# Chapter 28

# Data Science and Credit Analysis

What's in this chapter:

- queries and databases
- regression
- probability and decision trees
- issues for data science and credit markets

DATA SCIENCE COMBINES statistical and scientific analysis with computer science. The use of statistical data and other quantitative tools for investing dates back at least to the 1950s. At that time, scholarly work in the area, often referred to as modern portfolio theory, was developed. The revolution in computing power has increased and expanded the use of these techniques in the investment world. Quantitative and statistical techniques in investment analysis encompass a vast area, and only a few brief concepts are discussed in this chapter. Data science can be descriptive or predictive. Descriptive techniques tend to analyze what happened in the past, such as portfolio performance attribution. Descriptive types of data can often be used to develop probabilities that can be applied to scenario analysis. Predictive techniques are used to try to make a statement about what might happen in the future. Regression analysis is a very common technique in predictive analytics.

## Queries and Databases

Computers can help to rapidly populate cells in spreadsheets by transmitting data straight from financial documents into an analyst's models and then highlighting where there are significant changes. Data science techniques can also use algorithms to do advanced word searches in company documents and highlight key words or changes in language from one quarter to the next. Data science tools can also be incredibly valuable in doing searches and rankings for relative value analysis. Many of these tasks involve queries and databases.

Credit analysis builds a plethora of data. The value of all this data can be enhanced if it is prepared in a usable format and stored in the right systems in such a way that it can be easily accessed. Not only should the analytical algorithms be easy to use, but it is important that the investment teams that are reacting to rapidly moving markets have easy access to them too.

A good database and query technology can allow an analyst, portfolio manager, or investment banker to input criteria for a data search and generate a list of options. For example, an investment banking analyst may want to generate a list of companies in an industry. The analyst may want them ranked by revenue growth and then by weakest EBITDA margins. The goal might be to find companies that would be attractive acquisitions for a more efficient operator. An analyst working for a portfolio manager may want to search for companies that have seen the biggest improvement in leverage ratios over the last three years. Good query systems should be able to generate lists that meet a number of prioritized selection criteria and simultaneously supply rankings for certain criteria as well. They can be valuable tools for relative value screens and analysis. Analysts and portfolio managers may also want to set certain automated queries. For example, these may be regularly generated reports that highlight changes in equity market values or if a key credit ratio has changed by more than a certain amount.

Corporate credit markets have numerous characteristics for both the companies that issue the debt and the actual debt instruments. All of these items need to be captured to make a database and query system valuable. Analysts need to be aware of how critical it is to design data fields correctly and enter the right data when building data science systems. Analysts have to try to be forward-thinking about what types of field might be of interest now and in the future when they collaborate with the data science team.

The quality of the data and the design of databases is important for all aspects of data science. If designed correctly, they can allow query systems to be linked with performance analytics and scenario analysis.

## Regression

Regression is a basic statistics technique that can be used to examine the strength of the relationship between two variables (e.g., leverage and YTW). Regression can also be used as a predictive tool. Simple regression shows how one data point (the dependent variable) will likely react in relation to change in another data point (the independent variable). This technique is based on how the two datasets have acted in the past. Ordinary least squares is the technique used to run regression analysis.

The simple form of linear regression produces a linear equation that can be used to create a line on a graph, sometimes called a fitted line. The equation and the line can, theoretically, be used to predict the dependent variable that would occur for each value of the independent variable. As an example, the line could produce a theoretical prediction of how much the yield on a bond would move if its leverage ratio moved from 3.5× to 2.5×. More advanced regression models can use multiple variables.

Combining regression techniques with computing power can be a very valuable tool to analyze data and generate some predictive models. Systems can run regressions on any number of combinations of variables to look for the most meaningful relationships and even which relationships are the strongest in different market and economic conditions.

