Отчет по контесту «Прогнозирование вероятности невозврата кредита»

Таскынов Ануар, 517 группа

Курс «Прикладные задачи анализа данных»

21 января 2018 г.

Даны:
Кредитные истории клиентов с четырёх разных источников. Данные содержат сумму кредита, дату кредита, историю платежей и т.п.

Задача:
Необходимо по этим данным предсказать вероятность невозврата кредита клиентом.

Функционал качества:
AUC.

Обработка данных:

- Сначала были удалены некорретные значение даты выдачи кредита.
- Пропущенные значения в сумме кредита были заменены на медиану.

Признаки:

- Различные разности между датами. В некоторых случаях, дата выдачи кредита была позже, чем дата возвращения. В таких случаях даты менялись местами.
- Добавим выплату по месяцам.
- Добавляем текстовые признаки: длина строки, первая/последняя попавшаяся выплата, первая/последняя просрочка выплаты, количество комбинаций символов; взвешенная сумма выплат.
- Счётчики по credit_active, credit_type и num_source.
- Добавим признак: (взвешенная сумма выплат) x (месячная выплата)

Агрегация:
Агрегация проводилась по источникам: брались всевозможные статистики.
Модель:
Десять LGBM, усредненных по сидам, обученных на агрегированных данных.
Итог: **0.713** на Public и **0.7057** на Private (14 место).