Прогноз размера будущего дохода банка от сотрудничества с клиентом

Райффайзен БАНК

Team: NewFolder2

ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ включала в себя:

- сортировку данных по дате и клиенту;
- обработка признаков с пропусками:
  - удаление признаков с пропусками более 50% значений;
  - признаки с небольшим количеством пропусков заполняем в категориальных признаках Unknown, в количественном – медианой
- обработка категориальных признаков:
  - Метод ONE HOT ENCODING – для категориальных признаков с небольшим числом вариаций признака
  - Метод MEAN TARGET – замена категориальных признаков групповыми средними по признаку gi_smooth_3m
- подготовка целевой переменной – CLTV как сумму gi_smooth_3m по месяцам с 7 по 12 в тренировочных данных
- ввиду избыточного размера выборки и ограниченного времени на выработку решения было принято решение о сокращении датасета в 4 раза путем отбора 25% клиентов в train-набор данных на вход модели

В итоговом виде датасет представил собой выборку 25% клиентов с 6 записями (за первые 6 месяцев 2018) для каждого из клиентов и указанием целевой переменной CLTV и наборов признаков.

ССЫЛКА НА ПОДГОТОВЛЕННЫЙ ДАТАСЕТ

https://drive.google.com/open?id=1-8bF7x_dN0kBm_oaz73_mmK_6vPJ73-Y

ВТОРИЧНАЯ ОБРАБОТКА ДАННЫХ включала в себя отбор признаков:

- использовали оценку важности признаков 2мя методами: f_regression и Ridge
- по усредненному показателю важности каждого из признаков были отобраны наиболее значимые:

Признаки	F test	Ridge	average_all_models
gi_smooth_3m	1	0.02	0.51
cur_quantity_mort	0.02	1	0.51
cur_quantity_pl	0.08	0.53	0.31
cur_quantity_cc	0.04	0.2	0.12
cc_cash_spend_c	0.01	0.19	0.1
cu_education_level_1	0.01	0.19	0.1
active	0.01	0.18	0.1
big_city_SPB	0	0.12	0.06
cur_quantity_mf	0	0.12	0.06
big_city_MLN	0	0.11	0.06
standalone_nonpayroll_dc_f	0.01	0.11	0.06
cu_education_level_Unknown	0.01	0.1	0.05
standalone_dc_f	0.01	0.09	0.05
cu_education_level_3	0	0.09	0.05
cur_quantity_dc	0.01	0.08	0.05
big_city_OTH	0	0.08	0.04
cu_education_level_2	0	0.08	0.04
big_city_MSK	0	0.08	0.04
ca_f	0	0.07	0.03
cur_quantity_deposits	0	0.05	0.03
ПРОЧИЕ ПРИЗНАКИ	-	-	Менее 0.03

ИСПОЛЬЗУЕМЫЕ МЕТРИКИ КАЧЕСТВА:

- **MAPE -** np.mean(np.abs((y_true - y_pred) / y_true)) * 100
- **MAE -** стандартная реализация Scikit-learn 0.22.2

ОТБОР МОДЕЛЕЙ ПРОИЗВОДИЛСЯ ИЗ:

- sklearn.linear_model.**LinearRegression**
- sklearn.linear_model.**Ridge**
- sklearn.neighbors.**KNeighborsRegressor**
- sklearn.ensemble.**RandomForestRegressor**(n_estimators=500, max_depth=7)
- sklearn.tree.**DecisionTreeRegressor**(max_depth=7)
- sklearn.linear_model.**SGDRegressor**

РЕЗУЛЬТАТЫ ОТБОРА МОДЕЛЕЙ ПРИ РАЗБИЕНИИ ДАННЫХ 70% TRAIN/ 30% TEST:

Модель	MAPE	MAE
LinearRegression	120.46	35364.61
Ridge	121.75	35364.62
KNeighborsRegressor	nan	36540.90
RandomForestRegressor	67.69	31528.46
DecisionTreeRegressor	68.86	31905.78
SGDRegressor	100.00	8.95e+18

В РЕЗУЛЬТАТЕ ОТБОРА БЫЛА ВЫБРАНА МОДЕЛЬ **RANDOMFORESTREGRESSOR**.

Для тестовых данных была произведена подготовка аналогично этапам предобработке тренировочных данных:

- Удаление/замена признаков с пропусками значений
- Обработка категориальных признаков методами ONE HOT ENCODING и MEAN TARGET

ГОТОВОЕ РЕШЕНИЕ ПО ССЫЛКЕ

https://drive.google.com/open?id=1Q3OcuKoJGMuEGb0d3KKscH-oQPlDlI5Z

ВОЗМОЖНОСТИ МОДЕЛИ:

- возможную модель можно использовать для предсказаний показателя CLTV в условиях доступности данных о 6 последних месяцев показателей клиента;
- подготовленный датасет

ССЫЛКА НА ПОДГОТОВЛЕННЫЙ ДАТАСЕТ

https://drive.google.com/open?id=1-8bF7x_dN0kBm_oaz73_mmK_6vPJ73-Y

можно использовать в качестве источника обучения более сложных моделей, например CatBoost, XGBoost, а также нейронные сети. К сожалению, в силу ограниченного времени были проверены только классические решения.

Команда **NewFolder2**

Денис Дубовицкий
e-mail: dubovitskyden@gmail.com
tel: +7-923-705-19-46

Владимир Викулов
e-mail: vikulov-vl@yandex.ru
tel: +7 (950) 733-42-31

Дарья Пирожкова
e-mail: pirozhkova-dasha@mail.ru
tel: +7-913-018-17-78

Кирилл Чертоганов
e-mail: chertoganov.kirill@gmail.com
tel: +7 (928) 041 51 40