Пермский филиал федерального государственного автономного
образовательного учреждения высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Сведение задачи оценки эффекта от воздействия к задаче регрессии. Метод двух моделей

Работу выполнил:
Студент группы ИАУП-17 НМ
Кучумов Артем

Пермь, 2018

Постановка задачи

Имеется некоторый набор данных вида <X, Y, T>, где

- X – набор признаков эксперимента
- Y – искомые значения в рамках эксперимента
- T – воздействие во время эксперимента

Необходимо оценить индивидуальный эффект от воздействия

Пример задачи

Исходная выборка

Время прибытия (X_1)	День недели (X_2)	Количество человек (X_3)	Чек (Y)	Предоставилась скидка (T)
17:30	Понедельник	1	445	0
12:23	Пятница	4	1567	0
21:48	Среда	2	1234	1
…	…	…	…	…
10:29	Понедельник	1	245	1
16:08	Суббота	3	751	0
14:02	Вторник	2	947	1

Кластеризация пользователей

Постоянные клиенты

Одноразовые клиенты

Любители наживы

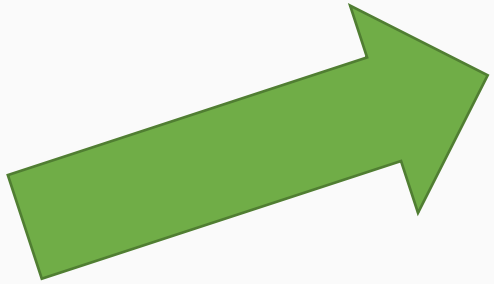
Недотроги

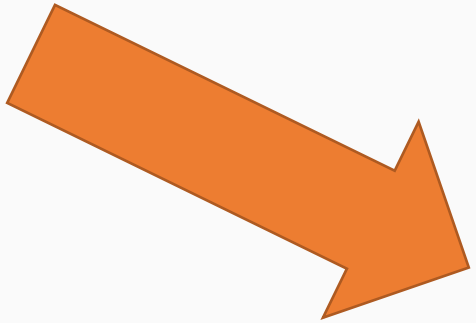
Метод двух моделей

Деление данных

Исходная выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	…
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	0





Выборка с воздействием

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1

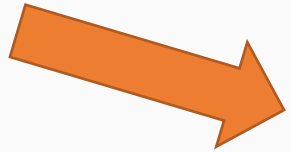
Выборка без воздействия

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0

Деление данных на test/train

Выборка с воздействием

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	…
…	…	…	…	…	1
…	…	…	…	…	1

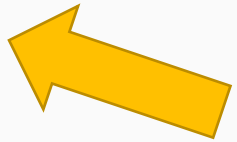


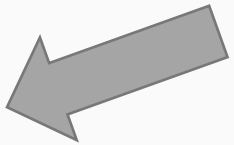
Обучающая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	…
…	…	…	…	…	1

Тестовая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	…
…	…	…	…	…	1





Обучающая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	…
…	…	…	…	…	0

Тестовая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	…
…	…	…	…	…	0

Выборка без воздействия

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	…
…	…	…	…	…	0
…	…	…	…	…	0

Наборы данных

Обучающая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0

Тестовая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0

Исходная выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	…
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	1
…	…	…	…	…	0
…	…	…	…	…	0

