

## Построение модели Такаги-Сугено

**Credit** – набор данных, содержит данные по среднему долгу по кредитной карте (balance) для ряда клиентов банка, а также несколько количественных предикторов: age (возраст), cards (количество карт), education (количество лет, потраченных на образование), income (доход, тыс. долларов), limit (кредитный лимит) и rating (кредитный рейтинг). Помимо этих количественных переменных, у нас есть также четыре качественные переменные: gender (пол), student (является ли клиент студентом), status (состоит ли клиент в браке) и ethnicity (белый, афроамериканец или азиат).

Разбиваем таблицу Credit на кластеры

```
x1<-data.frame(cbind(Income=Credit$Income, Rating=Credit$Rating, Student=Credit$Student,
Balance=Credit$Balance))
myclusters <- kmeans(x = x1, centers = 4)
myclusters$cluster
myclusters$centers
```

#Формируем кластеры

```
cluster1<-Credit[myclusters$cluster==1,-1]
cluster2<-Credit[myclusters$cluster==2,-1]
cluster3<-Credit[myclusters$cluster==3,-1]
cluster4<-Credit[myclusters$cluster==4,-1]
```

Центры кластеров

```
> myclusters$centers
  Income  Rating Student  Balance
1 92.17267 615.5581 1.348837 1380.48837
2 41.17069 354.7757 1.056075  481.89720
3 50.10497 452.3878 1.122449  900.43878
4 31.63538 218.5000 1.046053   58.15132
> |
```

4-й кластер с самой маленькой задолженностью, в среднем 58.15.

2-й – больше, в среднем 481,89.

3-й- в среднем 900.44.

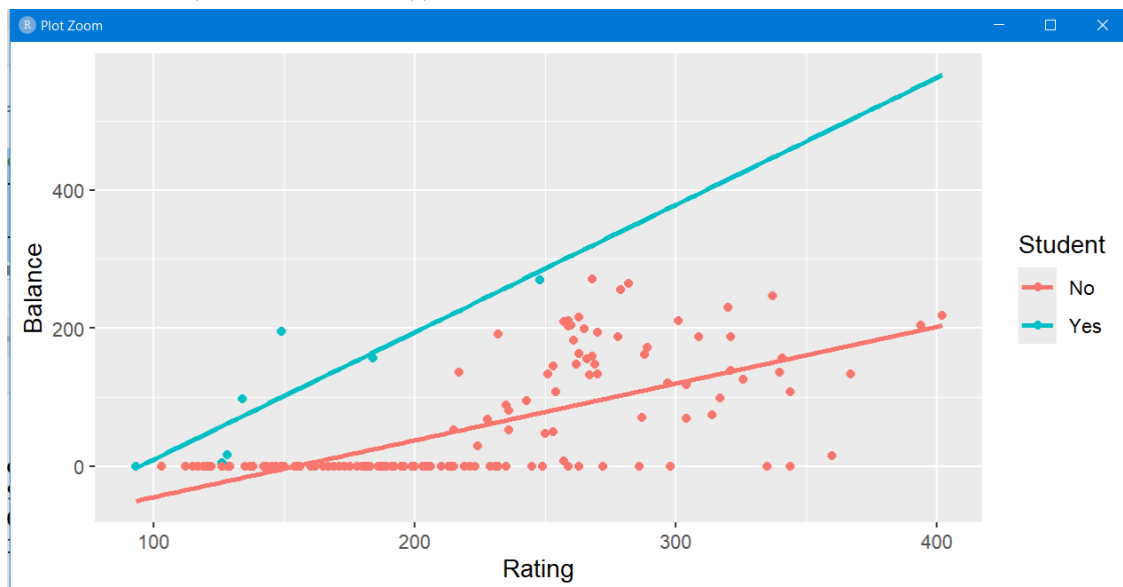
1-й кластер с самой большей задолженностью, в среднем 1380.49.

Для каждого кластера находим значимые количественные предикторы, с точки зрения влияния на Balance.

