УДК 338.482

Д. И. Коровин

ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: dikorovin@fa.ru

Е. Л. Золоторева

ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: elzolotareva@fa.ru

П. П. Радачинская

ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: 79152361074@mail.ru

ХАРАКТЕРИСТИКА ВЛИЯНИЯ РАЗЛИЧНЫХ СЦЕНАРИЕВ НА РАЗВИТИЕ ТУРИСТСКОЙ ИНФРАСТРУКТУРЫ

Ключевые слова: цифровые технологии, туристский рынок, туристские потоки, моделирование, цифровая экономика, туристская инфраструктура, оптимизационные модели, датасеты, устойчивое развитие, эконометрические модели, дестинация, эффективность.

В данной статье будет продемонстрирован механизм, отслеживающий динамику туристских потоков между регионами России под воздействием факторов средового влияния и, как следствие, отслеживающий динамику доходов/расходов дестинации от туристской деятельности и уровня развития инфраструктуры. Данный алгоритм основан на методах машинного обучения. В ходе статьи будет доказана важность построения двух типов моделей. Модели первого типа прогнозируют совокупный туристский поток, измеряемый числом ночевок и объемом доходов коллективных средств размещения (КСР), в регион с заданными характеристиками. Модели второго типа прогнозируют перекрестные потоки между регионами, при этом регион отправления и регион прибытия описываются разным набором признаков. В статье отмечено, что использование моделей двух типов обусловлено следующими причинами. Во-первых, стоит отметить, что мировой опыт прогнозирования туристских потоков представлен в большинстве случаев именно моделями прогнозирования туристских прибытий без анализа географической структуры спроса. Более того, представленные модели, за редким исключением, не предполагают использование факторов средового влияниях и уровня развития туристской инфраструктуры в качестве объясняющих переменных. Таким образом, поставленная в настоящей статье научная задача является новой. Во-вторых, имеют место существенные различия в методиках сбора статистической информации о туристском потоке.

D. I. Korovin

Financial University under the Government of the Russian Federation, Moscow, e-mail: dikorovin@fa.ru

E. L. Zolotoreva

Financial University under the Government of the Russian Federation, Moscow, e-mail: elzolotareva@fa.ru

P. P. Radachinskaya

Financial University under the Government of the Russian Federation, Moscow, e-mail: 79152361074@mail.ru

CHARACTERISTICS OF THE IMPACT OF VARIOUS SCENARIOS ON THE DEVELOPMENT OF TOURIST INFRASTRUCTURE

Keywords: digital technologies, tourist market, tourist flows, modeling, digital economy, tourist infrastructure, optimization models, datasets, sustainable development, econometric models, destination, efficiency.

This article will demonstrate a mechanism that tracks the dynamics of tourist flows between the regions of Russia under the influence of environmental factors and, as a result, tracks the dynamics of destination income/expenses from tourism activities and the level of infrastructure development. This algorithm is based on machine learning methods. In the course of the article, the importance of building two types of models will be proved. Models of the first type predict the total tourist flow, measured by the number of overnight stays and the amount of income of collective accommodation facilities (DAC), to a region with specified

characteristics. Models of the second type predict cross-flows between regions, while the departure region and the arrival region are described by a different set of features. The article notes that the use of two types of models is due to the following reasons. Firstly, it is worth noting that the world experience of forecasting tourist flows is represented in most cases by models of forecasting tourist arrivals without analyzing the geographical structure of demand. Moreover, the presented models, with rare exceptions, do not assume the use of environmental factors and the level of development of tourist infrastructure as explanatory variables. Thus, the scientific task set out in this article is a new one. Secondly, there are significant differences in the methods of collecting statistical information about the tourist flow.

Введение

Принятая в Российской Федерации периодичность сбора данных на квартальной основе не соответствует мировым трендам (так, в большинстве развитых стран сбор данных ведется на ежемесячной основе, а в отдельных дестинациях Китая – на ежедневной и даже почасовой). Кроме того, статистика прибытий в разрезе регионов не ведется. Поскольку объем и качество статистических данных играют решающую роль для построения модели машинного обучения, было принято решение о поэтапном моделировании туристского потока – сначала на уровне региона в целом (модели совокупного потока), далее – в разрезе регионов, из которых прибывают туристы (модели перекрестного потока). По мере усложнения моделей возрастают требования к наборам данных и увеличивается количество допущений и импутаций, необходимых для восстановления отсутствующих значений. Для обеспечения воспроизводимости результатов моделирования все указанные преобразования подробно описываются в настоящем разделе с указанием источника данных. Основными источниками являются базы данных Росстата, Ростуризма, Банка России и проекта «Инфраструктура научно-исследовательских данных». Важным ноу-хау исследования является использование Google Trends (статистика поисковых запросов), для оценки вклада каждого региона в совокупный поток дестинации, так как подобного рода информация в РФ централизованно не собирается. Обработка данных и моделирование производились с помощью программного кода на языке Python 3.6 с использованием специализированных библиотек Pandas, Numpy, Scikit-learn, XGBoost, Matplotlib,pytrends и ряда других.

