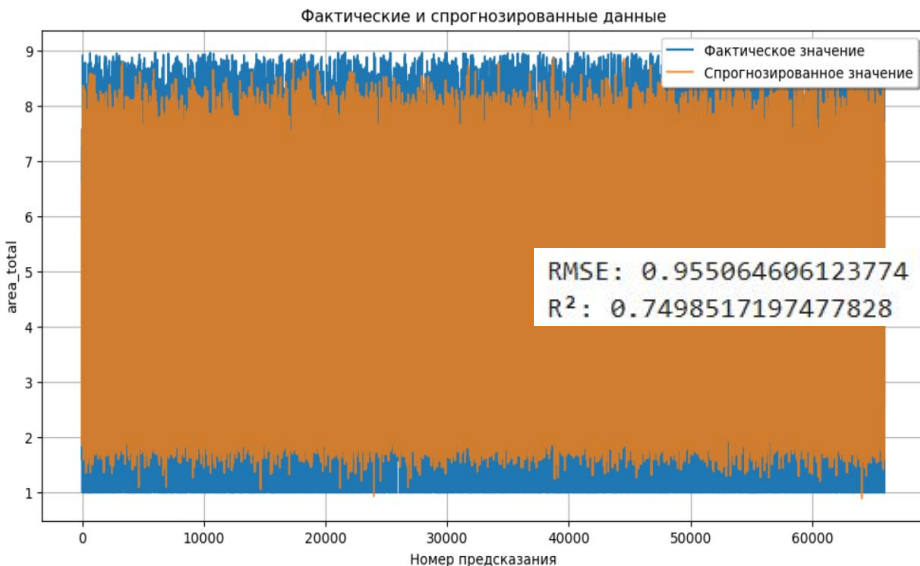
Прогноз площади: видимая ошибка снизу и сверху - в среднем 0.5 ед. С учётом нормализации данных по \log_2 , при приближении данных к максимальным значениям в 500 га, и минимальным в 2 га, ошибка сверху может составить до ~8.75 гектаров, ошибка снизу - до ~1.75 гектаров соответственно.

```
np.pow(9, 2) - np.pow(8.5, 2)
```

```
np.pow(2, 2) - np.pow(1.5, 2)
```

```
np.float64(8.75)
```

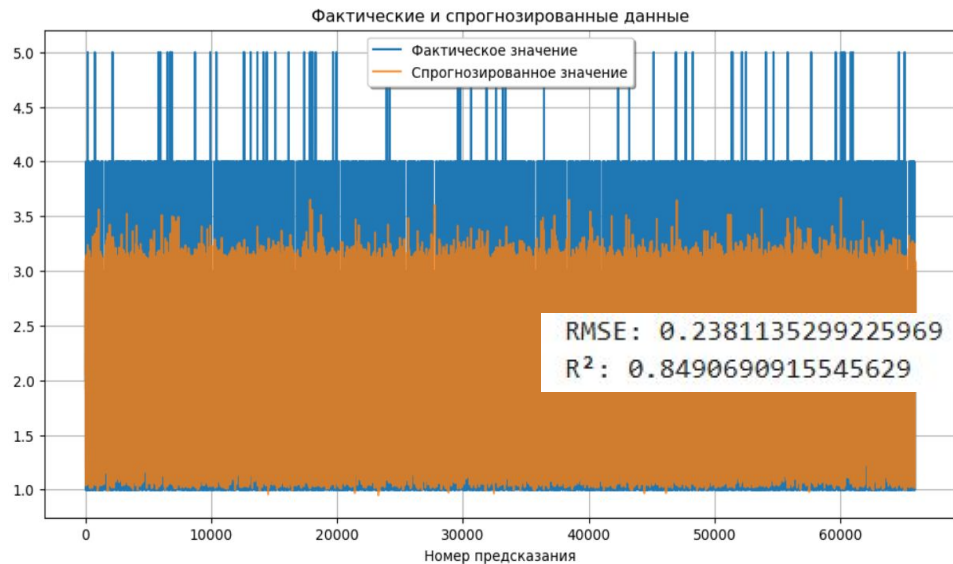
```
np.float64(1.75)
```



Прогноз длительности: видимая ошибка сверху - в среднем 0,75 ед. С учётом нормализации данных, при приближении данных к максимальным значениям в 16 дней, ошибка сверху может составить до ~1.75 дней.

```
np.pow(4, 2) - np.pow(3.25, 2)
```

```
np.float64(1.75)
```



