Cash Valuation of Black Tea in the Nuwara Eliya District based on Sensory Quality Attributes: A Case Study.

Rodrigo W.A.R.H.[1], Magalla C.H.[2], Basnayake W.M.N.D.[3]
Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka[1,2], La Trobe University, Melbourne, Australia[3]
[1]ramalee.rodrigo@gmail.com [2]champa@stat.cmb.ac.lk [3]N.Basnayake@latrobe.edu.au

***Abstract*—The cash valuation of tea; estimated price of tea, decided prior to the auctions, is influenced by sensory assessments from tea brokers and tasters up to a certain degree, although market conditions and customer preferences decide the auction price. This study aims to predict the cash valuation of black tea using sensory quality measures and to identify the key factors that impact these valuations overall and grade specific. This information is crucial for stakeholders to maintain the quality standards of Sri Lankan Tea, supporting the export economy. While past research mainly focused on auction price prediction, few studies have modeled estimated prices using sensory quality parameters. However, using categorical sensory measures is significant in the current study as it was not previously implemented by past researchers. The study analyzed 1,119 tea samples with 13 attributes, finding that tea grade and ordinal sensory attributes are important for cash valuation. Due to the attributes being ordinal, numerical encoding was used to predict prices for the entire dataset and for each grade, using statistical and machine learning regression methods. With advanced analysis showing gradient boosting regression as the best predictive model for overall cash valuation of tea, the model achieved a minimum RMSE of 79.75. The study identified that tea grade, average weight of a tea sample and dry leaf color are essential in cash valuation, with DUST1 and BOPF being the most expensive grades. Cash valuations for these tea grades were observed to be higher when the dry tea leaves were in shades of black.**

***Index Terms*—tea, sensory quality measures, estimated price, machine learning**

I. Introduction

Tea is one of the major export crops in Sri Lanka, playing a vital role in the country's economy by contributing significantly to foreign exchange earnings and the Gross Domestic Product (GDP). As the fourth largest tea producer and third largest exporter globally, Sri Lanka's tea, known for its unique flavor and quality, is highly esteemed in international markets. Therefore, it is important to identify the features that contribute to the cash valuation of tea to uplift the country's export economy. Sri Lankan tea varies significantly in quality based on the geographical and climatic conditions of its growing regions, which are categorized by elevation into high, medium, and low-grown teas. High-grown teas, particularly from the Nuwara Eliya district at elevations above 4,000 feet, are known for their premium quality and delicately fragrant aroma and are highly sought after in the global market. This study focuses on the black tea prices specifically in the Nuwara Eliya district, renowned for producing some of the finest Ceylon tea.

In Sri Lanka, tea is sold mainly through public auctions like the Colombo Tea Auction (CTA). Brokers assess tea using sensory quality measures and their specialized terminology [1] to set estimated prices because the quality of tea is synonymous with its prices [2]. These prices guide buyers, while final auction prices are determined through competitive bidding, reflecting market perceptions of tea quality. Evaluation involves a thorough examination of dried tea leaves, infused leaves, and tea liquor to ensure that only the highest quality teas proceed to auction. Hence, a proper price prediction method based on the quality measures of tea is required. Therefore, the price estimation process used by the tea brokers is applied, as it is a process that requires the quality evaluation of tea using sensory quality measures. This process does not have a standardized method of cash valuation because it is a tedious process that requires experience, domain knowledge, and is based on the personal merit-demerit system used by each tea broker.

Hence, the main objective of this study is to formulate a model to predict the cash values of tea based on sensory quality parameters, while the secondary objectives are:

- To identify the most important sensory quality parameters associated with the cash valuation of tea.
- To identify the most important sensory quality parameters associated with the cash valuation for each tea grade.

The significance of this research lies in the use of terms employed by stakeholders in the industry (categorical variables) to develop a predictive model for cash valuation, rather than providing ranks or gradings [3], [4], [5], [6] (numerical variables) for sensory tea quality measures. Manufacturers will be able to comprehend these measures more easily if they are provided in the terminology they use (in a qualitative format). Additionally, the predictive cash valuation model developed in this study addresses the subjectivity of the brokers. Fur-

Correspondence: Ramali Rodrigo (E-mail: ramalee.rodrigo@gmail.com)
Received: 16-06-2024 **Revised:** 12-08-2024 **Accepted:** 09-09-2024
Ramali Rodrigo and Champa Magalle are from Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka (ramalee.rodrigo@gmail.com, champa@stat.cmb.ac.lk) and Nadeeka Basnayake is from La Trobe University, Melbourne, Victoria Australia (N.Basnayake@latrobe.edu.au)
DOI: https://doi.org/10.4038/icter.v18i2.7295
The 2025 Special Issue contains the full papers of the abstracts published at the 24th ICTer International Conference.

This is an open-access article distributed under the terms of the Creative Commons Attribution License , which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

thermore, this research aims to identify the important sensory quality measures affecting the cash valuation of tea, both overall and grade-specific, through the best-fit models. The paper is structured as follows: Section II describes related work, Sections III and IV describe the methodology and results, respectively, and Section V contains the conclusion of the paper.

II. Related Work

The Sri Lanka Tea Board acts as the monitoring authority of the teas that are distributed to the market to maintain the established global quality standard of 'Ceylon Tea', represented by the Lion logo. This is why Sri Lankan tea is globally known for its excellence in processing and purity of product, as reflected by world-class certification. Hence, the future sustainability of the industry depends on improving the quality and increasing the yield of tea.

