Разработка модели машинного обучения для предсказания склонности клиента к покупке услуги на основе предоставленных и открытых данных

Команда "НейроДантес"
Дмитрий Курочкин, Артём Каличенко
Никита Гордеев, Тимофей Цвирко

1_техническое_задание

1) Цель

- Разработать модель машинного обучения, способную предсказывать склонность клиента к покупке тарифа домашнего интернета «Игровой»

2) Задачи

- Проанализировать ЦА
- Обработать тренировочные и тестовые данные
- Обогатить модель внешними данными
- Построить модель машинного обучения, используя внутренние и внешние данные

3) Технологии

- Язык Python
- Open Source библиотеки

4) Тариф "Игровой"

- Мощный интернет
- Игровые бонусы
- Выделенная команда поддержки

2_анализ_целевой_аудитории

Внутренние факторы

портрет игрока	WarGaming	My.Games	Форгейм
пол	98% мужчины 2% женщины	79% мужчины 21% женщины	------
возраст	28% 35-44 26% 45-54	42% 14 - 18 30% 19 - 30	------
аккаунтов	160 000 000	19 000 000	2 000 000
покупки в месяц	2000 рубей	------	------
время игры в день	2 часа	------	------

2_анализ_целевой_аудитории

Внешние факторы

репутация оператора	сбои в работе интернета	стоимость услуг	известность
осведомлённость клиента	TV / YouTube реклама	Рекламные баннеры	Контекстная реклама
политические новости	стабильность в стране	государственные праздники	конференции
технологические новости	видеоигры	комплектующие для ПК	криптовалюты
род деятельности	загруженность на работе	престиж	онлайн или офлайн
погодные условия	время года	температура на улице	развлечения
уровень жизни	переезд в новый дом	сидячий или активный образ жизни	средняя ЗП

3_обработка_и_обогащение_данных

1) Данные

- Данные от Ростелеком
- Миграция по городам
- Средняя зарплата по населенному пункту
- Инвестиции в основной капитал города

2) Преобразование признаков

- Label Encoder
- One-Hot Encoder
- Binary Encoder

3) Обоснование

- Числовое представление символьных данных приводит к повышению точности модели.

4) Конкурентное преимущество

- Использование многих внешних факторов
- Прогнозирование отсутствующей информации
- Высокая скорость обработки новых данных

4_построение_модели_машинного_обучения

1) Методы машинного обучения

- Ансамбль из градиентных бустингов над решающими деревьями

2) Обоснование

- Один из популярных методов решения проблемы классификации и выявлений аномалий
- Подходит для использовании в представленной задаче

3) Программное обеспечение

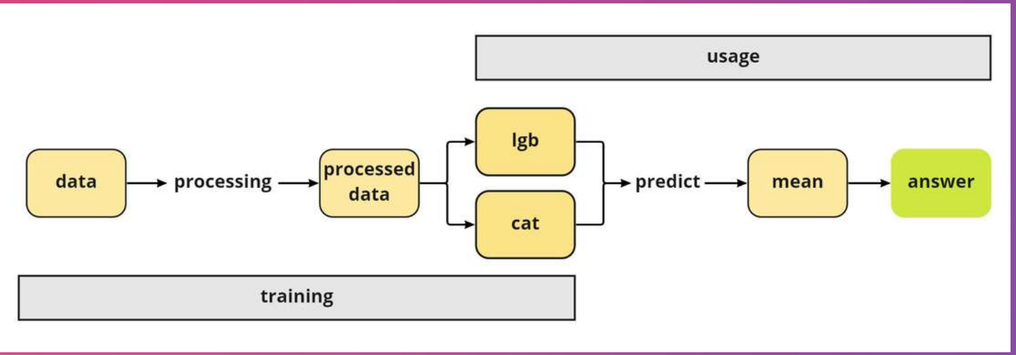
- Google Collab
- Python
- Open-source библиотеки: Pandas, NumPy, sklearn, CatBoost, LightGBM

