Learning for learning

Выполнил: Клейман Вадим

Руководитель: Шпильман Алексей

Stepik

Stepik - образовательная платформа и конструктор бесплатных открытых онлайн-курсов и уроков.

Что было сделано?

Построен классификатор, который может, с относительно высокой точностью, определять пользователей, которые не справятся со степом.

Цель и задачи

- **Цель**:

Научиться предсказывать время прохождения step-а.

- **Задачи**:

1. Подобрать признаки.

2. Построить регрессионную модель, для наиболее точного прогнозирования времени выполнения степа.

Классификация

Используемые классификаторы

- RandomForestClassifier
- GradientBoostingClassifier
- AdaBoostClassifier
- XGBoost
- LGBMClassifier

Признаки

Список признаков, описывающих конкретный step:

- **Среднее время**, затраченное пользователями на step;
- **Количество** пользователей, **прошедших** step;
- **Количество** пользователей, **проваливших** step.

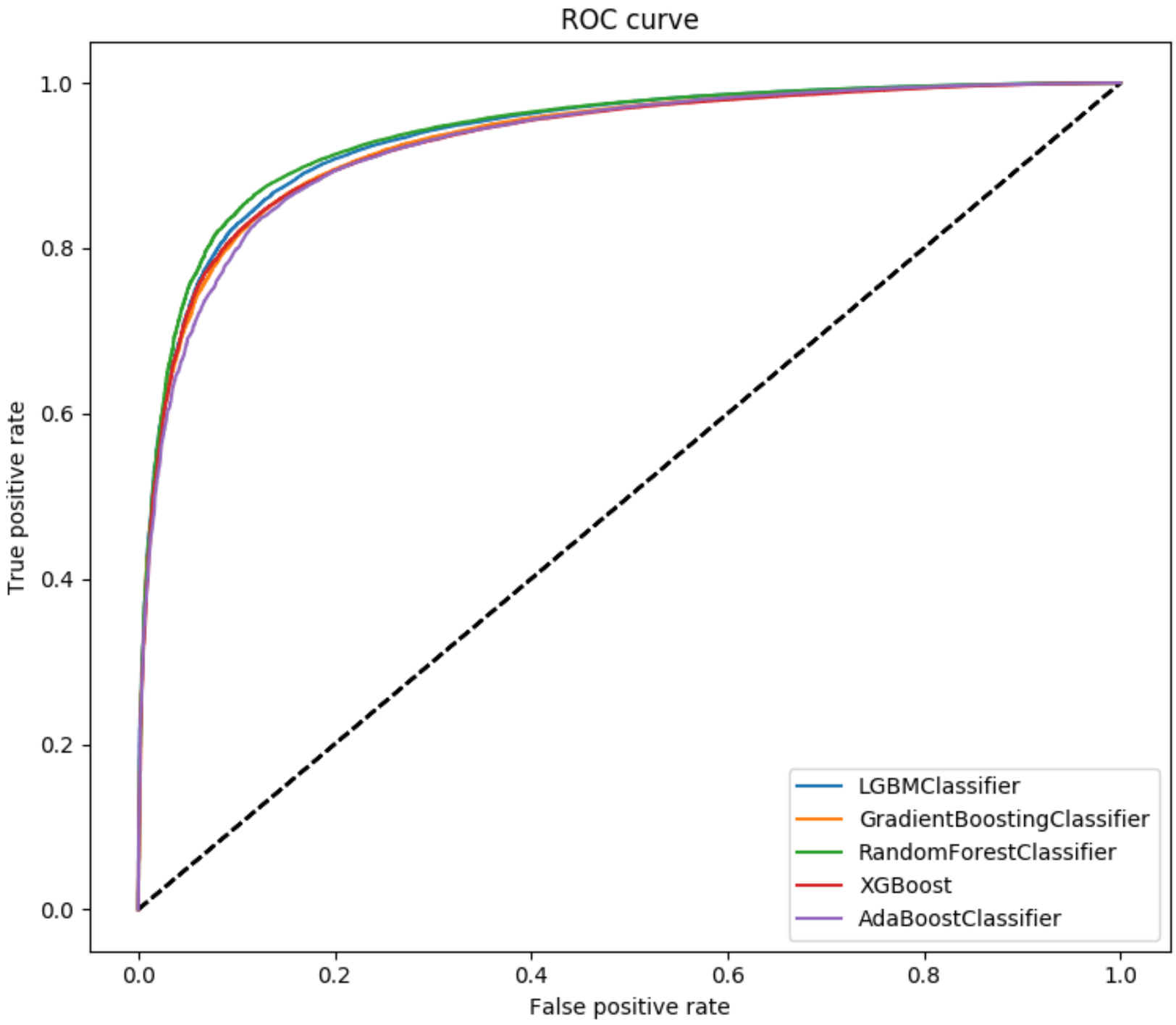
Признак, описывающий пару «step-пользователь»:

- **Время**, затраченное конкретным пользователем **на данный step.**

Список признаков, описывающих конкретного пользователя:

- **Среднее время**, затраченное пользователем на его step-ы;
- **Количество** step-ов, **проваленных** пользователем;
- **Количество** step-ов, **пройденных** пользователем.

Результаты



Classifier	AUC
LGBM	0,9355739
GB	0,9297582
RandomForest	**0,9389669**
XGBoost	0,9283769
AdaBoost	0,9264367

Регрессия

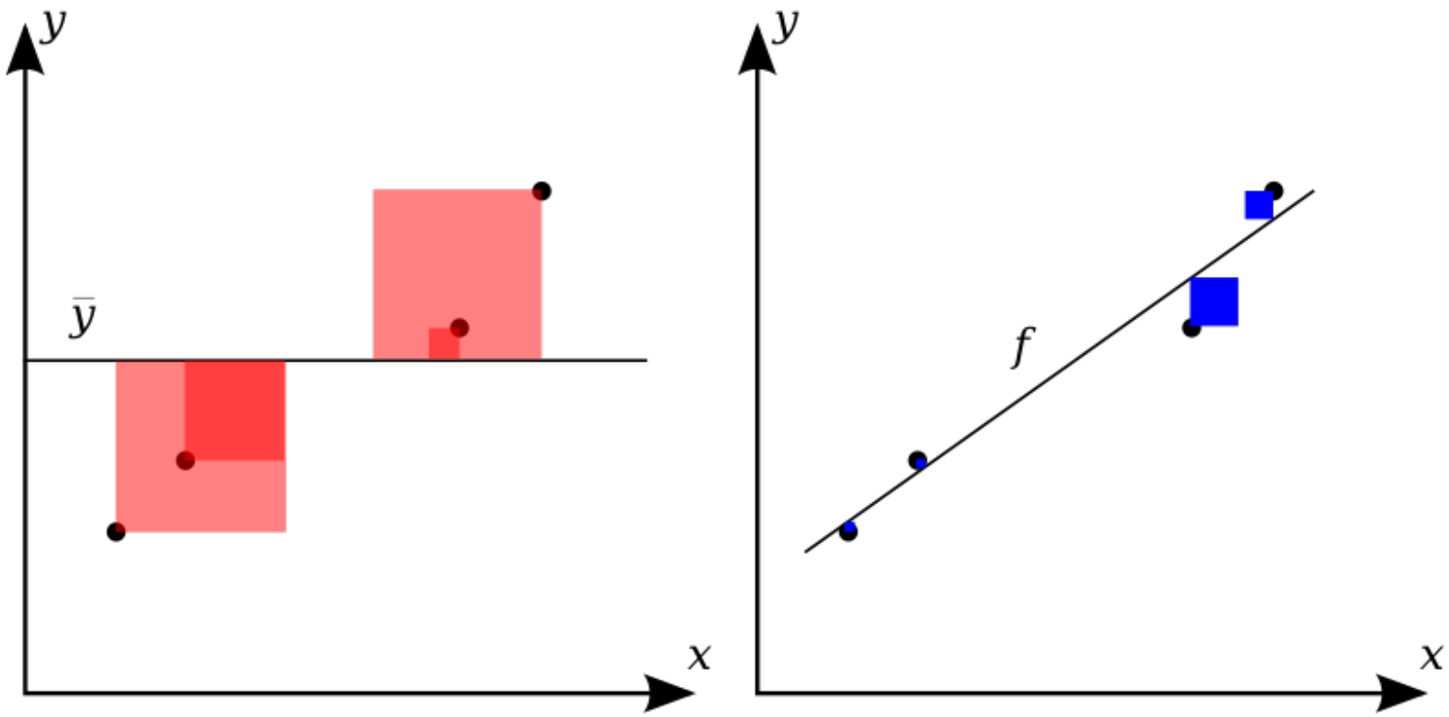