Past research has been conducted to identify the quality of green tea leaves. [7] suggested that the quality of tea leaves can be maintained if the harvesting interval is kept between 10-13 days. These plucking standards are considered important for maintaining tea leaf quality, even though the normal plucking standard yields a significantly higher quantity of production [8]. The quality of tea is significantly increased by using the plucking standards B+2 and B+3. Therefore, it is vital to pluck 2-3 tender leaves from the top to produce quality tea while keeping the damage to a minimum, as the tender leaf contains 20% more polyphenols than the mature leaf [9]. Considering leaf quality in different factories, the overall findings by [10] suggest that tea leaf freshness (enhanced by implementing proper plucking standards) significantly influences the output quality of tea. Hence, differences in tea leaves at the factory level can significantly affect the output quality of tea. Implementing limitations by considering one factory will help overcome quality variations due to differences in factory plucking standards.

According to past research, the quality assessment of tea has been conducted via two methods: the traditional human sensory assessment and chemical analysis. However, sensory quality evaluation is a difficult task performed by the tea taster, who uses a unique terminology in assessing the taste, that is often difficult for the consumer to understand. These experts judge tea quality by inspecting smell, flavor, and taste through oral examinations. Past research on grading tea quality using sensory aspects was conducted by [11] and [12]. However, this method was prone to errors due to the contradiction in test results caused by the use of different tasters. The trained panel may be influenced by psychological, physiological, and environmental factors. Hence, the need for standardized methods such as NIR (near-infrared) technology as an internationally acceptable tea quality evaluation measure has been studied by [6], involving different sets of tasters in identifying quality grades for different teas produced in various countries. However, this method was conducted on a smaller scale (using a few different countries). Therefore, it is difficult to construct a standard procedure that can depict this sensory evaluation of tea in different countries. Due to the lack of quantification of tea grade [13], the need to develop a method to overcome these limitations of experimental sensory assessment [14], [15], [16] was evident in [14].

Research has considered the composition of tea, such as mineral elements, as an identifier of tea grade [13]. However, considering chemical methods of addressing the issue was viewed as a proper scientific basis by the researcher [17]. The theaflavin (TF) content was identified as a significant chemical compound in assessing black tea quality was found by [18][18]. Furthermore, research conducted by [19], [20] confirmed the relationship between TF content and the broker's quality valuation of Central African black tea. Hence, TF is considered an important chemical aspect contributing to the brightness in tea infusion [21].

Without focusing on the chemical aspects, researchers have tried to incorporate the application of sensory aspects in assessing sensory quality evaluation. However, in the Sri Lankan context, most of the previous studies focused on the individual chemical properties of tea. The correlation of the chemical composition and color differences of black tea infusions studied by [22] showed significant variation with their sensory quality as assessed by tea tasters. Also, individual TF, total color and soluble solid content were researched by [21]. The total polyphenol content, total free amino acid content, total liquor color and brightness with respect to sensory aspects, infused leaf color, color of liquor, strength, and quality, as evaluated by [17], revealed that the polyphenol content had a positive correlation with brightness, which was responsible for the liquor quality in tea. This study also highlighted the significance of sensory properties, liquor color and brightness as reliable quality parameters to determine the quality of black tea.

In the international context, [23] used morphological features and chemometric tools in the evaluation of Congou black tea quality. [3] used the chemical composition and color differences of black tea infusions in assessing sensory quality, determining that certain chemical compounds affected both tea taste and color. It further implied the complexity of the composition of tea and the volatile components [24]. Hence, it is important to use technologies that provide accurate, objective, stable and effective means of measuring these chemical parameters. Methods such as NIR spectroscopy [6], lab-made computer vision systems (CVS) [23], electronic nose [15], [25], [26], electronic tongue [14], [16], [26] , and mass-spectrometry [13], [24], [26] have been used in determining the quality of tea. Majority of these techniques have been implemented in international tea research, but only a few in the Sri Lankan context.

In the above literature, the quantification of sensory variables was employed when assessing tea quality, either by using tea tasters' quality gradings, machines, or producing standardized quality measures [6]. However, the terminology used by tea tasting experts is qualitative, and the use of qualitative parameters seems to have been rare in past research. The present study attempts to use these categorical aspects in assessing the quality of tea.

A. Tea auction price

The auction price of tea is the selling price determined at tea auctions. It is useful for tea producers to have an idea about the prices for their produced tea, as the tea auction price is inherently noisy, non-stationary, and chaotic in nature. Hence, research considering all the factors that contribute to auction price variation has been rare. However, a few studies have been conducted in the context of forecasting the auction price of tea, considering certain parameters. These studies were time series-based, such as forecasting prices at auction centers using VAR models by [27]. However, it was found that ANN (Artificial Neural Networks) performed better than the time series regression approach when considering prices at different auction centers [28]. When modeling the auction price with respect to prices at different auction centers, these prices reflect the supply and demand balance, which is important in securing good prices in tea auctions, rather than being influenced by individual buyers or sellers [29].

Through data mining, [30] showed that the auction price has no relationship with the production volume and has little correlation with exports of all types of tea. According to [2], tea grade, elevation, previous auction price, valuation price, and the net quantity of the tea significantly affected auction price formation. It was also found that tea price is sensitive to changes in tea grade, elevation, and broker, according to [9]

[31] incorporated production and weather variables in modeling tea prices. Since temperature and humidity play a role in the production of tea [32], it was found that these variables contribute to altering the auction price of tea, even if the impact is not straightforward.

Adulteration is considered the main problem in reducing CTA prices, as it tarnishes the reputation built by Sri Lankan tea in the world market with its Lion Logo. This emphasizes the need to determine tea prices by considering tea quality as more important than other contributing factors. This idea is supported by [27], [28] who also mentioned the importance of considering quality characteristics in price determination.

