Machine Learning in the Real World

Vineet Chaoji
vchaoji@amazon.com

Rajeev Rastogi
rastogi@amazon.com

Gourav Roy
gouravr@amazon.com

Machine Learning
Amazon.com
Bangalore, India

ABSTRACT

Machine Learning (ML) has become a mature technology that is being applied to a wide range of business problems such as web search, online advertising, product recommendations, object recognition, and so on. As a result, it has become imperative for researchers and practitioners to have a fundamental understanding of ML concepts and practical knowledge of end-to-end modeling. This tutorial takes a hands-on approach to introducing the audience to machine learning. The first part of the tutorial gives a broad overview and discusses some of the key concepts within machine learning. The second part of the tutorial takes the audience through the end-to-end modeling pipeline for a real-world income prediction problem.

1. INTRODUCTION

Broadly, ML [3, 4] involves use of statistical techniques for the discovery of relationships or patterns in data. ML algorithms construct mathematical models of the data to discover patterns. The models are subsequently used to make decisions or predictions on future data. For example, you can use ML models - based on past customer purchases, browsing history, and search patterns - to predict if a customer will purchase a product.

Machine Learning is being used extensively by companies across a broad spectrum of applications. Search, online advertising, product recommendations, spam detection, speech recognition, and object identification are but a few examples of applications where ML has had significant impact. There are many other areas such as game playing, unmanned cars, and automated question answering where ML is poised to drastically change the way technology affects our lives.

Not too long ago, Machine Learning as a field was primarily a subject of academic research or was limited to few domains (e.g., hand-written digit recognition). Multiple factors have helped popularize Machine Learning. In recent years, the proliferation of online applications and reduction in data storage costs have resulted in organizations collecting and storing large volumes of data. Simultaneously, researchers have developed ML algorithms that can crunch large volumes of data, either in a streaming or distributed fashion. Lastly, popular distributed computing systems/platforms such as Hadoop and Spark have embraced and integrated these scalable ML algorithms. The confluence of these factors has empowered anyone with access to commodity hardware to experiment with ML techniques to solve their business problems. The success of open challenges such as the Kaggle and Netflix competitions bears testimony to the widespread proliferation of Machine Learning.

We believe that understanding basic ML concepts and tools is as fundamental to data science as design patterns are to software development. With that goal in mind, this tutorial is designed to provide an introduction to ML concepts while focusing on applying the concepts to build models for a real-life problem.

2. TARGET AUDIENCE

The tutorial is targeted towards industry practitioners, students and researchers who have limited knowledge of ML, and people who want to learn more about the practical aspects of how ML is used to solve problems in the real world.

The tutorial does not have any specific requirements, although a basic understanding of algorithms, linear algebra, and convex optimization could help better appreciate the material. Familiarity with Python or an equivalent programming language would also simplify the hands-on section of the tutorial.

3. SCOPE AND STRUCTURE

The first part of the tutorial will cover basic ML concepts and techniques such as Classification and Regression, Overfitting and Underfitting, Linear Models and Random Forests. The second part, which constitutes a bulk of the tutorial, will cover the end-to-end modeling process right from ML problem definition to data preparation and cleaning to data visualization to feature engineering to model training and performance evaluation. Participants will actually train and evaluate models for a real-world *income prediction* problem using Spark ML Pipelines [1] via a Jupyter Notebook interface, as shown in Figure 1.

Thus, participants will not only get an exposure to key ML concepts and methods, but will also get their hands dirty with state-of-the-art ML systems and technologies.

