

Статья **Big Data и ИИ против лесных пожаров**

Заголовок статьи	Big Data и ИИ против лесных пожаров
Тема статьи О чем обязательно нужно рассказать	Использование технологий больших данных и искусственного интеллекта (ИИ) для прогнозирования лесных пожаров
Для кого статья Опишите аудиторию: что это за люди, какие они	ИТ-специалисты, интересующиеся решениями на базе ИИ Инвесторы Государственные органы и экологические организации, которые отвечают за предотвращение / тушение лесных пожаров
Какие проблемы есть у аудитории в этой теме	ИТ-специалисты: интерес скорее профессиональный «А как работает это решение? Для чего еще можно использовать ИИ?» Инвесторы: Для чего еще можно использовать ИИ? Какая выгода от этого решения, каковы риски и сложности? Для каких проектов можно пригласить этого специалиста? Гос.органы и экология: Как внедрить? Сколько стоит? Как использовать?
Полезное действие статьи Зачем читать эту статью? Помогаем читателю увидеть/понять...	Помогаем ИТ-специалистам понять алгоритм внедрения и разработки. Рассказываем, с помощью каких инструментов реализовано решение, приводим примеры кода. Остальным показываем, что решение в принципе есть и при определенных условиях они могут его использовать.

Чем сможем проиллюстрировать статью	Скрины, код
Какие продукты или услуги упоминаем	Алгоритм разработки решения, + в лиде или наоборот в заключении общая / контактная информация. Если будет оформляться целое портфолио (как задумано), можно сделать перекрестные ссылки «Читайте также о решении...» <ссылка на другую страницу портфолио>
Структура и тезисы Какая структура будет у материала в формате: Заголовок Лид Подзаголовок Тезисы	Заголовок (название решения?) Лид: вводная информация о решении и разработчике <i>Специалисты BAUM-Inform (Григорий Соколов в составе группы разработчиков + краткий professional profile) BAUM-Inform разработали систему на базе ИИ для определения степени риска лесных пожаров.</i> Тезисы: 1. Подробнее как и для чего можно использовать решение 2. Подготовка датасета 3. Обучение нейронной сети 4. Тестирование 5. Результаты Выводы: что сейчас с решением, можно ли его приобрести и использовать или ему требуется доработка, или может быть его уже кто-то использует. Возможно, планы по доработке и развитию решения. Footer: перекрестные ссылки на другие статьи

Big Data и ИИ против лесных пожаров



Фото: [Michael Chacon](#) на [Unsplash](#)

Меня зовут Григорий Соколов, я Senior data scientist & architect, автор и архитектор платформы прикладного искусственного интеллекта «BAUM AI». Я создал более 20 системных решений с применением искусственного интеллекта. Помимо этого, веду научную и образовательную деятельность в МГТУ им. Н.Э. Баумана. Мой телеграм-канал Lasttrader посвящен методам и способам применения искусственного интеллекта в современном мире — там мы с коллегами делимся полезными ссылками и участвуем в мастер-классах с live кодингом.

В последние годы весь мир, от российской тайги до Австралии, все чаще сталкивается с серьезными пожарами, уничтожающими миллионы гектаров леса. Мы в BAUM-Inform решили создать нейронную сеть для профилактики возгораний в пожароопасных зонах и оперативного реагирования на только что начавшийся пожар — ведь это сведет ущерб от стихийного бедствия к минимуму.

Алгоритм определения вероятности возникновения лесных пожаров

Для обучения модели нейронной сети нужна, прежде всего, обучающая выборка, для которой мы решили использовать архив космоснимков с отмеченными

термоточками, архив данных по последствиям лесных пожаров и архив погодных условий. Используя эту информацию, мы отобрали параметры, которые, на наш взгляд, влияют на вероятность возникновения новых пожаров:

1. Координаты термоточек;
2. Возникновение лесных пожаров в этих термоточках;
3. Последствия каждого конкретного лесного пожара.

Мы исключили из выборки все неподтвержденные термоточки и пожары без последствий.

