Predicting Total Auction Values in the Cape Colony

Tessa Hubble[a]

[a]*Stellenbosch University, South Africa*

Abstract

Predicting total auction values from a Cape Colony auction dataset between 1720 and 1820.

1. Introduction

Over the first 150 years of the Cape Colony, deceased estate auctions were an important means of exchange for the burgeoning population. The only other means of purchasing goods was through ships that docked along the coast. From livestock and furniture to slaves, a wide variety of goods were obtained through auctions. Although auction rolls from this time period have been transcribed, inconsistent spelling and errors are pervasive. 30 different types of goods have been identified and tagged throughout the auction rolls. This project aims to use this subset of goods within auctions to predict total auction values. Focusing on approximately 1400 auctions between 1720 and 1820, two different machine learning techniques will be used to predict total auction values. Being able to assess the total auction value based on a subset of goods can be an important tool within historical data where there is limited information. Exploratory data analysis will provide context and support for the features included in the models. Throughout the exploratory data analysis and subsequent models, it becomes clear that the number of slaves, the date and the auction size consistently place upward pressure on total auction values. The performance of a k-nearest neighbours model and random forest model are compared to that of a linear regression to assess accuracy.Ultimately, it is concluded that the subset of goods offer limited predictive power.

Email address: 21559953@sun.ac.za (Tessa Hubble)

2. Auction data

Table 1 below provides an example of a portion of sales from one auction in 1770. This auction roll shows goods sold, the price and the type of good sold. A buyer_id column exists as well. Within this column, some names include titles such as “de Wede” (widow) or “mijnheer” (mister) which offers the possibility to determine how many titled men or women attend an auction. Table 1 shows that only a few goods are tagged.There are 26 different tagged goods in total. The number of different types of tagged goods per auction will act as predictors. It is therefore possible that the accuracy of the machine learning models included may be hampered by limited presence of these goods across auctions.

mooc_id	purchase_id	price	goods	type
MOOC10/10.1	MOOC10/10.1/1	0.7	1 partij flessen, bottels en aardwerk	
MOOC10/10.1	MOOC10/10.1/10	2.4	2 steenen vaderl:s seep en 2 steenen Caapse seep	
MOOC10/10.1	MOOC10/10.1/11	2.1	1 casje met thee en 1 bijbel	books
MOOC10/10.1	MOOC10/10.1/12	3.1	6 kussens	
MOOC10/10.1	MOOC10/10.1/13	3.2	5 kussens	
MOOC10/10.1	MOOC10/10.1/14	25.4	1 bed, 1 peul, 4 cussens en 2 combaarsen	beds
MOOC10/10.1	MOOC10/10.1/15	5.2	1 bijbel en 1 nagtmaalboek met silvere beslag en 4 boeken	books
MOOC10/10.1	MOOC10/10.1/15	5.2	1 bijbel en 1 nagtmaalboek met silvere beslag en 4 boeken	books
MOOC10/10.1	MOOC10/10.1/16	0.5	1 partij olijtijten	
MOOC10/10.1	MOOC10/10.1/17	1.6	1 partij canten en lind	
MOOC10/10.1	MOOC10/10.1/18	9.4	2 swarte chitsen	
MOOC10/10.1	MOOC10/10.1/19	9.7	2 swarte chitsen	
MOOC10/10.1	MOOC10/10.1/2	1.7	6 porc:ne boterpotten	china

3. Exploratory Data Analysis

Figure 1 provides a scatter plot of the total auction values between 1720 and 1820. There is a general upwards trend in the value of auctions over time. However, a large proportion of auctions remain under 50 Rix-dollars (currency at the time). Grouping the goods according to household, kitchen, farming and select goods allows for the total auction value to be broken down according to a different collections of goods present at the auction.

