Применение статистического аппарата

При Анализе Данных

Роман
Гараев

Задача #1. Данные о Кредитном Лимите и Сравнение Различных Групп Относительно Возврата/Невозврата

Применить статистический аппарат проверки гипотез, составления доверительных интервалов и визуализации в целях изучения данных и построения выводов.

LIMIT_BAL: размер кредитного лимита (в том числе и на семью клиента) SEX: пол клиента (1 = мужской, 2 = женский)

EDUCATION: образование (0 = доктор, 1 = магистр; 2 = бакалавр; 3 = выпускник школы; 4 = начальное образование; 5 = прочее; 6 = нет данных).

MARRIAGE: (0 = отказываюсь отвечать; 1 = замужем/женат; 2 = холост; 3 = нет данных).

AGE: возраст в годах

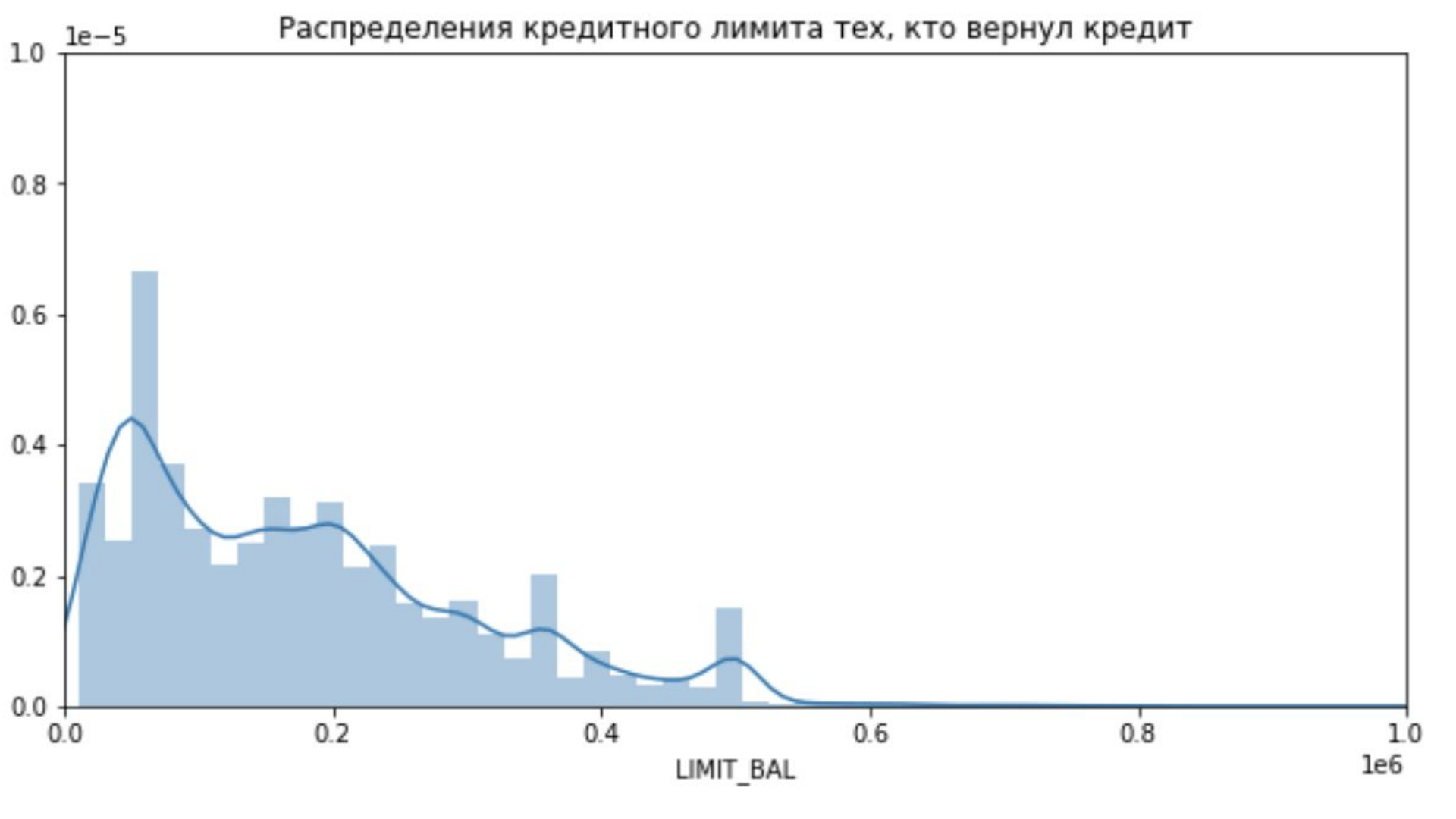
PAY_0 - PAY_6 : История прошлых платежей по кредиту. PAY_6 - платеж в апреле, ..., PAY_0 - платеж в сентябре. Платеж = (0 = исправный платеж, 1 = задержка в один месяц, 2 = задержка в 2 месяца ...)

