VECTOR INSTITUTE | INSTITUT VECTEUR

1000AIMs: 1,000 AI-Related Master's Graduates per Year, Within Five Years

Vector Institute Guidance for 1000AIMs AI-Related Master's Programs

Table of Contents

VECTOR INSTITUTE GUIDANCE FOR 1000AIMS AI-RELATED MASTER'S PROGRAMS

BACKGROUND AND CONTEXT

"1000AIMs" is an initiative of the Vector Institute (Vector) established to address the province of Ontario's goal of accelerating the number of artificial intelligence (AI)-related master's graduates to 1,000 per year, within five years. The province has committed $30 million of funding to work toward this target in partnership with Vector. More information about 1000AIMs is provided in the Appendices.

This document has been prepared to provide guidance that supports and enables universities who are interested in offering new and/or enhanced AI-related master's programs. The document integrates the results of consultations that took input and advice from over 150 individuals, including a working group established by Vector comprising representatives from universities, government ministries, Vector industry sponsors and health sector organizations.

The consultations and working group identified eight priority skills/competencies for **core technical AI-related master's graduates**:

1. Technical computing skills with, where relevant to the AI-related program's field of study, a particular focus on:
    - Visualizing data and results using appropriate software tools and libraries
    - Exhibiting good software engineering practices (e.g., code optimization, version control, software testing)
    - Programming in more than one computer language, choosing the most suitable language for the task
2. Computational thinking skills with, where relevant to the AI-related program's field of study, a particular focus on:
    - Formulating problems in a way that enables the use of computers and other tools to help solve them
    - Logically organizing and analyzing data
    - Representing data through abstractions such as models and simulations
3. Mathematics and statistics related to AI
4. Integration skills (i.e., the ability to integrate technical computing skills, computational thinking and knowledge and skills related to the domain in which the AI-related work is performed)
5. Practical application skills related to AI
6. Core methodological skills related to AI
7. Collaboration and communications skills
8. Knowledge of ethics and societal implications of AI

The consultation participants and working group members agreed that 1000AIMs should also support **complementary AI-related master's graduates**, i.e., master's graduates who have complementary knowledge and skills (including business skills, implementation skills and sector knowledge), who can work effectively on interdisciplinary teams helping to design, evaluate, refine and implement practical AI-related solutions and technologies that meet business and end user needs in real world settings.

GUIDANCE

GENERAL CONSIDERATIONS

AI-related fields include, but are not limited to: data science, machine learning and deep learning, computer vision, natural language understanding, intelligent robotics, knowledge representation and reasoning, intelligent agents, intelligent user interfaces, and intelligent medicine.

The working group has determined that, at this time, a guidance document alone cannot ensure that the essential requirements for AI-related master's programs are fulfilled because (a) it is not possible to issue guidance that speaks to all the potential variations in program design and (b) one or more expert panels will need to be involved to ensure that varying approaches to the AI-related content of new master's programs are of sufficient substance to meet employer needs.

Accordingly, Vector will establish one or more panels to operationalize the guidance presented in this document. The panel(s) will assess programs proposed by universities with demonstrable strength related to AI to determine whether the programs will be recognized by Vector as meeting the 1000AIMs essential requirements. The process for institutions to work with the panel(s) will be established as soon as practicable and will require institutions to submit a brief structured narrative foreground statement describing how the program will produce the required essential program level learning outcomes complemented with selected materials – that generally exist already – which support the claims in the narrative (e.g., course outline, curriculum vitae for instructors, syllabus, materials for within-institution quality assurance, completed degree level expectations tables in the case of programs undergoing Ontario Universities Council on Quality Assurance review, etc.). For clarity, the Vector panel(s) will not duplicate, or require as much information as, formal quality assurance or accreditation processes. Institutions should contact 1000AIMs@vectorinstitute.ai to ensure that they have the most current version of the structured narrative template before they begin submissions for panel consideration.