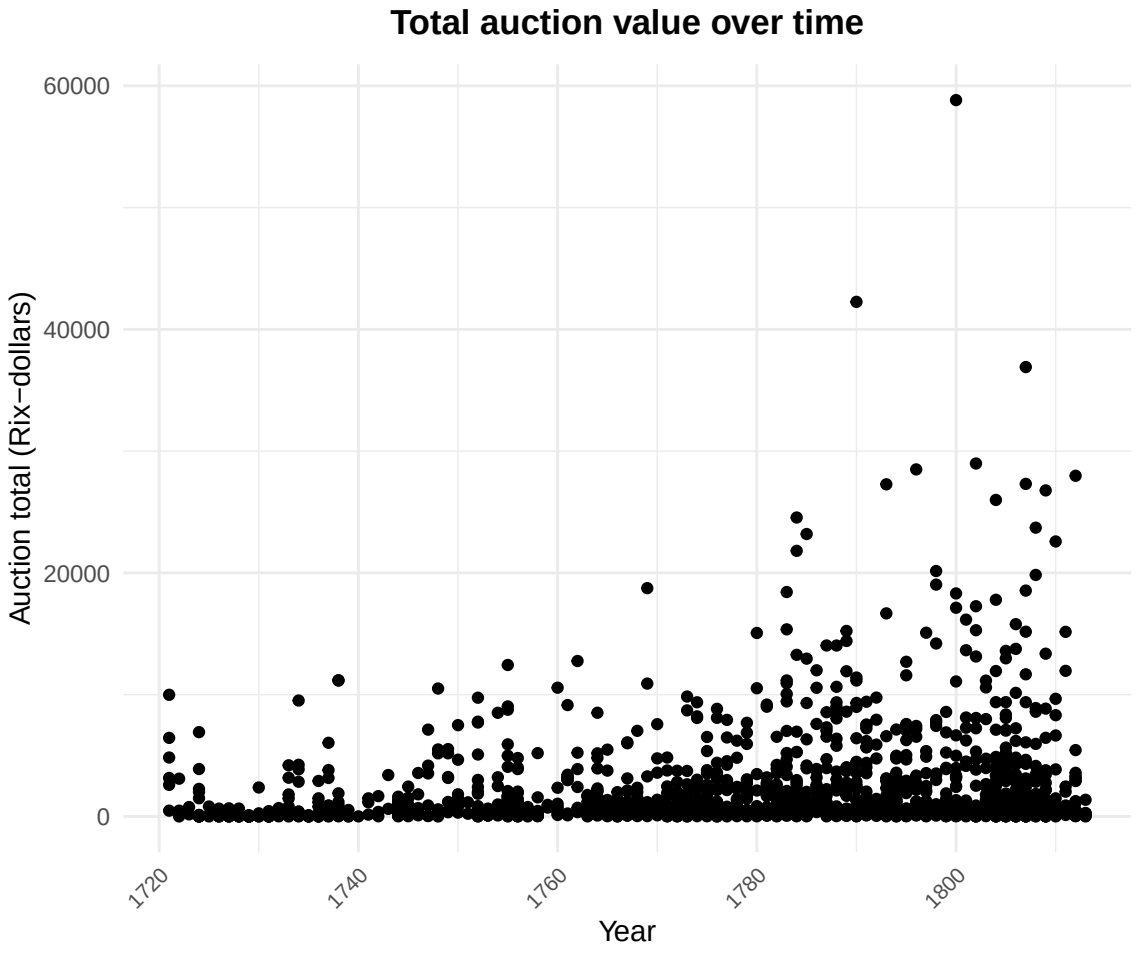


Figure 2 makes it clear that china, cups and chairs make up a large amount of the household goods sold per auction. However, the number of household goods, in absolute terms, does decline over time.

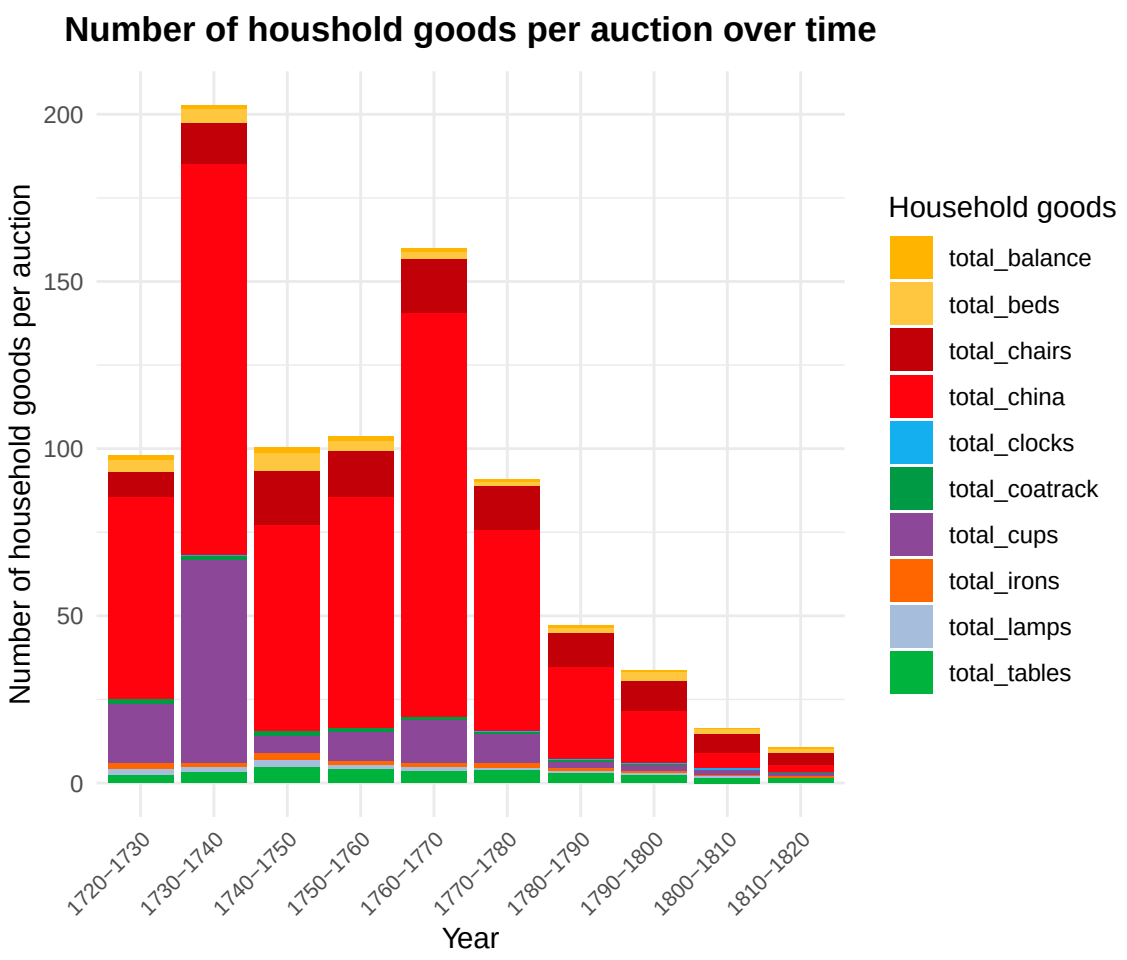


Figure 3 shows that utensils and plates make up the largest amount of kitchen goods sold per auction. As with household goods, kitchen goods also decline in absolute terms over time.

Figure 4 shows that the total number of farming goods is much higher than that of household and kitchen goods which peak at 200 and 120, respectively. Cattle and sheep are the most prominent farming goods sold per auction.