R^2-score

$$R^2 = 1 - \frac{u}{v}$$

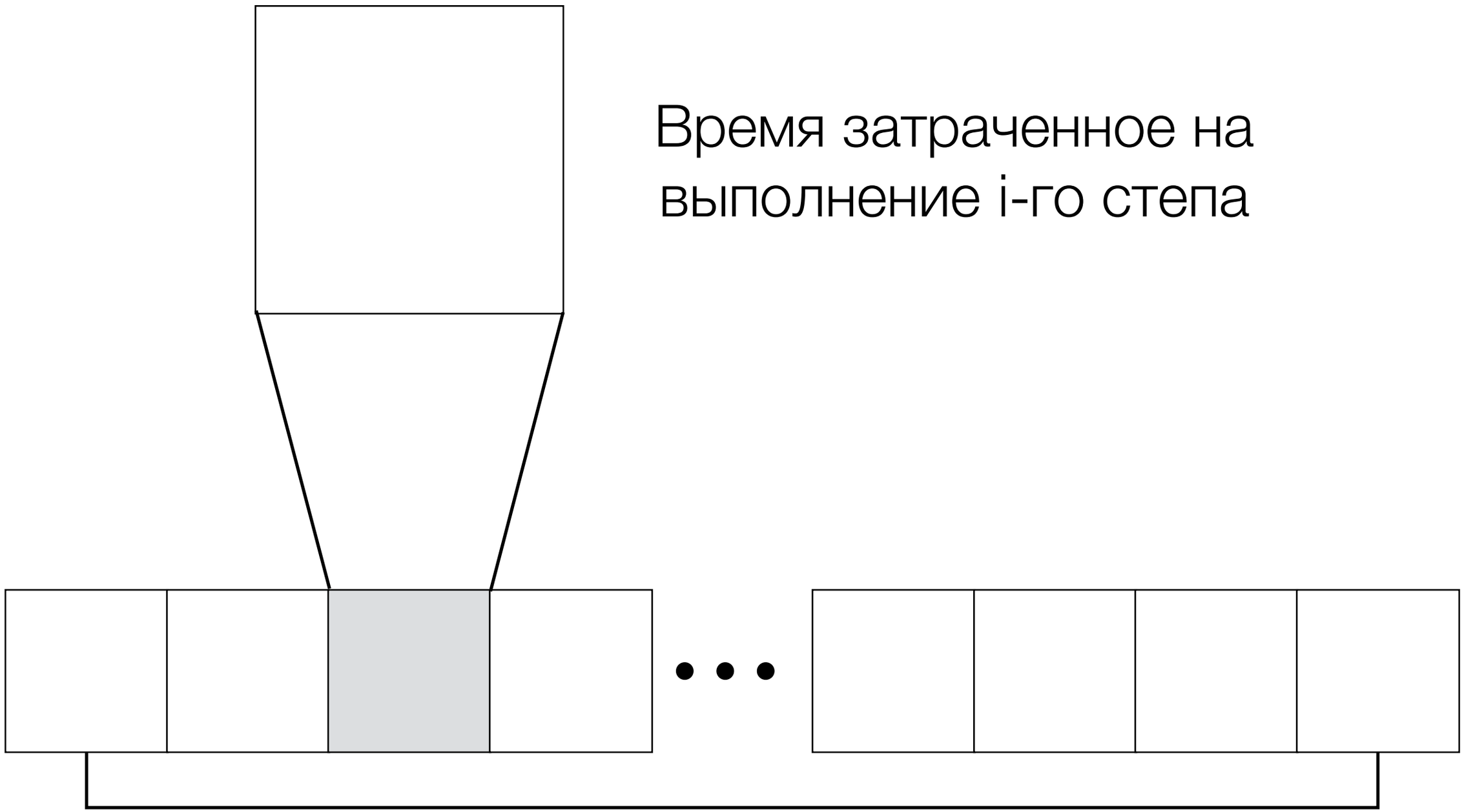
$$u = \sum (f(x_i) - y_i)^2$$

$$v = \sum (\bar{y} - y_i)^2$$

$$\bar{y} = \frac{1}{N} \sum y_i$$



Признаки

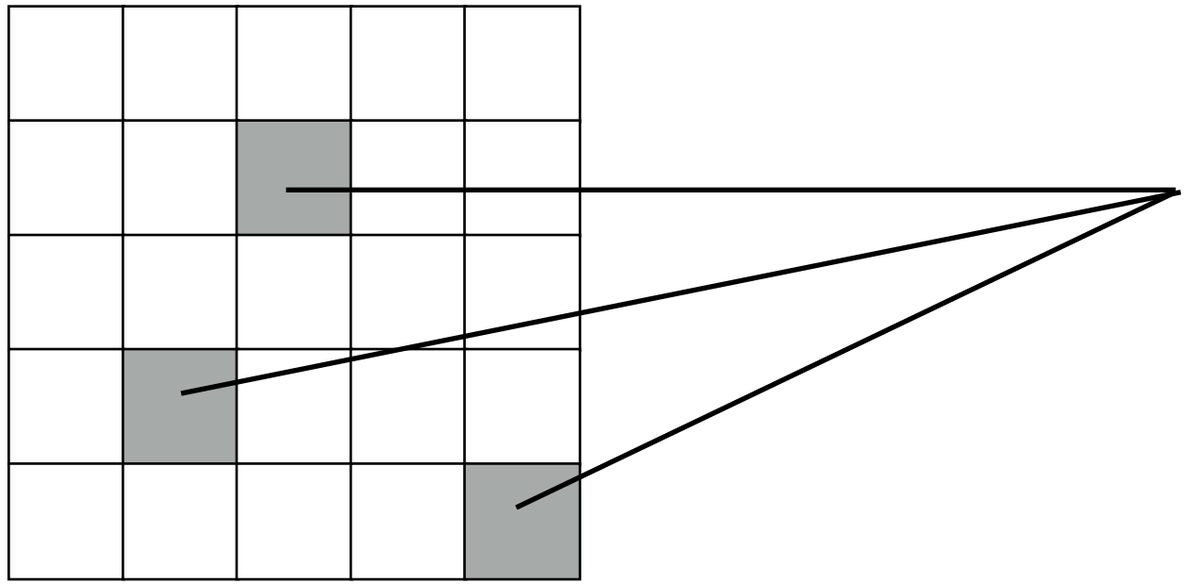


