Lesson 2 – CRISP-DM

Slide 1

Hi everyone. This week we will discuss the cross-industry standard process for data mining, or CRISP-DM, for short.

CRISP-DM was designed in the 1980s and is used by companies today to structure data science or analytics projects.

CRISP-DM unifies the common steps involved in all analytics projects and helps data science teams develop successful data products.

It consists of 6 core activities, starting with business understanding and ending with the deployment of a data solution, which could be anything ranging from a report or dashboard, to a machine learning system.

Slide 2

The first step in CRISP-DM is Business Understanding. This stage is focused on understanding project requirements from a business standpoint. It includes three important steps.

The first is identifying the business objective. Before beginning any analytics project, it is important to understand exactly what problem a business is trying to solve.

Imagine you are a data scientist at a company where customers are spending too much time on your website searching for products.

This is leading to customers leaving the site without making a purchase.

In this situation, a business objective might be to find a way to display relevant products to customers based on their website activity. The goal would be to increase sales by helping customers find products faster.

Slide 3

Once a business objective has been identified, we must determine the analytical goals of the project. In this step, we clearly define what success looks like.

From our example, one metric might be tracking the average time for a product purchase for each customer. Minimizing this and increasing revenue is the ultimate goal. Therefore, any data solution to the business objective must focus on obtaining these results. Without clearly defined analytical goals, it is impossible to determine if a project was successful.

We also need to determine which machine learning algorithms are appropriate for solving the business objective. This depends on whether our goal is inference, where we are interested in uncovering factors that drive our customers to purchase products, or if we are simply interested in prediction accuracy.

In many cases, it's a combination of both and we will discuss how to choose machine learning algorithms as we progress through the course.

The last step is producing a project plan that incorporates all business objectives and success metrics and communicating this throughout the business.

Slide 4

The second step of CRISP-DM is Data Understanding where we are focused on collecting, describing, and exploring data that may be useful in solving our business problem.

In this stage, we are getting familiar with the data through exploratory data analysis, or EDA.

This involves calculating summary statistics, such as the mean and standard deviation of numeric data. Making histograms or boxplots to study the distribution of data values.

Looking for associations among variables.

Creating data visualizations to study complex relationships between variables in our data.

And transforming variables to create new features that might improve the performance of machine learning algorithms.

The Business Understanding and Data Understanding steps are iterative. Usually, when we learn something new about our data, we go back to the business understanding step and adjust our objectives and metrics.

Slide 5

Next is Data Preparation. This stage includes all steps in the process of converting raw data into a numeric data matrix for machine learning applications. Many refer to this step as ETL, which stands for extract-transform-load.

This stage starts with cleaning the raw data. It can involve processing unstructured text data, formatting numeric variables, imputing missing values, and many more tasks. An example of what this process looks like is displayed on the righthand side of the slide. Here we start with our raw customer data and use technologies such as SQL and R to convert it into the numeric data matrix at the bottom on the slide. Many of these preprocessing tasks are easy to perform but others can be extremely complex. One example would be determining whether a customer has a master's degree from their LinkedIn profile. This step alone would involve hundreds of lines of code.

Once the data is processed, we must maintain the code that takes new data and automatically converts it to the format we need. This step is critical for making sure we don't get unexpected results in the future.

When we are confident that our preprocessing code is working correctly, it must be integrated into a production environment where it automatically processes new data.

Slide 6

Once we have explored our data and prepared it for machine learning, we can begin the Modeling step. In this step, we train and assess the accuracy of different machine learning algorithms and determine which ones work best for solving the business objective.

All models are assessed based on their ability to optimize the business objective.

This is why it is critical to have clear, quantifiable success metrics before proceeding with the Modeling stage.

Data preparation and modeling are iterative steps. During modeling, it is common to discover new features that are important for prediction accuracy. When this happens the data preparation steps must be updated.

Slide 7

The next step is Evaluation. This step is concerned with determining whether our final machine learning algorithm from the modeling stage meets the business objective on new data.

Every machine learning algorithm should be tested on new data before it is deployed throughout a business.

Machine learning algorithms are developed with training data from the data preparation step. The prediction accuracy on training data is generally an overly optimistic estimate of how the algorithm will perform on new data sources.

The evaluation step is designed to find out if our trained algorithm will generalize well to new incoming data.

Many companies will conduct a series of A/B tests, where the machine learning algorithm is applied to a subset of new data and the rest is left as a control group. From our earlier example, we would split new customers into two groups, apply our algorithm to one group, and see if that group had larger amounts of product purchases. If so, we could proceed to the next step.

Slide 8

In the deployment phase, the final data solution is deployed throughout the business.

This may be as simple as producing monthly dashboards with model results to drive business decisions.

It may be the complex process of deploying a machine learning algorithm in a production environment, such as Microsoft Azure or AWS, where the results of the algorithm must be stored and passed between several business applications with customer interactions in real-time

Slide 9

Once a model is deployed into production, it needs to be monitored and periodically re-trained as new data is available. This ensures that new customer activity, for example, is captured by the machine learning algorithm.

Data governance is also important during model maintenance. Strict rules must be put in place to ensure the raw data is not unknowingly altered, resulting in a crash of the machine learning system

We also have to monitor new data regulations, such as GDPR from the European Union and customer privacy laws. We must ensure our algorithms are compliant and do not exhibit any bias based on protected consumer data such as gender, age, and ethnicity.

Now that we have gone through the steps of CRISP-DM, we see that data science projects are much more than just training machine learning algorithms. In fact, this typically only represents 20% or less of the total project activities. To ensure success of any analytics project, whether simple or complex, it’s important to follow these steps throughout the process.