Modeling and Simulation 建模与仿真, 2021, 10(4), 1023-1031
Published Online November 2021 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2021.104102

Hans汉斯

基于随机森林算法的鳄梨价格预测

陈梦凡，张　涛*

广西科技大学，广西　柳州

收稿日期：2021年9月19日；录用日期：2021年10月29日；发布日期：2021年11月5日

摘　要

为了能够更好地预测鳄梨的价格走向趋势，解决在大量特征和大数据下价格预测精度低的问题。本研究在随机的基础上提出了一种基于Pearson系数的随机森林新的组合模型方法。首先，利用Pearson系数进行相关性检验，来进行特征筛选；对随机森林参数调优；最后利用剩余特征进行建模回归预测，并得出最终结论。实验结果表明：改进后的随机森林预测值的平均绝对误差(MAE)和均方误差(MSE)都得到了较大的提高。经研究发现，本文建立的新的组合模型，可以实现对鳄梨价格的短期预测，并且可以达到不错的预测效果。

关键词

鳄梨价格，决策树，随机森林，预测

Avocado Price Prediction Based on Random Forest Algorithm

Mengfan Chen, Tao Zhang*

Guangxi University of Science and Technology, Liuzhou Guangxi

Received: Sep. 19th, 2021; accepted: Oct. 29th, 2021; published: Nov. 5th, 2021

Abstract

In order to be able to better predict prices to the trend of avocado, and solve the problem of low price prediction accuracy under a large number of features and big data, this study on the basis of random puts forward a random forest new combination model based on coefficient of Pearson method. Firstly, Pearson coefficient was used for correlation test to carry out feature screening; tuning random forest parameters; finally, residual features were used for modeling regression

*通讯作者。

文章引用: 陈梦凡, 张涛. 基于随机森林算法的鳄梨价格预测[J]. 建模与仿真, 2021, 10(4): 1023-1031.
DOI: 10.12677/mos.2021.104102

prediction, and the final conclusion was drawn. The experimental results show that the improved random forest predicted mean absolute error (MAE) and mean square error (MSE) got improved. The study found that through a new portfolio model, this paper can realize the avocado price short-term prediction, and can achieve good prediction effect.

Keywords

Avocado Prices, Decision Tree, Random Forest, Prediction

1. 引言

目前随着科学技术的迅速发展，数据库的信息量也在不断增大，随之提升的还有数据挖掘的技术，即人们从大量繁琐而复杂的信息中收集数据的能力也得到了很大的提高[1]。大量数据的获取和存储非常方便，但若想从这些数据中发现潜在的规律，预测未来的发展趋势是存在一定困难的。因此如何有效地利用这庞大的数据，并从中分析数据的价值成为当下研究的热点之一。

据报道，从 2013 年到 2020 年七年的时间内，美国市场上的鳄梨数量翻了两番，达到了 123.1 万吨。在此期间，欧洲鳄梨市场的规模增长了 5 倍，达到了 66.7 万吨，增长主要是由秘鲁和其他鳄梨来源国推动的[2]。2020 年，欧洲的人均鳄梨消费量为 1.11 公斤，而令人难以置信的是 2003 年欧洲的鳄梨人均销量竟然只有区区 0.23 公斤。尽管欧洲的鳄梨消费量也一直在稳步增长，但仍仅为美国的三分之一。2003 年，美国的人均消费量为 1.05 公斤，目前的鳄梨人均销量为 3.71 公斤。2019 年哈斯鳄梨消费研究(Haas Avocado Consumption Study)显示，尽管在这样的情况下，仍有近 50%的美国家庭不吃鳄梨，这意味着如果加以推广美国鳄梨市场的销量可能翻倍[3]。然而，要增加产量和保持价格稳定，是存在一定难度的，仍有很多工作要做。如果欧洲的消费量与美国相当，那么欧洲市场的规模将增加两倍，达到近 200 万吨[4]。但是要实现这一目标将是非常困难的，需要做好推广工作。在这样的背景下，对鳄梨价格的预测就显得尤为重要了。本文便根据往年的鳄梨价格数据，建立了相应的模型，进行了对未来鳄梨价格趋势的预测。