B. The tea price with quality characteristics of tea

In the tea auction market, the final price of tea results from a complex process due to the intervention of many variables. During sales, consumers often believe that a high price is synonymous with high quality [2]. To what extent this assumption is true remains a problem, as there has been no standardized procedure for assessing tea quality worldwide. Generally, the individual quality gradings by tea tasters and brokers determine the price of the tea before it is introduced to the market [2]. According to past research, the market price (auction price) is decided upon by the interplay of these parameters. However, the final price often differs from the estimated price set by the tea broker.

Research has attempted cash evaluations for different tea quality parameters [3]. For instance, [4] discovered that odor attributes can be used as a measure in deciding the market price in Dianhong tea grade classification. However, tea tasters primarily consider the tea liquor characteristics when determining tea quality [22], which is why, in the beverage industry, quality assessments are typically based on sensory analysis. In econometrics, the method of predicting prices by considering sensory assessments is known as the 'Hedonic price function.' Using this concept, the research conducted by [5] focused on determining the price of tea with respect to biochemical quality parameters and assessed the accuracy of predictions with the help of tea tasters' sensory assessments. However, it was observed that there was an issue of incoherent subjectivity in the tasters' assessments.

However, research had attempted cash evaluations for different tea quality parameters [3] as [4] who discovered that the odor attributes can be used as a measure in deciding the market price in Dianhong tea grade classification. However, the tea taster mainly considers the tea liquor characteristics in deciding the tea quality [22], which is why in the beverage industry, the quality assessments are only based on sensory analysis. In econometrics, the method of predicting prices by considering the sensory assessment is called 'Hedonic price function'. Using this concept, the research conducted by [5], consisted of determining the price of tea with respect to biochemical quality parameters and had assessed the accuracy of predictions with the help of tea tasters' sensory assessment but, it was seen that there was an issue of incoherent subjectivity of the taster's assessments.

Instead of directly considering quality parameters, some studies have found the overall quality of tea by using tea tasters' evaluations and analyzing their impact on market prices [6]. These studies found that some teas were predicted to have high quality but had low market prices, and vice versa. This showed that high prices do not always reflect the quality of tea samples.

[19] developed an equation including TF content and the color of the tea liquor, which successfully explained the variation of the market price at a particular time. However, the study conducted by [21] failed to identify the correlation between the considered tea constituents (both chemical and sensory measures), such as TF, total color or solids to the market price.

Studies that incorporate only sensory quality aspects in price estimation are rare in the existing literature. Some quality parameters used in previous studies were factory-specific, as they depend on the manufacturing processes carried out by producers. Against this background, the present study was conducted to achieve the objective of examining the association of sensory quality aspects of tea to predict its cash valuation. Hence, the incorporation of factory-specific parameters is not required, as the study focuses on assessing sensory measures of tea, which are not affected by factory specifications.

C. Methods of Assessment of tea Quality

The tea quality grade is determined by the tea taster. According to studies conducted on analyzing tea quality, the rank given by the tea taster was used as a measure of quality [3], [4], [5], [6]. However, the methods employed by these researchers did not consider the estimated price (given by the tea brokers after evaluating the quality of tea), even though many studies showed positive correlations between the quality scores (given by the tasters) and the market price. The quality measures assessed by the taster and the tea broker are the same. Unlike the tea taster, the tea broker assigns an estimated price for tea through sensory quality assessment instead of a score. Hence, the estimated price given by the tea broker can be considered an alternative measure for assessing the quality of tea, in place of the tea taster's score. However, this score (or price) can be highly subjective [11], [12], depending on the tea taster (or tea broker). The current study intends to conduct cash valuation using the estimated price provided by the tea broker instead of the score given by tea tasters, addressing the problem of subjectivity in the tea broker's price estimation by using data prepared by several tea brokers.

There have been numerous efforts to correlate the quality of tea by considering its chemical and sensory aspects [33]. In analyzing tea quality, researchers have applied methods to quantitative sensory and chemical quality attributes. Principal component analysis (PCA) was employed [3], [5], [6] to identify sensory and biochemical measures for the assessment of quality. Techniques such as PCA [14], [16], [17], linear discriminant analysis (LDA) [16], [33], and linear regression coefficients [21] were used to identify patterns of association and for variable selection. Furthermore, regression coefficients were used to identify correlations among these quality measures with the tasters' scores and the price. However, without considering the statistical assumptions such as data normality and homoscedasticity, the relationship between these measures in linear regression cannot be deemed accurate. As a feature selection method, [5] used a data mining approach based on the robust analysis of variance (ANOVA) technique, which employs method of moment estimators (MM-estimators), concluding that this was a more powerful method than using linear regression models. Association rule mining was also employed to identify underlying relationships [34].

However, the current study plans to use the sensory quality measures provided by the tea broker in qualitative format to conduct the cash valuation of tea. Therefore, the application of the above-mentioned measures in analyzing the impact of sensory quality characteristics on the estimated price is unlikely.

D. Modelling the Tea Prices from quality measures

The goal of the present study is to predict the realized prices of teas based on sensory assessments. Published studies on the cash evaluation of tea quality have considered the use of quantitative sensory quality factors. The study conducted by [22] employed a multiple linear regression (MLR) technique to identify the influence of different quality characteristics on cash valuation. The research by [35] used regression analysis to assess the correlation of quality parameters with quality attributes and the price of black tea. This research was carefully executed by verifying the statistical assumptions before implementing the model.

Most studies have considered the cash valuation of quality attributes by evaluating both biochemical properties and sensory quality measures that are quantitatively defined. However, the nonlinear relationships between these quality parameters were influenced by the complex dynamics of chemical quality constituents, making linear models unsuitable. Consequently, [5] utilized nonlinear methods for the cash valuation of tea quality attributes. The predictability of price based on inherent quality in tea has been explored using a non-parametric regression technique called Multivariate Adaptive Regression Splines (MAR Splines). This method was successful in obtaining predictions.

