

Unit 1: Lecture



What is data mining and how is it different from predictive modeling?

Data mining is the exploration and analysis of data to find patterns that enable us to make better decisions. It is the flip side of the coin to predictive modeling. In the latter, the goal is predicting the future above all else. In the former, we are primarily concerned with explaining phenomena in our data.

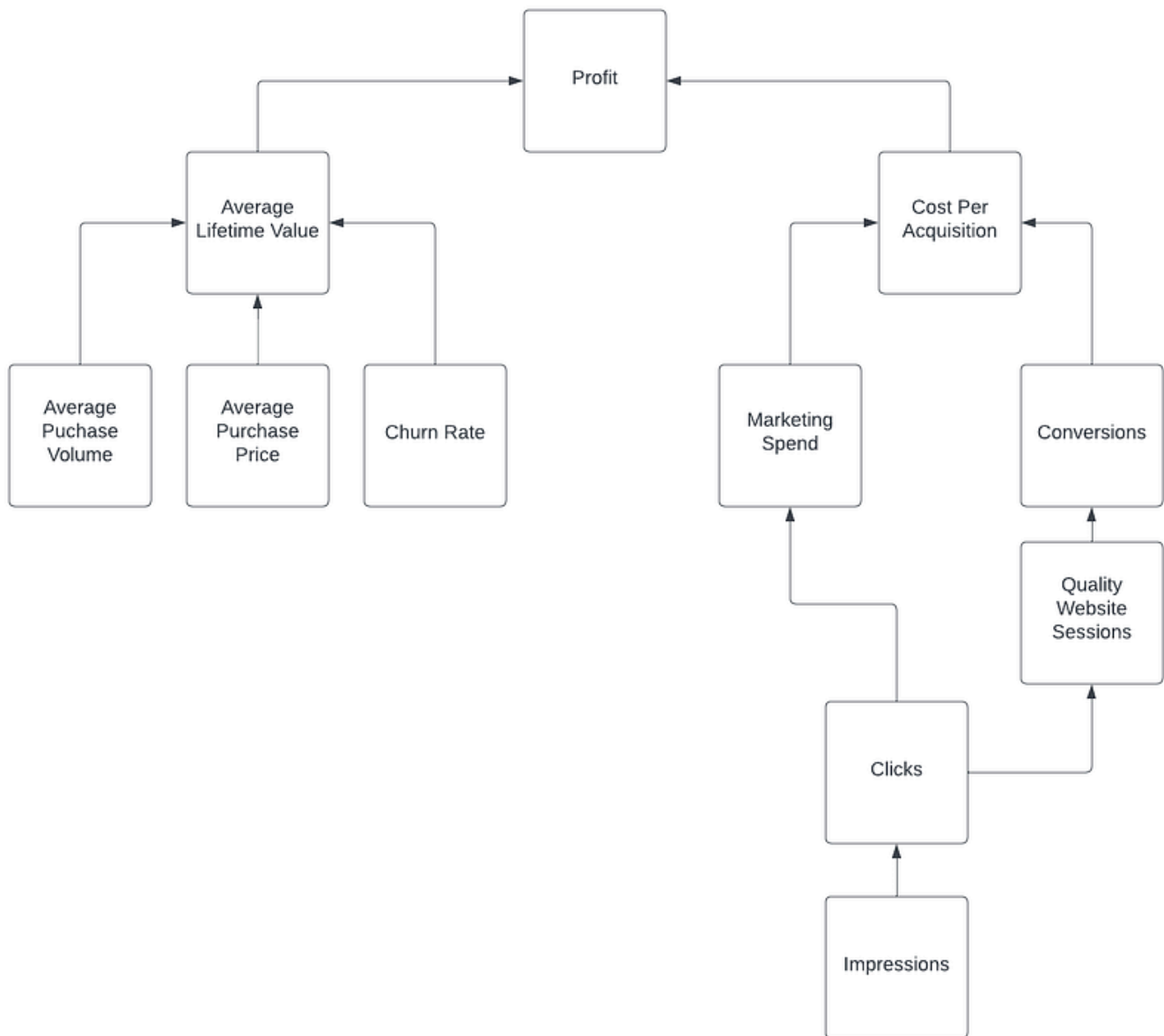
This distinction may appear small, but it is crucial. When we are building a model with solely prediction in mind, we might sacrifice some level of interpretability to squeeze out more predictive power. This is likely not a tradeoff we want to take in the realm of data mining, where explanation is king. Likewise, there are statistical techniques we might employ in data mining that we would not use when building a model purely for prediction. Certain techniques lend themselves to explainability, though they are inferior for strictly prediction tasks.

