

## Аннотация

Целью данной работы является изучение возможности определять неработающие видео по статистике просмотров. В ходе работы был предложен ряд моделей и проведены численные эксперименты. Наилучшей точности классификации удалось достичнуть с использованием классификатора из библиотеки CatBoost<sup>1</sup>. Наилучший классификатор показал точность в 80.3 % и полноту в 51.4 %. Точность классификации удалось улучшить на 9 % с помощью добавления в данные событий плеера и создания отдельной модели для каждого плеера. В данном случае удалось достичнуть точности классификации в 89.3 % при той же полноте. На основании данной работы рекомендуется внедрение классификатора в сервис Yandex Video, а также дальнейшее изучение свойств модели и ее последующее улучшение.

---

<sup>1</sup>Liudmila Prokhorenkova и др. “CatBoost: unbiased boosting with categorical features”. B: (28 июня 2017). arXiv: <http://arxiv.org/abs/1706.09516v5> [cs.LG].

## Содержание

<b>1. Аннотация</b>	<b>2</b>
<b>2. Используемые определения</b>	<b>4</b>
2.1. Бинарная классификация . . . . .	4
2.2. Видеоплеер . . . . .	4
2.3. Yandex Video . . . . .	5
2.4. Yandex Toloka . . . . .	6
2.5. Неработающие видео . . . . .	6
<b>3. Введение</b>	<b>9</b>
<b>4. Предыдущие работы</b>	<b>10</b>
<b>5. Подготовка данных</b>	<b>12</b>
5.1. Исходные данные . . . . .	12
5.2. Выбор признаков . . . . .	12
<b>6. Обучение классификатора</b>	<b>14</b>
6.1. Выбор модели . . . . .	14
6.2. Обучение модели . . . . .	15
6.3. Анализ модели . . . . .	19
6.4. Одноплеерная модель . . . . .	19
<b>7. Заключение</b>	<b>23</b>
<b>Список литературы</b>	<b>24</b>

## Используемые определения

### 2.1 Бинарная классификация

В данной работе рассматривается задача бинарной классификации, то есть разбиения множества объектов на два класса. Были выбраны следующие обозначения: под классом 0 (нулевым классом, отрицательным классом) подразумевается класс работающих видео, а под классом 1 (положительным классом) подразумевается класс неработающих видео, детектирование которых и является целью данной работы.

При анализе предсказания модели в задаче бинарной классификации принято говорить о четырех группах объектов: объекты, верно отнесенные к отрицательному классу (true negatives); объекты, верно отнесенные к положительному классу (true positives); объекты, неверно отнесенные к отрицательному классу (false negatives) и объекты, неверно отнесенные к положительному классу (false positives).

Для оценки модели в задаче бинарной классификации принято использовать метрики точности (precision) и полноты (recall), которые вводятся следующим образом:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN},$$

где с помощью  $P$  обозначена точность (precision),  $R$  — полнота (recall),  $TP$ ,  $FP$  и  $FN$  — числа объектов, относящихся к той или иной группе из четырех указанных выше. Так,  $TP$  — это число объектов, верно отнесенных к положительному классу (true positives),  $FP$  — это число объектов, неверно отнесенных кциальному классу (false positives), а  $FN$  — это число объектов, неверно отнесенных к отрицательному классу (false negatives). Здесь и далее под точностью и полнотой классификации подразумеваются именно введенные выше метрики.

### 2.2 Видеоплеер

Под видеоплеером (плеером) в данной работе подразумевается элемент web страницы, позволяющий пользователю смотреть видеозаписи (рисунок 1). Один видеоплеер, как правило, умеет воспроизводить ряд видеозаписей, хранящихся на сервере провайдера контента. Различные видеохостинги могут использовать один общий плеер, однако внутри одного хостинга видео, как правило, воспроизводятся с помощью одного и того же плеера.

Помимо предоставления конечным пользователям возможности просматривать видеозаписи, видеохостинги часто предоставляют возможность другим интернет ресурсам интегри-

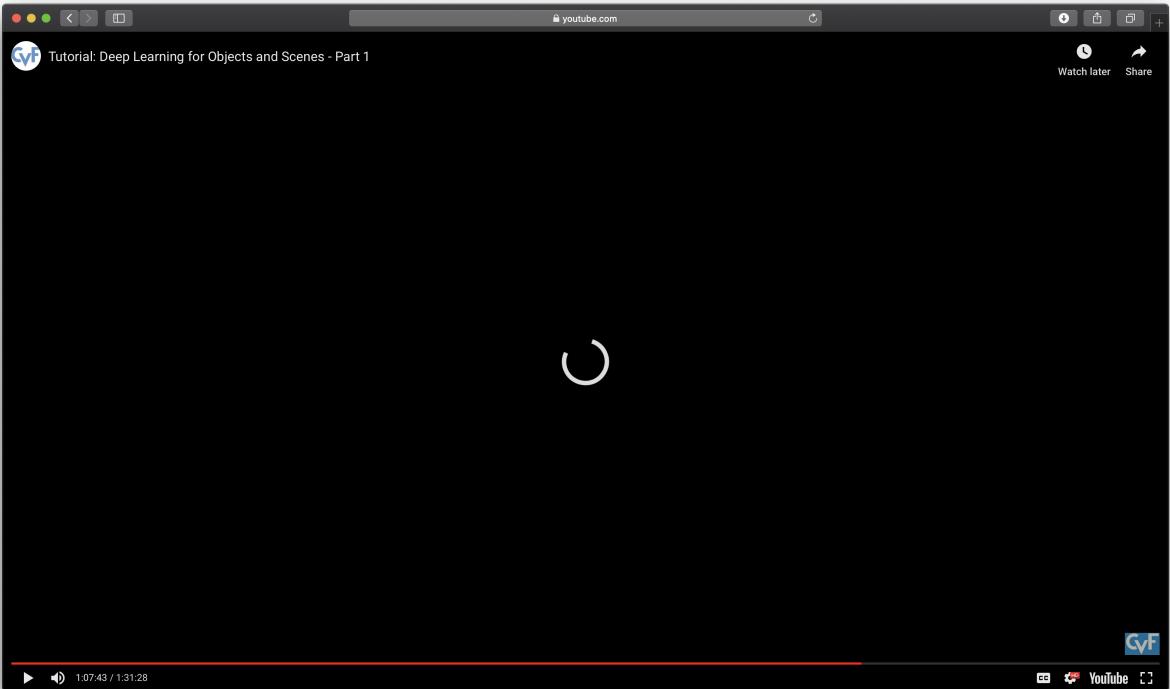


Рисунок 1 — Видеоплеер YouTube.

ровать этот контент в свои страницы. Для этого, как правило, используется тэг iframe языка разметки HTML. Данный тэг, согласно спецификации<sup>1</sup>, позволяет web странице открывать внутри себя вложенные страницы, например, с видеоплеером.

Кроме того ряд крупных видеохостингов предоставляют более обширное API для интеграции контента. В частности, некоторые плееры открывают доступ к так называемым событиям плеера, то есть оповещениям о некоторых событиях, которые дают возможность оценивать состояние контента. К таким событиям чаще всего относятся события старта, паузы и ошибки плеера. Примером плеера, который предоставляет подобную функциональность может служить плеер видеохостинга YouTube<sup>2</sup>.

### 2.3 Yandex Video

Yandex Video<sup>3</sup> (яндекс видео, видеопоиск) — это сервис, позволяющий пользователям смотреть видеозаписи, собранные с множества различных видеохостингов (рисунок 2). Из-за большого объема контента и невозможности моментально реагировать на его изменения, хранить видео на локальных серверах не представляется возможным, а потому видеозаписи показываются пользователям в виде интегрированных с помощью тэга iframe видеоплееров.

<sup>1</sup> Scott O'Hara и др. *HTML 5.3. W3C Working Draft*. <https://www.w3.org/TR/2018/WD-html53-20181018/>. W3C, окт. 2018.

