*Тарарин А.В.*

*Рыбалко Н.А.*

*студенты*

***Научный руководитель****: Денисова А.И.,*

*канд.экон.наук, ст.преподаватель*

(ГУУ, г. Москва)

**МОДЕЛИРОВАНИЕ ВЕРОЯТНОСТИ ВОЗНИКНОВЕНИЯ ЛЕСНЫХ ПОЖАРОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ**

***Аннотация.*** *В данной работе исследуются открытые данные о местах и типах природных пожаров в России за 2012–2021 гг., собранные по оперативным сведениям МЧС. Проведен анализ факторов, влияющих на вероятность возникновения лесных пожаров. С целью прогнозирования подобных ситуаций построены и оценены модели логистической регрессии и градиентного бустинга.*

***Ключевые слова:*** *лесные пожары, логистическая регрессия, градиентный бустинг, предсказание вероятности.*

Лесные пожары представляют серьезную экологическую угрозу и могут привести к негативным последствиям как для природы, так и для человечества. Поэтому моделирование вероятности возникновения лесных пожаров имеет большое значение для прогнозирования и предотвращения таких бедствий. Исследование факторов, влияющих на вероятность возникновения лесных пожаров, является актуальным и востребованным направлением, и особенно будет полезна в сфере пожарной безопасности.

При анализе вероятности возникновения лесных пожаров актуальны методы машинного обучения, так как они позволяют обрабатывать большие объемы данных и выявлять сложные взаимосвязи между различными переменными. В моделях машинного обучения могут быть учтены множество факторов. Эти методы могут помочь создать точные и надежные модели предсказания возникновения пожаров, что в свою очередь позволит своевременно принимать меры по их предотвращению и борьбе с ними. Таким образом, использование методов машинного обучения в анализе вероятности лесных пожаров позволяет повысить эффективность и точность прогнозирования, что является крайне важным для обеспечения безопасности природы и людей.

Моделирование вероятности наступления каких-либо событий стало предметом многих исследований: этот подход используется многими банками для выявления риска дефолта, например, такое исследование проводилось Банком России в 2022 году [1], в этой работе автор с помощью логистической регрессии и случайного леса строил прогноз вероятности дефолта российских компаний с использованием транзакционных данных платежной системы Банка России. Подобный подход регулярно используется при моделировании поведений клиентов, например, для предсказания ухода клиентов в телекоммуникационной отрасли [2], а также статья, которая исследует вероятности противоправных действий [3].

Классический подход для решения такой задачи – бинарная классификация. В контексте методов машинного обучения здесь применяются следующие методы принятия решений: метод k-ближайших соседей (k-NN), логистическая регрессия, алгоритм Байеса, ансамблевый метод бэггинга, алгоритм бустинга, случайный лес и деревья решений [4].

Для моделирования вероятности лесных пожаров было выбрано два метода: логистическая регрессия и градиентный бустинг.

Логистическая регрессия - это метод классификации, который используется для прогнозирования вероятностей принадлежности объектов к определенным классам. Он основан на логистической функции и может работать с числовыми данными [5].

Градиентный бустинг - это метод ансамблирования, который объединяет несколько слабых предсказывающих моделей (чаще всего деревьев решений) в сильную предсказывающую модель. Он основан на принципе пошагового обучения, при котором каждая новая модель исправляет ошибки предыдущих моделей (иначе говоря, бустинге), причем оптимизация осуществляется методом градиентного спуска. Градиентный бустинг обычно демонстрирует весьма высокую точность предсказаний и хорошую обобщающую способность, однако, в отличие от логистической регрессии, гораздо сложнее интерпретируется [6].

Для построения модели предсказания вероятности возникновения лесных пожаров были использованы открытые данные о местах и типах природных пожаров в России за 2012–2021 гг. из источника [7]. Этот набор данных представляет собой оперативные сведения МЧС о произошедших пожарах, включая в себя тип пожара (лесной, торфяной, контролируемый, неконтролируемый), местоположение (координаты широты и долготы) и дату.

Согласно [8], основными факторами, ведущими к возникновению лесных пожаров, являются: климатические условия, наличие сухой и горючей растительности, человеческая деятельность, недостаток мер профилактики, недооценка рисков возникновения пожаров, географические особенности местности. В виду представленной информации становится понятно, что факторы для анализа не могут дать полную информацию для построения модели, поэтому было принято решение создания новых переменных из уже имеющихся факторов.

Мы можем рассматривать следующие группы новых факторов.

