ПОДХОДЫ В МАШИННОМ ОБУЧЕНИИ ДЛЯ ЗАДАЧ КРЕДИТНОГО СКОРИНГА

Шичков Д.В.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Фролов И.И. – канд. техн. наук, доцент

В данной работе рассматриваются базовые методы и подходы, которые существуют в машинном обучении для решения задач кредитного скоринга: линейная модель, логистическая и решающие деревья.

Линейная регрессия – модель зависимости переменной от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости. Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей [1].

Пусть n = 0.7 – 0.01 * a + 0.5 * k – вероятность невозврата кредита, где a – возраст заёмщика; k – кредитная нагрузка заёмщика. В представленной выражении сделано предположение того, что чем старше заёмщик, тем более вероятнее он вернёт кредит и чем больше у него уже имеется долгов, тем менее вероятнее он вернёт кредит. На рисунке 1 наглядно показано, что линейная модель оценки риска дефолта не подходит для решение данной задачи, т.к. с увеличением возраста заёмщика, вероятность возврата становится отрицательной: возраст больше 70, означает вероятность меньше 0.

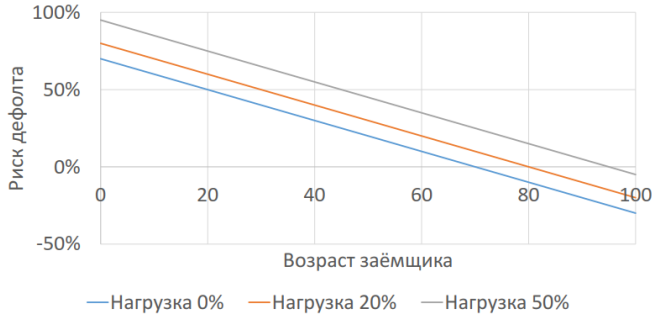

Рисунок 1 – График зависимости риска дефолта от возраста заёмщика с определённой нагрузкой, который иллюстрирует проблему линейной регрессии

Логистическая регрессия представляет собой статистическую модель, которая используется для прогнозирования вероятности возникновения определённого события путём подгонки данных к некой логистической кривой. Главной идей логистическая регрессией является тот факт, что пространство исходных значений может быть разделено на две части соответствующих классам области [2].

Пусть n = 1 / (1 + exp(-Z)), где Z представляет собой любую линейную функцию. Можно заметить, что чем больше эта линейная функция, тем меньше exp(-Z) и тем выше вероятность невозврата кредита.

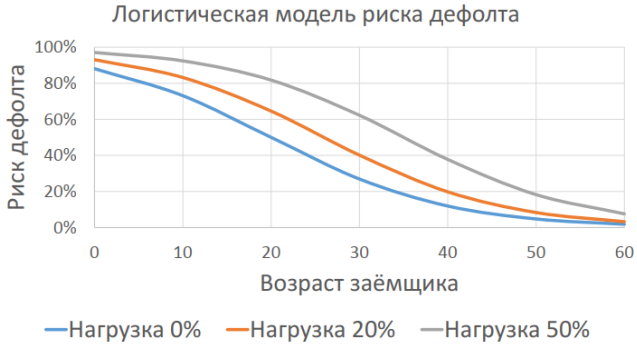

Рисунок 2 – График зависимости риска дефолта от возраста заёмщика с определённой нагрузкой для логистической функции

На рисунке 2 видно, что логистическая функция более подходит для решения задачи оценки кредитных рисков, т.к. значения вероятности дефолта не становится отрицательной, а стремиться к нулю при возрасте старше 60 лет.

Недостатки логистической регрессии:

1) довольно сложная формула (не рассчитать в уме):

$p\,(y = 1) = 1 + \exp 0.02 \times \text{срок} + 0.001 \times \text{сумма} - 0.09 \times \text{возраст} + \cdots - 1;$

2) не умеет учитывать нелинейности;
3) не умеет учитывать взаимодействие признаков;
4) на рисунке 3 показан пример.

Рисунок 3 – Пример недостатка логической регрессии

Можно рассмотреть задачу оценки кредитного риска на примере решающего дерева [3]. На рисунке 4 показан пример решающего дерева на основе желаемого результата и полученного.

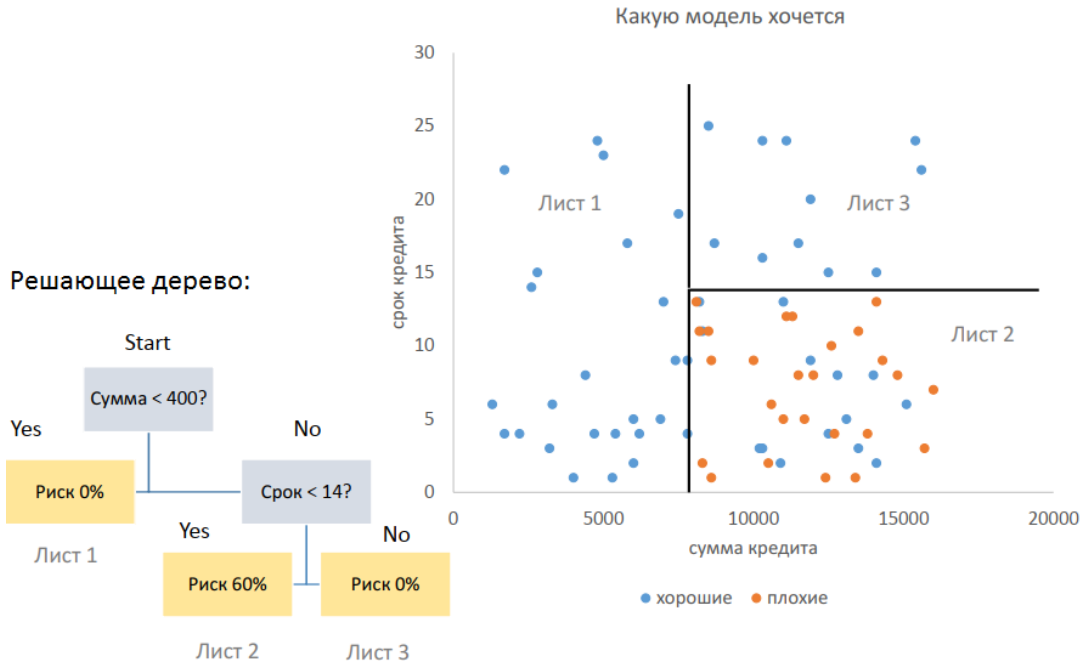

Рисунок 4 – Пример решающего дерева для кредитного скоринга

Алгоритм построения решающего дерева:

1) Перебрать все признаки.
2) Для каждого перебранного признака перебрать все пороги.
3) Выбрать признак и порог, которые приводят к наибольшему росту точности модели
4) Разделить в этом месте выборку на две части.
5) Для каждой разделённой половине повторить шаг 1.
6) Если таковых не оказалось, вернуться.

Список использованных источников:

1. Линейная регрессия в машинном обучении [Электронный ресурс] – Режим доступа: https://neurohive.io/ru/osnovy-data-science/linejnaja-regressija/ – Дата доступа 20.03.2021.

2. Логистическая регрессиям [Электронный ресурс] – Режим доступа: http://statistica.ru/theory/logisticheskaya-regressiya/ – Дата доступа: 20.03.2021.

3. Соколов Е.А. Решающие деревья [Электронный ресурс] – Режим доступа: https://www.hse.ru/mirror/pubs/share/215285956. – Дата доступа: 20.03.2021.