<sup>2</sup> *YouTube Player API Reference for iframe Embeds*. Май 2018. URL: [https://developers.google.com/youtube/iframe\\_api\\_reference](https://developers.google.com/youtube/iframe_api_reference).

<sup>3</sup> *Yandex Video*. 2019. URL: <http://yandex.ru/video/>.

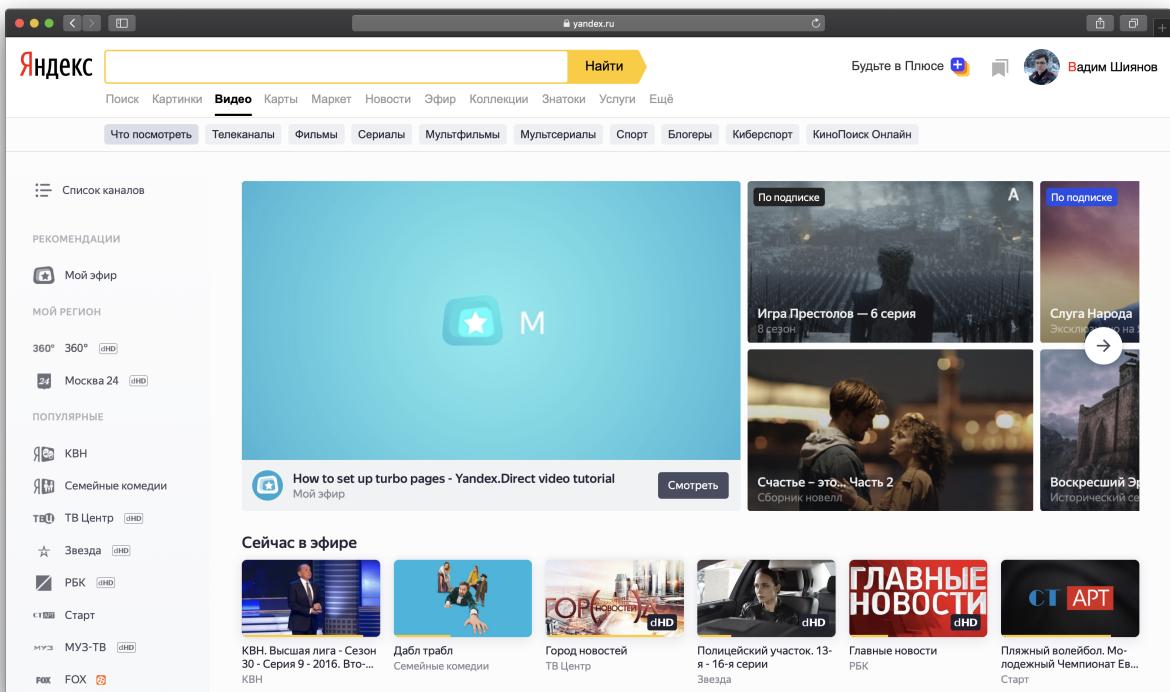


Рисунок 2 — Сервис Yandex Video.

## 2.4 Yandex Toloka

Yandex Toloka<sup>4</sup> (яндекс толока, толока) — это сервис, позволяющий публиковать некоторые несложные задания, которые другие пользователи могут выполнять за материальное вознаграждение (рисунок 3). Данный сервис использовался в данной работе для получения экспертной разметки данных: пользователям, взявшимся за выполнение задания, показывалась iframe с видеозаписью и требовалось ответить, проигрывается эта видеозапись или нет.

## 2.5 Неработающие видео

Под неработающим видео (рисунок 4) в данной работе поднимается любая видеозапись, которую пользователь не может просмотреть. Videозапись может прекратить работу по ряду причин: отключение сервера видеохостинга, на котором данная видеозапись хранилась или блокировка просмотра видеозаписи на основании обращения правообладателя контента.

<sup>4</sup>Yandex Toloka. 2019. URL: <https://toloka.yandex.com>.

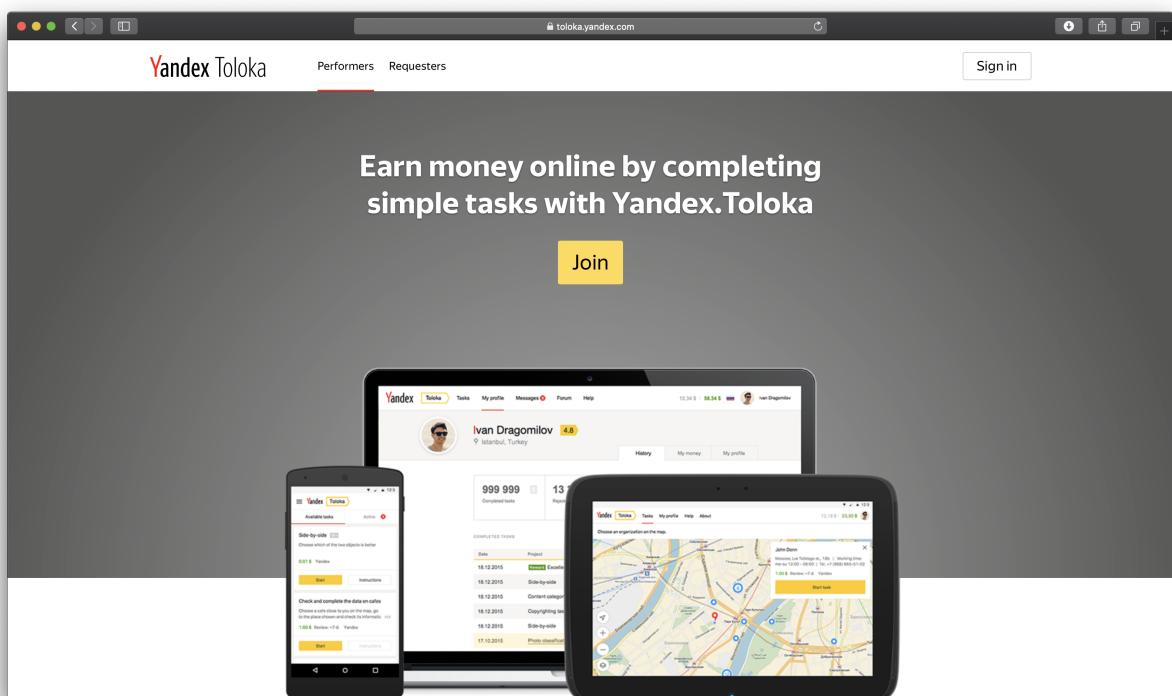


Рисунок 3 — Сервис Yandex Toloka.

