Research on COE Program with Machine Learning Algorithms

Jiashi Li
Department of Statistics and Applied Probability, UC Santa Barbara, CA, U.S.A.

Keywords: COE Coffee, Machine Learning, One-Hot Coding, Hedonic Pricing Theory.

Abstract: The market price of COE coffee depends on its heterogeneity characteristics, and hedonic model theory is the dominant research approach for coffee prices. In this paper, the correlation between coffee prices and the New York C-futures index is first verified. Further, machine learning algorithms such as Support Vector Regression (SVR), Multilayer Perceptron (MLP) models are used to study the factors affecting prices. A more general coding of coffee types is designed and improved the one-hot coding of the regression model. The performance shows that the improved model is better in terms of performance. The prediction accuracy of the model improved by 24.65% after generalized coding of coffee categories. The study further explores the implementation of the hedonic pricing theory.

1 INTRODUCTION

The growing demands for specialty coffee have led to a rapidly growing market for specialty coffee in many countries. The percentage of adults who consume specialty coffee has increased in recent years. Specialty coffee quality is a key factor in stabilizing market development (Traore, 2018, Wilson, 2018, Fields, 2018).

Many Latin American countries participate in the Cup of Excellence (COE) program. Every year an auction of coffee is held. A jury tastes the coffee based on samples of brown beans submitted by the farms. Each cup is given a score from 0 to 100, those scoring 84 or more quality points in the competition are awarded the prestigious Cup of Excellence Award (Bacon, 2004). The winning coffees are ranked according to their scores, and the coffee with the highest score in a given category is awarded first place, followed by the highest quality.

Scholars have used extensively COE dataset to predict the price of specialty coffee. They studied the role of product differentiation and quality production in the world coffee market (Teuber, 2010, Ferreira, 2016, Liska, 2016, Cirillo, 2016). There is a growing literature on the relationship between coffee quality and regional environmental characteristics, especially for the so-called specialty coffees, but consumer price analysis can provide useful information on coffee quality differences. A comprehensive analysis is possible if the datasets can cover both objective and subjective quality attributes.

Donnet analyzes the importance of sensory and reputational attributes in the origin markets (Donnet, 2007, Weatherspoon, 2007, Hoehn, 2007). They found that country of origin effect is evident except sensory quality and scores.

The hedonic price model has been used to study the relationship between prices and attributes of agriculture, food, and real estate. This model is inspired by Waugh's publication "Quality Factors Affecting Vegetable Prices article and the work of Rosen. This approach is used to measure and analyze the contribution of a product's attributes (Hu, 2019, He, 2019, Han, 2019, Gu, 2011, Zhu, 2011, Jiang, 2011). The price of the product is usually modeled as a parameter. Thus, the regression model is used to predict the quality fraction and the price based on various attributes.

However, achieving healthy, sustainable price is a challenge. The framework presented in this paper integrates machine learning and consumption models. Aggregate multiple regressions on various subsamples of the dataset to improve the prediction accuracy. Improving prediction accuracy and thus controlling overfitting. Tree regression algorithms are considered as non-linear, non-parametric methods with high generalization.

Through the analysis of historical data, the original purpose of establishing COE coffee is validated. It is to keep the futures price from affecting

Li, J.
Research on COE Program with Machine Learning Algorithms.
DOI: 10.5220/0011191800003440
In *Proceedings of the International Conference on Big Data Economy and Digital Management (BDEDM 2022)*, pages 563-568
ISBN: 978-989-758-593-7

the price of coffee, so as not to affect the income of the farms and to ensure the stability of the coffee market. Further, the historical data is modeled and analyzed using machine learning algorithms. The factors that have a strong influence on the price are explored to provide suggestions for a good operation of COE coffee in the future.

This paper is structured as follows: Section 2 describes the boutique coffee market and the fundamentals of the Hedonic method. Section 3 introduces the one-hot coding approach with a generalization that improves the accuracy of the regression model. Simulation experimental results are analyzed and modeling implications of the improved model and the classical and regression models: support vector regression and multilayer perceptron models are compared in Section 4. Finally, the conclusion is made in Section 5.

2 RELATED WORKS

2.1 The Hedonic Price Model

The hedonic price model is used to determine the effect of the substance and its properties on the quality fraction or price of specialty coffee. Through hedonic price analysis, the intrinsic and symbolic attributes of each buyer for various coffees can be expressed through price (Niklas, 2020, Rinke, 2020, Oladunni, 2017, Sharma, 2017).

Regression models are commonly used to predict the quality score and price of coffee. It is used to predict the quality fraction and price of coffee based on various attributes.

