Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления
Кафедра Интеллектуальных информационных технологий

Индивидуальная практическая работа №1
по дисциплине «Статистические основы индуктивного вывода»
на тему

Построение бинарного классификатора средствами MS EXCEL

Выполнил: Е. С. Колосовский

Студент группы
021703
Проверил: А. А. Ефремов

Минск 2023

1 Оценка стоимости домов

1.1 Задачи

1. Подобрать в открытых источниках data set, состоящий из результативного признака (заданного бинарной переменной) и нескольких факторных признаков (не менее 3)
2. Построить бинарный классификатор, пользуясь методическими указаниями из примера ниже.
3. В отчёте представить: постановку задачи с описанием переменных (А), фрагмент таблицы с исходными данными (Б), уравнение логистической регрессии (В), значение $Z_{\text{гр}}$ (Г), оценку надёжности классификатора через расчёт процента ошибок (Д).

1.2 Суть задачи

В качестве датасета будет использоваться `https://www.kaggle.com/datasets/yasserh/housing-prices-datas`
Известны следующие параметры из датасета:

1. Рентабельность покупки
2. Площадь (коэффицент x_1)
3. Этажность (коэффицент x_2)
4. Количество комнат (коэффицент x_3)
5. Стоимость (коэффицент x_3)

Коэффициенты зависимости рентабельности от факторов x_1, x_2, x_3, x_4 являются следующими - 54.0 , 37.0 , 42.0, 72.0 соотвественно Требуется:

1. Построить линейную регрессионную модель для оценки стоимости жилья;
2. Построить регрессионную дискриминантную модель, найти граничное значение и отнести потенциальную недвижимость благоприятной к покупке или нет.

Информация по жилью приведена в следующей таблице.

Площадь (x_1)	Этажность (x_2)	Комнаты (x_3)	Стоимость (x_4)	Рентабельность(Z)	Номер п/п
7420	3	6	13300000	0	1
8960	4	8	12250000	0	2
9960	2	5	12250000	1	3
7500	2	6	12215000	1	4
7420	2	5	11410000	0	5
7500	1	6	10850000	1	6
8580	4	7	10150000	1	7
16200	2	8	10150000	1	8
8100	2	5	9870000	1	9
5750	4	5	9800000	1	10
13200	2	4	9800000	1	11
6000	2	7	9681000	1	12
6550	2	6	9310000	1	13
3500	2	6	9240000	0	14
7800	2	5	9240000	0	15
2787	2	6	2380000	1	16
1836	1	3	2275000	0	17
5300	1	4	2233000	1	18
4600	2	5	1960000	1	19
3850	2	4	1750000	1	20
2400	1	4	1750000	1	21

Таблица 1 – Характеристики недвижимости.

1.3 Выполнение

1.3.1 Пункт 1

Введем таблицу с данными в Excel. Линейная регрессионная модель для вычисления стоимости жилья в данном случае имеет вид: $Z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$.

Для оценки коэффициентов $\beta_k, k = \overline{0,3}$, , будем использовать модуль «Анализ данных», который вызывается из «Сервиса» в главном меню. В «Анализе данных» найдем инструмент «Регрессия» и вызовем его. В появившемся окне укажем входные интервалы Y и X.

Входной интервал Y – это массив ячеек (в таблице исходных данных), содержащих значения объясняемой переменной Z. Входной интервал X – это массив ячеек, содержащих значения объясняющих переменных x_1, x_2, x_3 и x_4 .

После ввода входных интервалов, нажмем на кнопку «OK». В результате появится новый лист с параметрами регрессионной модели. Оценка коэффициента $\widehat{\beta_0}$ равна значению коэффициента β_0 для «Y-пересечения», а оценки $\widehat{\beta_1}$,$\widehat{\beta_2}$,$\widehat{\beta_3}$, $\widehat{\beta_4}$ коэффициентов β_1, β_2, β_3 β_4 равны значениям коэффициентов для переменных x_1, x_2, x_3 и x_4.

Стоимость жилья высчитывается по формуле:$\widehat{Z} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_3 + \widehat{\beta_4} x_4$, (2) где x_1, x_2, x_3 и x_4 и – заданные значения коэффициентов для потенциальной стоимости.

В результате вычислений мы получили следующие значения:

1. $\widehat{\beta_0} = 0$
2. $\widehat{\beta_1} = 0{,}705456542$

3. $\widehat{\beta_2} = 0{,}004544756$
4. $\widehat{\beta_3} = 0{,}000545785$
5. $\widehat{\beta_4} = 0{,}000024846$

При этом коэффицент значимости равен 0,0000356544, что даёт нам уверенность в том ,что построенная регрессия несёт в себе смысл.

1.3.2 Пункт 2

В качестве регрессионной дискриминантной модели можно взять модель из п.1. Для каждого наблюдения вычисляются прогнозные значения показателя по формуле:

$$\widehat{Z}^i = \widehat{\beta_0} + \widehat{\beta_1} x_1^i + \widehat{\beta_2} x_2^i + \widehat{\beta_3} x_3^i + \widehat{\beta_4} x_4^i i = \overline{1, N}$$

Затем с помощью функций СРЗНАЧ и СТАНДОТКЛОН нужно найти средние значения $\overline{Z_1}$ и $\overline{Z_2}$, и стандартные отклонения σ_1 и σ_2 для наблюдений со благоприятной покупкой (1-й массив) и для наблюдений с неблагоприятной покупкой. (Для этого предварительно следует упорядочить таблицу соответствующим образом.)
$\overline{z_1}$=0,531511212
$\overline{z_2}$=0,215154864
σ_1=0,451215178
σ_2=0,147548456

Граничное значение вычисляется по формуле: $Z_{\text{гр}} = \frac{\sigma_1 \overline{Z_2} + \sigma_2 \overline{Z_1}}{\sigma_1 + \sigma_2}$ равное 0,501049885 в нашем случае.
Поскольку $\overline{z_2} > \overline{z_1}$, благоприятность покупки дома оценивается как высокая, если $\widehat{Z} < Z_{\text{гр}}$, и как низкая, если $\widehat{Z} > Z_{\text{гр}}$.
Для нашего потенциального дома вычислим $\widehat{Z}_{\text{пот}} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_3 + \widehat{\beta_4} x_4$=0.75567
Следовательно наш дом благоприятен к покупке.