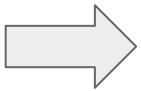
Learning for learning (Stepik)

Выполнила: Лапицкая Людмила
Научный руководитель: Шпильман А.А.

Что такое Stepik?

Stepik («Стэпик») – образовательная платформа и конструктор бесплатных открытых онлайн-курсов и уроков.

Пользователи Стэпика пытаются решать задачи (step-ы) 

Большое количество **статистических данных**

Постановка задачи

- **Мотивация:**

Научиться предсказывать «неудачи» пользователя, чтобы предлагать степы, соответствующие его уровню

- **Задача:**

Построить классификатор, который смог бы с высокой точностью определять несправляющихся пользователей

Данные

Dataset sample:

step_id	user_id	attempt_time	submission_time	status	course_id
6550	1	1409659242	1409659262	correct	67
6539	1	1409828065	1409828067	wrong	67
13940	1	1415569549	1415605943	wrong	67
13940	1	1415605956	1415605964	correct	67
13943	1	1415634165	1415638097	wrong	67

Инфографика, описывающая данные

46211

Количество пользователей

6

Количество курсов

581

Количество степов

3585451

Размер датасета

Что было использовано

- Статьи по **disengagement prediction**
  - “Predicting Player Disengagement in Online Games” (Xie et al., 2014)
  - “To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns”. (Mills et al., 2014)

- В своих работах авторы использовали **decision trees**

➔ В данной работе был использован классификатор **random forest**

Рассматриваемые признаки (features)

Список признаков, описывающих конкретный step:

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

Признак, описывающий пару «step-пользователь»:

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

Рассматриваемые признаки (features)

Список признаков, описывающих конкретного пользователя:

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

Проблемы: Дисбаланс выборки

Дисбаланс количества пользователей, которые справились со степом, относительно тех, кто справиться не смог

Ухудшение качества классификации

Решение: SMOTE

- SMOTE – Synthetic Minority Over-sampling Technique
- Были сгенерированы новые значения, дополняющие исходные данные
- По новой выборке был построен классификатор

Недостаток классификатора

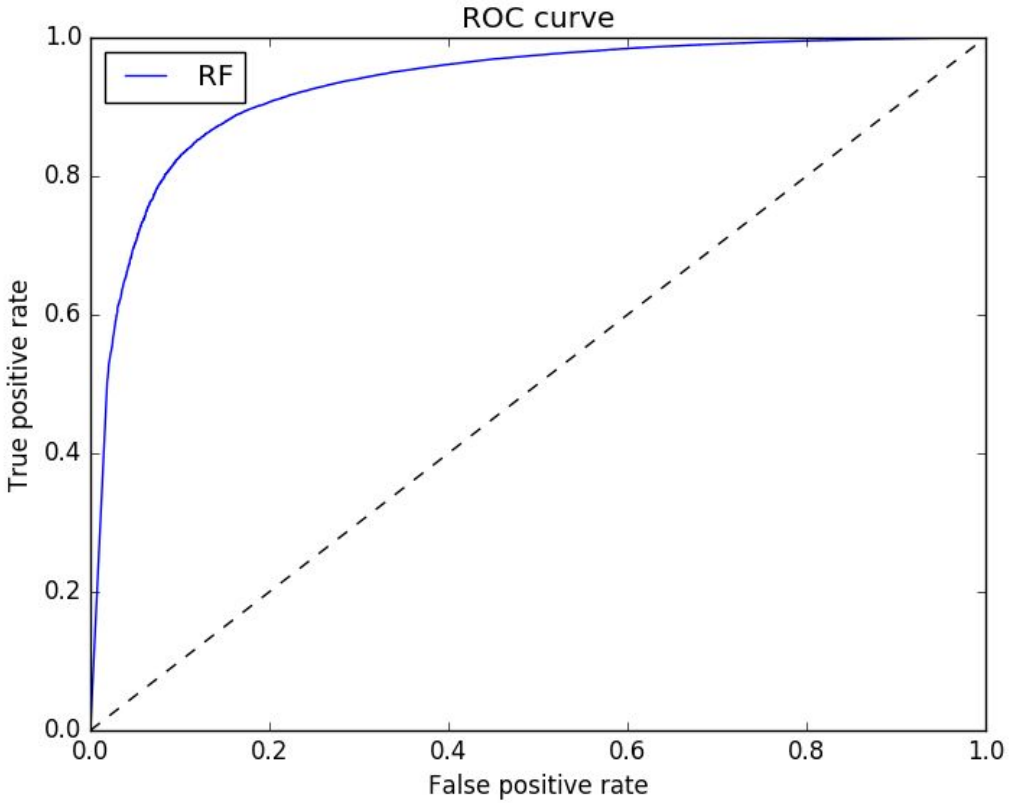
Большой показатель False Positive - плохо определяем пользователей, которые не могут пройти step