The formula is as follows Eq. (1).

$$a = (a, a, \dots, a_n) \tag{1}$$

where a_i denotes the number of characteristic attributes. Therefore, the price of p is given as Eq. (2).

$$p(a) = (a_1, a_2, \dots, a_n) \tag{2}$$

The hedonic price of feature i can be defined as Eq. (3).

$$\frac{\partial p}{\partial a} = p_i(a_1, a_2, \dots, a_n) \tag{3}$$

In the hedonic pricing model, some variables can be measured in the theoretical model. The framework of the hedonic model can be used to find the effect of a certain characteristic on the price while keeping certain variables. It is also a price model nowadays mainly used to calculate special goods.

2.2 The Regression Model

Regression is the prediction of continuous type data. Studying the relationship between the dependent and independent variables, such as sales prediction or manufacturing defect prediction, the goal of a regression model is to get each sample in the training set to fit as close as possible to a linear model. While general regression models use mean square error MSE as the loss function.

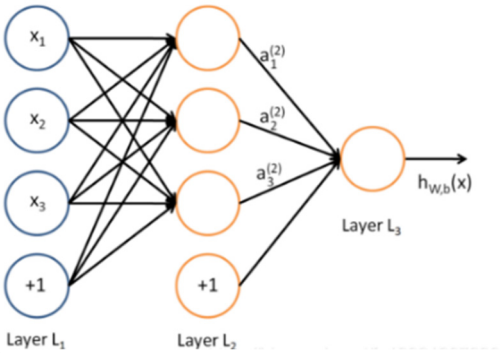

Figure 1: MLP model.

Multilayer perceptron (MLP) is an artificial neural network with a three-layer model. In addition to the input layer and output layer, the middle layer can have multiple hidden layers. The simplest MLP contains only one hidden layer, that is, a three-layer structure, as shown in figure 1.

2.3 The Support Vector Regression Model

Support Vector Regression (SVR) implements different models depending on the input data, and does regression if the input labels are continuous values. By seeking to minimize the structured risk to improve the generalization ability of the learning machine, the empirical risk and the confidence range are minimized, so as to achieve the purpose of obtaining good statistical laws even in the case of a small statistical sample size.

The SVM is formulated as the following linear estimation function Eq. (4):

$$f(x) = \left(w.\emptyset(x)\right) + b \tag{4}$$

where w denotes the algorithm weight vector, b denotes a constant, and $\emptyset(x)$ denotes a mapping in the feature space function.

$$L_\varepsilon(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon, & |f(x) - y| \geq \varepsilon \\ 0, & otherwise \end{cases} \tag{5}$$

The formula Eq. (5), where ε is a precision parameter representing the radius of the tube located around the regression function $f(x)$.

In the SVR, the coefficients w and b can be estimated by the minimization method.

3 COE COFFEE PRICE FORECASTING MODEL

Economic factors interact with coffee prices in a complex and non-linear way. The machine learning algorithms are used to analyze historical data or past experiences to optimize the performance criteria of the model. For the problem of predicting the value or number of variables, typical machine learning algorithms (MLAs), mainly supervised learning methods, can be divided into the following categories: kernel learning methods, tree-based methods, distance-based methods, and neural network methods. To compare the MLA performance between different theories and architectures, three commonly used machine learning algorithms are selected. Multi-layer perceptron models, Support vector regression and multiple regression models are used to improve prediction accuracy and avoid over-fitting.

Higher quality coffee gets better bids and prices in the market. Altitude is an easily available and commonly used proxy indicator. A systematic comparison of price and altitude reveals a statistically insignificant correlation between altitude and price. The correlation between altitude and price was not significant. Certification has a greater impact on price than altitude. These relationships between price, quality, and certification deserve further study. The market attempts to establish a healthy coffee price market to achieve coffee quality, farm profitability sustainability and healthy development. Refined monitoring of coffee prices should provide an important influence for a fair market.

Table 1: One-hot coding.

	year	country	minp	rank	score	C_Price	Farm_Size	Coffee_growing_area	Variety	Process	lb_price
0	2004	El Salvador	1.20	19	95.76	0.8047	17.5	17.5	Paca	wet	2.31
1	2004	Honduras	3.50	1	95.69	0.8047	30.0	20.0	Paca	wet	13.00
2	2010	Colombia	8.90	1	94.92	2.3002	8.0	3.0	Caturra, Castillo	wet	40.09
3	2007	Nicaragua	3.95	1	94.84	1.2355	90.0	30.0	Caturra	wet	47.06
4	2010	Nicaragua	5.25	1	94.14	1.9090	8.0	5.0	Maragogipe	wet	35.65
...	...	...	...	...	...	...	...	...	...	...	...
1115	2005	Colombia	1.85	28	84.40	1.1573	3.0	3.0	Caturra	ecological	2.05
1116	2005	Colombia	1.85	30	84.33	1.1573	7.5	2.0	Caturra	ecological	1.90
1117	2005	Colombia	1.85	31	84.28	1.1573	5.0	4.0	Caturra	ecological	2.15
1118	2005	Colombia	1.85	32	84.18	1.1573	6.5	2.5	Caturra	ecological	2.25
1119	2005	Colombia	1.85	33	84.13	1.1573	2.0	1.0	Caturra	ecological	1.85

1120 rows × 11 columns

The one-hot codes are shown in table 1. The classical one-hot coding method for coffee categories is improved, from one coding for each category to a combined coding method for each category. The new method reduces the number of categories and reconstructs the original sparse category features to make the features more dense. The generalized category is 32 categories. The encoding after generalization is shown in table 2.

