Building and Leading Machine Learning teams

Ajinkya More
ajinkya.more@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

ABSTRACT

Machine Learning (ML) is a growing area in the software industry, occupying a niche role in a handful of high tech companies a decade ago to becoming a critical and ubiquitous function across a large number of industries. As ML functionality matures and team sizes increase, having a standardized playbook around ML team management can be crucial to help with decision making around hiring, career growth, team culture, stakeholder management and measuring success. In this paper, we will discuss a few practices of running ML teams in one of the largest Data Science/Machine Learning organizations at Walmart Global Tech.

KEYWORDS

machine learning, data science, team management, hiring, stakeholder management, leadership, recruiting

1 INTRODUCTION

Machine Learning is a relatively new but rapidly expanding functional addition to many industries, especially in software. While there is some peer-reviewed literature around managing software engineering teams [5–7] and a lot of publications on ML model development [4], the literature around managing ML teams is comparatively scant. In this article, we outline the processes and frameworks for managing ML teams at one of the largest companies in the world.

2 HIRING

Evaluating ML skill sets systematically is a challenging proposition since there are many sub-areas within ML in which different candidates might specialize. These include but are not limited to Deep Learning (DL), Reinforcement Learning (RL), Natural Language Processing (NLP), Computer Vision (CV), Speech, Econometrics, Causal Inference (CI). In addition, candidates also need to demonstrate an understanding of common Data Structures and Algorithms, Software Design Patterns, Databases (SQL and NOSQL). Knowledge of statistics is also essential, especially in teams that do a lot of online experimentation. There are also specific domains (e.g., Search, Recommendations, Forecasting, Ads, Fraud, etc) across a variety of companies and demonstrating expertise in one or more of these can be important for certain senior-level roles.

Having a standardized interview process helps calibrate different candidates, set expectations for both candidates/interviewers and makes it easier for interviewers to have a consistent set of questions and/or evaluation criteria. Towards this end, we will describe the various components of our hiring process and the considerations with which they were designed.

2.1 Job description

Typically, the entry point for most candidates into the hiring funnel is a job application. This makes it extremely important to have a well-crafted job description (JD). We aim to keep JDs clear and concise since there is a limited amount of time candidates will spend looking at any given JD. This is in contrast with many job postings that are too verbose and include a large universe of skills to attract maximum candidates [2]. Our job descriptions are typically structured in the following way:

(1) A brief introduction to the team and the problem space.
(2) Enumeration of the key responsibilities for the role.
(3) Expected skills from the candidate.
(4) Skills and experience that might be ideal to have but not necessarily required.

2.2 Job advertisement

For every 100 resumes we receive, 1-2 might lead to an offer. Anecdotally, these numbers don't strike as unusual across a number of large tech companies. As such, we need to ensure we have as wide a reach for any job advertisement as possible. A wider funnel also helps improve the diversity and inclusiveness of the candidate pool. We target the following avenues to market the job postings:

(1) Company careers website (external-facing as well as internal)
(2) LinkedIn job posting
(3) LinkedIn posts from hiring managers

2.3 Resume vetting

Every job posting might receive hundreds of applications. The first step in the evaluation process is to select candidates for interview based on their resume and answers to some questions (e.g., having legal authorization to work). The candidates in this stage are chosen by a combination of hiring managers, recruiters and senior members of the team taking into account their education, prior industry experience, technical skills (programming languages, frameworks, knowledge of common ML models), any publications, etc. Roughly 5% of candidates make it through this stage.

2.4 Initial screening

The first step of the interview process is an initial screen to understand the candidate's background as well as their ML breadth and depth. It is also an opportunity to give the candidate an overview of the team's work. The interviewer might ask the candidate to go over one of the recent relevant projects that the candidate has worked on and walk through the following dimensions:

(1) The problem context and overall goals for the project
(2) Quantitative problem formulation
(3) The kind of ML techniques used to solve the problem
(4) Model metrics and baselines to assess model performance

(5) Business/project success criteria
(6) Deployment strategy and infrastructure (senior candidates)

This may be viewed as a thesis defense style interview, where the candidate presents their past work and the interviewer might probe at various points to understand the work in more depth.

