УДК 311.21

АНАЛИЗ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ В ЦЕНООБРАЗОВАНИИ

Хуснияров И.Ф.
ООО «Тур», Уфа, e-mail: Tour@fangid.com

Определение цен на товары и услуги всегда зависит от множества факторов, при этом во время определения цены товара и/или услуги необходимо учитывать, что завышенная цена может принести разовую прибыль или отпугнуть потенциальных покупателей. В связи с этим в рамках ценообразования устанавливаются тарифы, которые позволяют «завоевать» большую часть рынка и обеспечить максимизацию прибыли. В рамках данной статьи рассматриваются подходы к ценообразованию с точки зрения получаемой прибыли с использованием инструментов анализа данных и моделей машинного обучения. Машинное обучение – это одна из наиболее актуальных областей, которое позволяет решать задачи прогнозирования, классификации или кластеризации данных. Применение анализа данных и машинного обучения к процессу ценообразования билетов на концерты артистов позволяет обеспечить стратегию динамического ценообразования, снизить трудозатраты на определение стоимости билета на концерт, обеспечить учет нестандартных факторов, которые оказывают влияние на стоимость. Полученные модели являются частью музыкально-аналитического сервиса Fanstat – платформы музыкальной аналитики, которая позволяет отслеживать динамику и рейтинговать популярность артистов в региональном разрезе, как ключевого фактора ценообразования в рамках организации концертной деятельности.

Ключевые слова: моделирование, машинное обучение, анализ данных, ценообразование, популярность артистов, сервис, социальные сети

DATA ANALYSIS AND MACHINE LEARNING IN PRICING

Khushiyarov I.F.
Tour LLC, Ufa, e-mail: Tour@fangid.com

Determination of prices for goods and services always depends on the preferences of buyers, while choosing the price of goods and / or services, it must be taken into account that an overpriced price can bring a one-time profit or scare off buyers. In connection with the establishment of pricing tariffs that allow you to “win” a large part of the market and maximize profits. This article discusses approaches to pricing in terms of profit received using data analysis tools and machine learning models. Machine learning is one of the most relevant areas that allows you to solve the problems of predicting, classifying or clustering data. The application of data analysis and machine learning to the process of pricing tickets for concerts of artists allows us to provide a dynamic pricing strategy, reduce labor costs for determining the cost of a concert ticket, and ensure that non-standard factors that affect the cost are taken into account. The resulting models are part of the Fanstat music analysis service, a music analytics platform that allows you to track the dynamics and rate the popularity of artists in the regional context, as a key pricing factor in the organization of concert activities.

Keywords: modeling, machine learning, data analysis, pricing, artist popularity, service, social media

Стремительное развитие социальных сетей за последние несколько лет связано в первую очередь с технологическими прорывами и пандемией. Подобное развитие технологий детерминировало медиасреду и медиапотребление [1]. Согласно данным статистики [2], социальные сети используют более 4,5 млрд чел., а темпы прироста пользователей за 2021 г. составили почти 10%.

Самыми популярными социальными сетями являются Youtube (более 90 млн пользователей в России), ВКонтакте (более 70 млн пользователей в России), Tik-Tok (более 35 млн пользователей в России) [3]. При этом по данным ВЦИОМ за сентябрь 2021 г. 54% жителей России проводят в социальных сетях более часа в день.

Различные социальные сети могут использоваться как средство общения, прослушивания музыки, просмотра видео и картинок, чтения новостей, обмена мнениями. Новым направлением использования социальных сетей в последние годы стала рассылка рекламной информации, которая обеспечивается механизмами таргета с точки зрения продавцов/производителей товаров или услуг и, соответственно, поиск информации, касающейся отзывов и рекомендаций, перед покупкой товаров или услуг с точки зрения потребителей.

Внедрение и распространение подобных механизмов использования социальных сетей с точки зрения культурно-развлекательных мероприятий и концертной деятельности делает возможным проведение оценки популярности артистов на основе вовлеченности аудитории социальных сетей и использовании полученных результатов для решения различных задач.

Важным аспектом при организации концертной деятельности является определение стоимости билета на мероприятие, которое должно соответствовать уровню популярности и востребованности артиста

в определенном городе или регионе, месту проведения и ряду других факторов, которые определяют динамичность процесса ценообразования. Использование инструментов анализа данных и машинного обучения в процессах динамического ценообразования позволяет максимизировать получаемую организаторами мероприятий прибыль и избавиться от большого количества непроданных билетов, минимизировав при этом трудозатраты.