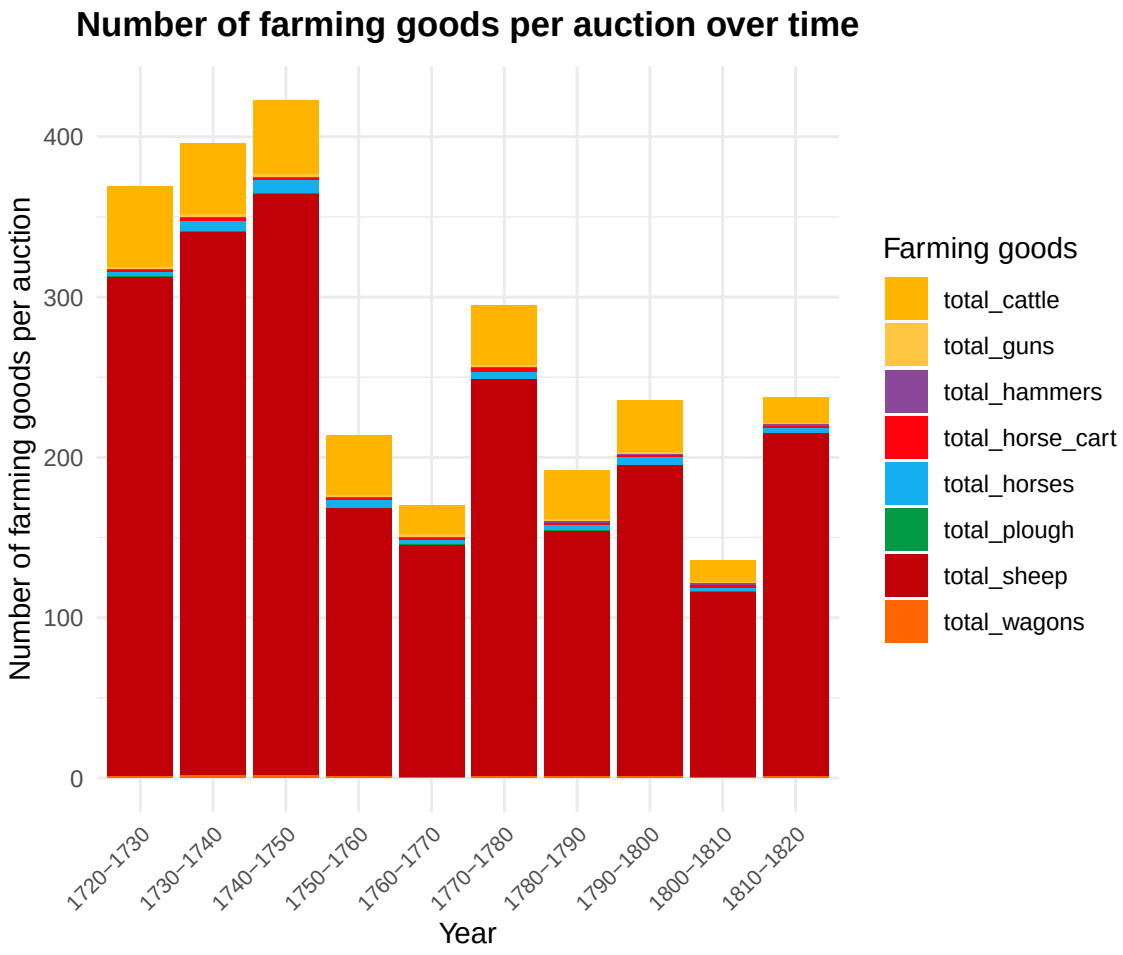
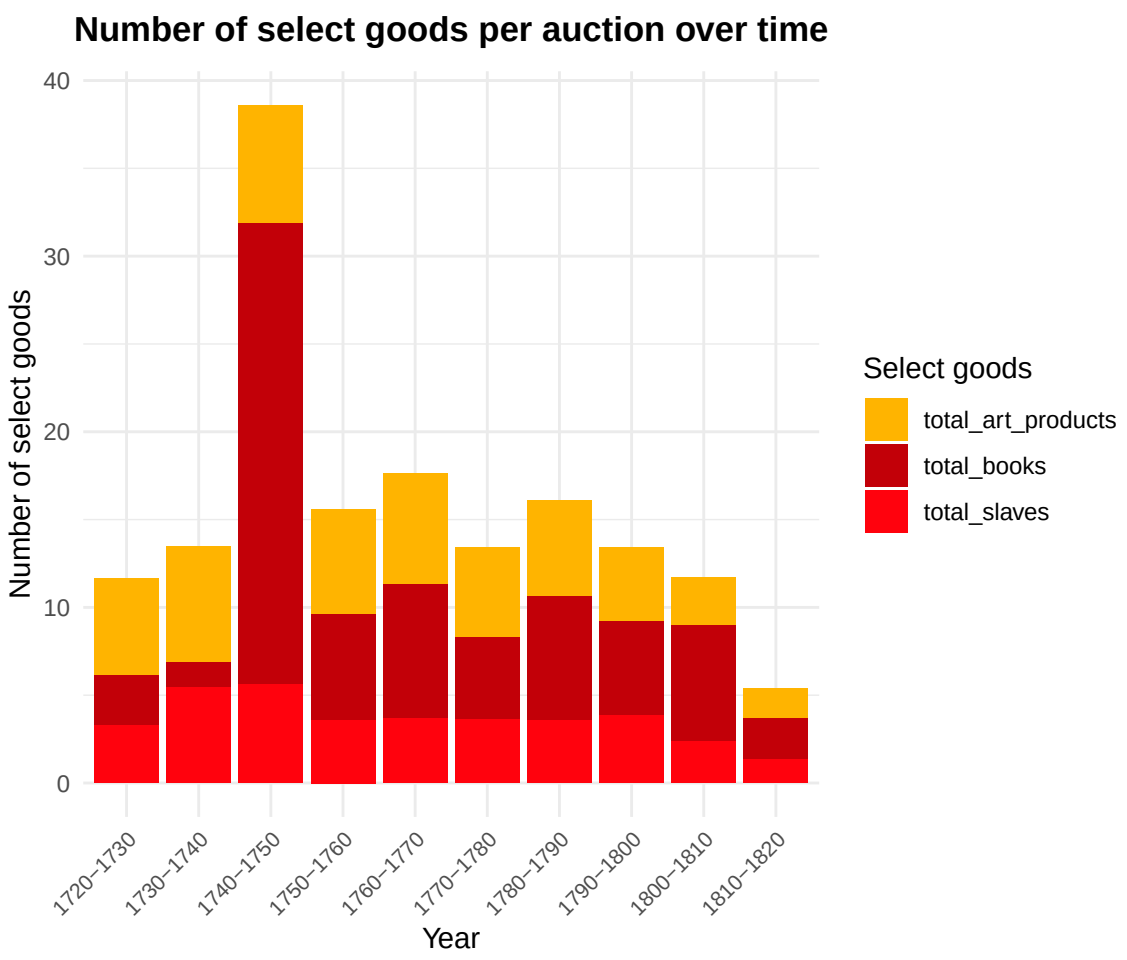
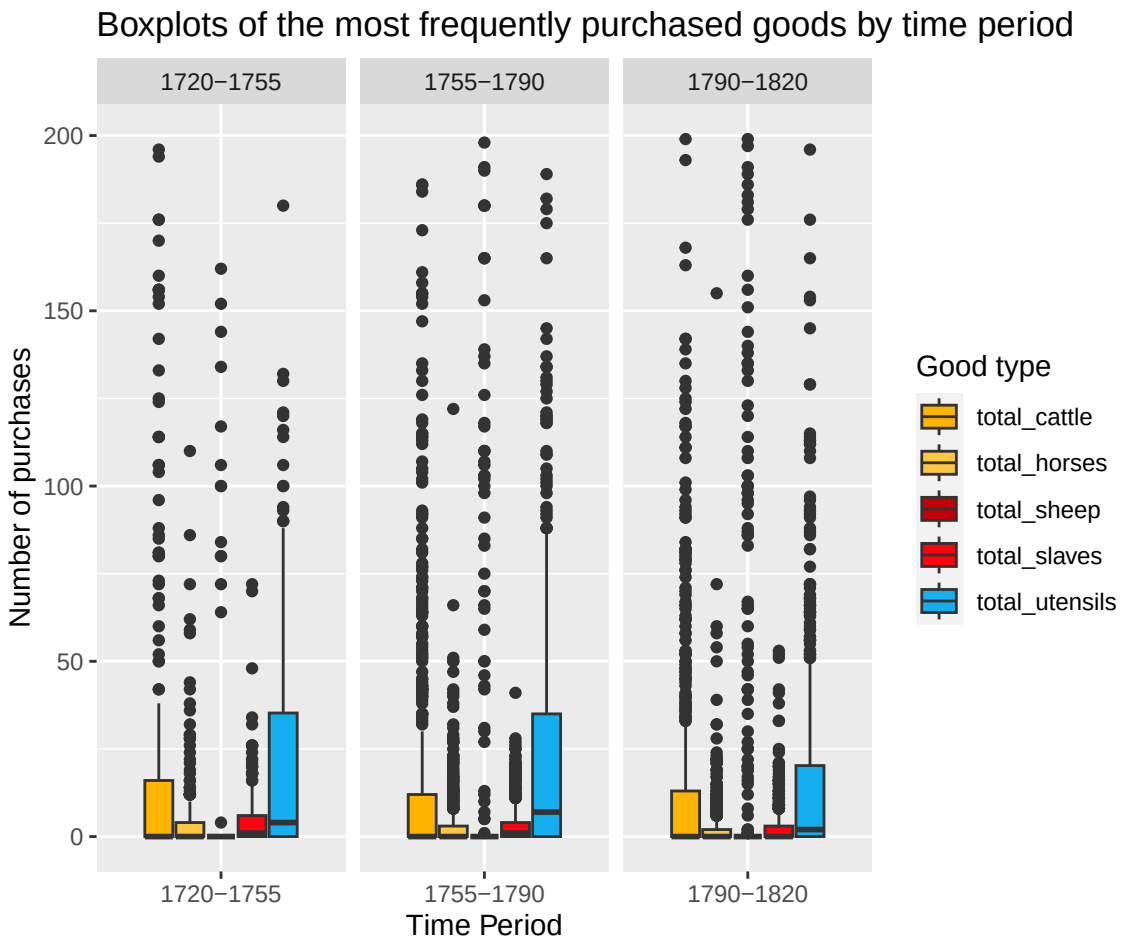