4) Параметры

- Catboost: 300 деревьев
- LightGBM: 500 деревьев

5_точность_работы_алгоритма

1) Схема решения

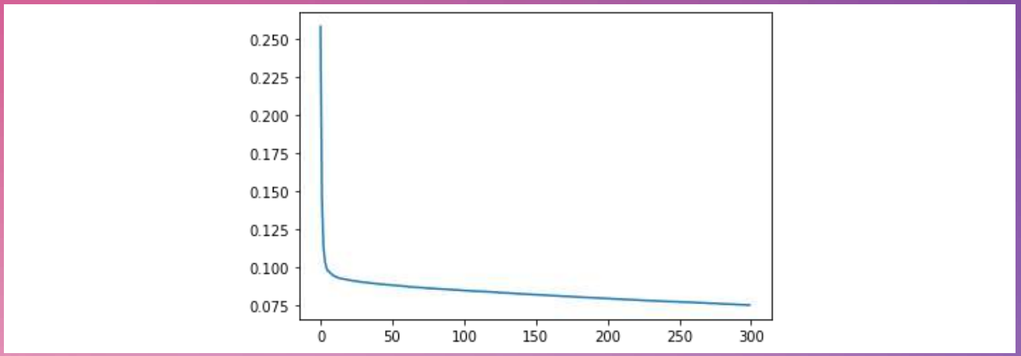


2) Кривая ошибок AUC-ROC

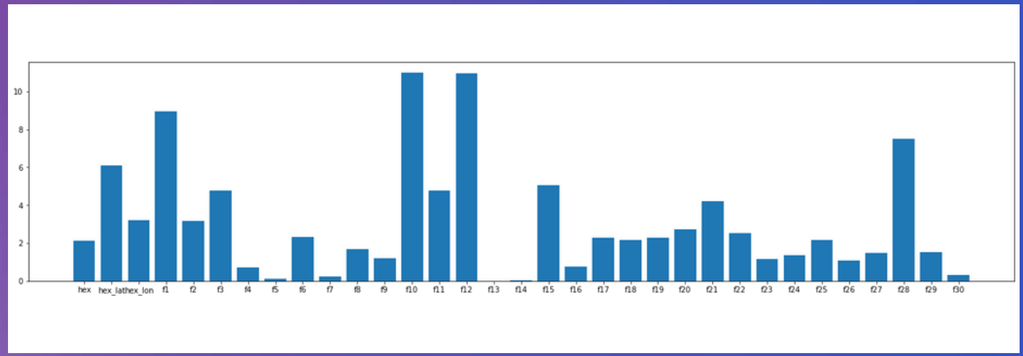
```
[ ]  roc_auc_score(df_test['label'], ensemble_pred)

     0.7470760534794361
```

3) Значение ошибки Logloss



4) Важность атрибутов



6_выводы

1) Адаптируемость

- Используемые нами внешние факторы универсальны, поэтому если обучить модель на другом тарифе, то её можно будет применить к этому тарифу.
- Если добавить новые внешние данные, то это повысит точность модели.

2) Результаты

- Мы смогли:
  - Проанализировать ЦА
  - Обработать тренировочные и тестовые данные
  - Построить модель машинного обучения, используя внутренние данные

3) Запускаемость кода

- Код запускается без ошибок и предупреждений.

4) Рекомендации

- Использовать больше внешних и внутренних факторов для предсказания склонности клиента к покупке и анализа целевой аудитории

7_роли_в_команде

Дмитрий Курочкин

- Программист
- ПетрГУ, Программная инженерия, 3 курс
- Поиск информации, работа над нейронной сетью, работа с таблицами

Артём Каличенко

- Программист / Спикер
- ПетрГУ, Прикладная математика и информатика, 3 курс
- Обработка данных, анализ ЦА, защита проекта

Никита Гордеев

- Программист / Дизайнер
- ПетрГУ, Программная инженерия, 2 курс
- Очистка данных, поиск информации, работа над нейронной сетью - варианты обучения 1 и 3, презентация

Тимофей Цвирко

- Программист
- ПетрГУ, Прикладная математика и информатика, 3 курс
- Поиск библиотек для работы с данными; разработка нейронной сети, варианты обучения 2, обучение моделей ML.

8_использованные_материалы

1. Работа с геоданными в Python и Jupyter // Proglib URL: https://proglib.io/p/rabota-s-geodannymi-v-python-i-jupyter-2021-03-22 (дата обращения: 26.08.2022).
2. Разбор маркетинга World of Tanks. Танки в маркетинге // ETDBOX URL: https://etdbox.ru/blog/6 (дата обращения: 27.08.2022).
3. Рекламный кабинет ВКонтакте // ВКонтакте URL: https://vk.com/adscreate (дата обращения: 27.08.2022).
4. Категориальные признаки // Хабр URL: https://habr.com/ru/post/666234/ (дата обращения: 27.08.2022).
5. Основы линейной регрессии // Хабр URL: https://habr.com/ru/post/514818/ (дата обращения: 27.08.2022).
6. Классификация с многими метками // Хабр URL: https://habr.com/ru/company/piter/blog/488362/ (дата обращения: 26.08.2022).
7. Ежедневная аудитория Warface превысила 700 тысяч человек // Игры Mail.Ru URL: https://games.mail.ru/pc/news/2014-04-18/ezhednevnaja_auditorija_warface_prevysila_700_tysjach_chelovek/ (дата обращения: 26.08.2022).
8. Use Machine Learning to Predict Horse Racing // towardsdatascience URL: https://towardsdatascience.com/use-machine-learning-to-predict-horse-racing-4f1111fb6ced (дата обращения: 28.08.2022).
9. Цифра дня: Сколько пользователей по всему миру зарегистрировались в World of Tanks? // ferra.ru URL: https://www.ferra.ru/news/games/cifra-dnya-skolko-polzovatelei-po-vsemu-miru-zaregistrirovalis-v-world-of-tanks-20-08-2020.htm (дата обращения: 28.08.2022).
10. Количество предварительных регистраций Blade & Soul 2 достигло 2 млн менее чем за день // mmo13 URL: https://mmo13.ru/news/post-15184 (дата обращения: 28.08.2022).

Свяжитесь с нами!

Telegram:

- @timofffey
- @nikitagordeev10
- @Skand21
- @Dima_Supchik