УДК 004

Таратухина Ю.В.[1]
Блескина И.А.[2]

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ПРИТОКА ОПЛАТ НА МАССОВЫХ ОТКРЫТЫХ ОНЛАЙН-КУРСАХ (МООК) НА ПЛАТФОРМЕ COURSERA

Аннотация. Данная работа посвящена анализу МООК на примере МООК НИУ ВШЭ на платформе Coursera. Предполагается, что поведение пользователя на МООК зависит от ряда факторов, таких как название курса, язык, тематическая область и др. Таким образом, в данной статье предпринимается попытка проанализировать МООК и спрогнозировать прибыльность того или иного онлайн-курса с определенными заданными параметрами, что позволит заранее определить параметры для создания МООК, востребованного аудиторией.

Ключевые слова: МООК, анализ оплат на МООК, прогнозирование поведения функций оплат на МООК, прогнозирование прибыльности МООК

Taratukhina J.
Bleskina I.

ANALYSIS AND FORECASTING THE FLOWS OF PAYMENTS ON THE MASSIVE OPEN ONLINE CURSES (MOOC): COURSERA'S CASE

Abstract. This paper is devoted to the analysis of the MOOCs by the example of the NRU HSE MOOCs on Coursera. It is assumed that the user's behavior at the MOOCs depends on a number of factors, such as the name of the course, language, theme, etc. Thus, in this article we attempt to analyze the MOOCs and predict the profitability of an online course with certain parameters that allow to predetermine and configure the parameters for creating a MOOCs demanded by the audience.

Keywords: MOOC, analysis of payments on MOOCs, forecasting the functions of payments on MOOCs, forecasting profitability of MOOCs

Использование информационных технологий в процессе обучения студентов порождает взаимодействие нового типа: традиционный формат обучения (преподаватель в аудитории) становится необязательным условием образовательного процесса, поскольку новые технологии предлагают обучение посредством интерактивных ресурсов, в частности, массовых открытых онлайн-курсов (МООК).

Термин **массовые открытые онлайн-курсы** (МООК) ещё в 2008 ввел исследователь Дейв Кормье. Данный термин был использован для описания курса *Connectivism and Connective Knowledge* [1, 2]. Курс был открыт для регистрации для всех, заинтересованных в участии. Обучающимся было предложено установить свои индивидуальные цели обучения, и была предоставлена возможность использо-

[1] *Таратухина Юлия Валерьевна*, кандидат филологических наук, доцент кафедры инноваций и бизнеса в сфере информационных технологий, Национальный исследовательский университет «Высшая школа экономики», г. Москва, e-mail: jtaratuhina@hse.ru
Julia Taratukhina, Ph.D. in Philology, Associate Professor, Department of information and Communication Technologies, National Research University Higher School of Economics, Moscow, e-mail: jtaratuhina@hse.ru
[2] *Блескина Ирина Александровна*, выпускник магистерской программы «Электронный бизнес», начальник отдела координации образовательных проектов, Дирекция по онлайн обучению, Национальный исследовательский университет «Высшая школа экономики», г. Москва, e-mail: iableskina@mail.ru
Irina Bleskina, MSc in Business Informatics, Head of Education Projects Coordination Unit, eLearning Office, National Research University Higher School of Economics, Moscow, e-mail: iableskina@mail.ru

вать разные инструменты (социальные сети и т.п.) и принимать решение об уровне участия в различных активностях [3, 4]. Дальнейшее развитие МООК в университетах обусловлено массовым спросом на образовательные услуги и реализацией тенденции «образование в течение всей жизни» (*education through life*), а также возможностью, формировать свои индивидуальные учебные планы [5].

1. **Отличительные особенности МООК.** Исследователи предлагают различные определения МООК. Согласно одному из них, «МООК – это ставшая популярной тенденция в области онлайн-обучения, относящаяся к открытым образовательным ресурсам (*Open Educational Resources* (OER)) и технологиям дистанционного обучения» [6].

Отличительными особенностями МООК являются: открытость, интерактивность, мультимедийность, в большинстве случаев бесплатное обучение и возможность «учиться» у ведущих мировых преподавателей и экспертов.