Время затраченное на выполнение i-го степа

581 признак

Проблемы

Сильно разреженная таблица

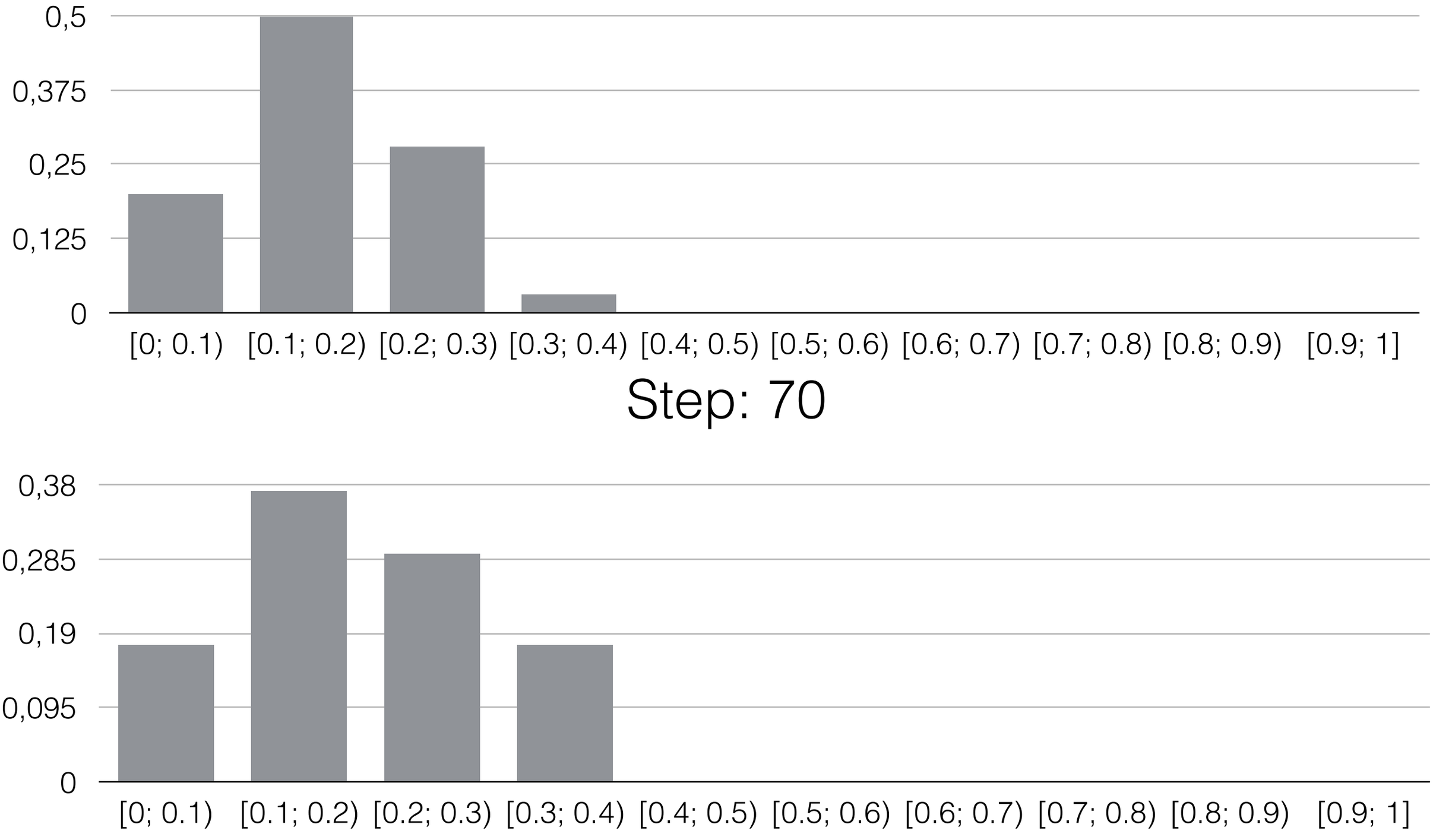


5% ненулевых значений

Выберем только те step-ы, которые принадлежат конкретному курсу.