Кроме того, в качестве параметров мы проанализировали следующие данные:

1. История изменения погодных условий в термоточках;
2. Расстояние от термоточек до ближайших населенных пунктов по наикратчайшему пути и по путям транспортного сообщения;
3. Характеристики населенных пунктов, их жителей, наличие в этих городах вузов, театров и других объектов культуры;
4. Удаленность термоточек от ближайших железнодорожных путей и автодорог.

И, наконец, учли еще ряд моментов:

1. Удаленность термоточек от мест складирования или переработки ТБО;
2. Наличие промышленных предприятий, пастбищ, сельхозугодий, линий ЛЭП, трубопроводов и пр. рядом с термоточками;
3. Наличие в населенных пунктах предприятий и их характеристики, в том числе отрасль и численность сотрудников.

Последствия пожаров – это цели создания решения; все остальные данные мы использовали как входные параметры обучения нейронной сети.

Основные входные параметры для модели можно получить из открытых источников: это топографические карты, открытые БД о характеристиках населенных пунктов и открытые архивы погодных условий. Труднее всего получить доступ к данным о последствиях пожаров и архиву термоточек – его нужно приобрести у одного из российских поставщиков информации.

Что еще можно учесть в модели – и для чего это нужно

Если включить в модель прогноз погодных условий или перемещение населения, например, выезды на дачу или шашлыки на майские праздники, модель станет динамической

Если применить классическую модель распространения лесного пожара с учетом вероятности направления и силы ветра (по статистике) можно зонировать территорию по степени риска

Если добавить в модель фактические данные с сети метеостанций можно задействовать интернет вещей

Если добавить историю действия сил и средств и численные характеристики их результативности у нас получится экспертная интеллектуальная система, способная рекомендовать управленческие решения

Задачи

В терминах ИИ мы поставили перед собой две возможные задачи:

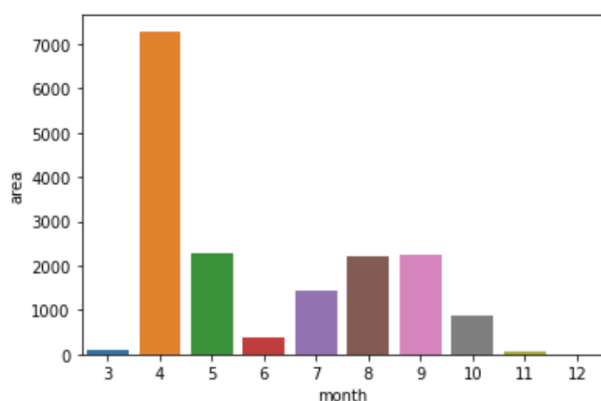
1. Вычислить площадь термоточки, её широту и долготу (задача регрессии для трёх искомых переменных).
2. Установить вероятность принадлежности к классу (то есть отсутствие или наличие термоточки по бинарной классификации). В этом случае мы получаем ноль при отсутствии термоточки и единицу при её наличии.

Для решения поставленных задач, конечно же, необходим датасет.

Датасет

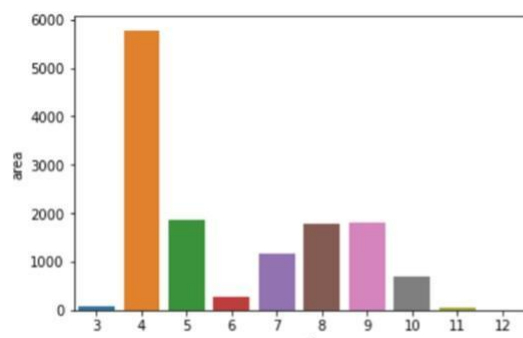
В качестве выборки мы используем 230000 термоточки по всей России (с 01.01.2018 по 02.04.2020) и дополнительно 16857 термоточек по Самарской области (с 01.01.2009 по 03.11.2020).

Данные о погоде за период с 01.01.2009 по 03.11.2020 мы получаем от погодных станций населенных пунктов: Самара, Авангард, Сызрань, Челно-Вершины, Большая-Глушица и Бугуруслан. В качестве входных данных принимаем вектор данных из числовых и категориальных признаков погодных условий ([посмотреть все признаки](#)).

