Б-тесты
в антифрод-скоринге

КБ «Ренессанс Кредит»
Афанасьев Сергей

История

1881г.
Ньюком
астроном

Обнаружил, что в логарифмических таблицах единиц больше чем двоек, двоек больше чем троек и т.д.

1938г.
Бенфорд
физик

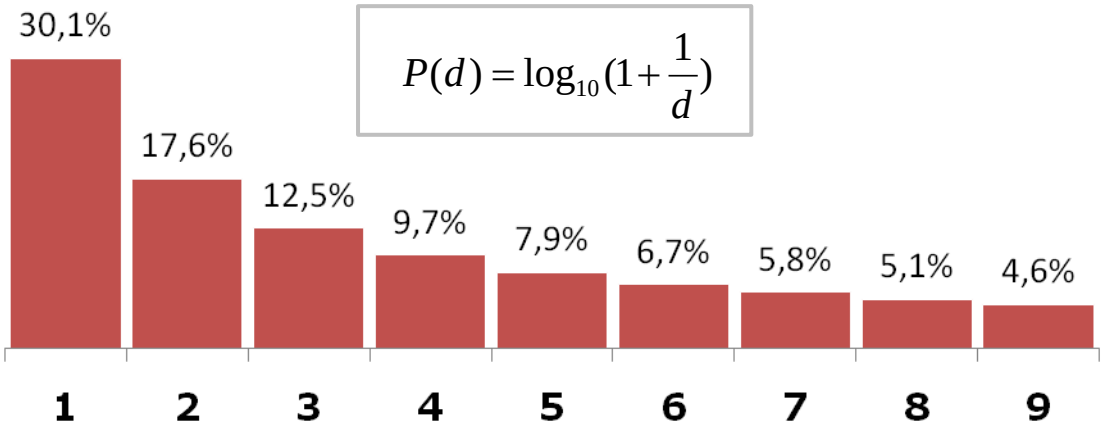
Проанализировал различные справочники с данными и построил эмпирический закон распределения цифр

1997г.
Ernst & Yang
компания

Разработали и начали использовать 6 тестов для аудита отчетности компаний и выявления поддельных данных