However, in the tea industry, tea brokers rely on sensory analysis for price realization of tea. Therefore, it is important to address the subjectivity inherent in sensory measures during price estimation. The application of machine learning for predicting 'realized' prices, considering human taste perception (sensory), appears to be rare in literature. Hence, this study aims to employ various machine learning models to accurately predict prices for different tea grades.

III. Methodology

This section initially describes the dataset acquisition and pre-processing steps. Then, the exploratory analysis and advanced analysis conducted to achieve the objectives of the study are presented alongside the methods used to identify the most important features, the best predictive model, and interpretations of model behavior.

A. The Dataset

Understanding the data required to achieve the research aims involved consultations with domain experts and industry stakeholders, including the Assistant Tea Commissioner of the Sri Lanka Tea Board, Tea Tasters from the Sri Lanka Tea Tasting Unit, and tea brokers. Publications recommended by these experts were reviewed to gain a comprehensive understanding of the study area. Due to the unavailability of data in a tabulated format and the prevailing COVID-19 situation, relevant documents from a factory in the Nuwara Eliya district were collected in person from one of the eight tea broker companies. These documents included weekly data on orthodox black tea samples sold at the Colombo Tea Auction (CTA) for the years 2019 and 2020, such as muster reports and sales reports.

- Muster Reports (2019-2020) – Consist of details regarding the tea brokers' quality assessments for each tea sample sent for each weekly sale. After a close examination of the dry leaf, infused leaf, and the liquor properties of tea, the evaluations are provided in detailed descriptions. Each tea sample can be individually identified by its invoice number.
- Sales Reports (2019-2020)- Contain the tea samples from the muster report and new or additional tea samples that were either not mentioned in the muster report or incorrectly reviewed in the muster report. These also contain the estimated prices from the tea broker. Each tea sample can be individually identified by its invoice number.

These reports, prepared by several tea brokers, were essential for addressing the subjectivity inherent in tea valuation practices. Tea auctions are held weekly, resulting in one report per sale for each tea factory, totaling 204 reports in PDF format for the two years under study. The descriptive data from these reports were manually extracted into a tabulated format using Microsoft Excel. Key variables extracted included Sale Number (Sale_no) and Tea Sample Index (Inv_no), which uniquely identify each tea sample sold. Using these indexes, data from the reports were combined to create a comprehensive dataset. The datasets for 2019 and 2020 were imported into R Studio and merged, resulting in a final dataset comprising 11 variables with 1,119 observations, after omitting the index variables.

B. Dealing with missing values

Initially, data pre-processing was carried out, and missing values were found for four attributes. To assess the nature and extent of the missing data, an aggregation plot was generated for further evaluation, as shown in Figure 1 below.

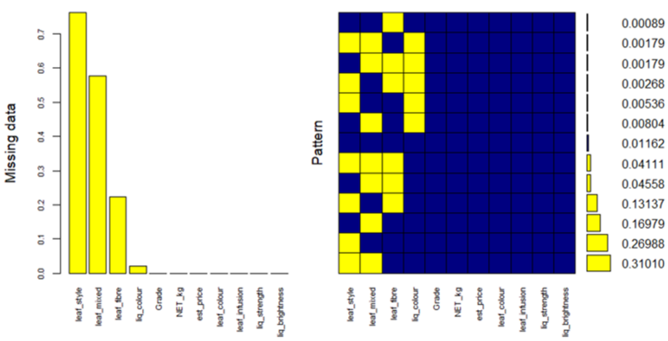

Fig. 1: Visualize Missing Values

Two variables were imputed using the missForest package in R, as it provided the lowest imputation error percentage for 10% to 30% missingness, according to [36], [37]. This method was verified for the data used in the study by employing several imputation methods, such as mean, mode, knnImpute, MICE, and missForest packages available in R Studio, while considering the lowest proportion of falsely classified (PFC) values. The results of this verification are provided in Table I below.

TABLE I: Imputation Method and their Imputation Error Percentage.

Imputation Method	PFC (%)
Mean Mode	30
knnImpute	33
MICE	37
missForest	28

The remaining two variables were removed because their missing value percentages were 57.7% and 75%, which were too high to be imputed without introducing bias to the data.

C. Descriptive and Advanced analysis

Next, to get an insight to the data used by the study, an exploratory descriptive analysis was conducted for the sensory measures tea grade-wise and price-wise using graphical methods. Here, each of the independent variables were analyzed with the response variable estimated price for price-wise analysis whereas, for grade-wise analysis, each sensory quality characteristic was analyzed with the available tea grades to identify if there are attributes unique to certain tea grades.

Furthermore, to identify multicollinearity among these sensory quality measures, correlations were calculated using Kendall's Tau-b rank correlation coefficients and a test of significance, as suggested by to [38]. This analysis helped identify variables that may have associations with cash valuation and confirmed the ordinal nature of the variables.

The advanced analysis aimed to develop the best predictive model for cash valuation and identify key sensory quality measures. Eight regression techniques were applied to the dataset, with the training and testing sets consisting of 80% and 20% of the data, respectively. To address the presence of six ordinal exploratory variables and preserve the relationships exhibited by the quality measures from the exploratory analysis, two encoding methods (numerical and dummy encoding) suggested by [39], [40] were considered and applied to each regression method. Parameter tuning was performed with 10-fold cross-validation (except for multiple linear regression) to optimize the models. The best predictive model for cash valuation of tea was identified based on the lowest RMSE, and the encoding method used by that model was selected as the best encoding method for the current data. From the best-fit model, important features associated with the cash valuation of tea were identified using variable importance plots (VIP) [41].