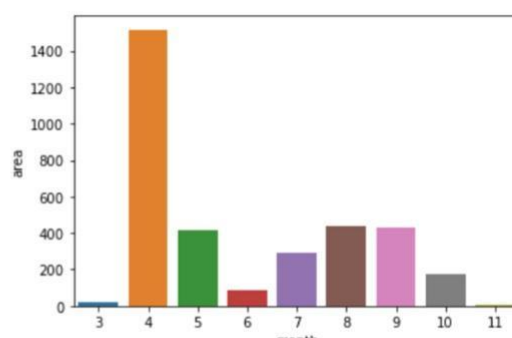


Распределение по количеству пожаров в Самарской области за указанный период

Далее мы разделили термоточки по Самаре на обучающий датасет (80% наблюдений – 13485 строки) и тестовый датасет (20% наблюдений – 3372 строки).

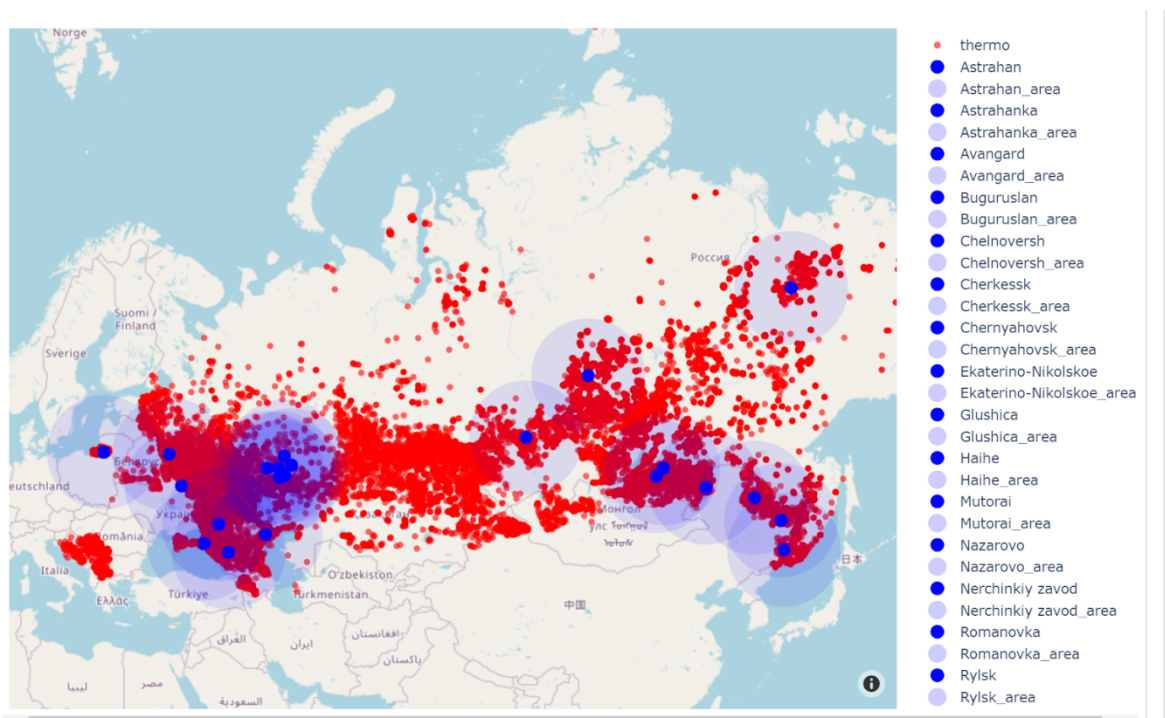


ТЕРМОТОЧКИ ДЛЯ
ОБУЧАЮЩЕЙ ВЫБОРКИ

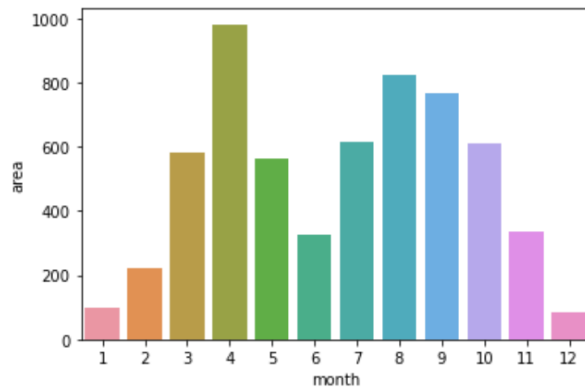


ТЕРМОТОЧКИ ДЛЯ
ТЕСТОВОЙ ВЫБОРКИ

Полученные 243485 термоточек мы соотнесли с погодными станциями в радиусе 100 км. В результате по погодным станциям разнеслось 49749 термоточек.



После объединения термоточек с погодными данными мы получаем обучающий датасет с распределением по количеству пожаров.



Обучение модели

Признаки

В качестве целевого признака мы берем **area** – бинарный признак, показывающий, был пожар или нет.

Список признаков для обучения модели:

T – температура воздуха
P – атмосферное давление, приведенное к уровню моря
U – относительная влажность
Ff – скорость ветра
Td – температура точки росы (Td)
RRR – Количество выпавших осадков
DD – направление ветра
N – общая облачность
day – день