Преимущества применения моделей машинного обучения в таких процессах связаны в первую очередь с возможностью обработки большого количества данных, самостоятельной коррекции модели с течением времени под влиянием изменения ключевых факторов ценообразования и получения быстрого результата.

Данные для построения моделей ценообразования в рамках данного исследования собираются из разных источников: для оценки популярности и востребованности артистов с помощью парсеров собираются данные из социальных сетей, стриминговых сервисов, запросов, чартов и радиостанций. Дополнительные параметры, связанные с экономическим положением регионов, собираются из статистических таблиц и сборников, ежегодно публикуемых Росстатом. Для формирования цены и ее оценки используются экспертные оценки, полученные путем опросов организаторов концертной деятельности.

Цель данного исследования заключается в применении моделей машинного обучения и их применимости к динамическому ценообразованию. В рамках данного исследования выбрана наиболее целесообразная модель для данной прикладной задачи, алгоритм работы которой стал одной из основных моделей сервиса музыкальной аналитики. Непосредственными задачами при этом являются:

1) анализ существующих моделей и методов машинного обучения;

2) анализ методов сбора и учета данных;

3) разработка математической модели формирования популярности артиста на основе данных социальных сетей, как одного из факторов ценообразования;

4) проработка алгоритмов и выбор оптимального решения.

Для реализации исследования использована интегрированная среда разработки для языков R и Python RStudio с открытым исходным кодом. Основным инструментом сбора данных служат парсеры веб-страниц и классы, работающие с API обрабатываемых сервисов. Для каждого из обрабатываемых сервисов написан отдельный парсер, работа которого определяется структурой исследуемой страницы. Для обеспечения безопасного доступа к данным и решения возможных проблем доступа используются такие инструменты, как:

− библиотеки, предоставляющие функции для удобной работы с API некоторых сервисов, например Apple Music API, VK API;

− Selenium – этот инструмент также позволяет выполнить скрипт и обратиться к HTML-элементу с помощью css-селектора;

− прокси.

Используемые модели и алгоритмы

Для построения динамического ценообразования в рамках исследования использовались такие алгоритмы и модели машинного обучения, как:

− модель XGBoost – модель градиентного бустинга, которая реализует последовательно уточняющие друг друга деревья решений. Данная модель служит для решения задач классификации и регрессии с реализацией последовательного обучения;

− модель дерева решений – модель решения задачи классификации и прогнозирования на основе поиска показателей, дающих наилучшую классификацию или предсказание на основе разбиения данных на подгруппы в ходе рекурсии;

− модель Random forest, применяемая для решения задач классификации, кластеризации и регрессии на основе генерации ансамбля деревьев решений, каждое из которых строится на основе случайной подвыборки из обучающего набора;

− модель SMOTE – модель, решающая проблемы стандартизации выборки с несбалансированными классами за счет дополнения подвыборок сгенерированными данными, похожими на данные подвыборки.

Для реализации моделей использована выборка из данных по прошедшим в 2019 г. концертам. Данная выборка разделена на обучающую и тестовую в отношении 0.7/0.3 случайным образом.

Алгоритм проведения исследования

Для построения моделей ценообразования было проведено поэтапное исследование, включающее в себя:

1. На первом этапе проводится сбор данных из различных источников. Проводится предобработка данных.

2. На втором этапе проводится расчет популярности артиста на основе данных социальных сетей.

3. На третьем этапе проводится построение одной из моделей машинного обучения, включающей в себя два последовательных

применения одного и того же алгоритма. При первом применении оценивается уровень сборов относительно категорий плохой/средний. При втором применении относительно категорий средний/отличный.

4. На четвертом этапе проводится оценка и интерпретация полученных значений.

5. На пятом этапе отбирается наиболее подходящая модель.

Статистическая информация, необходимая для построения обучающей и тестовой выборок

Ключевой параметр, оказывающий влияние на ценообразование билета – это популярность артиста в определенном городе или регионе. Для оценки популярности артистов с учетом распространенности социальных сетей и медиа использованы показатели, характеризующие вовлеченность аудитории социальных сетей как подписчиков артистов. Данные параметры собираются по единому показателю популярности на основе средневзвешенных оценок распространенности каждой социальной сети на территории России. В качестве дополнительных параметров при построении моделей машинного обучения используются: бинарные показатели активности артиста в социальных сетях и уровень формации (старая формация до 2010 г. и новая формация с 2010 г. появления артиста на сцене), количество запросов в Яндексе и Википедии с учетом территориального расположения, финансово-экономические показатели региона, в котором запланирован концерт (количество населения, средний уровень расходов, валовой региональный продукт).

