Modeling

panko.aliaksandr

November 2020

Contents

1 What is Modeling

Modeling is a replication of a process or an object.

Example 1 (Object Modeling):

- Modeling of a building
- Modeling of a car

Example 2 (Process Modeling):

- Modeling of yearly temperature
- Modeling of stock returns

The result of modeling is a model, which replicates the required features/behavior of an object/process and allows to execute experiments and draw conclusions.

There are 2 model types:

1. Physical (Real object: small car, small version of a generator etc.)
2. Mathematical (Statistical Model represented as a set of Mathematical Equations)

Modeling in both cases required 2 essential steps:

1. Create a model
2. Validate a model (Make sure that the model indeed has all required features/ demonstrates required behavior)

After a valid model is created an analyst can execute experiments to get additional knowledge about an object/process (including predictions)

2 Statistical Modeling

Financial Modeling models financial processes using mathematical equations.

Example of Financial processes:

1. Model probability of default
2. Model stock returns

There are 2 types of Financial Data:

1. Cross-sectional (only one time point)
2. Time-series

Based on these 2 types of data 2 different types of model can be created, resulting into 2 types of analysis:

1. Cross-sectional analysis (Ex. probability of default with logistic regression)
2. Time-series analysis (Ex. stock returns with ARIMA)

Regardless of a model type an analyst needs to execute the same 2 operation: Create and Validate a model.

3 Model Creation

Every model consists of 2 parts:

1. Model Input
2. Model Output

Output variable contains the information an analyst is interested in. The goal is to be able to somehow predict the output variable. Let's consider an example. A seller needs to know how much items to buy/produce and than sell maximizing the profit. Output variable would be "Number of purchases")

The output variable is always known

The problem is to find Influencing Factors or Independent Variables.

3.1 Factors Selection

First of all the knowledge of a domain is very important. This knowledge helps to come up with a set of possible factors and later avoid overfitting checking if a suggested factor has some rational explanation.

First of all let's discuss the definitions:

3.1.1 Definitions

- **Feature Engineering** - is a process of designing more meaningful features(factors) from original raw data using the knowledge of a domain. Normally it also involves pre-processing. New features can be derived in any possible **mathematical** way from the original raw data. It is the most general term. However, there is another view. According to the other opinion, FE is a step between raw data and dimension reduction and consists of operations such as:
  - Indicator Values
  - Feature Interactions
  - Feature Representation (ex. convert datetime to dayOfaWeek or binding in intervals)
  - Adding completely new data
  - Log Transform or Power
- **Feature extraction** is usually used when the original data is very different. In particular, when you could not have used the raw data. E.g. original data were images. You extract the redness value, or a description of the shape of an object in the image. However, it also means any transformation with original data to reduce the dimension. The names exists to separate from Feature Selection.
- **Feature Selection** - the process of dimension reduction by selecting a sub-set of original raw data (no construction)
- **Factor Analysis** - dimension reduction process, which is aimed at finding independent latent variables that can explain the variance in dependent variable. Feature Analysis is a sub-set of Feature Engineering since Factor Analysis methods can create new features from raw data only in some **pre-programmed** way (E.g. PCA creates a **linear combination** of raw data). Feature Engineering process also uses expert knowledge of domain, so new factor can be designed from head (not be pre-programmed).

Hierarchy Resume:

- Feature Engineering is the most general term. Or a step before Feature Factor Analysis/Feature Extraction in more narrow sense.
- Feature Extraction = Factor Analysis (no difference found)
- Feature Selection - sub-set of FE and FA

3.1.2 Methods

The first step is Feature Engineering in narrow sense.

- Indicator Values
- Feature Interactions
- Feature Representation (ex. convert datetime to dayOfaWeek or binding in intervals)
- Adding completely new data
- Log Transform or Power

The result of this step is set (potentially big) of features which logically make sense and statistically have better properties.

The second step is dimension reduction, which is required to speed up the ML process and reduce the noise. The step is represented by first Feature Selection and then Factor Analysis (Feature Extraction in general sense)

Feature Selection is done first, because this method remains initially selected and fully understandable set of features. There are 2 main tasks:

1. Eliminate Redundant Factors (unsupervised models)
2. Delete Irrelevant Factors (supervised models)

The methods are:

1. Correlation Analysis (delete highly correlated) (Filter)
2. Variance Analysis (eliminate with lowest variance) (Filter)
3. Recursive Feature Elimination(RFE) (Wrapper)
4. Forward Feature Selection (Wrapper)
5. Backward Feature Elimination (Wrapper)
6. Exhaustive Feature Selection (tries all combination) (Wrapper)
7. LASSO Regularization (L1)
8. Random Forest Importance

The last step in dimension reduction can be Factor Analysis, namely farther feature compression. The methods are:

1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)

The is one problem with these methods: new features are not really intuitive...

Let's assume that we finally constructed a model, now it is time to validate it!

4 Model Validation

Statistical Validation involves:

1. Goodness of fit
2. Residuals analysis (must be random)

Methods:

1. KS Statistics (+ the confusion matrix)
2. Holdout (just split into train and test)
3. K-fold cross-validation (train and test are chosen many times from the same set of data points)
4. Bootstrapping

Backtesting is just a particular type of Holdout method (simply split in train and test), where observations are time-series elements. There is one special problem with time-series models - look ahead bias. It occures when one includes new information into a model and than test the model in past, when the included information was not available yet.