month – месяц

Далее, мы использовали метод ONE для кодирования категориальных признаков DD и N, после чего стандартизировали все признаки.

Вот как выглядит итоговый список признаков для обучения:

['T', 'P', 'U', 'Ff', 'Td', 'RRR', 'DD_Ветер, дующий с востока',
'DD_Ветер, дующий с востоко-северо-востока',
'DD_Ветер, дующий с востоко-юго-востока',
'DD_Ветер, дующий с запада',
'DD_Ветер, дующий с западо-северо-запада',
'DD_Ветер, дующий с западо-юго-запада',
'DD_Ветер, дующий с севера',
'DD_Ветер, дующий с северо-востока',

'DD_Ветер, дующий с северо-запада',
 'DD_Ветер, дующий с северо-северо-востока',
 'DD_Ветер, дующий с северо-северо-запада',
 'DD_Ветер, дующий с юга',
 'DD_Ветер, дующий с юго-востока',
 'DD_Ветер, дующий с юго-запада',
 'DD_Ветер, дующий с юго-юго-востока',
 'DD_Ветер, дующий с юго-юго-запада',
 'DD_Штиль, безветрие',
 'N_10% или менее, но не 0', 'N_100%.',
 'N_20–30%.', 'N_40%.', 'N_60%.',
 'N_70 – 80%.', 'N_90 или более, но не 100%',
 'N_Небо не видно из-за тумана и/или других метеорологических явлений.',
 'N_Облаков нет.', 'day', 'month', 'N_50%.']