На сегодняшний день существует большое число разных платформ-провайдеров МООК таких, как Coursera, Edx, Udacity, Национальная платформа открытого образования и другие. Самой крупной платформой является Coursera, аудитория которой составляет 37 млн. пользователей по всему миру [7, 8].

Одним из самых активных участников данного проекта является Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ). ВШЭ входит в Топ-10 по числу разработанных и размещенных МООК на Coursera и охватывает аудиторию около 2 млн. слушателей. Таким образом, задачей настоящей работы является анализ МООК и попытка спрогнозировать поведение слушателей на примере анализа данных по МООК НИУ ВШЭ. Мы предполагаем, что существует определенная закономерность выбора МООК разными пользователями, которую можно представить в виде алгоритма. С помощью данного алгоритма можно прогнозировать поведение слушателей, а также некоторые показатели МООК, например, количество оплат за сертификаты. Данные алгоритмы позволят определить параметры, которые оказывают значимое влияние на выбор слушателем курса и, соответственно, определить потребности аудитории и разрабатывать МООК с учетом данных потребностей.

2. Предварительный анализ данных и выбор релевантных решений.

Для проверки гипотезы необходимо провести анализ данных по МООК на Coursera. В качестве базы исследования была выбрана база данных МООК НИУ ВШЭ на Coursera, которая представляет собой «*сгенерированные массивы подмножеств данных*». Экспортированная база состоит из набора логов, содержащих информацию о контенте МООК, описание взаимодействия слушателей с контентом, детализацию оценок слушателей, информацию о форумах, основную информацию об аудитории, демографическую информацию и т.д.

В данном исследовании наибольший интерес представляют данные о МООК (количество слушателей; количество оплат; язык МООК, количество заданий и др.). Также предполагается разбиение курсов на некоторые кластеры: принадлежность курса к специализации (**Специализация** – это связка курсов и практико-ориентированный итоговый проект, основанный на знаниях, которые слушатели получают на всех курсах специализации и доступный только при условии успешного освоения всех курсов специализации.); *специализированные* – МООК, которые предназначены для формирования определённых компетенций, направленных на повышение квалификации или переквалификации или на приобретение знаний в той или иной области; и *общеобразовательные* – МООК, направленные на приобретение общих представлений об определенной тематической области и саморазвитие индивида; МООК, направленные на приобретение слушателем *hard skills* и *soft skills*. (*Soft skills* предполагают развитие социальных навыков индивида, таких как навыки коммуникации, навыки ведения переговоров, принятия решений и других, и способствуют поддержанию актуальности *Hard Skills* в изменяющихся обстоятельствах. *Hard Skills* позволяют индивиду приобретать профессиональные навыки, например, необходимые чтобы

стать инженером, физиком, философом. *Soft skills* развивают индивида в направлении, которое не зависит от роли индивида и выходит за рамки жестких требований профессии [9]. Различие между данными навыками также может быть обусловлено с точки зрения личных и социальных навыков. Личные навыки в основном соответствуют когнитивным навыкам, таким как знания и навыки мышления, в то время как социальные навыки относятся к отношениям с другими людьми [10].)

Ограничение выборки обусловлено кейсом, который был использован для проведения исследования: количеством МООК НИУ ВШЭ на платформе Coursera – 77 курсов (на момент проведения исследования).

Основной задачей анализа является исследование различий между кластерами и группами индивидов, оплачивающими МООК каждого кластера, характеристика данной аудитории, а также прогнозирование количества оплат по курсам.

3. Анализ данных, описание моделей и экспериментов. Модели строятся на основании имеющихся данных и заданных таргетов, однако первый этап требует проведения предварительного анализа данных. Этот этап включает в себя ряд тестов на проверку качества данных, а также позволяет установить некоторые взаимосвязи между переменными. Далее рассмотрим некоторые описательные статистики, которые позволяют выявить характер взаимосвязи между зависимой переменной (количество оплат) и независимыми (признаки МООК).

Так, были проанализированы данные по оплатам по всем МООК и построены графики оплат, с помощью которых было выявлено, что график оплат каждого курса ведет себя как временной ряд с явно выделенным трендом. На примере курсов A, B, C и D (См. Рис. 1) можно заметить, что данную функцию можно аппроксимировать линейной функцией и какой-либо из моделей временного ряда, например, ARIMA.