Обучающая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1

Тестовая выборка

X_1	X_2	…	X_n	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1

Обучение моделей

Обучающая выборка

X_1	X_2	…	X_N	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1

Модель с воздействием

Обучающая выборка

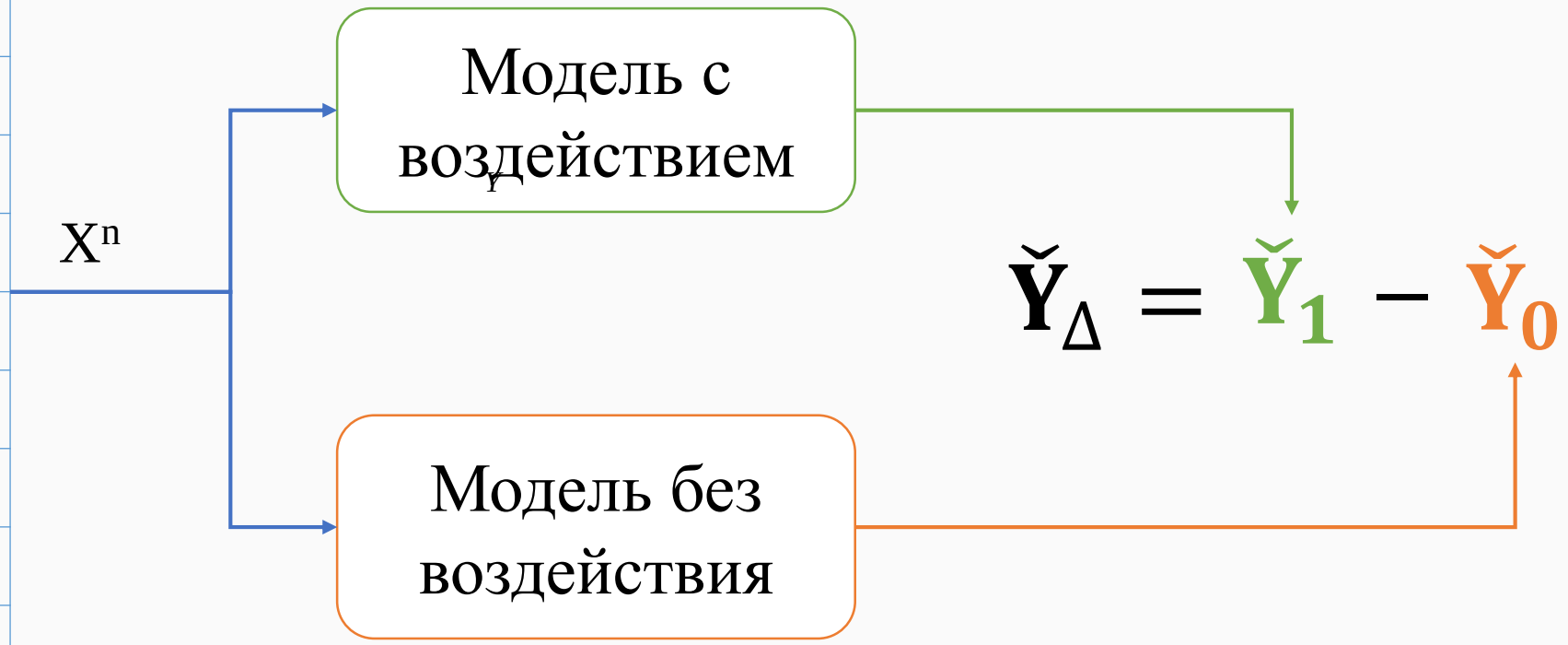
X_1	X_2	…	X_N	Y	T
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0

Модель без воздействия

Вычисление эффекта

Тестовая выборка

X_1	X_2	…	X_N	Y	T
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	1
…	…	…	…	…	…
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0
…	…	…	…	…	0



Алгоритмы регрессии

- Линейная регрессия
- Метод опорных векторов
- Наивный байесовский подход
- Деревья решений
- Случайный лес
- Градиентный бустинг

Данные для тестирования

- Синтетические данные
- Mine That Data 2008

Синтетические данные

- Всего 100 000 наблюдений
- $Y = X_1 + X_2 * T + E$
- $X_1 = N(0, 10)$
- $X_2 = N(0, 1)$
- $X_3 = N(0, 10)$
- $T = R_{0.5}(0, 1)$
- $E = N(0, 10)$

x1	x2	x3	t	y
-4.682088	-0.479330	5.853140	1	-16.907571
-8.228249	0.152597	-9.560731	1	6.168450
-0.653801	0.796189	7.909909	1	-1.915019
-7.133619	-1.365698	-10.525031	1	-29.735355
9.063509	-1.006977	4.531767	1	14.568520
7.662367	0.458075	-20.877253	1	-10.879150
8.260541	1.691872	2.016003	1	11.659834
13.236828	0.388031	4.557115	0	-8.579580
17.524445	-1.531967	-16.693999	1	-18.016732
10.024491	-0.058524	11.687139	1	12.767626
5.448095	2.381058	5.377824	0	-3.349772
18.951609	-0.995676	-3.666407	0	12.528117
-7.693575	-0.714523	-10.293597	0	-16.799552
14.030959	-0.668677	-7.914176	1	-30.511036
-6.324675	1.528842	-10.155746	0	-0.233012

Вычисление среднего эффекта

X_1	…	X_N	Y_*	Y	T
…	…	…	12.4	…	1
…	…	…	4.4	…	1
…	…	…	-12.1	…	1
…	…	…	-2.22	…	1
…	…	…	0.4	…	1
…	…	…	…	…	…
…	…	…	13.3	…	0
…	…	…	-12.4	…	0
…	…	…	1.5	…	0
…	…	…	-4.57	…	0
…	…	…	0.98	…	0