**Начинаем с 4-го кластера** (самого благополучного, с точки зрения задолженности).

**Средний баланс Balance 58,15, Income 31.63, Rating 615.56**

```
#Исследуем четвертый кластер с самой маленькой задолженностью,  
#Исключаем при этом качественные переменные  
model_4<-lm(data=cluster4[, -c(7:10)], Balance~.)  
#Большие значения коэффициент vif (>5) указывает на сильную зависимость  
признаков от других признаков  
vif(model_4)  
# Убираем переменную Limit  
model_4<-lm(data=cluster4[, -c(7:10)], Balance~.-Limit)  
vif(model_4)  
# Включаем дополнительно качественные переменные  
model_4<-lm(data=cluster4, Balance~.-Limit)  
summary(model_4)  
# В итоговую модель включаем только значимые переменные  
model_4<-lm(data=cluster4, Balance~Income+Rating+Student)  
summary(model_4)  
#Нарисуем зависимость Balance от Rating и Student  
library(ggplot2)  
ggplot(data=cluster4, aes(x=Rating, y=Balance, colour = Student))+geom_point()+  
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE,  
    aes(color=Student))
```



**Кластер 4: Income, Rating, Student.**

Из-за того, что много нулевых значений переменной Balance, проведем дополнительные исследования

#Более детальное исследование кластера 4

# Много нулевых значений у переменной Balance

# Итоговая модель для наблюдений, у которых рейтинг превышает 210

```
model_4_1<-lm(data=cluster4[cluster4$Rating>210,], Balance~Income+Rating+Student)
```

```
summary(model_4_1)
```

# Итоговая модель для студентов, у которых рейтинг меньше 210

```
model_4_2<-lm(data=cluster4[cluster4$Rating<=210 & cluster4$Student=="Yes",],
```

```
Balance~Income+Rating)
```

```
summary(model_4_2)
```

#Баланс равен нулю для не студентов с рейтингом, меньшим 210

Правила:

Если  $Rating > 210$ , то  $Balance = -362.08 - 4.89Income + 2.41Rating + 299.34Student$

$Rating \leq 210$  и  $Student = "Yes"$ , то  $Balance = -271,2 - 4,25Income + 3.24Rating$

Для не студентов с  $Rating \leq 210$   $Balance = 0$

**Исследуем кластер 1.**

*Здесь средний долг по карте в среднем равен 1380.48. Income примерно 92.17, Rating примерно 615.56. Age – 56.8 лет*

$$Balance = 616,98 + 1,4Rating - 3.39 \cdot Age + 268.49 \cdot Student$$

Правило: если студент, то  $Balance = 885,47 + 1,4Rating - 3.39 \cdot Age$

если не студент, то  $Balance = 616,98 + 1,4Rating - 3.39 \cdot Age$

**Исследуем кластер 2**

*Здесь средний долг по карте в среднем равен 481.9. Income примерно 41.17, Rating примерно 354.8. Age – 53.4 года*

$$Balance = -561.1 - 8.67Income + 4.12 \cdot Rating - 1.161Age + 399.6 \cdot Student - 33.3Married$$

**Исследуем последний кластер 3**

$$Balance = -524 - 8.45Income + 4.08 \cdot Rating - 0.91Age + 412.46Student$$

### Входные переменные

**Income:** Примерно 32, Примерно 41, Примерно 50, Примерно 92

**Rating:**

Меньше 210, Больше 210, но до примерно 355, Примерно 355, Примерно 452, Примерно 616

**Age:** Примерно 56 (кластер 4), примерно 57 (кластер 1), примерно 54 (кластер 2), примерно 58 (кластер 3)

**Student:** 0 или 1

**Правила:**

Income	Rating	Age	Student	Married	Уравнение
условия правил					уравнения для заключений правил
Кластер 4					
Примерно 32	Около 210 или меньше. Можно задать с помощью трапециевидной функции принадлежности с параметрами (0;0;210;220)		0		0
Примерно 32	Около 210 или меньше.		1		$Balance = -271,2 - 4,25Income + 3.24Rating$
Примерно 32	Примерно от 200 до 350. Можно задать в виде треугольного числа (200; 300; 350)				$Balance = -362.08 - 4.89Income + 2.41Rating + 299.34Student$
Кластер 1					
Примерно	Примерно	Примерно			$Balance = 616,98 + 1,4Rating - 3.39 \cdot Age + 268.49 \cdot Student$

о 92	616	о 57			
<b>Кластер 2</b>					
Примерн о 42	Примерно 355	Примерн о 54			$Balance = -561.1 - 8,67Income + 4,12 \cdot Rating - 1.161Age +$ $+ 399.6 \cdot Student - 33.3Married$
<b>Кластер 3</b>					
Примерн о 50	Примерно 453	Примерн о 57			$Balance = -524 - 8,45Income + 4,08 \cdot Rating - 0.91Age +$ $+ 412.46Student$