**README**

**→ основные цели и задачи проекта:**

**Цель:** построить ml модель для предсказания дефолта.

**Задача проекта:** улучшить предсказательную силу модели банковского скоринга для принятия решения о том, выдавать ли кредит заемщику или нет (понять, возможен ли дефолт по кредиту). Это задача бинарной классификации, поэтому анализа вероятности дефолта является задачей логистической регрессии.

**→ краткая информация о данных:**

Данные представляют собой датасет из 73799 записей о кредитной истории клиентов. Всего 18 признаков в датасете, из них 6 признаков в виде строк, а 12 числовые, целевая переменная 'default'.

**Описание датасета:**

1. client\_id - идентификатор клиента

2. app\_date - дата подачи заявки

3. education - уровень образования

4. sex - пол заемщика

5. age - возраст заемщика

6. car - флаг наличия автомобиля

7. car\_type - флаг автомобиля иномарки

8. decline\_app\_cnt - количество отказанных прошлых заявок

9. good\_work - флаг наличия “хорошей” работы

10. score\_bki - скоринговый балл по данным из БКИ

11. bki\_request\_cnt - количество запросов в БКИ

12. region\_rating - рейтинг региона

13. home\_address - категоризатор домашнего адреса

14. work\_address - категоризатор рабочего адреса

15. income - доход заемщика

16. sna - связь заемщика с клиентами банка

17. first\_time - давность наличия информации о заемщике

18. foreign\_passport - наличие загранпаспорта

→ **этапы работы над проектом:**

* EDA
* очистка и обогащение данных
* оценка важность признаков для модели
* обучение модели логистической регрессии и прогноз вероятности дефолта
* методы машинного обучения для улучшения прогнозирования вероятности дефолта
* настройка гиперпараметров модели
* участие в соревновании на Kaggle

→ **ответы на вопросы саморефлексии:**

**Какой частью своей работы вы остались особенно довольны?**

Получилось обработать числовые, бинарные и категориальные признаки и подготовить их для модели. Определила значимость признаков. Довольна логарифмированием распределения числовых признаков.

**Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать?**

Чем тщательнее я обрабатывала данные, например, удаляла выбросы, тем хуже работала модель. Почему-то новые признаки также не привели к улучшению показаний метрик (некоторые добавленные мной признаки удалила).

Запуталась с гиперпараметрами, взяла предложенное решение в slack и с платформы, но видела более продвинутое решение и настройки гиперпараметров.

Не получилось сильно улучшить модель. Повторяла одни и те же шаги при построении модели, запуталась с порядком действий.

В процессе  построения модели постоянно зависал компьютер, не поняла, почему это происходило.

**Что интересного и полезного вы узнали в этом модуле?**

Submit на Kaggle, принцип логистической регрессии, матрицу ошибок.

**Что является вашим главным результатом при прохождении этого проекта?**

Попробовала нормализовать выборку данных.

**Какие навыки вы уже можете применить в текущей деятельности?**

EDA, dummies, значимость непрерывных переменных, ROC-анализ.

**Планируете ли вы дополнительно изучать материалы по теме проекта?**

Хочу попробовать поучаствовать в других соревнованиях. Продолжаю изучать https://scikit-learn.org/