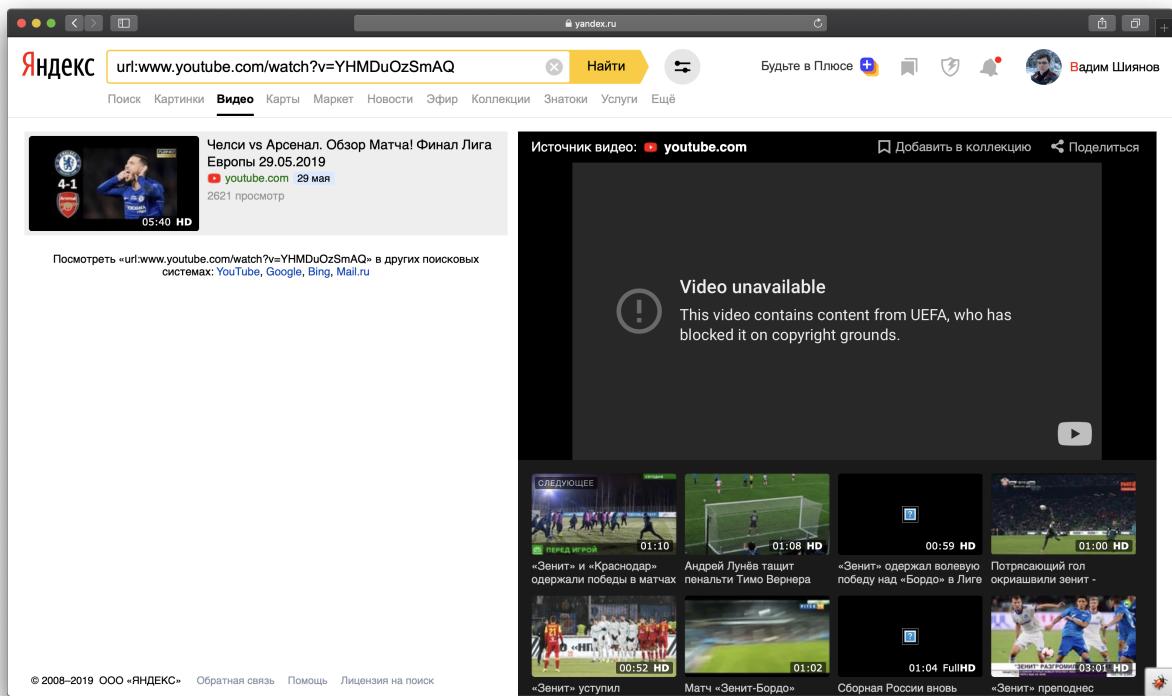
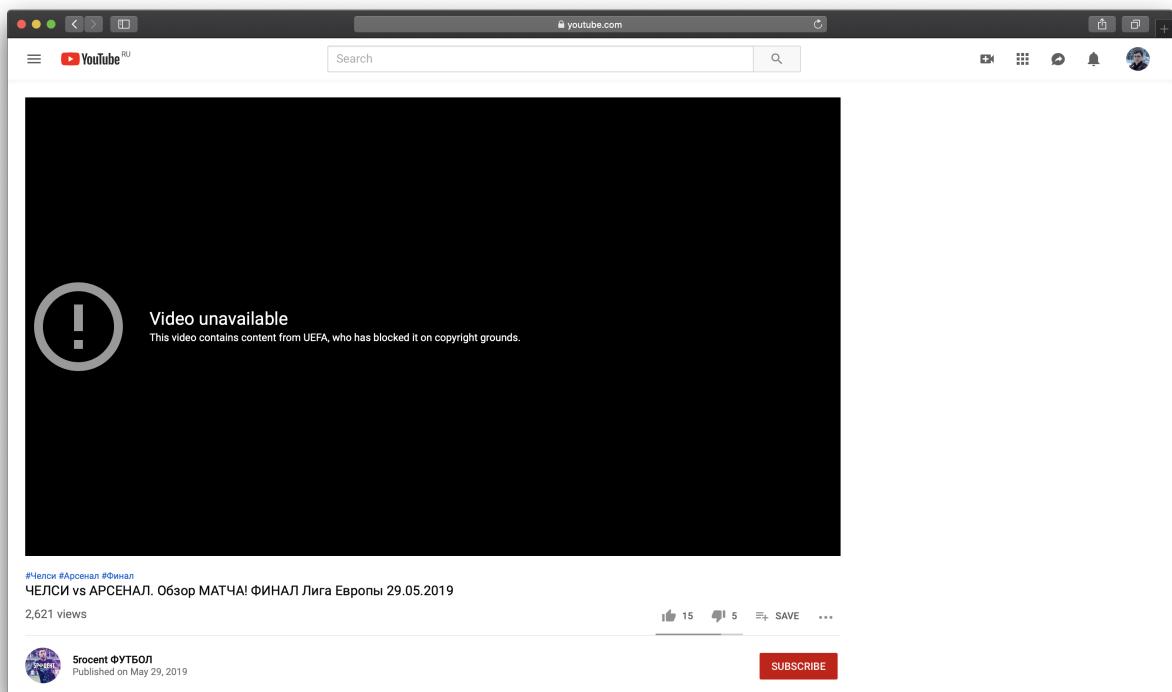


Рисунок 4. Пример неработающего видео на сайте [www.youtube.com](http://www.youtube.com) и на сайте Yandex Video.

## Введение

Выбор темы обусловлен спецификой работы сервиса Yandex Video. В силу того, что контент данного сервиса выдается пользователю в виде плееров, интегрированных в сайт с помощью тэга `iframe`, у данного сервиса возникает необходимость вовремя детектировать неработающие видеозаписи и убирать их из выдачи. На данный момент эта задача решается с помощью периодического обхода видеозаписей с помощью специального механизма, умеющего эмулировать работу клиентского web браузера, а также умеющего нажимать на кнопку начала проигрывания видеозаписи и оценивать, началось ли воспроизведение. Данный механизм позволяет получать достаточно достоверный сигнал о недоступности видео, однако обладает рядом существенных минусов, например, невозможность проверить доступность видео в странах кроме России (так как все запросы территориально отправляются из России), то есть учесть специфику локальных блокировок сайтов и контента. Для устранения подобных недостатков было предложено использовать статистику просмотров для детектирования неработающих документов.

С этой целью по сессиям пользователей собирается анонимная статистика: когда пользователь начинает просмотр видеозаписи, включается таймер, когда же пользователь переключает видео или закрывает страницу сервиса, время просмотра записывается для дальнейшего анализа. Вместе со временем просмотра сохраняются и события плеера, если они доступны. Основной проблемой таких данных является высокая степень загрязненности. Если во время просмотра или переключения видео у пользователя пропадет интернет соединение, запись о просмотре может не прийти или содержать неверные данные. Также возможен случай, когда пользователь включает проигрывание видеозаписи, а затем отходит от компьютера. В это время таймер считает время просмотра, несмотря на то, что видео могло оказаться неработающим и так и не загрузиться.

Задачей данной работы ставится создание модели, которая смогла бы предсказывать неработающие видео с высокой точностью, чтобы данный механизм мог дополнить, а в перспективе и заменить механизм обхода видеозаписей.

В качестве метода исследования был выбран экспериментальный подход: сбор и разметка выборки, оценка точности и полноты классификации с помощью метода перекрестной проверки и анализ предсказаний полученной модели.

## Предыдущие работы

Наиболее близкой задачей к поставленной является задача детектирования soft 404 страниц. В соответствии с протоколом HTTP<sup>1</sup>, если при обращении к серверу клиент запрашивает документ, который не доступен по той или оной причине, сервер должен возвращать ошибку. Как правило это ошибка 403 (Forbidden Error) или 404 (Not Found). Тем не менее многие интернет ресурсы стремятся предоставить пользователю более дружественный интерфейс и вместо возврата ошибки имитируют нормальную работу, возвращая код возврата 200 (OK), а вместо запрашиваемого документа отображают страницу, оповещающую пользователя о причинах недоступности контента.

Подобное поведение делают задачу детектирования недоступных документов нетривиальной. Первая попытка борьбы с данной проблемой была предпринята в статье от 2004 года<sup>2</sup>. Авторы данной статьи предлагают спровоцировать сервер вернуть заведомо недоступный документ путем прибавления к имеющемуся названию искомого документа произвольного окончания. После этого предлагается сравнивать поведение сервера в случае обращения к исходному документу и к заведомо недоступному. Второе упоминание о проблеме soft 404 датируется 2012 годом<sup>3</sup>. В отличии от первой статьи, данная работа целиком посвящена проблеме детектирования soft 404 страниц. Авторы этой статьи предлагают использовать классификатор, который использует в качестве данных лексические сигнатуры, содержащиеся в заголовке или в тексте страницы.

