DATA FUNDAMENTALS: AI AND ML

Data Readiness: Lessons From The Field For Machine Learning Data Prep

Creating a data foundation that's AI and ML ready

The field of Machine Learning (ML) is not new, yet marketers are still discovering new ways to apply ML methods on their large, complex and expanding data sets. Demand for data science talent continues to grow, but the problems of collecting and normalizing clean, meaningful data for machine learning are snowballing faster than most firms can respond.

Tealium's unique positioning within our clients' organizations has provided us with insight into each step of the machine learning process. The Universal Data Hub was designed to produce ML-ready data, in real time. This guide is a result of strategic learnings across the lifecycles of ML projects based on successfully working with top brands across a wide variety of industries. It will help marketers, data scientists, engineers, and developers work together to take the steps needed to create a solid foundation that can support ML and AI initiatives.

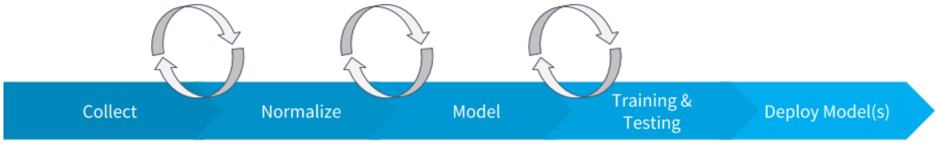
> **"In order for brands to take advantage of the avalanche of artificial intelligence functionality, it's critical that they first establish a data foundation that's ready for the task."**
> *- Brian Moore, Digital Strategist, Tealium*

Machine Learning (ML) and Artificial Intelligence (AI)

ML is the process of analyzing data in an automated way. It is a part of the AI process and provides machines and systems with the technical knowledge to be able to learn and make decisions off of data, without any human interaction.

AI is what makes it possible for machines to learn how to process and analyze data in an automated way. AI is what is behind the creation of the sophisticated machines and technology that are able to perform ML initiatives.

Understanding The 5 Steps Of A Machine Learning Project Lifecycle



1. Data Collection

Preparing customer data for meaningful ML projects can be a daunting task due to the sheer number of disparate data sources and data silos that exist in organizations. In order to build an accurate model, select data that is likely to be predictive of the target: the outcome which you hope the model will predict based on other input data.

For a typical consumer brand, desirable input data may include web activity, mobile app data, historical purchase data, and/or customer support interaction data. Traditionally, these data sources may be difficult to access or configure. **With Tealium's Universal Data Hub, deploying data collection on new platforms or devices is expedited to give brands a central point of collection across different data sources.**

For Tealium, data collection goes beyond reacting to abandoned carts or recommending a favorite category, but rather has a truly revolutionary capability (not just taking last actions and extrapolating or averaging) to actually predict the future.

2. Data Normalization

The next step in the ML process is where analysts and data scientists typically spend most of their time on analysis projects: cleaning and normalizing dirty data. This oftentimes requires data scientists to make decisions on data they may not understand, like what to do with missing data, incomplete data, and outliers.

This data may not be easily correlated to the proper unit of analysis: the customer. In order to predict if a single customer will churn (not a segment or whole audience), for example, siloed data from disparate sources can't be relied on. A data scientist will prepare and aggregate all of the data from those sources into a format that ML models can interpret. This can end up being a lengthy process and requires a lot of work before any ML can even occur.

Core features of Tealium's Universal Data Hub illustrate how these challenges can be solved: Data Sources and Event Specifications. Together, these features set the foundation for data collection and normalization. **In a matter of minutes, developers and analysts can test and verify that their data is clean and meets expectations prior to writing a single query.**

Tealium has a marketplace of common Event Specifications to configure with Data Sources, or clients can add their own. These capabilities offer analysts and data scientists a quick and easy way to inspect data and gain familiarity with the business challenge they wish to address with machine learning - while simultaneously saving time and automating the data collection and normalization process.

3. Data Modeling

The next phase of an ML project is to model the data that will be used for prediction. **Part of modeling data for a prediction about customers is to combine disparate data sets to paint a proper picture of a single customer.** This includes blending and aggregating silos of data like web, mobile app, and offline data.

For example, below are three examples of data from a single customer across 3 different channels; the desktop website, a mobile app, and in-store transactions:

```
Website engagement data:
visitor_id |   device   |   time   |   date   |  url_path
-----------+------------+----------+----------+-----------------------
 fD39xn3x  |   Chrome   | 07:18:39 | 09/14/18 | /homepage
 fD39xn3x  |   Chrome   | 07:18:51 | 09/14/18 | /shoes
 fD39xn3x  |   Chrome   | 07:22:19 | 09/14/18 | /hi-top-runners-white
 fD39xn3x  |   Chrome   | 11:05:03 | 09/16/18 | /login
 xYmz49gA  |  iOS 11.3  | 07:22:19 | 09/18/18 | /fall-shoe-sale
-----------------------------------------------------------------------

Mobile app data:
  app_uid  |   device   |   time   |   date   |  event_name
-----------+------------+----------+----------+---------------
 nv-7sn3n  |   iPhone   | 14:20:28 | 09/17/18 | install
 nv-7sn3n  |   iPhone   | 14:20:55 | 09/17/18 | launch
 nv-7sn3n  |   iPhone   | 14:21:19 | 09/17/18 | sign-in
 nv-7sn3n  |   iPhone   | 14:21:34 | 09/17/18 | my-account
 nv-7sn3n  |   iPhone   | 14:22:13 | 09/17/18 | checkout
 nv-7sn3n  |   iPhone   | 14:26:03 | 09/17/18 | confirmation
---------------------------------------------------------------

In-store transaction data:
  cust_id  |   date   |  trans_id  | trans_total
-----------+----------+------------+--------------
 nv-7sn3n  | 09/17/18 | 902-12322  | $247.68
 nv-7sn3n  | 09/17/18 | 902-29377  | ($75.44)
--------------------------------------------------
```

