Костенко Віталій Леонідович – доктор технічних наук, професор, Одеський національний політехнічний університет, Кафедра металорізальних верстатів, метрології та сертифікації, пр. Шевченка, 1, м. Одеса, Україна, 65044; e-mail: kvl777@ukr.net.

Kostenko Vitaliy – professor, Odessa National Polytechnic University, Department of Metal-cutting machines Metrology and Certification, Shevchenko 1, Odessa, Ukraine, 65044; tel.: 063-169-62-49; e-mail: kvl777@ukr.net.

Поперека Екатерина Дмитриевна – аспирант, Одесский национальный политехнический университет, Кафедра металлорежущих станков, метрологии и сертификации, пр. Шевченко, 1, г. Одесса, Украина, 65044;

Поперека Катерина Дмитрівна – аспірант, Одеський національний політехнічний університет, Кафедра металорізальних верстатів, метрології та сертифікації, пр. Шевченка, 1, м. Одеса, Україна, 65044;

Popereka Kateryna – postgraduate, Odessa National Polytechnic University, Department of Metal-cutting machines Metrology and Certification, Shevchenko 1, Odessa, Ukraine, 65044; e-mail: popereka2013.prof@mail.ru.

УДК 004.021

А. Ю. САВЧЕНКОВА

ПРОГНОЗ ЭФФЕКТИВНОСТИ КОЭФФИЦИЕНТА КОНВЕРСИИ НА ОСНОВЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

В статье рассмотрено оптимальное хранение прошлых данных. Рассмотрены алгоритмы для лучшей конверсии предложены в будущем более точные результаты вероятностей той или иной конкретной рекламы. Рассмотрены сочетания оценок CTR с помощью логистической регрессии. Приведены основные сведения про CTR оптимизацию. Дается описание иерархической модели данных. В иерархической модели автоматически поддерживается целостность ссылок между предками и потомками. Основное правило: никакой потомок не может существовать без своего родителя. Также рассматриваем расчёт вероятности с помощью логистической регрессии. С помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических переменных от независимых переменных, имеющих любой вид шкалы.

Ключевые слова: логистическая регрессия, RTB, CTR, деревья данных, коэффициент конверсии, CPC, рекламные сети, иерархии данных, DSP, CPA.

Введение. В современном мире, где существует интернет, рекламодатели пытаются продать свои продукты публикуя свою рекламу в виде графического объявления на различных веб-страницах, пользующихся популярностью среди потенциальных потребителей, например, на страницах новостных порталов. Основной целью рекламодателя является достижение наиболее подходящую аудиторию в данной тематике, которая будет взаимодействовать с отображаемыми объявлениями, это и называется контекстной рекламой. Контекстная реклама - Реализация этой цели является достаточно сложной, в следствии чего рекламодатели должны использовать такое технологические решение, как DSP. Demand Side Platform (DSP, автоматизированная система покупки) – технологическая система организации аукциона для рекламодателей, которая торгуется с SSP (платформами для RTB-торгов со стороны площадок), управляет несколькими рекламными сетями (Ad Networks) и рекламными биржами (Ad Exchanges), обменивается прочими данными в интересах рекламодателя в цифровой экосистеме RTB. Цель DSP — как можно дешевле купить показы аудитории, максимально соответствующей запросам рекламодателя. По сути, DSP позволяет рекламодателям покупать аудиторию, а не конкретные места для размещения рекламы. Когда пользователь кликает на ссылку, SSP-система запускает торги на DSP-площадке. На основании данных SSP, собственной информации с сайта рекламодателя и купленных сведений у DMP (Data Management Platforms — поставщика профилей пользователей и систем управления ими), DSP формирует ставки и проводит RTB-аукцион.

Рекламодатели ищут оптимальную цену на торгах для каждого объявления, чтобы улучшить эффективность их кампании. Оптимальная цена за баннер зависит от CPC (цена за клик) или CPA (цена за действие). Если CPC или CPA установлены правильно, то показатель кликабельности (CTR) будет высокий. Например, ваш рекламный блок показан 1 раз и на него кликнул один человек, значит его CTR — 100 %.

