

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

Задание.

Проанализировать набор данных о клиентах банка и предсказать, будет ли просрочка 90 дней и более при выдаче кредита.

Решение.

Шаг 1: Прочтем данные из файла *data.csv*.

```
import pandas as pd
```

```
# Читаем данные из файла
```

```
data = pd.read_csv('data.csv', delimiter=',')
```

```
data.head()
```

| | Id | SeriousDlqin2yrs | RevolvingUtilizationOfUnsecuredLines | age | \ |
|---|----|------------------|--------------------------------------|-----|---|
| 0 | 1 | 1 | 0.766127 | 45 | |
| 1 | 2 | 0 | 0.957151 | 40 | |
| 2 | 3 | 0 | 0.658180 | 38 | |
| 3 | 4 | 0 | 0.233810 | 30 | |
| 4 | 5 | 0 | 0.907239 | 49 | |

| | NumberOfTime30-59DaysPastDueNotWorse | DebtRatio | MonthlyIncome | \ |
|---|--------------------------------------|-----------|---------------|---|
| 0 | 2 | 0.802982 | 9120.0 | |
| 1 | 0 | 0.121876 | 2600.0 | |
| 2 | 1 | 0.085113 | 3042.0 | |
| 3 | 0 | 0.036050 | 3300.0 | |
| 4 | 1 | 0.024926 | 63588.0 | |

| | NumberOfOpenCreditLinesAndLoans | NumberOfTimes90DaysLate | \ |
|---|---------------------------------|-------------------------|---|
| 0 | 13 | 0 | |
| 1 | 4 | 0 | |
| 2 | 2 | 1 | |
| 3 | 5 | 0 | |
| 4 | 7 | 0 | |

| | NumberRealEstateLoansOrLines | NumberOfTime60-89DaysPastDueNotWorse | \ |
|---|------------------------------|--------------------------------------|---|
| 0 | 6 | 0 | |
| 1 | 0 | 0 | |
| 2 | 0 | 0 | |
| 3 | 0 | 0 | |
| 4 | 1 | 0 | |

| | NumberOfDependents |
|---|--------------------|
| 0 | 2.0 |
| 1 | 1.0 |
| 2 | 0.0 |

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

```
3          0.0
4          0.0
```

Шаг 2: Выведем описание прочтенных данных.

Описание данных

```
data_description = data.describe()
```

```
data_description
```

```
count    Id  SeriousDlqin2yrs  RevolvingUtilizationOfUnsecuredLines  \
mean     675.500000          0.060000          3.577895
std      389.855743          0.237575          84.914699
min       1.000000          0.000000          0.000000
25%      338.250000          0.000000          0.031140
50%      675.500000          0.000000          0.156891
75%     1012.750000          0.000000          0.543145
max     1350.000000          1.000000         2340.000000
```

```
count    age  NumberOfTime30-59DaysPastDueNotWorse  DebtRatio  \
mean     52.048889          0.257778          356.123363
std      15.009875          0.751718         1156.603074
min      22.000000          0.000000          0.000000
25%      40.000000          0.000000          0.175125
50%      52.000000          0.000000          0.367049
75%      63.000000          0.000000          0.807001
max      97.000000          10.000000         15466.000000
```

```
count    MonthlyIncome  NumberOfOpenCreditLinesAndLoans  \
mean     6438.473492          8.434074
std      7849.754675          5.129287
min       0.000000          0.000000
25%      3300.000000          5.000000
50%      5222.500000          8.000000
75%      8055.250000         11.000000
max     208333.000000         31.000000
```

```
count    NumberOfTimes90DaysLate  NumberRealEstateLoansOrLines  \
mean           0.080000          0.986667
std           0.376634          1.008401
min           0.000000          0.000000
25%           0.000000          0.000000
50%           0.000000          1.000000
75%           0.000000          2.000000
max           5.000000          8.000000
```

```
count    NumberOfTime60-89DaysPastDueNotWorse  NumberOfDependents
1350.000000          1307.000000
```

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

| | | |
|------|----------|----------|
| mean | 0.062222 | 0.737567 |
| std | 0.306555 | 1.086949 |
| min | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 |
| 50% | 0.000000 | 0.000000 |
| 75% | 0.000000 | 1.000000 |
| max | 5.000000 | 8.000000 |

Шаг 3: Отобразим несколько первых и последних записей.

