Business Applications of Predictive Modeling at Scale

Qiang Zhu[1], Songtao Guo[1], Paul Ogilvie[2], Yan Liu[1]
Business Analytics[1] and Engineering[2] at LinkedIn Corporation
2029 Stierln Ct, Mountain View, CA 94043 USA
{qzhu, soguo, pogilvie, yliu}@linkedin.com

ABSTRACT
Predictive modeling is the art of building statistical models that forecast probabilities and trends of future events. It has broad applications in industry across different domains. Some popular examples include user intention predictions, lead scoring, churn analysis, etc. In this tutorial, we will focus on the best practice of predictive modeling in the big data era and its applications in industry, with motivating examples across a range of business tasks and relevance products. We will start with an overview of how predictive modeling helps power and drive various key business use cases [5]. We will introduce the essential concepts and state of the art in building end-to-end predictive modeling solutions, and discuss the challenges [6], key technologies, and lessons learned from our practice, including case studies of LinkedIn feed relevance [1-4] and a platform for email response prediction. Moreover, we will discuss some practical solutions of building predictive modeling platform to scale the modeling efforts for data scientists and analysts, along with an overview of popular tools and platforms used across the industry.

Keywords
predictive modeling; business analytics; machine learning; machine learning platforms

1. INTENDED AUDIENCE
This tutorial is suitable for researchers, students, and practitioners of predictive modeling who are interested in the industry applications. Advanced techniques in data mining and statistical modeling are not required but some background in statistics and big data is expected.

2. OUTLINE
The tutorial consists of three main sections: an introduction, an overview of predictive modeling, and considerations when choosing a framework for predictive modeling.

2.1 Introduction
In the introduction, we will briefly survey the audience to better understand their goals and adapt the depth of information presented during the tutorial.

We motivate the tutorial with prominent examples of predictive models and demonstrate how actionable prediction scores can fuel better ROI.

- Motivating examples
- Understanding of audience and learning objectives

2.2 Predictive Modeling Overview
Those applying predictive modeling in a business environment must carefully consider a wider range of aspects than commonly discussed in academic literature. Statistical methods and machine learning algorithms are just one component of full solution. This section describes the full range of considerations, details common challenges, and provides a concrete example of applying predictive modeling to the feed ranking problem at LinkedIn.

- End-to-end walkthrough of a production modeling solution
  - Label preparation
  - Data integration and feature engineering
  - Machine learning algorithms
  - Model management
  - Performance measurement through A/B test
- Common pitfalls and challenges
- Case Study - LinkedIn Feed Ranking

2.3 Choosing a Framework
A practitioner in industry may be faced with the challenge of choosing or building a framework for predictive modeling within their company. Deciding whether to build or buy depends on a range of considerations, which the tutorial presents. We also present an overview of existing platforms and open source software, closing with a concrete example of the decisions made for a propensity modeling developed at LinkedIn.

- Considerations when choosing a framework
- Platforms
  - Amazon Machine Learning[1]
  - Databricks[2]
  - Microsoft Azure Machine Learning[3]
  - Google Cloud Machine Learning Platform[4]
  - H2O[5]
  - Dato[6]

KDD '16, August 13-17, 2016, San Francisco, CA, USA
ACM 978-1-4503-4232-2/16/08.
http://dx.doi.org/10.1145/2939672.2945388

[1] https://aws.amazon.com/machine-learning

[2] https://databricks.com

[3] https://azure.microsoft.com/en-us/services/machine-learning

[4] https://cloud.google.com/ml/

[5] http://www.h2o.ai

[6] https://dato.com

- Open source software
  - Vowpal Wabbit[7]
  - Spark MLlib[8]
  - DMLC[9]
  - Scikit-learn[10]
  - R[11]
- Modoop - example of a scaled framework

3. PRESENTER INFORMATION

Qiang Zhu is a Staff member of Business Analytics Data Mining team at LinkedIn. He and his team apply advanced Data Mining techniques to drive LinkedIn's monetization efforts, ranging from a machine learning platform which powers member Email Marketing, to Sales Intelligence tools while help salespeople sell smarter. Prior to joining LinkedIn, he worked at StumbleUpon as a Data Scientist. Qiang holds a PhD in Computer Science from University of California, Riverside. His work has appeared in many top tier Data Mining conferences and journals, including the one which won the Best Paper Award in SIGKDD 2012.

Songtao Guo is a Principal Data Scientist and tech lead of Data Mining team at Linkedin where he leads many of data driven products and analytics systems. His work involves building large-scale knowledge base as one of the foundations of LinkedIn's Economic Graph, inventing data mining platforms to scale business analytics and partnering with product, sales, and marketing to deliver impactful solutions. Before joining LinkedIn, Songtao was a senior researcher at AT&T interactive, focusing on improving data quality and search relevancy for local business search. He holds a PhD in computer science from University of North Carolina at Charlotte where he studied privacy preserving data mining.

Paul Ogilvie manages the Machine Learning Algorithms team in the Engineering organization of LinkedIn. The team's mission is to research and develop the learning algorithm libraries and datasets that help research scientists more productively build state-of-the art relevance models. He earned his PhD in Language and Information Technologies from Carnegie Mellon University in 2010, where he studied semi-structured information retrieval with applications to web search, XML element retrieval, and question answering systems. He has previously worked on news recommendation at a startup (mSpoke) and at LinkedIn.

Yan Liu manages the Data Mining team at LinkedIn Analytics group. She leads various data mining initiatives and efforts in building advanced intelligence solutions and scalable data mining platforms to create leverage and drive business impact across functions. Before joining LinkedIn, she worked on search relevance and personalization at NexTag. Yan holds a Ph.D. in statistics from University of Virginia and B.S. in computer science from China.

4. REFERENCES

[1] Agarwal, D., Chen, B.C., Gupta, R., Hartman, J., He, Q., Iyer, A., Kolar, S., Ma, Y., Shivaswamy, P., Singh, A., and Zhang, L. 2014. Activity ranking in LinkedIn feed. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. ACM, New York, NY, USA, 1603-1612.

[2] Agarwal, D., Chen, B.C., He, Q., Hua, Z., Lebanon, G., Ma, Y., Shivaswamy, P., Tseng, H.P., Yang, J., and Zhang, L. 2015. Personalizing LinkedIn Feed. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1651-1660.

[3] Lebanon, G. 2015. Making Your Feed More Relevant – Part I. November 17, 2015. Retrieved June 12, 2016 from https://engineering.linkedin.com/blog/2015/11/making-your-feed-more-relevant--part-i

[4] Lebanon, G. 2016. Making Your Feed More Relevant – Part 2: Relevance models and features. March 15, 2016. Retrieved June 12, 2016 from https://engineering.linkedin.com/blog/2016/03/making-your-feed-more-relevant--part-2--relevance-models-and-fea

[5] Rosenberg, C. 2015. B2B Predictive Analytics Technology Report: Best practices, tools, and vendor evaluations to help marketing and sales organizations adopt predictive analytics. July, 2015. Retrieved June 12, 2016 from Infer: https://www.infer.com/wp-content/uploads/2015/08/TOPO-Predictive-Analytics-08-03-15.pdf

[6] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., and Young, M. 2014. Machine Learning: The High-Interest Credit Card of Technical Debt. *Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

[7] https://github.com/JohnLangford/vowpal_wabbit/wiki

[8] http://spark.apache.org/docs/latest/mllib-guide.html

[9] http://dmlc.ml

[10] http://scikit-learn.org

[11] https://www.r-project.org