Цель исследования: комплексный учет основных характеристик влияния различных сценариев на развитие туристской инфраструктуры при создании концептуальной модели прогнозирования внутренних туристских потоков для определения приоритетов финансовой, административной и информационной поддержки проектов создания новых объектов исследуемой инфраструктуры

Материалы и методы исследования

Модели совокупного потока

Целевые переменные

Для оценки объема туристского потока могут использоваться различные количественные показатели. В первую очередь, это непосредственно количество лиц, прибывших в регион с туристической целью (с ночевкой или без, остановившихся в коллективных средствах размещения или в «частном секторе»). Во-вторых, объем туристского потока может измеряться в денежном выражении – например, таким показателем могут служить совокупные траты лиц, прибывающих в регион, на приобретение товаров и услуг, связанных с индустрией гостеприимства или же, наоборот, доходы гостиниц, санаториев, предприятий общественного питания, транспортных компаний и т.п. Вопрос выбора оптимального показателя для измерения объема туристского потока в настоящее время является открытым, поскольку существуют проблемы как методологического, так и практического характера. Так, в ходе дискуссии на Круглом столе при Аналитическом центре при Правительстве РФ [1] было отмечено, что централизованно собираемые в настоящее время показатели не дают полного представления об объеме туристского потока. В частности, по оценкам компании Мегафон, построенным на основе исследования динамики передвижений абонентов сети, лишь 20% отдыхающих в Краснодарском крае останавливаются в зарегистрированных средствах коллективного размещения и, соответственно, отражаются в статистике. В то же время, следует учитывать, что с практической точки зрения сбор статистических данных нужного уровня детализации невозможен без межведомственной интеграции информационных систем (и, возможно, сотрудничества с частными

компаниями из телекоммуникационного и банковского сектора) и нового уровня цифровизации туристической отрасли, о чем заявлено в Стратегии 2035[2].

В настоящем исследовании основным источником данных об объеме туристского потока является ресурс «Открытые данные Ростуризма», в частности, группа наборов данных «Статистическая информация в сфере туризма» [3]. Из доступных в данной группе наборов, индикаторами объема туристского потока в разрезе регионов, с определенными допущениями, могут служить показатели, приведенные в таблице (табл. 1).

Из характеристик наборов видно, что наибольший объем данных доступен по двум показателям – «Информация о количестве ночевок в коллективных средствах размещения» (далее – «Ночевки») [4] и «Доходы коллективных средствах размещения от предоставляемых услуг без НДС, акцизов и аналогичных платежей» (далее – «Доходы») [5], поэтому именно они были выбраны в качестве целевых переменных для моделирования туристского потока.

Ниже в качестве иллюстрации приведен срез набора данных о ночевках (рис. 1). Полный набор содержит 109 строк и 61 колонку. Структура датасета «Доходы» полностью аналогична.

Следует отметить, что, во-первых, информация о регионах приводится в текстовом формате, что чревато техническими ошибками (опечатки, лишние пробелы, различные названия регионов – например, «Чувашская республика» и «Республика Чувашия»). Кроме того, субъекты разных уровней не выделяются – например, данные по Центральному федеральному округу включают в себя информацию по входящим в него областям и т.п. Во избежание дублирования записей и технических ошибок, для дальнейшей работы потребуется унифицировать данные о регионах с помощью специального справочника регионов, содержащего цифровой код региона, варианты написания названия и сведения об иерархии (принадлежность к тому или иному федеральному округу). Во-вторых, сведения о временных интервалах приведены также в текстовом формате, а квартальные данные и вовсе публикуются накопленным итогом (например, «январь-март 2011 года»). Это в свою очередь, потребует дополнительных манипуляций для выделения информации по каждому кварталу в отдельности и унификации дат в формате «дата/время». В то же время, описанные преобразования являются техническими и не вызывают логических затруднений.