All master's programs that are recognized as being AI-related by Vector will need to fulfill the essential requirements and articulate program level learning outcomes that are specific to the field of study. Fulfilling the requirements set out in this guidance document, including panel approval, will not necessarily result in an approved program. Standard approval and quality assurance processes for new or modified programs will continue to be in effect.

A subset of new and enhanced AI-related master's programs will far exceed the essential requirements presented in this document and will produce world-class graduates. Vector will work with employers and AI experts to develop guidance and supports for world-class AI-related master's programs in the near future.

Vector will have a role in disseminating examples of programs and program elements that meet the essential requirements and programs that are deemed to be world-class, so that these programs can serve as models for institutions looking to enhance their own programs and/or develop new AI-related content.

ESSENTIAL REQUIREMENTS FOR CORE TECHNICAL AI-RELATED MASTER'S PROGRAMS

Vector will recognize new and enhanced core technical AI-related master's programs that are science, technology, mathematics and engineering (STEM) programs. These core technical AI-related master's programs will often be in computer science, engineering, mathematics or statistics, but can also be in other fields such as physics or informatics. Core technical AI-related programs must fulfill the following essential requirements in order to be recognized by Vector:

1. The program includes at least three (3) curriculum components[1] with learning outcomes focused on AI-related methodologies and applications
    a. At least one (1) curriculum component with intended learning outcomes focused on the application of AI-related methodologies to ensure that graduates have knowledge and skills related to algorithms and representations regardless of their application area. The focus of the AI-related methodology component(s) will vary depending on the master's program but it is strongly recommended that a machine learning curriculum component be offered. Other options could include: neural networks, deep learning, graphical models, reasoning under uncertainty, pattern recognition, planning, logic, and other topics.
    b. At least one (1) curriculum component that involves studying an AI-related application area in-depth to ensure that graduates are able to apply AI-related methodologies and have knowledge of their limits in solving problems. The focus of the AI-related application component(s) will vary depending on the master's program, but could include: computer vision, computational linguistics (NLP), intelligent robotics, intelligent agents, intelligent medicine, and other topics.
2. The program has learning outcomes related to communication, teamwork and interdisciplinary practice related to AI. This requirement could be satisfied through different forms of learning including, but not limited to, an internship with academic supervision and an industrial supervisor, or a capstone or culminating project.
3. The program has learning outcomes related to the ethics and societal implications of AI. This requirement could be satisfied through different forms of learning including a dedicated curriculum component or a module within an AI-related methodology or AI-related application course, or a thread that is integrated through and across courses.

[1] A curriculum component will often be a course (in-person or online) but a program could prepare master's graduates through various combinations of: classroom learning, distance learning, culminating or capstone projects, summer school, faculty-supervised research projects, etc. For clarity, a curriculum component could be a topic or competency that is developed as a thread through multiple courses as opposed to a stand-alone component.

ESSENTIAL REQUIREMENTS FOR COMPLEMENTARY AI-RELATED MASTER'S PROGRAMS

Vector will recognize new and enhanced complementary AI-related master's programs that may be STEM, business, social sciences or humanities master's programs. These complementary AI-related master's programs will integrate AI-related content into master's programs with a focus of study that applies AI technologies (e.g., business, public health, environmental sciences). Complementary AI-related programs must fulfill the following essential requirements in order to be recognized by Vector:

1. The program includes at least three (3) AI-related curriculum components with program level learning outcomes that are focused on AI-related applications and/or methodologies related to the primary field of study. The learning outcomes will ensure that graduates have sufficient knowledge of the relevant AI-related applications and methodologies to be able to contribute to AI-related work, providing essential input to the development, evaluation, refinement and implementation of AI-related methods, tools, products and services in real world settings, particularly those in the primary field of study.
2. The program has learning outcomes related to communication, teamwork and interdisciplinary practice related to AI. This requirement could be satisfied through different forms of learning including, but not limited to, an internship with academic supervision and an industrial supervisor, or a capstone or culminating project.
3. The program has learning outcomes related to the ethics and societal implications of AI. This requirement could be satisfied through different forms of learning including a dedicated component or a module within an AI-related methodology or AI-related application component, or a thread that is integrated through and across courses.

