Кредитный скоринг.

Линейные модели *vs* современные модели data science

Руководитель
Воробьева Мария Сергеевна

Задача кредитного скоринга

Идея

Оценка кредитоспособности клиента для принятия решения о выдаче ему кредита.

Как?

На основе данных из анкеты и доступной информации о предыдущих кредитах строится модель, предсказывающая вероятность того, что кредит будет выплачен.

Зачем?

Минимизация кредитных рисков банка.

Аналоги

Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.

Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L. C.
2015

- Сравнивалось 41 классификаторов по 6 разным параметрам
- Ансамблевые методы показали себя значительно лучше стандартного подхода
- Утверждается, что случайный лес должен быть принят как базовый алгоритм вместо логистической регрессии

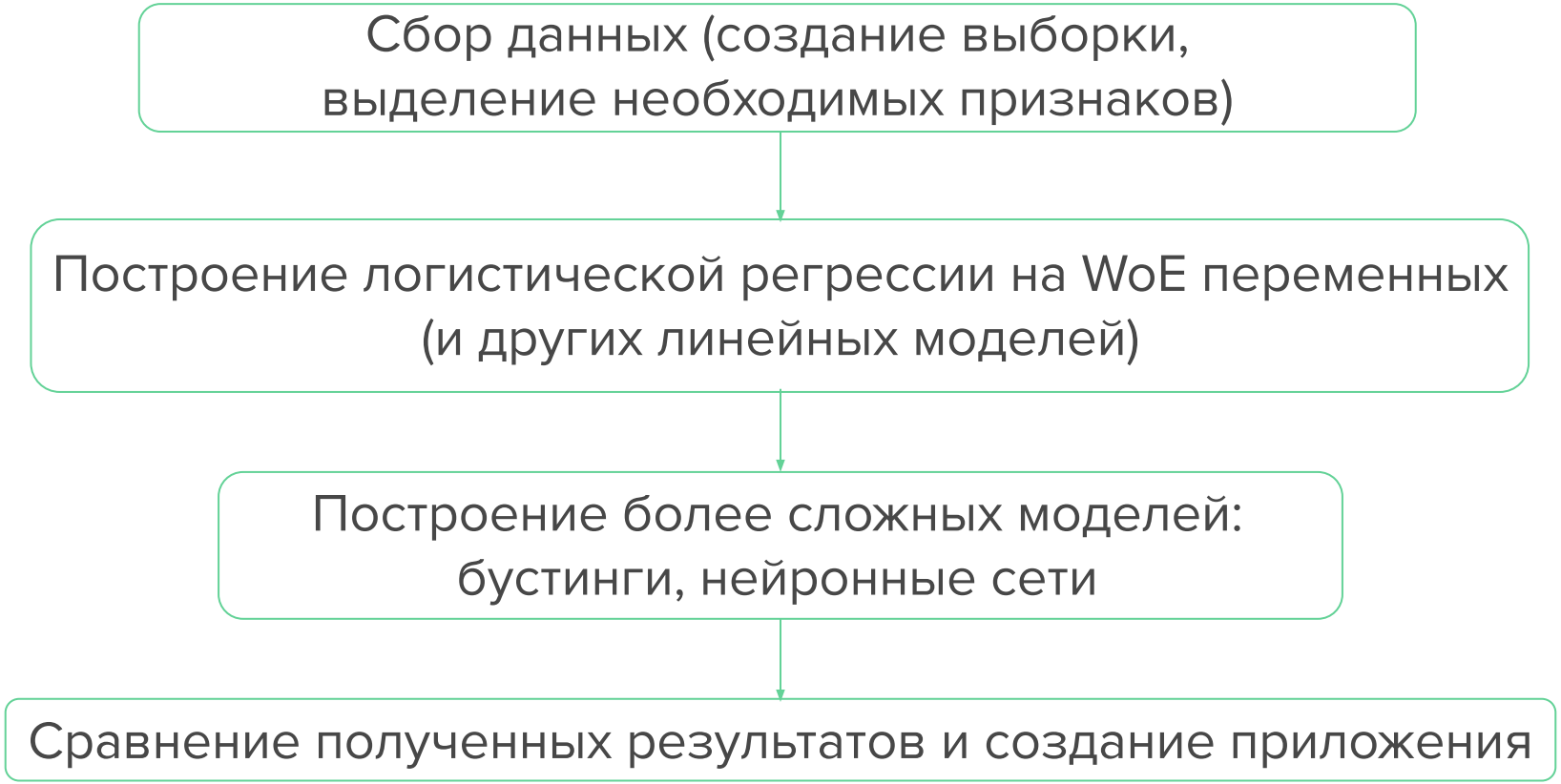
Аналоги

Deep Learning for Credit Scoring: Do or Don’t?

Bjorn Rafn Gunnarsson, Seppe vanden Broucke, Bart Baesens, Maria Oskarsdottir, Wilfried Lemahieu
2019 (accepted: 2021)

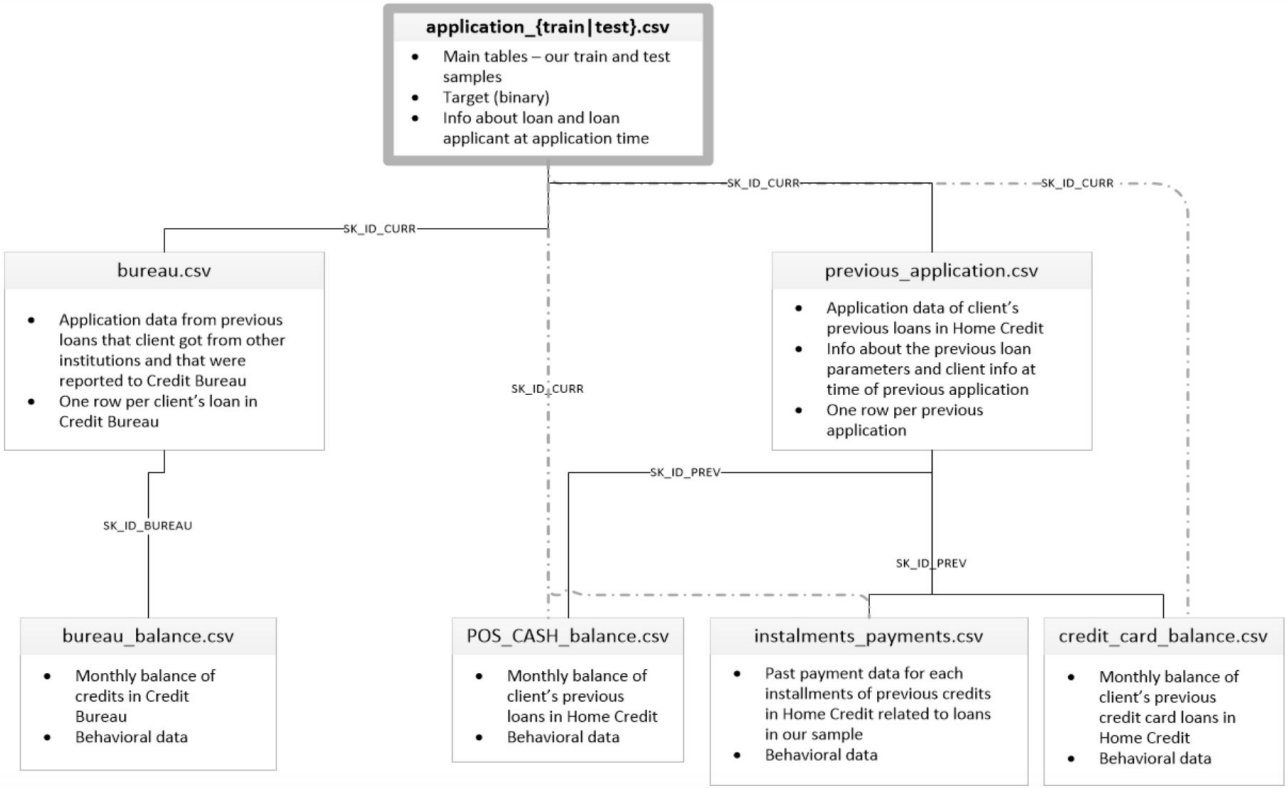
- Нейронные сети сравнивались с XGBoost, случайным лесом, деревом решений и логистической регрессией
- Оценивание проводилось по 4 показателям (в том числе AUC-ROC) и по каждому из них выставлялись “баллы” по 10-балльной шкале
- XGBoost показывает результаты примерно в 2 раза лучше, чем лог. регрессия и в 3 раза лучше, чем DBN
- Увеличение количества слоев в сети приводило к ухудшению результата

