Применение методов машинного обучения в задачах скоринга клиентов банка

Хасанова Кристина

гр. 22.М03-мм
Санкт-Петербургский государственный университет
Кафедра статистического моделирования

5 июня 2024 г.

Используемые определения

- *Скоринг клиентов* — определение вероятности возврата кредита заемщиком (значение от 0 до 1),
- *Дефолт* — кредит считается в дефолте, если выплаты по нему просрочены более чем на 90 дней,
- *Бенчмарк* — градация алгоритмов после тестирования на множестве реальных данных,
- *Гиперпараметры* — заданные параметры модели, не меняющиеся в процессе обучения модели.

Постановка задачи, цели работы

Цель дипломной работы - построение широко используемых и недавно опубликованных алгоритмов бинарной классификации и проведение их оптимизации на реальных данных из домена кредитного скоринга. Задачи дипломной работы:

- Изучить бенчмарки по моделям кредитного скоринга и бинарной классификации;
- Изучить методы отбора признаков в итоговую модель, методы подбора гиперпараметров;
- Поиск реальных данных банка;
- Применение моделей и методов их оптимизации на найденных данных;
- Подсчет финансового эффекта от применения лучшей модели.

Модели. Логистическая регрессия

$$P = \frac{1}{1 + e^{-z}}$$

- P — вектор столбец вероятностей не возвращения кредита заемщиками,
- z — произведение матрицы признаков и вектора коэффциентов

Модель: $[P(B)]^{Y}[1 - P(B)]^{1-Y}$ (распределение Бернулли)
Подбор коэффициентов: минимизация логирифма функции правдоподобия:

$$L^* = \sum_{i=1}^{k} [Y_i ln P_i(B) + (1 - Y_i) \ln(1 - P_i(B))].$$

- B — вектор коэффициентов регрессии,
- $P_i(B)$ — оценка вероятности дефолта заемщика,
- Y — метки класса.

Модели. Градиентный бустинг

Алгоритм будет состоять из ансамбля:

$$F_m = \sum_{m=1}^{M} w_m a(x, b_m), w_m \in \mathbb{R}, b_m \in B.$$

Подбор w_m осуществляется линейным поиском:

$$w_m = argmin \sum_{i=1}^{N} L(F_{m-1}(x_i) - w \bigtriangledown Q_i).$$

- Q_i — ошибки до m алгоритма, например, $y_i - F_{m-1}(x_i)$,
- L — функционал ошибки

Найдем a_m минимизируя функционал ошибки:

$$a_m = argmin \sum_{i=1}^{N} L(F_{m-1}(x_i) - wa(x_i, b_m)).$$

В случае бинарной классификации w_i - веса, определяющие важность объектов в выборке.

Модели. AutoInt слой эмбеддингов

Категориальные признаки:

$$e_i = \frac{1}{q} V_i x_i.$$

- e_i — эмбеддинг категориального признака,
- q — количество значений, которое принимает категориальная переменная,
- V_i — матрица эмбеддингов,
- x_i — one-hot вектор.

Для взаимодействия категориальных и числовых признаков, последние представлем тоже в низкоразмерном пространстве:

$$e_m = V_m x_m.$$

- e_i — эмбеддинг числового признака,
- V_i — матрица эмбеддингов,
- x_i — вектор признака.

Модели. AutoInt, слой взаимодействия [1/2]

Найдем корреляцию эмбеддингов e_m и e_k:

$$\alpha^h_{m,k} = \frac{exp(\psi^h(e_m, e_k))}{\sum_{l=1}^{M} exp(\psi^h(e_m, e_l))}.$$

здесь $\psi^h(a, b)$ — функция внимания, определяющая похожесть эмбеддингов, задается следующим образом:

$$\psi^h(e_m, e_k) =< W^h_{\text{QUERY}} e_m, W^h_{\text{KEY}} e_k > .$$

- W^h_{QUERY} — матрица запросов из механизма внимания
- W^h_{KEY} — матрица ключей из механизма внимания

Модели. AutoInt, слой взаимодействия [2/2]

Далее находим обновленный признак e_m для головы h:

$$\widetilde{e}_m^h = \sum_{k=1}^{M} \alpha_{m,k}^h (W_{\text{VALUE}}^h, e_k).$$

Для получения признака, учитывая все головы, находим:

$$\widetilde{e}_m = \widetilde{e}_m^1 \oplus \widetilde{e}_m^2 \oplus ... \widetilde{e}_m^H.$$

- H — количество голов,
- $\oplus$ — оператор конкатенации,
- $\widetilde{e}_m^i$ обновленный признак по i голове.

