慧科 - 香港科技大学旅游指数

WISERS–HKUST Tourism Index

Dr. Jingxin Zhao[1], Yu Xiong[1], Guancong Ren[1], Dr. Helena Lau[1], Dr. Chao He[1]

Prof. Allen Huang[2], Prof. Kai-Lung Hui[2]

1. Wisers AI
2. HKUST Business School

Last modified on: [2023-07-30]

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

执行摘要

本白皮书介绍了由慧科讯业与香港科技大学商学院合作研究的香港旅游业项目，即基于社交媒体大数据及线下旅游业统计数据而计算得出的实时预测性指数，慧科－香港科技大学旅游指数（简称：慧科－港科大旅游指数）。

香港旅游业是香港四大支柱产业之一。然而，受到 2019 年社会运动和 2020 年新冠疫情的影响，香港的旅游业遭受了前所未有的考验，2023 年 2 月香港与内地通关，旅客往来香港内地无需入境隔离，香港政府也推出一系列鼓励游客到访香港的措施，香港旅游业迎来新的转机，在此之际，分析能够影响未来香港旅游业的因素，不管是对香港政府还是旅游商业，都十分重要。做为香港旅游业的主要市场，在 2019 年，中国内地的旅客占全部访港旅客的 78.29%。此次项目也以中国内地旅客为研究对象。

本项目旨在通过构建一系列慧科－港科大旅游指数来对未来香港旅游业进行预测。不同于以往旅游业学术论文使用旅客问卷或专家小组的方式进行研究，我们首次使用了大数据建模的方式，利用来自慧科社交媒体大数据及线下旅游业统计数据来构建预测模型。

通过对大量以往旅游业相关学术论文的研究，并结合香港旅游业的本地特点，我们定位了 6 大旅游因素，分别为到访方法，设施，景点，娱乐活动，当地社群和附带因素。在慧科全景智能大数据平台上，我们可以获取内地旅客经常使用的社交媒体数据，包括新浪微博数据，线上旅行机构数据，旅行相关论坛数据和其他热门社交媒体的数据，如百度贴吧，马蜂窝，小红书，抖音等。这些媒体的数据总量超过 1000 万条/天。为囊括不同的时期和各类重要事件，我们选择了 2018 年 1 月至 2023 年 3 月做为此次研究的时间范围。通过对这五年多的数据进行分析研究和比较，我们对香港旅游业经历的变化获得了一个系统的认知，也计算出了全面且稳定的一系列预测性慧科－港科大旅游指数。

经过初步的数据清洗和分析后，我们将上述 6 大因素进一步细分为 11 个模型自变量， 分别为到访方法–交通，到访方法–签证，设施，景点，当地社群–正面，当地社群–负面，娱乐活动–购物，娱乐活动–运动，娱乐活动–文娱，娱乐活动–展览，以及附带因素。同时，我们选择了 9 个可能影响香港旅游业的外因变量，分别为公众假期，学生寒暑假期，港币人民币汇率，香港社会运动–实际发生，香港社会运动–内地媒体大量报 导，香港内地通关政策，月份，气温，以及环境–空气质量。我们将对香港酒店入住率、香港访港旅客人数和香港酒店平均价格分别进行分析预测。

本项目利用慧科自主研发的自然语言处理技术和机器学习模型，将以上数据进行处理和量化，最终采用随机森林回归预测模型。同时，我们研究了对不同时间点的未来数据进行预测的效果，得到兼顾预测性和实时更新性的一系列慧科－港科大旅游指数，最终预测结果的误差率可低至 4%。

通过本项目的成果，香港政府及旅游业界将获得对未来香港旅游业相关数字和趋势的精准预测，调整落实紧跟时事的旅游业相关政策，及时更新商业决策及计划。

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

目录

1. 项目任务介绍 4
1.1 背景介绍 4
1.2 研究目标：慧科－港科大旅游指数 4
1.3 研究团队 4
2. 项目方法论 5
2.1 数据理解和采集 5
2.2 变量设计和提取 7
2.3 数据处理和量化 9
2.4 模型建立 10
3. 结果展示 12
3.1 预测结果 13
3.2 模型评估 14
3.3 模型解释 15
3.3.1 酒店入住率预测模型 15
3.3.2 到访人数预测模型 17
3.3.3 酒店价格预测模型 20
3.4 慧科－港科大旅游指数 21
4. 总结 23

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

1. 项目任务介绍

1.1 背景介绍

旅游业是香港的核心产业之一，其中来自中国内地的游客更是占全部访港旅客的绝大多数；2019 年，从中国内地入境香港的旅客人数达到了总旅客人数的 78.29%。

受到疫情封关的影响，相比于 2019 年，2020 年的访港旅客人数骤减 99.8%，香港旅游业遭受严重打击，许多服务于旅游业的香港人失业。 2023 年 2 月香港与内地通关，游客往来相关和内地无需隔离，政府亦推出一系列吸引游客的措施， 在此情况下，能准确的定位和分析影响香港旅游业的因素显得尤为重要。

分析和预测离不开数据的支持，社交媒体平台是内地访港旅客最常用的分享和交流旅行信息的平台，而慧科拥有丰富的社交媒体数据源和成熟的自然语言处理及数据分析技术。在此项目中，我们意在根据社交媒体大数据来构建一个具备预测能力并能实时更新的旅行热度指标，为香港政府和旅游商业洞悉重要信息，帮助决策。

1.2 研究目标：慧科 - 港科大旅游指数

慧科－港科大旅游指数是一个直观便捷，并具备实时月度更新性质的预测性指标。慧科－港科大旅游指数拥有完整的理论框架支持，囊括超过 20 种旅行相关的影响因素（如交通，景点等），依靠用户原创内容作为大数据支持，能够为我们的客户提供准确的预测，帮助客户获得及分析未来的旅行相关数字和趋势（如下个月的旅客到访人数等）。

慧科－港科大旅游指数的主要受众为政府和酒店等旅游相关商业。对于政府，慧科－港科大旅游指数的预测能力能够协助其调整未来的旅游业相关政策并落实到相关部门（如入境处等）。对于旅游相关商业，慧科－港科大旅游指数的实时更新性能够紧跟时事，季节，及假期的变化，为客户提供第一手的预测性商业信息，让其能够根据预测更新或提前安排商业计划。

1.3 研究团队

本项目由慧科AI 部门（Wisers AI）与香港科技大学商业及社会资讯分析研究中心合作完成。

Wisers AI，前身为 Wisers AI Lab（慧科AI 实验室）2014 年 7 月成立于香港，专注于以人工智能技术解决中文全媒体资讯自动化分析与大数据情报挖掘。目前拥有 20 余位毕业于欧美及中港台知名院校的 AI 及计算语言学专家。所有成员均拥有硕士以上学位，其中 35% 的成员拥有博士学位。 Wisers AI 依托慧科 20

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

多年积累的全球数一数二且不断增长的中文全媒体信息数据库及语义资源，以人工智能与大数据技术为驱动，专注研发面向实际应用的开放领域多元化数据 AI 分析技术，实现从跨媒体的海量数据流中及时发掘与识别对客户最重要、最有价值、及最相关的资讯情报。自主研发的全面涵盖自动化媒体情报处理与挖掘的自然语言处理及人工智能技术包括：命名实体识别、情感分析、话题分类、文章聚类、图像识别等，拥有 10 多项国际发明专利及奖项。