BILL_AMT1 - BILL_AMT6: задолженность, BILL_AMT6 - на апрель, ..., BILL_AMT1 - на сентябрь

PAY_AMT1 - PAY_AMT6: сумма уплаченная в PAY_AMT6 - апреле, ..., PAY_AMT1 - сентябре

default - индикатор невозврата денежных средств

Проверка равенства медианных значений 2 групп

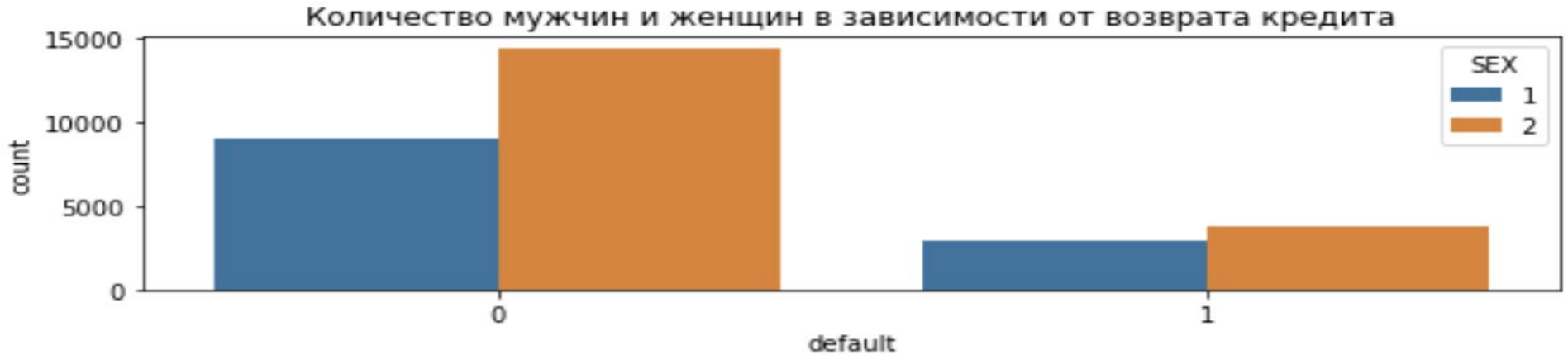

Распределения кредитного лимита тех, кто вернул кредит
1e−5
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
LIMIT_BAL

Использованные техники и Вывод

Проверка равенства медианных значений с помощью сравнения медианных значений, доверительных интервалов и применения перестановочного критерия.

Вывод: Заемщики, которые не возвращают кредит берут большую сумму.

