DecisionNext Candidate

Data Assignment

Description

The following assignment is intended to evaluate some basic skills of data manipulation, exploratory analysis, and simple ML concepts oriented towards forecasting commodities. We will use beef prices as an example. We expect you to spend between 1 and 2 hours on this to get as far as you can. If you do not understand a question (e.g., “long tails”) or feel that it would take you too much time, that’s ok - just skip it and get as far as you can. Send a copy of your work, and then we will meet with you in an informal discussion where you can present your findings. Do not feel like you have to do all parts of the exercise or spend time creating a formal presentation. We do encourage visualization of individual results in charts and tables, however. Use of Python is encouraged for Science candidates (e.g., a Jupyter notebook), but you are free to use Excel and/or any modeling or presentation tools that you are comfortable with.

A. Data extraction

A.1. Get list of reports from USDA datamart

The following link documents market reports accessible via the datamart of the Agricultural Marketing Service of the US Department of Agriculture (USDA). Let's call this URL the base URL:

https://mpr.datamart.ams.usda.gov/services/v1.1/reports

Exercise: Form a table or list of dicts with the content of the web page.

A.2. Get report content

We will work with the report titled: "National Daily Boxed Beef Cutout & Boxed Beef Cuts - Negotiated Sales - PM (PDF) (LM_XB403)". This report is published daily under the Beef section of the Datamart website and API.

Exercise: Describe the contents of this report.

- What is the business significance?
- How are these values measured?
- What are the units?

- What is the difference between Choice and Select?

Exercise: Find the LM_CT115 report under the Cattle section of the Datamart. Describe the contents of this report, including the business significance and units of the fields.

Exercise: Define startDate as Jan 1, 2016 and endDate as Dec 31, 2023. Download the content of LM_XB403 as CSV and save as a text file. Extra credit if coding: use the API rather than the website.

A.3. Extract and format series

Exercise: Load the data into your chosen analysis framework.

Exercise: There are a lot of cuts. Filter down to 3 Choice cuts only (6 series including price and volume) and rename to a more readable format:

- Loin, tndrloin, trmd, heavy (189A 4) => tndrloin
- Loin, top butt, bnls, heavy (184 1) => butt_bnls
- Loin, top butt, CC (184B 3) => butt_CC

Exercise: format dates to ISO standard YYYY-MM-DD

B. Data exploration

We are ready for some exploratory data analysis. Let's get a feel for the series at hand.

B.1. Aggregation and missing values

Exercise: Aggregate the 6 series to Weekly. Explain your chosen method for each of the 6 series.

Exercise: Examine if the series have missing values.

- Which would you discard? Why?
- Which would you fill? How?

Exercise: Fill out everything at least since 2017 following your chosen method.

- Discard history before 2017.

B.2. Visualization:

Exercise: using a simple line plot, chart the 6 series.

- Are there periods of extreme values?
- Outside of those periods, is there any visible seasonality?

Exercise: Visualize the distribution of price values for the series tndrloin.

- Does the distribution of values have “long tails”? What does that mean?

B.3. Correlation:

Exercise: Generate a table of correlation values for the 6 series.

- Describe the meaning of these values.
- Discuss causality and whether you have any hypotheses for significance.

Exercise: Generate scatter plots and trend lines for each price-volume pair.

Exercise: Generate scatter plots and trend lines among the 3 price series.

B.3. Extreme values

The USDA data has been vetted for outliers caused by reporting errors, but there are periods where beef prices have been truly extreme.

Exercise: Using some basic descriptive statistics, can you identify the periods with values that seem extreme or abnormal?

C. Statistical Modeling

C.1. Linear Regression with full history

Construct a basic statistical model based on linear regression.

Exercise: Standardize all series and rename with suffix _scaled.

Exercise: Regress series _tndrloinscaled against _butt_bnlsscaled,_butt_CCscaled and an intercept.

- Question 1: Is the resulting model better than fitting a constant?
- Question 2: Are all regressors significant? Does the answer contradict Question 1?
  Exercise: Drop one of the regressors and fit again.
- Question 3: Has the goodness of fit improved, deteriorated, or stayed approximately unchanged?
- Question 4: Are all regressors more or less significant than before? Hint: check how the condition number has changed.
- Question 5: Are there any known deficiencies with the prior model with respect to bias or efficiency with this model?

Exercise: Create a simple plot of the fit vs actual.

C.1. Seasonal model with full history

Exercise: Create a seasonality and trend model for tndrloin.

Exercise: Describe the pros and cons of your chosen seasonality and one or two alternatives.

- Question 1: Is the resulting model better than fitting a constant?
- Question 2: Are all regressors significant? Does the answer contradict Question 1?
  Exercise: Drop one or more of the regressors and fit again.
- Question 1: Has the goodness of fit improved, deteriorated, or stayed approximately unchanged?
- Question 2: Are all regressors more or less significant than before? Hint: check how the condition number has changed.

Exercise: Create a simple plot of the fit vs actual.

C.2. Estimation using holdouts

Let's use one of the previous models for out-of-sample predictions. You can choose any of them.

Exercise:
1. Split data in two subsets: before 2023 (train set) and 2023 (test set).
2. Fit OLS model on train set.
3. Predict on validation test set.
4. Compare to actual in 2023 to calculate the out-of-sample error.

C.3.(Optional) Cross-validation loop.

Question: How would you document the expected accuracy of the model going forward?. Describe succinctly a strategy.

D. Submission

If you are executing this exercise in Python or R, please create a repository on the Github free service, make the repo private and give us access. We will pull the code from there and run the solution. We will review together your approach on a call.

If you are using another tool, please format the results and email to us so we can review before and during the call.