Figure 5 shows the average number of art products, books and slaves sold per auction over time. The absolute number of select goods is lower than the number of household, kitchen and farming goods. The umber of select goods per auction is more consistent over time than the other collections of goods.



The figures above indicate that the following goods the most frequently appear per auction: china, plates, utensils, sheep cattle, books, slaves and art products. Figure 6 shows box plots for some of these frequently purchased good types. For visual purposes, the y-axis was capped at 200 but there are a few outliers exceeding this number. Cattle and utensils appear to be some of the most frequently purchased goods throughout auctions yet the average number of goods per auction for each type remains very low.



In terms of attendance by titled men and women, Figure 7 shows the trends over time. There is a downward trend in attendance by titled purchasers from 1750 onwards and titled men always outnumber titled women at auctions. Although this univariate analysis identifies potentially important good types, their relationship with auction total values is more important.

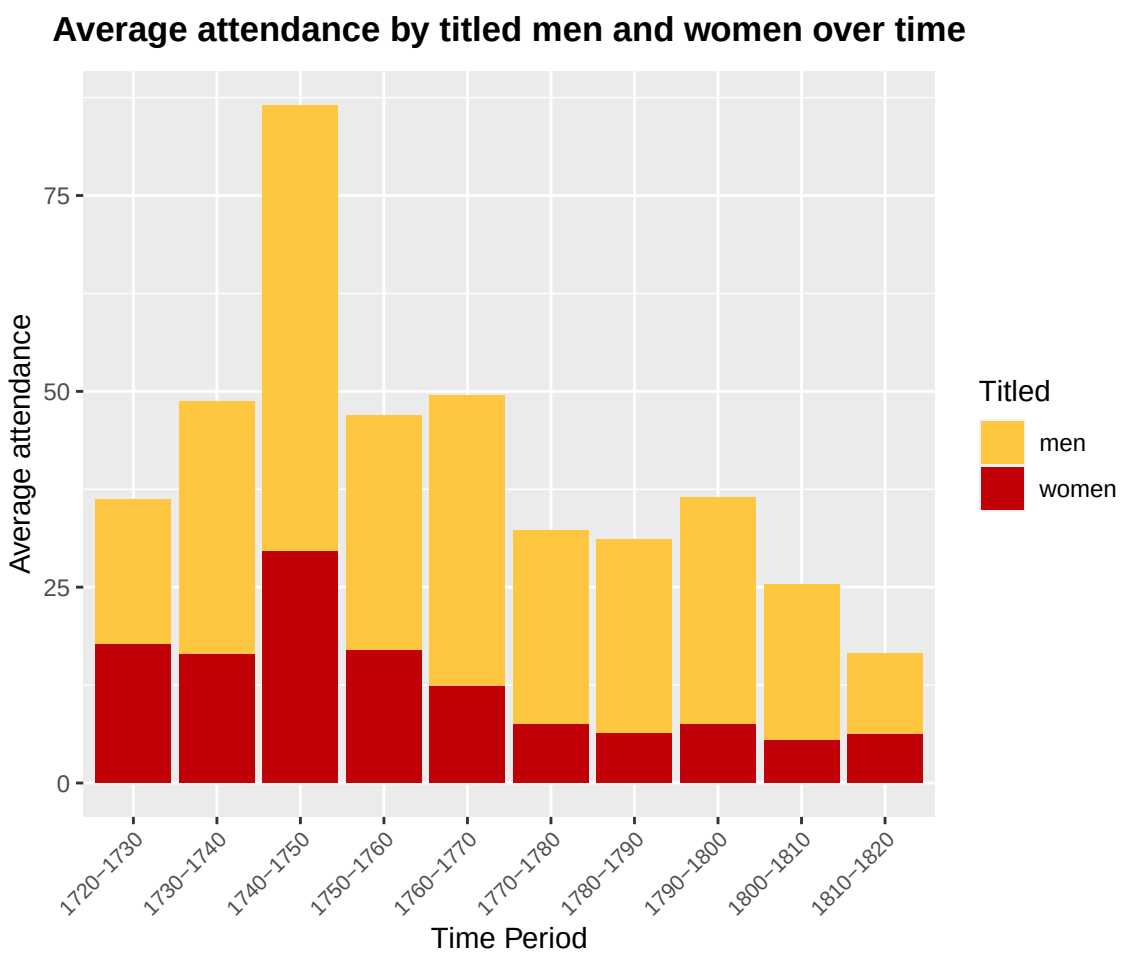
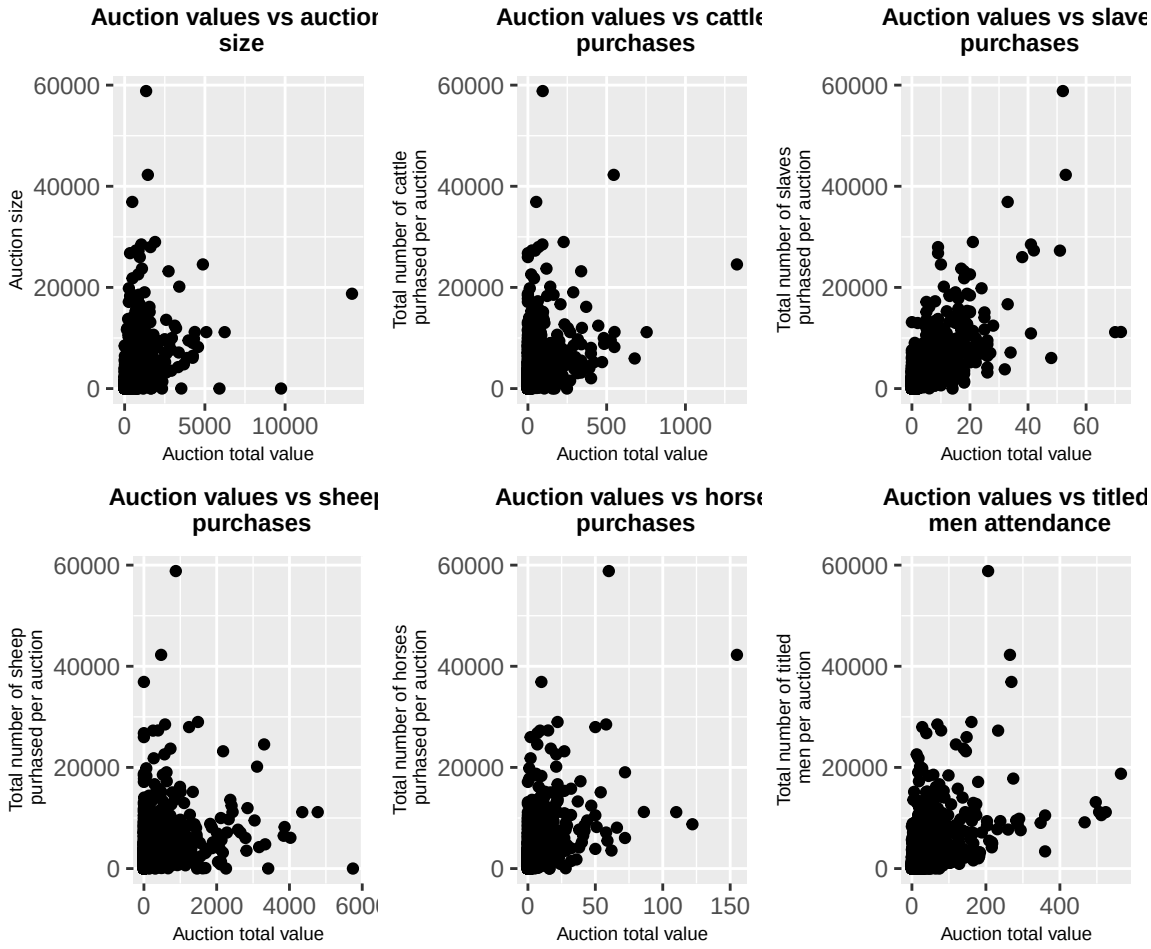
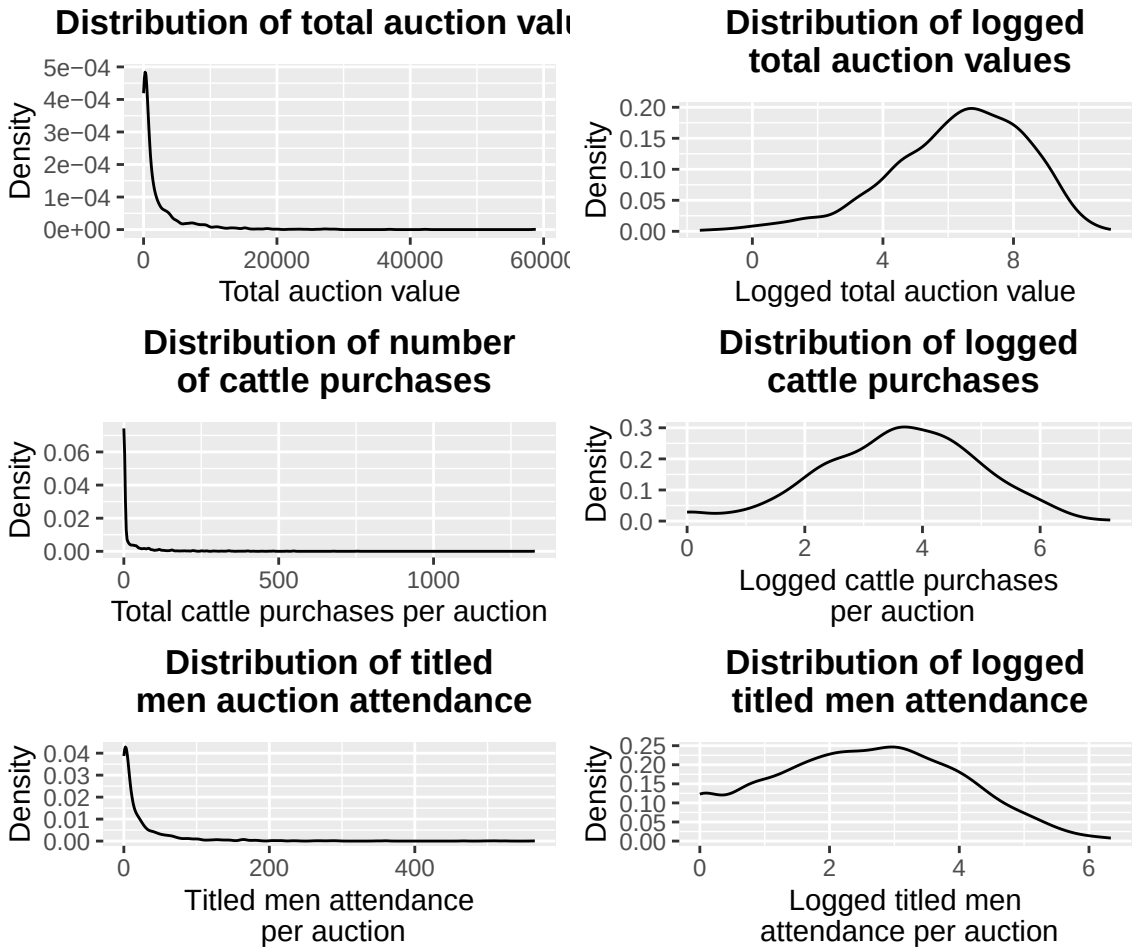