香港科技大学商业及社会资讯分析研究中心（CBSA）应用最先进的统计、计量 经济学、机器学习和人工智能工具于分析新兴的大数据趋势以产生商业和社会洞察力，并作为协作大学学者、商业组织、非政府组织和智库的研究人员之间的平台，为专业和普通受众分析商业和社会历史,案例及问题。具体来说中心将收集和分析来自传统媒体、文献、及社交网络的用户生成数据，适时进行行业调查和民意调查及分析以提供创新的商业和社会见解。这些见解和建议将有助于香港和大中华地区的经济和社会发展。

2. 项目方法论

2.1 数据理解和采集

在预测旅游业数据的研究里，最基本也是最重要的步骤之一为定位和分析能够影响未来旅游业数据的关键因素。在此次项目中，慧科团队对大量过往相关文献进行了研究，并注意到，在过去的旅游产业相关学术报告中，尽管不同研究所提出的能够影响旅游业的因素不尽相同，但最为关键的几个因素基本一致。2017 年 Reitsamer 等人以旅客问卷方式进行研究，将到访方法，设施，景点，娱乐与当地社群作为关键因素进行分析。在 Lee 等人于 2010 年的研究中，以专家小组分析的方式，将到访方法，设施，景点和辅助服务作为关键旅游因素。同样以专家小组分析进行研究的，Deng 等人在 2002 年的研究中，将交通，便利设施，景点，当地社群和周边景点定位为能够影响旅游业的关键因素。值得注意的是，在 1998 年 Kim 以旅客调查采访的方式进行的研究中，除去上述因素外， 还将环境洁净程度，季节性景观等因素加入研究；同时将设施，娱乐等因素进一步细化为休闲设施品质，家庭型旅行设施及其安全性等细分因素。而在最早的 1974 年 Gearing 等人的专家小组研究中，则是以较为概括的概念，分析了食物，自然，社会，历史，及购物五大因素。

综上我们可以看到，尽管在过往的学术研究中，时间和空间的跨度很大，但这些研究提出的能够影响旅游业的因素大体类似。在此次项目研究中，我们总结了各文献中通用的旅游业因素，并结合了香港本地特点，定位了六个能够对香港旅游业起到重大影响的关键因素，分别为到访方法，设施，景点，娱乐活动，当地社群，和附带因素。

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

因素名稱	定义	包含内容
到訪方法	旅客到訪香港的方式	旅行簽證,交通
設施	住宿,餐廳,及基礎設施	酒店,餐廳,物價,旅客服務及旅行指示,水電 Wi-Fi 等
景點	自然景觀,人文景點,及季節性景點等	歷史景點,傳統漁村,戶外行山,主題公園,博物館,當地名吃等
娛樂活動	廣义社會文化性旅游資源,所有室内室外活動	购物,體育運動,文化娛樂,展會展覽,夜生活等
當地社羣	訪港旅客對香港本地人/社羣的觀感	安全/治安,當地人友好程度
附帶因素	除去上述五大類外,吸引旅客訪港的附帶因素	其它訪港原因,如注射 HPV 疫苗、购買保險等

在收集数据的方法上，我们注意到文献基本都采用了旅客问卷调查或专家小组分析。这两种方法相对比较简便直观，但无法为一个预测指标来提供更实时客观的信息。随着大数据技术的不断发展，学者们对大数据的应用也展开了丰富的研究。 Song 和Liu (2017)在用大数据分析旅游业的研究中指出，大数据可以提供更可靠的，最前沿的和最实时的信息。 Hendrik 和Perdana (2014)在研究中提出大数据对于分析理解旅客需求具有重要意义。 Varian (2014) 以Google 搜索为例，认为大数据的实时性极大地方便了捕捉消费者行为动态。因此，我们在此次的项目研究中，充分利用了慧科拥有的数据资源，用大数据分析的方法了解预测香港旅游业变化。

首先，我们选择了新浪微博做为我们的第一信源。做为内地最主流的社交媒体，新浪微博无疑是舆论影响的重要平台，同时，新浪微博的相对匿名性也为用户 原创内容提供了更多的空间，我们可以从中更全面地了解内地旅客对香港旅游 的看法。其次，我们亦搜集了专门针对旅游业的社交媒体，如马蜂窝，去哪儿， 携程以及各大旅游论坛。这些线上旅行机构及旅游论坛有专门的目的地版块， 以供网民咨询和分享当地的旅游信息和见闻。最后，我们也注意到在一些其他 流行社交媒体中(如抖音，小红书等)，一些网络意见领袖(KOL)会分享一些旅游目的地，吸引到关注者去旅行。我们将这些媒体也放入到我们的信源列表中。在 上述信源中，每日产生的数据总量超过 1000 万条。

我们将此次研究的时间范围确定为 2018 年 1 月至 2023 年 3 月。在这四年中，香港经历了相对平稳的 2018 年，出现香港社会运动的 2019 年，爆发新冠疫情的2020 年，伴随限聚令和入境强制核酸检测等措施的 2021 年和2022 年，以及迎来全面通关的 2023 年。不论是社会运动还是新冠疫情，都对香港旅游业造成了不可忽视的影响，在社交媒体中也引发了大量讨论，充分保证了数据的多样性。

我们使用关键词、标签和版面过滤的方法，获得 2018 年至 2023 年关于香港旅游的超过 18 万条数据。为方便使用，下文中我们用「Weibo」表示新浪微博数据，用「OTA」表示线上旅行机构、旅游相关论坛及其他流行社交媒体。

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

社交媒體類型	新浪微博	旅行論壇及線上旅行机構	其他流行社交媒體
社交媒體	新浪微博	百度貼吧,馬蜂窩,磨坊論壇,窮游論壇,去哪兒,攜程,知乎	小紅書,哔哩哔哩,抖音
時間范圍	2018 年 1 月至 2023 年 3 月		
數據總量	每日 1000 萬條左右		
過滤方法	關鍵詞/標签過滤	版面過滤	關鍵詞/標签過滤
過滤后數據總量	119,391 條	30,021 條	34,612 條
標注表示方法	Weibo	OTA	

2.2 变量设计和提取

我们根据上一节中定义的六大旅游因素导出此次项目预测模型的自变量。尽管到访方法，设施，景点，娱乐活动，当地社群，和附带因素这六大因素已经能全面的概括影响旅游业的变量，但经过对实际数据的分析之后，我们发现这六大因素所包括的细分因素对旅游业的影响不尽相同。例如当地社群因素可以进一步细分为当地社群–正面（友好）以及当地社群–负面（敌视），而这两个细分因素将会很明显地为预测结果带来相反的影响。从数据本身来看，我们也发现某些同一大类下的细分因素并不存在较高的线性相关性。因此，我们将这六大因素做更为细化的拆分，形成 11 个自变量，并同时应用于「Weibo」和「OTA」数据。 11 个自变量为到访方法–交通，到访方法–签证，设施，景点，当地社群–正面，当地社群–负面，娱乐活动–购物，娱乐活动–运动，娱乐活动–文娱，娱乐活动–展览，以及附带因素。综上所述，媒体数据源方面，我们总共获得 11 * 2 = 22 个自变量。

除了社交媒体数据，我们亦考虑到一些可能影响到旅游业，但不受或较少受旅游业影响的外生变量。这些外生变量来自一些非社交媒体的数据源（即非「Weibo」 和「OTA」 数据的数据源，如香港天文台。）经过多轮分析和筛选，我们最终选出了 9 个我们认为与香港旅游业密切相关的外生自变量，分别为公众假期， 学生寒暑假期，港币人民币汇率，香港社会运动–实际发生，香港社会运动–内 地媒体大量报导，香港内地通关政策，月份，气温，以及环境–空气质量。