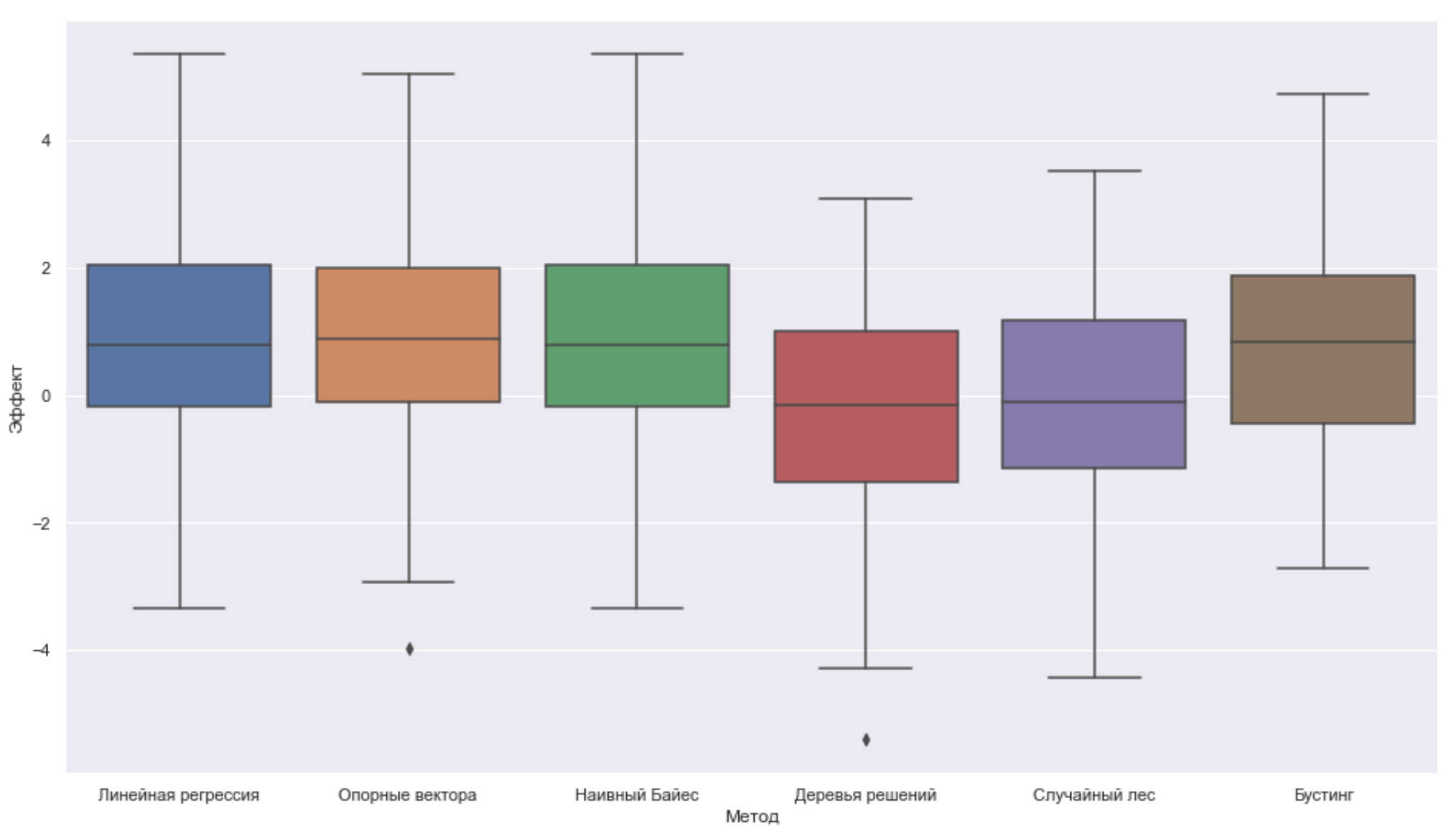
Sort By Y_*

X_1	…	X_N	Y_*	Y	T
…	…	…	22.4	…	1
…	…	…	14.4	…	0
…	…	…	12.5	…	1
…	…	…	12.7	…	1
…	…	…	9.4	…	0
…	…	…	…	…	…
…	…	…	0.3	…	0
…	…	…	-1.4	…	1
…	…	…	-1.5	…	1
…	…	…	-4.57	…	0
…	…	…	-6.98	…	0