1. Временные параметры: месяц, день недели, выходной или праздничный день. Каждый из параметров был преобразован в соответствующий набор фиктивных переменных.
2. Параметры местоположения: мы можем рассматривать в качестве факторов как координаты широты и долготы в исходном виде, так и поставить им в соответствие наименования субъектов Российской Федерации.
3. Параметры взаимосвязи: предположим, что пожары других типов (в частности, торфяной) могут стать причиной возникновения лесного пожара. Мы проверяем, имелся ли факт другого типа возгорания максимум за сутки до возникновения лесного пожара в радиусе до километра. Таким образом, для каждого из типов возможных типов пожаров была образована новая фиктивная переменная.

Таким образом, данные были подготовлены для дальнейшего построения моделей машинного обучения с учетом различных характеристик пожаров.

Далее был проведен дескриптивный анализ данных и сделаны следующие выводы: доля лесных пожаров среди всех пожаров в данных составляет 46%, а наиболее часто пожары происходят в апреле и июле.

Для построения логистической регрессии мы выбрали объясняемую переменную "Лесной пожар”, которая равна 1 в случае, если в это время и в этом месте был зафиксирован лесной пожар.

Для обучения модели мы случайным образом разделили данные на обучающую и тестовую выборки в соотношении 80% к 20%. Использование кросс-валидации требует дополнительных вычислительных ресурсов и может быть затратным по времени, поэтому в данном случае было принято решение использовать случайное разделение на обучающую и тестовую выборки для оценки качества модели.

Модель логистической регрессии была построена на языке Python с использованием библиотеки scikit-learn. После построения модели был сделан вывод, что факторами, которые больше всего влияют на вероятность, возникновения лесных пожаров является факт возгорания или в апреле, или в июле или в августе. Полученная имеет характеристики на тестовой выборке ценность позитивного прогноза, равная 0,54 и чувствительность, равная 0,5. Точность модели логистической регрессии равна 0,5. Теперь построим модель градиентного бустинга.

Модель градиентного бустинга также была построена на языке Python с использованием библиотеки scikit-learn. Она также была оценена характеристиками ценность позитивного прогноза, равная 0,87 и чувствительность, равная 0,85. Точность модели градиентного бустинга равна 0,85. Используя характеристику “значимость признака” (это информация, которая показывает, насколько значим каждый признак влияет на прогнозы модели) получили, что самыми значимыми факторами оказались широта, долгота и принадлежность даты июлю и августу [9].

После моделирования можно сделать вывод, что модель логистической регрессии имеет не высокую точность, но она помогает оценить влияние каждого фактора на факт возникновения лесного пожара, в свою очередь модель градиентного бустинга является более точной, но не даёт возможности посмотреть на влияние каждой переменной на зависимую.

Для построения лучшей модели предсказывания вероятности возникновения лесных пожаров нужны более развёрнутые данные, которые не содержаться в источники. Например: средняя температура воздуха, влажность воздуха, наличие на этих координатах предприятий, которые увеличивают риск возникновения лесного пожара.

*Библиографический список*

*1. Модель вероятности дефолта с использованием транзакционных данных российских компаний // Банк России. URL: https://cbr.ru/StaticHtml/File/138732/wp\_97.pdf (дата обращения: 16.04.2024).*

*2. Kavitha V. & Kumar G. & Kumar S. & Harish M. Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms // International Journal of Engineering Research and. -2020. -V9.*

*3. Денисова А. И. Моделирование рисков лицензируемых товарных рынков российской Федерации // Научные труды Вольного экономического общества России. – 2017. – Т. 206, № 4. – С. 133-145.*

*4. Донцова Ю. С. Анализ методов бинарной классификации // Известия Самарского научного центра Российской академии наук. -2014. -№16 (6-2). -С. 434-438.*

*5. Как легко понять логистическую регрессию // HABR. URL: https://habr.com/ru/companies/io/articles/265007/ (дата обращения: 16.04.2024).*

*6. Ансамблевые методы машинного обучения // HABR. URL: https://habr.com/ru/articles/571296/ (дата обращения: 16.04.2024).*

*7. Пожары в России: данные о местах и типах природных пожаров за 2012–2021 гг. // Инфраструктура научно-исследовательских данных. URL: https://data.rcsi.science/data-catalog/datasets/202/ (дата обращения: 16.04.2024).*

*8. Главное управление МЧС России URL:* [*https://63.mchs.gov.ru/deyatelnost/poleznaya-informaciya/dopolnitelnye-stranicy*](https://63.mchs.gov.ru/deyatelnost/poleznaya-informaciya/dopolnitelnye-stranicy)*/staticheskie-stranicy/poleznaya-informaciya/osnovnye-prichiny-vozniknoveniya-lesnyh-pozharov (дата обращения: 16.04.2024).*

*9. Feature importances with a forest of trees // scikit-learn URL: https://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_importances.html (дата обращения: 16.04.2024).*