



Факультет компьютерных наук

Прикладная математика и
информатика

Исследовательский проект

Анализ и визуализация данных с использованием Yandex DataLens: исследование по прогнозированию CTR

Data Analysis and Visualization Using Yandex DataLens: A Case Study of Click-Through Rate Prediction

Выполнил:

студент группы БПМИ-247
Грицан Максим Андреевич

Принял руководитель проекта:

Доцент, Базовая кафедра Т-Банка
Иванов Андрей Александрович



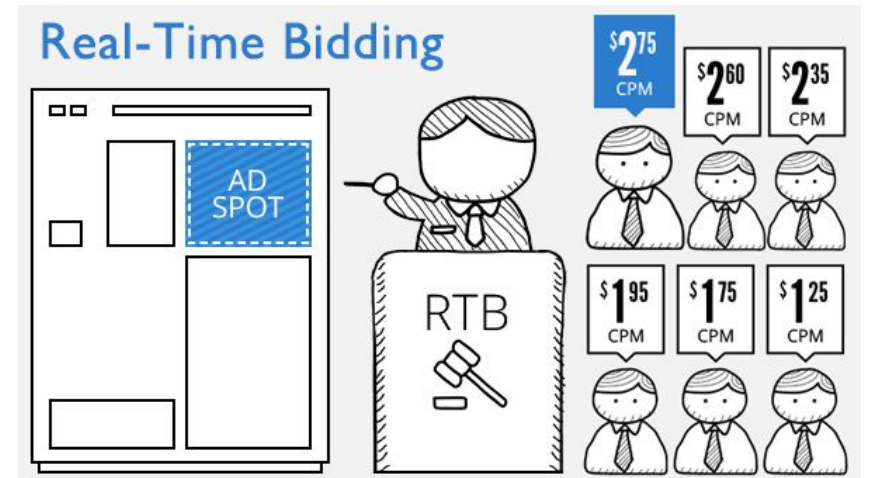
Предметная область и актуальность работы

В современном мире интернет-реклама — ключевой инструмент продвижения товаров и услуг. В частности рынок мобильной рекламы.

Этот рынок функционирует на базе аукционов реального времени (RTB), где за доли секунды принимается решение о показе конкретного рекламного баннера конкретному пользователю.

Эффективность рекламных кампаний напрямую зависит от точности прогнозирования кликабельности (CTR) объявлений.

Разработка методов анализа и визуализации данных для оценки факторов, влияющих на CTR, может позволить оптимизировать рекламные стратегии и повысить возврат инвестиций.





Цель и задачи работы

Цель: проведение разведочного анализа данных (EDA) с использованием платформы Yandex DataLens для выявления зависимостей, влияющих на вероятность клика (CTR).

Задачи:

Исследовательская часть:

- Исследовать набор данных
- Провести обзор источников, использующих набор данных.
- Пройти курс «DataLens: анализ и визуализация данных»

Программная часть:

- Применить алгоритм CatBoost для предсказания вероятности клика
- Разработать интерактивный дашборд



Набор данных Avazu Click-Through Rate Prediction

Набор данных представляет собой коллекцию данных о показах мобильной рекламы, собранных компанией Avazu в течение 10 дней.

Основные характеристики набора данных:

- Размер: 40 428 967 строк.
- Каждая запись содержит информацию об одном показе рекламного объявления и факте клика по нему.
- Период сбора: 10 дней (21 октября – 30 октября 2014 года).
- Целевая переменная: бинарная (клик/не клик).
- Количество признаков: 23.