План работы


Сбор данных (создание выборки, выделение необходимых признаков)
Построение логистической регрессии на WoE переменных (и других линейных моделей)
Построение более сложных моделей: бустинги, нейронные сети
Сравнение полученных результатов и создание приложения

Исходные данные

Используются данные компании Home Credit из соревнования на Kaggle



Добавление признаков

- Применялись различные виды агрегации данных из исходных таблиц (на данный момент задействованы *bureau, bureau_balance* и *previous_application*)
- Агрегация проводилась по временным срезам (последний год, три года и за все время) и по типам кредитов (потребительские кредиты, кредитные карты и т. д.)
- В итоге получено 312 новых признаков + 121 изначальный признак из таблицы *application_train*
- Создан документ с подробными описаниями всех добавленных признаков

Weight of Evidence и Information Value

Weight of Evidence - это некая предсказательная сила переменной (разная для разных значений переменной)

$$WoE_i(q) = ln(\frac{Distr\ Good_i(q)}{Distr\ Bad_i(q)})$$

Information Value - мера предсказательной силы всей переменной

$$IV_i = \sum_q (Distr\ Good_i(q) - Distr\ Bad_i(q)) * WoE_i(q)$$

Интерпретация значений Information Value

IV	Предсказательная сила
< 0.02	Не подходит для предсказания
0.02 to 0.1	Слабая
0.1 to 0.3	Средняя
0.3 to 0.5	Сильная
> 0.5	Очень сильная

Отбор признаков и построение модели

	All features	IV >= 0.02	|corr| < 0.8	interm. AUC-ROC	Forward selection	AUC-ROC
Cash loans	433	134	78	0.742	17	0.730
Revolving loans	433	173	110	0.729	16	0.740
final AUC-ROC for whole test sample = 0.733						

Скоринговая карта

- Таблица баллов, на основании которых принимается финальное решение по кредиту

$$балл = -\left(WOE_j \cdot b_i + \frac{b_0}{n}\right) \cdot R + \frac{A}{n},$$

$$R = \frac{D}{\ln(2)}, \qquad A = B - R \cdot \ln(C),$$

D - количество баллов для удвоения шансов получить кредит (принимаем равным 40)

B - значение на шкале баллов, в которой соотношение шансов составляет C:1 (в точке 600 баллов соотношение составляет 72:1)

Скоринговая карта

Cash Loans
EXT_SOURCE_3
EXT_SOURCE_2
EXT_SOURCE_1
AMT_GOODS_PRICE
ORGANIZATION_TYPE
b_active_Consumer credit_dur_max
CODE_GENDER
REGION_POPULATION_RELATIVE
NAME_EDUCATION_TYPE
b_Microloan
DAYS_ID_PUBLISH
p_yield_high
AMT_ANNUITY
OWN_CAR_AGE
p_cnt_avg
NAME_FAMILY_STATUS
YEARS_BUILD_MODE

Revolving loans
EXT_SOURCE_2
EXT_SOURCE_3
EXT_SOURCE_1
OCCUPATION_TYPE
b_start_Credit card_avg
NAME_EDUCATION_TYPE
REGION_POPULATION_RELATIVE
b_active_all_dur_max
p_Consumer loans_sum_app_avg
COMMONAREA_MEDI
p_all_sum_app_avg
NAME_FAMILY_STATUS
p_cnt_avg
p_all_high_percent
p_prod_group_POS household_percent
FLAG_DOCUMENT_3