2.5 Final interview loop

The final part of the interview process is a set of 3 to 4 additional interviews (alternately called an on-site or a virtual on-site depending on how the interviews are conducted). These include:

2.5.1 ***Machine Learning Theory/Applications.*** In this round, we aim to understand ML theory breadth and depth of the candidates and how they can apply ML algorithms in practical settings. These questions are typically tailored around the areas of expertise of the candidate such as NLP, DL, CI, etc. The questions tend to be open-ended with follow-ups that depend on the candidate's responses. Some examples include:

(1) How can we tackle vanishing/exploding gradients problem in RNNs? (Deep Learning)
(2) How can we modify the Inverse Propensity Score (IPS) estimator to reduce variance? (Causal Inference)
(3) How can we generate confidence intervals for a quantity of interest from a sample of data? (Statistics)

2.5.2 ***Coding.*** The success of most ML work in industry and in particular in our teams is predicated on integrating it into research and/or production models. This requires implementing any designed algorithm in a high-level programming language (e.g., Python, Scala). The interviews are conducted in a language agnostic manner, with candidates having the choice of selecting a programming language they are comfortable with. The expectation is to write clear, readable, modular, testable code that can compile. The interview is conducted in one of several popular platforms for coding interviews.

Coding interviews are common across many specializations in the tech industry. In many companies, leetcode [8] style questions are generally used to conduct these interviews. For companies and roles where a standardized hiring process is suitable, this presents an effective way to structure interviews, train interviewers and calibrate feedback. One downside of this standardization is that candidates often complain that these interviews are not reflective of the work they do on a daily basis. Since we do team-based hiring, we made the conscious decision of moving away from these sorts of algorithmic puzzles and instead focusing on designing problems that are representative of the kind of work our team members do.

2.5.3 ***Business Case Study.*** Most problems ML teams at Walmart work on have a clear business use case. Usually, the problem originates in discussions with business and product teams to solve a customer pain point. Often, at the early stages, the problem is not well-defined and only vaguely understood. Data science teams work to translate this use case into a quantitative problem that can potentially be solved using ML techniques. However, this journey requires many iterative conversations with stakeholders that have very varying degrees of technical understanding. As such, the ability to communicate with and understand the requirements of a non-technical audience becomes critical. Being able to drive consensus/adoption based on model recommendations is equally crucial. We test these skills in this round using a specific business use case. Candidates are expected to ask clarifying questions and determine the most appropriate way to cast it into an ML problem taking into account or making reasonable assumptions about data availability and distribution (e.g., unsupervised vs supervised, classification vs regression, point estimates vs interval estimates, etc).

2.5.4 ***Machine Learning System Design.*** For candidates with industry experience, this round focuses on their ability to build or integrate within an ML system. ML models in the industry are not built in a vacuum. At Walmart, they are usually part of a large-scale system and understanding the considerations to build models at this scale is an important signal for experienced candidates. These include batch vs real-time processing, monolith vs microservices, database design, latency and throughput expectations, etc.

2.5.5 ***Feedback.*** We use a standard feedback form to structure candidate assessment. The interviewer calibrates the candidate's performance across dimensions such as technical and soft skills (communication, ability to work with feedback, etc.) and whether the candidate raises the bar. They also provide an overall decision on a 4-point scale: Strong hire, hire, no hire and Strong no hire. This prevents interviewers from being on the fence. Additionally, interviewers capture the questions asked and the candidate's responses in a feedback form to add nuance to any overall evaluation.

2.5.6 ***Decision-making framework.*** In cases where the interview committee is in complete alignment (all yes vs all no), the decision is relatively straightforward. If the committee is split, we perform committee debriefs to align on the next steps. This may result in a decision or another follow-up round to get more resolution on a particular aspect of the candidate's experience.

2.5.7 ***Other considerations.*** The entire interview loop thus consists of 4-5 rounds, each typically lasting 45 minutes for a well-rounded evaluation of the candidate's skills and experience. The process lasts less than 4 hours and can be flexibly scheduled according to the needs of the candidate. This allows us to get a strong signal without imposing extreme demands on the candidates' time. We also experimented with or considered other formats but decided against them.

(1) Presentations: Many candidates we interview hold Ph.D. degrees or have built ML models in earlier jobs or competitions. Some may have published papers or given talks about them. Having candidates present about a topic they have developed expertise in seemed very appealing. However, after trying to host these interview rounds, the logistical challenges of aligning calendars of multiple panelists seemed untenable for the rate at which we were hiring.
(2) Take-home challenge: Another common interview format employed by many companies is a homework problem or a take-home challenge. Many candidates find it hard to perform under interview pressure and this might potentially help them work on a problem at their convenience without someone peering over their work in real-time. This can also help make interview problems more representative, giving

the candidate a chance to iterate through multiple hypotheses, perform data exploration, try out different models and present the final results from the best model. However, many candidates are unable to budget the necessary time to work on such open-ended challenges. This can result in a reduced and biased hiring pool. To avoid this, we decided to keep all our interview rounds time-boxed to 45 minutes.