細分變量名稱	細分變量定义	細分變量性質
到訪方法-交通	訪港旅客到訪香港時的交通方式	淨值變量,社交媒體上與訪港交通相關的月度數據量
到訪方法-签證	訪港旅客到訪香港時的相關签證	淨值變量,社交媒體上與訪港签證相關的月度數據量
設施	訪港旅客的住宿,餐廳,及基礎設施	淨值變量,社交媒體上與旅游設施相關的月度數據量
景點	訪港旅客在香港到訪的景點	淨值變量,社交媒體上與香港旅游景點相關的月度數據量
當地社羣-正面	訪港旅客對香港本地	淨值變量,社交媒體上對香港本

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

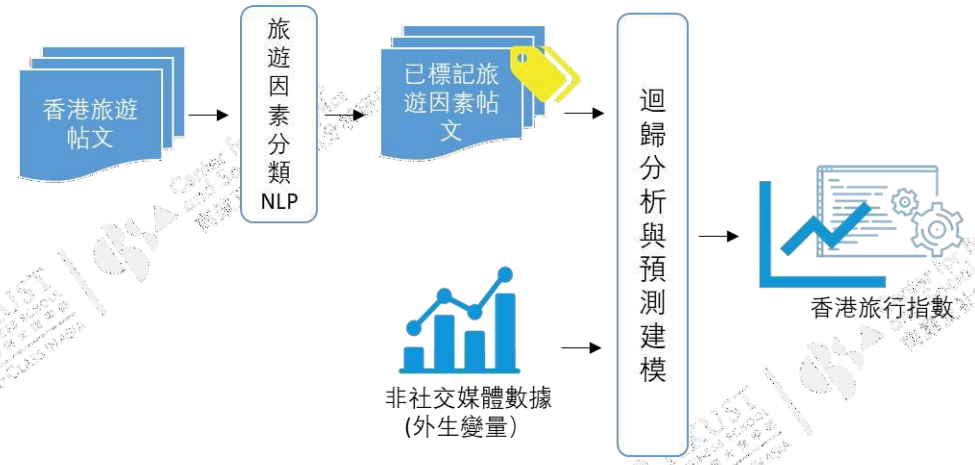
	社羣的正面觀感	地社羣正面觀感的月度數據量
當地社羣-負面	訪港旅客對香港本地社羣的負面觀感	淨值變量,社交媒體上對香港本地社羣負面觀感的月度數據量
娛樂活動-購物	訪港旅客在香港的購物活動	淨值變量,社交媒體上與在港購物相關的月度數據量
娛樂活動-運動	訪港旅客在香港的體育類活動	淨值變量,社交媒體上與在港體育活動相關的月度數據量
娛樂活動-文娛	訪港旅客在香港的文娛類活動	淨值變量,社交媒體上與在港文娛相關的月度數據量
娛樂活動-展覽	訪港旅客在香港的文化展覽類活動	淨值變量,社交媒體上與展覽相關的月度數據量
附帶因素	除去上述 10 个變量外,吸引旅客訪港的附帶因素	淨值變量,社交媒體與其它訪港原因相關的月度數據,如注射 HPV 疫苗、购買保險等

外生變量名稱	外生變量定义	外生變量性質
公衆假期	内地公共假期時間	01 變量,如該月有内地公共假期則爲 1(5 月有五一勞動節公共假期),否則爲 0
學生寒暑假期	内地學生的寒暑假期時間	01 變量,如該月有内地學生寒暑假期則爲【如 6 月爲學生暑假】 否則爲 0
港幣人民幣汇率	港幣兌人民幣月度平均汇率	淨值變量,汇率隨時間變動的月度時間序列
香港社會運動-實際發生	香港社會運動的實際發生時間	01 變量,如該月有實際發生香港社會運動則爲 1(2019 年 7 月), 否則爲 0
香港社會運動-内地媒體大量報導	内地媒體大量報導香港社會運動的時間	01 變量,如該月内地媒體有大量報導香港社會運動則爲 1(2019 年 8 月),否則爲 0
香港内地通關政策	受疫情影響香港内地是否通關	01 變量,如該月内香港内地由于疫情原因尚未通關則爲 1,否則爲 0
氣溫	月度平均氣溫	淨值變量,平均氣溫隨時間變動的月度時間序列
環境-空氣質量	月度平均 pm2.5 指數	淨值變量,平均 pm2.5 指數隨時間變動的月度時間序列
月份	當時月份	淨值變量

© 2023 Wisers Information Limited

The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

2.3 数据处理和量化



从社交媒体中爬取到同香港旅游有关的帖文后，我们应用慧科自主研发的自然语言处理技术（NLP）---文本话题分类，通过机器学习自动分析帖文内容有提及哪些旅游因素，并打上相应标签。

在同一篇帖文中，很可能会提及多个旅游因素，如下图中的帖文，既提及了与购物相关的优惠减价活动，也提及了打卡主题景点，那么在「景点」和「娱乐活动–购物」这两类因素中，都包括这一篇帖文。

旅遊因素類別	與帖文相關度
娛樂活動 – 購物	0.7955
景點	0.2047

在一段时间内（以一个月为例），会有许多帖文提及不同的旅游因素。对于每一类旅游因素，我们可以计算此因素在这个月有被多少篇帖文提及，这样我们便获得了每一类旅游因素的月度数据。对于外生变量，我们可以在相应的官方网站上，获取月度数据。

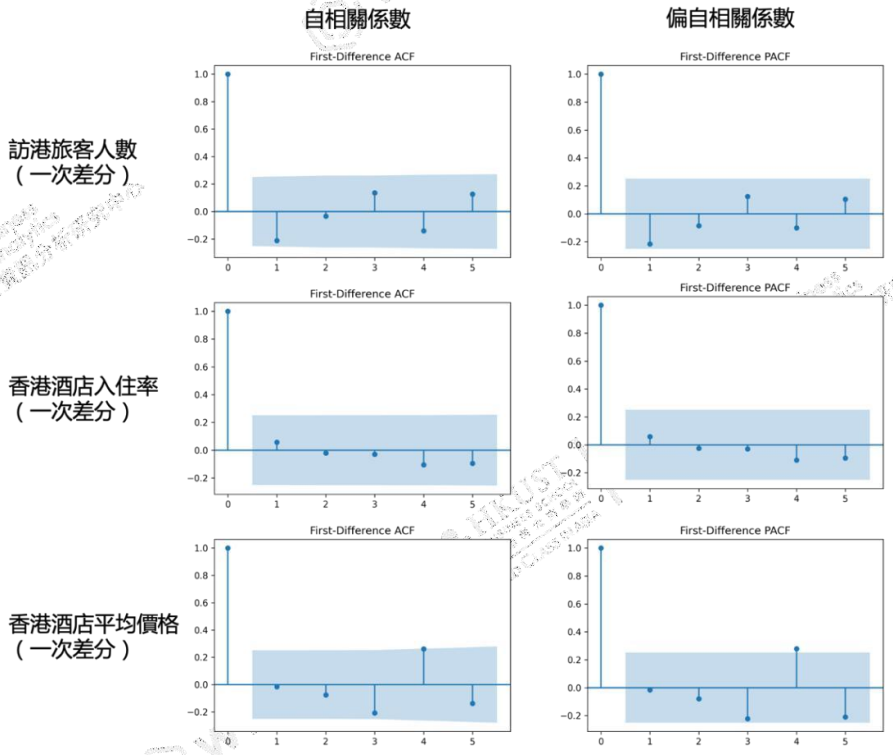
© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