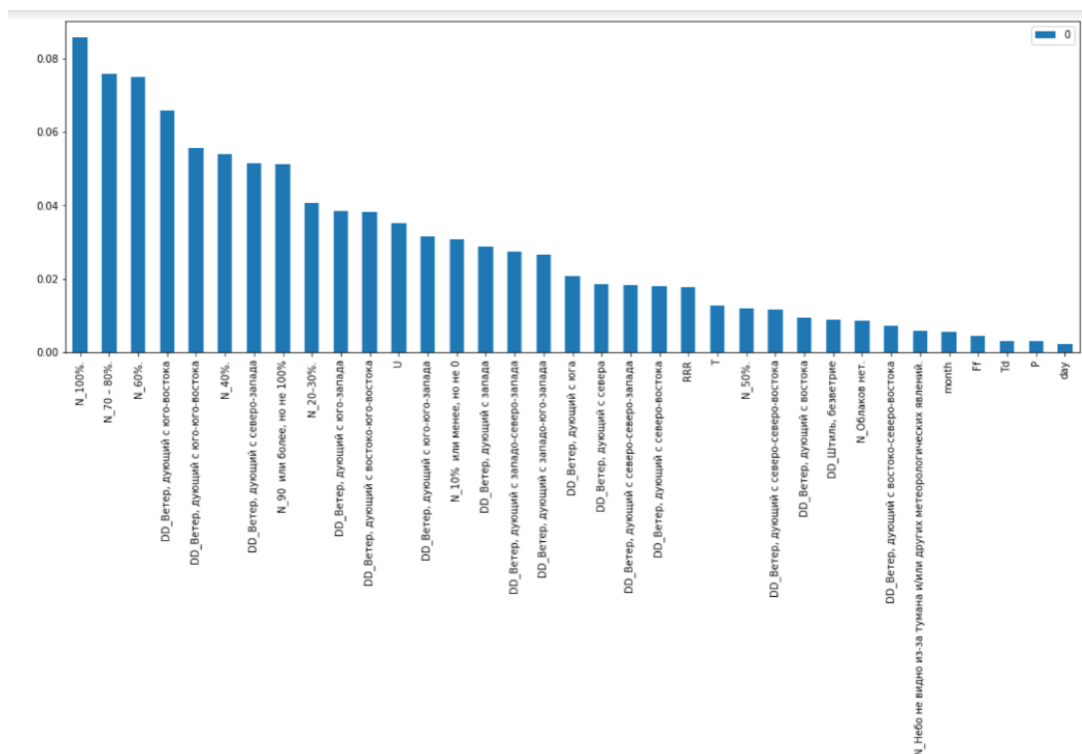


График важности признаков

Этапы обучения

Для начала пробуем добавить синтетические данные без перемешивания порядка записей, чтобы выровнять количество классов.

```
[ ] # объединяем X_train, y_train
XY = pd.concat([X_train, y_train], axis=1)
#расширяем датасет (мажоритарные данные)
area_0 = XY[XY.area==0]
area_1 = XY[XY.area==1]

area_upsampled = resample(area_1,
                           replace=True,
                           n_samples=len(area_0),
                           random_state=42)

upsampled = pd.concat([area_0, area_upsampled])

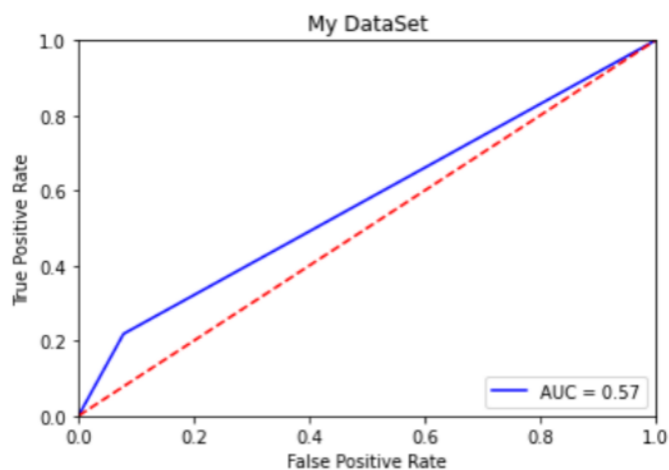
print(upsampled.area.value_counts())

y_train_up = upsampled.area
X_train_up = upsampled.drop('area', axis=1)

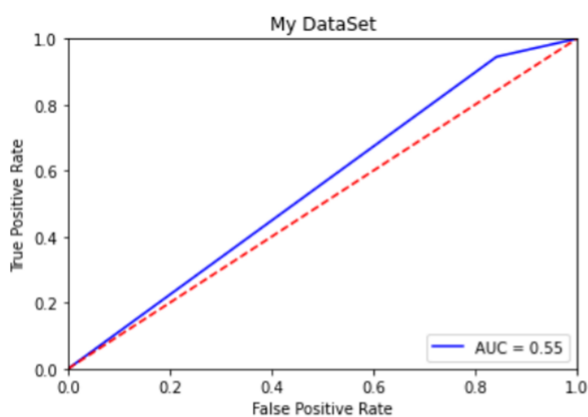
1      5757
0      5757
Name: area, dtype: int64
```

Далее для обучения модели пробуем методы dense neural networks, decision trees, random forest и метод Байесовских классификаторов (смотреть [программный код](#) по определению вероятности принадлежности к классу).