Идентификатор курса	**Ненулевые значения(%)**
67	0.19
73	0.17
76	0.26
94	0.15
129	0.18
70	0.24

Результат



Step: 70

Step: 76

Признаки

Список признаков, описывающих конкретный step:

- **Среднее время**, затраченное пользователями на step;
- **Количество** пользователей, **прошедших** step успешно;
- **Количество** пользователей, **проваливших** step.

Список признаков, описывающих конкретного пользователя:

- **Среднее время**, затраченное пользователем на его step-ы;
- **Количество** step-ов, **проваленных** пользователем;
- **Количество** step-ов, **пройденных** пользователем.

Target:

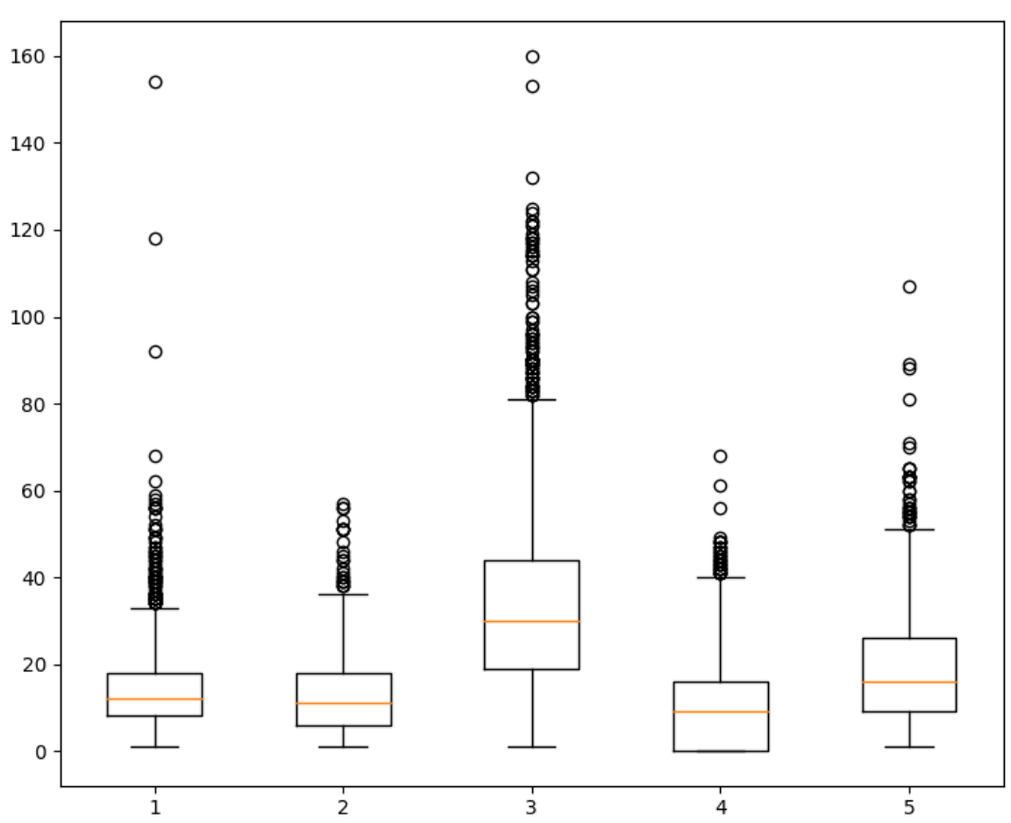
- **Время**, затраченное конкретным пользователем **на данный step**.