Appendix A: 1,000 AI-related master's graduates per year, within five years ("1000AIMs") Context and Consultation Results

March 29, 2018

Alison Paprica, PhD PMP, VP Health Strategy and Partnerships
Andrea Proctor, PhD, Director Health Care Engagement
1000AIMs@vectorinstitute.ai

VECTOR INSTITUTE | INSTITUT VECTEUR

Vector: Vision & Mission

Vision

Vector will drive excellence and leadership in Canada's knowledge, creation, and use of artificial intelligence (AI) to foster economic growth and improve the lives of Canadians.

Mission

- Vector will lead Ontario’s efforts to build and sustain AI-based innovation, growth and productivity in Canada by focusing on the transformative potential of deep learning and machine learning
- Vector, together with its AI partners in other parts of Canada, will work with Canadian industry and public institutions to ensure that they have the people, skills, and resources to be best in class at the use of artificial intelligence
- Vector will support Canada's innovation clusters in artificial intelligence and focus on helping start-ups grow to become Canadian-based global leaders
- Vector will attract the best global talent focused on research excellence; Vector’s researchers and academic partners will be part of a vibrant community of innovative problem solvers, working across disciplines on both curiosity-driven and applied research

Provincial Government Announcement

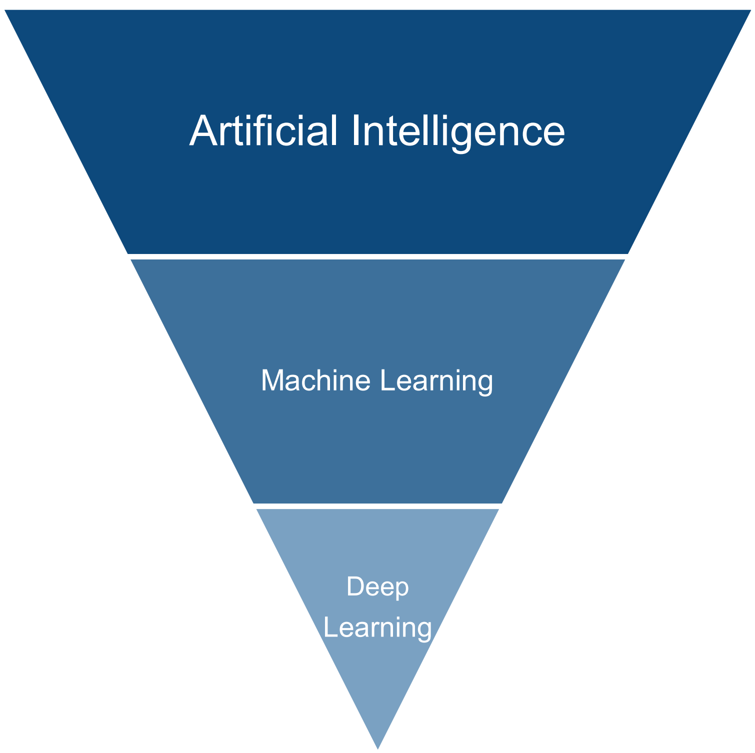
$30 million announced on October 18, 2018

…The goal is to graduate 1,000 applied masters students in AI-related fields per year, within five years

Ontario Boosting the Number of Graduates in Science, Tech, Engineering, Mathematics and Artificial Intelligence. (Ontario Government Newsroom, October 18, 2017). Retrieved January 10, 2018, from https://news.ontario.ca/medg/en/2017/10/ontario-boosting-the-number-of-graduates-in-science-tech-engineering-mathematics-and-artificial-inte.html

"AI-related" Master's Graduates

– AI refers to computers that can learn about the world flexibly, make inferences about what they see and hear, and achieve human-like understanding of information



– Vector's research focus is on machine learning and deep learning, but the scope of the "AI-related" master's programs of 1000AIMs is much broader.