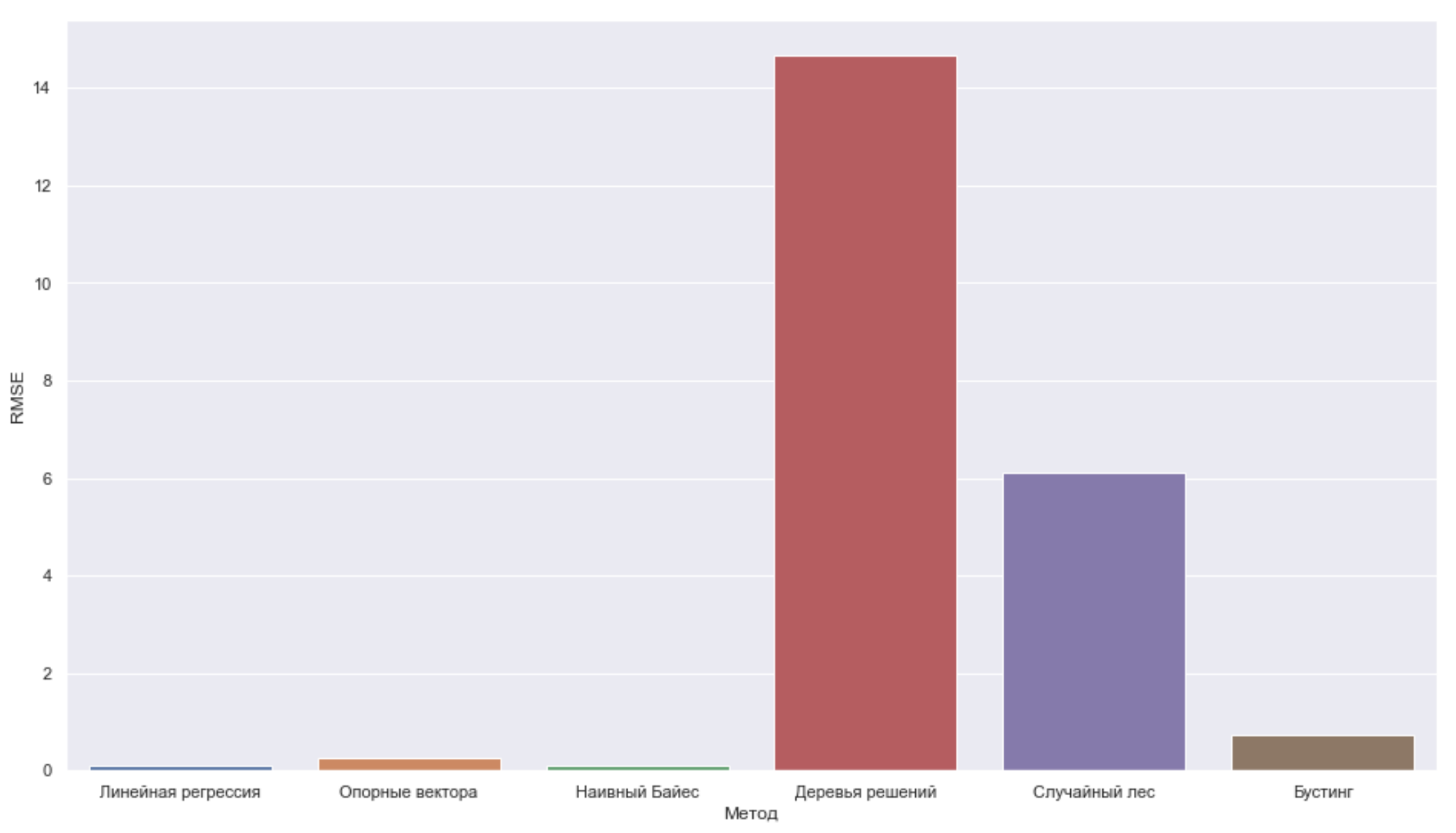
N%

$$\tilde{Y} = \bar{Y}_1 - \bar{Y}_0$$

Эффект от воздействия



Точность


RMSE
0
2
4
6
8
10
12
14
Линейная регрессия
Опорные вектора
Наивный Байес
Деревья решений
Случайный лес
Бустинг
Метод