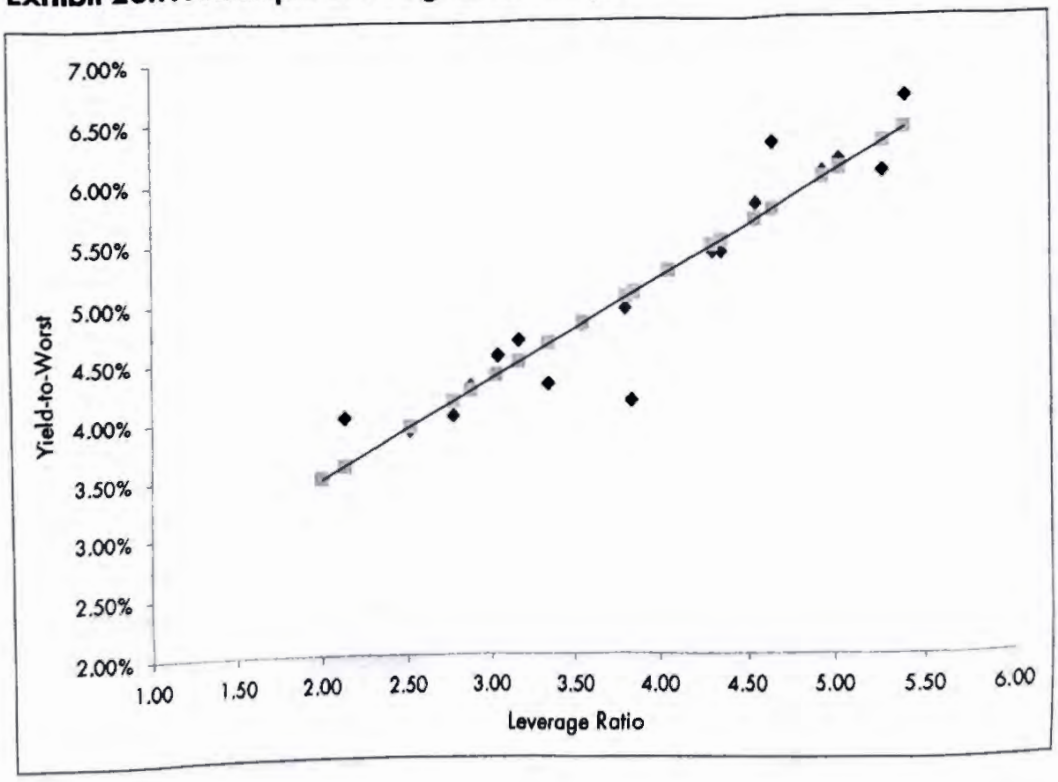
Correlation is an output generated from regression. It is a measure of the strength of the relationship between the variables in the regression. If the relationship is particularly weak, there are techniques that can be used to improve the relationship between the variables, such as using exponents or logarithms to adjust the datasets into a stronger prediction.

Exhibit 28.1 shows a graph for a hypothetical dataset of bonds where the leverage ratio and the YTW were regressed against each other. In this example, the YTW is the dependent variable and the leverage ratio is the independent variable. The outcome that the regression predicts is represented by the squares

along the fitted line. The diamonds show the actual results from the data, so the diamond on the upper far right represents a bond with approximately 5.5× leverage and a yield of 6.75%. Theoretically, a leverage ratio could be picked, and a line drawn until it hits the predicted line and then a perpendicular line to the Y axis, and it would predict where a bond should trade from the Y axis. This technique can be used as relative value tool as well. If the diamond is below the predictive line, it implies that the yield on the bond is too low relative to its leverage and the price of the bonds should move down so that the yield is more in line with the predicted yield. The difference between the actual datapoint and the predicted data point is called a residual. Most regression programs will print out the actual residual values. In this case, the residuals could be used to predict how rich or cheap a bond was based on its leverage.

The closer the dots cluster around the predicted line, the stronger the quality of the relationship between the two variables and the greater the implication that the independent variable is a strong predictor of the dependent variable. Exhibit 28.1 shows strong evidence of dependence on the independent variable.

**Exhibit 28.1: A Sample of a Regression Output**



## Probability and Decision Trees

Scenario analysis can be a large part of decision making, especially when analyzing the potential outcomes of an event, such as merger and acquisition activity, or when doing total return analysis. Scenario analysis is used by investors, investment bankers, and corporate financial officers. Once an analyst starts to lay out various scenarios and outcomes, the next logical step is to apply probabilities to each case, and perhaps, probability-weight the outcome. Probability is used all the time in decision making, though often it is done informally and intuitively. Probability is also the basis of many data science techniques.

One easy way to explore probabilities and scenarios is to lay out the possible outcomes in a decision tree. Decision trees are a form of flow chart, and each step in the decision process is a node. There are different types of nodes, including those that involve making a decision—these are often the root node or starting point from which branches extend to other nodes. Many of the other nodes will have uncertain outcomes and are often called chance nodes. When there are no more options and a conclusion for one decision branch, or path, is reached, there is an end node.