Сравнение данных о возврата/невозврате относительно пола



Использованные техники и Вывод

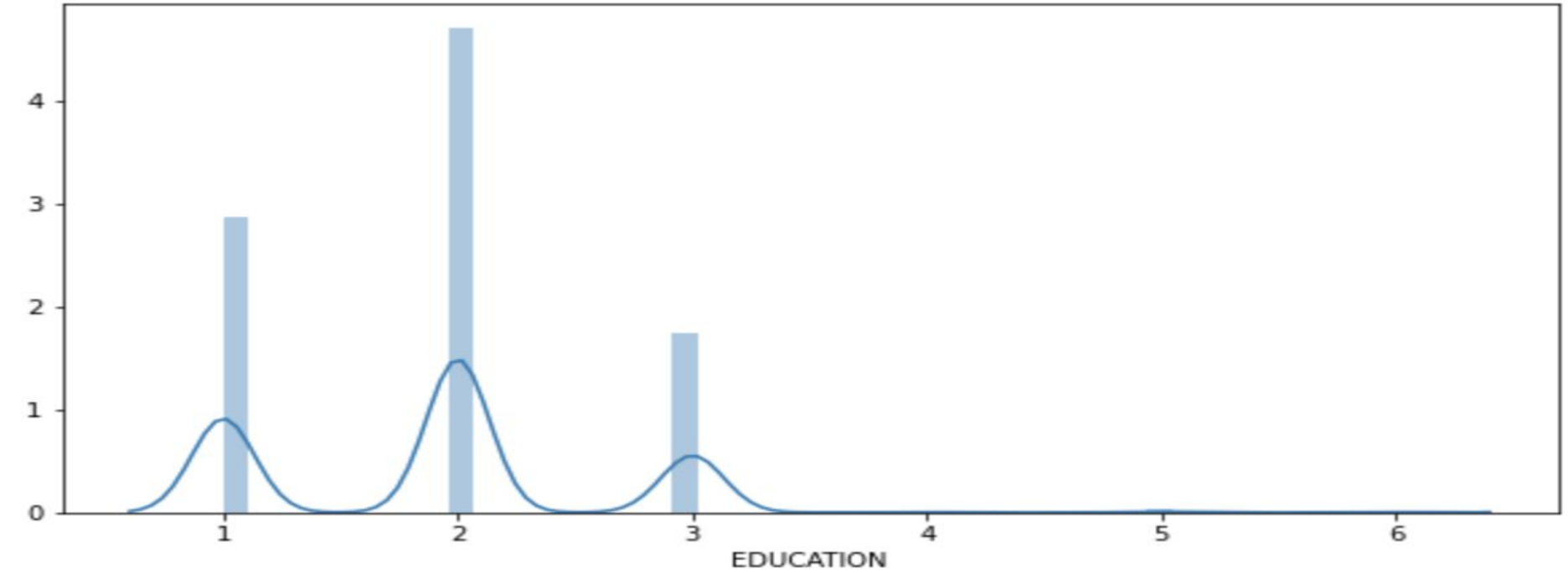
Построение доверительных интервалов для долей и разности долей

Вывод: Доверительный интервал для разности долей не включает ноль, а значит можно говорить о различных долях возврат в зависимости от пола.