К сожалению ни один из предложенных способов детектирования soft 404 документов не может быть применен к задаче детектирования неработающих видео. Это вызвано тем, что, в отличии от других интернет ресурсов, видеохостинги при обращении к несуществующему или уже удаленному документу чаще всего предпочитают показывать страницу, полностью идентичную странице, которая была бы отображена в случае, если видео было доступно для просмотра. Сообщение о недоступности видеозаписи как правило располагается в самом видеоплеере. Таким образом подход, основанный на детектировании переадресаций на заранее заготовленные документы, содержащие оповещение о недоступности видео не применим так как в данном случае переадресаций не происходит. Анализ лексических сигнатур, содержа-

---

<sup>1</sup>Roy T. Fielding, Julian Reschke. *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. RFC 7231. Июнь 2014. DOI: 10.17487/RFC7231. URL: <https://rfc-editor.org/rfc/rfc7231.txt>.

<sup>2</sup>Ziv Bar-Yossef и др. “Sic transit gloria telae”. B: *Proceedings of the 13th conference on World Wide Web - WWW '04*. ACM Press, 2004. DOI: 10.1145/988672.988716.

<sup>3</sup>Luis Meneses, Richard Furuta, Frank Shipman. “Identifying “Soft 404” Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections”. B: *Theory and Practice of Digital Libraries*. Springer Berlin Heidelberg, 2012, с. 197—208. DOI: 10.1007/978-3-642-33290-6\_22.

щихся на странице и ее заголовке также неприменим в силу того, что и тот и другой текст не зависит от доступности видеозаписи, а само сообщение о возникшей проблеме находится в плеере. Таким образом для детектирования таких сообщений требуется технически сложный механизм, который умеет эмулировать доступ к странице из пользовательского web браузера и попытку воспроизведения видеозаписи.

## Подготовка данных

### 5.1 Исходные данные

В качестве исходных данных имеется таблица, каждая строка которой содержит некоторый идентификатор плеера, идентификатор видеозаписи, а также время просмотра и события плеера, если они доступны. Также из разметки сайта провайдера контента берется продолжительность видеозаписи. Пример исходных данных представлен в таблице 1.

Плеер	Видео	Длительность, с	Просмотр, с	Старт	Ошибка
Плеер1	Видео1	100	10	True	
Плеер2	Видео2	200	180	True	
Плеер3	Видео3	0	5		True
...	...	...	...	...	...

Таблица 1 — Пример исходных данных.

### 5.2 Выбор признаков

Так как данные сильно зашумлены, клики необходимо усреднять. При этом события плеера — достаточно надежный источник информации, тем не менее они имеют несколько существенных недостатков. К этим недостаткам относится тот факт, что события плеера доступны не для всех плееров, а также каждомуциальному плееру характерен некоторый профиль событий для различных состояний видеозаписи, однако эти профили могут существенно отличаться для двух разных плееров. Существуют два различных способа борьбы с данными недостатками: во-первых можно сделать больший упор на время просмотра и производимые из него признаки, как на более универсальный источник информации, во-вторых можно попытаться сделать отдельную модель для каждого плеера, чтобы она смогла в большей мере использовать сигнал событий плеера, если они доступны. В ходе данной работы были опробованы оба подхода.

Также в силу высокого уровня зашумленности исходных данных понятно, что их необходимо каким-то образом усреднять. В случае с данным событием видеоплеера усреднение очевидно — если событие произошло, то данной записи ставится в соответствие единица, в противном случае ноль и в качестве признака используется среднее значение, то есть доля просмотров, в которых данное событие наблюдалось.

Из времени просмотра было выбрано сгенерировано большее количество признаков исходя из некоторых предположений о природе данных. Первое предположение заключается в

том, что, если видеозапись не работает, большинство пользователей заметят это и переключать видео за довольно короткий промежуток времени, близкий к 30-40 секундам. Поэтому в качестве первой группы признаков были выбраны доли пользователей смотревших видео не менее 15, 30 и 45 секунд. Следующая идея говорит о том, что информация о том, как хорошо пользователи смотрят данное видео хранится в распределении доли просмотра контента. Для простоты и однообразности расчета распределение было введено в выборку в виде долей пользователей, досмотревших видео до 5, 10, 20, 30, 40, 50, 60, 70, 80 и 90 % длительности видеозаписи. Последнее наблюдение говорит о том, что если видеозапись попала на выдачу, то когда-то она была обнаружена обходом, а значит в какой-то момент времени она работала. Таким образом чаще всего мы занимаемся не поиском неработающего контента, а пытаемся поймать момент, в который видеозапись перестала работать. В таких случаях принято говорить о данных в форме временного ряда, то есть последовательных значениях, зависящих от времени. Однако для удобства работы с данными и расширения круга потенциальных алгоритмов классификации хотелось бы иметь данные в виде вектора фиксированной размерности. Также важно помнить о шуме в данных, а значит и необходимости усреднения данных. Для соблюдения всех ограничений было принято решение воспользоваться методом скользящего среднего. В таком случае мы получаем, например, такие признаки как среднее время просмотра за 5 последних просмотров заканчивая последним записанным просмотром и такое же среднее за 5 просмотров заканчивая предпоследним просмотром и так далее. Хотелось бы заметить, что подход со скользящим средним применим не только к времени просмотра. Скользящее среднее также считалось для событий плеера и долей смотревших видео хотя бы 15, 30, 45 секунд и так далее.

В итоге для каждого видео мы получили вектор признаков, размерность которого составила порядка 150. После того как были выбраны признаки и собрана выборка, разметка видеозаписей производилась с помощью сервиса Yandex Toloka. Таким образом был собран датасет из порядка 100000 размеченных объектов.

# Обучение классификатора

## 6.1 Выбор модели

Заметим, что пространство признаков имеет очень большую размерность. К тому же многие из признаков сильно взаимосвязаны с другими (действительно, доля досмотревших до 90 % видео не может быть выше доли досмотревших до 50 %). Если с размерностью можно справиться, например, воспользовавшись методом главных компонент<sup>1</sup>, то избавиться от зависимости между признаками гораздо сложнее. В данной работе было принято решение использовать модели машинного обучения, основанные на решающих деревьях<sup>2</sup>. Это позволяет исключить из предобработки уменьшение размерности признаков, а также перестать беспокоиться о зависимостях в признаках. Однако решающие деревья в чистом виде используются редко из-за высокой чувствительности к значениям гиперпараметров и низкой обобщающей способностью модели. Однако решающие деревья показывают очень хорошие результаты, когда они используются в виде разных ансамблей.

В качестве базовой модели был выбран случайный лес<sup>3</sup>. Данный выбор обусловлен тем что метод зарекомендовал себя как неприхотливый алгоритм, который тем не менее показывает довольно высокие результаты практически во всех задачах. Также было замечено, что выборка в 100000 объектов не очень велика. Также важно понимать, что среди этих объектов только около 1 % видеозаписей были размечены как неработающие. Таким образом выборка содержит довольно мало сигнала, который мы хотим детектировать. Поэтому в качестве дополнительной модели был выбран метод чрезвычайно случайных деревьев (extremely randomized trees)<sup>4</sup>, который позволяет в некоторых случаях уменьшить разброс предсказаний за счет небольшого увеличения смещения предсказания. В качестве последней модели был выбран метод градиентного бустинга (gradient boosting)<sup>5</sup> на решающих деревьях, как

<sup>1</sup>Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. B: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (нояб. 1901), с. 559—572. DOI: 10 . 1080 / 14786440109462720.