If the situation being examined is similar to past situations, data and outcomes from these other case studies could be downloaded, and an analyst could develop some probability guidance based on past experiences. For example, if two PE firms are making competing bids to buy a company with publicly traded stock and bonds outstanding, an analyst may want to analyze the probability that the company has been bought, how the PE firm will likely finance the company (e.g., 20% equity, 60% subordinated debt, 20% secured debt), and what will happen to the existing bonds. A decision tree with probabilities for each outcome could be helpful in analysis. It could get fairly complex rather quickly, with numerous outcomes, including third-party bidders and all the varied financing options. If a database had details of the PE bids that had been made in the past, including the success of the bids, how they were financed, and how the previously existing bonds traded, an analyst could start to build probabilities for each outcome. Artificial intelligence algorithms have been designed to process probabilities for these types of problem, using decision trees. These algorithms can calculate the probabilities for these outcomes based on past data

Conditional probability, and specifically Bayesian probability, is the basis behind much of this logic used to develop predictive models that are based on multiple decision levels. Even in its simplest form, understanding conditional probability can help with decision making. When designing decision trees or simpler scenarios, conditional probability helps adjust probabilities as new information becomes available. Using the acquisition example above to outline conditional probability, assume a decision tree with probabilities for each outcome has been built with the goal of determining what the likely pro forma capitalization will be. When there are multiple bidders for the firm, each decision branch will end with a different buyer and the various likely capitalizations they will use for the acquisition, with a probability assigned to each outcome. Once it is clear that one of the PE firms will be buying the company, the probabilities for what the final capitalization will be all change because of this new condition.

Bayes's theorem helps show how to adjust the probability of an event as new data is received.

In this case, the goal is solving for the probability of event A occurring if event B has occurred.

The theorem needs three key pieces of data:

1. the probability of an event A occurring, designated by P(A)
2. the probability of a second event B occurring, designated by P(B)
3. the probability of event B occurring given that A has occurred. This is designated by P(B|A). The vertical line stands for the word *given*, so in the formula below, P stands for *probability*, and P(A|B) reads: *the probability of A given that B has occurred.*

$$P(A|B) = \{P(B|A) \times P(A)\}/P(B)$$

## Issues for Data Science and Credit Markets

There are many aspects of credit markets that can add to the complexity of using data science. Each constituent in the market requires a significant amount of descriptive data, such as issuer entity, industry, currency, country of risk, coupon, ranking, maturity, credit rating, call prices, and any special call features. There is also significant market-related data that has to be fed into the database and calculated, including price, spread, yield, and duration, all of which need to include various scenarios. As an example, for yield, there needs to be data for current yield, YTW, YTM, and YTC. The list above is not complete and does not include covenant differences or public stock, or ownership data.

Having good data can also be vital to a database. Unfortunately, pricing data in the credit market adds a level of complexity to creating a good data series. Some debt instruments in the market trade regularly, but others may only trade once a week or once a month, which makes some pricing data more readily available than others. The lack of consistent pricing data can make statistical analysis more difficult and less accurate.

Another factor that adds to the complexity of using data science techniques is the transitional nature of many constituents in these markets, and that requires constant updates. The constituents in the market are constantly changing through new entrants, maturities, calls, upgrades, and downgrades from investment grade and defaults. Most statistical techniques are based on historical data. Therefore any data scientist has to be very cautious—when examining data on the fixed-income market—that the historical data is still relevant to the current market in which investment decisions are being made.

There are also numerous problems that develop in statistical analysis, regardless of which market is being analyzed. Any number of biases can creep into analysis either intentionally or unintentionally. Selection bias and the base rate fallacy are common examples. Other common problems can include designing the analysis poorly, bad data, or gaps in data sets, all of which can distort probabilities. These risks are magnified if the market being analyzed goes through rapid changes. It can be dangerous to blindly follow the results of data output without applying logic and thought. Improper use of even the simplest probability model can result in poor decision making.

## Closing Comment

Regression and probability analysis can be incredibly valuable tools to enhance analysis and decision making. They can also improve modeling and forecasting techniques. Data science can enhance them and improve the ability to quickly analyze changes in data. Analysts should embrace these tools and not be worried about them. Credit analysis is more than crunching numbers; it is also about designing how those numbers should be crunched, how they should be analyzed, and how people will react to them. Data analysis is only as good as the data that is being used. In a market that changes so rapidly, always be wary of inputs that might be outdated.

The introduction of more quantitative tools increases the value of interaction with corporate management teams. This is a skill that algorithms still have not yet developed. An understanding of motivations and strategies at the companies that are being analyzed can be a material differentiator in analytical work, and the best insights come from interaction with management.