Part 1: Understanding Business Problems

By Josh Houlding

The most important part of creating a machine learning model is having sound business knowledge of the problem you're trying to solve. In 500-750 words, complete the following:

1. Formulate a prediction and an inferential question that you want to answer by applying predictive modeling. Ex: Prediction Question: How accurately and how far into the future can I predict the price of a house given the values of all the variables. Inferential Question: How accurately can I estimate the effect of each variable on the house price?

After brainstorming some ideas for questions that could be answered with predictive modeling, I settled on prediction of stock prices for a particular company given their historical stock prices and other relevant financial metrics.

Prediction question: How accurately can the future stock prices of a particular company be predicted based on historical data, market trends, and relevant financial metrics?
Inferential Question: How much of an impact does each financial metric (eg. earnings per share, price-to-earnings ratio, dividend yield, etc.) have on the stock price?

2. Search and locate a dataset that is relevant to the question(s) you created in the previous step. You may search repositories such as Data.gov, UCI Machine Learning, Kaggle, or Scikit-Learn. Find a dataset with no less than 10 variables and 10,000 observations, mostly quantitative.

Dataset chosen: Apple Stock Prices (2015-2020) | Kaggle

This dataset contains daily Apple stock prices and other relevant financial metrics for the company from May 2015 to May 2020, and is thus useful for time-series forecasting.

In [21]:

```
import pandas as pd

# Load data
df = pd.read_csv("AAPL.csv")

# Drop redundant columns
df = df.drop(columns={"Unnamed: 0"})
```

```python
# View data
df.sample(5, random_state=42)
```

Out[21]:

	symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCas
561	AAPL	2017-08-16 00:00:00+00:00	160.95	162.51	160.150	161.94	27321761	154.872237	156.373329	154.102447	155.824853	27321761	0.
101	AAPL	2015-10-19 00:00:00+00:00	111.73	111.75	110.110	110.80	29759153	103.426850	103.445364	101.927240	102.565963	29759153	0.
51	AAPL	2015-08-07 00:00:00+00:00	115.52	116.25	114.500	114.58	38670405	106.935199	107.610949	105.990999	106.065054	38670405	0.
63	AAPL	2015-08-25 00:00:00+00:00	103.74	111.11	103.500	111.11	103601599	96.030623	102.852925	95.808458	102.852925	103601599	0.
1073	AAPL	2019-08-29 00:00:00+00:00	209.01	209.32	206.655	208.50	21007652	207.343067	207.650595	205.006849	206.837135	21007652	0.

Dataset info (from Kaggle page):

symbol - Apple Stock

close - Closing price

high - Highest price of the day

low - Lowest Price of the day

open - Opening price of the day

volume - Volume of stock traded

adjClose - Closing stock price in relation to other stock attributes/actions

adjHigh - Highest stock price in relation to other stock attributes/actions

adjOpen - Opening Stock price in relation to other stock attributes/actions

adjVolume - Trading volume in relation to other stock attributes/actions

divCash - Cash dividend

splitFactor - Stock split

3. Discuss the origin of the data and assess whether it was obtained in an ethical manner.

The Kaggle page for this dataset is lacking in information about how the data was obtained, but considering that Apple is one of the most valuable companies currently, stock prices for the company are publicly available and easily obtained. Potential sources include stock market databases, financial data providers, or financial news websites. As a publicly-traded company, Apple is required to disclose its financial information, including stock prices, to the Securities and Exchange Commission (SEC), which then trickles down to public data sources like Kaggle. Because this data was publicly available, the person who posted it on Kaggle was simply redistributing it, and thus their acquisition was well within ethical bounds.

4. Explain your dataset's variables. List your dependent and independent variables, and identify which scale is used to measure each variable (interval, ordinal, or nominal). Hint: interval is the most appropriate scale for regression analysis.

In [22]:
```
df.head(1)
```

Out[22]:

	symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCash	spli
0	AAPL	2015-05-27 00:00:00+00:00	132.045	132.26	130.05	130.34	45833246	121.682558	121.880685	119.844118	120.11136	45833246	0.0	

In [23]:
```
# Check values of "divCash" and "splitFactor" to determine which kind of scale they use
print("Values for \"divCash\"")
print(df["divCash"].value_counts().to_frame().transpose())
print()
print("Values for \"splitFactor\"")
print(df["splitFactor"].value_counts().to_frame().transpose())
```

```
Values for "divCash"
divCash  0.00  0.57  0.63  0.73  0.77  0.52  0.82
count    1238     4     4     4     4     3     1

Values for "splitFactor"
splitFactor   1.0
count        1258
```

Interval: numeric variables, ordinal: categorical variables with a clear order/ranking, nominal: categorical variables without a clear order/ranking.

- **Dependent Variable:** `adjClose` (scale used: interval)
- **Independent Variable:** `symbol` (scale used: nominal)
- **Independent Variable:** `date` (scale used: interval)
- **Independent Variable:** `close` (scale used: interval)
- **Independent Variable:** `high` (scale used: interval)

- **Independent Variable:** `low` (scale used: interval)
- **Independent Variable:** `open` (scale used: interval)
- **Independent Variable:** `volume` (scale used: interval)
- **Independent Variable:** `adjHigh` (scale used: interval)
- **Independent Variable:** `adjLow` (scale used: interval)
- **Independent Variable:** `adjOpen` (scale used: interval)
- **Independent Variable:** `adjVolume` (scale used: interval)
- **Independent Variable:** `divCash` (scale used: interval)
- **Independent Variable:** `splitFactor` (scale used: nominal)