



Международная конференция по новым
образовательным технологиям

**ED
CRUNCH**
TOMSK

Прогнозирование финансовых показателей онлайн-курсов на Coursera

Блескина Ирина
Ведущий аналитик

Центра психометрических исследований
в онлайн-образовании
НИУ ВШЭ



**ED
CRUNCH**
UNIVERSITY
Инициатива НИТУ «МИСиС»

skyeng

**ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ**

Современная
цифровая
образовательная
среда в РФ

Томск, 29-31 мая 2019

Проблема

1. Coursera и EdX (2012) – самые крупные мировые провайдеры онлайн-обучения
2. Основная идея: доступное образование для всех вне зависимости от социальных и географических признаков
3. Конференция Coursera в Гааге в 2016 г. – активная монетизация платформы
4. Инвестиции в Coursera – 313.1 Million \$¹

¹ По данным Crunchbase

Кейс НИУ ВШЭ

- НИУ ВШЭ – 92 онлайн-курса на Coursera
- Топ-5 на Coursera по количеству курсов
- Инвестиции в каждый курс



Выбор предикторов

Зависимая переменная:

- средний приток оплат по курсу в неделю

Объясняющие переменные:

- язык,
- курс/специализация,
- уровень (Beginner, Intermediate, Advanced),
- тематическая область,
- длительность,
- стоимость сертификата/подписки,

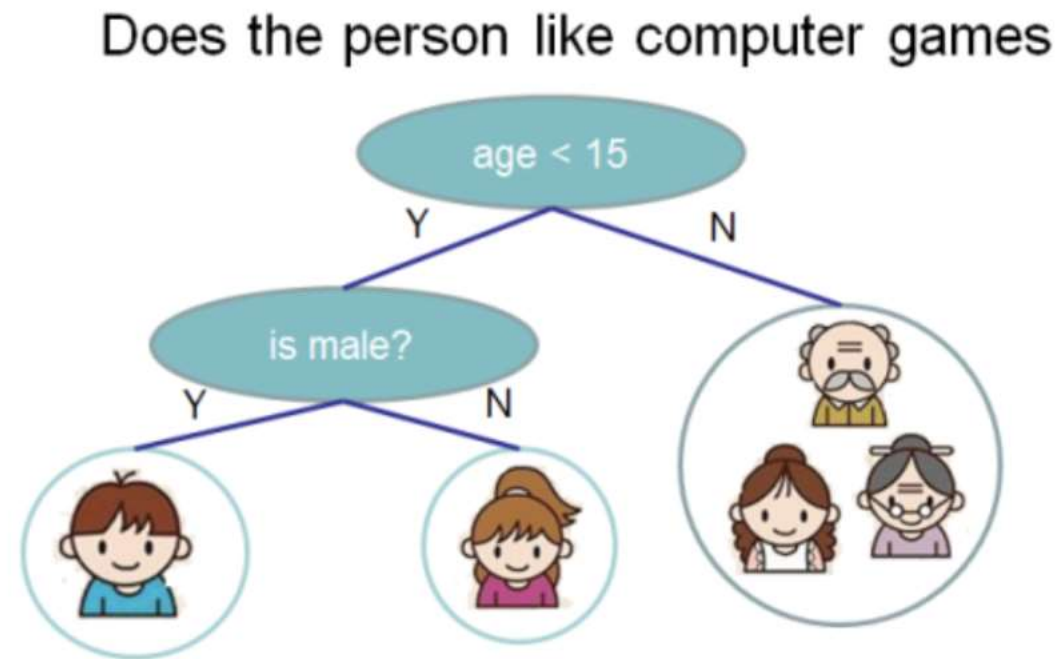
- нагрузка в неделю,
- наличие партнера,
- данные о заданиях:
 - количество заданий,
 - их формат,
 - общий процент в формуле оценивания, который выделяется на определенный тип заданий,
 - наличие итогового проекта или экзамена

Методы

- Обучение по прецедентам – конечное множество прецедентов, по каждому из которых измерены данные
- Supervised learning
- 3 метода:
 - Random forest
 - Ridge regression
 - XGBoost
- 3 метода отбора предикторов:
 - Все предикторы
 - Feature importance
 - Greedy algorithm

Методы

- Random forest и XGBoost – ансамблевые методы в машинном обучении
- Метод ансамблей – объединение некоторых базовых моделей для получения модели ансамбля с лучшими характеристиками
- Отличие:
 - Random forest – каждое дерево обучается независимо, используя случайную выборку данных
 - XGBoost – деревья строятся по одному, где каждое новое дерево помогает исправлять ошибки, допущенные ранее обученным деревом