Распределение Бенфорда

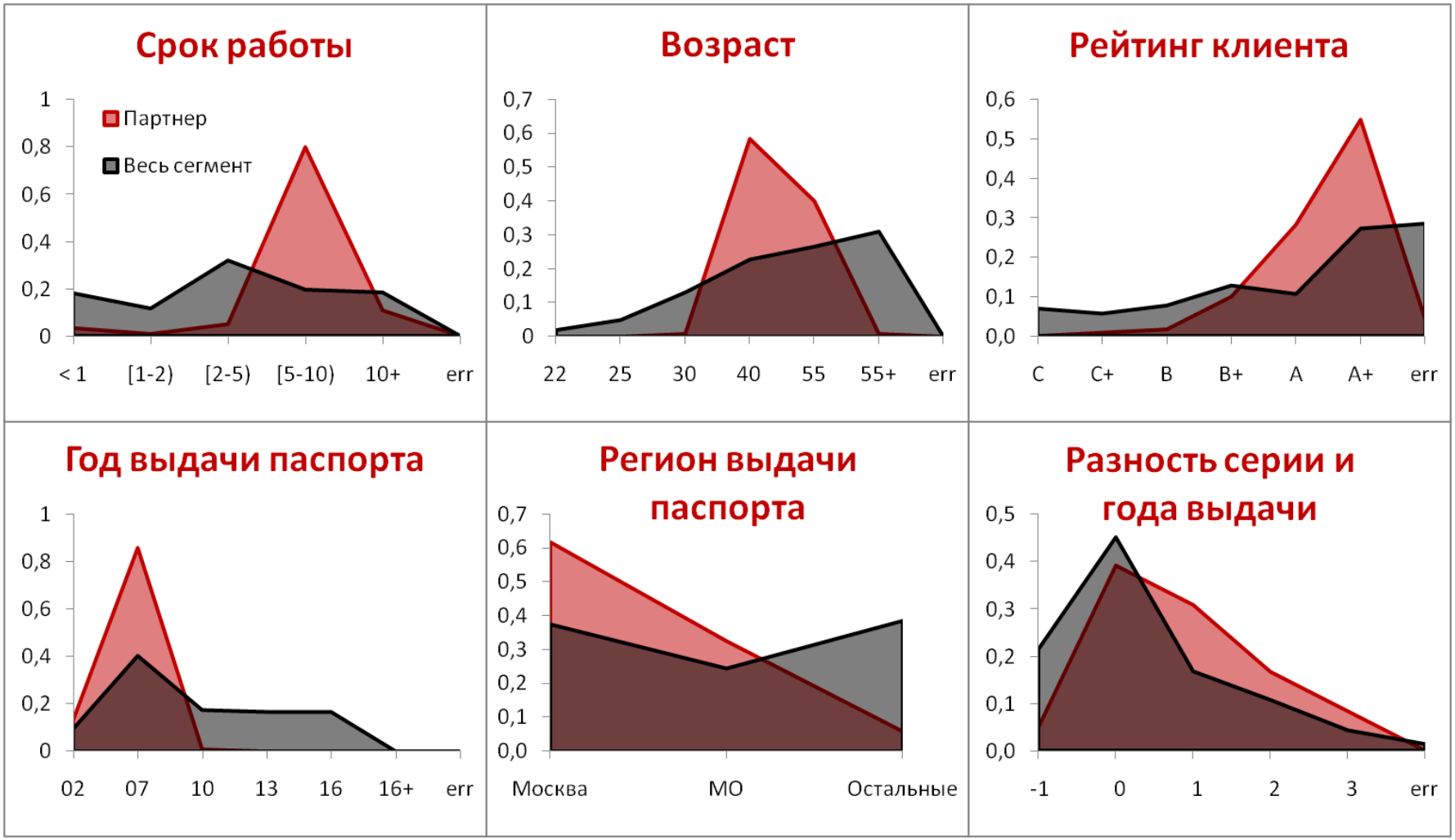
$$P(d) = \log_{10}(1 + \frac{1}{d})$$



Мошенничество

POS-партнер оформил 120 мошеннических кредитов по поддельным документам, используя ПО «Генератор скана паспортов»

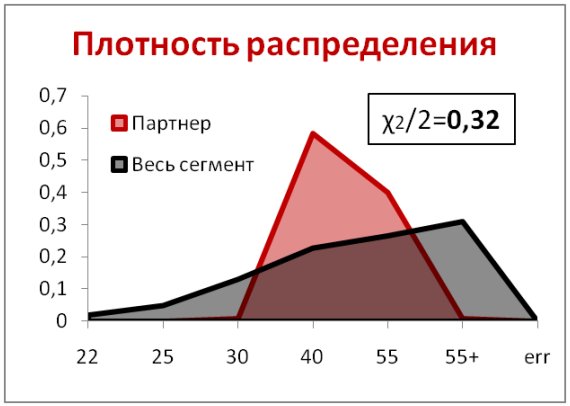
Аномалии


Срок работы
Партнер
Весь сегмент
1
0,8
0,6
0,4
0,2
0
< 1
[1-2)
[2-5)
[5-10)
10+
err
Возраст
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
22
25
30
40
55
55+
err
Рейтинг клиента
0,6
0,5
0,4
0,3
0,2
0,1
0,0
C
C+
B
B+
A
A+
err
Год выдачи паспорта
1
0,8
0,6
0,4
0,2
0
02
07
10
13
16
16+
err
Регион выдачи паспорта
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
Москва
МО
Остальные
Разность серии и года выдачи
0,5
0,4
0,3
0,2
0,1
0,0
-1
0
1
2
3
err

Статистики

Хи-квадрат

$$\chi^2 = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{a_i + b_i}$$



Минусы:

Значения нормированной статистики Хи-квадрат чаще сосредоточены у нуля и реже у единицы

Колмогорова-Смирнова

$$KS = \max_i |F(a_i) - F(b_i)|$$

Минусы:

Для распределений с несколькими локальными максимумами статистика KS может быть заниженной

S-статистика

$$S = \sum_{i=1}^{n} \frac{|a_i - b_i|}{2}$$

Плюсы:

S-статистика просто интерпретируется – это половина площади под непересекающимися областями