Data mining is a part of “decision science”. Decision science is the practical application of data, math, and psychology to aid in structured, informed decisions. “Data mining” represents the quantitative part of decision science. In some cases, we might be fine with a purely data-driven approach to decision-making. In other cases, we want to blend data with expert opinion. In yet others, especially when we have data quantity or quality issues, we would like to only rely on experts.

What are the goals of data mining?

Fundamentally, the goal of data mining is to positively impact business metrics. In the world of business, the main driving metric is profit.

To understand how to pull profit levers, the data science team must grasp the company’s hierarchy of metrics. The overarching goal is to increase profitability, but this aim cannot be performed directly. It can only be accomplished by driving metrics that ladder up to profitability. Below is a classic example of a simple metric tree for many digital sales and marketing organizations. (To note, an actual metric tree could be many layers deep).



All data mining projects should target one part of the metric tree. Understanding the business's mechanics will help to mitigate the “we solved the wrong problem” issue.

In a business with the foregoing metric tree, all data science endeavors should focus on improving either average lifetime value or cost per acquisition. On the lifetime value side, more specifically, any effort should center on bettering one of the three driving metrics: average purchase volume, average purchase price, and churn rate. Similarly, we can isolate the underlying, driving metrics to influence cost per acquisition. As one can see, even a simple metric tree can aid in focusing a data science group on the right tasks.

That said, even when we have a metric tree, the impact of some projects won't be straightforward to measure. For example, we might develop a tool that automatically suggests and generates search engine keywords on which to bid. This tool would have the function of making the marketing team more efficient. However, it might be difficult to get to a hard dollar figure on the effect as efficiency can be tricky to measure. That doesn't mean we should not pursue such projects. Rather, we should make the case that this “soft impact” can still positively influence parts of our metric tree by enabling others to more directly impact it.

Where is data mining used?

Data mining is ubiquitous. It's leveraged in business, non-profits, and government. In this course, we will primarily focus on applications in business. Over the past several years, organizations have focused on collecting increasing volumes of data. This reality has been made possible by the storage costs of data plummeting. Likewise, organizations now understand that data "exhaust" can drive business value. For example, sensor data from airplane parts can be used for predictive maintenance.


A survey of data mining methodologies

Methods for data mining can range from simple to advanced. Below is a non-exhaustive list of common methodologies you might encounter in data mining.

- Data visualization
- Significance testing
- Linear regression
- Tree-based models
- Tree ensemble models
- Unsupervised machine learning, such as clustering
- Association rules mining

One of the key traits of a good "data miner" is knowing which techniques to use for a given problem. The key is to know the tradeoffs of certain techniques and when each is appropriate.

Risks involved in data mining

Data mining is not all fun and games. Several risks exist. First, we might be pressured to present findings when none exist. We must remember that "no finding" is perfectly valid. Second, sooner or later, we will deal with spurious correlations - variables that are only related by chance (see some examples: [Spurious Correlations: Correlation is not Causation](https://www.tylervigen.com/spurious-correlations)  <https://www.tylervigen.com/spurious-correlations>). Third, data in the real world is messy. It can be incomplete, missing documentation, or plain incorrect. If we don't catch and correct these items, we will doom our data mining project. This list is certainly not exhaustive - many more risks can be at play.




Importance of clearly stating assumptions.

Related, in data mining, we will often make assumptions. For example, "we equate user happiness with how long they spend on the website." Is this the best definition? Maybe. Maybe not. Either way, we must state this assumption so that others are aware. Likewise, we might have some missing data and employ an imputation strategy to handle it. Again, documenting and clearly stating this assumption is important.

How to recognize and combat bias.

Data can be biased. Therefore, if we are not careful, our data mining exercises could be biased against certain groups of people. As analysts, we must always be aware of this risk. First, consider if the nature of the data might be unfair in any way. Second, question if the outcome of your analysis might disproportionately impact a certain group of people. Third, go back to steps one and two and repeat. We must be thorough and vigilant in this process. If we see bias in data, however, we must not ignore it. Ignoring it means that someone else might use it, even innocently. We must report it and work to correct the underlying cause.

Readings

- **Chapter Introduction: Data Science Definition and Ethics**  (<https://endtoenddatascience.com/chapter2-defining-data-science>)
- **Chapter Introduction: Data Science Theories**  (<https://endtoenddatascience.com/chapter3-applied-data-science-theory>)
- **Operationalization: the art and science of making metrics**  (<https://towardsdatascience.com/operationalization-the-art-and-science-of-making-metrics-31770d94998f>)