目前国内外价格预测的研究主要有以下两个方向：

1) 市场模拟间接预测方法，利用市场价格由供需关系和市场主体行为形成的机制，通过预测市场的供给和需求情况模拟市场交易得到市场价格，这种方法考虑系统条件和约束来模拟实际市场情况，可以得到更加符合实际的有效的出清价格信息[5]。但是要由市场模拟得到精确的市场价格预测结果，市场的基础数据必须充分详细准确，但目前市场的数据规模巨大、复杂，同时精确预测是目前仍待研究的问题，因此市场模拟间接预测的方法难以应用到实际生活中。

2) 数据分析直接预测方法，该方法利用大量数据分析挖掘市场价格的自身规律以及和其他相关因素的数据联系并进行数学建模，得到市场价格预测模型，以数学建模方法不同的主要有时间序列分析法、多元回归方法、人工神经网络、支持向量机、组合方法等类别。

为了得到一种具有较强通用性和较高预测效果的价格预测方法，通过阅读文献发现近年来部分学者将随机森林回归方法应用在股票价格预测等领域，取得了较好的效果，而股票价格预测与水果市场价格预测具有较强的相似性，同时大量的研究表明随机森林模型的泛化能力很强、对输入数据的误差不敏感

且具备分析输入特征重要程度等优点，有很强的通用性，因此本文利用随机森林方法对鳄梨市场价格进行预测。

2. 数据来源及预处理

本文使用的数据来自 kaggle 数据库(https://www.kaggle.com)，该数据集是关于美国各个地区的鳄梨价格数据，一共包含了 18,249 行 13 列数据，13 列数据分别对应着 13 个变量，其中每列数据分别为 Date (日期)、Average Price (平均价格)、Total Volume (总成交量)、4046 (PLU 为 4046 的鳄梨售出总数)、4225 (PLU 为 4225 的鳄梨售出总数)、4770 (PLU 为 4770 的鳄梨售出总数)、Total Bags (总包)、Small Bags (小包)、Large Bags (大包)、XLarge Bags (超大包)、type (类型)、year (年)和 region (地区)，由于数据量过大，本文只展示原始数据的部分情况，在 Python 软件中实现显示数据前几行的操作，如表 1 所示，显示的是原始数据的前五行。

Table 1. The original data
表 1. 原始数据

	Date	Average Price	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	X Large Bags	type	year	region
0	2015/12/27	1.33	64,236.62	1036.74	54,454.85	48.16	8696.87	8603.62	93.25	0	conventional	2015	Albany
1	2015/12/20	1.35	54,876.98	674.28	44,638.81	58.33	9505.56	9408.07	97.49	0	conventional	2015	Albany
2	2015/12/13	0.93	118,220.22	794.7	10,9149.7	130.5	8145.35	8042.21	103.14	0	conventional	2015	Albany
3	2015/12/6	1.08	78,992.15	1132	71,976.41	72.58	5811.16	5677.4	133.76	0	conventional	2015	Albany
4	2015/11/29	1.28	51,039.6	941.48	43,838.39	75.78	6183.95	5986.26	197.69	0	conventional	2015	Albany
5	2015/11/22	1.26	55,979.78	1184.27	48,067.99	43.61	6683.91	6556.47	127.44	0	conventional	2015	Albany

通过对数据的查看，我们可以发现，其中鳄梨的类型(type)分为两类，conventional (传统的鳄梨)和 organic (有机的鳄梨)，地区(region)分为美国的 54 个地区，分别为 Albany，Atlanta，Baltimore Washington，Boise，Boston，Buffalo Rochester，California，Charlotte，Chicago，Cincinnati Dayton，Columbus，Dallas Ft Worth，Denver，Detroit，Grand Rapids，Great Lakes，Harrisburg Scranton，Hartford Springfield，Houston，Indianapolis，Jacksonville，Las Vegas，Los Angeles，Louisville，Miami Ft Lauderdale，Midsouth，Nashville，New Orleans Mobile，New York，Northeast，Northern New England，Orlando，Philadelphia，Phoenix Tucson，Pittsburgh，Plains，Portland，Raleigh Greensboro，Richmond Norfolk，Roanoke，Sacramento，San Diego，San Francisco，Seattle，South Carolina，South Central，Southeast，Spokane，St Louis，Syracuse，Tampa，Total US，West，West Tex New Mexico。