Таблица 1

Наборы данных, характеризующих туристский поток

Название набора данных	Характеристики набора
Информация о численности граждан Российской Федерации, размещенных в коллективных средствах размещения (чел.)	Ежегодные данные – с 2013 по 2019 гг., поквартальные – с 2019 по март 2021 гг.
Информация о численности иностранных граждан, размещенных в коллективных средствах размещения (чел.)	Ежегодные данные – с 2013 по 2019 гг., поквартальные – с 2019 по март 2021 гг.
Информация о численности лиц, размещенных в коллективных средствах размещения	Ежегодные данные – с 2013 по 2019 гг., поквартальные – с 2019 по март 2021 гг.
Информация об объеме платных туристских услуг (млн руб.)	Только данные за 1-е полугодия 2017 и 2018 гг.
Информация об объеме платных услуг гостиниц и аналогичных средств размещения (млн руб.)	Ежегодные данные с 1993 по 2018 гг., а также данные за январь-сентябрь 2019 г.
Информация о количестве ночевок в коллективных средствах размещения	Ежегодные данные с 2002 по 2010 гг., поквартальные – с 2011 по март 2021 гг.
Доходы коллективных средствах размещения от предоставляемых услуг без НДС, акцизов и аналогичных платежей	Ежегодные данные с 2002 по 2010 гг., поквартальные – с 2011 по март 2021 гг.
Доходы санаторно-курортных организаций от предоставляемых услуг без НДС, акцизов и аналогичных платежей	Ежегодные данные с 2003 по 2010 гг., с 2011 по март 2021 – поквартальные
Средства, поступившие от реализации туристского продукта (за минусом налога на добавленную стоимость, акцизов и аналогичных обязательных платежей)	Ежегодные данные с 2013 по 2019 год

	Регион	2003 г.	2003 г.	2004 г.	2005 г.	2006 г.	2007 г.	2008 г.	2009 г.	2010 г.	январь–март 2011	январь–июнь 2011	январь–сентябрь 2011	январь–декабрь 2011	2011 г.
1	Российская Федерация	158246455	151018882	154555258	167373138	172041146	172319961	177945166	162772633	162987818	23534143	52977791	94646294	125915081	166197118
2	Центральный федеральный округ	40451931	39424304	40182688	41715827	44790040	42444525	43034226	38454136	39923422	7247843	15637233	25577910	34444845	41089911
3	Белгородская область	774953	617867	692799	718658	797281	905140	886123	793041	853283	118488	304941	523819	688052	778636
4	Брянская область	1117693	987149	979548	948039	790392	792047	1132888	836715	792601	150016	322108	490679	655742	797043
5	Владимирская область	1246776	1094898	1034999	1241849	1369687	1325198	1559655	1237068	1266376	121075	297901	580699	767996	1333338
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Рис. 1. Срез набора данных «Ночевки»

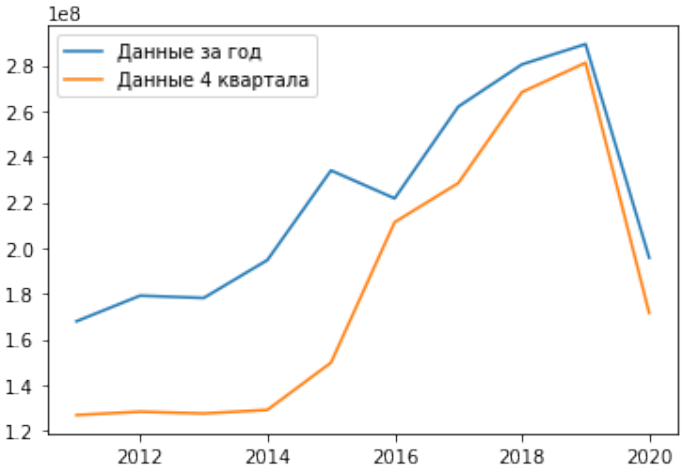

Рис. 2. Расхождение между годовыми и квартальными данными о числе ночевок (все регионы суммарно)

Более серьезная проблема заключается в том, что квартальные данные доступны только с 2011 года, при этом, согласно информации из паспорта набора данных, квартальные сведения за 2011-2015 год приведены без учета субъектов малого предпринимательства, в то время как годовые данные приводятся по всему кругу хозяйствующих субъектов. Это приводит к несовпадению статистики за 12 месяцев (период январь-декабрь) и за год, причем расхождение может быть весьма существенным. Так, в целом по Российской Федерации количество ночевок за период январь-декабрь 2011 года составило 125 915 081, а за 2011г. (годовые данные) – 166 197 118, то есть на 32% больше. Более того, детальный анализ показывает, что несмотря на то, что с 2016 года квартальные данные должны охватывать уже весь круг хозяйствующих субъектов, расхождения по-прежнему сохраняются. Так, например, за период январь-декабрь 2020 года количество ночевок в КСР Российской Федерации составило 167 476 435 шт., а за 2020 год (годовые данные) – 191 175 546 (+14%). Имеющиеся расхождения проиллюстрированы на графике (рис. 2) видно, что после 2016 года разрыв снижается, но не исчезает полностью.