CTR = (количество кликов / количество показов) * 100

CTR напрямую связаны с намерением пользователя, взаимодействующего с объявлением в данном контексте и его трудно моделировать и предсказать, в чем и заключается самая главная сложность.

Иерархия хранения прошлых данных. В основе иерархической модели данных лежит один главный элемент (главный узел), с которого все и начинается, такой элемент называет корневым элементом, в теории графов это называется корнем дерева. Вообще, по сути, что сетевая база данных, что иерархическая база данных имеет древовидную структуру. Все элементы или узлы, которые находятся ниже корневого узла иерархической модели, являются потомками корня. Стоит сказать, что и иерархическая база данных, и сетевая база данных оптимизированы на чтение информации из БД, но не на запись информации в базу данных, эта особенность обусловлена самой моделью данных.

Узлы дерева, которые находятся на одном уровне, обычно называются братьями. Узлы, которые находятся ниже какого-то определенного уровня, являются дочерними узлами по отношению к нему. Иерархическую модель данных можно сравнить с файловой системой компьютера. Компьютер умеет очень быстро работать с отдельными файлами: удалять конкретный файл, редактировать файл, копировать или перемещать файл. Но

операция проверки компьютера антивирусом может происходить достаточно длительное время.

В контекстной рекламе рекламные компании, рекламодателей и непосредственно объявления можно рассматривать придерживаясь некоторой иерархической структуры. Структура представляет собой деревья. Каждое объявление в DSP относится к определенной рекламной кампании, которая, в свою очередь принадлежит рекламодателю. Например, рекламодатель: Acme Cars, кампания: 2011 Year End Sales, объявления: Incredible Year End Sales Event. Точно так же, сайт, на котором будет отображаться объявление, принадлежит издателю и непосредственно издатель может принадлежать какой-то категории. Например, тип издателя: News, издательство: Acme City Times, страница: Auto News. Пример таксономии демонстрирует вышесказанное. Иерархическая структура для пользователя (рис. 1) и издателя (рис. 2), рекламодателя (рис. 3).