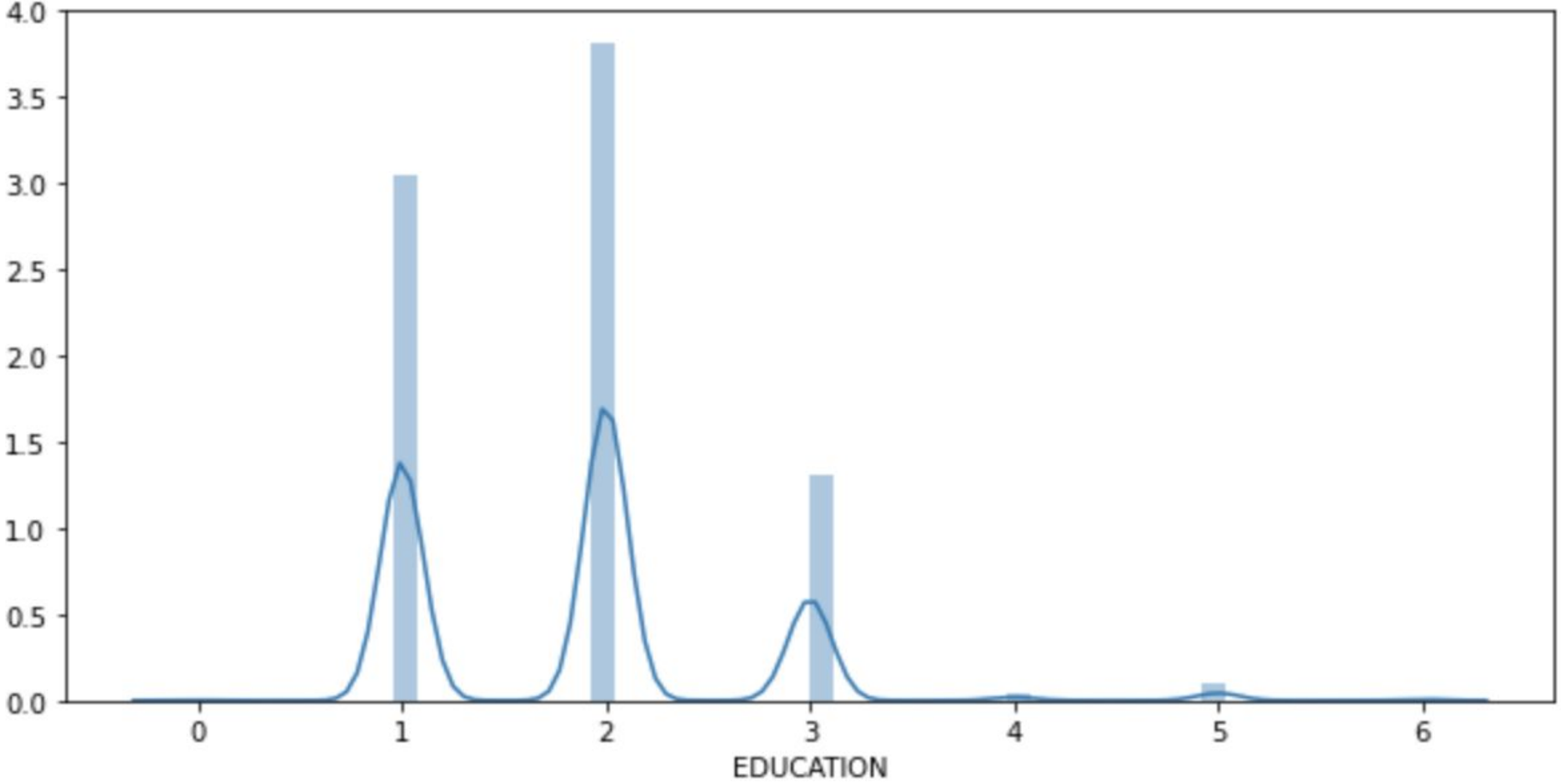
P.S. Данный вывод справедлив только для данного датасета. На самом деле данный признак никак нельзя использовать для реальной оценки по причине неэтичности.