Для формирования обучающей выборки добавлены показатели стоимости билета на прошедшие концерты, вместимость и тип площадки, на которой проводился концерт (клубы, дворцы спорта, концертные залы) и экспертная оценка уровня сборов по прошедшим мероприятиям с точки зрения их окупаемости (плохой, средний, отличный). Ключевые показатели и метрики для построения моделей представлены в таблице.

Ключевые показатели

Показатель	Тип и состав
Популярность артиста	Вещественный, рассчитан путем агрегирования показателей социальных сетей с учетом их распространенности на территории России (по количеству пользователей) по проработанной математической модели на основе логарифмирования. Использованы такие социальные сети и показатели, как: 1. ВКонтакте (Количество записей артиста, Количество репостов записей, Количество лайков, Количество комментариев, Количество просмотров, Количество подписчиков). 2. Tik-Tok (Количество видео с треками артиста, Количество лайков, Количество комментариев, Количество просмотров, Количество подписчиков). 3. YouTube (Количество видео артиста, Количество дизлайков, Количество лайков, Количество комментариев, Количество просмотров, Количество подписчиков). 4. GoogleAds, Shazam (Количество запросов). 5. Радиостанции (Количество треков артиста на радиостанции, Количество воспроизведений в эфире). 6. Чарты (Количество дней в чарте, Средняя позиция в чарте, Дата наивысшей позиции в чарте). 7. Стриминговые сервисы (Количество подписчиков плейлистов, Средняя позиция плейлиста, Количество плейлистов, Количество подписчиков артиста)
Яндекс, Википедия	Целочисленный, Количество запросов
Награды	Целочисленный, Количество наград артиста
Активность	Бинарный, Активен в социальных сетях / Не активен в социальных сетях
Формация	Бинарный, Старая/новая
Экономические показатели	Целочисленный, Численность населения в регионе Целочисленный, Средние расходы населения Вещественный, Валовой региональный продукт
Вместимость площадки	Целочисленный, Количество мест
Тип площадки	Целочисленный, 1 – Дворец спорта, 2 – Сидячий зал, 3 – Клуб
Цена билета	Целочисленный, Стоимость
Сбор	Целочисленный, 1 – плохой сбор, 2 – средний сбор, 3 – отличный сбор

Сбор необходимой статистической информации проводится с использованием программы сбора и систематизации информации в несколько этапов.

1. На первом этапе проводится сканирование исходного массива информации страницы в социальных сетях, базы данных стримингового сервиса, чарта, запросов.

2. На втором этапе проводится конвертация полученных данных в необходимый формат и агрегация полученных результатов в единой таблице.

Проверка качества исходной информации

Собираемые для проведения исследования данные должны отвечать определенным требованиям [4]:

1) достоверности – соответствию данных тому, что есть на самом деле. В настоящем исследовании методика, техника и организация проведения статистического наблюдения направлены на обеспечение достоверных данных. Как известно, общим условием обеспечения достоверности является полнота охвата наблюдаемого объекта, то есть полнота и точность регистрации данных по каждой единице наблюдения [5]. Это условие выполняется на основе обновления данных в режиме реального времени.

2) возможности обобщения данных об отдельных явлениях или их сопоставимости друг с другом, то есть, чтобы данные собирались в одно и то же время и по единой методике. Для выполнения данного условия все показатели должны быть приведены к стандартизированному виду, что обеспечивается работой программы сбора и систематизации информации перед занесением в базу данных.

Для организации процесса сбора данных разработаны парсеры для социальных сетей, которые обеспечивают автоматический сбор данных о подписчиках, упоминаниях и статистике в социальных сетях, а также обеспечивают наличие данных об их вовлеченности и активности на страницах артистов (лайки, дизлайки, комментарии, просмотры). Для обеспечения качества информации при таком подходе решаются следующие задачи:

1. Требования к конечным данным. При наличии четкой структуризации социальных сетей был определен набор показателей (таблица) и временные диапазоны.

2. Периодический анализ структуры социальных сетей для отслеживания изменений и соответствия полей.

