**Topics/areas you can cover in your note:**

* How did you approach this challenge & Any other insights/features you created to solve the problem

To be honest, I only started working on the problem in last hour (which I am going to repeatedly mention many time further to look less stupid for not trying out many interesting or obvious methods :-) ), which means valuable loss of time for exploratory analysis and any chance to try lot of different methods. So it pushed me to work on only one approach, which I can think of as best under the constraints.

1. First of all, considering it is a multi feature problem I ruled out the attempt to go for any ARIMA models.

2. Next, it was pretty clear that I had to create date related features. As it was trading data, based on my experience weekday & quarter features might prove more effective. Month day was tempting, but I felt only beginning and end of the month might uncover some patterns. So I didn't tried that.

3.I start any problem with summarizing the problem details. Under the wraps of all details, I strongly believe there will be simplest form of problem statement to be framed. I approached this problem also in same way.

This is how I summarized this problem.

There is a demand and associated supply value with it. So the key question is what is the driver? Volume or Price?

In reality both factors pull & push each other to work for equilibrium over the period of time.

1. So one approach might have been (which is again a hypothesis to be tested!) to model both ways, once predicting volume first followed by price

2. Then other way around. Check how they perform and choose best one in cross validation.

3. Another approach is we can ensemble them and check how they work.

Now as I was racing against time, I had to little bit modify the way I ask question. “Who might be the bigger driver?” Here is what I thought will be the most suitable hypothesis to check.

Volume is the indicator of market demand. I tried to check generic hypothesis that “volume indicates drive in price”. It means, I will create model 1 which will be model for volume without price, then model b for price with volume. Scoring follows the same chain.

* Data pre-processing steps (univariates /multi-variates / data cleaning & data preparation)

Pre-processing:

Reordering: Test & Train data were not alligned columnwise to create combi, so had to re arrange it.

Data Type Fixes: First task was as per data definition make sure all variables are in appropriate data types.

Removal of redundant column with no functional value: I removed the redundant ID column. After creation of derived features from DateTime, I even removed it from final model features.

NA handling: I also found that Category 2 had NA. So initially I tried with best approach of predictive replacement using missforest package. But unfortunately the missforest package can't handle the levels present in other columns. I might have tried other algorithms, but I didn't sweat on that here.

So I left with the approach of replacement of NA with Mean. I missed the part of digging in to NA case analysis here due to time constraint.

Standardization temptation: The temptation to standardize the volume before modeling for price lingered around for a while to use and test, but I left it to live some other day as it made me to ask the sanity of the question, considering we have time varying factors which might be uniquely behaving each values. So what is the effect of overall standardization in this case or does that question even makes sense? Felt better to not do it, instead of blindly doing it. Marked it for stack overflow questions, if not found answer myself.

* How did you select an algorithm to predict the day and item wise price & number of products sold

There are multiple parts here.

1. Which package I used?: H2o is my go-to package for its ability to handle large quantity of data in short time.

2. Which algorithm I used?: GBM has been my favorite choice of algorithm when working with the large data. The reason is its ability to be accurate with trying its best to maintain bias-variance trade off with the help of some hyper parameter tuning.

3. Cross Validation(CV): I have seen in this problem CV RMSE closely followed the train RMSE. It saved some valuable time on computation for me.

4. Hyper parameter Tuning: I am aware that there is nothing called one size fits all. But as per h2o documentations & based on my past experience, I have given values here. Ideally I go for random grid search for finding the parameter values.

As given in hypothesis, I first created model to predict price leaving out volume and next created the model

After competition I tried again doing multifold cross validation in datascienceworkbench lab which provides me computing power that I need to have for my h2o. CV behavior remained consistent.

When I uploaded I was is 41'st rank in leaderboard. I still had some time to tryout new things, fit more. But I trusted my CV scores. After the competition, I found I was in second place.

Last notes:

Not bad, but was not satisfied by the fact of large gaps I left behind to try out. I will work on them. I am curious to see new things from the legend, Jason of machine learning mastery on time series. Hopefully soon I might be able to learn and apply them, if they prove to be beneficial.