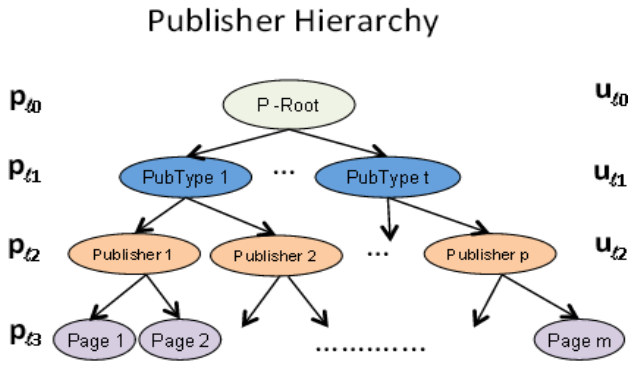

Рис. 1 – Таксономия издателя

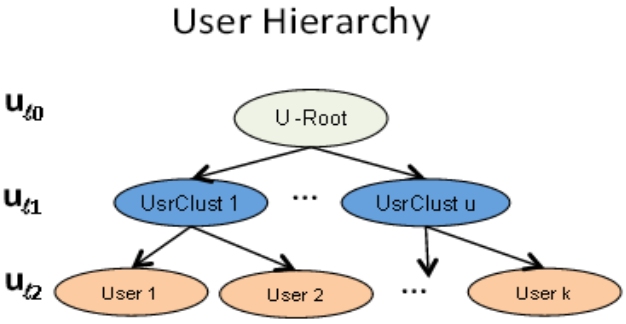

Рис. 2 – Таксономия пользователя

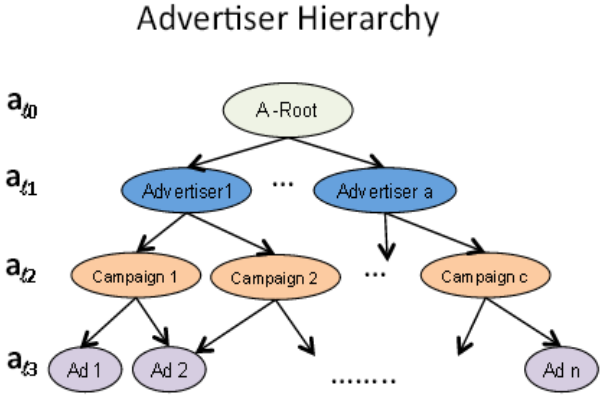

Рис. 3 –Таксономия рекламодателя

Иерархии данных позволяют определить явные или неявные пользовательские кластеры. Явная пользовательская кластеризация, основана на представлении каждого пользователя с помощью набора функций (например, демографическая информация, географических особенностей, уровня доходов, типа веб-сайтов, которые часто посещают, уровня активности и т.д.) и кластеризации на основе некоторых сходство метрики, такие как евклидово расстояние. С другой стороны, неявная кластеризация основана на использовании иерархии данных, а не пользовательских функциях. Например, группа пользователей, которые посещают сайты в определенной категории, например спорт, может рассматриваться в качестве неявного кластера. Мы можем представить эту группу, как декартово произведение {Пользователь x Издатель}. Или же по-другому, мы можем также рассмотреть всех пользователей, которым было показано объявление кампании на определенном сайте: {Пользователь x Издатель x Кампания}.

Если мы предположим, что пользователь, издатель, рекламодатель - это l_u, l_p, l_a. Тогда их соответствующие иерархии данных l_u x l_p x l_a. После чего имея {user: u_i, page: p_j} мы можем определить подходящие явные и неявные кластеры пользователей с иерархией данных и использовать прошлые данные подсчета (то есть, количество показов и число переходов) каждого уровня, чтобы получить различные оценки (вероятности), представленной в формуле (1).

$$P_{ijk} - p\,(Y - 1|\, u_i, p_j, a_k) \qquad (1)$$

Логистическая регрессия. Логистическая регрессия (Logistic regression) — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Основная идея логистической регрессии заключается в том, что пространство исходных значений может быть разделено линейной границей (т.е. прямой) на две соответствующих классам области. Итак, что же имеется ввиду под линейной границей? В случае двух измерений — это просто прямая линия без изгибов. В случае трех — плоскость, и так далее. Эта граница задается в зависимости от имеющихся исходных данных и обучающего алгоритма. Чтобы все работало, точки исходных данных должны разделяться линейной границей на две вышеупомянутых области. Если точки исходных данных удовлетворяют этому требованию, то их можно назвать линейно разделяемыми.

Если одна из переменных (её ещё называют объясняемой) зависит от других факторов (объясняющих переменных), то можно построить уравнение, коэффициенты которого будут свидетельствовать о вероятности для объясняемой переменной принять одно из двух альтернативных значений.

Такое уравнение называется бинарной логистической регрессией.

Использование бинарной логистической регрессии (собственно, как и любого другого метода) не сводится только к выполнению вычислений. Важную роль играет построение модели (какие объясняющие переменные включать в уравнение), осмысление результатов анализа, корректировка исходного уравнения и формулирование выводов.

Что касается интерпретации результатов, то она похожа на интерпретацию значений OR: если их диапазон при заданном доверительном интервале не включает в себя единицу, то рассматриваемый фактор значимо влияет на объясняемую переменную.