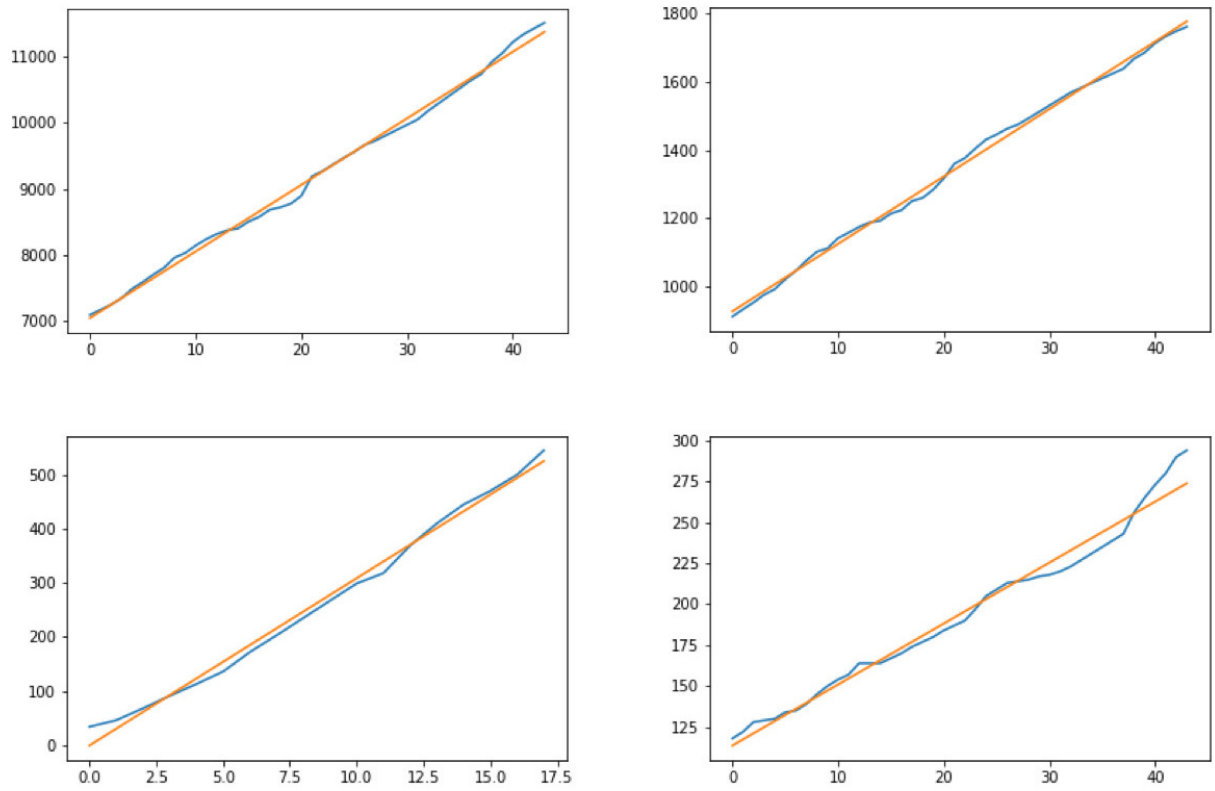


Рисунок 1 – Поведение функций оплат курсов A, B, C и D во времени

Цель настоящего исследования – анализ поведения функций оплат МООК в зависимости от их параметров и прогнозирование финансовых показателей курсов. В этой связи необходимо построить модель временного ряда с линейной регрессией. Формула 1, представленная ниже, будет аппроксимировать функцию оплат от времени:

Формула 1

$$y_t = k * t + const + \in_t = f(x) * t + const + \in_t,$$

y –количество оплат за курс,

t – момент времени,

k – средний приток оплат в неделю,

const – свободный член, который не равен 0 в случаях, когда МООК заранее был открыт на pre-enrollment (*Pre-enrollment* – публикация целевой страницы курса до запуска, чтобы учащиеся могли найти курс и зарегистрироваться. Предварительная регистрация является хорошим инструментом продвижения курса до его запуска на платформе [19].), и до старта его оплатило некоторое количество слушателей,

f(x) – функция прогнозирования,

$\in_t$ – шумы, которые можно аппроксимировать через модель временного ряда (ARIMA) (см. Формула 2):