Для функций `head()` и `tail()` можно передать параметр `n`, который указывает на количество строк, которые необходимо отобразить. Если параметр не указан, по умолчанию будет показано 5 строк.

Выводим первые 5 записей

`first_records = data.head()`

Выводим последние 5 записей

`last_records = data.tail()`

`first_records, last_records`

| | Id | SeriousDlqin2yrs | RevolvingUtilizationOfUnsecuredLines | age | \ |
|---|----|------------------|--------------------------------------|-----|---|
| 0 | 1 | 1 | 0.766127 | 45 | |
| 1 | 2 | 0 | 0.957151 | 40 | |
| 2 | 3 | 0 | 0.658180 | 38 | |
| 3 | 4 | 0 | 0.233810 | 30 | |
| 4 | 5 | 0 | 0.907239 | 49 | |

| | NumberOfTime30-59DaysPastDueNotWorse | DebtRatio | MonthlyIncome | \ |
|---|--------------------------------------|-----------|---------------|---|
| 0 | 2 | 0.802982 | 9120.0 | |
| 1 | 0 | 0.121876 | 2600.0 | |
| 2 | 1 | 0.085113 | 3042.0 | |
| 3 | 0 | 0.036050 | 3300.0 | |
| 4 | 1 | 0.024926 | 63588.0 | |

| | NumberOfOpenCreditLinesAndLoans | NumberOfTimes90DaysLate | \ |
|---|---------------------------------|-------------------------|---|
| 0 | 13 | 0 | |
| 1 | 4 | 0 | |
| 2 | 2 | 1 | |
| 3 | 5 | 0 | |
| 4 | 7 | 0 | |

| | NumberRealEstateLoansOrLines | NumberOfTime60-89DaysPastDueNotWorse | \ |
|---|------------------------------|--------------------------------------|---|
| 0 | 6 | 0 | |
| 1 | 0 | 0 | |
| 2 | 0 | 0 | |
| 3 | 0 | 0 | |
| 4 | 1 | 0 | |

©МатБюро – Консультации по математике, программированию, экономике, праву, естественным наукам

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

```

NumberOfDependents
0          2.0
1          1.0
2          0.0
3          0.0
4          0.0 ,

      Id SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age \
1345 1346                0                0.000000        39
1346 1347                0                0.045694        49
1347 1348                0                0.022780        53
1348 1349                0                0.036934        56
1349 1350                0                0.000000        62

      NumberOfTime30-59DaysPastDueNotWorse DebtRatio MonthlyIncome \
1345                0      0.055916        4166.0
1346                0      0.300175        4000.0
1347                0      0.323068       10000.0
1348                0      0.287935        8362.0
1349                0 1463.000000           NaN

      NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate \
1345                5                0
1346               14                0
1347               14                0
1348                8                0
1349                5                0

      NumberRealEstateLoansOrLines NumberOfTime60-89DaysPastDueNotWorse \
1345                0                0
1346                1                0
1347                2                0
1348                1                0
1349                1                0

      NumberOfDependents
1345                0.0
1346                1.0
1347                1.0
1348                2.0
1349                0.0 )
    
```

Шаг 4: Прочитаем описание столбцов из файла DataDictionary.ru.txt и определим тип каждого столбца.