В логистической регрессионной модели предсказанные значения зависимой переменной или переменной отклика не могут быть меньше (или равными) 0, или больше (или равными) 1, не зависимо от значений независимых переменных; поэтому, эта модель часто используется для анализа бинарных зависимых переменных или переменных отклика, указываем в формуле (2).

$$y=\exp(b_0+b_1*x_1+...+b_n*x_n)* \\ *[1+\exp(b_0+b_1*x_1+...+b_n*x_n)]^{-1} \quad (2)$$

Легко увидеть, что независимо от регрессионных коэффициентов или величин *x*, предсказанные значения (*y*) в этой модели всегда будут лежать в диапазоне от 0 до 1.

Термин логистической регрессии произошел от того, что эту модель легко линеаризовать с помощью логит преобразования. Предположим, что бинарная зависимая переменная *y* является непрерывной вероятностью *p*, лежащей в диапазоне от 0 до 1. Тогда можно преобразовать эту вероятность *p* следующим образом (формула (3))

$$p' = \log_e \{p/(1-p)\} \quad (3)$$

Это преобразование называется логит или логистическим преобразованием.

Заметим, что *p'* теоретически может принимать любые значения от минус до плюс бесконечности. Поскольку логит преобразование решает проблему 0/1 границ для исходной зависимой переменной (вероятности), то можно использовать эти (логит преобразованные) значения в обычном линейном уравнении регрессии.

Фактически, при проведении логит преобразования обеих частей логит регрессионного уравнения, приведенного выше, мы получим стандартную линейную модель множественной регрессии (формула(4))

$$p' = b_0+ b_1*x_1+ b_2*x_2+ ... + b_n*x \quad (4)$$

Решив уравнение, мы получим значения регрессионных коэффициентов, по которым затем можно восстановить вероятность *p*.

Теперь снова вернемся к нашему алгоритму.

Вместо того, чтобы выбирать лучшую оценку среди всех оценок, мы стремимся оптимально совместить эти оценки.

Для некоторой функции f(.): $[0,1]^M \rightarrow [0,1]$, который имеет набор параметров, обозначенных β. Для каждого представления, исход всех оценок может быть рассчитан. Кроме того, на данном этапе мы можем наблюдать представления с различными объявлениями и запоминать результаты. Каждое объявление подается другому пользователю на новом сайте, на основе чего высчитывается CTR. Вероятность этих обучающих выборок можно записать в след виде.

Мы можем использовать вероятности и выбрать набор параметров, что они будут максимальными в обучающей выборке. При таком наборе f(.) оптимальный набор параметров можно найти решая следующую задачу оптимизации:

Еще один способ интерпретировать эту формулировку, чтобы воспринимать отдельные оценки как факторы (или особенности) в модели классификации и оптимизации процесса в поиске оптимальных коэффициентов линейной комбинации, что будет классифицировать обучаемые данные как можно точнее. Тем не менее, следует отметить, что, поскольку мы не на столько заинтересованы в классификации кластера, сколько, в оценки его вероятность преобразования, нам необходимы только вероятностные оценки, присвоенные логистической регрессии и нам не требуется выбрать порог классификации.

Мы также можем измерить производительность обработки логистической регрессии в качестве вывода оценки классификации.

Выводы. В результате предложен гибкий и принципиальный подход для оценки коэффициентов пересчета для использования в RTB алгоритме. Рассмотренный алгоритм обеспечивает последовательное улучшение в оценке скорости преобразования над некоторыми базовыми оценками. Представленный подход также предлагает простые, но эффективные рецепты для обработки практических вопросов стойких в реальном DSP, такие как отсутствие данных и дисбаланс между данными. Обучение для предлагаемого подхода в течение большого количества кампаний работает быстро благодаря распараллеливания на отдельные части кластера.