通过对传统鳄梨和有机鳄梨的销量及平均价格的关系绘制相应的图形，得到的分布图如图 1 所示，从分布图中可以看出，有机的鳄梨平均价格相较传统的鳄梨价格要偏高一些，而且有机鳄梨的销售量也要比传统鳄梨的销售量高一些，由此可以看出，随着全球经济的不断发展，人们的生活水平也在不断的提升，追求生活品质的同时人们对食品的质量要求也在不断提高，人们普遍认为有机食品是更健康的选择，越来越多的人开始追求食用有机食品。因此，有机鳄梨尽管价格要高于传统鳄梨，但其销量却也能高于传统鳄梨。

再将不同地区的两种鳄梨平均价格情况用条形图展示出来，其分布如图 2 所示，从图中我们可以看出有机的鳄梨价格偏高一些，其中 San Francisco 和 Hartford Springfield 以及 New York 这三个地区的无论是传统的还是有机的鳄梨价格相较于其他地区都要更高一些。这可能是由不同地区的经济发展状况存在一定的

差异导致的。而 San Francisco 和 Hartford Springfield 及 New York 这三个地区的经济水平要高于其他地区，所以可能会造成其物价也高于其他地区的情况，毕竟不同地区的物价和当地的经济发展状况是息息相关的。