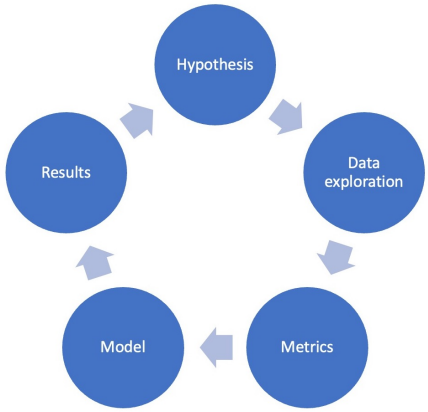
ML interviews are not an exact science and different formats may work for different companies. Considering the large volume of hiring, the kinds of problems we work on, the skill sets we need and keeping the process inclusive, we crafted the above interview process. We have been pleased with the candidates hired via this process. The process has been fine-tuned for precision rather than recall so we may occasionally miss strong candidates.

3 CULTURE

Culture is an effective way to promote best practices within a team. It also helps set expectations for team members to succeed in their roles. We will discuss a few aspects of building a strong ML culture.

3.1 ML rigor

Building ML models is an iterative cycle and it can be helpful to have a recipe to structure this process. We will illustrate this process with an example.



- **Hypothesis**: We may begin with a concrete hypothesis (which may eventually get redefined). An example would be "Including image-based features can improve search relevance for an e-Commerce catalog".
- **Data exploration**: This is usually followed (or even preceded by) a data exploration phase to evaluate model feasibility and potentially available features. In the above example, we may extract product brand and color from the image or a low dimensional vector space embedding representation of the image as features.
- **Metrics**: We need to define how we will measure improvements to the model. In our example, this might include online metrics like click-through rate. However, before we can test a model online, we may iterate through various models offline. Having offline metrics to track model performance is equally important. For the example above, we might choose to have a pointwise [1] model for search relevance learnt using binary labels and use AUC as a metric to monitor and compare performance of models. We also need to establish appropriate baselines. For a mature model in production, a natural baseline would be the current production model.
- **Model development**: Once we have aligned on clearly defined metrics, we may experiment with different modeling strategies and select the best-performing model according to our chosen metrics. Care must be taken to ensure model development is reproducible for validation and debugging.
- **Results**: For online experimentation, we may choose several candidate models to pick the best model according to online metrics. Often, we may have a suite of metrics for deciding a model rollout strategy. This may necessitate well-defined rules for prioritizing different metrics to determine a rollout candidate. This stage might also include communicating the results to various stakeholders in business as well as product partners. ML scientists might need to appropriately tailor their presentations to non-technical audiences. The whole process is iterative and at any stage, it may be necessary to go back to a previous step to check our assumptions based on the findings of that stage.

3.2 Experimentation

Online A/B testing is the gold standard for conducting randomized controlled experiments to understand the impact of any user-facing changes. Gating ML model updates via an A/B test is a common paradigm in the industry. That said, given we cannot A/B test all potential hypotheses, selecting appropriate candidate models for testing is paramount. This can be driven by domain experience and results from past experiments while aligning with business goals. We also need to be cognizant of the gap between online and offline metrics and how to use offline metrics to guide decisions around what models to test. In some domains, ML models might not be used for user-facing products or A/B testing might not be feasible for other reasons. In these cases, it's essential to develop validation techniques that work within the governing constraints.

3.3 Knowledge dissemination

Our teams work across a variety of domains under ML: Forecasting, NLP, CV, CI, etc. Across different projects, there can be varying degrees of overlap. Having a forum for knowledge dissemination can help ML scientists learn from each other, get feedback on their work and also gain visibility across the broader team. To this end, we run a weekly seminar to discuss modeling progress, results, infrastructure improvements, new datasets and even new business areas amenable to ML-based techniques.

In addition to presenting to internal audiences, team members are encouraged to showcase their work to the broader ML community by publishing in various academic and industry ML conferences (KDD, Neurips, INFORMS, MLConf, etc.) as well as on our company tech blog. These avenues tend to also help with our recruiting efforts where potential candidates can get to see the diversity and impact of the work our teams are doing.

4 PERFORMANCE MANAGEMENT

ML model development process is inherently stochastic. It is likely that not every hypothesized model improvement leads to success.

With mature products, the A/B testing success rate can be as low as 25-30% [3]. As such, assessing the performance of individuals based on the success rate of the models they developed appears very limited. We evaluate performance of ML scientists using a three-dimensional framework:

(1) **Competencies**: These include the various skills that are necessary to succeed in the role within our teams including: business acumen, quantitative problem formulation, data exploration, modeling, coding/testing, data visualization, model selection and model deployment/scaling. It can be seen that these competencies tie directly to the areas we cover in our interview process.
(2) **Proficiency**: Within each competency, we have multiple degrees of proficiency that are expected at different levels on the individual contributor ladder. For example, an entry-level ML scientist might be expected to take an existing model and add a few features to potentially improve model performance with a lot of guidance. In contrast, those at more senior levels will be expected to operate with increasing levels of autonomy and independence to own the entire model development process from conception to production.
(3) **Aspects**: Within each proficiency under a competency, multiple aspects can be demonstrated with concrete examples from an individual's work. For example, for model deployment, this might include: selecting appropriate metrics for model monitoring (data drift, model drift, etc.), model lifecycle management (release cadence, training triggers, etc.) and failure management (how to recover from training or publishing failures).