Список литературы:1. *Agarwal, D.* Estimating rates of rare events at multiple resolutions [Text] / *D. Agarwal, A. Broder, D. Chakrabarti, D. Diklic, V. Josifovski, M. Sayyadian* // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2007 – P. 6–20. **2.** *Ahmed, A.* Scalable distributed inference of dynamic user interests for behavioral targeting [Text] / *A. Ahmed, Y. Low, M. Aly, V. Josifovski, A. J. Smola* // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2011. – P. 15–30. **3.** *Bax, E.* Comparing predicted prices in auctions for online advertising [Text] / *E. Bax, A. Kuratti, P. Mcafee, J. Romero* // Int. J. of Industrial Organization, 2012. – P. 4–5. **4.** *Blei, D.* Latent dirichlet allocation [Text] / *D. Blei, A. Ng, M. Jordan* // J. of Machine Learning Research, 2003. – P. 20–23. **5.** *Cai, J. F.* A singular value thresholding algorithm for matrix completion [Text] / *E. J. Candes, and Z. Shen*// SIAM J. on Optimization, 2008. – P. 20–23. **6.** *Cerrato, D.* Classication of proxy labeled examples for marketing segment generation [Text] / *R. Jones, A. Gupta* // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2011. – P. 15–20. **7.** *Chen Y.* Real-time bidding algorithms for performance-based display ad allocation [Text] / *P. Berkhin, B. Anderson, N. R. Devanur* // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2011. – P. 15–20. **8.** *De Leeuw, J.* Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods [Text] / *K. Hornik, P. Mair* // J. of Statistical Software, 2009. – P. 3–24. **9.** *Williams, D* On classication with incomplete data [Text] / *X. Liao, Y. Xue, L. Carin, B. Krishnapuram* // On Pattern Analysis And Machine Intelligence, 2007. – P. 29. **10.** *Menon, A* Response prediction using collaborative Itering with hierarchies and side-information [Text] / *K. Chitrapura, S. Garg, D. Agarwal, and N. Kota* // ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2011. – P. 21–34. **11.** *Visa, S.* Issues in mining imbalanced data sets - a review paper [Text] / *S. Visa* // Cognitive Science Conf, 2005. – P. 67–73.

Bibliography (transliterated): 1. *Agarwal, D., Broder, A., Chakrabarti, D., Diklic, D., Josifovski, V., Sayyadian, M. (2007).* Estimating rates of rare events at multiple resolutions. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. **2.** *Ahmed, A. Low, Y., Aly, M., Josifovski, V., Smola, A. J. (2011).* Scalable distributed inference of dynamic user interests for behavioral targeting. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. **3.** *Bax, E., Kuratti, A., Mcafee, P., Romero, J.* (2012). Comparing predicted prices in auctions for online

advertising. Int. J. of Industrial Organization, 30:80{88. **4** *D. Blei, A. Ng, Jordan, M.* (2003). Latent dirichlet allocation. J. of Machine Learning Research, 3:993{1022. **5.** *Cai, J.-F., Candes, E. J., Shen, Z.* (2008). A singular value thresholding algorithm for matrix completion. SIAM J. on Optimization, 20:1956{1982. **6.** *Cerrato, D., Jones, R., Gupta, A.* (2011). Classication of proxy labeled examples for marketing segment generation. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. **7.** *Chen, Y., Berkhin, P., Anderson, B., Devanur, N. R. (2011).* Real-time bidding algorithms for performance-based display ad allocation. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. **8.** *De Leeuw, J., Hornik, K., Mair, P.* (2009). Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. J. of Statistical Software, 32(5):1{24. **9.** *Williams, D. Liao, X., Xue, Y., Carin, L., Krishnapuram, B. (2007).* On classication with incomplete data. IEEE Trans. On Pattern Analysis And Machine Intelligence, 29. **10.** *Menon, A., Chitrapura, K., Garg, S., Agarwal, D., Kota. N. (*2011). Response prediction using collaborative ltering with hierarchies and side-information. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. **11.** *Visa, S., Ralescu, A. (2005).* Issues in mining imbalanced data sets - a review paper. Proc. of the 16th Midwest AI and Cognitive Science Conf., 67{73

