Data Science

Week 6

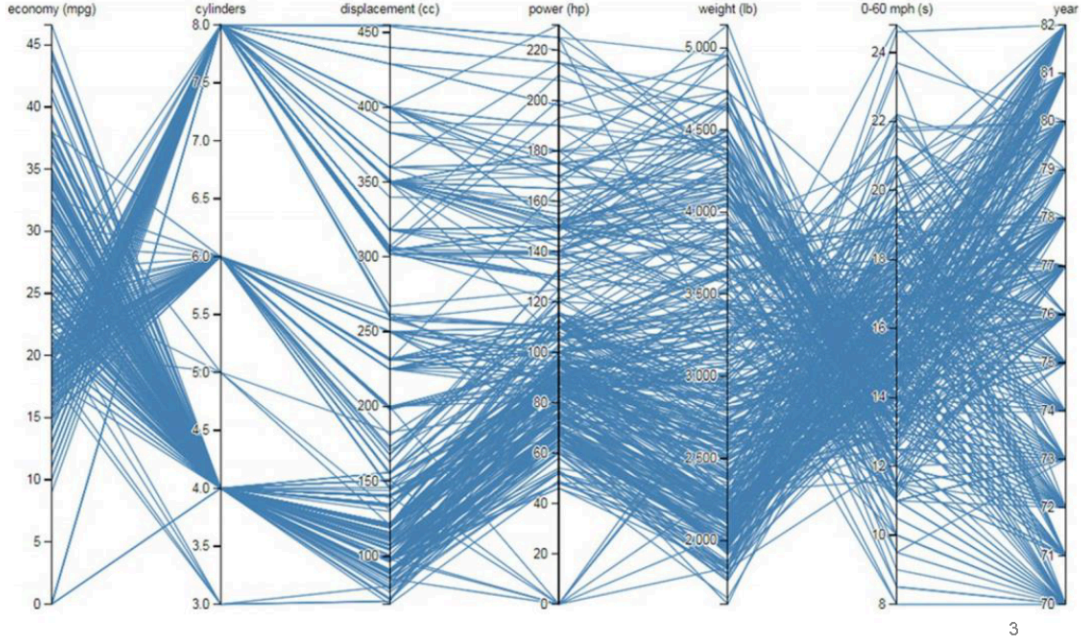
Introduction to Predictive Modeling

Review for Week 5

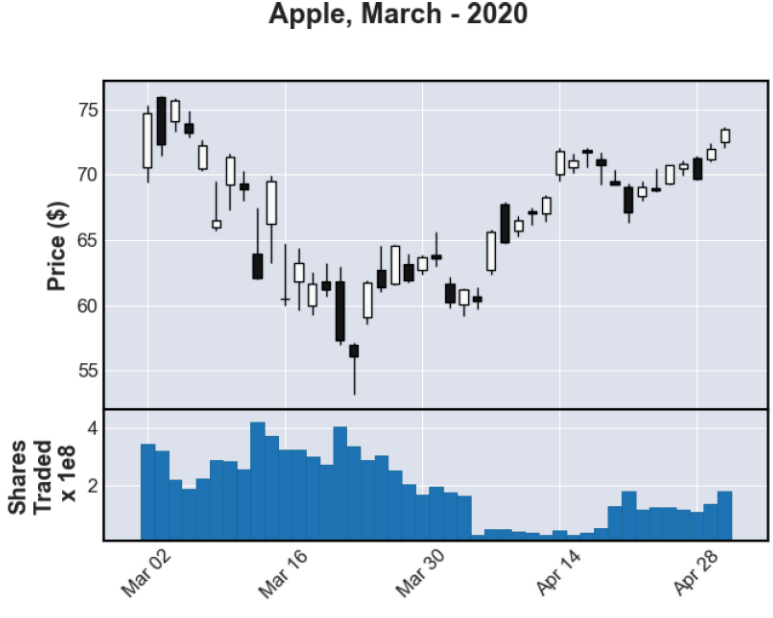
1. Basics of Data Visualization

2. One more step to application for data science
   - Visualization with parallel coordinates plot
   - Visualization for time series data (stock data)

Parallel Coordinates Plot



Stock Data Visualization



Example in Week 2: Predicting Customer Churn

- MegaTelCo, a largest telecommunication firm, has a major problem with customer **retention** in the wireless business.
- 20% of cell phone customers leave when their contracts expire, and it is getting increasingly difficult to acquire new customers. Customers switching from one company to another is called *churn.*
- You have been called in to help understand the problem and to devise a solution. Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to prevent churn. Marketing has already designed a special retention offer.

Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

Converting "Predicting Customer Churn" to *Supervised* Segmentation

- Following our example of data mining for churn, we will begin by thinking of predictive modeling as *supervised* segmentation—how can we segment the population into groups that differ from each other with respect to some quantity of interest.
- In particular, how can we segment the population with respect to something that we would like to predict or estimate.
- The target of this prediction can be something we would like to avoid (negative light), such as ***which customers are likely to leave the company when their contracts expire***, which accounts have been defrauded, or which web pages contain objectionable content.