Next, using the best encoding method identified earlier, the eight regression models were fitted for each tea grade, and the best predictive model for price prediction for each tea grade was identified by the lowest RMSE. These best-fit models for each tea grade were then used to identify the tea grade-specific important variables associated with cash valuation using variable importance plots. Due to the difficulty in interpreting machine learning models, known as "black box" methods, partial dependence plots (PDP) and individual conditional expectation (ICE) plots were employed to understand the relationships of the most important features with the predicted response. These plots, obtained using the pdp package of R Studio [42], aided in the successful interpretation of the behavior of the best-fitted models.

IV. Results and Evaluation

Descriptive analysis is an important part of the study, and it will support advanced analysis. Figure 2 represents the distribution of tea grades used for the current study.

The estimated prices of teas showed a bimodal distribution (Figure 3) due to differences in price between BOPF and DUST1 grades (majority grades) and BOP, DUST, and FGS1 grades (minority grades), indicating the importance of tea grades in cash valuation. Expensive majority grades were produced more, while cheaper minority grades were produced less, as clearly depicted in Figure 2. This suggests that producers tend to pay more attention to highly priced tea grades rather than to lower-priced teas in the market.

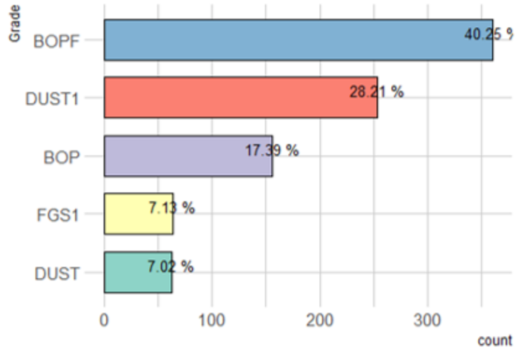

Fig. 2: Tea Grade Categories

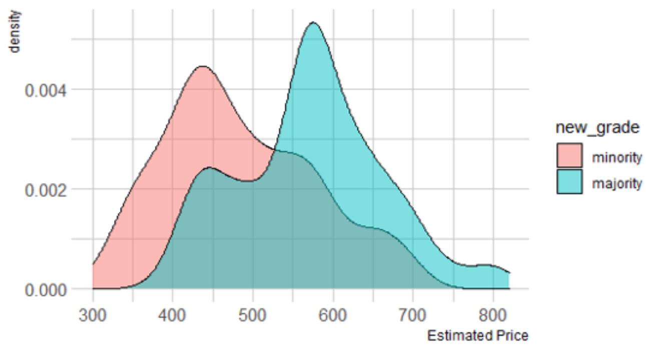

Fig. 3: Distribution of majority and minority grades with Estimated tea price

The exploratory analysis was conducted for sensory quality measures grade-wise and price-wise (cash valuation). In identifying quality characteristics for cash valuation, each of the independent variables was analyzed with the response variable, estimated price. Analyzing the behavior of the variables with the response variable, estimated price, revealed that some sensory quality measures, such as tea grade, leaf color, leaf fiber, and liquor color, were important in cash valuation. For expensive tea (tea with a high estimated price), the sensory quality measures identified in the analysis were:

- Leaf colour tend to be reasonably black.
- Leaf fiber content tends to be few or fairly clean.
- Liquor colour ten to be light coloured.

Considering the quality characteristics grade-wise, Table II shows that some sensory characteristics, namely leaf color, leaf fiber, liquor strength, and liquor brightness, exhibit clear differences between the off-grades and main grades.

TABLE II: Important Sensory quality measures specific to Tea Grades

Variable	**main-grades** **BOPF, BOP**	**off-grades** **DUST1, DUST, FGS1**
leaf colour	black shades	brown shades
leaf fiber	few, fairly clean	noticeable, prominent
liquor strength	Some strength (creamier)	little strength
liquor brightness	little brightness	bright liquor

However, apart from grade and estimated price, Kendall's tau-b correlation indicated significant associations and multicollinearity among some sensory quality measures, such as liquor strength, liquor brightness, leaf fiber, and leaf color.

The advanced analysis aimed at predicting the cash valuation of tea (estimated price) revealed that for multiple linear regression models, the encoding method did not alter variable significance. Average weight, tea grade, dry leaf color, leaf fiber content, and liquor color were identified as important predictors.

In advanced analysis, multiple linear regression models for both encoding types concluded that the significance of the variables did not change due to the encoding method. It was further observed that average weight, tea grade, dry leaf color, leaf fiber content, and liquor color were important variables. All regression assumptions were satisfied, but there were signs of multicollinearity, which had been previously identified. To address this problem, regularized regression was used. Regularized regression methods for both encoding types performed better than multiple linear regression, providing lower RMSEs. However, for dummy encoding, predictions were better than for numerical encoding. Ridge regression with dummy encoding performed best, with an RMSE of 83.54 and an R square of 0.27 Table III].

Since dummy variables increased the degrees of freedom, regularized feature selection was necessary. Lasso and Elastic-Net models did not perform feature selection as required (all variables were included in the final model), and their prediction models were relatively poor compared to ridge regression. In contrast, lasso regression with numerical encoding was the best predictive model with an RMSE of 84.35, consisting of 10 variables, where the dummy variable for a particular grade was not included. Hence, this model was discarded. Ridge regression with dummy encoding provided better predictions of cash valuation (lowest RMSE of 83.54 and R square of 0.27). However, there was no substantial decrease in test RMSE compared to MLR models. Therefore, machine learning techniques were needed to obtain better predictive models.