Скоринговая карта

Feature	IV	Value	WoE	Score
EXT_SOURCE_3	0.326824	(-∞; 0.284]	-0.899550	68.990114
		(0.284; 0.444], NaN	-0.212029	41.214654
		(0.444; 0.618]	0.360558	18.081236
		(0.618; ∞]	0.847568	-1.596127
NAME_FAMILY_STATUS	0.035	Civil marriage	-0.282	37.714
		Single / not married	-0.229	36.694
		Unknown	0.0	32.298
		Married	0.103	30.318
		Separated	0.119	30.014
		Widow	0.498	22.752

Градиентный бустинг

Были рассмотрены такие виды бустингов как:

- **LightGBM**
- **XGBoost**
- **CatBoost**

Отбор признаков

- Так как нам необходимо создать калькулятор в который можно будет вбивать данные по клиенту, а вбивать 434 признака слишком неудобно было принято решение снизить количество признаков до 30
- Для это сначала был использован параметр feature_importance из построенной модели с помощью которого количество признаков уменьшилось до 117
- После этого были откинуты все сильно коррелирующие признаки и признаков осталось 75
- И в конце был использован метод rfe(recursive feature elimination) оставивший 30 признаков

Подбор гиперпараметров и построение модели

- Для подбора гиперпараметров использовался фреймворк Optuna
- Полученные значения AUC-ROC:

	LightGBM	CatBoost	XGBoost
До отбора признаков	0.762	0.7689	0.772
После отбора признаков	0.7481	0.7475	0.753

Нейронные сети

Рассматривались 2 различных типа моделей классификации, от простых нейронных сетей до более сложных нейросетевых архитектур, а именно MLP и **TabNet**.

MLP

Элементы сети:

- Количество слоев и нейронов на каждом слое
- Механизм обучения или оптимизатор
- Функция активации
- Регуляризации

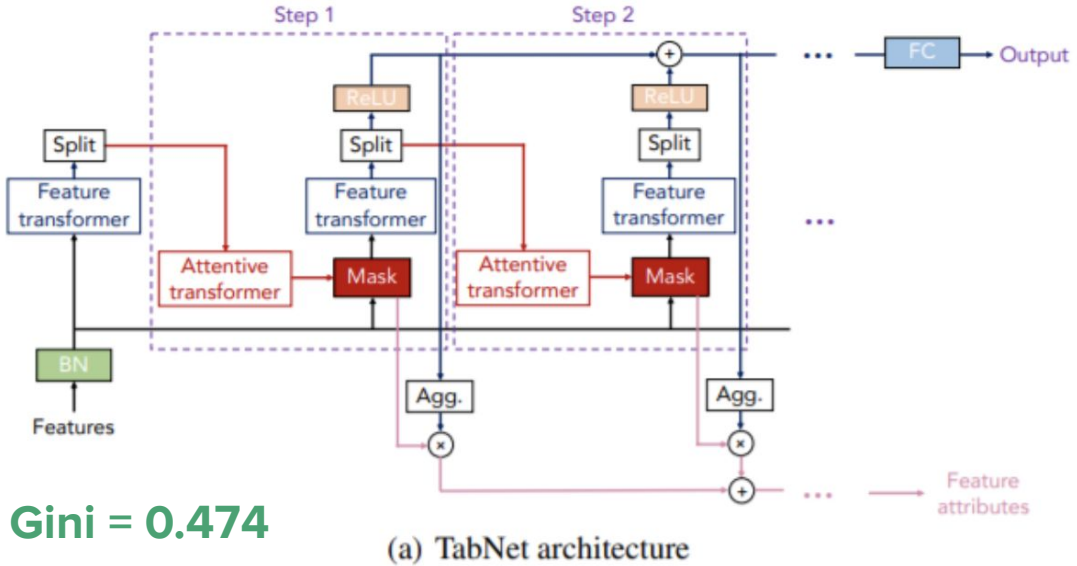
batch_size, n_epochs

Keras Tuner для настройки архитектуры сети и гиперпараметров.

AUC-ROC = 0.745, Gini = 0.49

TabNet

TabNet – архитектура глубокого обучения на основе табличных данных. Нейросеть состоит из полносвязных слоев с последовательным механизмом внимания.