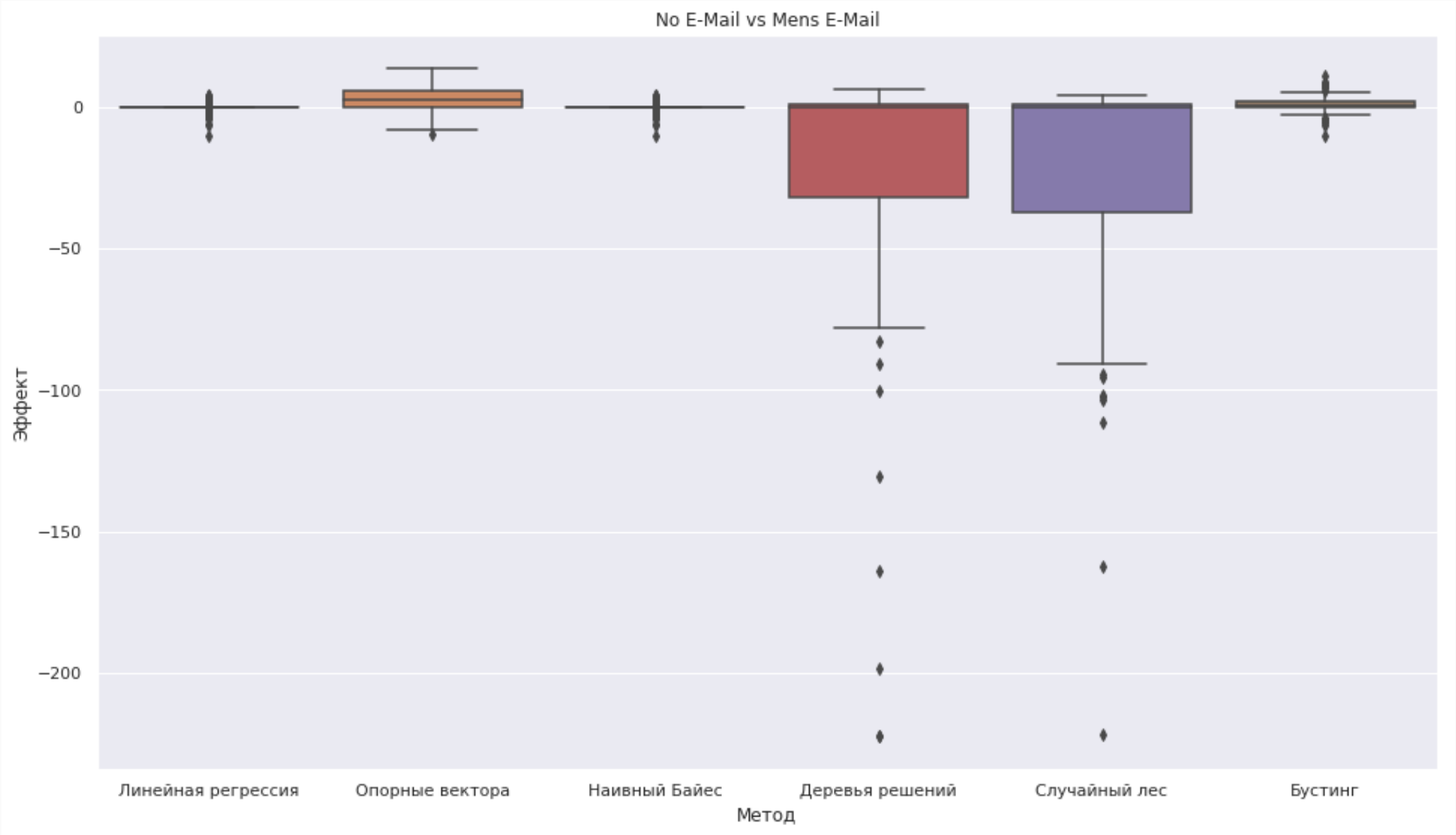
Mine That Data 2008

- Всего 64 000 наблюдений
- 578 наблюдений, где значение spend не равно 0
- 21307 - Mens E-Mail
- 21387 - Womens E-Mail
- 21306 - No E-Mail