– 1000AIMs AI-related fields include, but are not limited to: data science, machine learning and deep learning, computer vision, natural language understanding, intelligent robotics, knowledge representation and reasoning, intelligent agents, intelligent user interfaces, and intelligent medicine

Opportunities for 1000AIMs

- Support and enable universities that are interested in offering new and enhanced AI-related master's programs
- Build interest in AI-related education by demonstrating the return on investment (ROI) of AI-related master's programs, including to students and their families
- Help employers find AI-talent and integrate interns and new graduates of AI-related master's programs into the workforce in meaningful ways, thus building Ontario's AI ecosystem

Broad Strokes of 1000AIMs

- Two categories of master's programs will be recognized by Vector:
  - Core Technical AI-related Master's Programs
  - Complementary AI-related Master's Programs

- Three "tracks" of work:
  - Track 1: enhanced existing master's programs (mostly "on book", regulated tuition) – first will launch in 2018
  - Track 2: bespoke new AI-focused master's programs (mostly "off book", full cost recovery) – first may launch in 2019
  - Track 3: distributed delivery model of AI-related master's programs (i.e., models through which institutions that are not positioned to deliver an entire AI-related master's program make specific contributions toward degrees awarded in partnership with other institutions) – planning will initiate once Track 1 and Track 2 are established

Broad Strokes of 1000AIMs cont.

- The majority of the $30 million will be used for scholarships.
- 1000AIMs will be flexible, support innovation, and be developed and implemented in partnership with universities and employers:
  - Essential requirements for core technical AI-related master's programs and complementary AI-related master's programs were developed based on extensive consultation and by a working group with diverse representatives
  - Essential requirements are defined in the least prescriptive way possible ("*min specs*" approach) so that institutions are free to innovate in varying ways to fulfill the requirements (i.e., no restrictions are placed on modes of learning)
  - Noting most AI-related master's graduates will immediately join the workforce, master's programs that produce graduates the proceed to PhDs are also in scope
  - Flexibility and commitment to work with employers to update essential requirements as employer needs change

Process for Developing 1000AIMs Essential Requirements

- Identify starting point list of essential skills and competencies that graduates would need to have*
- Modify list based on advice from Vector faculty and other university-based AI experts
- Broad consultation (150+ participants) on essential skills and competencies including live polling, online survey and meetings
- Establishment of working group to produce guidance through a series of three meetings, considering consultation findings and their own experience
- Near final guidance shared with stakeholders

* Several lists were considered, Vector would like to thank the Compute Ontario Highly Qualified Personnel committee who agreed to have Vector use a list they developed as the best starting point

Working Group Members

Alison Paprica (Chair) – Vector Institute

Ayse Bener – Ryerson University

Brenda Brouwer* - Queen's University/Vector Institute+

Mark Daley – University of Western Ontario

Kevin Deluzio – Queen's University

Sven Dickinson – University of Toronto

Afsaneh Fazly – Thompson Reuters

Paul Fieguth – University of Waterloo

Michael Hillmer – Ministry of Health and Long-Term Care

Murat Kristal – York University

Muhammad Mamdani – Li Ka Shing Knowledge Institute

Bill Mantel – Ministry of Research, Innovation and Science/Ministry of Economic Development and Growth

John McLaughlin – Public Health Ontario

Tomi Poutanen – Layer 6 AI/TD

Andrea Proctor* – Vector Institute

Ladislav Rampasek~ – University of Toronto

Cindy Robinson – Council of Ontario Universities

Langis Roy – University of Ontario Institute of Technology

Marny Scully – Ministry of Advanced Education and Skills Development

Graham Taylor – University of Guelph/Vector Institute

* Non-voting member ~ Student representative

+ Incoming Head, Academic Partnerships at Vector, secondment from Queen's beginning July 3 2018