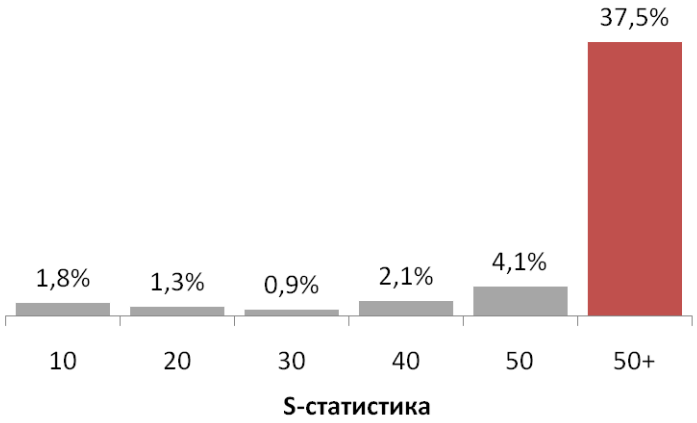
Fraud-статистики

Название	$\chi^2/2$	KS	S
Отрасль	70%	48%	**80%**
5-я цифра дохода (с конца)	43%	55%	**61%**
Срок работы	39%	53%	**60%**
Должность	40%	54%	**59%**
Совпадение адресов	37%	56%	**56%**
Год по серии паспорта	31%	50%	**50%**
Год выдачи паспорта	33%	50%	**50%**
Возраст	32%	30%	**49%**
Рейтинг	25%	24%	**45%**
Регион по паспорту	16%	33%	**33%**
Пол	11%	32%	**32%**
Год по серии минус дата	9%	23%	**24%**
4-я цифра дохода (с конца)	7%	20%	**22%**

Б-тесты

Триггер

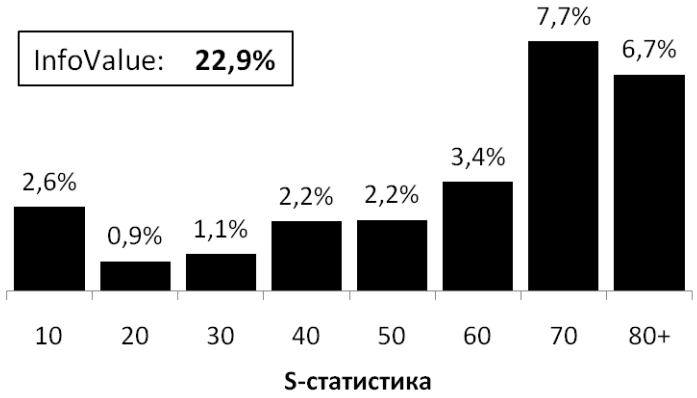
"Год выдачи паспорта по серии"
(доля fraud-точек)



Триггер позволяет выделить высокорисковую группу ТО/сотрудников, которую эффективно отправлять на расследование. По оставшейся выборке пересчитывается InfoValue для использования Б-теста в антифрод-скоркарте.

Предиктор

"4-я с конца цифра дохода"
(доля fraud-точек)



Б-тесты, с помощью которых нельзя выделить высокорисковую группу ТО/сотрудников, используются как предикторы. Предикторы с высоким InfoValue включаются в антифрод-скоркарту.

Результаты

Название	Тип	S-порог %	fraud-group	IV	IV_short
Год по серии паспорта	trigger	60+	37,5%	12,2%	8,2%
Год выдачи паспорта	trigger	50+	30,0%	10,6%	7,1%
5-я цифра дохода (с конца)	predictor	50+	3,9%	34,0%	-
Совпадение адресов	predictor	50+	3,5%	28,5%	-
4-я цифра дохода (с конца)	predictor	70+	7,5%	22,9%	-
Срок работы	predictor	50+	5,7%	15,2%	-
Регион по паспорту	predictor	90+	5,9%	13,6%	-
Возраст	predictor	50+	8,2%	12,5%	-
Регион ТО = Регион по паспорту	predictor	60+	3,9%	11,2%	-
Пол	-	30+	3,8%	8,3%	-
Рейтинг	-	40+	2,4%	7,6%	-
Год по серии минус дата	-	30+	2,4%	4,1%	-
Отрасль	-	-	-	-	-
Должность	-	-	-	-	-

Спасибо за внимание!