Для начала прочитаем содержимое файла.

```

# Читаем содержимое файла с описанием столбцов
with open('DataDictionary.txt', 'r', encoding='utf-8') as file:
    data_dictionary_content = file.readlines()
    
```

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

data_dictionary_content

```
['SeriousDlqin2yrs    \n',
 'Произойдет ли просрочка платежа на 90 или более дней в ближайшие 2 года
(Y/N)\n',
 '\n',
 'RevolvingUtilizationOfUnsecuredLines    \n',
 'Процент оставшихся денег на всех кредитных картах    (%) \n',
 '\n',
 'age \n',
 'Возраст (integer)\n',
 '\n',
 'NumberOfTime30-59DaysPastDueNotWorse    \n',
 'Количество кратковременных просрочек платежа (30-59 дней) за последние 2
года (integer)\n',
 '\n',
 'DebtRatio    \n',
 'Выплачиваемые в месяц проценты от долга, алименты, бытовые расходы,
деленные на месячный доход (если месячный доход неизвестен, то абсолютная
величина выплат) (%) \n',
 '\n',
 'MonthlyIncome    \n',
 'Месячный доход (real)\n',
 '\n',
 'NumberOfOpenCreditLinesAndLoans \n',
 'Количество кредитов (или кредитных карт) за исключением ипотеки
(integer)\n',
 '\n',
 'NumberOfTimes90DaysLate \n',
 'Кредитная история: количество серьезных просрочек платежа на 90 дней или
более (integer)\n',
 '\n',
 'NumberRealEstateLoansOrLines    \n',
 'Количество ипотек (integer)\n',
 '\n',
 'NumberOfTime60-89DaysPastDueNotWorse    \n',
 'Количество среднесрочных просрочек платежа (60-89 дней) за последние 2 года
(integer)\n',
 '\n',
 'NumberOfDependents    \n',
 'Количество иждивенцев в семье, включая самого клиента (integer)\n']
```

- SeriousDlqin2yrs: Произойдет ли просрочка платежа на 90 или более дней в ближайшие 2 года (Y/N) - категориальный
- RevolvingUtilizationOfUnsecuredLines: Процент оставшихся денег на всех кредитных картах (%) - вещественный
- age: Возраст - целый

©МатБюро – Консультации по математике, программированию, экономике, праву, естественным наукам

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

- NumberOfTime30-59DaysPastDueNotWorse: Количество кратковременных просрочек платежа (30-59 дней) за последние 2 года - целый
- DebtRatio: Выплачиваемые в месяц проценты от долга, алименты, бытовые расходы, деленные на месячный доход - вещественный
- MonthlyIncome: Месячный доход - вещественный
- NumberOfOpenCreditLinesAndLoans: Количество кредитов (или кредитных карт) за исключением ипотеки - целый
- NumberOfTimes90DaysLate: Кредитная история: количество серьезных просрочек платежа на 90 дней или более - целый
- NumberRealEstateLoansOrLines: Количество ипотек - целый
- NumberOfTime60-89DaysPastDueNotWorse: Количество среднесрочных просрочек платежа (60-89 дней) за последние 2 года - целый
- NumberOfDependents: Количество иждивенцев в семье, включая самого клиента - целый

Теперь перейдем к следующему шагу.

Шаг 5: Исправление данных в столбце DebtRatio

Значения в столбце DebtRatio, соответствующие известному месячному доходу, являются отношениями. Остальные значения являются абсолютными значениями месячных выплат процентов. Нам необходимо преобразовать все значения этого столбца в абсолютные значения.

Для этого мы можем умножить значения в DebtRatio на MonthlyIncome для тех строк, где MonthlyIncome известен.

Обновляем значения в столбце DebtRatio

```
data['DebtRatio'] = data['DebtRatio'] *  
data['MonthlyIncome'].where(data['MonthlyIncome'].notnull(), 1)
```

```
data['DebtRatio'].head()
```

```
0    7323.197016  
1     316.878123  
2     258.914887  
3     118.963951  
4    1584.975094
```

```
Name: DebtRatio, dtype: float64
```

Шаг 6: Изменим имя столбца DebtRatio на Debt.