<sup>2</sup>California USA) Breiman Jerome (Stanford University California USA) Friedman Charles J. (University of California Berkeley USA) Stone R. A. (Stanford California USA) Olshen Leo (Consultant Berkeley. *Classification and Regression Trees*. Taylor & Francis Ltd, 1 янв. 1984. 368 с. ISBN: 0412048418. URL: [https://www.ebook.de/de/product/3606994/leo\\_consultant\\_berkeley\\_california\\_usa\\_breiman\\_jerome\\_stanford\\_university\\_california\\_usa\\_friedman\\_charles\\_j\\_university\\_of\\_california\\_berkeley\\_usa\\_stone\\_r\\_a\\_stanford\\_california\\_usa\\_olshen\\_classification\\_and\\_regression\\_trees.html](https://www.ebook.de/de/product/3606994/leo_consultant_berkeley_california_usa_breiman_jerome_stanford_university_california_usa_friedman_charles_j_university_of_california_berkeley_usa_stone_r_a_stanford_california_usa_olshen_classification_and_regression_trees.html).

<sup>3</sup>Tin Kam Ho. “Random decision forests”. B: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press, 1995. DOI: 10 . 1109/icdar . 1995 . 598994.

<sup>4</sup>Pierre Geurts, Damien Ernst, Louis Wehenkel. “Extremely randomized trees”. B: *Machine Learning* 63.1 (март 2006), с. 3—42. DOI: 10 . 1007/s10994-006-6226-1.

<sup>5</sup>Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine.” B: *The Annals of Statistics*

наиболее точной хотя и наименее неприхотливый алгоритм.

Численные эксперименты производились с помощью языка программирования Python<sup>6</sup>. Реализации первых двух алгоритмов были взяты из библиотеки scikit-learn<sup>7</sup>, реализация последнего была взята из библиотеки CatBoost<sup>8</sup>.

## 6.2 Обучение модели

Для подбора гиперпараметров и получения оценки точности и полноты использовался сеточный поиск с оценкой точности и полноты с использованием 5-кратной перекрестной проверки. Для моделей random forest и extremely randomized trees подбиралось количество деревьев в ансамбле. Для классификатора CatBoost кроме того подбиралась максимальная глубина дерева.

Результаты сеточного поиска представлены в таблице 2, а также на рисунках 5, 6 и 7. При этом в таблице 2 модели extremely randomized trees и классификатору из библиотеки CatBoost в таблице соответствуют по две строки каждому. Это сделано для того, чтобы показать метрики и для модели с наибольшей точностью, и для модели с наивысшей полнотой. Для модели random forest разница между точностью для этих двух классификаторов несущественна, поэтому представлены только значения метрик для модели с наивысшей полнотой.

Модель	Точность, %	Полнота, %	Число деревьев	Глубина
RandomForestClassifier	72.0	50.8	500	
ExtraTreesClassifier	74.3	47.0	1400	
ExtraTreesClassifier	73.9	47.2	1100	
CatBoostClassifier	<b>80.3</b>	51.4	650	3
CatBoostClassifier	77.5	<b>53.2</b>	500	4

Таблица 2 — Результаты сеточного поиска

Как можно видеть, модель extremely randomized trees действительно смогла немного увеличить точность классификации по сравнению с результатами базовой модели. Однако при этом модель оказалась слегка более смещенной из-за чего снизилась полнота. Лучше всех себя показал классификатор из библиотеки CatBoost. Именно с использованием данного алгоритма была достигнута наилучшая в данном эксперименте точность классификации в 80.3 %. При этом полнота классификации у данной модели оказалась выше, чем у моделей RandomForestClassifier и ExtraTreesClassifier. Наибольшей полноты классификации в данном эксперименте (51.4 %) также удалось достичь именно классификатору из библиотеки CatBoost, правда, при других значениях гиперпараметров.

29.5 (окт. 2001), с. 1189—1232. DOI: 10.1214/aos/1013203451.

<sup>6</sup>*Python Programming Language*. 2019. URL: <https://www.python.org>.

<sup>7</sup>F. Pedregosa и др. “Scikit-learn: Machine Learning in Python”. В: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.

<sup>8</sup>Prokhorenkova и др., “CatBoost: unbiased boosting with categorical features”.