Формула 2

$$\in_t = c + \sum_{i=1}^{p} a_i * \in_{t-i} + \sum_{J=1}^{q} b_j * \theta_{t-j} + \theta_t,$$

По результатам анализа функции оплат было принято решение о том, что прогнозировать необходимо не количество оплат, а средний приток оплат в неделю (*k)*. Таким образом, при заданном *t* и известном *k* возможно вычислить среднее количество оплат по курсу в заданный момент времени.

Данная модель будет следующей:

– на входе подаются заданные параметры, такие как язык МООК, принадлежность к специализации, относится курс к общим или специфичным областям знаний, уровень, тематическая область, направленность на развитие «hard/soft skills», длительность, стоимость сертификата или подписки, данные по заданиям (количество заданий, их формат, общий процент в формуле оценивания, который выделяется на определенный тип заданий, наличие итогового проекта или экзамена), средняя нагрузка в неделю, количество преподавателей, наличие индустриальных партнеров и т.п.

– на выходе: угол наклона функции (или число оплат в неделю).

На первом этапе строится модель линейной регрессии для каждого МООК, на выходе которой выдаются значения *k* для всех МООК, входящих в выборку. Затем на основе полученной оценки переменной *k* конструируется так называемая ***предсказательная модель***, в которой данная переменная используется как трагет. Для создания данной модели необходимо исключить данные по оплатам, и попытаться построить модель для планируемых к разработке МООК, которая будет предсказывать *k*, поскольку, зная *k* можно примерно оценить успешность (прибыльность) нового курса по заданным признакам.

На этапе анализа данных и построения модели прогнозирования оплат по МООК было решено преобразовать таргеты в натуральные логарифмы (*ln*), поскольку функция оплат является положительной неубывающей функцией и, соответственно, *k* также не может принимать отрицательные значения. Таким образом, была введена переменная *K*, которой было присвоено значение *ln(k)*.

При проверке взаимосвязи между переменными, была преобразована выборка и построена корреляционная матрица, которая позволила сделать вывод о том, что существует некоторая линейная связь между зависимой переменной и объясняющими переменными.

На следующем этапе строится ***модель случайного леса*** (*Random Forest Regressor*) и проверяется ее качество с помощью процедуры кросс-валидации. (**Кросс-валидация** – процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам. В рамках данной процедуры фиксируется некоторое множество разбиений исходной выборки на обучающую и тестовую подвыборки. Для каждого разбиения алгоритм настраивается по обучающей подвыборке, далее происходит оценка средней ошибка на тестовой подвыборке [11].) При построении модели средняя квадратичная ошибка (MSE)

по *K* равна 1.55 – в среднем модель с таргетом *K* ошибается на 1.55 логарифма оплат.

Далее можно рассмотреть значимость признаков для данной модели с помощью процедуры *future importance*. По результатам анализа после применения процедуры *future importance* были выбраны следующие наиболее значимые признаки: частота запуска когорт, количество заданий, создание МООК при участии индустриального партнера, отдельный МООК / МООК специализации, длительность, количество преподавателей, средняя нагрузка в неделю, стоимость сертификата/подписки, область знаний.

Далее необходимо построить модель, учитывая только те признаки, которые были отобраны процедурой *future importance*. При построении данной модели MSE стало равно 1.49, соответственно, можно заметить, что модель стала лучше по метрике MSE. Далее необходимо рассчитать стандартное отклонение и представить его в более интерпретируемом виде (*k*) – стандартное отклонение для модели по среднему количеству оплат в неделю. Для этого была введена метрика **стандартное отклонение** (Std. Deviation *k*), которое после использования *future importance* стало равным 10.37: в среднем модель ошибается на 10 оплат.

Для построения данной модели также можно использовать ***метод жадного алгоритма*** (*Greedy algorithm*), который заключается «в выборе локально оптимальный решений на каждой итерации и предполагает оптимальность итогового решения» [12]. При построении модели стандартное отклонение получилось равным 6.42, что указывает на худший результат по сравнению с предыдущим методом.