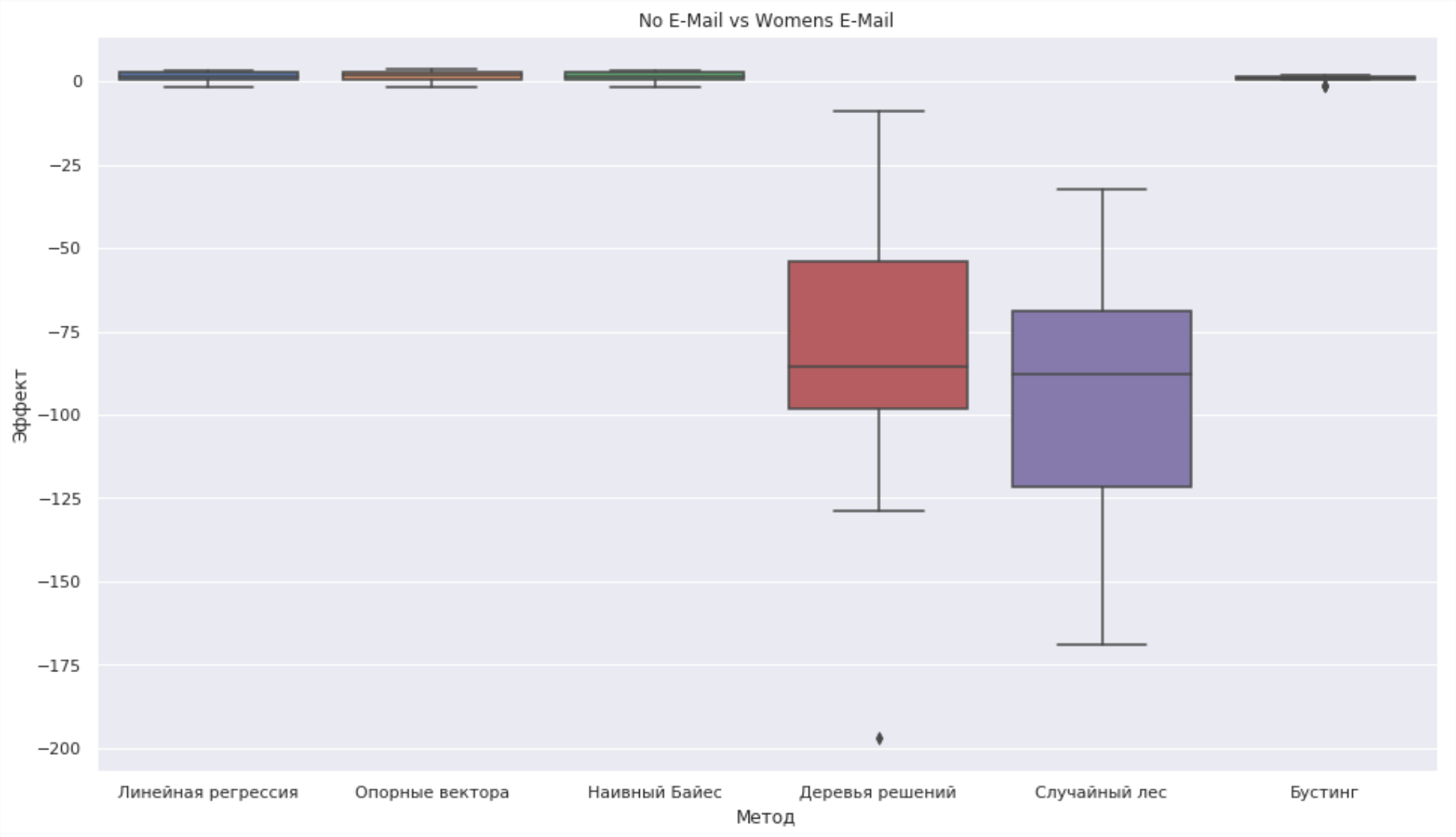
recency	history_segment	history	mens	womens	zip_code	newbie	channel	segment	visit	conversion	spend
10	2) 100–200	105.54	1	0	Urban	0	Web	Mens E-Mail	0	0	0.0
5	1) 0–100	38.91	0	1	Urban	1	Phone	Mens E-Mail	0	0	0.0
6	1) 0–100	29.99	1	0	Urban	1	Phone	Mens E-Mail	0	0	0.0
1	5) 500–750	552.94	1	0	Surburban	1	Multichannel	Womens E-Mail	0	0	0.0
1	4) 350–500	472.82	0	1	Surburban	0	Web	Mens E-Mail	0	0	0.0

https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

Эффект от воздействия



Эффект от воздействия



Выводы

- Для получения хороших результатов достаточно использовать линейные модели
- Линейные модели и бустинг дают стабильные результаты
- Метод двух моделей отлично работает на данных с простыми зависимостями
- Метод двух моделей плохо работает на реальных данных со сложными зависимостями

Спасибо за внимание!

Работу выполнил:
Студент группы ИАУП-17
Кучумов Артем
kuchumov7@gmail.com