Proceedings of the VLDB Endowment, Vol. 9, No. 13

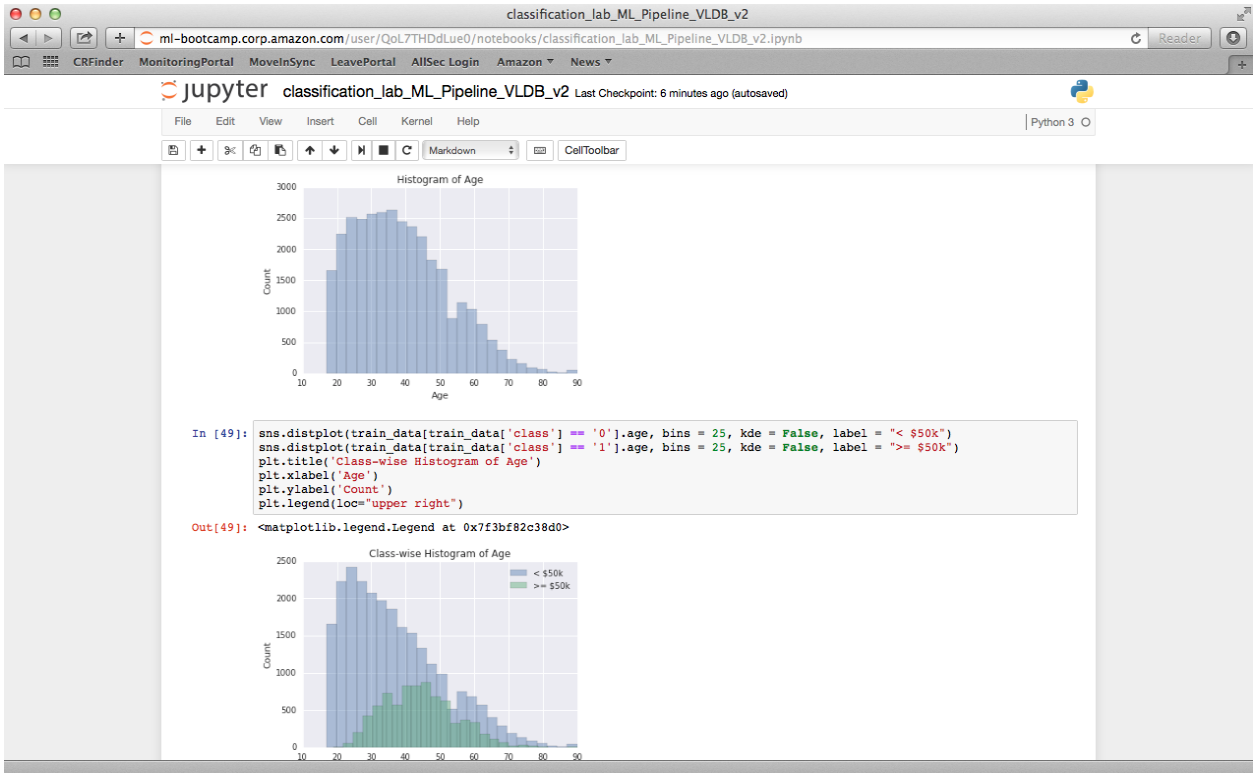

Figure 1: Screenshot of Jupyter Notebook for model building and visualization.

The tutorial will have the following structure:

Basic ML concepts and techniques: (45 minutes)
Within this part of the tutorial we will discuss the following concepts.

- **Supervised learning**: We will start with "What is ML?", "Where is ML useful?" and move on to supervised learning. From a practitioner's perspective, supervised learning is one of the most applicable area within ML. Based on the type of the output variable, supervised learning is further divided into Classification and Regression. We will introduce the concepts behind Classification and Regression.
- **Overfitting and Underfitting**: ML model's health can be determined by looking at the fit of the model to the data. Overfitting and underfitting are concepts that relate to the fit of the model and they drive actions to improve the model. For instance, overfitting occurs when the model fits the training data well but does not generalize to unseen or test data. One can combat overfitting by increasing the amount of data or making the model less complex. On the other hand, underfitting occurs when the model lacks the expressive power to capture the target. One can combat underfitting by increasing model complexity.
- **Linear Models**: These are the simplest and at times the best class of models for handling extremely large datasets. We will go into the details of linear models.
- **Random Forests**: Random Forests fall within the class of non-linear models. While they perform better than linear models for many use cases, they are computationally much more expensive to train.