Machine learning techniques (except for regression trees) performed better, as RMSE values were further reduced. However, for numerical encoding, predictions were better than for dummy encoding. In the k-NN models, dummy encoding required five data points to determine the cash values of tea, while numerical encoding required a higher number (i.e., seven) of data points for predictions. The RMSE was considerably high for k-NN, making it less suitable as a prediction method.

The change in encoding method did not affect regression trees, as they produced the same decision tree for both instances, with an RMSE of 85.33. This performance was poor compared to k-NN regression. Random forest was employed to address the underperformance of regression trees. The best predictive model was achieved with numerical encoding, resulting in an RMSE of 81.01. Due to the correlation among trees in the random forest, the gradient boosting technique was applied. It was successful, as it performed better in price prediction, with the lowest RMSE (79.75) among all machine learning models [Table III].

TABLE III: Results of Regression Models for both Encoding Methods

Model type	Encoding method	Tuned Parameter	RMSE	R^2
Ridge Regression	Dummy	$\lambda = 2.857$, $\alpha = 0$	83.54	0.27
	Numeric	$\lambda = 3.702$, $\alpha = 0$	84.39	0.26
Lasso Regression	Dummy	$\lambda = 0.061$, $\alpha = 1$	83.76	0.27
	Numeric	$\lambda = 0.744$, $\alpha = 1$	84.35	0.26
E-net Regression	Dummy	$\lambda = 0.376$, $\alpha = 0.1$	83.71	0.27
	Numeric	$\lambda = 0.487$, $\alpha = 1$	84.45	0.25
k-Nearest Neighbours	D	$k = 5$	82.74	0.295
	N	$k = 7$	81.59	0.308
Regression Trees	Dummy	$C_p = 0.01089277$	85.33	0.248
	Numeric	$C_p = 0.01089277$	85.33	0.248
Random Forest	Dummy	$mtry = 4$, $ntree = 2000$	82.79	0.298
	Numeric	$mtry = 3$, $ntree = 2000$	81.01	0.317
Gradient boost	Dummy	$n.trees = 1000$, $iteration.depth = 4$, $shrinkage = 0.01$	80.48	0.325
	Numeric	$n.trees = 1000$, $iteration.depth = 4$, $shrinkage = 0.01$	79.75	0.335

The best predictive model incorporated the use of numerical encoding. The gradient boost model with numerical encoding identified important sensory measures for cash valuation as those attributes that showed a considerable difference in variable importance (VI) scores (shown in Figure 4) compared to other measures. Namely, tea grade, average weight and the dry leaf colour. The relationships of these attributes were further assessed via PDP and ICE plots.

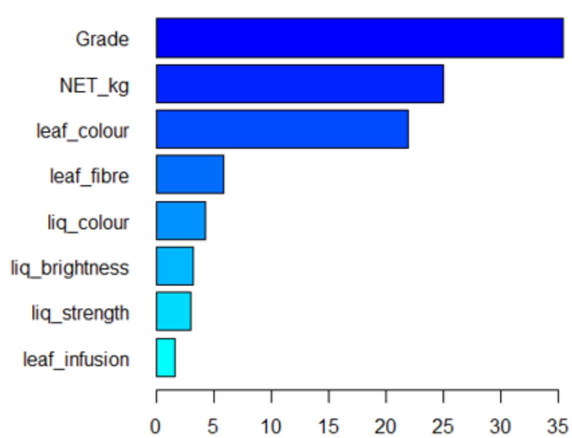

Fig. 4: VIP for the best predictive (gradient boost) model

From the results obtained, it was observed that, overall (considering all tea grades), the machine learning models performed better with the numerical encoding method. Hence, preserving the order of the levels of the factor variables was necessary for the study. Feature selection was not employed, as the number of quality attributes was limited to 11 due to the use of numerical encoding. Furthermore, tea grade was identified as an important variable associated with overall cash valuation, and descriptive analysis revealed certain grade-specific sensory quality measures. To achieve the objective of identifying the important variables associated with cash valuation, regression models were fitted for each tea grade, and the best predictive model was used to determine the most important variables for each grade. The results are provided in Table IV.

TABLE IV: Summarized Results for Grade-wise predictions

Tea grade	Most important variable/s	Best predictive model	RMSE	R^2
BOPF	leaf colour	Gradient boost	89.50	0.127
BOP	liquor colour and liquor strength	Lasso Regression	74.80	0.221
DUST	liquor brightness, leaf fibre, average weight, leaf colour	Multiple Linear Regression	61.52	0.329
DUST1	leaf colour	Random Forest	66.28	0.357
FGS1	average weight (NET_kg)	Random Forest	62.05	0.043

V. Conclusion

This study focused on developing a well-performing predictive model for the cash valuation of tea using sensory quality measures and identifying the most important sensory quality measures that affect the cash valuation of tea overall, as well as for each tea grade. Based on the results obtained from advanced analysis, it can be concluded that the gradient boosting regression method with numerical encoding on ordinal variables was the best predictive model for the overall cash valuation of tea, as it had the lowest RMSE of 79.75 compared to all other fitted models and explained 33.5% of the total variation in estimated price (Table IV). The model indicated that tea grade was important for determining the cash valuation of tea, with DUST1 and BOPF grades being the most expensive. Additionally, maintaining the average weight of a tea sample between 500-600 kg and 1500-2000 kg results in a high cash valuation of teas. Teas with dry leaf color in shades of black are considered high quality and more expensive. Since grade was the most important factor for cash valuation, it was concluded that each grade had specific sensory quality measures that contributed to the high prices:

- BOPF grade: having black shades of dry tea leaf
- BOP grade: having light liquor colour and 'strong' liquor strength
- DUST grade: having little liquor brightness, low leaf fiber low average weight and black shades of dry leaf colour
- DUST1 grade: having black shades of dry tea leaf colour
- FGS1 grade: having the average weight of a tea sample between 200-400kg

These characteristics resulted in a high cash valuation for tea. To produce high-quality teas, these parameters should be maintained, as they are the most important sensory measures associated with the estimated prices of tea (cash valuation).