在上文中提到的慧科自主研发的AI 文本话题分类技术，以慧科超过 20 年积累的大数据和行业知识辞典，通过深度学习训练出的词向量模型为基础，对于给定的文本，计算词向量，话题向量和文本向量间的相关度，获得文本在各话题上的相关度分布，来最终确定与文本相关的话题。

2.4 模型建立

由于量化之后的数据（自变量）都和我们想要预测的旅游业统计数字（因变量）都是时间序列，我们首先考虑建立一个时间序列模型。但是，利用一次差分将其平稳化后，我们发现所有自相关和偏自相关都不显著。如果直接用一次差分的数据做线性回归，结果也并不理想。

© 2023 Wisers Information Limited



The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

此外，基于线性假设的预测模型，尽管可解释性强，但是仅能关注有无影响和其正负性，无法考虑非线性关系，精确度不足，而覆杂机器学习模型的引入可以大大提高精度。 随机森林(Random Forest) 模型是一种基于决策树的算法，通过对大量决策树的汇总提高了模型的预测精度，不需要顾虑一般回归分析面临的多元共线性的问题，通过算法自动发现具有关联性的变量。另外，随机森林便于计算变量的非线性作用，而且可以体现变量间的交互作用(Interaction)，即一个自变量 X1 的变化导致另一个自变量 X2 对 Y 的作用发生改变，这种作用在其他模型中(如Logistic Regression) 因其覆杂性经常被忽略。对于集成学习方法，效果虽好，但一直无法解决可解释性的问题（整个模型有 N 棵决策树组成），属于“黑盒模型”，解释性较差，只能验证某一指标体系有效性，对单个指标不能提供具体解释。为此，与模型无关的解释性机器学习被提出，大大拓展了其应用场景。SHAP 是最新解释性机器学习方法之一，是一种可以将单个样本的预测结果描述为所有特征效应之和的归因方法，与其它解释性方法相比其最大优势在于既可以进行基于单个样本的局部解释也可以进行基于全体样本的全局解释。

有鉴于此，本研究基于旅游数据进行定量研究，先使用随机森林模型对未来趋势进行预测，再使用 SHAP 解释方法对模型结果进行分析，探索不同变量对于预测值的影响大小以及其背后的原因，为香港旅游业在运营策略和政策制定方面的优化提供依据。

基于获取到的 2018 年 1 月到 2023 年 3 月的月度数据，我们分别对访港旅客人数，香港酒店入住率，及香港酒店平均价格进行了预测建模。

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

3. 结果展示

在本节中，我们将展示对三种不同目标预测的结果，分别为：

1. 对未来香港酒店入住率的预测
2. 对未来内地到访香港旅客人数的预测
3. 对未来香港酒店平均价格的预测

慧科拥有全景智能大数据平台，每天实时汇聚千万条各类媒体数据，并结合数十个行业的知识图谱，为海量数据进行结构化处理，可以在我们构建模型及预测结果时，提供最实时的社交媒体数据。但香港旅游业数字的统计因为存在实际发生–数据生成–数据披露的时间差，无法实时更新应用到预测模型中。

举例来说，假设现在时间为2021年12月10日，那么慧科的大数据平台已经完成了2021年11月的所有社交媒体数据的收集和处理。然而对于香港旅游业统计数字，虽然理论上，2021年11月的旅游业数据在2020年12月1日之前就已经结算，但实际上，我们现在只能够拿到2021年10月的旅游业数据，拿到2021年11月旅游业数据的最早时间为2021年12月末。总结来说，最早能拿到X月旅游业统计数据的时间为X+1月末。

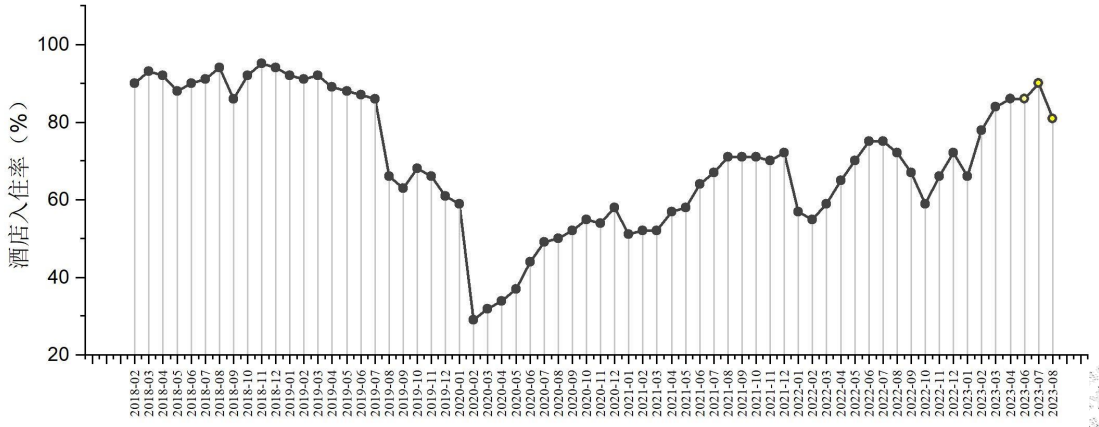
无论是慧科的社交媒体大数据，还是香港旅游业统计数据，都是慧科一港科大旅游指数不可或缺的关键因素，而上文提到的旅游业数据更新时间差会对预测结果造成影响。因此，我们计算出了不同时间点的预测结果，既能够保持慧科一港科大旅游指数的预测性， 也可以不断更新校准预测结果，最小化数据时间差给模型带来的影响。

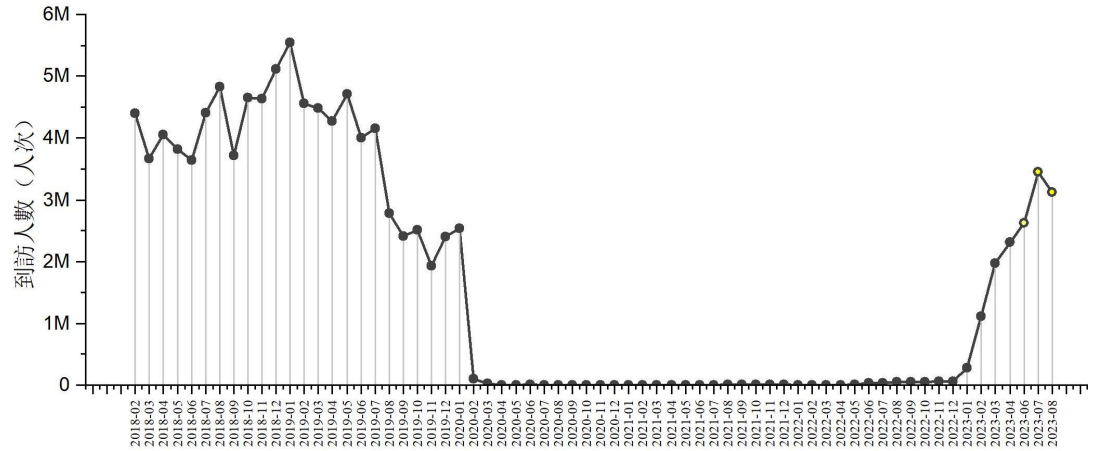
在X月10日，慧科大数据平台已产出X–1月的社交媒体数据，能拿到的旅游业数据为（在X–1月末拿到的）X–2月的数据。所以我们将使用X–1月的社交媒体数据和X–2月的旅游业数据，对X月(本月)，X+1月（下月）和X+2月（下下月）进行预测。