Скриншоты работы ПО

Предсказание пожаров

Предсказание площади пожаров

Широта: 12 Долгота: 7

Расстояние до ближайшего населенного пункта: 10

Территория пожара при обнаружении: 5

Месяц начала: Апрель

Регион: Камчатский край

Год пожара: 2026

Тип пожара: Не лесной

Примерная площадь пожара:

Предсказать **Назад**

Предсказание пожаров

Предсказание площади пожаров

Широта: 12 Долгота: 7

Расстояние до ближайшего населенного пункта: 10

Территория пожара при обнаружении: 5

Месяц начала: Апрель

Регион: Камчатский край

Год пожара: 2026

Тип пожара: Не лесной

Ожидание результатов...

Примерная площадь пожара:

Назад

Предсказание пожаров

Предсказание площади пожаров

Широта: 12 Долгота: 7

Расстояние до ближайшего населенного пункта: 10

Территория пожара при обнаружении: 5

Месяц начала: Апрель

Регион: Камчатский край

Год пожара: 2026

Тип пожара: Не лесной

Примерная площадь пожара: 4.106811135211324

Предсказать **Назад**

Спасибо за внимание!

Сбор и накопление данных

Был найден набор данных по пожарам в России

Его составляющие - информация об отдельных очагах пожаров: точке регистрации (приведены ее координаты и положение относительно ближайшего населенного пункта), датах первого и последнего наблюдения и площади, пройденной огнем. Данные доступны за период с 2000 года по 12 сентября 2024 года.

Ссылка на данные: <https://tochno.st/datasets/fires>

Набор данных представлен в виде плоской таблицы, содержащей 23 атрибута, 1 459 464 наблюдения.

№	Атрибут	Описание	Число пропусков	Единица измерения	Тип
1	region	Наименование субъекта	0	-	String
2	oktmo	Код ОКТМО ¹ региона. Трем регионам, которые перестали существовать, присвоены специальные коды: Агинский Бурятский автономный округ — 76100000, Читинская область — 76650000, Усть-Ордынский Бурятский автономный округ — 25657000.	0	-	String
3	okato	Код ОКАТО ² региона. Трем регионам, которые перестали существовать, присвоены специальные коды: Агинский Бурятский автономный округ — 76100000, Читинская область — 76650000, Усть-Ордынский Бурятский автономный округ — 25657000.	0	-	String
4	year	Год	0	-	Integer
5	type	Тип пожара (Лесные, Нелесные)	0	-	String
6	code	Номер пожара (административный код пожара в информационной системе)	0	-	String
7	latitude	Широта точки регистрации пожара	0	градус	Float
8	longitude	Долгота точки регистрации пожара	0	градус	Float
9	zone_beginning	Зона мониторинга точки регистрации	5511	-	String
10	landmark_settlement	Ближайший к точке регистрации населенный пункт	369	-	String

¹ Общероссийский классификатор территорий муниципальных образований / Russian Classification of Territories of Municipal Formations

² Общероссийский классификатор объектов административно-территориального деления / Russian Classification on Objects of Administrative Division

№	Атрибут	Описание	Число пропусков	Единица измерения	Тип
11	landmark_azimuth	Азимут точки регистрации относительно ближайшего населенного пункта	369	градус	Float
12	landmark_distance	Расстояние от точки регистрации до ближайшего населенного пункта	369	километр	Float
13	forestry	Лесничество	6942	-	String
14	date_beginning	Дата первого наблюдения	0	-	String
15	area_beginning	Площадь пожара в момент регистрации	0	га	Integer
16	date_end	Дата последнего наблюдения	0	-	String
17	current_state	Состояние на момент формирования выгрузки	0	-	String
18	area_total	Площадь, пройденная огнем в субъекте РФ, всего	0	га	Integer
19	area_forest	Площадь, пройденная огнем в субъекте РФ, покрытая лесом	0	га	Integer
20	area_fund_total	Площадь на территории лесного фонда, пройденная огнем в субъекте РФ, всего	0	га	Integer
21	area_fund_forest	Площадь на территории лесного фонда, пройденная огнем в субъекте РФ, покрытая лесом	0	га	Integer
22	comment	Примечания	1 017 058	-	String
23	zone	Зона(ы) мониторинга	805 824	-	String