End-to-End ML Modeling process: (2 hours)

The second part of the tutorial will be the hands-on session where participants will build and evaluation ML models using Spark ML Pipelines via a Jupyter Notebook (Figure 1). Traditionally, interfaces such as R and Matlab were popularly used for exploratory analysis and subsequently model development. The new era of interfaces are browser-based interactive webpages (popularly called as *notebooks*) that support powerful visualization constructs and can invoke scalable ML algorithms that are hosted on a cloud service (e.g., Amazon EC2). In this tutorial we will use the Jupyter Notebook interface backed by ML algorithms from the Spark MLlib library, running on an Amazon EC2 instance. Figure 2 shows the typical steps involved while building an end-to-end ML modeling pipeline, even for a production quality setup. The figure also shows the inputs and outputs of each building block in the modeling pipeline. Some of the steps are described in further detail below.

- **ML Problem Definition**: This step helps formulate a business problem as a Machine Learning problem by translating components of the business problem into *observations* that comprise *features* and a *target label*. Defining elements that constitute observations and labels is a prerequisite for the following steps within the modeling pipeline. In addition to framing the business problem as an ML problem, this step also maps business metrics to ML model performance metrics.
- **Data Preparation and Cleaning**: Once observations and features are generated from raw data, one has to ensure that the data is ready to be fed into the algorithm. This step typically involves identifying (and eliminating) erroneous values and outliers, filling in missing values, and splitting the data into train and validation sets. Finally, for

some models the data might have to be randomly shuffled to ensure proper convergence of the algorithm. This is true, especially for algorithms that learn from one observation at a time, rather than algorithms that learn from the entire batch of training examples.

• **Data Visualization**: Understanding the data is essential for developing an intuition for the problem. Visual inspection of the data develops the necessary intuition. Visualization typically involves slicing and dicing the data along multiple dimensions to plot data distributions and histograms for features, and determine the correlations between the features and the target. Visualization can also provide cues for missing or erroneous data.

• **Feature Engineering**: Often raw features are not in a form that is best suited for a learning algorithm. Yet, appropriate transformations of the raw input features can result in new features with significantly higher predictive power and models that are more accurate. In many instances, linear models with simple features may be inadequate for capturing complex correlations between the data and the target label. One way to improve the expressive power of linear models is to introduce non-linearity through feature transformations. Feature engineering has multiple forms – Non-linear data transformations, domain-specific features, data-driven features and feature selection.

• **Model Training and Parameter Tuning**: The fundamental goal of Machine Learning is to generalize beyond the examples in the training set – because, no matter how much data we have, it is unlikely that we will see those exact examples again at prediction time. However, since future examples are unknown, learning algorithms typically train a model by maximizing its performance relative to a specified objective function on the training dataset. The objective function value on the training data essentially serves as a surrogate for the error on unknown examples. We want to minimize these errors. So, the training phase involves identifying a model that maximizes its performance (or minimizes its loss) in the context of the training data. Further, the quality of trained models can be improved by appropriately setting the model parameters.

• **Model Performance Evaluation**: Evaluating the model's performance is essential to selecting the best-performing model that meets business objectives. There are multiple aspects to consider while evaluating the model – which data should the model be evaluated on, what metric should be used for evaluation, etc. The tutorial will address some of these questions and discuss commonly used evaluation metrics such as AUC, precision, recall, etc.

What is not covered? Machine Learning is a a broad field and a single tutorial cannot do justice to the entire domain. We have deliberately excluded topics such as unsupervised and reinforcement learning. We have also not covered detailed theoretical analysis of ML concepts and techniques. Other advanced techniques such as multi-class classification, semi-supervised learning, latent factor models, and sequential models have also been ommitted. We hope that practitioners will follow up on these topics as the need arises in their line of work. Books to get a deeper understanding of ML topics include Bishop [3], Mitchell [5], Hastie et al. [4], Murphy [6], Barber [2].