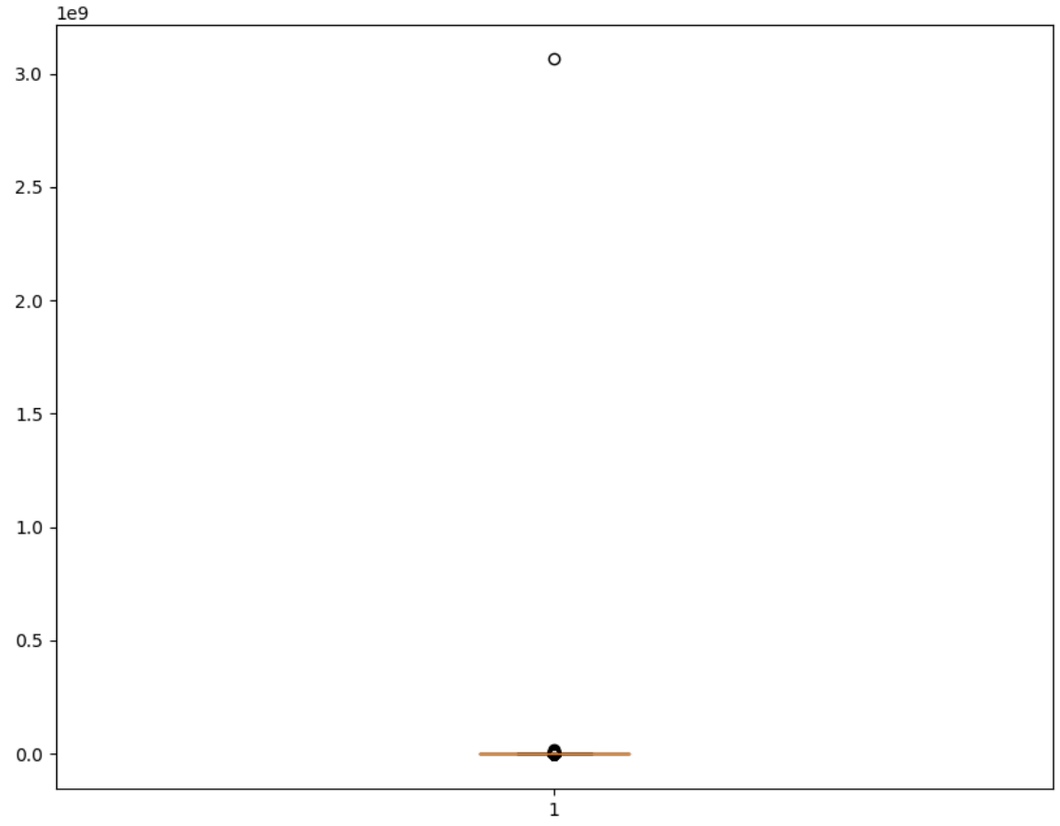
Результат

XGBoost: max_depth=6, learning_rate=0.01, n_estimators=400

Accuracy: -0.47 (+/- 6.71)

Причина: выбросы





Step: 22105. User: 3485.

Результат

XGBoost: max_depth=6, learning_rate=0.01, n_estimators=400

Accuracy: 0.66 (+/- 0.33)

Признаки

Список признаков, описывающих конкретный step:

- **Среднее время**, затраченное пользователями на step;
- **Относительное количество** пользователей, **прошедших** step;
- **Относительное количество** пользователей, **проваливших** step.

Список признаков, описывающих конкретного пользователя:

- **Среднее время**, затраченное пользователем на его step-ы;
- **Относительное количество** step-ов, **проваленных** пользователем;
- **Относительное количество** step-ов, **пройденных** пользователем.

Target:

- **Время**, затраченное конкретным пользователем **на данный step**.