Figure 8 shows how certain good type purchases relate to total auction values. This shows that the data is heavily clustered close to 0 due to the small number of tagged goods present at each auction.



4. Target and feature engineering

The target in this project is the auction total value. The features used to predict the target include: the total number of items within a good type per auction, total number of items within an auction (auction size), number of titled men and women in attendance per auction and the year of auction. Figure 9 provides evidence that the features are skewed left. This is solved by applying a log transformation to all features (except for the date).

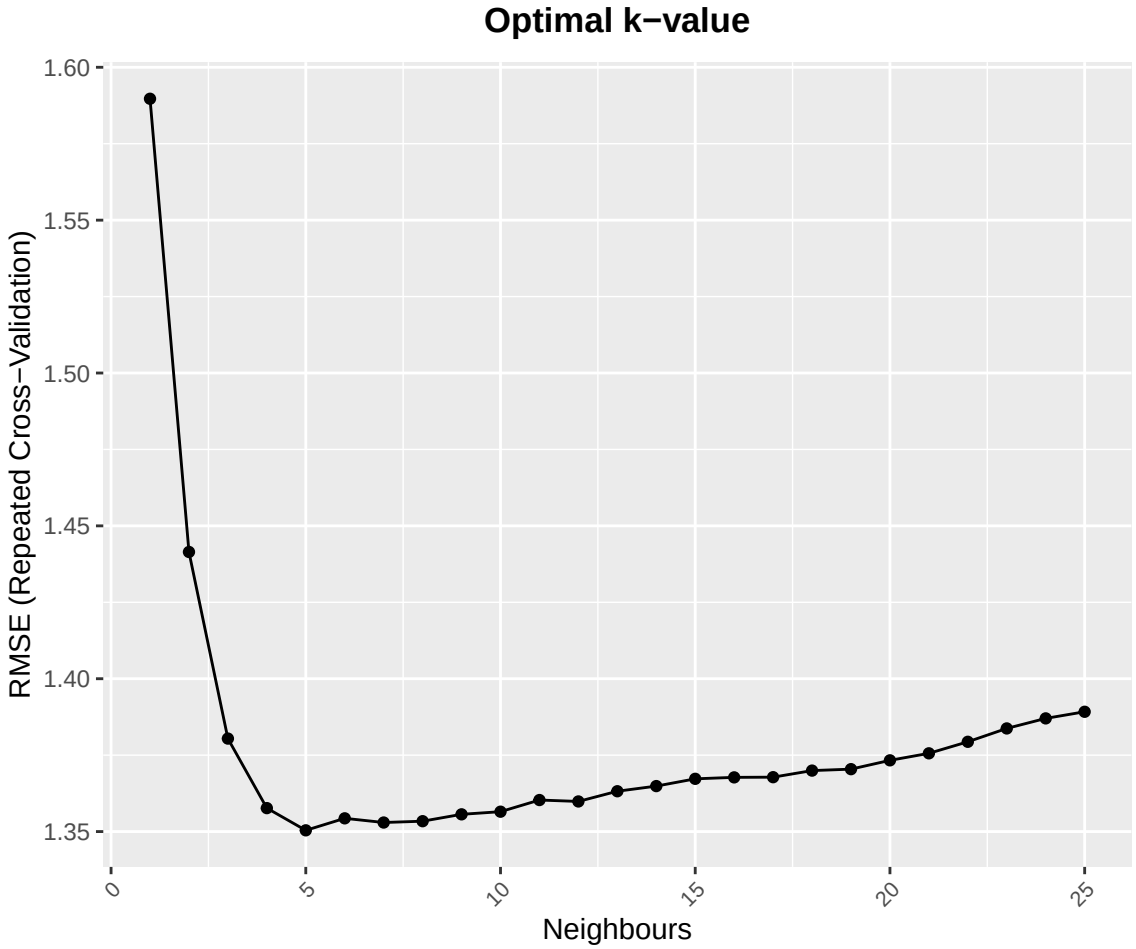


5. Methodology

This project aims to predict auction values based on different types of goods present at an auction. Predicting a continuous, numerical value presents a regression problem. As a result, the k-Nearest Neighbours (KNN) model and a random forest model will be used. A linear regression model will be used to determine how well machine learning models compare to a regression. The metric used to assess the quality of the machine learning predictions will be root-mean-squared-error (RMSE). This metric indicates how far predictions are from actual values on average. Additionally, r^2 will be used to determine how well the model fits the data. A default model will be run for each model before conducting hyperparameter tuning to improve the model.