To summarize this data, the following may be derived data points that represent customer-level behavior:

- First Activity: 9/14/2016
- Last Activity: 9/27/2016
- Lifetime Web Visit Count: 2
- Lifetime Mobile Visit Count: 1
- Lifetime Transactions Count: 2
- Lifetime Value: $172.24
- Favorite Category: Shoes

These are just a handful of meaningful derived data points that a brand might define for a customer. But this is no easy feat. Consider when the brand sees this customer again in the future and wants to predict, for example, the customer's probability of converting at the start of their next web session - in real time.

Tealium's Customer Data Platform, AudienceStream, provides visit and visitor-level enrichment to allow marketers, in tandem with developers, data scientists, or analysts, to define the business rules for these aggregate data points. One or more of those data points may be leveraged for Visitor Stitching – Tealium's patented method for merging visitor profiles together across devices in real time. **The result of these capabilities is a clean and correlated single view of the customer, providing the robust data foundation required for machine learning.**

4. Model Training and Feature Engineering

After a brand has deployed collection and enrichment of meaningful input data, it's time to put the predictive power of that data to the test. To do so, data scientists take a representative sample of the population (i.e. all customers, anonymous visitors, or known prospects) and set aside a portion for training models. The remainder is used to validate the models after training is complete.

A key component of this phase is to iterate rapidly, continuously testing new data points that can be derived from the data source. This process is called feature engineering.

To continue with the earlier example, the following engineered features may be tested:

- Customer Age in Days (the difference between first and last activity date): 13
- Average Order Value (lifetime value divided by total transactions): $86.12

These attributes and others can be easily calculated from the aggregate visitor data, providing data scientists a quick way to iterate on training models to compare accuracy. Tealium's DataAccess products (for example, AudienceStore) provide a seamless way to export visit and visitor data, in real-time, for training ML models. If an engineered feature is thought to be beneficial, users can add new AudienceStream attributes in a matter of minutes with the flexible enrichment options and attribute data types.

This is how data scientists can be technologically enabled to produce better insights instead of requiring them to complete mundane data management tasks repeatedly and manually.

5. Deploying Models to Production

All work to this point culminates in the final step of deploying a model to production where the ability to predict outcomes in the real world is tested. By this point, models should meet some threshold of accuracy that warrants deploying them to production. For this reason, it's important to interpret model performance with stakeholders to agree on what level of risk is acceptable for inaccuracy. Some customer behaviors may not be sufficiently predictable, and thus a model may never achieve accuracy to justify deploying to production.

> **"With the right technology, teams can establish ML viability with their data, with complete transparency when interpreting their models (instead of black box prevalence in many AI solutions). This flexibility allows brands to apply their learnings right back into their business."**
>
> *- Brian Moore, Digital Strategist, Tealium*

Once models are live, marketers and stakeholders can finally capitalize on their predictions. This might include serving a promotion to a suspected high-value prospect or suppressing marketing for predicted low-value visitors. Work with stakeholders and martech owners to think through the marketing applications of your predictions.

In the end, Machine Learning isn't going to replace a digital marketing strategy, but instead, it will augment and enable it. Successful brands will put their customer at the center of what they do and Machine Learning is one tool (among many) to optimize decision making as part of that larger initiative.

What does deploying models to production look like with Tealium? We offer multiple methods for ingestion that import predictions back into the Universal Data Hub and can also upload predictions via offline import or in real time (i.e. in response to a live visitor session, using our inbound API).

Quick Tips for ML Success

- Valuable ML applications don’t require billions of data points. Even small data problems can be supported by ML.
- Not every business problem can or should be solved with ML. Work with stakeholders to educate internal teams on the costs and benefits of ML proactively.
- Set up workflows to fail fast: collect data thought to be predictive, perform exploratory data analysis and check feasibility for ML models.

Notes

Collect, Enrich, and Take Action On Customer Data

Tag Management | API Hub | Customer Data Platform | Data Management

Tealium revolutionizes today's digital businesses with a universal approach to customer data orchestration – spanning web, mobile, offline and Internet of Things devices. With the power to unify customer data into a single source of truth, Tealium offers a turnkey integration ecosystem supporting more than 1,000 client-side and server-side vendors and technologies. The Tealium Universal Data Hub encompasses tag management, API hub, customer data platform, and data management solutions that enable organizations to leverage real-time data to create richer, more personalized digital experiences across every team, technology, and customer touchpoint. More than 800 businesses worldwide trust Tealium to power their customer data strategies.

For more information, visit www.tealium.com.