This framework is used in our annual performance evaluation and promotion decisions. Managers are encouraged to have ongoing performance review discussions with their team members to assess any gaps at the current or target level.

5 STAKEHOLDER MANAGEMENT

Typically, an ML scientist works with 3 stakeholders: business partners, product partners and the immediate team. The business team generally owns the P&L (profit and loss statement) for the product and is responsible for making investment decisions in the area. The product team understands customer needs and uses them to create a product roadmap. The immediate ML team is in charge of executing on the roadmap by delivering ML-based solutions.

5.1 Project scoping

This might include discussions on creating a new ML product or adding features to an existing one to enhance its performance. This step involves many conversations with business and product partners, data exploration, feasibility considerations and dependency checks to decide how the new project fits in the product roadmap.

5.2 Prioritization

Even at large tech organizations, there are usually more interesting problems than people to solve them. Thus, prioritization is key. Given a list of potential projects to include in the roadmap, the product partners come up with a relative prioritization of projects in discussion with the ML team by analyzing return on investment (ROI) vs effort. High impact, low effort projects usually get the highest priority while low impact, high effort projects might get put in the backlog. Everything in between requires nuanced deliberation.

5.3 Communication

As mentioned in the hiring section, communicating with stakeholders with varying levels of technical expertise is a key trait we look for in ML scientists. This is important to set appropriate stakeholder expectations during project scoping. One common pushback we have seen from non-technical audiences is the discomfort of launching probabilistic solutions where in any given instance, the model may not predict the right answer even if on average the model does very favorably on metrics of interest. We have found that carefully selecting appropriate business metrics to show the impact of model performance helps alleviate these concerns. Similar, considerations apply while communicating model results.

6 SUMMARY

In this article, we outlined the processes and learnings from building and managing ML teams in a large organization. Understanding the business and then being deliberate about creating structures for hiring, culture building, growth and management can help teams scale up and use ML to make a massive impact that aligns with the company's goals and values.

7 AUTHOR BIO

Ajinkya More is a Director of Data Science, Machine Learning at Walmart Global Tech, currently managing the teams focused on Pricing Algorithms, Clearance Algorithms, Item relationships and Shipping Speed Elasticity. Earlier, he managed the teams working on Merchandizing recommendations for Apparel, Item Matching and Universal demand quantification and an ML infrastructure team. Prior to this role, he worked on Personalization and Recommendations at Netflix. Earlier he worked in Search at WalmartLabs and also managed the Product Catalog Data Science team. He has a Ph.D. in Mathematics from University of Michigan, Ann Arbor.

REFERENCES

[1] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. 2009. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems* 22 (2009).
[2] Nancy Collamer. 2014. Why So Many Job Postings Are So Ridiculous. (2014). https://www.forbes.com/sites/nextavenue/2014/09/22/why-so-many-job-postings-are-so-ridiculous/?sh=108041995ad0
[3] Jen Dante. 2018. Consumer Science and AB testing. (Nov. 2018). https://www.youtube.com/watch?v=pDjDMuwQuWo
[4] Karen Hao. 2019. "We analyzed 16625 papers to figure out where AI is headed next". (Jan. 2019). "https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/"
[5] Eirini Kalliamvakou, Christian Bird, Thomas Zimmermann, Andrew Begel, Robert DeLine, and Daniel M. German. 2019. What Makes a Great Manager of Software Engineers? *IEEE Transactions on Software Engineering* 45, 1 (2019), 87–106. https://doi.org/10.1109/TSE.2017.2768368
[6] Petri Kettunen. 2013. The Many Facets of High-Performing Software Teams: A Capability-Based Analysis Approach. In *Systems, Software and Services Process Improvement*, Fergal McCaffery, Rory V. O'Connor, and Richard Messnarz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 131–142.
[7] Jil Klünder, Carolin Unger-Windeler, Fabian Kortum, and Kurt Schneider. 2017. Team Meetings and Their Relevance for the Software Development Process Over Time. 313–320. https://doi.org/10.1109/SEAA.2017.57
[8] Erin Schaffer. 2021. Is LeetCode the best way to prepare for interviews? (March 2021). https://www.educative.io/blog/leetcode-interview-prep