The research faced several limitations and challenges during its execution. Due to the COVID-19 pandemic, data collection was restricted to a single factory in Nuwara Eliya, limiting the study to orthodox black teas and excluding other tea types. Additionally, there was a scarcity of prior studies using categorical sensory quality measures for price estimation in the tea industry, necessitating reliance on insights from domain experts and literature reviews. High levels of missing data posed another significant challenge, resulting in the omission of two variables and the imputation of the remaining missing data. The study also encountered difficulties with numerous ordinal variables, prompting the exploration of two encoding methods to handle them effectively. Despite efforts, the R-squared statistic for the best predictive model of cash valuation was lower than expected, likely due to the limited inclusion of sensory quality attributes. Future research is suggested to improve model performance by incorporating a broader range of sensory quality measures.

This study proposes several avenues for enhancing the predictive cash valuation model developed using sensory quality measures for tea. It suggests generalizing the model to include all tea factories in the Nuwara Eliya district and extending it to other tea districts across Sri Lanka, incorporating the district as a variable. Additionally, exploring elevation-wise (high, medium, low) estimated price predictions as a separate variable is recommended. The study advocates expanding the model to encompass a wider range of tea types and grades beyond orthodox black tea, considering additional sensory quality attributes and interaction terms to improve predictive accuracy. It also suggests exploring advanced machine learning techniques like Neural Networks and Support Vector Machines for better predictive models compared to gradient boost regression. Furthermore, employing data mining techniques such as association rule mining could uncover hidden relationships and patterns influencing high cash valuations, thus further refining the predictive capabilities of the model.

REFERENCES

[1] E. L. Keegel and T. R. I. of Ceylon, *Tea Taster's Terms*, 2nd ed. Colombo: Tea Research Institute of Ceylon, 1958, pp. 134–137.

[2] H. S. A. S. B. Gajanayake and K. H. S. K. Kithsiri, "Price formation, seller satisfaction and degree of competition at the colombo tea auction," *Appl. Econ. Bus.*, vol. 2, no. 1, pp. 36–44, August 2018.

[3] Y. Liang, J. Lu, L. Zhang, S. Wu, and Y. Wu, "Estimation of black tea quality by analysis of chemical composition and colour difference of tea infusions," *Food Chem.*, vol. 80, no. 2, pp. 283–290, 2003.

[4] X. Pang *et al.*, "Development of regression model to differentiate quality of black tea (dianhong): Correlate aroma properties with instrumental data using multiple linear regression analysis," *Int. J. Food Sci. Technol.*, vol. 47, no. 11, pp. 2372–2379, 2012.

[5] K. P. Sanjoy, "Predictability of price of tea from sensory assessments and biochemical information using data-mining techniques," *J. Sci. Food Agric.*, vol. 88, no. 8, pp. 1354–1362, 2008.

[6] S. H. Yan, "Evaluation of the composition and sensory properties of tea using near infrared spectroscopy and principal component analysis," *J. Near Infrared Spectrosc.*, vol. 13, no. 6, pp. 313–325, 2005.

[7] H. Sylvestre, N. Jacqueline, M. Alphonsine, and T. Olive, "Impact of plucking interval on tea productivity and green leaf quality," *Int. J. Soil Crop Sci.*, vol. 2, no. 10, pp. 68–72, 2014.

[8] W. S. Botheju, I. S. B. Abeysinghe, and C. K. Illeperuma, "Effect of plucking standard on quality and profitability of made tea produced in the up country dimbulla region during the cropping season," *Trop. Agric. Res.*, vol. 13, pp. 90–99, 2001.

[9] D. D. R. M. Karunaratne, R. A. P. I. S. Dharmadasa, R. M. S. D. Rathnayaka, and E. A. J. K. Edirisinghe, "Determinants of net sale average (nsa) of orthodox black tea factories in high elevation, sri lanka," *Kelaniya J. Manag.*, vol. 3, no. 1, p. 63, 2014.

[10] R. Samanga and V. Sachitra, "Impact of factory model cultivation on tea input and output quality: A study in selected tea factories in ruhunu region, sri lanka," in *International Conference Proceedings*, December 2017.

[11] A. Bakhshipour, A. Sanaeifar, S. H. Payman, and M. de la Guardia, "Evaluation of data mining strategies for classification of black tea based on image-based features," *Food Anal. Methods*, vol. 11, no. 4, pp. 1041–1050, 2018.

[12] H. Zhu, Y. Ye, H. He, and C. Dong, "Evaluation of green tea sensory quality via process characteristics and image information," *Food Bioprod. Process.*, vol. 102, pp. 116–122, 2017.

[13] L. Meng *et al.*, "Linear and nonlinear classification models for tea grade identification based on the elemental profile," *Microchem. J.*, vol. 153, p. 104512, August 2020.

[14] Q. Chen, J. Zhao, and S. Vittayapadung, "Identification of the green tea grade level using electronic tongue and pattern recognition," *Food Res. Int.*, vol. 41, no. 5, pp. 500–504, 2008.

[15] H. Liu, D. Luo, F. Li, and G. Xie, "Quality evaluation for anxi tieguanyin tea based on electronic nose and pca-lda method," in *Proceedings of the 2013 International Conference on Information Science and Cloud Computing Companion (ISCC-C)*, 2014, pp. 543–549.

