

Report

Question

Use the information in the tables to reduce the number of dummies that will be used in the model.

Answer

Based on the distribution of mean value of competitive auctions we find that:

The categories "Books", "Clothing/Accessories" and "Toys/Hobbies" can be merged into one.

Then categories "Antique/Art/Craft", "Collectibles", and "Music/Movie/Game" can be merged into one.

Then categories "Automotive", "Coins/Stampss", "Pottery/Glass", and "Jewelry" can be merged into one.

Then categories "Business/Industrial", "Computer", and "Home/Garden" can be merged into one.

Then categories "Electronics", "Photography", and "SportingGoods" can be merged into one.

Finally categories "Health/Beauty", and "EverythingElse" can be merged into one.

So, we can select one dummy from each of the 6 different sets formed above
We will choose - "Books", "Collectibles", "Automotive", "Computer", "Electronics" and "EverythingElse" as our reduced features

Now, for currency, EUR and US can be merged. So we choose US, and GBP as our reduced features

Now, for endDay, {Fri, Sat, Sun and Wed} can be merged. Then {Tue, Thu} can be merged

So we choose Fri, Thu and Mon as our reduced features

Finally for Duration, {1, 10}, {3, 7} and {5} are different sets
So we choose 1, 3 and 5 as our reduced features

Question

How does this model compare to the full model with respect to predictive accuracy?

Answer

This model performed poor (with ~62% accuracy) in comparison to the full model (with ~78% accuracy)

Question

Interpret the meaning of the coefficient for closing price. Does closing price have a practical significance? Is it statistically significant for predicting competitiveness of auctions? (Use a 10% significance level.)

Answer

Now, for the ClosePrice, the coefficient value was 1.396e-01 with a p-value of 2e-16 which is way less than 0.1 significance level. This means that the ClosePrice is highly statistically significant for predicting competitiveness of auctions

Question

Find the model with the best fit to the training data. Which predictors are used?

Answer

"Category_Automotive", "Category_Electronics", "Category_EverythingElse", "currency_US", "currency_GBP", "endDay_Fri", "endDay_Mon", "Duration_5", "sellerRating", "OpenPrice", "ClosePrice" are the predictors which are used to find the model with the best fit to the training data.

Question

Find the model with the lowest predictive error rate. Which predictors are used?

Answer

For validation models, following predictors were selected:

Category_Collectibles + Category_EverythingElse + currency_GBP + endDay_Thu + endDay_Mon + Duration_5 + sellerRating + OpenPrice + ClosePrice

Question

What is the danger of using the best predictive model that you found?

Answer

The danger of using the best predictive model is that the true performance is different (can be greater or less) because you need to test it on data that has not been seen before. The performance on validation data can be overly optimistic when it comes to predicting performance on data that it has not been exposed to the model at all. When the validation data is used to select a final model, you are selecting based on how well the model performs with those data and not incorporating some of the randomness of those data into the judgement about the best model. There are also practical issues such as costs of collecting variables, error-proneness, and model complexity into the selection in the final model.

Question

Why the best-fitting model and the best predictive models are the same or different?

Answer

They are different because In a best-fitting model, the "best" model is one that satisfies the criteria of lowest AIC values. In other words, it is only "best" in terms of lowest AIC. As soon as the AIC criteria is removed, "best" no longer exists.

Now, for the best predictive models, the "best" model is one that satisfies the criteria of lowest predictive error rate. So, the best-fitting model and the best predictive models are different Predictive performance wants to know how well the model predicts new data. Meanwhile, the best fitting model wants to know how well the model fits the data it was trained with.

Question

If the major objective is accurate classification, what cutoff value should be used?

Answer

When picking the cutoff value, you can look at the ROC curve. If the threshold is really high, we will classify very few cases as the factor we are looking for which is competitive. Based on the ROC curve, have a sensitivity of 0.80 and 1-specificity of 0.20 will be optimal. This responds to a cutoff of 48%.

Question

Based on these data, what auction settings set by the seller

(duration, opening price, ending day, currency) would you recommend as being most likely to lead to a competitive auction?

Answer

Based on these data, we would recommend "opening price" as being most likely to lead to a competitive auction.