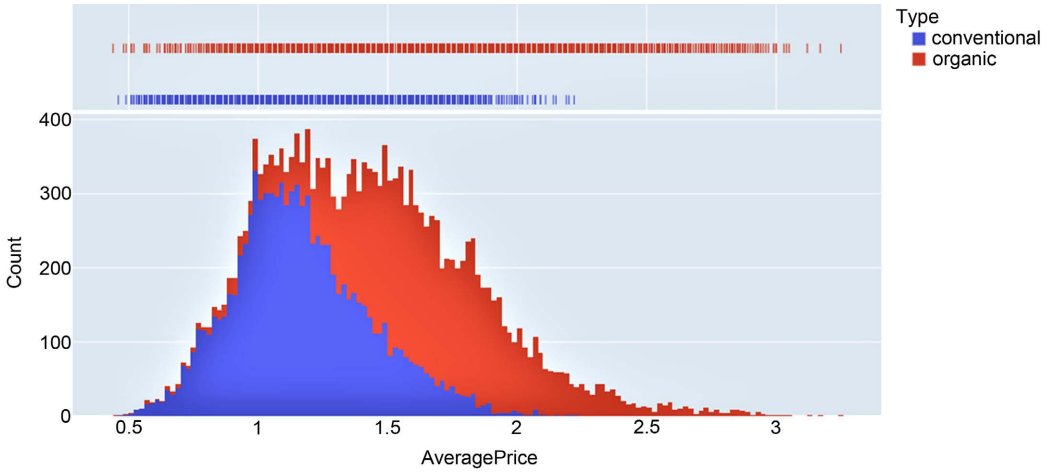

Figure 1. The average price distribution of avocados
图 1. 鳄梨的平均价格分布

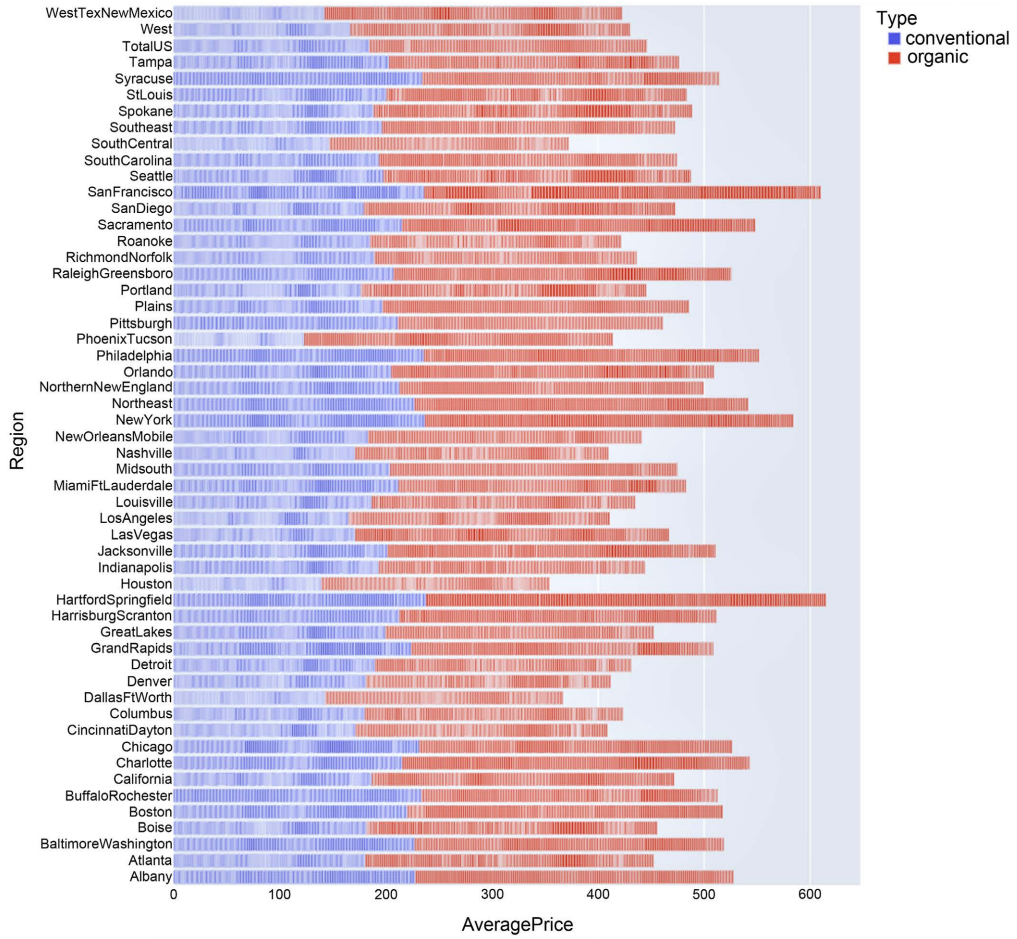

Figure 2. Average price distribution of avocados by region
图 2. 不同地区的鳄梨平均价格分布图

3. 数据建模

3.1. 模型介绍

决策树算法是一种用于进行决策分析的方法，其通过对概率的计算，并进行判断最终获得的所计算的概率期望值大于等于零的情况，以此来评估在已知项目存在的风险和项目的可行性的情形下，事情发生的概率及各种的问题出现的可能性[6]。这是一种很直观的方法，其使用概率来对事物进行分析，并最终通过图形展示出来。因为它的绘制方式与树的分支类似，所以将这个用来决策的分支称为决策树。决策树学习在资料探勘的过程中也十分常见。在这个过程中我们可以采用递归式的方法来进行树的剪枝。在剪枝过程中如果遇到不能再进行分割，或者已经形成一个单独的可以被应用于某一分支的类的，就表明该递归过程已完成了。而随机森林分类器实际上就是应用许多结合起来的决策树的分类结果，从而达到提升分类的正确率的目的[7] [8] [9]。下面是决策树算法的具体介绍。

决策树由根节点，叶子节点以及非叶子节点组成[10] [11] [12]。通过对训练集进行回归分析，生成从根节点到叶子节点的路径分析出路径规则。根据路径规则对新数据进行分类或预测。CART 是基于信息熵，通过 Gini 系数最小原则指标来进行节点分裂，对训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 所输入得空间划分相应的区域，利用递归式的方法将每个样本划入其所对应的区域，并以此得出确定的输出值，其算法步骤如下：

假设自变量特征为 j，该特征的取值为 s。假设取值 s 将特征 j 的空间划分两个区域，其表达式如下：

$$R_1(j,s)=\left\{x\middle|x^{(j)}\le s\right\}$$

1) 依次遍历计算每个切分点 (j,s) 的损失函数(loss function, LF)，并选取损失函数最小的切分点。

$$LF=\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

其中，c_1，c_2 分别为 R_1，R_2 区间内的输出平均值。

2) 将划分的两部分迚行计算切点，依次进行，直到不能继续划分。

3) 将输入空间划分成 M 个部分 $R_1,R_2,\cdots,R_M$，最终生成的决策树为：

$$f(x)=\sum_{m=1}^{M}c_mI(x\in R_m)$$

而 Pearson 相关系数是取值范围在−1 和 1 之间的一种用来描述自变量与因变量之间的相关程度大小的方法，相关性大小可通过 Pearson 相关系数的绝对值来进行比较，其绝对值越大，则代表相关性越强。大于 0 代表正相关，小于 0 代表负相关[13]。其计算公式为：

$$\frac{\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}}{\sqrt{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}}}$$