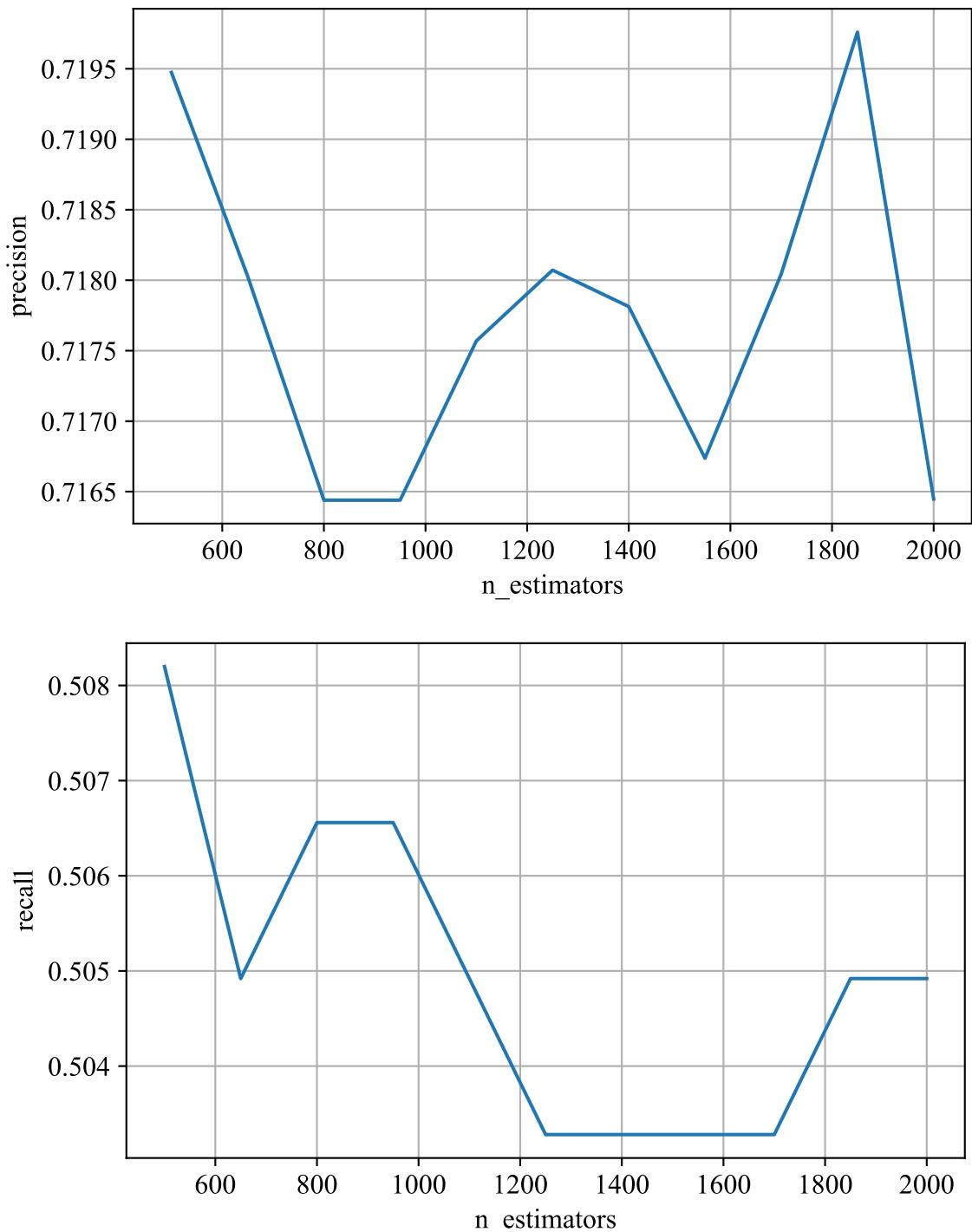


Рисунок 5 — Результаты сеточного поиска для модели RandomForestClassifier

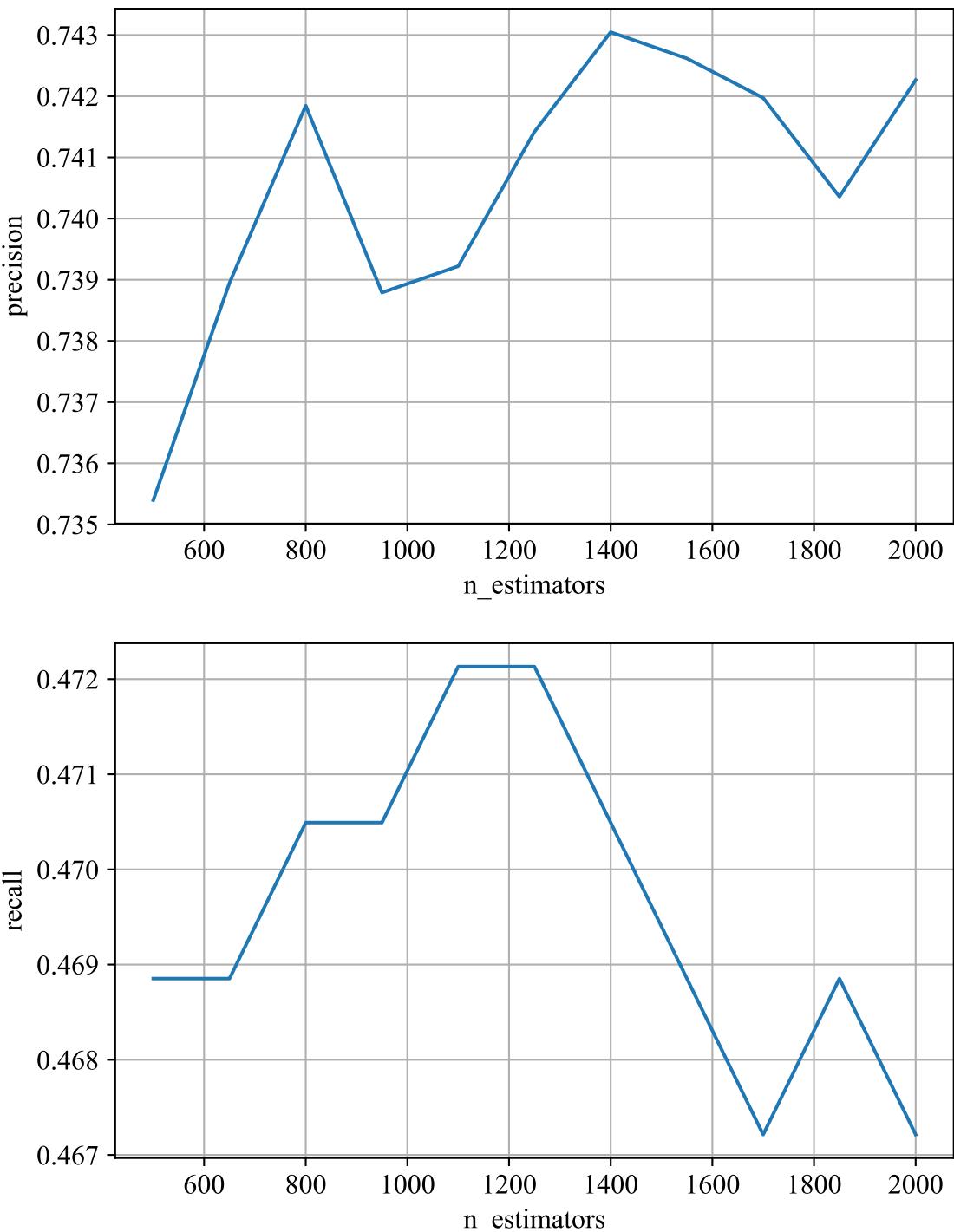


Рисунок 6 — Результаты сеточного поиска для модели ExtraTreesClassifier

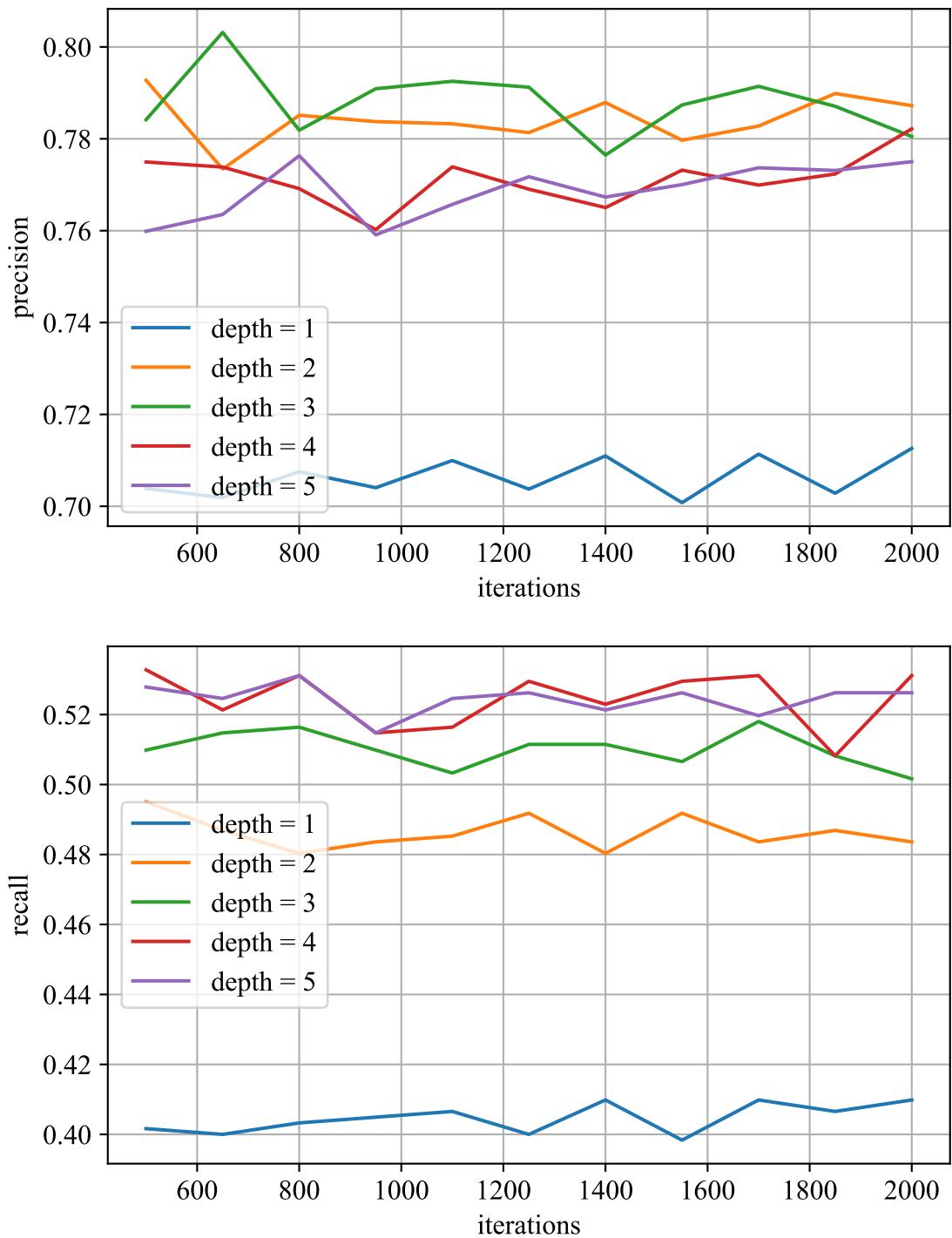


Рисунок 7 — Результаты сеточного поиска для модели CatBoostClassifier

### 6.3 Анализ модели

После численного эксперимента был произведен анализ наилучшей модели. В результате этого анализа были выявлены две существенные закономерности. Во-первых не смотря на преобладающее количество признаков, полученных из времени просмотра, события плеера, а именно усредненные события старта и ошибки, обладают высоким весом в финальной модели. Вторым важным фактом, обнаруженным во время анализа модели, является то, что выборка содержит в себе заметную долю видеозаписей, неверно размеченных как работающие. Как выяснилось в дальнейшем, это вызвано спецификой работы сервиса Yandex Toloka, который склонен верить хорошо зарекомендовавшим себя пользователям в случае, если они размечают видеозапись, как работающую (в силу того, что достоверно известно, что неработающих видеозаписей всего около 1 %).