Сравнение данных об образовании относительно возврата/невозврата

Возврат



Невозврат



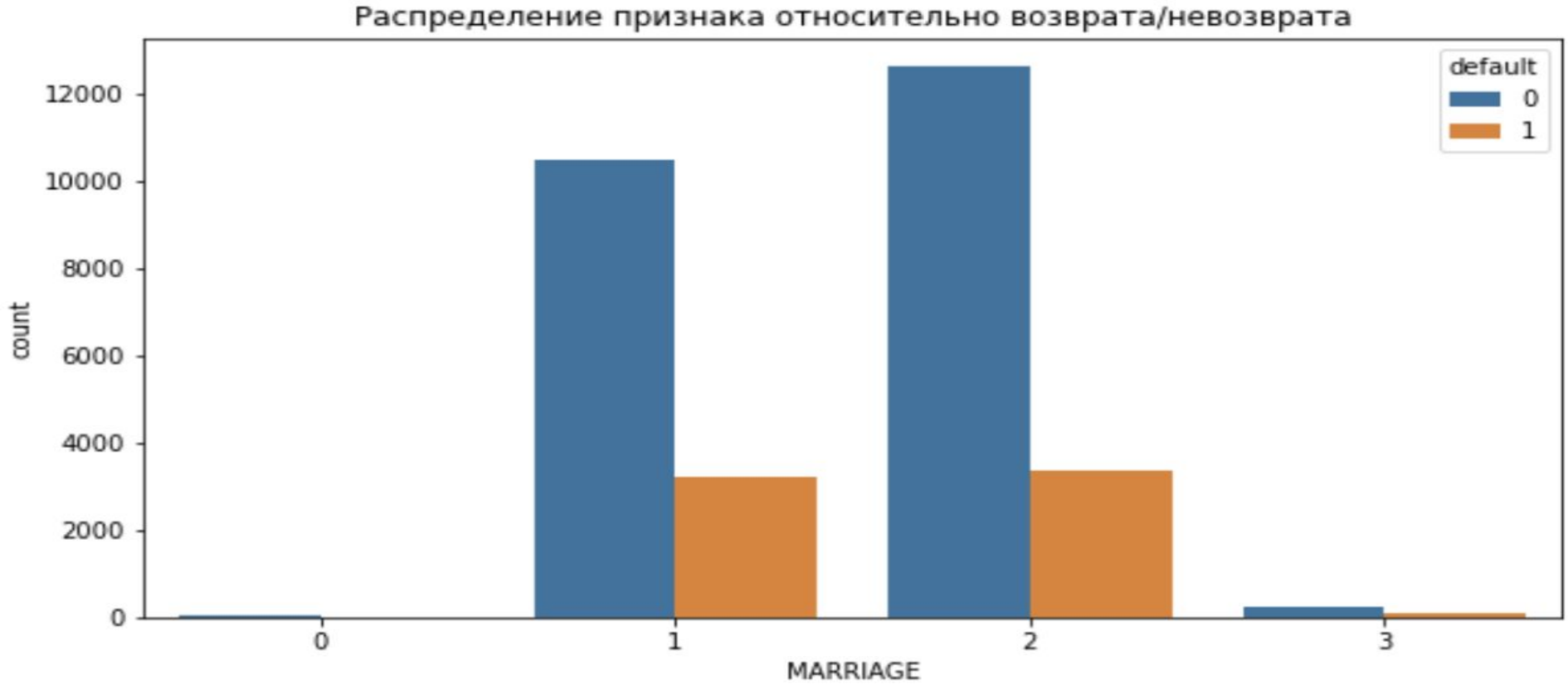

Пропорция возвратов/невозвратов среди разных уровней образования
1.0
0.9
0.8
0.7
0.6
0.5
0
1
2
3
4
5
6

Использованные техники и Вывод

Проверка на соответствие распределения нормальному с помощью критерия согласия Пирсона ХИ-квадрат

Вывод: Распределение признака отличается от нормального. Уровень образования влияет на вероятностью возврата/невозврата

