SOLUTIONS FLYER

USING MACHINE LEARNING TO FILL GAPS IN LARGE DATASETS

One of the essential problems involved in managing large datasets is ensuring they're complete. Many use cases, including materials databases, can use machine learning (ML) and artificial intelligence (AI) algorithms to accurately identify and fill gaps with data extrapolated from other data in the set. The datasets might contain time series data which, for example, may track the movement of components through a supply chain and/or static data like a parts inventory or test results. Altair's data science tools are well suited to this task.

Computing missing values with the right ML algorithms improves the accuracy of predictive models, even when the source data contains a substantial number of missing values.

Substitute Missing Values

File corruption, failure to record data points, and an array of other causes can all lead to missing values. In fact, they're quite common and there are often patterns in missing values that help data scientists select the best method for filling them. Altair's Knowledge Studio® AI and ML software offers a node that addresses this requirement. This makes it easy to identify missing values and generate new substitute values based on a variety of substitution algorithms.

Approaches for handling missing data effectively include:

- Using a computed mean (average of all non-missing values in a set), median (middle number in a sorted set), or mode (most common value) value. This type of algorithm works well when the missing values are likely to be randomly distributed.
- Replacing missing values with a user-defined value. For example, if all known values are positive, replace missing values with zero or other specific value.
- Replacing missing values always introduces a bias in the data; however, that bias is often useful. For example, in credit risk models when the income is unknown, it is assumed to be the minimum value (or no income at all) to reduce risk.
- For time series data, interpolated values based on adjacent values in the series is often sufficient. Using a moving average to calculate missing values is an example of this approach.
- The approach considered to be the most accurate is to use a machine learning model to infer the missing values based on known values.

Knowledge Studio provides everything you need to replace missing values with suitable substitutes to create complete datasets.

> "Most datasets are incomplete in the real world, especially in cases where the data has been collected from multiple sources over long periods of time — of course, these types of data sources are quite often the most useful. Using AI and ML to make accurate predictions based on such sources requires that data science teams have ready methods for filling in the gaps. Knowledge Studio provides an excellent set of tools to do just that."
>
> Sam Mahalingam, CTO, Altair

Learn more about
Altair Data Analytics
altair.com/data-analytics

Example Use Cases

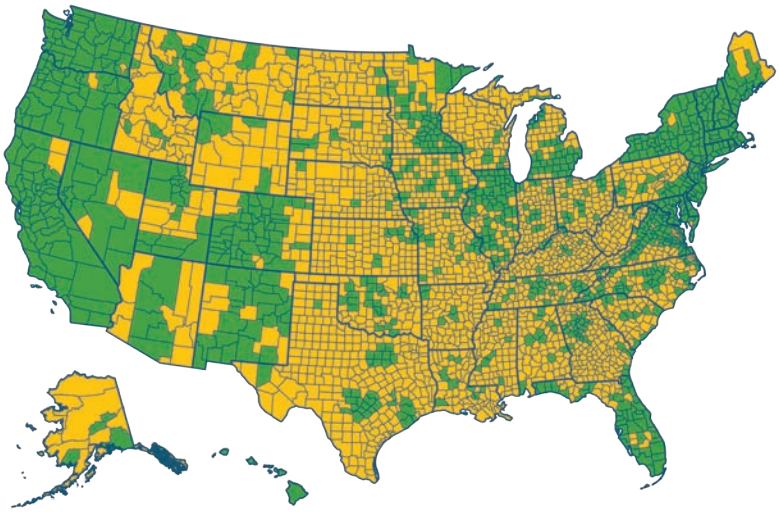
Predictive AI models are becoming ubiquitous in the business world – source data completeness is critical to model accuracy. Essentially, data scientists can use ML algorithms to "predict the past" and fill in missing values in source data as needed. Then, they can use other AI models to predict future performance or behavior in a variety of applications, including:

- Calculating consumer credit scores even if the dataset of previous payments and loans is incomplete;
- Adjusting retail prices even when the record of purchases for a particular product type is incomplete;
- Predicting when a component in factory machinery is likely to fail even if sensor data streaming in from the machine has been interrupted due to a technical fault;
- Building recommender systems for customers even in cases where the data on past purchases, customer demographics, and search histories are incomplete;
- Identifying consumer spending patterns even if complete data from retailers is unavailable; and
- Selecting short lists of materials that are most likely to meet specific technical requirements even when the materials database does not contain detailed information about the latest offerings.

Predicting EV Adoption Rates Accurately Using Incomplete Source Data

Let's look at a specific use case involving large, public data sources. Altair has worked closely with most major automotive manufacturers in the world for the past 35 years, and we develop new simulation and engineering software to help these manufacturers develop electric vehicles (EVs). Altair's data scientists recently sought to better understand EV adoption rates in the U.S., so they obtained publicly available databases related to EV adoption in 15 states and used Knowledge Studio to "fill in the blanks" and create complete historical data sets to predict the EV adoption level of counties in the remaining 35 states.

The team created two datasets in Knowledge Studio: the training set – which used 2019 data from 15 states for predictive modeling purposes (with 1141 records) – and the scoring dataset which predicted results for the remaining 35 states (with 2001 records). After some experimentation, the team discovered the variables with the most predictive power included the percentage of people working from home, median rent, and 2016 election results for a given geographic area. They built a variety of ML models in Knowledge Studio to make their predictions. Their approach concluded that there are three primary factors influencing EV adoption rates: availability of charging stations, housing affordability, and educational level. See the full story here.



Median EV Adoption = 0.61 Vehicles per 1,000 People. States with high EV = Counties in that state met the Median.

This chart displays EV adoption rates for all 50 states as predicted by the decision tree-based machine learning model the Altair team built using Knowledge Studio.