Методы отбора предикторов

- Жадный алгоритм (greedy algorithm) – алгоритм, который на каждом этапе делает выбор, который кажется самым наилучшим в данный момент (локально оптимальный выбор) в надежде, что данный выбор приведет к оптимальному итоговому решению.
- Feature importance – метод, позволяющий отобрать наиболее значимые предикторы

Значимые предикторы в Random forest

Future importance

- Длительность
- Нагрузка
- Стоимость сертификата/подписки
- Курс/специализация
- Наличие индустриального партнера

Greedy algorithm

- Данные о заданиях – количество заданий на программирование
- Тематические области: Social sciences

Результат: алгоритм в среднем ошибается на **6** оплат

- Тематические области: Life sciences, Business, Algorithms, Machine learning
- Данные о заданиях:
 - Количество заданий на программирование и их процент в формуле оценивания
 - Количество тестов и их процент в формуле оценивания
 - Количество заданий на взаимное оценивание и их процент в формуле оценивания

Результат: алгоритм в среднем ошибается на **10** оплат

Алгоритм, построенный на всех предикторах, в среднем ошибается на **12** оплат

Значимые предикторы в Ridge regression

Greedy algorithm

- Длительность
- Курс/специализация
- Наличие индустриального партнера
- Тематические области: Arts and humanities, Business, Language learning, Life sciences, Machine learning
- Данные о заданиях: количество заданий на программирование
- **Результат:** алгоритм в среднем ошибается на **6** оплат
- Алгоритм, построенный на всех предикторах, в среднем ошибается на **37** оплат

Значимые предикторы в XGBoost

XGBoost (самостоятельно отобранные предикторы)

- Длительность
 - Язык
 - Нагрузка
 - Курс/специализация
 - Тематические области: Social sciences, Business, Math and logic, Physical science and engineering, Economics, Arts and humanities, Marketing
- Данные о заданиях:
 - Количество заданий на программирование
 - Количество тестов и их процент в формуле оценивания
 - Количество заданий на взаимное оценивание
 - Формат экзамена (тест/взаимное оценивание) и процент за экзамен

Результат: алгоритм в среднем ошибается на 6 оплат

Greedy algorithm

- Данные о заданиях – количество заданий на программирование и их процент в формуле оценивания
- Тематические области: Math and logic и Physical science and engineering

Результат: алгоритм в среднем ошибается на 4 оплаты

Сравнение алгоритмов по средней ошибке оплат

Алгоритмы	Random Forest	Ridge Regression	XGBoost
Все предикторы	12	37	6
Процедура future importance	10	-	
Greedy algorithm	6	6	4

Bagging

Bagging – ансамблевый метод в машинном обучении, заключающийся в параллельном и независимом обучении алгоритмов, а затем объединением их, следуя некоторому детерминированному процессу усреднения

Тестирование и сравнение алгоритмов

59 курсов – обучающая подвыборка

18 курсов – тестовая подвыборка

Кросс-валидация — процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам (проверка того, насколько успешно разработанный алгоритм способен работать на независимом наборе данных). В рамках данной процедуры исходная выборка разбивается на две подвыборки: обучающую и тестовую N количество раз. На обучающей подвыборке происходит обучение модели, а тестовая подвыборка используется для тестирования.

Тестирование и сравнение алгоритмов

Курсы	Реальные значения	XGBoost (Greedy algorithm)	Bagging
Курс 1	2	0	1
Курс 2	0	1	0
Курс 3	2	1	1
Курс 4	0	2	1
Курс 5	0	1	0
Курс 6	9	1	5
Курс 7	2	1	1
Курс 8	0	2	1
Курс 9	0	0	0
Курс 10	0	0	0
Курс 11	0	1	1
Курс 12	0	0	0
Курс 13	1	0	1
Курс 14	0	0	0
Курс 15	16	4	12
Курс 16	9	3	5
Курс 17	64	83	65
Курс 18	4	11	9

Эксперимент на примере курса X

1. Заданы предикторы курса X, который недавно был запущен на платформе
2. Разработанный алгоритм спрогнозировал средний приток оплат равный 31
3. Статистика на платформе показала количество оплат, поступившее за последнюю неделю равное 38

Выводы

1. Задача решена успешно – количество оплат за сертификаты курсов на Coursera можно спрогнозировать, используя разработанный алгоритм
2. Ограничения: с появлением новых курсов алгоритм необходимо строить снова
3. На текущий момент мы в поисках оптимальных предикторов и оптимизации разработанного решения

Международная конференция по новым
образовательным технологиям

**ED
CRUNCH
TOMSK**

Спасибо за внимание!

Блескина Ирина

E-mail: ibleskina@hse.ru



Национальный
исследовательский
Томский
государственный
университет

**ED
CRUNCH**
UNIVERSITY
Инициатива НИТУ «МИСиС»

skyeng



ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ



Современная
цифровая
образовательная
среда в РФ

Томск, 29-31 мая 2019