Сравнение данных о семейном положении относительно возврата/невозврата

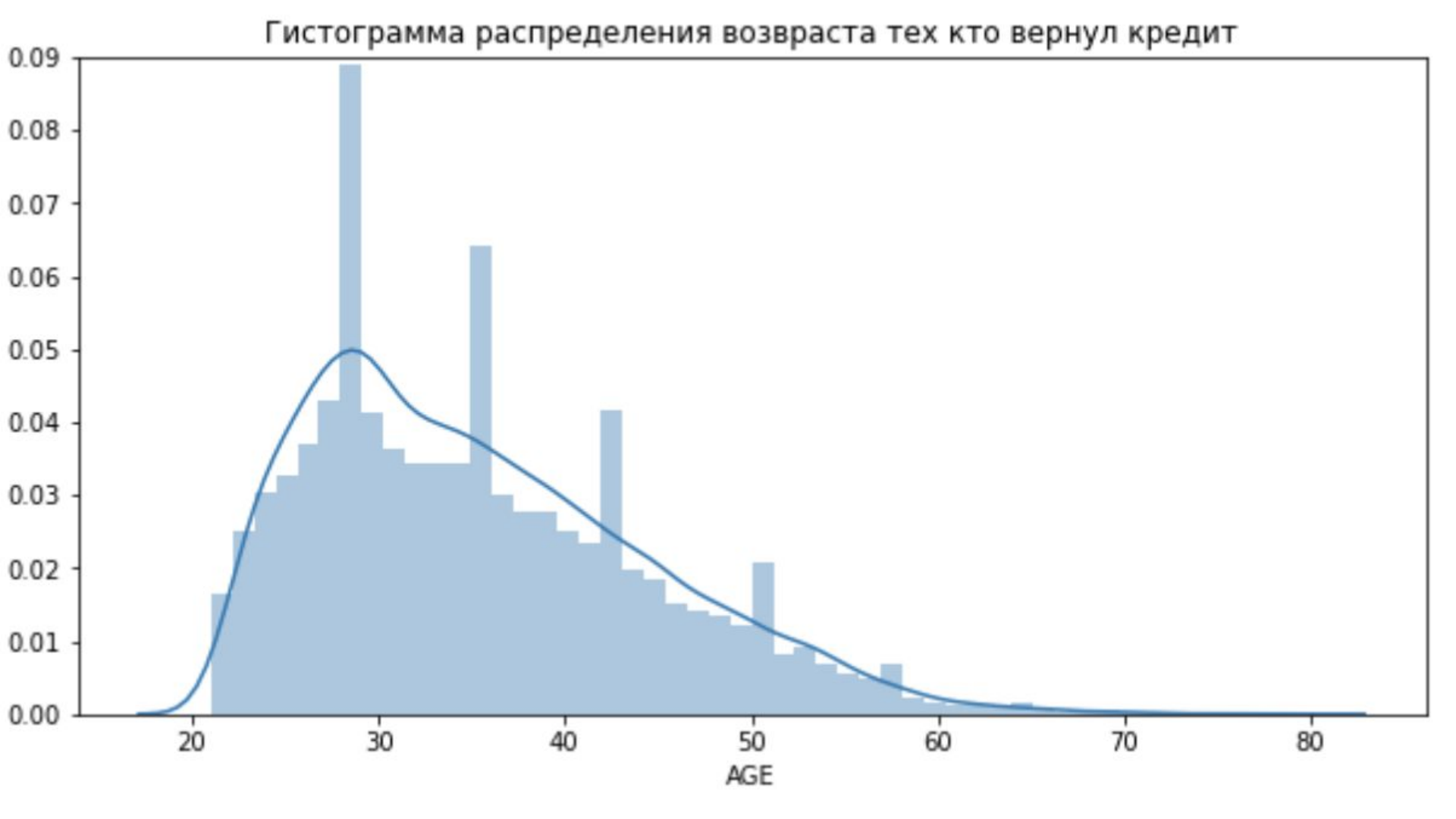


Использованные техники и Вывод

Коэффициент корреляции “V” Крамера

Вывод: Судя по коэффициенту Крамера линейной связи между полом и возвратом/невозвратом нет. Это соответствует здравому смыслу, а значит практическая значимость присутствует

Сравнение данных о возрасте относительно возврата/невозврата


Гистограмма распределения возвраста тех кто вернул кредит
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
80
AGE

Использованные техники и Вывод

Интервальная оценка, доверительные интервалы, проверка гипотезы о равенстве распределений при помощи непараметрического перестановочного критерия проверки средних.