[16] M. Palit *et al.*, "Classification of black tea taste and correlation with tea taster's mark using voltammetric electronic tongue," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 8, pp. 2230–2239, 2010.

[17] J. Kottawa-Arachchi, M. Gunasekare, M. Ranatunga, L. Jayasinghe, and R. Karunagoda, "Analysis of selected biochemical constituents in black tea (camellia sinensis) for predicting the quality of tea germplasm in sri lanka," *Tropical Agricultural Research*, vol. 23, no. 1, p. 30, 2012.

[18] E. A. H. Roberts, "The phenolic substances of manufactured tea. ix.—the spectrophotometric evaluation of tea liquors," *Journal of the Science of Food and Agriculture*, vol. 14, no. 10, pp. 700–705, 1963.

[19] P. J. Hilton and R. T. Ellis, "Estimation of the market value of central african tea by theaflavin analysis," *Journal of the Science of Food and Agriculture*, vol. 23, no. 2, pp. 227–232, 1972.

[20] J. B. Cloughley, "The effect of fermentation temperature on the quality parameters and price evaluation of central african black teas," *Journal of the Science of Food and Agriculture*, vol. 31, no. 9, pp. 911–919, 1980.

[21] K. Sivapalan and J. De Silva, "Tea constituents and export standards of sri lanka teas," *Journal of Tea Science*, vol. 54, no. 1, pp. 26–31, 1985.

[22] A. K. Biswas and A. K. Biswas, "Biological and chemical factors affecting the valuations of north-east indian plains teas: I.—statistical association of liquor characteristics with cash valuations of black teas," *Journal of the Science of Food and Agriculture*, vol. 22, no. 4, pp. 191–195, 1971.

[23] G. Ren, N. Gan, Y. Song, J. Ning, and Z. Zhang, "Evaluating congou black tea quality using a lab-made computer vision system coupled with morphological features and chemometrics," *Microchemical Journal*, vol. 160, p. 105600, 2021.

[24] R. Joshi and A. Gulati, "Fractionation and identification of minor and aroma-active constituents in kangra orthodox black tea," *Food Chemistry*, vol. 167, pp. 290–298, 2015.

[25] B. Tudu, A. Jana, A. Metla, D. Ghosh, N. Bhattacharyya, and R. Bandyopadhyay, "Electronic nose for black tea quality evaluation by an incremental rbf network," *Sensors and Actuators B: Chemical*, vol. 138, no. 1, pp. 90–95, 2009.

[26] J. Wang and Z. B. Wei, "The classification and prediction of green teas by electrochemical response data extraction and fusion approaches based on the combination of e-nose and e-tongue," *RSC Advances*, vol. 5, no. 129, pp. 106 959–106 970, 2015.

[27] K. A. S. D. B. Dharmasena, "International black tea market integration and price discovery," *Research and Management Journal*, vol. 1, no. 2, pp. 1–15, 2003.

[28] H. A. C. K. Hettiarachchi and B. M. S. G. Banneheka, "Time series regression and artificial neural network approaches for forecasting unit price of tea at colombo auction," *Journal of the National Science Foundation of Sri Lanka*, vol. 41, no. 1, pp. 35–40, 2013.

[29] M. R. Motha, K. G. M. C. P. B. Gajanayake, and Sarukkali, "Degree of competition at the colombo tea auction in sri lanka," in *3rd Agricultural Research Symposium, Wayamba University of Sri Lanka*, 2004, p. 5.

[30] H. C. Fernando, W. M. R. Tissera, and R. I. Athauda, "Gaining insights to the importation of food to sri lanka using data mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 (IMECS 2010)*, no. March, 2008, pp. 540–545.

[31] R. P. D. Gunathilaka and G. A. Tularam, "The tea industry and a review of its price modelling in major tea producing countries," *Journal of Management Strategy*, vol. 7, no. 1, 2016.

[32] E. L. Keegel, "Relation of temperature and humidity to made tea," pp. 60–68, 1962.

[33] Q. Zhang and J. Ruan, *Tea: Analysis and Tasting*, 1st ed. Elsevier Ltd., 2015.

[34] A. Caldera and D. R. Serasinghe, "Analyzing tea auction trends for beneficial seasonal tea production in sri lanka," *Journal of Advanced Management Science*, vol. 5, no. 2, pp. 152–157, 2017.

[35] L. P. Wright, N. I. K. Mphangwe, H. E. Nyirenda, and Z. Apostolides, "Analysis of the theaflavin composition in black tea (camellia sinensis) for predicting the quality of tea produced in central and southern africa," *Journal of the Science of Food and Agriculture*, vol. 82, no. 5, pp. 517–525, 2002.

[36] D. J. Stekhoven and P. Bühlmann, "Missforest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[37] A. K. Waljee *et al.*, "Comparison of imputation methods for missing laboratory data in medicine," *BMJ Open*, vol. 3, no. 8, pp. 1–7, 2013.

[38] H. Khamis, "Measures of association: How to choose?" *Journal of Diagnostic Medical Sonography*, vol. 24, no. 3, pp. 155–162, 2008.

[39] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, illustrated ed. Springer New York, 2013.

[40] C. Winship and R. D. Mare, "Regression models with ordinal variables," *American Sociological Review*, vol. 49, no. 4, p. 512, 1984.

[41] B. M. Greenwell and B. C. Boehmke, "Variable importance plots—an introduction to the vip package," *R Journal*, vol. 12, no. 1, pp. 343–366, 2020.

[42] B. M. Greenwell, "pdp: An r package for constructing partial dependence plots," *R Journal*, vol. 9, no. 1, pp. 421–436, 2017.