- The target might instead be cast in a positive light, such as which consumers are most likely to respond to an advertisement or special offer.

Attributes and Target

- One of the fundamental ideas of data mining: finding or selecting important, informative variables or "attributes" of the entities described by the data. For example, the age of customer maybe an attribute to predict customer churn.
- A key to supervised data mining is that we have some target quantity we would like to predict or to otherwise understand better. Often this quantity is unknown or unknowable at the time we would like to make a business decision, such as whether a customer will churn soon after her contract expires.
- Having a target variable crystalizes our notion of finding informative attributes: is there one or more other variables that reduces our uncertainty about the value of the target?
- This also gives a common analytics application of the general notion of correlation discussed above: we would like to find knowable attributes that correlate with the target of interest -that reduce our uncertainty in it.

Model

- A model is a simplified representation of reality created to serve a purpose (e.g. Predicting Customer Churn).
- For example, a map is a model of the physical world. It abstracts away a tremendous amount of information that the mapmaker deemed irrelevant for its purpose. It preserves, and sometimes further simplifies, the relevant information.
- For example, a road map keeps and highlights the roads, their basic topology, their relationships to places one would want to travel, and other relevant information.

Satellite image

Road map

Predictive Model

- In data science, a predictive model is a formula for estimating the unknown value of interest: the target. The formula could be mathematical, or it could be a logical statement such as a rule.

- Given our division of supervised data mining into classification and regression, we will consider classification models (and class-probability estimation models) and regression models.

Predictive Model - Prediction

- In common usage, prediction means to forecast a future event. In data science, prediction more generally means *to estimate an unknown value*. This value could be something in the future (in common usage, true prediction), but it could also be something in the present or in the past. Indeed, since data mining usually deals with historical data, models very often are built and tested using events from the past.
- Predictive models for credit scoring estimate the likelihood that a potential customer will default (become a write-off, future unknown value). Predictive models for spam filtering estimate whether a given piece of email is spam (*present unknown value*). Predictive models for fraud detection judge whether an account has been defrauded (*past unknown value*). The key is that the model is intended to be used to estimate an unknown value.

Predictive Model vs. Descriptive Modeling

- The primary purpose of the Descriptive Modeling is not to estimate a value but instead to gain insight into the underlying phenomenon or process.
- A descriptive model of churn behavior would tell us what customers who churn typically look like. Descriptive modeling often is used to work toward a causal understanding of the data generating process.
- The difference between these model types is not as strict as this may imply; some of the same techniques can be used for both, and usually one model can serve both purposes.

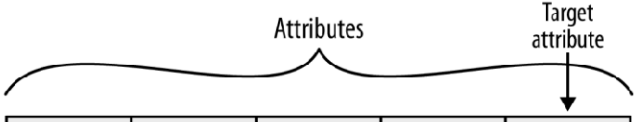
Terminology of Predictive Modeling

- Supervised learning is model creation where the model describes a relationship between a set of selected variables (attributes or features) and a predefined variable called the target variable.

- The model estimates the value of the target variable as a function (possibly a probabilistic function) of the features.

- For our churn-prediction problem we would like to build a model of the propensity to churn as a function of customer account attributes, such as age, income, length with the company, number of calls to customer service, overage charges, customer demographics, data usage, and others.

Example Problem of Credit Write-off Prediction



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

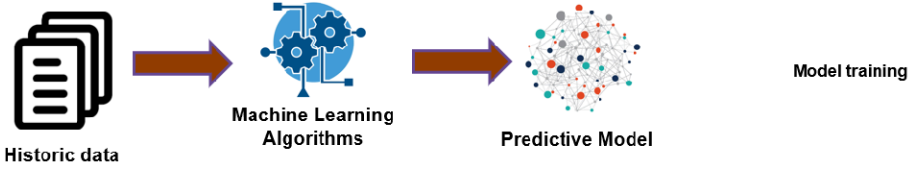
- An instance or example represents a fact or a data point—in this case a historical customer who had been given credit.
- An instance is described by a set of attributes (fields, columns, variables, or features).
- An instance is also sometimes called a feature vector, because it can be represented as a fixed-length ordered collection (vector) of feature values.

The problem is supervised because it has a target attribute and some "training" data where we know the value for the target attribute. It is a classification (rather than regression) problem because the target is a category (yes or no) rather than a number.

Training and Test in Predictive Modeling

Training: build a model

Target and attributes are known in training phase



Test: use the built model on the new data for prediction.

Target is unknown, and attributes are known in test (scoring) phase.

Reference

- Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., 2013.

Data Science

Week 7

Predictive Modeling and
Decision Tree