Добавляем признаки произведения первого порядка:

$$\widetilde{e}_m^{\text{Dim}} = ReLU(W_{\text{Dim}} e_m + \widetilde{e}_m).$$

здесь $ReLU$ – функция активации , W_{Dim} – матрица для получения совпадения по размерности, $\widetilde{e}_m^i$ обновленный признак по i голове.

Модели. AutoInt, выходной слой

Применим сигмоидную функцию активации к произведению матрицы эмбеддингов и вектора весов:

$$\hat{P} = \sigma(w^T(e_1^{Dim} \oplus e_2^{Dim} \oplus ... \oplus e_M^{Dim}) + b).$$

- σ — В нашей задаче сигмоидная функция активации,
- e_i^{Dim} — эмбеддинги из слоя взаимодействия,
- w — вектор весов для эмбеддингов,
- b — сдвиг.

При обучении минимизируем логлосс:

$$\text{loss} = \frac{1}{N}\sum_{j=1}^{N}(y_j log(\hat{P}_j) + (1 - y_j) log(1 - \hat{P}_j)).$$

- N — Количество клиентов в обучающей выборке,
- y_j — метка класса,
- $\hat{P}_j$ — предсказанная вероятность.

Методы отбора финальных признаков в модель

- Статистический: отбор признаков по shap значениям

$$\phi_i(p) = \sum_{S \supset N/i} \frac{|S|!(n - |S| - 1)!}{n!}(p(S \cup i) - p(s)).$$

  - $p(S \cup i)$ — предсказание модели с i-ым признаком,
  - $p(S)$ — предсказание модели без i-ого признака
  - n — количество признаков
  - S — набор признаков без i -ого признака

- Последовательный отбор признаков на основе ключевой метрики качества (SFS)

$$x^+ = argmax J(X_k + x), x \in X_d - X_k \quad (1)$$

$$X_{k+1} = X_k + x^+ \quad (2)$$

$$k = k + 1 \quad (3)$$

Отбор гиперпараметров модели

Проблема обычного байесовского подбора гиперпараметров - сильно локализованный поиск. Приведу следующее возможное решение:

- На заданном числе итераций собирается начальная статистика по наборам параметров и значений objective
- По заданному квантилю γ на значениях objective строится распределение вида

$$p(x|y) = \begin{cases} l(x), & \text{if } y < \gamma \\ g(x), & \text{if } y \geq \gamma \end{cases} \tag{4}$$

  где $l(x)$ - лучшие наблюдения, $g(x)$ - все остальные, x - значение оптимизируемого параметра
- Из $l(x)$ производится семплирование заданного числа кандидатов и для каждого x расчитывается величина Exprected Improvement $EI = \frac{l(x)}{g(x)}$
- Рассчитываются значения objective на отобранных кандидатах.

Результаты работы с данными

Данные АльфаБанка по 3 млн. клиентов с 62 признаками.
Таблица результатов по моделям:

model	val	test
xgboost_25_shap_tunned	0.775	0.745
catboost_all_default	0.759	0.677
AutoInt_tunned_hp	0.721	0.71.7
lgboost_all_tunned	0.71	0.686
logreg	0.629	0.624

Общая сумма долга заемщиков в дефолте 62922 денежных единиц, при применении лучшей модели xgboost_25_shap_tunned сумма потерь сократилась на 9438 денежных единиц

Заключение

- Изучена литература по моделям кредитного скоринга и методам их оптимизации;
- Усовершенствован байесовский подбор гиперпараметров (прирост + 1п.п - 1.5 п.п в итоговом качестве)
- Технически реализована нейронная сеть AutoInt с возможностью изменения конфигурации
- Найдены реальные данные Банка по домену кредитного скоринга и на них реализованы все модели и методы их оптимизации
- Получен финансовый эффект в 9438 ден.ед (15% от общей суммы в дефолте)