其中 x_i 是自变量，y_i 是因变量。

随机森林属于算法原理就是由多个弱分类器组合成一个强分类器[14]。其采用 bootstrap 抽样方法来对训练集进行随机有放回的抽样，这里假设抽取 m 个样本，并在 bagging 的基础上对每棵决策树进行随

机特征的选择，然后即可对这 m 个样本分别建立决策树模型，在决策树模型中先进行一次分类，可以得到 m 个分类结果，再对这 m 个结果进行投票，选择最高票数的分类结果作为最终的随机森林预测类别。随机森林算法的步骤分为如下几步：首先输入训练集 D，利用 bootstrap 抽样形成 k 个训练子集 D_k，再从原始全部特征中随机抽取 m 个特征，并利用特征对训练子集 D_k 进行训练，然后将随机选择的 m 个特征做出最优切分，分别得出这 k 颗决策树的预测结果，最后根据决策树分类得到的 k 个预测结果，进行预测结果的投票操作，将得票数最高的分类结果作为最终的随机森林预测类别。

为了提高随机森林算法的预测效果，本文尝试将特征选择方法与改进的网格搜索法相结合，从而实现对鳄梨价格的预测。在建立模型之前首先利用 Pearson 特征选择方法删除部分无关的数据特征，再利用改进的网格搜索法对决策树的参数进行调优，选择适当的参数进行建模，通过优化后的 k 颗决策树所构成的随机森林来得到预测结果。其算法过程如图 3 所示。

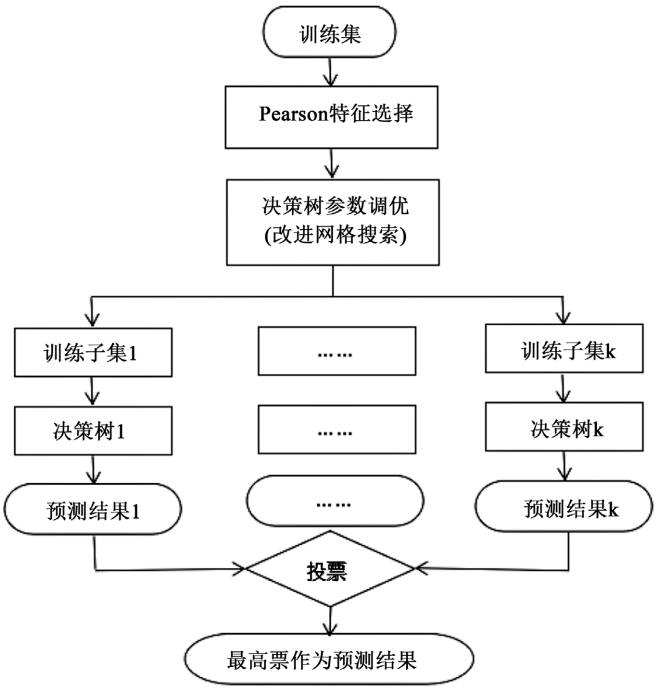

Figure 3. Random forest model based on Pearson feature selection
图 3. 基于 Pearson 特征选择的随机森林模型

3.2. 建立模型

在建立模型之前先对部分变量名称进行调整，将“4046”重新命名为“Small/Med Hass”，将“4225”重新命名为“Large Hass”，将“4770”重新命名为“Extra Large Hass”，然后通过求各变量之间的相关性系数，来判断变量之间的关系。

相关性系数是介于[−1, +1]之间的。当相关性系数介于[−1, 0]之间时，表明变量呈负相关；当相关性系数介于[0, 1]之间时，表明变量呈正相关关系；当相关性系数为 0 时，变量之间不存在相关性。相关性系数越接近 1，变量之间的相关性越强，相关系数越接近 0，则认为变量之间的相关性很弱。当相关性系数的绝对值处于不同的区间范围时，对应有不同的说法，当其介于 0.1~0.3 之间时，认为两变量呈弱相关；当其介于 0.3~0.5 之间时，认为存在中度相关；当其大于 0.5 时，则认为两变量具有强相关性。