Напомню, что изначальным желанием было избавиться или практически избавиться от влияния событий плеера на предсказание модели, так как, несмотря на аккуратность и информативность такого сигнала в целом, они доступны не для всех видеозаписей, а также характерный профиль событий различается от плеера к плееру. Тем не менее модель посчитала данные признаки информативными и смогла добиться достаточно высокой точности. В связи с этим было предложено добавить в данные идентификатор плеера, как категориальный признак, что позволило бы модели корректировать ожидаемый профиль событий в зависимости от их источника. Так как из всех используемых моделей с категориальными признаками умеет работать только CatBoost (для остальных моделей потребовалась бы дополнительная предобработка данных), было принято решение ограничиться им. Результаты эксперимента представлены в таблице 3, а также на рисунке 8.

Модель	Точность, %	Полнота, %	Число деревьев	Глубина
CatBoostClassifier	<b>72.6</b>	65.0	1250	2
CatBoostClassifier	71.3	<b>68.1</b>	800	5

Таблица 3 — Результаты эксперимента с категориальным признаком.

Как можно видеть, точность классификации упала на 7.7 % до 72.6 %, однако полнота возросла на 14.9 % и составила 68.1 %. Это говорит скорее всего о переобучении модели на новый категориальный признак, то есть модель просто запомнила, что видеозаписи некоторых плееров чаще всего не работают и прекратила попытки честно их классифицировать.

### 6.4 Одноплеерная модель

Для решения проблемы неаккуратной разметки данных, было принято решение попробовать размечать выборку с помощью данных обхода. В таком случае мы получаем очень достоверный сигнал о том, какие видеозаписи не работали. Однако возникает обратная проблема: у нас нет достоверной информации о том, какие работали. Для решения данной проблемы был выбран следующий подход: известно, что среди видеозаписей, которые представлены на выдаче поиска, не работает примерно 1 %. Исходя из этого, при составлении

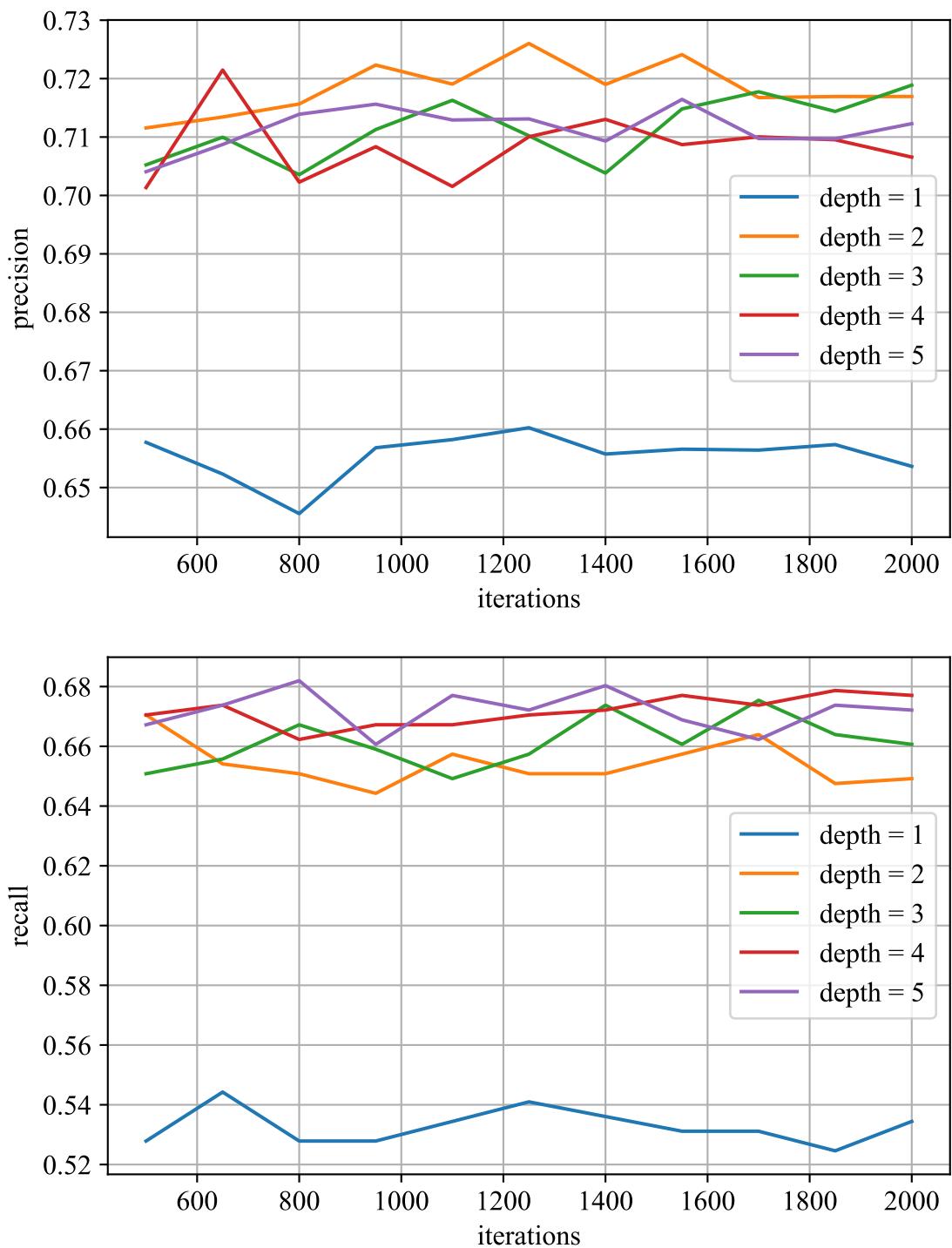


Рисунок 8 — Результаты эксперимента с категориальным признаком.

выборки мы считаем количество достоверно неработавших видео исходя из данных обхода и размечаем эти видеозаписи. Остальные видео сортируем в убывающем порядке по среднему времени просмотра и отбираем столько, чтобы баланс классов составлял 99 к 1. Таким образом мы получаем размеченную выборку большого размера, так как история обхода доступна за довольно длительный промежуток времени. Это позволяет ограничиться данными единственного плеера, что позволит модели в полной мере использовать данные событий плеера. Так как в результирующей выборке получилось около 1000000 объектов, время обучения модели стало занимать заметно большее время. Поэтому было принято решение ограничиться в данном эксперименте только хорошо зарекомендовавшим себя классификатором из библиотеки CatBoost. Результаты сеточного поиска представлены в таблице 4 и на рисунке 9.

Модель	Точность, %	Полнота, %	Число деревьев	Глубина
CatBoostClassifier	<b>89.3</b>	51.4	1550	3
CatBoostClassifier	88.2	<b>54.7</b>	1850	5

Таблица 4 — Результаты сеточного поиска для модели обученной на данных обхода.

Эксперимент показал прирост точности классификации в 9.1 %. Сама точность составила 89.3 %, что является хорошим результатом учитывая природу данных и их зашумленность. При этом модель показала полноту, которая не существенно отличается от случая модели, обученной на данных размеченных с помощью сервиса Yandex Toloka. Наилучшая полнота, которая была достигнута в данном эксперименте (54.7 %) и вовсе выше, чем в случае экспертной разметки данных.