Вывод: Распределения выборок отличаются, возраст влияет на вероятность возврата/невозврата

Задача #2. Анализ эффективности удержания

state — штат США

account_length — длительность использования аккаунта

area_code — деление пользователей на псевдорегионы, использующееся в телекоме

intl_plan — подключена ли у пользователя услуга международного общения

vmail_plan — подключена ли у пользователя услуга голосовых сообщений

vmail_message — количество голосых сообщений, который пользователь отправил / принял

day_calls — сколько пользователь совершил дневных звонков

day_mins — сколько пользователь проговорил минут в течение дня

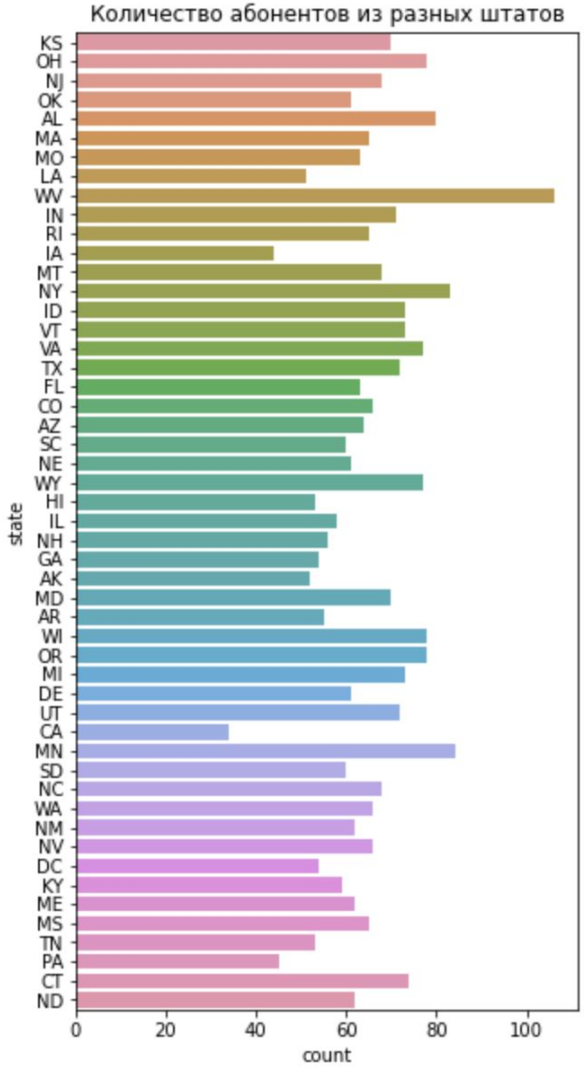
day_charge — сколько пользователь заплатил за свою дневную активность

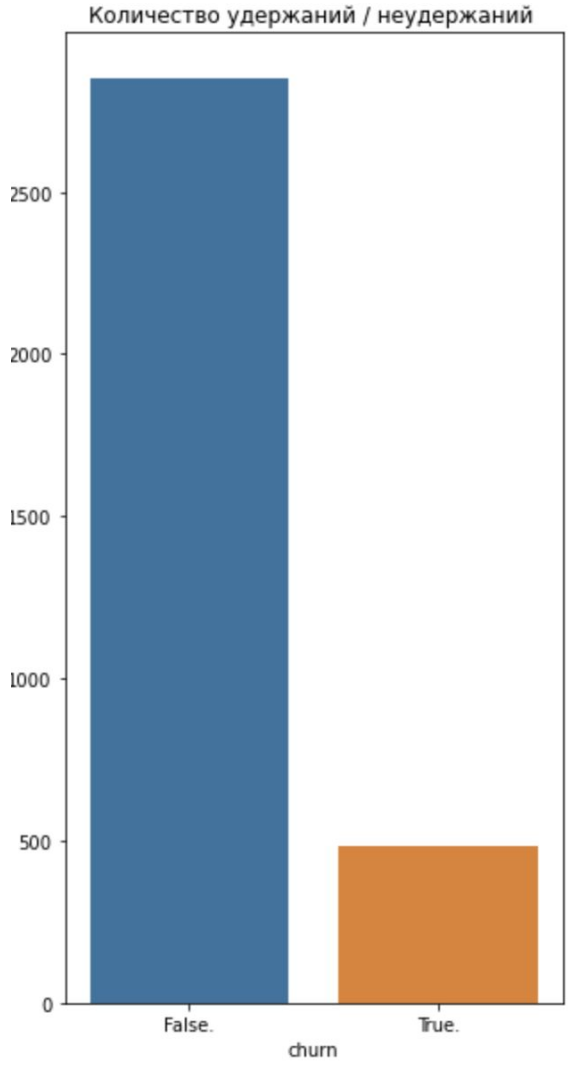
custserv_calls — сколько раз пользователь позвонил в службу поддержки