True label		
True	0.976 (True Positive)	0.023 (False Negative)
False	**0.521 (False Positive)**	0.478 (True Negative)
	True	False

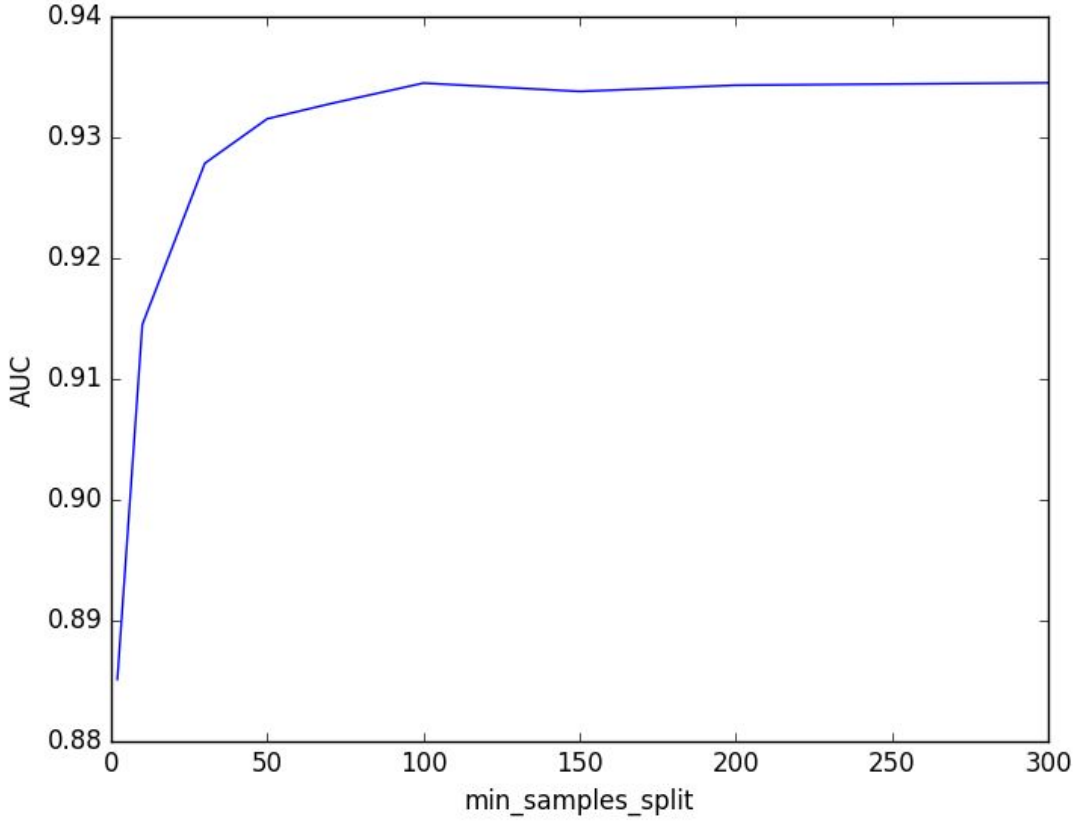
Predicted label

Решение: анализ ROC-кривой

Подбираем threshold классификатора



Подбор параметров классификатора



Результат работы

- Получилось классифицировать 92% пользователей, испытывающих трудности

True label		
True	**0.821** (True Positive)	**0.178** (False Negative)
False	**0.074** (False Positive)	**0.925** (True Negative)
	True	False

Predicted label

Что нового я узнала?

- Основы ML
- Алгоритмы классификации:
  - decision trees
  - random forest
- SMOTE и др.

Планы на будущее

В дальнейшем планируется развивать работу в этих направлениях:

- Проверка на полных данных
- Определение, насколько быстро пользователь справится со степом
- Определение вероятности того, что пользователь бросит курс
- Определение его возможной оценки

Спасибо за внимание!