Results of Consultations

- Of a list of six high-level skills/competencies, five were identified as essential for AI-related master's graduates by at least one of the four stakeholder groups, with notable variation in what different groups identified as essential
  - Technical computing skills (identified as essential by almost all Vector faculty and university stakeholders)
  - Computational thinking skills (identified as essential by almost all Vector faculty and university stakeholders)
  - Integration skills (identified as essential by almost all health stakeholders and Vector industry sponsors)
  - Mathematics and statistics (identified as essential by almost all Vector faculty)
  - Team work and communications skills (identified as essential by almost all Vector industry sponsors)

Results of Consultations cont.

	Core Technical AI-related Master's	Comple-mentary AI-related Master's	Comments entered as "Other Advice"	Total
Total number of text responses	159	138	14	311
% of comments that made reference to ethics and societal implications of AI	6%	6%	0%	5%
% comments that made reference to practical application of AI, "real world" etc.	19%	13%	21%	17%
% of comments that made reference to collaboration including multi-disciplinary AI teams	7%	17%	21%	12%
% of comments that directly referenced the need for core technical AI-related master's grads to have complementary knowledge/skills (e.g., related to business and field of application) or complementary AI master's grads needing technical knowledge/skills	9%	22%	0%	14%

Examples of Advice on Most Needed Skills

What I hear from employees all the time is that business students need better technical skills while comp sci students need better business skills.

Practical, project-based application of deep learning and data analysis techniques to support marketing, sales and general business analytics.

Understand basics of AI/ML, but also what makes useful/valid data, ethics and governance related to data, and understanding opportunities to partner beyond just with AI experts.

Timeline and Next Steps

- ✓ Vector compiles advice on essential requirements through live facilitated sessions with follow up via open online survey distributed more broadly
- ✓ Working group with university, ministry and employer representatives develop guidance on essential requirements
- ✓ Convert input on essential requirements into a plain language document that articulates employer needs and learning outcomes by the end of March
- ❑ Develop process for submission to panels and disseminate via COU and web meetings (as soon as practicable)
- ❑ Provide institutions with information about support for direct delivery costs of new and enhanced programs (as soon as practicable)
- ❑ Establish scholarship process for Track 1 (as soon as practicable)
- ❑ Communications, marketing and registrations for Track 1 programs (as soon as practicable)
- ❑ Begin pairing students with internship opportunities with intent to integrate learnings into coordinated internship process established in 2018/19
- ❑ First Track 1 students receive scholarships and commence studies (fall 2018)
- ❑ Establish coordinated internship working group (2018)
- ❑ Review first group of Track 1 programs, make improvements (including updated and refined essential requirements), add Tracks 2 and 3 (spring 2019)

Email 1000AIMs@vectorinstitute.ai to join the distribution list and receive notifications as additional information becomes available

Appendix B: Forthcoming Master of Engineering in AI Program

University of Waterloo

Proposed Master of Engineering in Artificial Intelligence

Faculty of Engineering, University of Waterloo

Preamble: The area of Artificial Intelligence (and related areas of Machine Intelligence, Machine Learning) is rapidly growing in every respect – in terms of research interest, research impact, student interest, and industry. Waterloo Engineering has significant research experience in the area, coupled with extensive industry connections via its co-op programs and similarly extensive experience in design pedagogy.

Program Features: Students must complete eight Engineering graduate courses satisfying the following criteria:

- Core courses: AI 601 Introduction to Artificial Intelligence, AI 671 Design Capstone 1, AI 672 Design Capstone 2
- The eight-month capstone projects will include R&D / industry sponsored projects at Waterloo or a company
- At least one of the complementary AI 611 (The Ethics of AI) or AI 612 (AI and Society) courses
- At least three of the listed technical courses (to ensure depth and breadth)
- Interdisciplinarity and Industry-Relevance ensured via AI 601 and core design courses