Table 2: Improved coding method.

	year	country	minp	rank	score	C_Price	Farm_Size	Coffee_growing_area	Variety	Process	lb_price
0	0	0	1.20	19	95.76	0.8047	17.5	17.5	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0	2.31
1	0	1	3.50	1	95.69	0.8047	30.0	20.0	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0	13.00
2	1	2	8.90	1	94.92	2.3002	8.0	3.0	[0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0	40.09
3	2	3	3.95	1	94.84	1.2355	90.0	30.0	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0	47.06
4	1	3	5.25	1	94.14	1.9090	8.0	5.0	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0	35.65
...	...	...	...	...	...	...	...	...	...	...	...
1115	5	2	1.85	28	84.40	1.1573	3.0	3.0	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4	2.05
1116	5	2	1.85	30	84.33	1.1573	7.5	2.0	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4	1.90
1117	5	2	1.85	31	84.28	1.1573	5.0	4.0	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4	2.15
1118	5	2	1.85	32	84.18	1.1573	6.5	2.5	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4	2.25
1119	5	2	1.85	33	84.13	1.1573	2.0	1.0	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	4	1.85

1120 rows × 11 columns

4 EXPERIMENTAL STUDIES

This study utilizes the data set of the 2004-2011 Cup of Excellence sponsored by the Alliance for Coffee Excellence (Wilson 2015, Wilson 2015). The dataset includes information on the final auction price of each coffee, the grade of the coffee, the quantity of coffee, farm data, elevation of the coffee trees, processing method and origin.

Coffee price distribution is shown in fig.3, which conforms to the cumulative Gaussian distribution. The Gaussian mixture model assumes that the data of each sub-data is consistent with the Gaussian distribution, and the current data presents a distribution that is the result of superimposing the Gaussian distributions of each cluster together.

The relevance between the price and the futures index is verified by the Pearson coefficient.

Fig. 2 is a heat map of the relationship between the fields of the dataset. It can be found that the relationship between the parameters of the coffee, the highest correlation between the score and the price, is 0.568. These data indicate that all other data of the coffee have an impact on its price, such as the taste of the coffee, flavor, etc., which are to be further investigated.

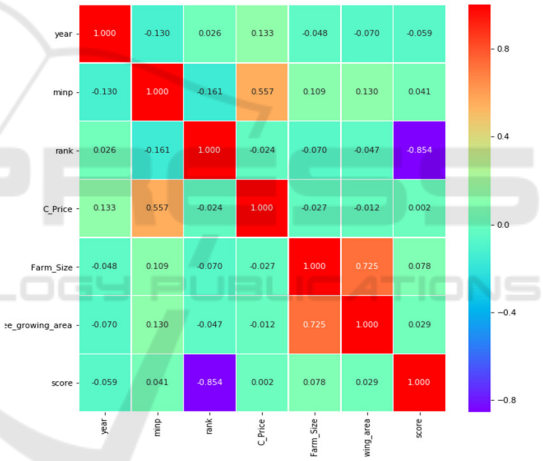

Figure 2: Parameter coefficient heat map.

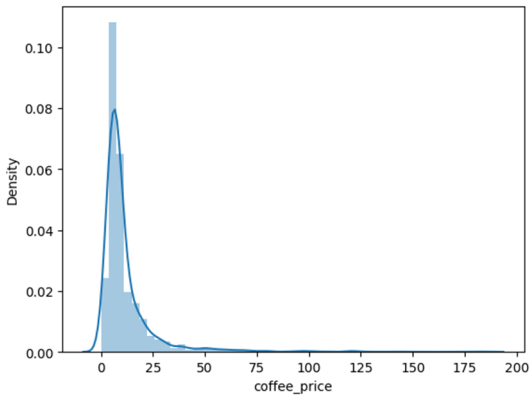

Figure 3: Distribution of the coffee_price.

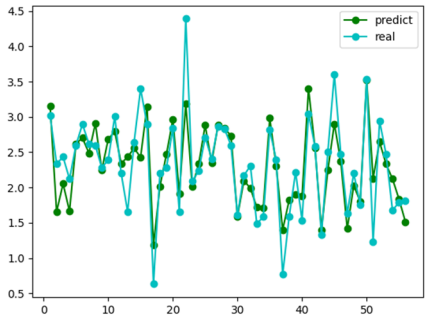

Figure 4: Improved Regression.