Поле	Тип	Описание
id	string	Уникальный идентификатор записи
click	integer	Целевая переменная: 0 (нет клика), 1 (клик)
hour	integer	Временная метка в формате YYMMDDHH
C1	integer	Анонимизированная категориальная переменная
banner_pos	integer	Позиция баннера на странице
site_id	string	Идентификатор сайта
site_domain	string	Домен сайта
site_category	string	Категория сайта
app_id	string	Идентификатор приложения
app_domain	string	Домен приложения
app_category	string	Категория приложения
device_id	string	Идентификатор устройства
device_ip	string	IP-адрес устройства
device_model	string	Модель устройства
device_type	integer	Тип устройства
device_conn_type	integer	Тип подключения устройства
C14	integer	Анонимизированная категориальная переменная
...	integer	Анонимизированная категориальная переменная
C21	integer	Анонимизированная категориальная переменная



Обзор академических источников

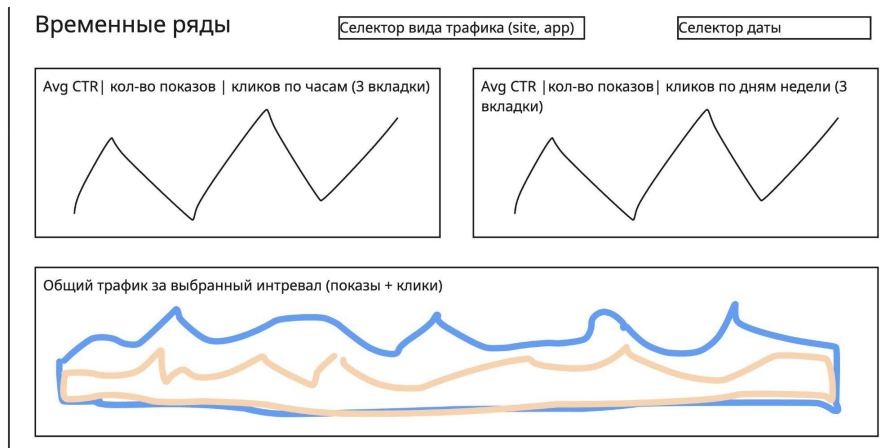
Авторы	Название статьи	Краткая выжимка	Роль набора данных в работе
Li Li	Click-Through-Rate Prediction Using Deep Neural Networks and Efficient Channel Attention Mechanisms	Предложена модель предсказания CTR на основе глубоких нейронных сетей (DNN) в сочетании с механизмом эффективного канального внимания (ECANet).	Использовался в качестве одного из основных наборов данных для валидации предложенной модели.
Zhu Jieming, Liu Jinyang, Yang Shuai, Zhang Qi, He Xiuqiang	Open Benchmarking for Click-Through Rate Prediction	Представили первый открытый бенчмарк для предсказания CTR, направленный на решение проблем воспроизводимости и несогласованности результатов в существующих исследованиях.	Был взят одним из основных набор данных для бенчмаркинга.
Damaševičius Robertas, Zailskaitė-Jakštė Ligita	Usability and Security Testing of Online Links: A Framework for Click-Through Rate Prediction Using Deep Learning	Предложена модель Human-Centric Cyber Security (HCCS): фреймворк 3U (User, Usage, Usability) с методами искусственного интеллекта для предсказания CTR в контексте кибербезопасности и юзабилити веб-приложений.	Набор данных Avazu использовался для экспериментальной валидации.



Проектирование и разработка макета интерактивного дашборда

При проектировании архитектуры дашборда учитывались возможные потребности трех основных групп потенциальных пользователей:

- Менеджеры проекта, руководители, рекламные площадки и рекламодатели
- Дата-инженеры и дата-аналитики
- ML-инженеры



Для удовлетворения потребностей всех групп пользователей была выбрана многостраничная (модульная) архитектура. Для этого макет дашборда концептуально разделен на три независимые, но логически связанные группы вкладок:

- Продуктовые показатели рекламы и бизнес-метрики
- Структура набора данных
- Оценка модели CatBoost и важность признаков

В свою часть каждая группа разделялась на вкладки по принципу «от общего к частному»

Модель CatBoost, ее результаты и важность признаков

Log Loss модели

Roc Auc модели

Score модели на
kaggle

Результат на
kaggle



Обучение модели CatBoost для предсказания CTR

Модель CatBoost необходима для получения результатов в Kaggle соревновании, а также для построения чартов с третьей вкладки макета дашборда.

