SAS competitions
Предсказание вероятности невозврата кредита

Тимошилов Иван
(2 место)

Ход решения

1. Работа с источниками
2. Фича-инжиниринг
3. Построение и валидация моделей
4. Ансамбль

Источники

- Уникальный ключ кредита = id + дата + сумма кредита
- Удаляем дубликаты (оставляем новейший)
- Из всех источников выбираем новейший

Новые признаки

- Текстовая строка
- Разницы дат
- ОНЕ для категорий
- Курсы валют
- Запросы и просрочки – суммируем

Агрегирование

- Сортировка по дате
- Groupby по ID
- 'min', 'mean', 'median', 'max', 'std', 'sum', 'count', 'first', 'last', 'mean'/ 'std'
- ИТОГО: 1500+ признаков

Отбор признаков

1. XGBoost на всех признаках
2. Отбор 2х лучших, база
3. Добавление по одному
4. Отбор по CV score

ИТОГО: 700+ признаков

Модели

	CV Score	Public (solo)	Public (30-70 seeds)	Private
XGBoost	0,7050	0,7138	0,7173	?
LightGBM	0,7070	0,7150	0,7190	?
XGB+LGB	?	0,7194		?
Stack	0,7115	0,7175		?
Average	?	0,7190		0,7121