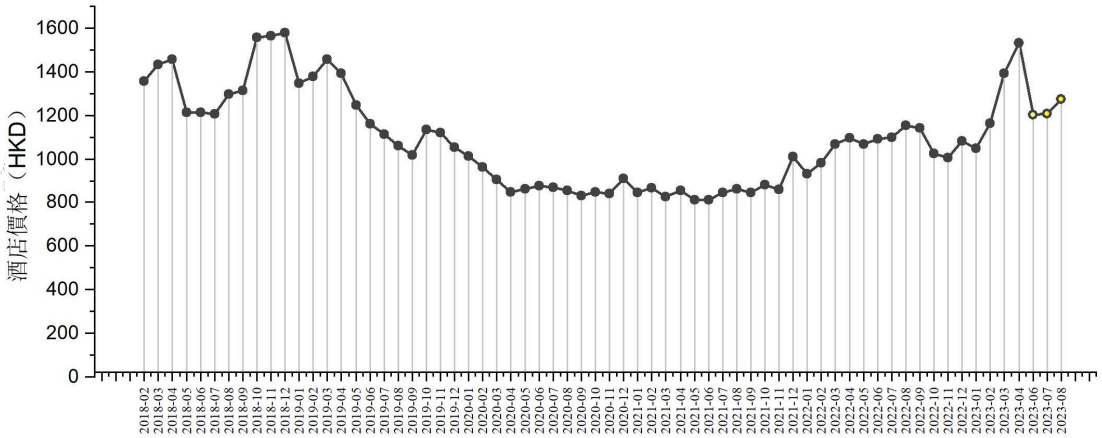
预测时间点	预测目标	可使用的慧科社交媒体数据	可使用的香港旅游业数据
X月10日	X月（本月）	X–1月	X–2月
X月10日	X+1月（下月）	X–1月	X–2月
X月10日	X+2月（下下月）	X–1月	X–2月

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

3.1 预测结果





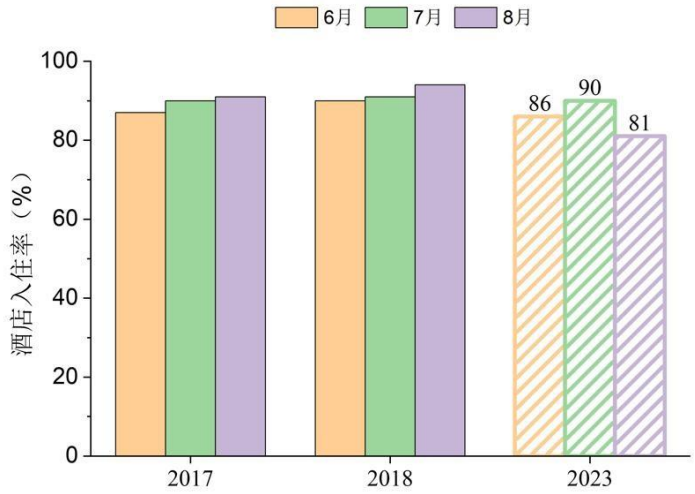


© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

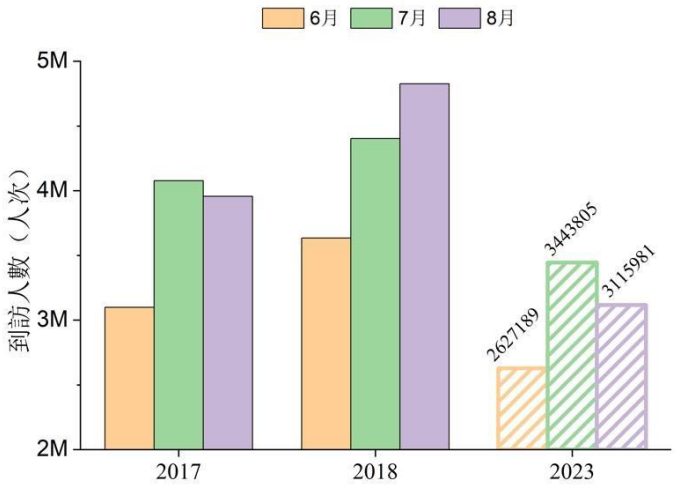
3.2 模型评估

模型	MASE	MAPE	SMAPE
预测本月酒店入住率	0.20	0.06	0.06
预测下月酒店入住率	0.30	0.08	0.08
预测下下月酒店入住率	0.18	0.05	0.05
预测本月到访人数	0.12	0.14	0.19
预测下月到访人数	0.14	0.16	0.18
预测下下月到访人数	0.18	0.19	0.23
预测本月酒店价格	0.12	0.04	0.04
预测下月酒店价格	0.30	0.06	0.06
预测下下月酒店价格	0.23	0.05	0.05

酒店入住率 6、7、8 月预测值与往年同期酒店入住率对比：

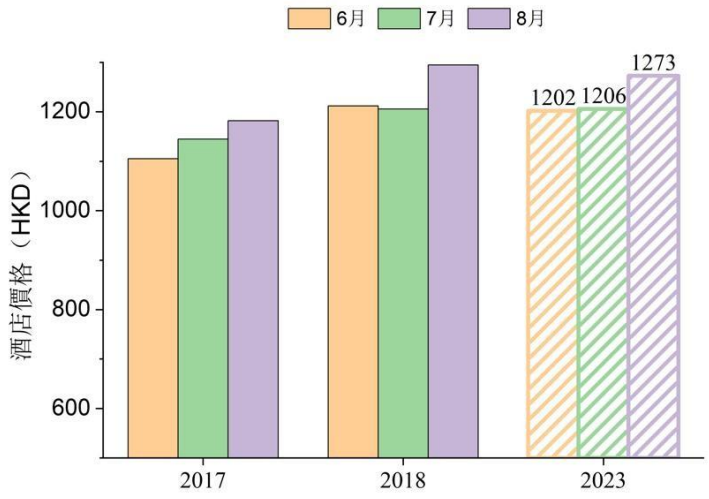


到访人数 6、7、8 月预测值与往年同期到访人数对比：



© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

酒店价格 6、7、8 月预测值与往年同期酒店价格对比：



3.3 模型解释

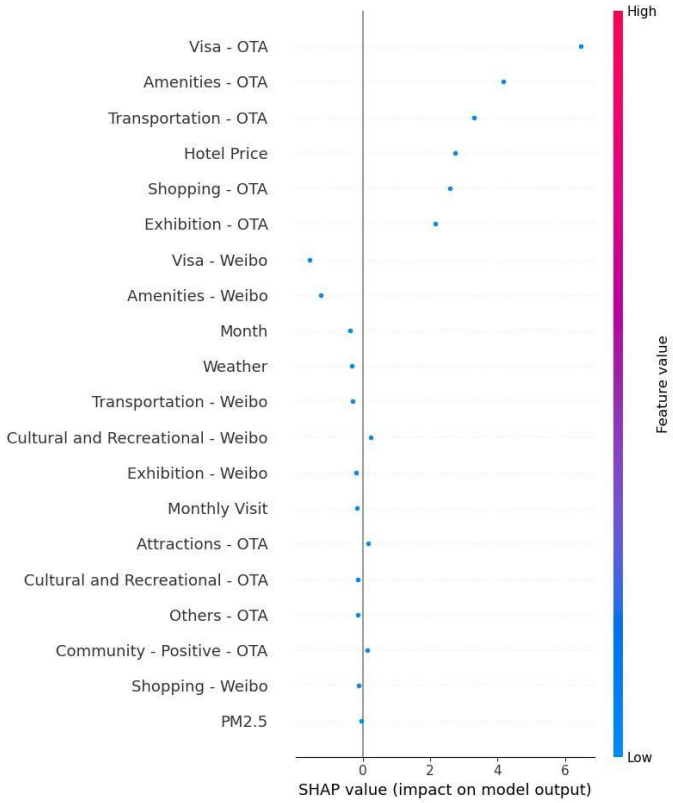
特征影响全局解释主要根据 SHAP value 分析模型中每个特征对模型预测值产生的重