В целом, однако, можно отметить, что, несмотря на погрешность, годовые и квартальные данные после 2016 все же высоко коррелированы (коэффициент линейной корреляции года составляет 98,6 %), а среднее абсолютное процентное отклонение составляет не более 12,5%. Расхождения, вероятно, объясняются различиями в методологии сбора. Данные за год представляются более полными, однако квартальные данные содержат компонент сезонности, который играет значительную роль в туристической отрасли. Кроме того, использование квартальных данных увеличивает объем датасета, что важно для построения моделей машинного обучения. Учитывая изложенное, для дальнейшей работы с датасетом был принят ряд допущений.

Таблица 2

Сезонная структура спроса, Краснодарский Край, 2019 г.

Период	Количество ночевок (квартальные данные)	Доля в общем количестве за год
1 кв. 2018	4 227 488	9,5%
2 кв. 2018	9 971 759	22,5%
3 кв. 2018	21 716 823	48,9%
4 кв. 2018	8 474 171	19,1%
1-4 кв. 2018	44 390 241	100,0%

Таблица 3

Интерполированные данные

Период	Количество ночевок (интерполированные данные)	Доля в общем количестве за год
1 кв. 2018	4 591 326	9,5%
2 кв. 2018	10 829 976	22,5%
3 кв. 2018	23 585 875	48,9%
4 кв. 2018	9 203 498	19,1%
2018 год	48 210 675	100,0%

Таблица 4

Интерполяция данных до 2011 года

Период	Количество ночевок (интерполированные данные)	Доля в ообщем количестве за год
1 кв. 2010	2 639 856	10,8%
2 кв. 2010	5 039 198	20,6%
3 кв. 2010	11 863 899	48,6%
4 кв. 2010	4 867 156	19,9%
2010 год	24 410 109	100,0%

1) Квартальные данные (2011-2020) используются для определения структуры туристского потока в тот или иной регион в течение года. Расчет выполняется отдельно для каждого региона за каждый год. В качестве примера приводятся расчеты для Краснодарского края в 2018 году (табл. 2). Количество ночевок в январе – декабре 2019 года составило 44 390 241.

2) Годовые данные (2011-2020) интерполируются на основе структуры, выявленной по квартальным данным. Это преобразование также выполняется для всех регионов отдельно за каждый год (2011-2020). Так, согласно годовым данным, количество ночевок в Краснодарском крае в 2018 году составило 48 210 675, и это значение, используя рассчитанные пропорции, можно распределить по кварталам следующим образом (табл. 3).

3) Для периодов за 2002-2010 годы, где квартальные данные отсутствуют, интерполяция годовых значений выполняется на основе усредненных пропорций за 2011-2020 годы. Например, для Краснодарского края туристкий поток в 2010 году предположительно распределился следующим образом (табл. 4). Как видно, примерно половина всех ночевок приходится на третий квартал, что характерно для региона, где пляжный отдых является основным видом туризма.

4) Аналогичный подход (интерполяция годовых данных на основе усредненных пропорций) применяется и для записей, относящихся к периоду 2011-2021 гг., в которых по каким-то причинам отсутствуют квартальные данные, но есть годовые – это данные по Республике Крым и г. Севастополю за 2014 год, а также данные по Республике Ингушетия за 2015 год.

Структура датасета «Доходы» полностью повторяет «Ночевки», и для нее характерны те же проблемы. После аналогичных преобразований был получен объединенный датасет, количество строк в котором составило 6545 (рис. 3). Каждая запись в нем содержит значения двух целевых переменных (число ночевок и доходы КСР) определенного региона, обозначаемого кодом, за конкретный период (квартал) с января 2002 по март 2021 г.

Код	Date	Income	Year	Nights
1.0	2002-04-01	1.917269e+04	2002	75260.993793
1.0	2002-07-01	9.910530e+03	2002	47710.478367
1.0	2002-10-01	1.275538e+04	2002	91604.496182
1.0	2003-01-01	1.406150e+04	2002	66213.031659
1.0	2003-04-01	1.731074e+04	2003	64454.131669
...	...	...	...	...