3. Контроль охвата по элементам и охвата по полям.

Для моделирования использованы данные по реализованным концертам артистов за 2019 г. Для моделирования собрана информация о концертной деятельности артистов за 2019 г. Оценка популярности артиста в социальных сетях рассчитывается на основании данных, полученных за полгода до даты проведения концерта, для оценки уровня сборов применяется метод экспертных оценок. 27,5 % выборки – концерты с низким уровнем сборов, 33 % – со средним уровнем сборов, 39,5 % – с высоким уровнем сборов.

Результаты моделирования

1. Модель дерева принятия решений (рис. 1), реализованного на языке R с помощью команды rpart. Зависимая переменная – показатель сборов, независимые переменные – все остальные показатели из таблицы. После этого выстраиваем матрицу сопряженности. В рамках данного решения получена матрица сопряженности 21/0 0/16 с чувствительностью и специфичностью модели на уровне 1,0. Особенностью полученных результатов является изначальная несбалансированность выборки.

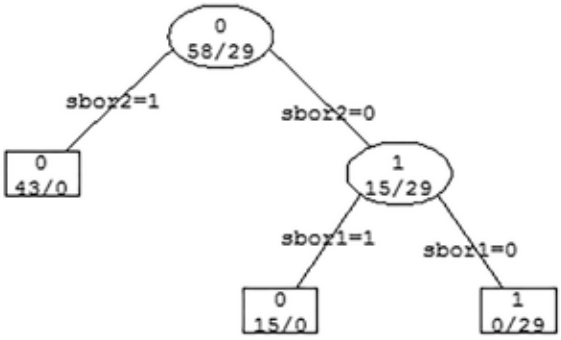

Рис. 1. Дерево принятия решений

2. Модель Random forest. Для работы данного алгоритма необходимо введение дополнительных параметров: количества деревьев и количества регрессоров. Для реализации исследования были выбраны параметры 300 (количества деревьев) и 4 (количество регрессоров) исходя из наибольшего падения индекса Джини (рис. 2).

Данная модель показывает низкую чувствительность и специфичность, при итоговой матрице сопряженности 6/16 15/0.

Для того чтобы не подбирать параметры для модели вручную, воспользуемся пакетом caret, который позволяет создать сетку гиперпараметров для перебора модели. В рамках моделирования варьируем параметр количества регрессоров с использованием кросс-валидации. Данный процесс дает наиболее высокие показатели при количестве регрессоров равном 4, т.е. качество модели не может быть улучшено.

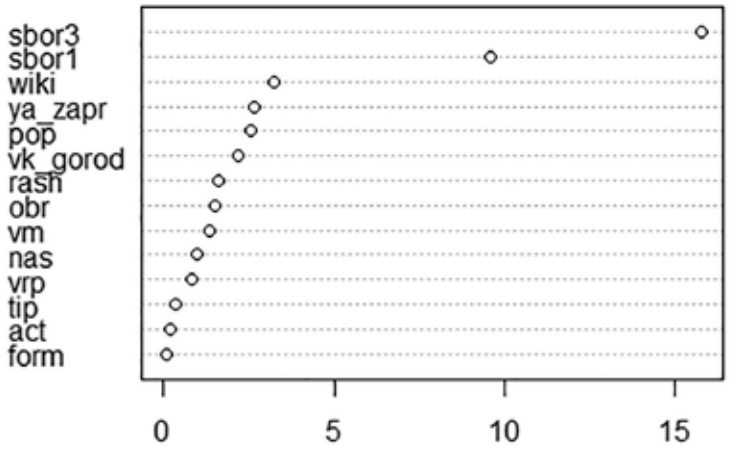

Рис. 2. Среднее уменьшение индекса Джини

Введем дополнительный параметр – количество случайных отсечений и построим модель Random forest by randomization на четырех вариациях. Результаты моделирования представлены на рис. 3.

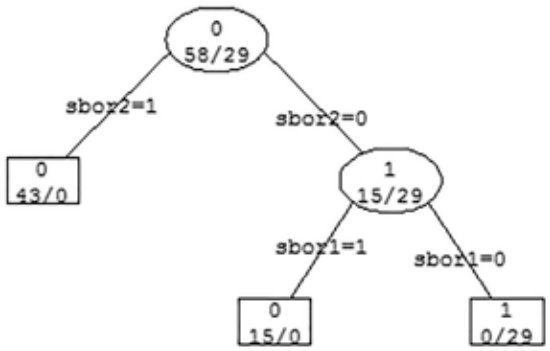

Рис. 3. Random forest by randomization

Лучшая модель построена на показателях 4/2, однако существенных улучшений качества модели с точки зрения чувствительности, специфичности и точности это не дало.