要性和影响正负性，各个特征的 SHAP value 根据所有样本的 Shapley value 绝对值的平均值求出。下图反映了特征重要性排名和特征对模型预测值的影响正负性，散点图中的点表示每个样本，每一行表示模型的一个特征，横坐标表示 SHAP value 的值，数值越大表示特征对于该样本的预测结果影响越大，正值代表正向影响，负值代表负向影响。

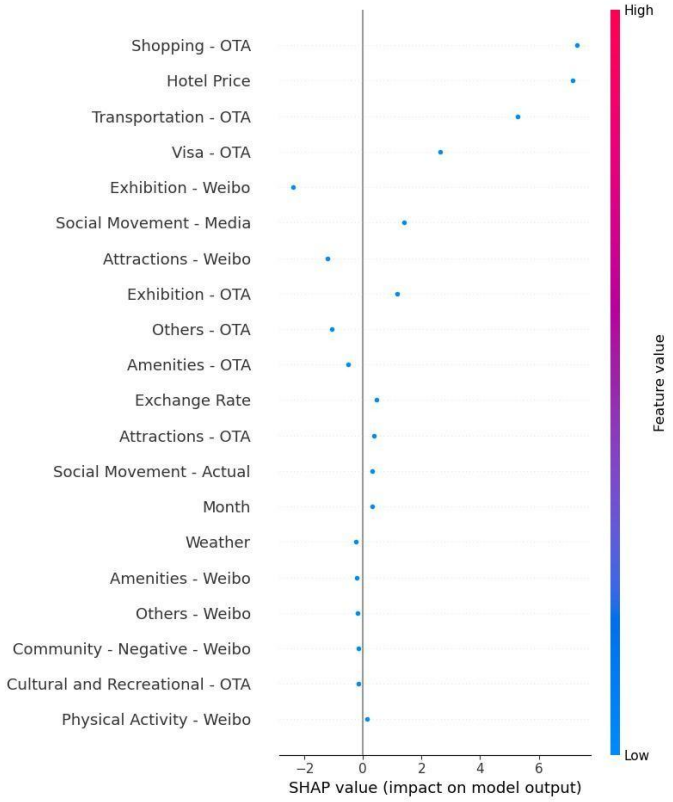
© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

3.3.1 酒店入住率预测模型

预测本月酒店入住率模型中各因子对于预测值的影响

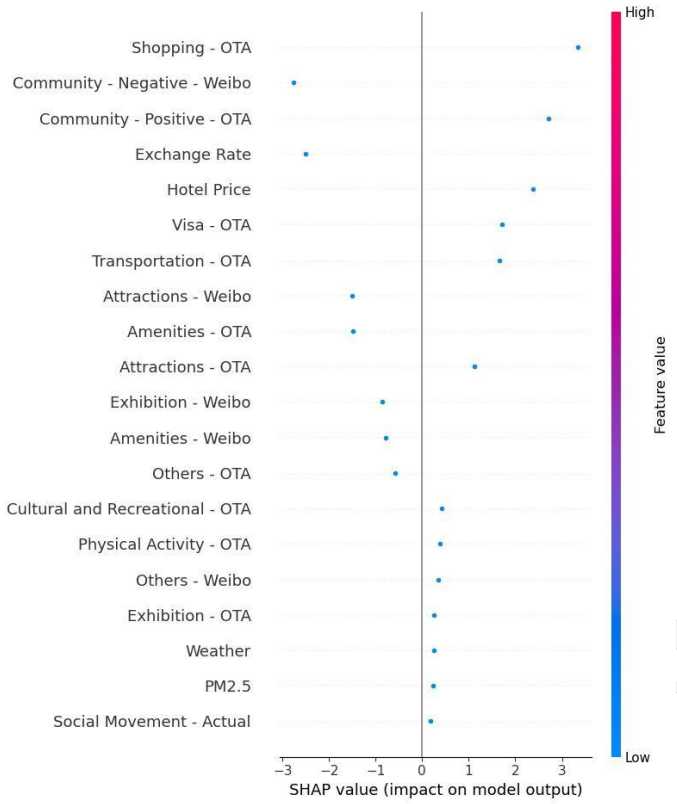


预测下月酒店入住率模型中各因子对于预测值的影响



© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

预测下下月酒店入住率模型中各因子对于预测值的影响

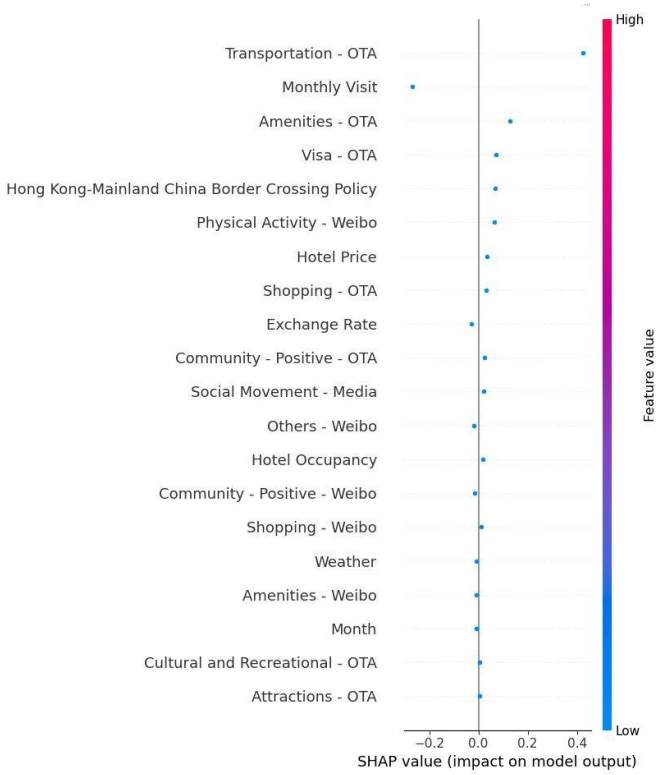


我们可以看到本月酒店入住率预测模型中最重要的特征为 OTA 社交媒体上对于签证的讨论，对与酒店入住率呈现正向影响。下月和下下月酒店入住率的预测模型中， OTA 社交媒体上对于购物的讨论是最重要影响因素，呈现对酒店入住率的正向影响。在三个模型中，酒店价格与酒店入住率呈现正相关，这一表现符合旅游高峰期酒店 价格上涨的趋势。预测结果显示 8 月酒店入住率出现下降，SHAP value 表明其主要受到 Weibo 平台对访港旅客对香港本地社群的负面观感的讨论和汇率的影响。