treatment — номер стратегии, которая применялись для удержания абонентов (0, 2 = два разных типа воздействия, 1 = контрольная группа)

mes_estim — оценка интенсивности пользования интернет мессенджерами

churn — результат оттока: перестал ли абонент пользоваться услугами оператора


Количество абонентов из разных штатов
KS
OH
NJ
OK
AL
MA
MO
LA
WV
IN
RI
IA
MT
NY
ID
VT
VA
TX
FL
CO
AZ
SC
NE
WY
HI
IL
NH
GA
AK
MD
AR
WI
OR
MI
DE
UT
CA
MN
SD
NC
WA
NM
NV
DC
KY
ME
MS
TN
PA
CT
ND
state
0
20
40
60
80
100
count


Количество удержаний / неудержаний
2500
2000
1500
1000
500
0
False.
True.
churn

Проверка гипотезы о том, что штат абонента не влияет на то, перестанет ли абонент пользоваться услугами оператора

Проверка проводилась с помощью критерия хи-2 квадрат для таблицы сопряженности, а в дальнейшем была использована поправка Йетса (коррекция точности аппроксимации на непрерывность)

Вывод: до проверки в 34 из 1241 пар штатов последний влиял на результат, однако после поправки 0/1241

Корреляция между дневными звонками и интенсивностью использования мессенджера

Для проверки величины линейной взаимосвязи между данными признаками были использованы коэффициенты корреляции Спирмена и Мэтьюса

Вывод: Значение коэффициентов корреляции не превышают 0.013 , а значит можно сказать об отсутствии линейной взаимосвязи

Сравнение методов удержания и контрольной группы

Для сравнения методов удержания были построены доверительные интервалы Уилсона для каждой пары групп. Далее также 3 пары доверительных интервалов для разности долей. Наконец с помощью p-value из проверки на разности долей и поправки на множественную проверку гипотез Холма-Бонфферони был сделал окончательный вывод

Вывод: На основании данных можно сказать об эффективности метода “2”

Задача #3 Анализ Результатов АБ теста

- userID: уникальный идентификатор пользователя
- browser: браузер, который использовал userID
- slot: в каком статусе пользователь участвовал в исследовании (exp = видел измененную страницу, control = видел неизменную страницу)
- n_clicks: количество кликов, которые пользоваль совершил за n_queries
- n_queries: количество запросов, который совершил userID, пользуясь браузером browser
- n_nonclk_queries: количество запросов пользователя, в которых им не было совершено ни одного клика

Сравнение среднего количества кликов в контрольной и экспериментальной группах

Сравнение проводилось с помощью средних значений, доверительных интервалов и их разностей

Вывод: Кликов больше было в экспериментальной группе

Сравнение доли кликов в обеих группах по разным браузерам

Сравнение проводилось с помощью непараметрического критерия сравнения независимых выборок Манна-Уитни и поправки Холма на множественную проверку гипотез.

Вывод: Тестируемое значение можно предложить внедрить для той доли пользователей, для которых улучшение показало себя практически и теоретически значимо