将各变量之间的相关关系绘制成热能图如图 4 所示。从图中我们不难发现，本研究中平均价格和年

份与其他变量的相关性不是特别的强，而其余变量之间却均存在很强的相关性，原因主要是由于这些其他的变量均与销售量的联系十分紧密。

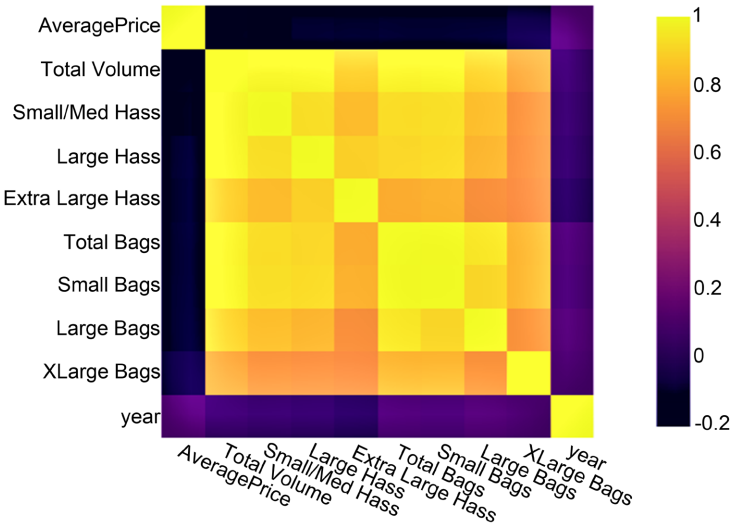

Figure 4. A heat diagram of the correlation between variables
图 4. 各变量之间的相关关系热能图

在建立模型之前本文首先对数据进行了特征的筛选，通过软件实现，选择得分最高的四个变量作为特征变量，分别为“Total Volume”(总成交量)，“Total Bags”(总包)，“type”(类型)，“region”(地区)。选择这几个特征变量之后需要进行数据集的划分，划分数据集之后对决策树进行调参，参数调整过程中 max_leaf_nodes 值对应的 Mean Absolute Error 值如表 2 所示。从表中可以看出，在当 max_leaf_nodes 值由 5 变 50 时 Mean Absolute Error 值减少了 0.033，当 max_leaf_nodes 值由 50 变 500 时 Mean Absolute Error 值减少了 0.05，而当 max_leaf_nodes 值由 500 变 5000 时 Mean Absolute Error 值仅减少了 0.003，变化并不大，说明此时提升 max_leaf_nodes 值效果不大，因此最终选择 max_leaf_nodes 值为 500。

Table 2. Decision tree callback
表 2. 决策树调参

max_leaf_nodes	5	50	500	5000
Mean Absolute Error	0.331	0.298	0.248	0.245

本研究分别采用了两种方法来对鳄梨价格进行预测，两种方法分别为决策树算法和随机森林算法，并对两种算法的预测结果进行了比较。其中决策树的 Mean Absolute Error 值为 0.247，决定系数 R^2 值为 0.646，而随机森林得到的 Mean Absolute Error 值为 0.118，决定系数 R^2 值达到了 0.820。通过 Mean Absolute Error 值和决定系数 R^2 的对比结果来看随机森林的预测效果都明显优于决策树。采用五折交叉验证将数据集划分为五份，选取一份作为测试集，另外四份作为训练集，并重复五次，每次选取不同的训练集。对每次训练都求出平均绝对误差，并队伍此轮显得平均绝对误差求均值，结果如表 3 所示。

通过以上预测结果的平均误差的对比，我们可以看出本文所采用的经过特征筛选后的随机森林模型取得了较高的预测精度，较相同条件下的决策树平均误差而言相对减少了 12.9%。由多次测试的误差分布来看，随机森林模型的预测效果较为稳定，误差的标准差也比较小，而决策树的多次预测误差标准差比较大，预测效果的波动较大，认为其预测能力不够稳定。不难发现，经过特征筛选的随机森林模型的

预测精度相比一般模型有一定的提高，说明了特征筛选对提高模型的预测误差具有正向作用。另外从预测模型的训练时间来看，我们发现随机森林模型所需时间更短，具有更高的计算效率，更加适应大规模样本的训练。综上所述，从整体效果来看，经过特征筛选的随机森林回归模型具有较强的优越性。