Все данные были разбиты на 2 выборки: обучающую и валидационную в соотношении 74:26. Также был использован Out-of-Time (OOT) подход, так как наши данные времязависимые. Обучение происходило на вычислительных ресурсах Kaggle. Сам набор данных очень объемный и целиком хранить его в оперативной памяти не представлялось возможным. Обучающая и валидационная выборки передавались в модель через Pool'ы, которые представлены в библиотеке CatBoost.



Подготовка и загрузка данных для дашборда

В качестве СУБД был выбран ClickHouse, поскольку Яндекс сам рекомендует его для работы с Yandex DataLens. Целиком загрузить за раз весь набор данных не получилось даже в ClickHouse, поэтому он отправлялся по частям с помощью Python-библиотеки *clickhouse_connect*. Она предоставляет удобный интерфейс для подключения к базе данных, выполнения SQL-запросов и пакетной вставки данных.

Для третьей группы вкладок дашборда необходимо получить метрики и сводные таблицы от обученной модели CatBoost. Библиотека предоставляет широкий функционал для изучения зависимостей и закономерностей, которая модель нашла в процессе обучения.

Датасеты		
	Lift результаты модели	wip-62 4 дня назад
	Данные для диаграммы надежности	wip-62 4 дня назад
	Датасет Avazu Click-Through Rate Prediction	wip-62 несколько секунд назад
	Качество связей признаков	wip-62 4 дня назад
	Метрики модели	wip-62 4 дня назад
	Полезность признаков	wip-62 4 дня назад
	Средний предсказанный и реальный CTR	wip-62 4 дня назад



Демонстрация дашборда построенного в Yandex DataLens



[Ссылка](#)



Основные результаты и выводы

- Обученная модель CatBoost на валидационной выборке показала следующие результаты: **ROC-AUC = 0.7548, LogLoss = 0.3831, Accuracy = 0.8449**. Также модель выбила LogLoss = 0.37913 на тестовой выборке, что дает **1045 место Avazu Click-Through Rate Prediction**.
- Успешно **разработан многостраничный дашборд в системе Yandex DataLens**, предназначенный для мониторинга ключевых рекламных метрик и анализа эффективности кампаний.
- С помощью дашборда было определено, что **CTR сильнее всего зависит от конкретного пользователя (его IP-адреса) и площадки (ID сайта и приложения), а их комбинации — показывают максимальную эффективность**. При этом на большинстве площадок аудитория изначально не настроена на рекламу, что снижает общую кликабельность, поэтому для успешного показа критически **важно использовать оптимальные позиции баннеров**.



Направление дальнейшей работы

- Автоматизация подбора гиперпараметров: замена эмпирического подбора ключевых конфигураций модели на систематический автоматизированный поиск.
- Эксперименты различными с моделями для предсказания CTR: градиентный бустинг – стандарт классического машинного обучения, но он не всегда выдает лучшие результаты. Возможно, можно подобрать оптимальную архитектуру для нейронной сети, чтобы улучшить качество и интерпретируемость предиктивной модели.
- Самое перспективное направление развития: переход к потоковой аналитике. Модернизация архитектуры хранения данных ClickHouse для обработки не исторических статичных данных, а непрерывных real-time потоков логов, поступающих напрямую с рекламных аукционов (RTB) в режиме реального времени.



Список использованных источников

1. Li L. Click-Through-Rate Prediction Using Deep Neural Networks and Efficient Channel Attention Mechanisms // Informatica. Slovenian Society Informatika, 2025. Т. 49, № 19. Сс. 21–36.
2. Jieming Z. и др. Open Benchmarking for Click-Through Rate Prediction // Proceedings of the 30th ACM International Conference on Information and Knowledge Management. Virtual Event, Queensland, Australia: ACM, 2021. Сс. 2759–2769.
3. Robertas D., Ligita Z.-J. Usability and Security Testing of Online Links: A Framework for Click-Through Rate Prediction Using Deep Learning // Electronics. MDPI, 2022. Т. 11. С. 400.
4. Kaggle. Avazu Click-Through Rate Prediction [Электронный ресурс]. 2025. URL: <https://www.kaggle.com/competitions/avazu-ctr-prediction/data> (дата обращения: 12.12.2025).

И 18 других, все источники можно найти в отчете на стр. 68