4. INTENDED LENGTH OF THE TUTORIAL

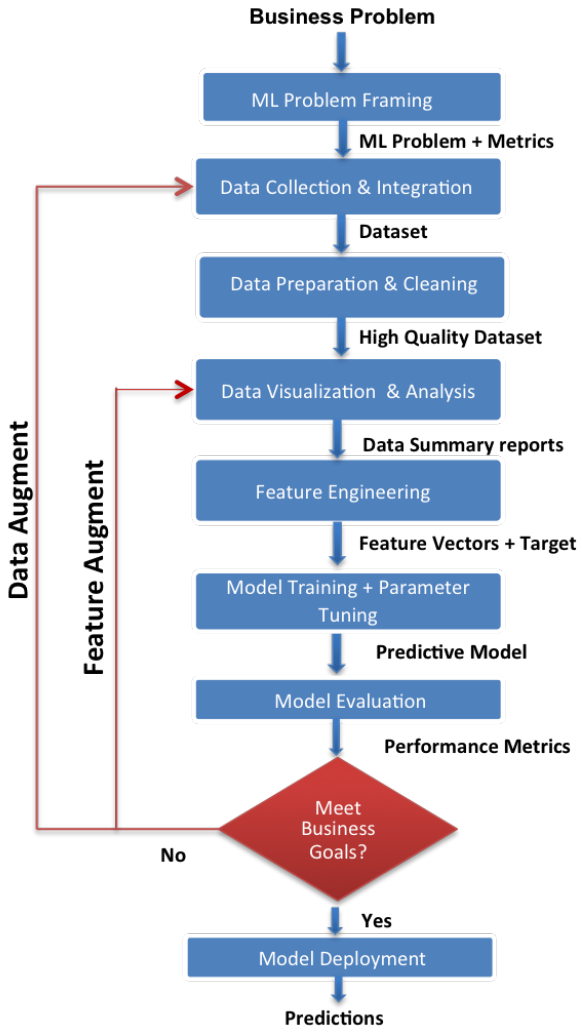

Figure 2: End-to-end Modeling Building Pipeline

The tutorial will be conducted over 3 hours. As mentioned above, it consists of two parts the first part covering ML concepts will take 45 minutes followed by a hands-on session that will take about 2 hours.

5. SPEAKER BIOGRAPHIES

Vineet Chaoji is an Applied Science Manager within the Machine Learning team at Amazon where he leads projects related to econometric models of customer behavior, customer targeting and malware detection. Prior to joining Amazon, he was a Scientist at Yahoo! Labs in Bangalore where his research focused on online advertising and social networks. Vineet obtained a PhD in Computer Science from Rensselaer Polytechnic Institute. He has published at top-tier data mining and database conferences and journals. Vineet has also served on the program committees of leading data and web mining conferences.

Rajeev Rastogi is the Director of Machine Learning at Amazon where he directs the development of machine learning platforms and applications such as product classification, product recommendations, customer targeting, and

deals ranking. Previously, he was the Vice President of Yahoo! Labs in Bangalore where he was responsible for research programs impacting Yahoo!s web search and online advertising products. He was named a Bell Labs Fellow in 2003 for his contributions to Lucent's networking products while he was at Bell Labs Research in Murray Hill, New Jersey. Rajeev was named an ACM Fellow in 2012 for his contributions to large-scale data analysis and management. He has published over 100 papers in top-tier international conferences and 33 papers in international journals. Rajeev has also been a prolific inventor with 57 issued US Patents. He is currently a member of the News editorial board of the CACM, and was previously an Associate editor for TKDE. He has served on over 50 program committees of the leading database and data mining conferences, and was a Program Co-chair for the Applied Data Science track of the KDD conference in 2016, the CIKM conference in 2013 and the ICDM conference in 2005.

Gourav Roy is a Software Engineer in the Machine Learning team at Amazon where he builds scalable machine learning platforms and applications. He is interested in streaming approximate algorithms and distributed systems. His work on streaming anomaly detection recently got accepted at the International Conference on Machine Learning. Prior to joining Amazon, he got a bachelors degree in Computer Science at BIT Mesra.

6. REFERENCES

[1] ML Pipelines. `https://amplab.cs.berkeley.edu/ml-pipelines/`, 2014. [Online; accessed 17-July-2016].

[2] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, New York, NY, USA, 2012.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[5] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[6] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.