На следующем этапе необходимо обратиться к другим моделям. В качестве первой модели была выбрана регрессионная модель (*Ridge Regression*). MSE *K* равен 8.7, что показывает худший результат по сравнению с моделью случайного леса: на кросс-валидации в пространстве ln MSE равен 8.7 по сравнению с 1.55 и 1.49 для случайного леса. Далее попробуем улучшить модель регрессии, применив процедуру жадного добавления выбора признаков. После применения данного алгоритма MSE на кросс-валидации в пространстве ln на отобранных методом «жадного алгоритма» признаках стал равен 1.8, соответственно, стандартное отклонение равно 6.57 – модель в среднем ошибается на 6–7 человек, что показывает худший результат по сравнению с моделью случайного леса.

Следующей моделью была выбрана ***модель градиентного бустинга*** (*XGBoost*). На первом этапе применяется ***метод жадного алгоритма*** для отбора признаков и построения модели. После использования данного алгоритма MSE на кросс-валидации в пространстве ln стал равен 1.65, соответственно, стандартное отклонение равно 4.6 – модель в среднем ошибается на 4-5 оплат, что показывает лучший результат по сравнению с другими моделями. Однако поскольку в ***модель градиентного бустинга*** встроена процедура *future importance*, на следующем этапе необходимо построить модель градиентного бустинга и проанализировать метрики. В данном случае MSE в пространстве ln равен 1.58, соответственно, стандартное отклонение равно 6.79 – модель в среднем ошибается на 6-7 человек, что показывает худший результат по сравнению с предыдущей моделью (см. Таблица 1).

Таблица 1. Значения ключевых метрик построенных моделей

Модели	**Случайный лес (Random Forest)**		**Регрессия (Ridge Regression)**		**Градиентный бустинг (XGBoost)**	
Метрики (среднеквадратичная ошибка и стандартное отклонение) на кросс-валидации	MSE ln(k)	Std. Deviation k	MSE ln(k)	Std. Deviation k	MSE ln(k)	Std. Deviation k
Все признаки	1,55	12,68	8,7	37,41	1,58	6,79
Процедура *future importance*	1,49	10,37	–	–		
Метод жадного алгоритма (*Greedy algorithm*)	1,52	6,42	1,8	6,57	1,65	4,6

После построения всех моделей и анализа метрик на кросс-валидации было принято решение о целесообразности использования не лучшей из моделей по метрике стандартного отклонения *k*, а прибегнуть к процедуре *Bagging*, которая предполагает усреднение результатов всех построенных моделей (в данном случае используется не усреднение, а присвоение веса каждой модели). Проведем тестирование полученной модели: часть МООК входят в обучающую выборку, остальные остаются в тестовой. Лучшая из полученных моделей – модель градиентного бустинга на тестовой выборке демонстрирует следующие результаты:

[0, 1, 1, 2, 1, 1, 1, 2, 0, 0, 1, 0, 0, 0, 4, 3, 83, 11];

после применения процедуры *Bagging* результаты получаются следующими:

[1, 0, 1, 1, 0, 5, 1, 1, 0, 0, 1, 0, 1, 0, 12, 5, 65, 9];

в то время как исходные значения *k* равны:

[2, 0, 2, 0, 0, 9, 2, 0, 0, 0, 0, 0, 1, 0, 16, 9, 64, 4],

соответственно, если сопоставить результаты, полученные в ходе эксперимента, можно сделать вывод о том, что наилучшим решением поставленной в исследовании задачи является использование процедуры *Bagging* (см. формула (3).