Поступила (received) 25.12.2015

Відомості про авторів / Сведения об авторах / About the Authors

Савченкова Анастасія Юріївна – Студентка, Кафедра Комп'ютерних наук, Донецький національний технічний університет, пл. Шибанкова, 2, м. Красноармійськ, Донецька область, Україна, 85300,

Савченкова Анастасия Юрьевна – Студентка, Кафедра Компьютерных наук, Донецкий национальный технический университет; пл. Шибанкова, 2, г. Красноармейск, Донецкая область, Украина, 85300;

Savchenkova Anastasija – Student, the Department of Computer Sciences , Donetsk National Technical University; Shybankova, 2,. Krasnoarmeysk, Donetsk region, Ukraine, 85300; e-mail: nasstya05@gmail.ru

УДК 338.24.01

Е. А. КОВАЛЕВА

РЕГРЕССИОННАЯ МОДЕЛЬ СЕБЕСТОИМОСТИ ЭЛЕКТРОННЫХ МУЛЬТИМЕДИЙНЫХ ИЗДАНИЙ

Данная статья посвящена построению эконометрических моделей себестоимости электронных мультимедийных изданий на основании статистических данных издательского центра "Академия". В статье автор рассматривает эконометрические модели двух типов – аддитивную и мультипликативную. Каждая из моделей построена прямым пошаговым методом, на каждой итерации которого методом наименьших квадратов оценивались значения параметров модели, анализировались статистическая значимость коэффициента при переменной, введенной на текущей итерации, и значение скорректированного коэффициента множественной детерминации.

Ключевые слова: эконометрическая модель, себестоимость, электронные мультимедийные издания.

Введение. На сегодняшний день ХНЭУ им. С. Кузнеца активно занимается разработкой и внедрением в учебный процесс электронных мультимедийных изданий (ЭМИ) [1]. Этот процесс осуществляется в русле общих проводимых в отечественном образовании реформ, обусловленных переходом к новой образовательной парадигме, приоритетами которой является повышение качества подготовки специалистов и их соответствие уровню требований интенсивно развивающегося общества. Одним из перспективных путей повышения качества подготовки специалистов признается широкое внедрение в учебный процесс ЭМИ, позволяющих управлять процессами образовательной деятельности. К числу ЭМИ в том числе относятся электронные учебники и пособия (ЭУ и ЭП). Вопрос стоимости создания ЭУ и ЭП – один из ключевых, ответ на который может предопределить судьбу электронного учебного издания. Каких-либо регламентированных методик расчета стоимости создания электронных учебных изданий пока не создано [2]. В данном случае каждая команда разработчиков вынуждена создавать собственную методику расчета стоимости создания ЭУ и ЭП.

Так как создание ЭМИ является достаточно новым видом деятельности, вопрос себестоимости продуктов такого рода весьма сложен. С одной стороны, создание ЭУ и ЭП – достаточно трудоемкий и специфический процесс, требующий приложения большого количества усилий [3]; с другой стороны, рынок диктует свои ограничения на цену ЭМИ.

Так как утвержденной или хотя бы общепринятой методики расчета стоимости создания ЭМИ нет, в данной статье автор предлагает свой вариант расчета себестоимости ЭМИ, используя эконометрические модели, концепции, приемы.

Анализ литературных данных и постановка проблемы. Для обоснованного выбора методологических подходов к моделированию себестоимости ЭМИ проанализировано достаточное количество работ, включающих опыт эконометрического моделирования большинства социально-экономических систем. Так, в работах [4 – 6] показано, что на сегодняшний день лучшие результаты дают именно эконометрические модели. Работы [7, 8] посвящены ценообразованию и себестоимости различных объектов так же используя эконометрический подход. Но ни одна из выше перечисленных работ не описывает себестоимость ЭМИ. Это объясняется тем, что ЭМИ достаточно новый вид продукции, себестоимость которого является не решенной задачей.

С другой стороны, существует достаточное количество статей, посвященных технологиям создания, развитию и практическому использованию ЭМИ как в учебном процессе, так и в коммерческой сфере [9 –11].