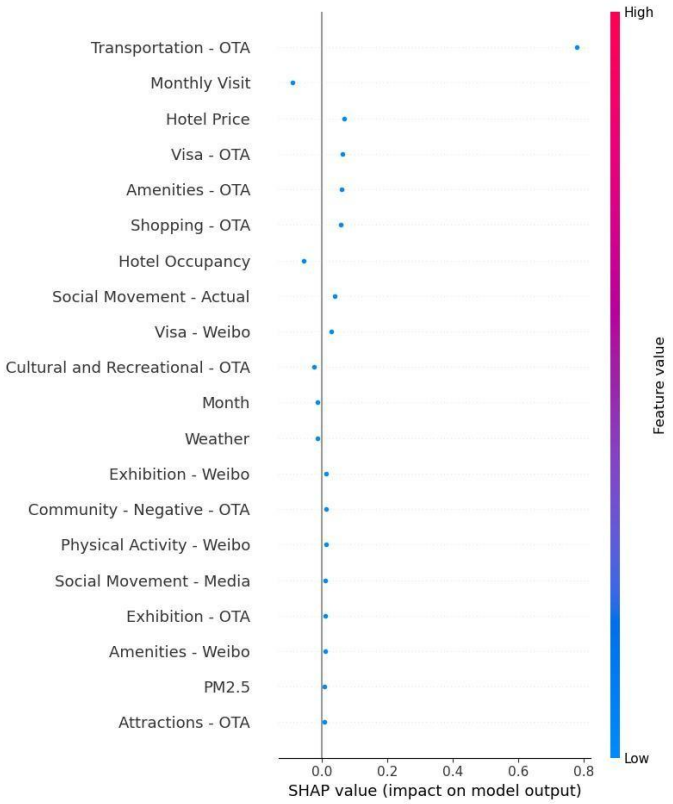
© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

3.3.2 到访人数预测模型

预测本月到访人数模型中各因子对于预测值的影响

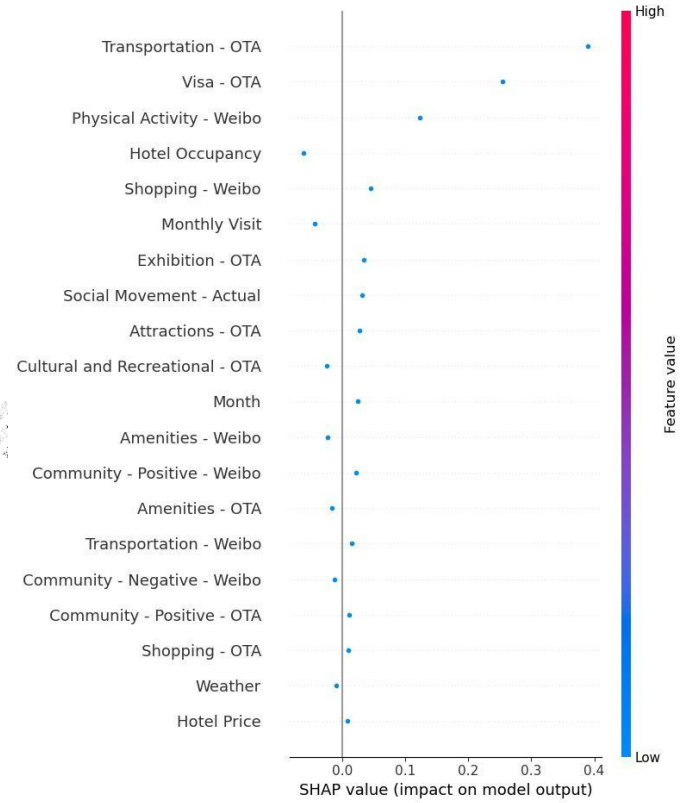


预测下月到访人数模型中各因子对于预测值的影响



© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

预测下下月到访人数模型中各因子对于预测值的影响

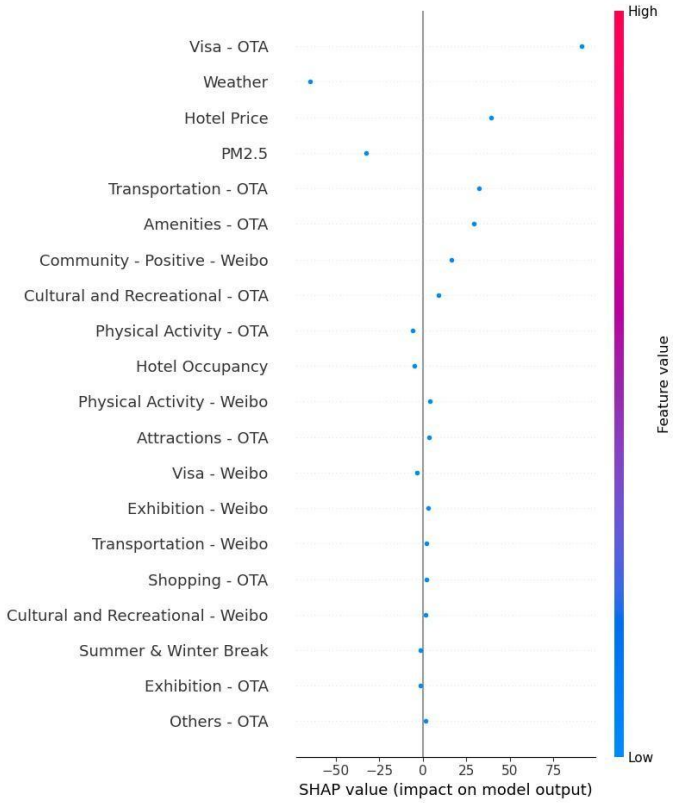


在预测本月、下月、下下月的到访人数模型中，OTA 社交媒体关于交通的讨论均为最重要的特种，呈现与到访人数正相关。

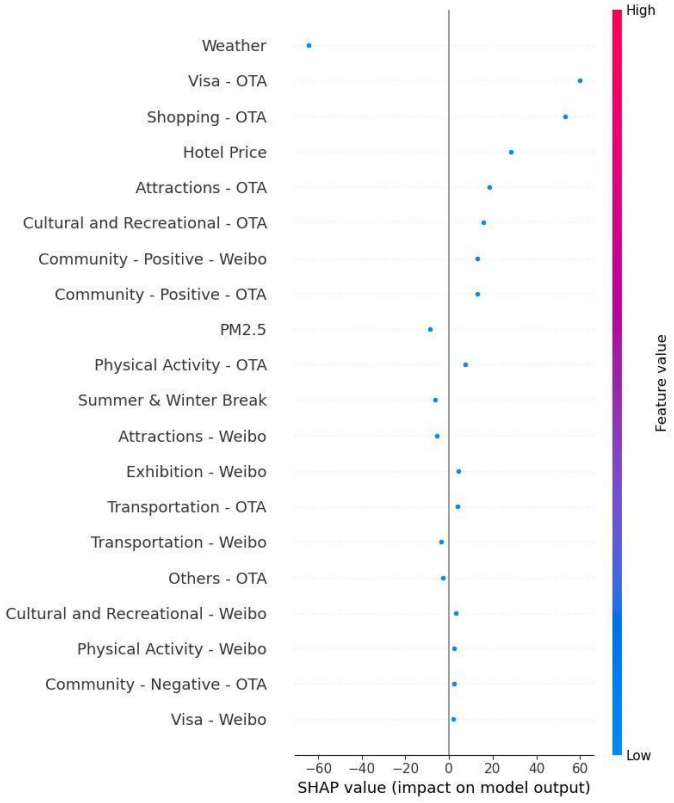
© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

