TMDB Box Office Prediction

Мирончук Б.С. (DA-6 Машинное обучение)

kaggle link

Постановка задачи

Задача: На основе предоставленных данных создать модель машинного обучения для предсказания общей выручки от показов фильмов в прокате по всему миру

Описание данных, проблемы и особенности

- **Тренировочный датасет содержит 22 поля и 3000 записей**
- **Среди исходных данных имелись:**
  - числовые (бюджет, популярность, продолжительность киноленты)
  - текстовые (Название фильма, описание фильма, ключевые слова)
  - временные (дата выхода)
  - данные в json формате (состав команды, актерский состав, языки перевода, страны и компании участвующие в проекте)
- **Проблемы исходных данных:**
  - json формат
  - много текстовых данных
  - числовые признаки с длинными хвостами и выбросами

Подходы в работе с данными

1. Удаление не несущих информацию признаков (различного рода id)
2. Создание бинарных признаков
3. Парсинг временных признаков
4. Создание новых количественных признаков
5. Логарифмирование признаков и таргета
6. Заполнение пропусков и выбросов медианными значениями (по причине распределений с длинными хвостами)
7. Для линейных моделей отсечение экстраполируемых "нереальных" хвостов (постпроцессинг)
8. Label encoding
9. Data scaling

Результаты

Модель	**Комментарий**	**Результат (ошибка на тесте)**
LinearRegression()	default по всем параметрам	3.41
Lasso()	default по всем параметрам	2.73
Lasso(alpha=0.5)	Уменьшена регуляризация чтобы задействовать больше признаков	2.63
Lasso(alpha=0.08)	Уменьшена регуляризация + из обучающего датасета удалены выбросы сверху (97.5 perc.) + заменены выбросы (97.5 perc.) в popularity на median()	2.33
RandomForestRegressor()	default по всем параметрам	2.17
CatBoostRegressor()	default по всем параметрам	2.10
CatBoostRegressor(random_state=42, max_depth=6, num_trees=1500)	Среднее предсказание обученных 3 CatBoostRegressor(max_depth=6, num_trees=1500)	2.018

Заключение

- В ходе работы были обучены несколько линейных моделей, случайный лес, градиентный бустинг (CatBoostRegressor)
- Наименьшей ошибки удалось достичь при усреднении предсказаний нескольких CatBoostRegressor, обученных на разном random_state. Размер ошибки - **2.018**