Рис. 3. Датасет с целевыми переменными

Результаты исследования и их обсуждение

Объясняющие переменные

Следующим шагом является наполнение датасета значениями объясняющих переменных (признаков). Следует отметить, что использование названий для обозначения регионов неизбежно приводит к ошибкам технического характера. Особенно это важно при объединении сведений из разных источников, где основным связующим звеном (ключом) выступает регион. Во избежание ошибок было принято решение отказаться от текстовых названий регионов и использовать их коды. За основу взяты коды субъектов Российской Федерации из Приложения N 1 к Порядку заполнения формы «Сведения о внесении в реестр филиалов и представительств международных организаций и иностранных некоммерческих неправительственных организаций, реестр представительств иностранных религиозных организаций, открытых в Российской Федерации, сведений о филиалах, представительствах международных, иностранных некоммерческих неправительственных, иностранных религиозных организаций (об изменениях, вносимых в реестры)», утвержденному приказом ФНС России от 22.05.2019 № ММВ-7-14/259@ [6]. Для индексации Федеральных округов использованы коды «Общероссийского классификатора экономических регионов. ОК 024-95» (утв. Постановлением Госстандарта России от 27.12.1995 № 640) (ред. от 10.02.2021). Структура созданного справочника представлена в таблице (табл. 5) на примере двух регионов – г. Москвы и Республики Чувашия.

Некоторую сложность представляет работа с данными по регионам, территориальная структура которых изменялась с течением времени (например, ряд автономных округов были упразднены в 2007 году). Такие данные требуют ручного редактирования.

Таблица 5

Справочник регионов

Наименование	Код	Код округа	Федеральный округ
Чувашская Республика – Чувашия	21	33	Приволжский федеральный округ
Чувашская	21	33	Приволжский федеральный округ
Chuvashia Republic	21	33	Приволжский федеральный округ
Чувашская республика	21	33	Приволжский федеральный округ
Чувашская Республика	21	33	Приволжский федеральный округ
…	…	…	…
г. Москва	77	30	Центральный федеральный округ
Город Москва столица Российской Федерации город федерального значения	77	30	Центральный федеральный округ
Москва	77	30	Центральный федеральный округ
Moscow	77	30	Центральный федеральный округ
г. Москва	77	30	Центральный федеральный округ

Таблица 6

Датасеты для моделей совокупного потока

Тип модели	Количество переменных	Количество строк (обучающая/тестовая выборки)
Совокупный поток	24	6 258 (5512/746)
Совокупный поток	40	5 458 (4850/608)

После формирования справочника регионов и перехода от текстовых названий к кодам, осуществляется объединение датасетов, содержащих признаки. Для моделирования совокупного потока было сформировано два датасета, характеристики которых приведены ниже (табл. 6). Оба датасета охватывают период с января 2002 года по апрель 2021. Второй датасет содержит меньшее количество строк, однако первоначальный набор переменных в нем дополнен данными о потребительских ценах (тарифах) на товары и услуги и информацией о численности населения.

Список переменных и их описание приводится в таблице ниже (табл. 7). В последнем столбце («Датасет») содержится указание на то, в каком из двух датасетов, содержащих, соответственно, 24 или 40 переменных, используется тот или ной показатель. Столбец «Регион» указывает, берется ли значение признака в целом по стране («Общий») или относится только к региону прибытия («То»).

Данные загружаются из разных источников и в ряде случаев требуют предобработки. В дополнение к выбранным переменным интересно было бы также использовать данные о мероприятиях, проводимых в различных регионах, для моделирования событийного туризма. Однако имеющаяся в открытом доступе информация либо требует сложной дополнительной обработки (например, данные с сайта Eventsinrussia.com [7]) либо содержат малый объем выборки требует перепроверки (так, данные «Реестра событий» за 2015-2018 гг. с портала АИС Туризм [8] содержат такие события, как «Тест», «Тест2» и проч.).

Таблица 7

Описание переменных в моделях совокупного потока

Название переменной	Регион	Описание признака	Датасет
Quater_num_To	Общий	Номер квартала поездки	24;40
Rate	Общий	Курс доллара	24;40
Population	Общий	Население РФ	40
Money	Общий	Среднедушевые доходы в целом по стране	24;40
Бензин автомобильный марки АИ-92, л	Общий	Потребительские цены в среднем по стране	40
Проезд в купейном вагоне скорого нефирменного поезда дальнего следования, в расчете на 100 км пути	Общий		40
Проезд в междугородном автобусе, в расчете на 50 км пути	Общий		40
Полет в салоне экономического класса самолета, в расчете на 1000 км пути	Общий		40
Поездка на отдых в Испанию, поездка	Общий		40
Экскурсионная поездка в Финляндию, поездка	Общий		40
Экскурсионная поездка во Францию, поездка	Общий		40
Экскурсионная поездка на автобусе по городам Европы, поездка	Общий		40
Экскурсионная поездка в Германию, поездка	Общий		40
Поездка в Китай, поездка	Общий		40