3.3.3 酒店价格预测模型

预测本月酒店价格模型中各因子对于预测值的影响

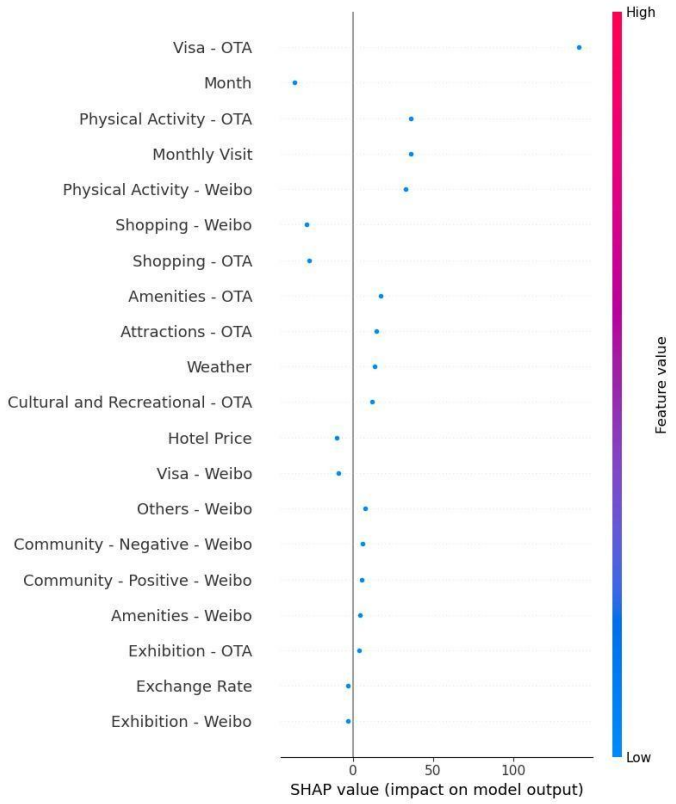


预测下月酒店价格模型中各因子对于预测值的影响



© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

预测下下月酒店价格模型中各因子对于预测值的影响



对于本月、下月和下下月酒店价格的预测，OTA 社交媒体关于签证的讨论度始终是重要影响因素，其呈现与酒店价格正相关。

3.4 慧科 - 港科大旅游指数

为了能够反映并记录香港旅游热度随时间的变化，我们将上述的回归预测模型进行标准化处理，构建出慧科一港科大旅游指数。

慧科一港科大旅游指数由三部分组成，即「慧科一港科大旅游指数：到访人数」，「慧科一港科大旅游指数：酒店入住率」和「慧科一港科大旅游指数：酒店价格」。我们以 2018 年的平均数做为基准点 100 点，分别计算这三部分的每月指数，以反映香港旅游市场不同方面随时间的变化。

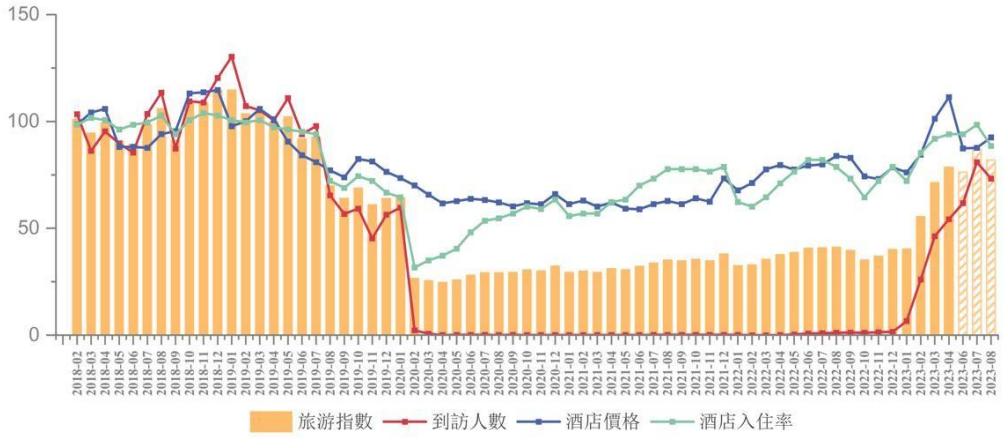
同时，我们将这三个成分指数进行加权平均，以得到一个可广义描述香港旅游热度的指数，「慧科一港科大旅游指数：综合指数」。在本白皮书的「慧科一港 科大旅游指数：综合指数」的计算中，我们分别给予「慧科一港科大旅游指数： 到访人数」，「慧科一港科大旅游指数：酒店入住率」和「慧科一港科大旅游指 数：酒店价格」50%，25%和 25%的权重。在实际应用中，使用者亦可根据需要对权重进行调整。

在每月的 10 号，我们可以公布上个月的最终各项指数，并对本月、下月及下下月的各项指数进行预测。如上图，在 2023 年 5 月 10 日，我们可更新 2023 年 3

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

月最终实际的「慧科一港科大旅游指数：到访人数」，「慧科一港科大旅游指数：酒店入住率」，「慧科一港科大旅游指数：酒店价格」及加权平均后的「慧科一港科大旅游指数：综合指数」。另外，我们对2023年5月、2023年6月和2023年7月的各项指数进行了预测，2023年二季度的香港旅行气氛预计逐渐升温。

慧科一港科大旅游指数可做为香港旅游市场经济活动与政府政策规划的一个重要参考指标。2018年访港旅客总数达6515万人次，与入境旅游相关的总消费为3317亿港元，占本地生产总值约10%。慧科一港科大旅游指数以2018年为基准点100点，大幅高于100的旅游指数表明香港旅游市场非常活跃，相关商业可增加旅游商品和相关服务人员的准备，政府则应加强相关商业及公共秩序的维护；大幅低于100的旅游指数表明香港旅游市场氛围惨淡，相关商业可策划各类促销宣传活动以吸引消费，政府可推出相关政策给予产业扶持。



© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.

4. 总结

在此项目中，我们利用自然语言处理技术和大数据技术，并综合慧科社交媒体大数据和线下旅游业数据，构建出可预测未来香港旅游业数据的一系列慧科－港科大旅游指数。通过慧科－港科大旅游指数，香港政府和旅游相关商业可判断未来香港旅游业的可能情况，调整资源和策略。

慧科－港科大旅游指数具有以下优势：

- 模型框架的设计参考了大量旅游业相关研究文献，并结合了香港本地特点，总结出 6 大旅游因素。
- 利用大数据方法获取数据进行研究，保证了数据的客观性、多样性和实时性。
- 考虑了实际生产中数据披露的延迟情况，研究了对不同时间点的预测结果。最终预测结果的误差率可达 4%，低于朴素预测法 (Naive approach) 误差率 80%左右。

目前，香港入境事务处已经开始披露每天的入境人数，同时新冠疫情和香港–内地通关情况已逐步缓和，所以在未来的研究中，我们可以获得更多的数据，并将尝试进行每周甚至每日的预测。另外，如果我们能够与香港政府或香港旅游业界有更加紧密的合作，那么数据披露延迟问题也可得到缓解，使预测更加及时准确。

欲了解更多信息，请联系
jessicazhao@wisers.com
williamwong@wisers.com

© 2023 Wisers Information Limited
The information contained herein is confidential and proprietary and should not be disclosed, copied or duplicated in any manner without written permission of Wisers Information Limited.