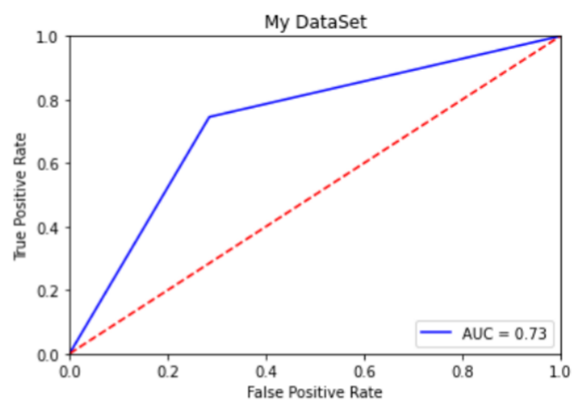
Визуально полученные результаты можно представить следующим образом:



DENSE NEURAL NETWORKS

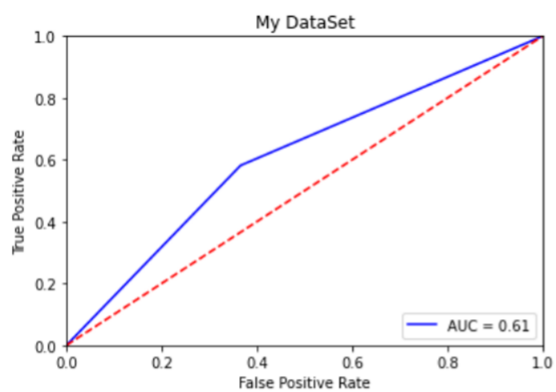


БАЙЕСОВСКИЕ



КЛАССИФИКАТОРЫ

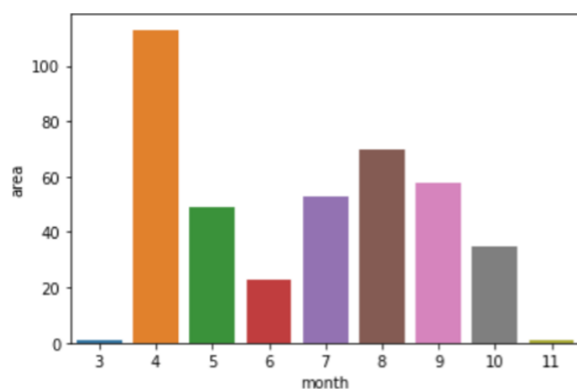
При определении классов наилучший результат мы получили при использовании Байесовских классификаторов. Далее дорабатываем полученные результаты с помощью grid search.



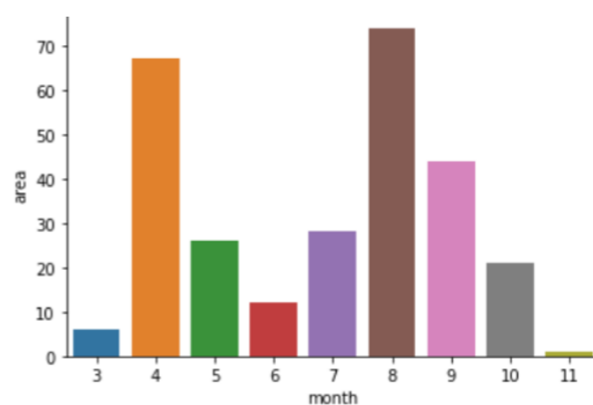
GRID SEARCH

Проверка на тестовой выборке

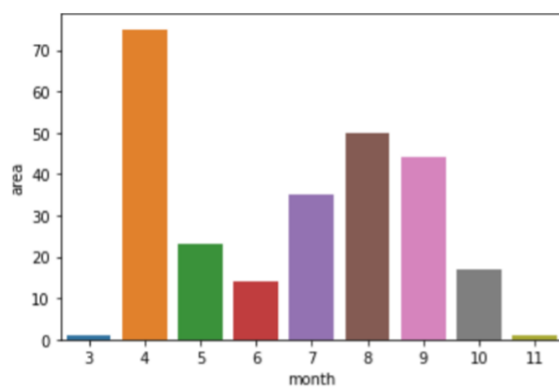
После обучения мы проверили модель на тестовой выборке термоточек. Распределения термоточек из тестовой выборки по выбранным погодным станциям:



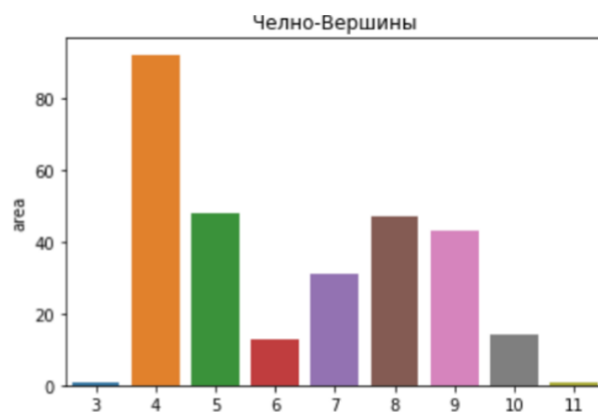
САМАРА



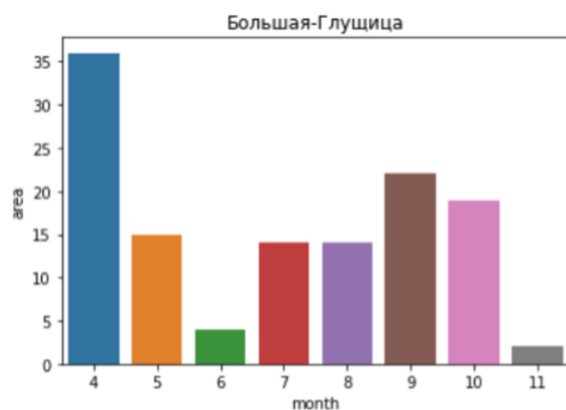
АВАНГАРД



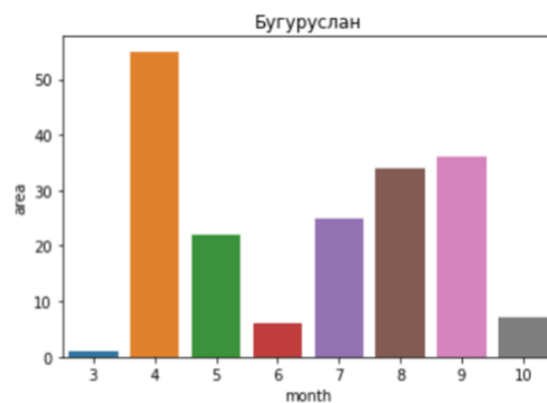
СЫЗРАНЬ



ЧЕЛНО-ВЕРШИНЫ

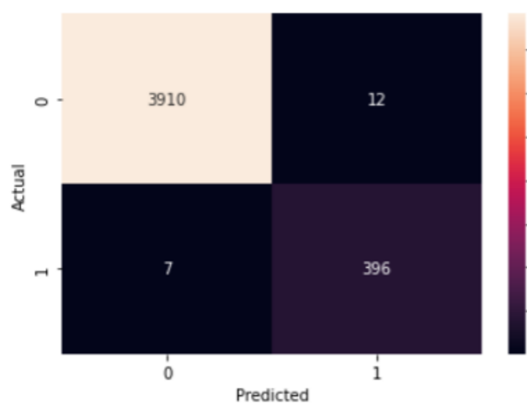


БОЛЬШАЯ ГЛУШИЦА

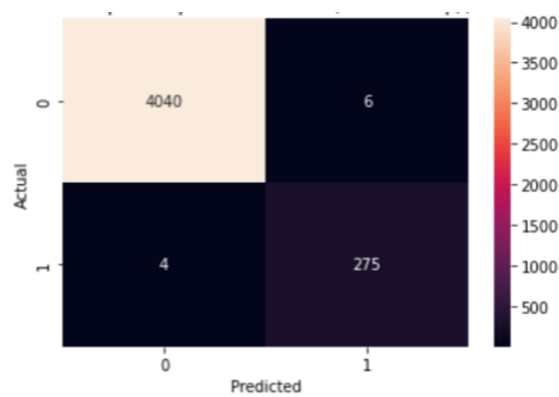


БУГУРУСЛАН

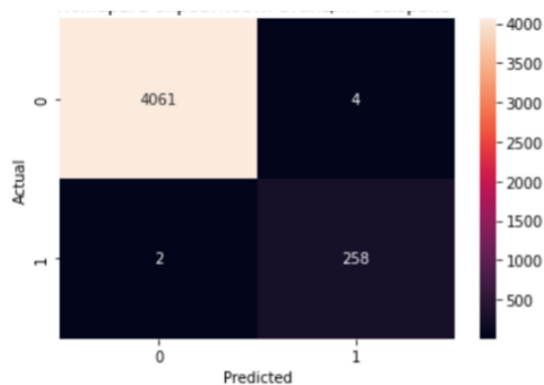
Ниже представлена матрица ошибок при пороге классификации 0,5. Как видно, мы получили модель, прогнозирующую лесной пожар в той или иной термоточке с достаточно высокой точностью, варьирующей в диапазоне от 0,11% до 0,43%.



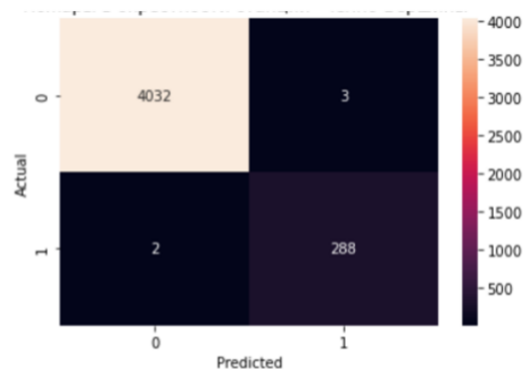
ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. САМАРА



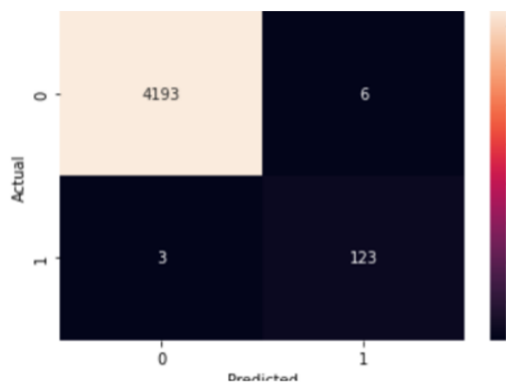
ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. АВАНГАРД



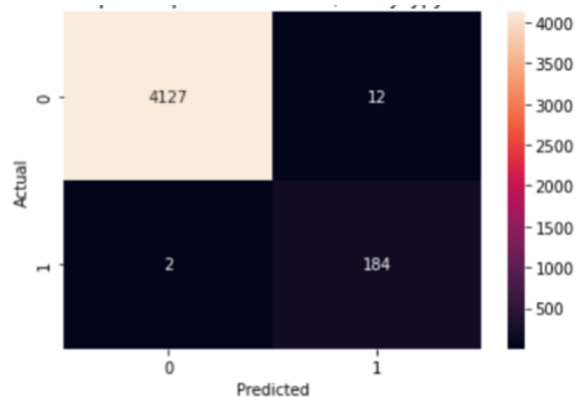
ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. СЫЗРАНЬ



ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. ЧЕЛНО-ВЕРШИНЫ



ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. БОЛЬШАЯ ГЛУШИЦА



ПОЖАРЫ В ОКРЕСТНОСТЯХ
СТ. БУГУРУСЛАН

Выводы

В целом, разработанная модель машинного обучения может прогнозировать возникновение лесных пожаров с высокой точностью.

С 2021 года решение используется заказчиками в ряде регионов. Чтобы обучить модель для других регионов, понадобятся координаты гидрометеостанций и данные по существующим термоточкам соответствующих территорий.