(a) TabNet architecture

AUC-ROC = 0.737, Gini = 0.474

Финальные значения AUC-ROC

	LogisticRegression + WoE	XGBoost	MLP
AUC-ROC	0.733	0.753	0.745

Сравнение экономического эффекта

Переплаты клиентов по кредитам с условием дифференцированных платежей:

- Для потребительских кредитов:

$$((S * 1.1 - \frac{2}{3}S) + (\frac{2}{3}S * 1.1 - \frac{1}{3}S) + \frac{1}{3}S * 1.1) - S = \boxed{0.2S}$$

- Для кредитных карт:

$$((0.7S * 1.2 - \frac{2}{3} * 0.7S) + (\frac{2}{3} * 0.7S * 1.2 - \frac{1}{3} * 0.7S) + \frac{1}{3} * 0.7S * 1.2) - 0.7S = \boxed{0.28S}$$

Сравнение экономического эффекта

Суммарная прибыль = 0.2(сумма выданных потребительских кредитов без дефолта) +

+ 0.28(сумма выданных возобновляемых кредитов без дефолта) -

- (сумма выданных потребительских кредитов с дефолтом) -

- 0.7(сумма выданных возобновляемых кредитов с дефолтом)

	LogisticRegression + WoE	XGBoost
Прибыль	3 157 483 895	3 351 374 100

Итоговые результаты для XGBoost

- AUC-ROC = 0.753
- Прибыль = 3.35 млрд
- Упущенная прибыль (сумма, которая могла быть заработана на клиентах без дефолта, которым не был выдан кредит) = 140 млн