Окончание табл. 7

Название переменной	Регион	Описание признака	Датасет
Временная разница с Москвой	To	Временная разница региона, в который едут туристы, с Москвой, в часах – косвенный показатель удаленности региона от центра	24;40
Активный отдых_To	To	Количество достопримечательностей соответствующего типа в регионе, в который едут туристы	24;40
Воинская слава_To	To		24;40
Гастрономический туризм_To	To		24;40
Детский отдых_To	To		24;40
Культура_To	To		24;40
Музеи_To	To		24;40
Необычные места_To	To		24;40
Оздоровительный туризм_To	To		24;40
Охота и рыбалка_To	To		24;40
Пляжный отдых_To	To		24;40
Приключения_To	To		24;40
Природа_To	To		24;40
Развлечения_To	To		24;40
Святыни и храмы_To	To		24;40
Сельский отдых_To	To		24;40
Театры_To	To		24;40
Традиции_To	To		24;40
hotel_places	To	Количество мест в гостиницах в регионе, в который едут туристы	24;40
health_places	To	Количество мест в санаториях в регионе, в который едут туристы	24;40
Обед в ресторане, на 1 человека	To	Потребительские цены в регионе, в который едут туристы	40
Ужин в ресторане, на 1 человека	To		40
Театры, билет	To		40
Музеи и выставки, билет	To		40
Санаторий, день	To		40
Проживание в гостинице, сутки с человека	To		40
Ave_Night	To	Средний доход КСР на одну ночевку (косвенный показатель стоимости ночевки для туриста)	24

Для всех переменных были рассчитана попарная ранговая корреляция Спирмена, а также их корреляция с целевыми переменными (табл. 8). Переменные в таблице упорядочены по абсолютному значение корреляции с переменной «Ночевки». Видно, что некоторые переменные демонтсрируют большую связь с переменной «Доходы», чем с переменной «Ночевки», при этом что большиство переменных имеют положительную связь с объемом туристского потока.

Ранговая корреляция более информативна по сравнению с традиционной линейной корреляцией Пирсона, поскольку сравниваются не сами значения переменных, а их ранги, однако все равно не учитывает сложные взаимодействия между признаками, которые могут быть выявлены моделью машинного обучения.

Таблица 8

Ранговая корреляция с целевыми переменными

Переменная	Ранговая корреляция с переменной Nights (Ночевки)	Ранговая корреляция с переменной Income (Доходы)	Ранговая корреляция с переменной Nights (Ночевки) – абсолютное значение	Ранговая корреляция с переменной Income (Доходы) – абсолютное значение
Nights	1,0000	0,8503	1,0000	0,8503
hotel_places	0,8849	0,8321	0,8849	0,8321
Income	0,8503	1,0000	0,8503	1,0000
health_places	0,8371	0,6407	0,8371	0,6407
Численность населения	0,8094	0,6775	0,8094	0,6775
Музеи	0,6747	0,5798	0,6747	0,5798
Театры	0,5877	0,4914	0,5877	0,4914
Развлечения	0,5321	0,4717	0,5321	0,4717
Детский отдых	0,5101	0,4319	0,5101	0,4319
Оздоровительный туризм	0,4625	0,3805	0,4625	0,3805
Культура	0,4488	0,3895	0,4488	0,3895
Святыни и храмы	0,4385	0,3556	0,4385	0,3556
Активный отдых	0,4204	0,4100	0,4204	0,4100
Приключения	0,4023	0,3614	0,4023	0,3614
Сельский отдых	0,3113	0,2495	0,3113	0,2495
Необычные места	0,2881	0,2483	0,2881	0,2483
Пляжный отдых	0,2698	0,2231	0,2698	0,2231
Театры, билет	0,2624	0,5707	0,2624	0,5707
Воинская слава	0,2535	0,2151	0,2535	0,2151
Музеи и выставки, билет	0,2255	0,5169	0,2255	0,5169
Обед в ресторане, на 1 человека	0,2212	0,4907	0,2212	0,4907
Ужин в ресторане, на 1 человека	0,1823	0,4983	0,1823	0,4983
Санаторий, день	0,1795	0,4941	0,1795	0,4941
Поездка на отдых в Таиланд, поездка	-0,1589	-0,1142	0,1589	0,1142
Охота и рыбалка	0,1491	0,1465	0,1491	0,1465
Традиции	-0,1368	-0,1065	0,1368	0,1065
Проживание в гостинице, сутки с человека	0,1112	0,4133	0,1112	0,4133
Ave_Night	0,0974	0,5376	0,0974	0,5376
Природа	0,0909	0,1102	0,0909	0,1102
Экскурсия автобусная, час	0,0899	0,4211	0,0899	0,4211
Гастрономический туризм	0,0741	0,0560	0,0741	0,0560
Проезд в купейном вагоне скорого нефирменного поезда дальнего следования, в расчете на 100 км пути	0,0537	0,3712	0,0537	0,3712
Поездка на отдых в Турцию, поездка	0,0425	0,3433	0,0425	0,3433
Полет в салоне экономического класса самолета, в расчете на 1000 км пути	0,0383	0,3182	0,0383	0,3182
Экскурсионная поездка на автобусе по городам Европы, поездка	0,0382	0,3584	0,0382	0,3584
Экскурсионная поездка в Финляндию, поездка	0,0356	0,3576	0,0356	0,3576