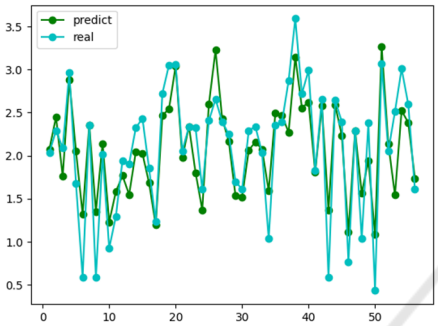

Figure 5: SVR model.

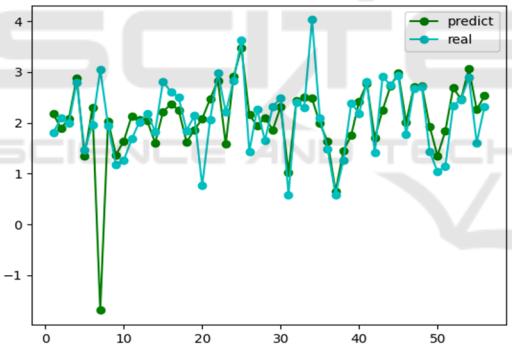

Figure 7: MLP model.

Fig.4, Fig. 5, and Fig. 6, show the prediction results for the dataset after modeling the three algorithms, and after statistics and calculations, the accuracy data of the four algorithms are shown in table 3. It can be found that the classical multiple regression is the least effective, at 49.51%. The improved algorithm of generalized coding is the best with 74.61%. It is higher than the two models, SVR and MLP. The statistics show that the classical multiple regression parameters vary greatly with an intercept of -100. while the other three tested models have an intercept error within -1 and the coefficient errors of the parameters are within a 0.1 range.

Table 3: Score of compared algorithm.

Algorithm	**SVR**	**MLP**
Score	***62.74%***	***56.98%***
Algorithm	C_Regression	I_Regression
Score	49.51%	74.16%

5 CONCLUSION

This paper investigates the COE coffee bidding system and designs a generalized coding of the categories that affect the price. The validation was carried out by three unsupervised learning algorithms of machine learning. Simulation results reveals that the improved regression algorithm outperforms support vector machine and multilayer perceptron models. The regression model after generalizing the coffee category ranked first in terms of prediction accuracy, followed by the SVR model. The third place is the MLP model, and the last is the classical regression model. Numerically, the proposed improved model has an accuracy 24.65 higher than the traditional regression algorithm, 11.42 higher than SVR, and 17.18% higher than MLPA.

Limited by the number of samples and the insufficient collection of factors about coffee, such as coffee taste, major buying countries, sales volume, etc., the accuracy of the algorithm can be further improved. In the future, as the number of samples increases and the quality of the dataset improves, the neural network algorithm will be used to do further prediction and analysis. The next step is to apply the model to other related fields.

REFERENCES

Bacon C. Confronting the Coffee Crisis: Can Fair Trade, Organic, and Specialty Coffees Reduce Small-Scale Farmer Vulnerability in Northern Nicaragua? [J]. Center for Global, International and Regional Studies, Working Paper Series, 2004, 33(3):497-511.

Donnet L, Weatherspoon D, Hoehn J P. Price Determinants in Top Quality E-Auctioned Specialty Coffees. 2007.

Ferreira H A, Liska G R, Cirillo M A. Selecting A Probabilistic Model Applied To The Sensory Analysis Of Specialty Coffees Performed With Consumer[J]. IEEE Latin America Transactions, 2016, 14(3):1507-1512.

Gu J, Zhu M, Jiang L. Housing price forecasting based on genetic algorithm and support vector machine[J]. Expert Systems with Applications, 2011, 38(4):3383-3386.

Hu L, He S, Han Z. Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies[J]. Land Use Policy, 2019, 82:657-673.

Niklas B, Rinke W. Pricing Models for German Wine: Hedonic Regression vs. Machine Learning[J]. Journal of Wine Economics, 2020, 15.

Oladunni T, Sharma S. Hedonic Housing Theory — A Machine Learning Investigation[C]// IEEE International Conference on Machine Learning & Applications, 2017, 522-527

R Teuber. Estimating the Demand for Sensory Quality – Theoretical Considerations and an Empirical Application toSpecialty Coffee[J]. Agrarwirtschaft, 2010.

Traore T M, Wilson N, Fields D. What Explains Specialty Coffee Quality Scores And Prices: A Case Study From The Cup Of Excellence Program[J]. Journal of Agricultural and Applied Economics, 2018, 50.

Wilson A P, Wilson N L W. The economics of quality in the specialty coffee industry: insights from the Cup of Excellence auction programs[J]. Agricultural Economics, 2015, 45(S1):91-105.