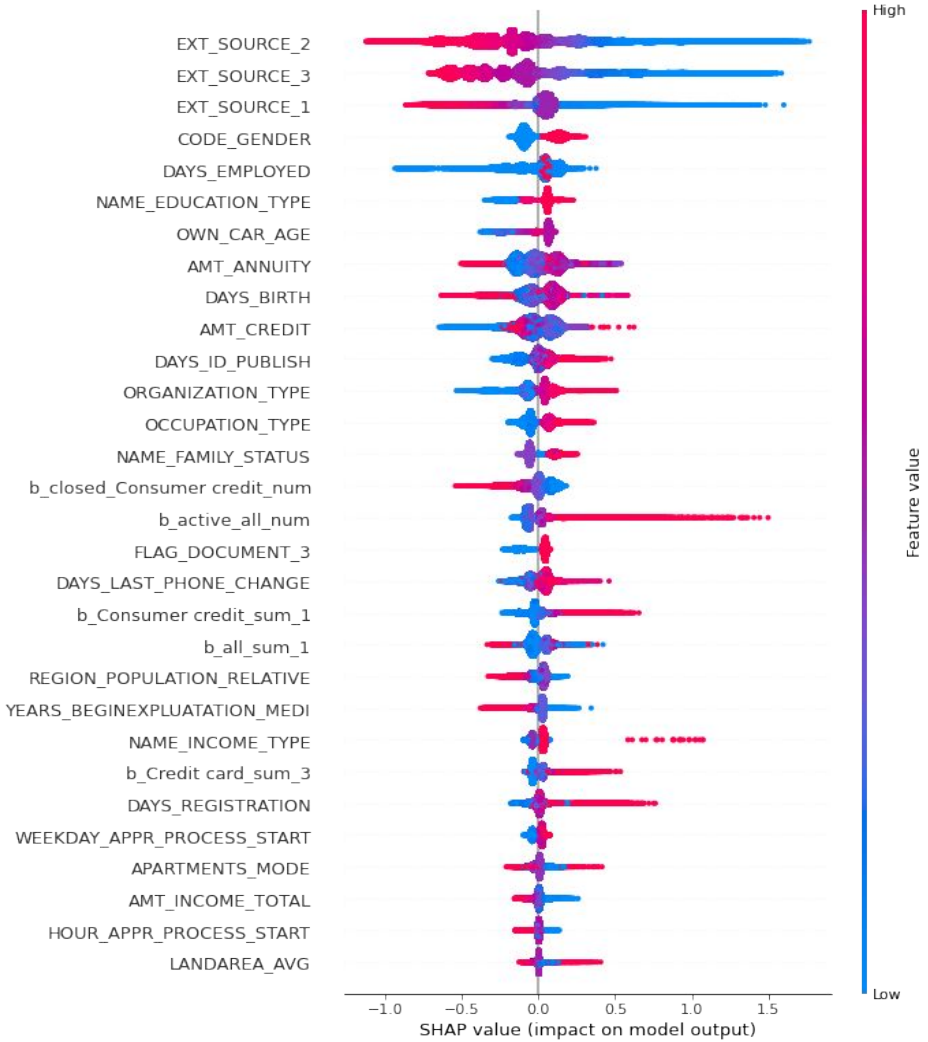
Метод интерпретации SHAP

Метод показывает влияние каждого признака на прогноз следующим образом:

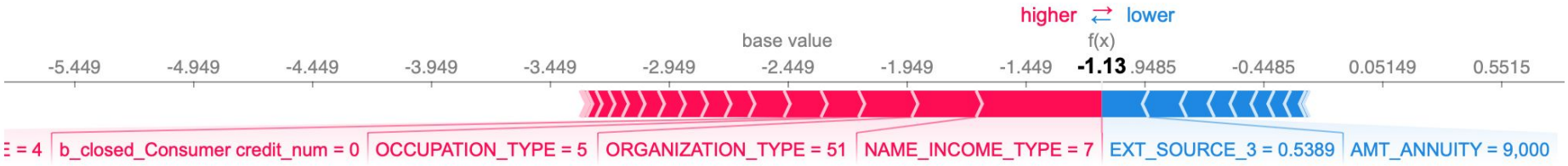
- Строится предсказание модели с заданным признаком i и без него
- Вычисляется значение Шепли для каждого возможного набора признаков без i, а затем суммируются для получения важности данной независимой переменной
- Значение Шепли вычисляется по формуле $\varphi_i(p) = \sum_{S \subseteq N / \{i\}} \frac{|S|!(n-|S|-1)!}{n!}(p(S \cup \{i\}) - p(S))$

  где p(S U {i}) - предсказание модели с i-м признаком, p(S) - предсказание без i-го признака, n - количество признаков, S - произвольный набор признаков без i-го

Глобальная интерпретация модели XGBoost



Локальная интерпретация модели XGBoost

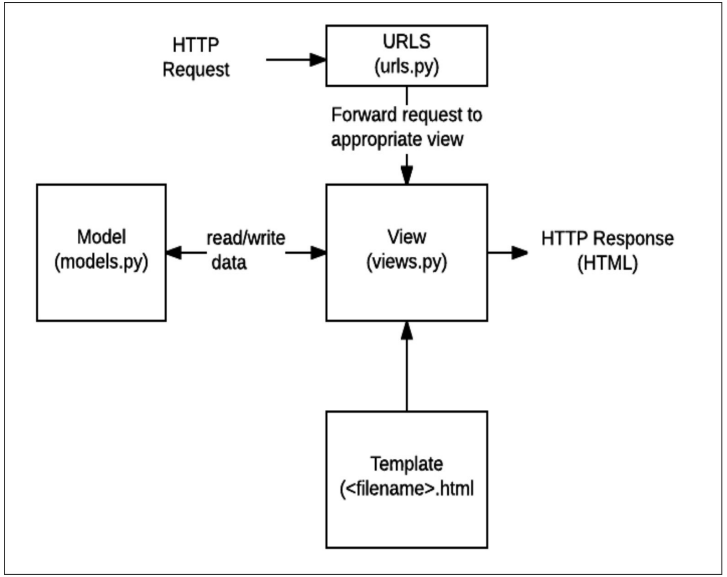


Пользовательский интерфейс

Django - серверный веб-фреймворк, написанный на Python.