Окончание табл. 8

Переменная	Ранговая корреляция с переменной Nights (Ночевки)	Ранговая корреляция с переменной Income (Доходы)	Ранговая корреляция с переменной Nights (Ночевки) – абсолютное значение	Ранговая корреляция с переменной Income (Доходы) – абсолютное значение
Поездка в Грецию, поездка	-0,0338	0,0000	0,0338	0,0000
Экскурсионная поездка во Францию, поездка	0,0317	0,3552	0,0317	0,3552
Population	0,0306	0,2258	0,0306	0,2258
Поездка на отдых в Испанию, поездка	0,0301	0,3535	0,0301	0,3535
Экскурсионная поездка в Германию, поездка	0,0279	0,3431	0,0279	0,3431
Поездка в Китай, поездка	0,0276	0,3349	0,0276	0,3349
Quater_num	0,0239	0,0197	0,0239	0,0197
Бензин автомобильный марки АИ-92, л	0,0235	0,3514	0,0235	0,3514
Проезд в междугородном автобусе, в расчете на 50 км пути	0,0205	0,3502	0,0205	0,3502
Money	0,0190	0,3843	0,0190	0,3843
Аренда однокомнатной квартиры у частных лиц, месяц	-0,0038	0,2147	0,0038	0,2147
Rate	-0,0021	0,2540	0,0021	0,2540
Временная разница с Москвой	0,0005	0,0572	0,0005	0,0572

Ярким примером является то, что переменные, отражающие уровень потребительских цен, положительно коррелированы с туристским потоком, хотя, казалось бы, с повышением стоимости отдыха спрос должен снижаться. Однако здесь играют роль и другие факторы – так, помимо непосредсвенно стоимости услуги необходимо учитывать и доступность комплиментарных услуг. Тем не менее, не все их перечисленных в таблице (табл. 7) переменных были включены модель.

Построение модели

Всего для каждой целевой переменной было построено 3 модели. Первоначально использовался датасет с 24 переменными, обучение было выполнено для логарифмированных и не логарифмированных значений переменных. Логарифмирование по основание натурального логарифма часто применяется в случае несимметричных распределений для снижения разброса между самыми большими и самыми маленькими значениями в выборке.

Для каждой из шести моделей были опробованы 8 алгоритмов (табл. 9). Все они относятся к классу алгоритмов для решения задач регрессии – то есть моделирования признаков, значения которых не ограничены каким-либо заранее определенным набором вариантов (в отличие от задачи классификации).

Подробный обзор алгоритмов, их преимущества и недостатки можно найти в книгах Бринка [9] и Соловьева [10]. Для воспроизводимости вычислений во всех случаях, где это применимо, фиксировалось значение random_state – параметра, отвечающего за генерацию случайных чисел.

Выбор наилучшего алгоритма и его гиперпараметров осуществлялся в два этапа. Первоначально для всех 8 алгоритмов, взятых с гиперпараметрами по умолчанию, была выполнена процедура 5-ти слойной кросс-валидации. Кросс-валидация, или перекрестная валидация, преставляет собой процедуру, при которой проводятся эксперименты с несколькими вариантами разбиений исходной обучающей выборки, для каждого из которого определяется качество прогноза на валидационной выбоке.