Формула 3

$$prediction = w_1 M_1 + w_2 M_2 + w_3 M_3 + w_4 M_4 + w_5 M_5 + w_6 M_6 + w_7 M_7$$

где

prediction – это $f(x_1, x_2, \ldots, x_n)$ – функция прогнозирования, зависящая от признаков,

$W_{(1,2,\ldots,7)}$ – веса, с которыми модели включаются в Bagging, которые в сумме образуют 1,

M_1 – модель случайного леса, построенная на всех признаках,

M_2 – модель случайного леса, построенная на отобранных признаках,

M_3 – модель линейной регрессии, построенная на всех признаках,

M_4 – модель линейной регрессии, построенная на отобранных признаках,

M_5 – модель линейной регрессии, построенная методом жадного добавления признаков,

M_6 – модель градиентного бустинга,

M_7 – модель градиентного бустинга, построенная методом жадного добавления признаков,

Для процедуры *Bagging* было проведено тестирование: всем построенным моделям присваивались разные веса, и было выбрано распределением, при котором получается результат, наиболее близкий к исходным значениям, как в продемонстрированном выше примере. Таким образом, было выбрано следующее распределение весов (см. Формула 4):

Формула 4

$$prediction = 0{,}1 * M_1 + 0{,}15 * M_2 + 0{,}15 * M_3 + 0{,}18 * M_4 + 0{,}1 * M_5 + 0{,}12 * M_6 + 0{,}2 * M_7$$

Значимость вышеуказанной модели в практическом плане заключается в возможности прогнозировать такой показатель, как прибыльность МООК. Таким образом, разработчики курсов на основе данной модели смогут дать прогноз об успешности курса и его востребованности аудиторией, а также заранее определить необходимые параметры МООК.

Заключение. Таким образом, в ходе настоящего исследования подтвердилось предположение о том, что поведение функции оплат МООК на Coursera может быть прогнозируемо по определенному разработанному алгоритму. Анализу подверглись МООК на платформе Coursera (на примере МООК НИУ ВШЭ). В результате была построена модель прогнозирования финансовых показателей МООК – прибыльность того или иного онлайн-курса с определенными заданными параметрами. Данная модель позволяет не только прогнозировать прибыльность МООК, но также определить параметры, которые имеют сильное влияние на рост функции оплат, что позволит заранее определить МООК с конкретными параметрами, который будет востребован ау-

диторией, а также определить параметры для создания курса.

Также в будущем появится возможность прогнозировать наиболее полезные для пользователя МООКи с точки зрения особенностей его личного и профессионального развития (выраженного в виде динамической компетентностной модели), а также трендов и запросов рынка труда.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1) Downes S. Places to go: Connectivism and connective knowledge. Innovate // Journal of Online Education. 2008. № 5.
2) The MOOC model for digital practice / A. McAuley, B. Stewart, G. Siemens, D. Cormier // University of Prince Edward Island. 2010.
3) Kaplan A.M. Collaborative projects (social media application): About Wikipedia, the free encyclopedia / A. M. Kaplan, M. Haenlein // Business Horizons. 2014. № 57(5). pp. 617–626.
4) Siemens G. Connectivism: A Learning Theory for the Digital Age // International Journal of Instructional Technology and Distance Learning. 2005. № 2 (1).
5) Технологии электронного обучения: учеб. пособие / Ю.В. Морозова, А.В. Гураков, В.В. Кручинин, Д.С. Шульц. Томск: ТУСУР. 2016. 68 с.
6) Learning Technologies and Globalization: Pedagogical Frameworks and Applications / N. Dabbagh, A.D. Benson, R. Denham, J. Maha, A. Ghania, H. Zgheib, Z. Fake // Springer. 2015. 40 p.
7) Coursera [Электронный ресурс] // Coursera. URL: http://www.coursera.org. (Дата обращения: 12.11.2017 г.)
8) Stone J. Awarding College Credit for MOOCs: The Role of the American Council on Education // Education Policy Analysis Archives. 2016. № 24 (38).
9) Grisi C.G.A. Soft Skills: a close link between enterprises and ethics. Speech at “Soft Skills and their role in employability // New perspectives in teaching, assessment and certification”, workshop in Bertinoro, Italy. 2014.
10) Engelberg S. A developmental Perspective on Soft Skills. Speech at “Soft Skills and their role in employability // New perspectives in teaching, assessment and certification”, workshop in Bertinoro, Italy. 2015.
11) Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // *14th international joint conference* on *Artificial intelligence*. 1995. pp. *1137*–1143.
12) Algorithmic toolbox [Электронный ресурс] // Coursera. URL: https://www.coursera.org/learn/algorithmic-toolbox/lecture/diKe3/review-of-greedy-algorithms (Дата обращения: 19.01.2018 г.)