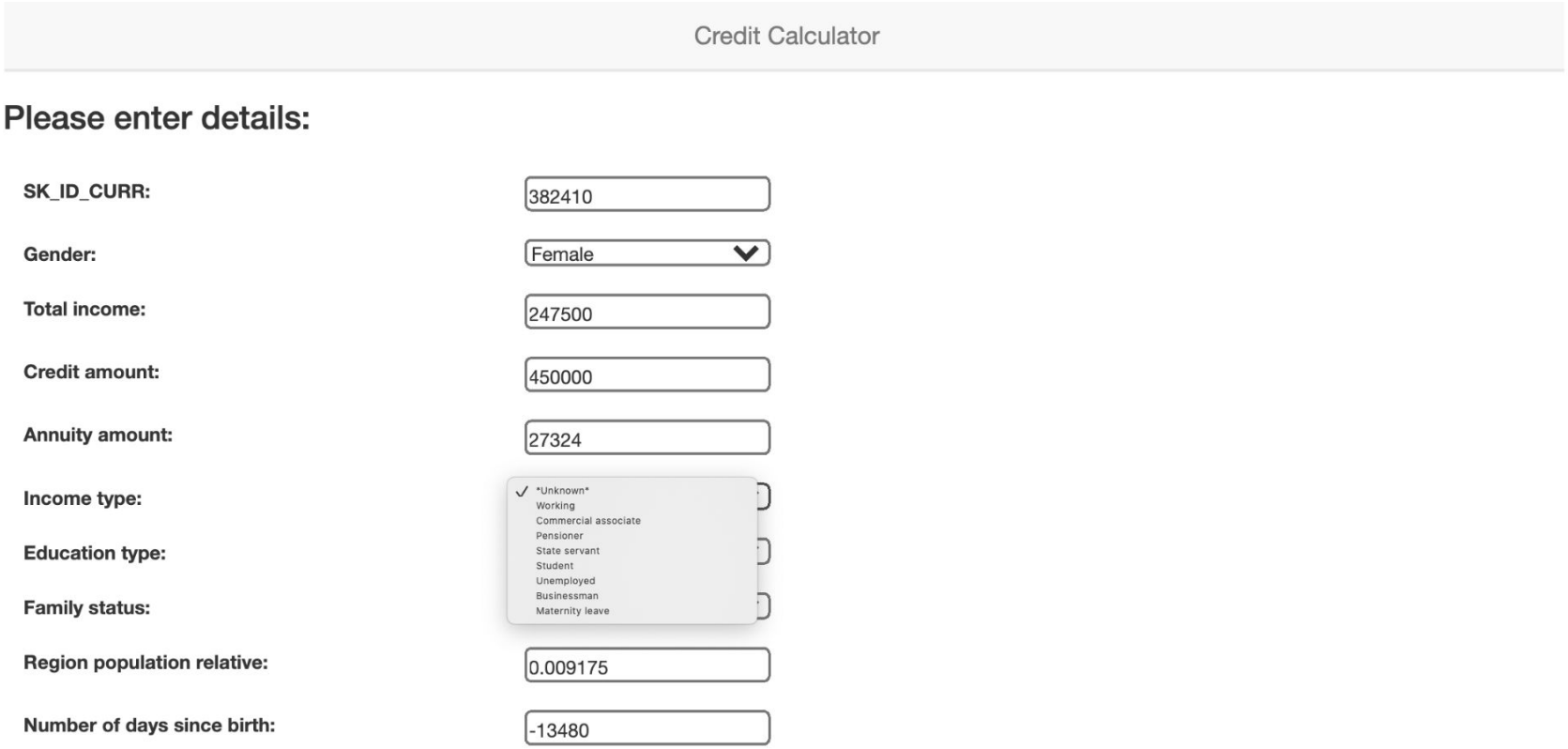
Frontend:
HTML - для создания страницы калькулятора и формы, в которую пользователь заносит детали заявки на кредит
CSS - для стилей

Backend:
Тренировочные данные и обученная на этих данных модель и encoders в joblib файлах
Чтение признаков, введенных пользователем, и возвращение решения о выдаче кредита, на основе выбранной модели



MVC

Пользовательский интерфейс



Полезные ссылки

- Исходные данные
  https://www.kaggle.com/c/home-credit-default-risk/data
- Описание признаков
  https://docs.google.com/spreadsheets/d/1LHIld12GG-WS2b7whc1NCa-oefVvZie0q-3Y_wbK5_M/edit#gid=210917708
- Github с ноутбуками и приложением

  https://github.com/sweetd0ve/credit-scoring-web-service