Table 3. Five fold cross validation results
表 3. 五折交叉验证结果

训练	训练 1	训练 2	训练 3	训练 4	训练 5	平均值
MAE	0.148	0.199	0.217	0.274	0.330	0.233

4. 结论与建议

本文将随机森林回归应用到鳄梨的价格预测中，利用随机森林特征重要度分析的功能对输入的特征进行筛选，从而降低了无关变量的干扰，进一步提高了模型的预测精度，根据鳄梨市场数据特点构建出价格预测模型。本文所采用经过特征筛选后的随机森林模型得到的 Mean Absolute Error 值为 0.118，决定系数 R^2 值达到了 0.820，该模型取得了较高的预测精度，较相同条件下的决策树平均误差而言相对减少了 12.9%。由多次测试的误差分布来看，随机森林模型的预测效果比较稳定，误差的标准差也比较小，通过验证分析得到的结果也表明本文建立的预测模型具有较高的稳定性和预测精度。因此本文提出的基于随机森林回归的特征筛选的预测模型是一种通用性较强且可行的思路和方法。

经研究分析，认为基于随机森林的预测模型具有较高效果的原因主要为：随机森林回归的子模型决策回归树具有简单的树结构能更好地适应市场价格和实时影响的大量输入；装袋算法这种集成学习方法的应用，避免了模型过拟合，相对其他模型减小了模型的泛化误差。因此进一步研究可以考虑：1) 结合随机森林回归树模型能同时接受输入多种类型数据的特点，充分考虑市场中各类难以量化的因素对市场价格的影响进行预测；2) 利用随机森林的装袋算法，以其他市场价格预测模型作为子模型构造新的集成学习模型并测试其预测效果。

同时，通过本文的研究，也能为除了鳄梨以外的其他商品提供价格预测的建议及方向，即在进行价格预测的时候可以考虑多种方法结合以提高预测模型的预测精度。在进行预测时，特征变量的选择也十分重要，选择合适的特征变量可以提高预测的精确度，甚至还能够加快模型的预测速度，从而达到提高模型的泛化能力及预测效率的目的。

参考文献

[1] 王鹏. 苹果品质评估与价格预测模型研究[D]: [硕士学位论文]. 泰安: 山东农业大学, 2019.

[2] Decker, B.K., Kelly, M.B., Mikolic, J., Walker, J.D. and Clancy, C.J. (2018) 1726 A Random Forest Prediction Model Accurately Identifies Periods at Increased Risk for Positive Legionella Cultures in a Hospital Water Distribution System. *Open Forum Infectious Diseases*, **5**, S54-S55. https://doi.org/10.1093/ofid/ofy209.132

[3] Ließ, M., Hitziger, M., Huwe, B. and Bradley, R.L. (2014) The Sloping Mire Soil-Landscape of Southern Ecuador: Influence of Predictor Resolution and Model Tuning on Random Forest Predictions. *Applied and Environmental Soil Science*, **2014**, Article ID 603132. https://doi.org/10.1155/2014/603132

[4] 王章章. 基于机器学习的价格预测模型研究与实现[D]: [硕士学位论文]. 西安: 长安大学, 2018.

[5] 乔麟婷. 决策树算法研究[J]. 课程教育研究, 2018(48): 224-225.

[6] 安葳鹏, 尚家泽. 决策树 C4.5 算法的改进与分析[J]. 计算机工程与应用, 2019, 55(12): 169-173.

[7] Zhang, H.z., Joshua, Z., Dan, N. and Nordman, D.J. (2020) Random Forest Prediction Intervals. *The American Statistician*, **74**, 392-406. https://doi.org/10.1080/00031305.2019.1585288

[8] Zhang, Q.-Y., Aires-de-Sousa, J. (2007) Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *Journal of Chemical Information and Modeling*, **47**, 1-8. https://doi.org/10.1021/ci050520j

[9] 李超. 机器学习模型在股票价格时间序列分析中的应用与比较[J]. 电子世界, 2021(9): 66-70.

[10] 张家通. 基于数据挖掘的笔记本电脑价格影响因素研究——以京东商城为例[J]. 信息技术与信息化, 2021(1): 58-62.

[11] 单文煜, 吴垠, 陈鹏. 基于机器学习的机票价格预测研究[J]. 现代计算机, 2020(22): 35-38.

[12] 焦岑. 基于随机森林与神经网络的汽车价格影响因素的研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2020.

[13] 曹洁. 基于随机森林模型的二手车价值评估研究[D]: [硕士学位论文]. 石家庄: 河北经贸大学, 2020.

[14] 万威. 灰色预测模型在材料价格预测中的应用[J]. 江苏科技信息, 2021, 38(15): 44-46.