Таблица 9

Алгоритмы машинного обучения

Алгоритм	Краткое описание
‘LR’: LinearRegression()	Линейная регрессия – классический, наиболее простой и наименее ресурсоемкий алгоритм машинного обучения, подходящий для моделирования взаимосвязей, близких к линейным.
‘SGD’: SGDRegressor (loss=’huber’, random_state=0)	Стохастический градиентный спуск с функцией потерь Хьюбера – модификация линейной регрессии для ускорения расчетов (стохастический градиентный спуск) и повышения устойчивости к выбросам (использование функции потерь Хьюбера)
‘SVR’:SVR(**kernel** =’rbf’)	Метод опорных векторов для регрессии. В отличие от линейной регрессии в машине опорных векторов применяется другой принцип нахождения оптимального решения (прямой или гиперплоскости), а использование специальных ядер (по умолчанию – радиальная базисная функция) дает возможность моделирования нелинейных взаимосвязей.
‘KNN’:KNeighborsRegressor()	Метод ближайших соседей. Прогнозирование выполняется за счет интерполяции известных значений целевой переменной в наблюдениях (строках), похожих на ту, в которой это значение требуется предсказать. Количество похожих строк (ближайших соседей) может быть разным, например, по умолчанию – 5.
‘DT’:DecisionTreeRegressor (random_state=0)	Дерево решений – иерархическая древовидная структура, состоящая из набора правил типа «Если значение признака..., то…». Относится к непараметрическим моделям, то есть построенная зависимость не может быть описана функционально.
‘RF’:RandomForestRegressor (random_state=0)	Случайный лес – ансамбль из большого (по умолчанию -100) количества деревьев решений, каждое из которых обучается независимо на случайно подвыборке данных. Прогноз строится путем усреднения предсказаний отдельных алгоритмов (деревьев решений). Каждое отдельное дерево должно быть достаточно глубоким, то есть содержать большое количество ветвлений, чтобы обеспечить точность прогноза.
‘XGB’: xgboost.XGBRegressor (random_state=0)	Усиленные деревья решений, градиентный бустинг – еще один тип ансамблевых моделей, основанных на деревьях решений. В отличие от случайного леса, деревья строятся не независимо, а последовательно, при этом каждый последующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. Ансамбль также состоит из большого количество деревьев, однако необходимости в большом количестве ветвлений нет.
‘MLP’:MLPRegressor (random_state=0)	Многослойный персептрон – классическая нейронная сеть. По умолчанию содержит 2 слоя, 100 нейронов и нелинейной активационной функцией (“ReLu”)

Закючение

Основная особенность моделей машинного обучение состоит в том, что они способны выявлять сложные плохо формализуемые закономерности, опираясь на большое количество примеров. На данный момент именно отсутствие необходимых для построения модели статистических данных является основной проблемой для применения алгоритмов машинного обучения для прогнозирования туристских потоков в России. Несмотря на большое количество ограничений и допущений, благодаря использования косвенных данных, в настоящем исследовании удалось построить несколько достаточно хороших моделей для прогнозирования квартального количества ночевок и доходов коллективных средств размещения, в том числе между регионами.

Статья написана в рамках НИР «Разработка концепции моделирования рынка туристических услуг России с применением методов экономико-математического моделирования и современных цифровых технологий».

Библиографический список

1. Запись Круглого стола при Аналитическом центре при Правительстве РФ. URL: https://www.youtube.com/watch?v=oEqDRmUuCwc (дата обращения: 07.10.2021).

2. Сайт Правительства РФ. О стратегии развития туристской сферы до 2035 года. URL: http://government.ru/docs/37906/ (дата обращения: 07.10.2021).

3. Сайт Открытые данные Ростуризма. URL: https://opendata.tourism.gov.ru/opendata/6 (дата обращения: 07.10.2021).

4. Сайт Открытые данные Ростуризма, группа показателей «Информация о количестве ночевок в коллективных средствах размещения». URL: https://opendata. tourism.gov.ru/7708550300-numberofovernightstays (дата обращения: 07.10.2021).

5. Сайт Открытые данные Ростуризма, группа показателей «Доходы коллективных средствах размещения от предоставляемых услуг без НДС, акцизов и аналогичных платежей». URL: https://opendata.tourism.gov.ru/7708550300-hotelsincomes (дата обращения: 07.10.2021).

6. Сайт Консультант плюс. URL: http://www.consultant.ru/document/cons_doc_LAW_327257/32556eb7b0d67c6d83dc341a91d4c263e84786a1/ (дата обращения: 07.10.2021).

7. Сайт продвижения туристского продукта России. URL: http://eventsinrussia.com/topevents (дата обращения: 07.10.2021).

8. Сайт АИС Туризм РФ. URL: http://fcp.russia.travel/kaprt/ ?guestMode=#$ROOT$:SplitPanel-78149EAED8C9D435_RCAS9JETBCBAL9KH: 8YSLW8LBMIOE6RNS (дата обращения: 07.10.2021).

9. Сайт Литрес. URL: https://www.litres.ru/dzhozef-richards/mashinnoe-obuchenie-25740052/ (дата обращения: 07.10.2021).

10. Сайт Литрес. URL: https://www.litres.ru/v-i-solovev/analiz-dannyh-v-ekonomike-teoriya-veroyatnostey-prikl-40089868/ (дата обращения: 07.10.2021).