FEATURE ENGINEERING

TOPICS

- What is Feature Engineering?
- Why Feature Engineering?
- How to apply Feature Engineering?
- Discussions on various scenarios

WHAT IS FEATURE ENGINEERING?

Feature engineering is the practice of using mathematical transformations of raw input data to create new features to be used in an ML model.

- Dividing total dollar amount by total number of payments to get a ratio of dollars per payment.
- Counting the occurrence of a particular word across a text document.
- Computing statistical summaries (such as mean, median, standard deviation, and skew) of a distribution of user ping times to assess network health.
- Joining two tables on user id.

WHY FEATURE ENGINEERING?

TRANSFORM ORIGINAL DATA TO RELATE TO THE TARGET

- A personal finance dataset that contains the current bank account balance and credit debt of each customer.
- If you are building a model to predict whether a customer will become delinquent in payments three months from now, then engineered feature of

  ratio-of-debt-to-balance = amount-of-debt / amount-of-balance

- In this case, the ML model will have an easier time of finding delinquency if the engineered feature is directly used as input.
- This will result in improved accuracy of predictions.

BRING IN EXTERNAL DATA SOURCES

- Imagine that you run an internet subscription service.
- The first time each customer logs in, you want to predict the lifetime value of that customer.
- Among other variety of metrics, you could capture the geographic location of each customer.
- You can do better by bringing in third-party demographic data.
- For example, this would allow you to compute the average income and population density of each user's location and to insert those factors directly into the training set.
- The feature engineering of location into income and population density enables you to asses which of these derivatives of location matter the most.

USE UNSTRUCTURED DATA SOURCES

- Feature engineering enables you to use unstructured data sources into ML models.
- Many data sources aren't inherently structured into feature vectors that can be directly inserted into the ML framework.
- Examples: text, time series, images, videos, audios, log data, etc.

CREATE FEATURES THAT ARE MORE EASILY INTERPRETED

- Feature engineering empowers ML practitioners to create features that are more interpretable and actionable.
- Consider a simple example of machines that manufacture computer hardware.
- You could use the raw machine data, such as measurement of signal response and other processing signals, to build ML models to predict part failure.
- But features such as time since the last machine tune-up and volume of hardware produced can provide insight into the changeable aspects of the manufacturing process.

ENHANCE CREATIVITY BY USING LARGE SETS OF FEATURES

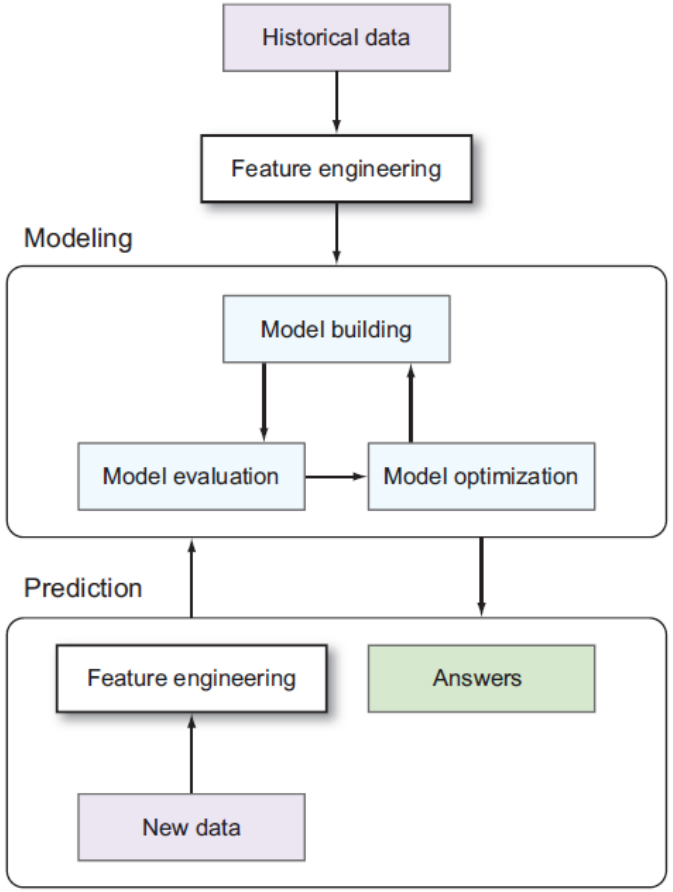
- Feature engineering empowers you to throw in large sets of features to see what sticks.
- You can create as many features as you can dream up and see which of them carries predictive power when thrown in to train a model.
- This allows ML practitioners to escape from a rigid mindset when creating and testing features and could result in newly discovered trends and patterns.

HOW TO APPLY FEATURE ENGINEERING?

FEATURE ENGINEERING AND DOMAIN EXPERTISE

- Another way to conceptualize feature engineering is as a mechanism that imbues domain expertise into a machine-learning model.
- For each problem at hand, knowledge about the data and systems under study is accumulated over time.
- The standard operating procedure for many companies is to use long lists of these ad hoc rules to make decisions and predictions.
- These business rules are a perfect set of engineered features on which to start building ML models!

BASIC FEATURE-ENGINEERING PROCESSES



EXAMPLES OF STATEMENTS OF DOMAIN EXPERTISE

- Web conversions are always higher on Tuesday (include the Boolean feature “Is it Tuesday?”).
- Household power consumption increases with higher temperature (include temperature as a feature).
- Spam emails typically come from free email accounts (engineer the Boolean feature “Is from free email account?” or email domain).
- Loan applicants with recently opened credit cards default more often (use the feature “Days since last credit card opened”).
- Customers often switch their cell-phone provider after others in their network also switch providers (engineer a feature that counts the number of people in a subscriber’s network who recently switched).

ACTIVITY

THE END