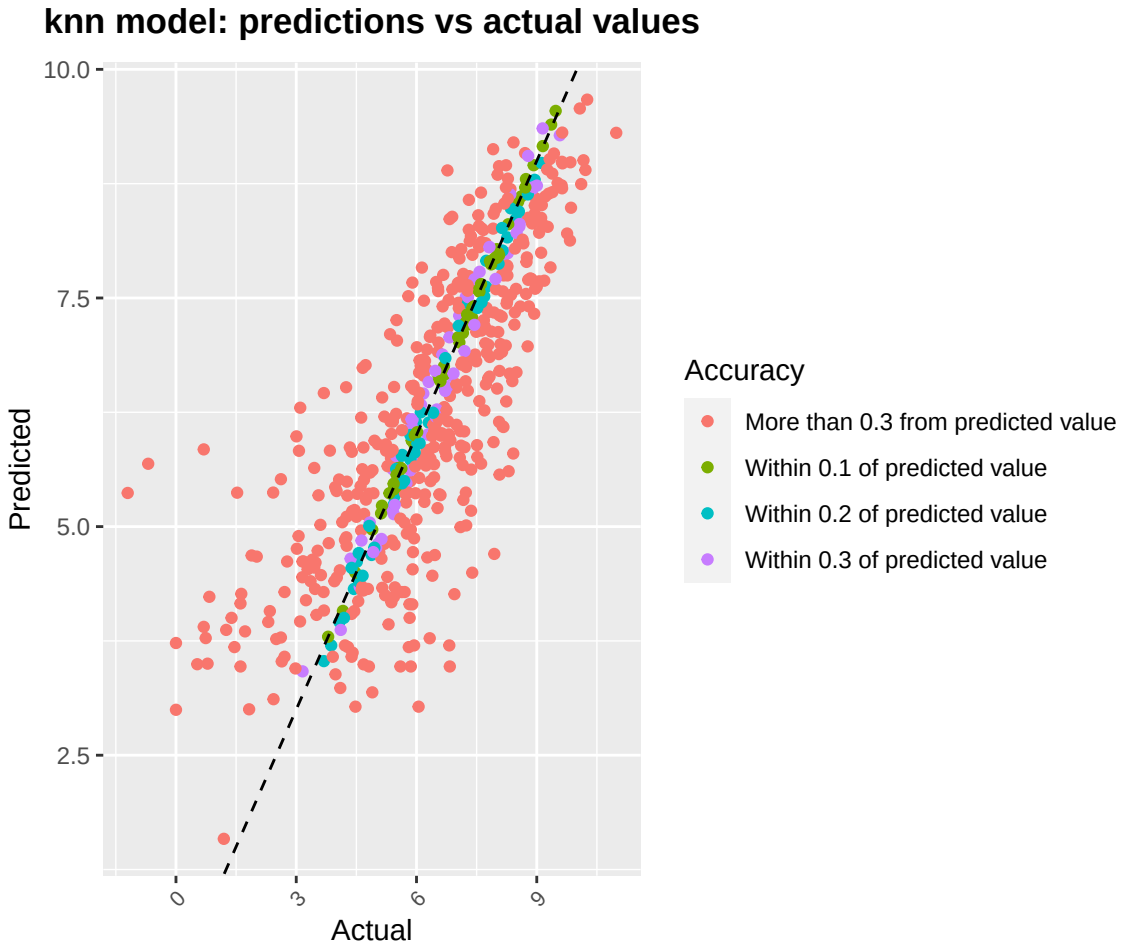
6. KNN model

The KNN model generates predictions for a new observation based on matching it to the most similar training observations and aggregating their values. When assessing a new observation (auction), it will compare the bundle of goods present to other auctions with the most similar bundles of goods sold and average their total values to determine a prediction. Hyperparameter tuning is required to set the optimal value of k. Figure 10 shows that k=5 is the optimal value. New observations predictions will be the average of its five most similar auctions.



With the training data set, RMSE is 1.35 and r^2 is 0.59. When the model uses the testing data,however, RMSE reduces to 1.17 and r^2 increases to 0.83. This would indicate that the predictive accuracy and model fit improved between the training and testing data. It is important to note that it is a problem that RMSE decreases and r^2 increases when moving from the training to the testing dataset. One would expect to see the opposite. Possible reasons for this may include data

leakage. After checking for this, I could not come to an explanation for why this was occurring.The mean logged auction total is 6.32 which means that the average difference between predictions and actual values is 1.35. As a proportion of the mean total auction value, this figure is quite large. Figure 11 shows that a very small proportion of predictions are within 0.3 points of the actual value. This indicates that knn is a poor model for predicting total auction values.