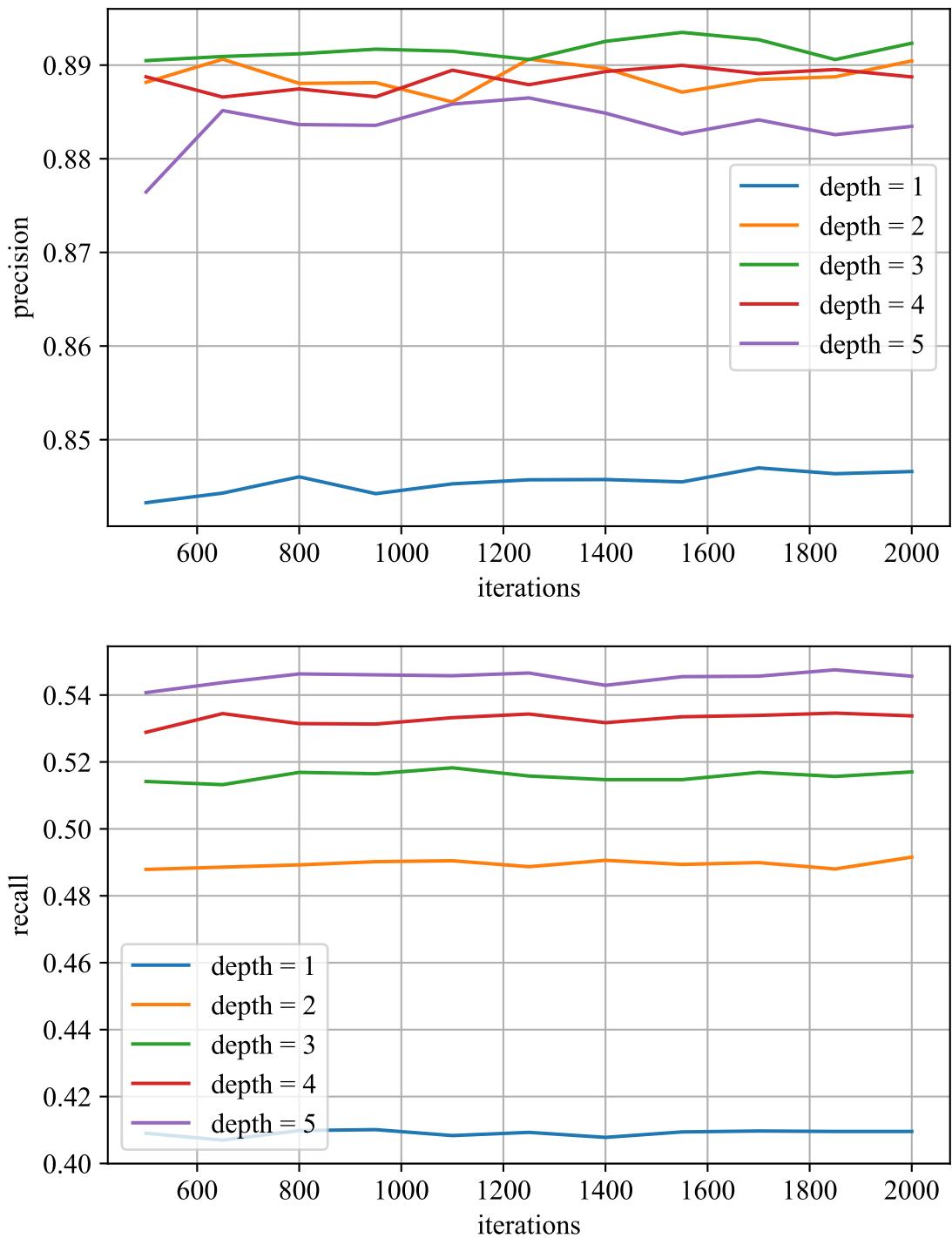


Рисунок 9 — Результаты сеточного поиска для модели обученной на данных обхода

## Заключение

Данная работа была посвящена изучению частного случая проблемы soft 404, а именно детектированию неработающих видеозаписей. Так как ни один из ранее предложенных способов решения данной проблемы не подходит в данном случае из-за специфики задачи, было предложено использовать статистику просмотров для детектирования неработающих видеозаписей. Благодаря такому подходу удалось обучить классификатор из библиотеки CatBoost, который показал точность классификации 80.3 % и полноту 51.4 %. Также был предложен способ ограничиться одним плеером, что позволило увеличить точность классификации на 9 % при практически неизменной полноте. Таким образом был построен классификатор, обладающий точностью предсказания 89.3 % и полнотой 51.4 %.

## Список литературы

- Bar-Yossef, Ziv и др. “Sic transit gloria telae”. B: *Proceedings of the 13th conference on World Wide Web - WWW '04*. ACM Press, 2004. DOI: 10.1145/988672.988716.
- Fielding, Roy T., Julian Reschke. *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. RFC 7231. Июнь 2014. DOI: 10.17487/RFC7231. URL: <https://rfc-editor.org/rfc/rfc7231.txt>.
- Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” B: *The Annals of Statistics* 29.5 (окт. 2001), с. 1189—1232. DOI: 10.1214/aos/1013203451.
- Geurts, Pierre, Damien Ernst, Louis Wehenkel. “Extremely randomized trees”. B: *Machine Learning* 63.1 (март 2006), с. 3—42. DOI: 10.1007/s10994-006-6226-1.
- Ho, Tin Kam. “Random decision forests”. B: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press, 1995. DOI: 10.1109/icdar.1995.598994.
- Leo (Consultant Berkeley, California USA) Breiman Jerome (Stanford University California USA) Friedman Charles J. (University of California Berkeley USA) Stone R. A. (Stanford California USA) Olshen. *Classification and Regression Trees*. Taylor & Francis Ltd, 1 янв. 1984. 368 с. ISBN: 0412048418. URL: [https://www.ebook.de/de/product/3606994/leo\\_consultant\\_berkeley\\_california\\_usa\\_breiman\\_jerome\\_stanford\\_university\\_california\\_usa\\_friedman\\_charles\\_j\\_university\\_of\\_california\\_berkeley\\_usa\\_stone\\_r\\_a\\_stanford\\_california\\_usa\\_olshen\\_classification\\_and\\_regression\\_trees.html](https://www.ebook.de/de/product/3606994/leo_consultant_berkeley_california_usa_breiman_jerome_stanford_university_california_usa_friedman_charles_j_university_of_california_berkeley_usa_stone_r_a_stanford_california_usa_olshen_classification_and_regression_trees.html).
- Meneses, Luis, Richard Furuta, Frank Shipman. “Identifying “Soft 404” Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections”. B: *Theory and Practice of Digital Libraries*. Springer Berlin Heidelberg, 2012, с. 197—208. DOI: 10.1007/978-3-642-33290-6\_22.
- O’Hara, Scott и др. *HTML 5.3*. W3C Working Draft. <https://www.w3.org/TR/2018/WD-html53-20181018/>. W3C, окт. 2018.
- Pearson, Karl. “LIII. On lines and planes of closest fit to systems of points in space”. B: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (нояб. 1901), с. 559—572. DOI: 10.1080/14786440109462720.
- Pedregosa, F. и др. “Scikit-learn: Machine Learning in Python”. B: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.
- Prokhorenkova, Liudmila и др. “CatBoost: unbiased boosting with categorical features”. B: (28 июня 2017). arXiv: <http://arxiv.org/abs/1706.09516v5> [cs.LG].

*Python Programming Language*. 2019. URL: <https://www.python.org>.

*Yandex Toloka*. 2019. URL: <https://toloka.yandex.com>.

*Yandex Video*. 2019. URL: <http://yandex.ru/video/>.

*YouTube Player API Reference for iframe Embeds*. Май 2018. URL: [https://developers.google.com/youtube/iframe\\_api\\_reference](https://developers.google.com/youtube/iframe_api_reference).