3. Модель XGBoost. Первый вариант модели на основе экстремального бустинга с сеткой поиска, с учетом дополнительных параметров: скорость обучения, минимального значения функции потерь и максимальной глубины дерева не дал существенного повышения качества модели.

Для реализации повышения качества модели и улучшения результатов моделирования рассмотрим возможные методы балансировки целевого показателя. К таким методам относят стратегии сэмплирования (undersapmling – основан на удалении некоторого количества примеров мажоритарного класса и oversampling – основан на увеличении количества примеров миноритарного класса) и метод SMOTE.

Метод SMOTE позволяет сбалансировать оба класса при формировании выборок. Данная модель применяется при несбалансированности классов в обучающей выборке. Общая идея заключается в искусственной генерации примеров миноритарного класса с использованием ближайших соседей этих случаев. Контроль количества случаев выбор по классам (класс меньшинства и класс большинства) контролируется параметрами. Формирование сбалансированной выборки данным методом напоминает стандартный формат построения моделей в R.

Для применения стратегии сэмплирования oversampling преобразуем выборку в разряженную матрицу и проведем процедуру сэмплирования, поскольку выборка не является сбалансированной. Результаты моделирования XGBoost после сбалансирования выборки дают чувствительность на уровне 0,79, при специфичности 0,82, уровень сбалансированности 0,8, со смещением в сторону ошибок первого рода.

Применение метода SMOTE в сочетании с экстремальным бустингом для оценки уровня сборов артистов позволяет получить модель с чувствительностью 1,0, специфичностью 0,88, при уровне сбалансированности 0,94 и высоком уровне точности, со смещением в сторону ошибок второго рода. Таким образом, результаты исследования показали, что SMOTE значительно превосходит другие методы с точки зрения повышения точности прогнозирования и качества моделирования.

Выбор модели ценообразования

Для выбора наиболее подходящей модели ценообразования из реализованных, воспользуемся полученными параметрами качества моделирования. Несмотря на самые лучшие показатели качества, первая модель не учитывает несбалансированность выборки, соответственно, не может быть принята в качестве результата. Наилучшим образом показали себя модели экстремального бустинга после применения методов сбалансирования выборок.

С точки зрения двух оставшихся моделей обратимся к особенностям, связанным с ошибками первого и второго рода.

С учетом специфики исследования ошибка первого рода подразумевает наличие среднего или отличного сбора (количества вырученных организаторами средств) при соответственно плохом или среднем уровне. То есть данная ошибка является критичной, поскольку организаторы концертной деятельности не получат заявленных средств и могут не окупить проведение мероприятия. Исходя из этого, предпочтительным назовем смещение в сторону ошибок второго рода и выбор модели XGBoost с применением метода SMOTE, как наиболее оптимального алгоритма ценообразования.

Список литературы

1. Щепилова Г.Г., Круглова Л.А. Телеканалы и социальные сети: специфика взаимодействия // Вестник Московского университета. Серия 10. Журналистика. 2018. № 3. С. 3–16.

2. Simon Kemp Digital 2021 October Global Statshot Report. [Электронный ресурс]. URL: https://datareportal.com/reports/digital-2021-october-global-statshot (дата обращения: 20.05.2022).

3. Аудитория социальных сетей и мессенджеров в 2021 году. [Электронный ресурс]. URL: https://blog.skillfactory.ru/auditoriya-soczialnyh-setej-i-messendzherov-v-2022-godu/ (дата обращения: 14.10.2022).

4. Хуснияров И.Ф. Сервис популярности артистов на основе анализа социальных сетей // Международный журнал экспериментального образования. 2022. № 3. С. 20–24.

5. Завьялова Н.Б., Головина А.Н., Завьялов Д.В., Дьяконова Л.П., Мельников М.С., Сагинова О.В., Сагинов Ю.Л., Семенков А.В., Скоробогатых И.И., Строганов И.А. Методология и методы научных исследований в экономике и менеджменте: пособие для вузов / Под ред. Н.Б. Завьяловой, А.Н. Головиной. М. – Екатеринбург, 2014. 282 с.