Переименовываем столбец

```
data.rename(columns={'DebtRatio': 'Debt'}, inplace=True)
```

Проверяем изменение

```
data.head().columns
```

©МатБюро – Консультации по математике, программированию, экономике, праву, естественным наукам

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

```
Index(['Id', 'SeriousDlqin2yrs', 'RevolvingUtilizationOfUnsecuredLines',  
      'age',  
      'NumberOfTime30-59DaysPastDueNotWorse', 'Debt', 'MonthlyIncome',  
      'NumberOfOpenCreditLinesAndLoans', 'NumberOfTimes90DaysLate',  
      'NumberRealEstateLoansOrLines', 'NumberOfTime60-  
89DaysPastDueNotWorse',  
      'NumberOfDependents'],  
      dtype='object')
```

Шаг 7: Вычислим средний ежемесячный доход и присвоим это значение всем клиентам с неизвестным доходом.

Вычисляем средний ежемесячный доход

```
average_monthly_income = data['MonthlyIncome'].mean()
```

Заполняем пропущенные значения средним доходом

```
data['MonthlyIncome'].fillna(average_monthly_income, inplace=True)
```

Проверяем, все ли пропущенные значения были заменены

```
data['MonthlyIncome'].isnull().sum()
```

```
0
```

Шаг 8: Используя метод *groupby*, оценим вероятности невозврата кредита (*SeriousDlqin2yrs=1*) для различных значений количества иждивенцев (*NumberOfDependents*) и для различных значений столбца *NumberRealEstateLoansOrLines*.

Оценка вероятности невозврата кредита для разных значений

NumberOfDependents

```
default_probability_dependents =  
data.groupby('NumberOfDependents')['SeriousDlqin2yrs'].mean()
```

Оценка вероятности невозврата кредита для разных значений

NumberRealEstateLoansOrLines

```
default_probability_real_estate =  
data.groupby('NumberRealEstateLoansOrLines')['SeriousDlqin2yrs'].mean()
```

```
default_probability_dependents, default_probability_real_estate
```

```
(NumberOfDependents
```

```
0.0    0.041397
```

```
1.0    0.089844
```

```
2.0    0.110465
```

```
3.0    0.057143
```

```
4.0    0.033333
```

```
5.0    0.000000
```

```
6.0    0.000000
```

```
8.0    0.000000
```

```
Name: SeriousDlqin2yrs, dtype: float64,
```

© МатБюро – Консультации по математике, программированию, экономике, праву, естественным наукам

Поможем вам с написанием программ: www.matburo.ru/sub_subject.php?p=pz

NumberRealEstateLoansOrLines

| | |
|---|----------|
| 0 | 0.056863 |
| 1 | 0.048729 |
| 2 | 0.063158 |
| 3 | 0.145455 |
| 4 | 0.105263 |
| 5 | 0.000000 |
| 6 | 1.000000 |
| 8 | 0.000000 |

Name: SeriousDlqin2yrs, dtype: float64)

Вот вероятности невозврата кредита (SeriousDlqin2yrs=1) для различных значений:

По количеству иждивенцев (NumberOfDependents):

- 0 иждивенцев: ~4.14%
- 1 иждивенец: ~8.98%
- 2 иждивенца: ~11.05%
- 3 иждивенца: ~5.71%
- 4 иждивенца: ~3.33%
- 5 иждивенцев: 0%
- 6 иждивенцев: 0%
- 8 иждивенцев: 0%

По количеству ипотек (NumberRealEstateLoansOrLines):

- 0 ипотек: ~5.69%
- 1 ипотека: ~4.87%
- 2 ипотеки: ~6.32%
- 3 ипотеки: ~14.55%
- 4 ипотеки: ~10.53%
- 5 ипотек: 0%
- 6 ипотек: 100% (это может быть аномалией или особенностью нашей выборки данных)
- 8 ипотек: 0%

Эти результаты могут помочь банку в оценке рисков при выдаче кредитов различным категориям клиентов.