Результат

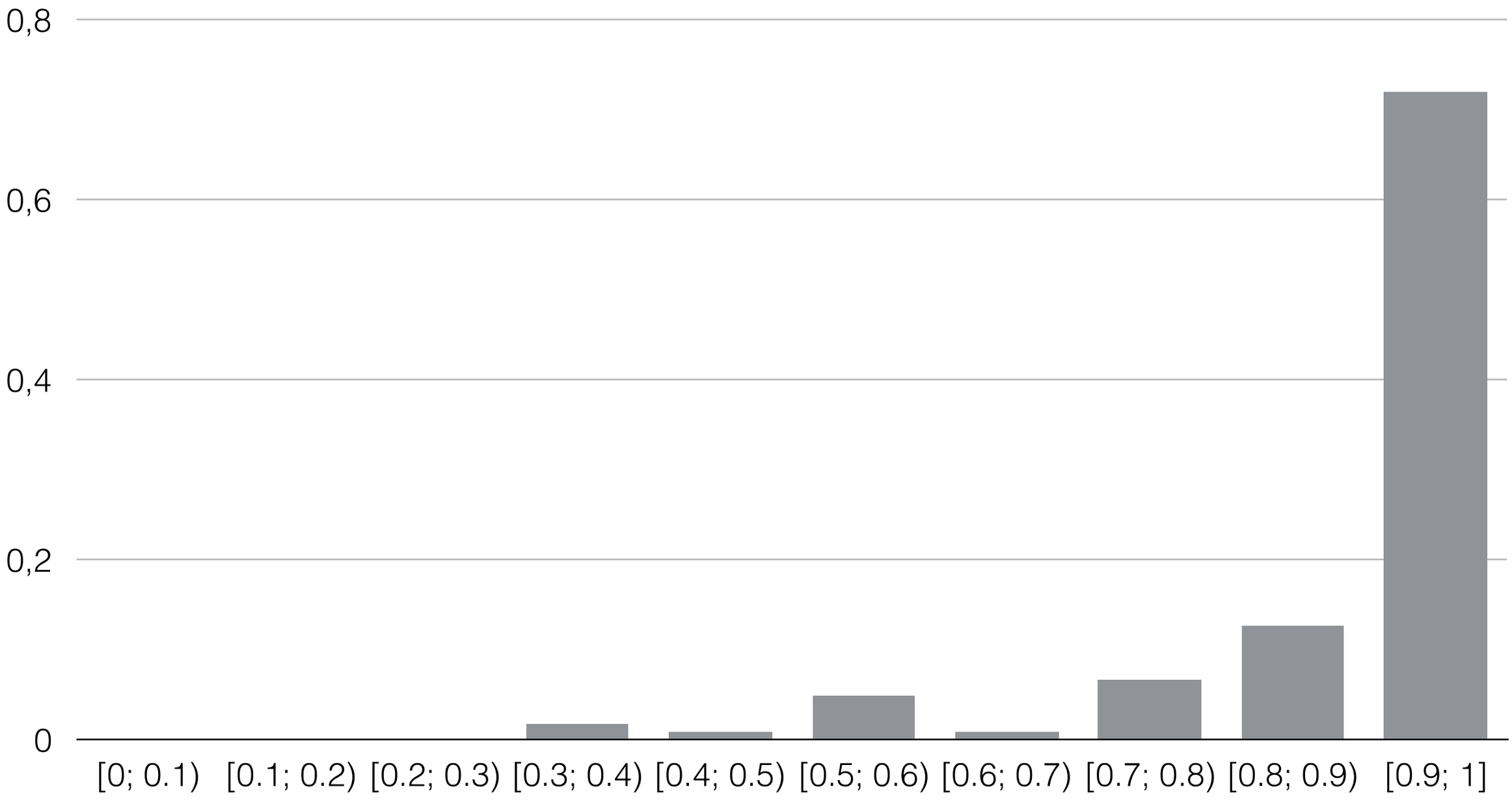
XGBoost: max_depth=6, learning_rate=0.01, n_estimators=400

Accuracy: 0.71 (+/- 0.28)

Посчитаем R метрику для каждого степа отдельно

Step ID	Accuracy
13074	0.94 (+/- 0.05)
13589	0.86 (+/- 0.11)
13129	0.92 (+/- 0.08)
13940	0.58 (+/- 0.18)

Результат



В планах

- Определить вероятность того, что пользователь бросит курс
- Предсказать оценку пользователя за курс
- Проверка на полных данных

Спасибо за внимание