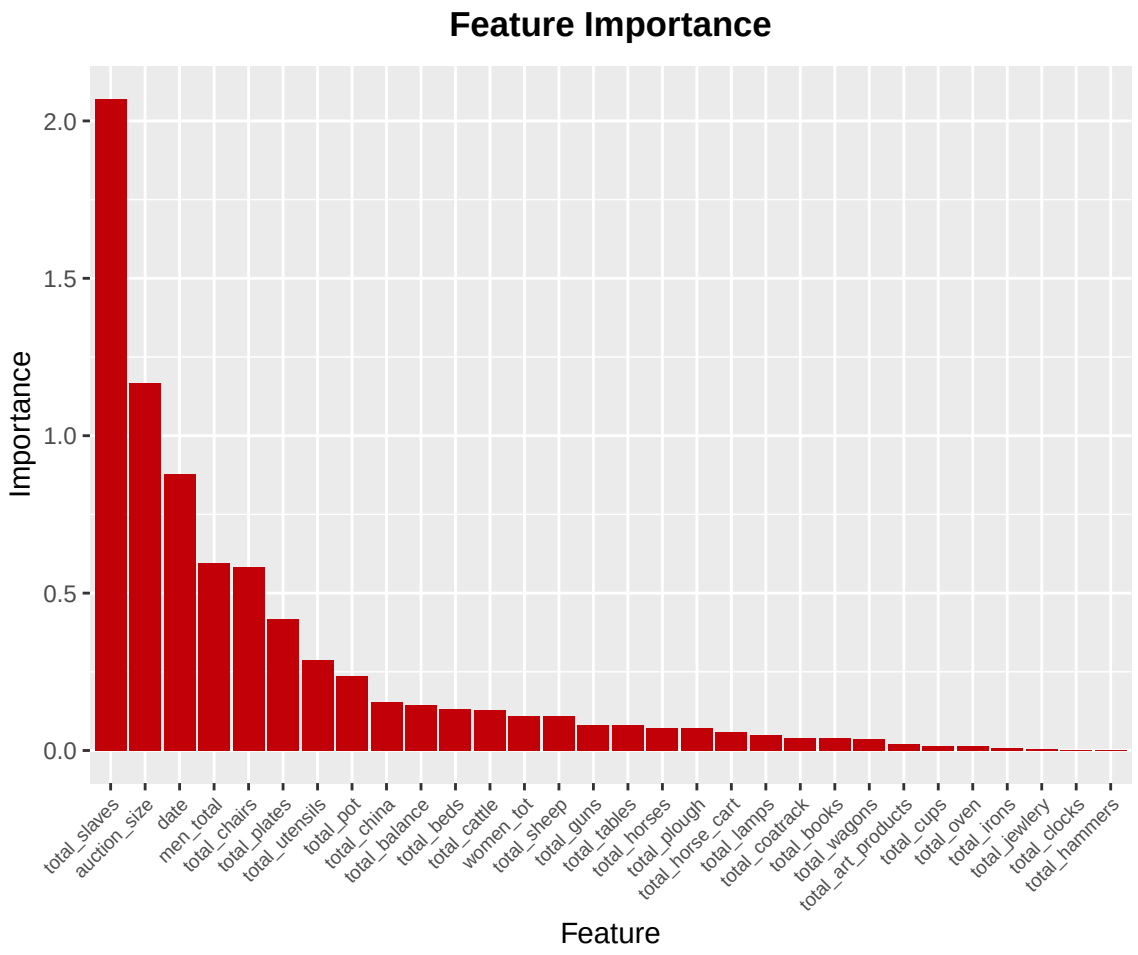


7. Random Forest Model

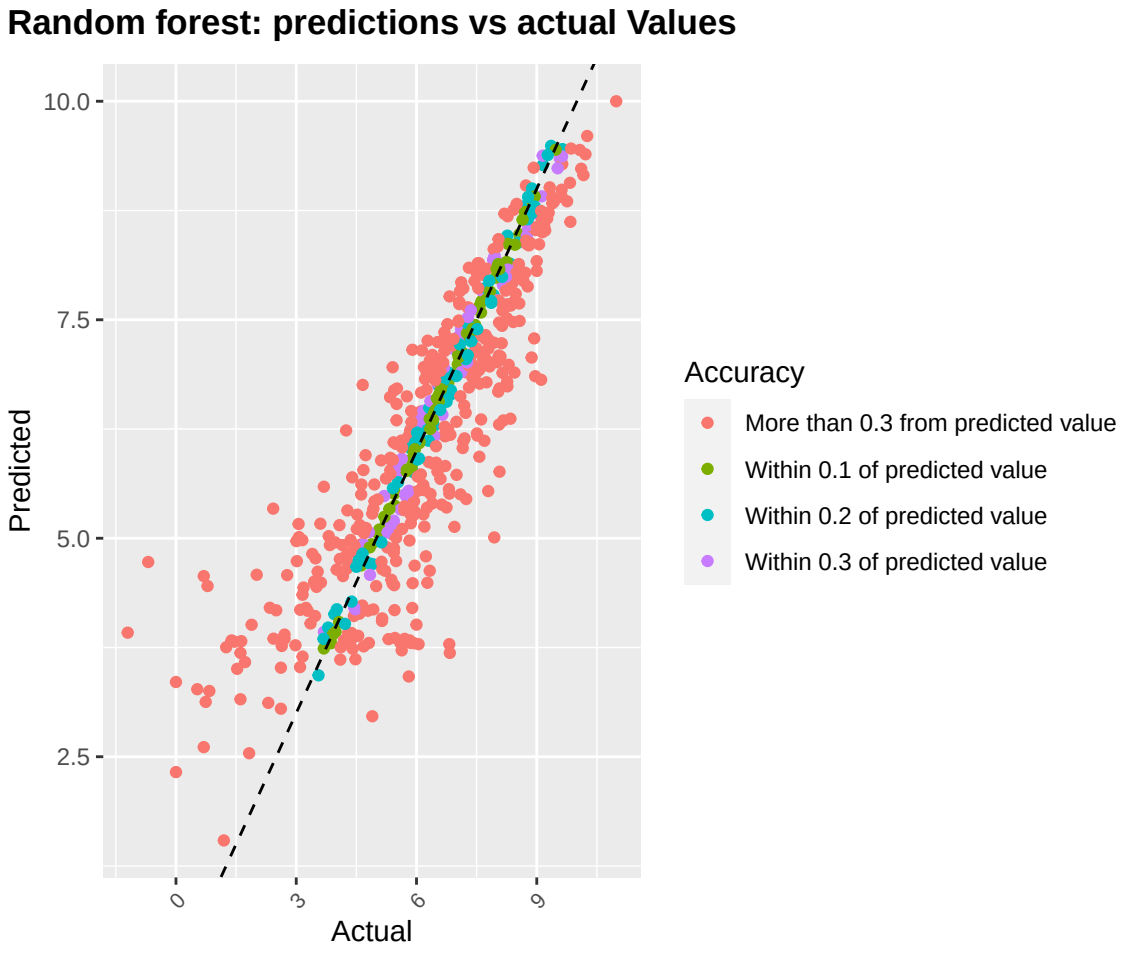
Random forest models include multiple decision trees to make predictions. Decision trees are built independently on a subset of data to introduce diversity among trees and reduce overfitting. Additionally, random features are selected at each node of the tree to cover a wider breadth of the data and include more robust predictions. Once the decision trees are trained, predictions of each tree are aggregated to generate predictions. The randomness of data and feature selection decreases variance in predictions and aggregating tree-level predictions minimises bias introduced within trees. Tuning

parameters such as the number of trees within the random forest and the number of features randomly selected at each node should improve predictions. Furthermore, the node size can also be tuned as it refers to the minimum number of observations required at a terminal node to ensure another split.

The default model includes 500 trees, 10 features randomly selected at each split and a target node size of 10. The RMSE is 1.1 and the r^2 is 0.73. A random forest model offers information on each features' importance in generating predictions. It is clear from Figure 12 that the number of slaves at an auction is the most important feature, followed by the size of the auction, the year the auction took place and the number of titled men in attendance. These features make sense within the context of the Cape Colony.



After completing a grid search to tune parameters, the default model is found to perform the best. When moving from the training to the testing data, RMSE increases to 2.68. Figure 13 below shows that only a small proportion of predictions are within 0.1 of the predicted value. Based on RMSE scores, the knn model performs better than the random forest model as its RMSE is 1.17.



8. Linear regression

A linear regression is shown below. The RMSE is 1.13 and the r^2 is 0.71. This linear regression model therefore performs better than the machine learning models with a higher prediction accuracy and better fit.

9. Critiques

Although the linear regression performed better than the knn and random forest model, all three models have low predictive accuracy. It is clear that the subset of goods used to predict total auction values, are insufficient predictors of total auction values.

10. Conclusion

Predicting total auction value using a subset of goods present at auctions resulted in very inaccurate models. The knn and random forest models were tested and performed poorly given the constraints presented by the features chosen. This highlights the importance of choosing appropriate features when using using machine learning techniques.