Automated time series project

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2019 — c4a1f0eb

0.1 Context

I have been working on time series related problems for the last year and a half. Here, I want to share my experience as to how we've been doing time series projects and how I envision automating the whole process.

0.2 Learned in this study

0.3 Things to explore

1 Overview

- Define data sets
- Define forecast horizons
- Define data traversal procedure
- Define metrics

2 Definitions

- Entity: An object to which we relate measurements and can be considered a unique identifier, generally of the categorical type, e.g., the SKU of a product, a person's full name, a UUID, a city, etc.

3 Process

Here I want to describe the whole process of a project in order to document it. The process of automating an existing process always starts by writing down what the existing process is.

- The client comes to the company (or the company goes to the client) because they want to start using ML or AI in one of their business practices
- Sales need to establish with the customer what their needs are, what data they have available, the level of technical expertise of the client, etc.
- A project is defined and agreed by both parties (the client and the company). Timelines are defined in order to give a rough idea of when the system is expected to be deployed in the client's environment.
- Different sources of data are **collected** by the client. These sources might be Excel documents, data in an SQL database or NoSQL database, data stored in a third-party provided such as Google Analytics/Ads, FB ads/social, Twitter, in their enterprise data lake, etc.
- Data is **documented** to prepare it to be used by individuals which may not have experience with it. Most companies do not document their data because this information is known within the context of the business.
- Data is **cleaned**. Some data may be invalid, data might be missing for certain fields, data we want to evaluate a model against might have been entered after the fact (a big no-no for time series forecasting)
- Data is **aggregated** and **merged** together. Most of the time it is not as straightforward as having a foreign key to associate two tables entries. It might be necessary to write complex rules to join two

datasets together. Information that should not have been available at certain point in time might become merged by the procedure, introducing errors in information availability.
- Data is **preprocessed** based on the objectives of the project. Preprocessing is the process of creating features using the data provided. Features can be as simple as lagged values (what was the value yesterday? two days ago?, last week?), moving average (what is the mean of the last 2 days? last week/7 days? last month?), converting categorical data to one hot encoded columns, min/mean/median/max/quantiles over all the previous data, either per entity or overall.

4 Automated steps

- Merge datasets into a single dataset
- Detect datetime-like columns
- Detect covariates and non-covariates columns
- Convert categorical columns into either one hot encoding or embeddings
- Generate features (lag, moving average, exponential moving average)
  - On covariates
  - On non-covariates
- Backtest multiple models

5 Notes

- In an ideal world, I want to do the following:
  - Drag-and-drop many csv files
  - Select the column that indicates the datetime index (out of columns detected as being datetime)
  - Include/Exclude columns from the data
  - Select the column(s) I want the model to learn to predict
  - Select the loss function (what to optimize)
  - Select the metrics to compute
  - Wait for the system to process my data and run it against multiple models
  - Show me the different models and their results
    - Let me visualize the worst offenders (the rows keeping the loss high/non-zero)

6 See also

7 References

- https://www.datarobot.com/