Houses. Predict sales prices

Kaggle competition

Team:

Roma_Ololo

Задача: предсказать цену домов (задача регресси)

Этапы:

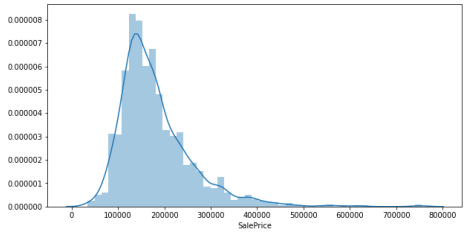
- EDA – анализ данных, работа с пропусками, удаление выбросов, кодирование и генерация новых признаков.
- MODEL – выбор базовых алгоритмов, настройка параметров
- STACKING and BLENDING – самблинг алгоритмов

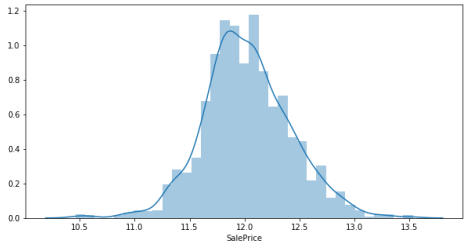
EDA

- Работа с пропусками:
  - Удалили признаки с пропусками и дисбалансом более 96% плюс незначащие по нашему мнению ('Utilities', 'Street', 'PoolQC', 'MiscFeature', 'Alley', 'Fence','GarageYrBlt’).
  - В категориальных признаках заменили пропуски модой.
  - В вещесвтенных признаках пропуски заменили медианой.
  - В части признаках пропуски заменили значениями в соответссвии с документацией.

EDA

- Нормализация данных:
  - Прологорифмирволи целевые значения – получили распределение близкое к нормальному





  - С помощью функции skew () обнаружили сильно скошенные признаки. Двойным преобразование box cox нормализовали часть.

EDA

- Future engineering:
  - Создали новые признаки, свидетельствующие о кач-ве объетка, например наличие камина - высший бал, отсутствие – низший.
- Удаление выбросов:
  - С помощью sm.OLS () проверили целевые метки и удалили выбросы на уровне значимости 0,01 (использована поправка Холма)
- Кодирование категориальных признаков:
  - Использовали get_dummies()

MODEL

Был использован grid search для поиска оптимальных параметров на 10 фолдах и скорингом r2.

Результаты базорвых алгоритмов ниже (скоринг на валидации - mean_squared_error , число фолдов - 5)

model	mean
ElasticNet	0.101
Lasso	0.101
Xgboost	0.109
LGBM	0.108

STACKING and BLENDING

- Были взяты базовые алгоритмы бустинга и регрессии:
  - ElasticNet, Xgboost, LGBM
- Металгоритм - Lasso
- Финальный результат был заблендeн по формуле:
  - 0.45 * stacking + 0.15 * Xgboost + 0.3 * LGBM + 0.1 * lasso