# Auction Price Predict

## Current Progress

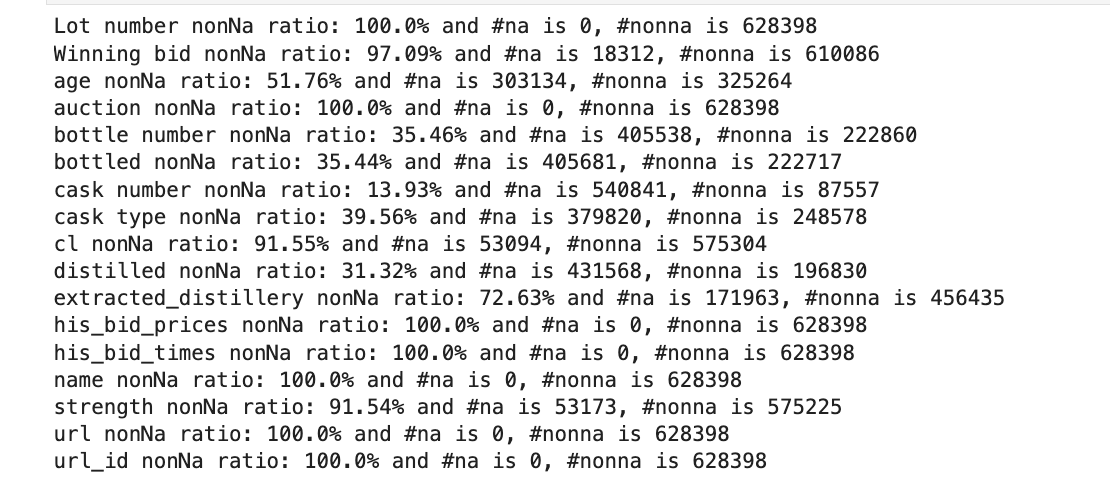
This paper focuses on the task of predicting auction prices for alcohol. Specifically, we need to collect openly accessible auction records and product information from the internet as data sources to predict auction prices. Accurate price prediction can facilitate transaction decision-making and auction management.

From the initiation of this project up till now, our primary focus has been on building a dataset by crawling open-source auction data. This process involves several steps including: webpage parsing, extracting relevant text, extracting information closely related to value from the text such as age, strength, volume, production location, historical price, brewing date, and the number of bottles, data cleansing, and expansion and merging of data sources. We have already collected over a million data samples, each with the above key features. In the next phase, our focus will be on further data cleaning to remove noise and sample duplication based on products to prevent leakage of label from historical order information.

For the price prediction task, if historical prices are disregarded, it can be modeled as a regression task. In this case, we can only use product information as features to predict numeric target using regression machine learning models or DNNs. However, a noticeable issue is that the same type of product can have different selling prices at different times. The only difference in their product information are a few features such as age and number of times auctioned, which may lead the model to excessively focus on a few time-dependent features. Also, fluctuations in prices may not be fully accounted for by product information alone. Hence, it is similar to stock price prediction where the value of the stock and market expectations i.e., historical prices are determinants of the stock price. Therefore, we also incorporate historical selling prices. With this data, the task is modeled as a sequence prediction where the model needs to handle both time sequence information and static features. This allows it to combine the value and market factors to complete price prediction, with CNN, LSTM, Transformer, and other models as potential options.

## Dataset

1. Data source website：scotchwhiskyauctions.com, whiskyauction.com
2. Number of samples：60w + 49w
3. Selected eatures：age, bottlee number, distilled time, bottled time, volume, strength, distillery, cask time, current auction time, history trading prices, history trading time, auction times.
4. Predict target: Current Winning bid
5. Partial numerical feature statistics for website1(scotchwhiskyauctions.com)

* 
* 2011688729525\_.pic

## 1981688729510_.pic

## Model

First, we will discuss the feature engineering before model training. For continuous numerical features, after removing outliers, they can either be used directly as continuous features or converted into discrete features through bucketing. For discrete categorical features, the number of categories can be reduced by appropriately classifying and merging, thus reducing the noise introduced during information extraction. After these processes, dimensionality reduction techniques like PCA, feature selection, etc., can be used to aid the model's learning.

Dataset Splitting: First, we need to avoid the label leakage mentioned earlier. Next, we can choose a more recent portion of the auctions as test data, or extract equal numbers of auction records from different years to create a time-balanced test set.

Baseline Model（ML&DNN）：Linear/Lasso Regression， Regression Decision Tree， Multil-layer Neural Network

Sequential Model（DNN）：CNN、RNN、LSTM、Transformer-based Model

For the regression task, we can choose the following metrics: R2 score, MAE, MSE. Additionally, eliminating extreme values can prevent unstable evaluation results. Furthermore, the task of predicting price can be turned into a classification task by binning the predicted prices, which reduces modeling difficulties. In this scenario, multi-class F1, ROC, ACC, etc., can be used as evaluation metrics. Moreover, due to differences in historical data, the difficulty of predicting will vary. Therefore, the prediction results can be reported in sections according to factors such as the number of transactions and the trading periods.