Задание

Вам предлагается набор данных, содержащий следующие переменные

Переменные, характеризующие кредитную историю клиента

cred_sum_cc_all - сумма кредитов по кредитным картам
mfo_inqs_count_month - количество запросов на кредиты в другие в МФО
all_closed_creds_sum_all количество закрытых кредитов
bank_inqs_count_quarter - количество запросов на кредиты в банки
cred_max_overdue_max_ly - максимальная просрочка за год
all_active_creds_sum_all - денежная сумма всех активных кредитов
mfo_last_days_all - Количество дней с последнего займа в МФО
cred_sum_cc_ly - Сумма лимитов кредитных карт, оформленных за последний год
cred_sum_debt_all_all - Сумма задолженности по всем кредитам
all_closed_creds_sum_ly - Сумма закрытых кредитов за последний год
mfo_cred_mean_sum_3lm - Средняя сумма МФО кредитов, выданных за последние 3 месяца
delay_more_sum_all - Количество просрочек более чем на 90 дней по всем кредитам
all_creds_count_all - Общее количество кредитов
cred_day_overdue_all_sum_all - Суммарное количество дней просрочки текущих активных кредитов
cred_max_overdue_max_3lm - Максимальная сумма просроченной задолженности, по кредитам взятым за последние 3 месяца
mfo_closed_count_ly - Количество закрытых МФО кредитов, взятых за последний год
cred_sum_overdue_cc_all - Сумма просрочек по кредитным картам
count_overdue_all_3lm - Количество кредитов на просрочке, взятых за последние 3 месяца
all_creds_count_lm - Количество кредитов, взятых за последний месяц
region - регион подачи заявки

Переменные характеризующие клиента
work_code - Профессия. 5 - рабочие профессии (слесарь, токарь). 3 - офисный работник (бухгалтер, программист). 1 - госслужащий (полицейский, медсестра)
month_income - доход

Целевые переменные:
bad - 1 - кредит просрочен, 0 - кредит возвращен, nan - отказ.
approved - 1 - одобрено, 0 - отказано.

Задача:
Построить две модели классификации на исходных признаках.
Модель 1 - обученная только на выданных заявках, целевая переменная bad.
Модель 2 - обученная на всех заявках, целевая переменная approved.

Для тренировки и валидации использовать файл train.csv

Сравнить распределения признаков на двух выборках. Первая - только выданные заявки (bad != nan). Вторая - все заявки. Выделить признаки, распределения которых сильно отличаются. Прокомментировать причину различий в распределении.

Сравнить модель 1 и модель 2
Методика сравнения:
1.1 Строим модель 1 и модель 2.
1.2 Считаем моделью 1 вероятность принадлежности к классу bad=1 на всем тестовом наборе test.csv
1.3 Выбираем выданные заявки (bad != nan), группируем заявки в 7 групп по скору. Считаем badrate в каждом интервале. Badrate - количество плохих (bad=1) кредитов в бине, поделенное на количество всех кредитов в бине. Строим график stackedbar. Границы интервалов необходимо подобрать так, чтобы badrate в соседних интервалах отличался.
2.
2.1. Выбираем невыданные заявки (bad=nan). Разбиваем на интервалы по скору, полученные в п.1.3
2.2. В каждом бине размечаем случайную часть заявок как “плохие” (bad=1). Количество плохих в бине должно соответствовать badrate, рассчитанному в п.1 для данного бина.
3. Смешиваем заявки, размеченные в п.2.2 с выданными заявками (bad != nan).
4. Сортируем полученную выборку по возрастанию вероятности принадлежности к классу bad=1, рассчитанной в п. 1.2. Выбираем первые 30% заявок и считаем по ним badrate.
5. Считаем моделью 2 вероятность принадлежности к классу approved=0 на всем тестовом наборе test.csv. Сортируем по возрастанию вероятности.
6. Выбираем первые 30% заявок и считаем по ним badrate.
7. Объяснить почему одна модель получилась лучше другой по целевой метрике. Целевая метрика - процент “плохих” среди одобренных заявок.
8. Придумать как можно улучшить целевую метрику - процент плохих в 30% одобренных заявках.

Подсказки:

- Комбинировать модель 1 и модель 2;
- Сэмплирование выборки;
- Применить алгоритм разметки отклоненных заявок.

При выполнении задания рекомендуем сделать упор на анализ данных, интерпретацию решений и ошибок моделей. Способ улучшения целевой метрики. Меньше усилий тратить на подбор гиперпараметров и выбор алгоритма.

На выполнение задания отводится 1 неделя с момента получения письма
Отчет прислать в файле .pdf

Данные
https://drive.google.com/drive/folders/1CPUMYtFsw5FC4lONB7K5NdTnLoCmtx7u?usp=sharing