Term	**1**	**2**	**3**	**4**			**Three courses, to be taken in any offered term, from any category**					
Core	**AI 601**	**AI 611** or **AI 612**	**AI 671**	**AI 672**		**Machine Intelligence Electives**	SYDE 522 Machine Intelligence	ECE 657 Tools of Intelligent Systems Design	SYDE 572 Introduction to Pattern Recognition	SYDE 675 Advanced Topics in Pattern Recognition	SYDE 556/750 Simulating Neurobiological Systems	SYDE 652 Deep Learning
Breadth	At least one of **AI 611: The Ethics of AI** **AI 612: AI and Society**					**Intelligent Systems Electives**	SYDE 575 Introduction to Image Processing	ME 640 Autonomous Mobile Robotics		AI 640 Computer Vision / Machine Perception		AI 630 Natural Language Processing
Electives	At least three electives, from those listed under **Machine Intelligence** **Intelligent Systems** **Data Analytics**					**Data Analytics Electives**	MSci 623 – Big Data Analytics		ECE 657A – Data and Knowledge Modelling and Analysis	Data 601 Fundamentals of Data Analytics	MSci 718 – Statistical Methods for Data Analytics	

Appendix C: Eight Month Applied Research Internship

Master of Science in Applied Computing (MScAC) Program

University of Toronto

Applied Research Industry Internships under the University of Toronto, Master of Science in Applied Computing (MScAC)

- The U of T Master of Science in Applied Computing (MScAC) program bridges the gap between academic research and firms wanting to deploy sophisticated technological methodologies.
- All graduates complete an eight (8) month internship project, following two semesters of graduate coursework.
- The internship project is proposed by the industry sponsor. The MScAC program reviews the nature of the project, and an academic supervisor provides the student with further support. The nature of the project could include:
    - A student-driven project investigating applied research of relevance to the organization.
    - A company-specific project involving applied research.
    - Evaluation of a new initiative for a specific company purpose.
- While the scope of an internship may involve coding or systems development that contributes to a company's product or services, intern sponsors should consider MScAC students for risk-tolerant projects that may fail or change course.
- During their internship, the student is mentored by both a company advisor and an academic advisor.
- At the conclusion of the internship, students prepare a project report that is approved by two readers (consistent with the requirements for an MSc degree in computer science).

Appendix D: Academic-Industry Partnership

Real-World Big Data Machine Learning Project as an Alternative to an Internship

University of Guelph

LSTM, RFM, LMFAO –Making Sense of Data Science Acronyms with a Deep Dive

POSTED IN DATA SCIENCE , LABS
BY ALEXANDER HINTON & CHRIS DULHANTY ON FEBRUARY 26, 2018

Hello, world! We are Alexander Hinton and Chris Dulhanty, two University of Guelph students who have recently gone down the machine learning rabbit hole through enrollment in Professor Graham Taylor's Introduction to Machine Learning class at the U of G last fall. The course provided us with a fantastic overview of the world of machine learning (ML) algorithms and techniques. In addition, we had the opportunity to get our hands dirty in a real-world machine learning problem via a unique academic-industry partnership, fostered between our professor and Rubikloud. The final project for our course was a retail prediction problem, made possible through access to an anonymized, proprietary dataset of transactional data, from a major health and beauty retailer, hosted by Rubikloud on the RubiOne platform. This blog documents our experience with the project: the research problem, our approach and results, and lessons learned along the way!

THE DATASET

Rubikloud provided a rich dataset of anonymized transactional data for one million customers enrolled in a major health and beauty retailer's loyalty program for the time period January 1st, 2015 to December 31st, 2016. Data on 18,198,302 customer purchases was provided; each entry in the dataset represented a single product purchased within an order from a customer, along with the date, quantity, price and contents of the transaction. Product data was provided with price, brand, and four levels of hierarchical categorization and customer data was provided with the date of registration in the retailer's loyalty program. Five top brands at the retailer were identified that represented 25.3% of all transactions for the one million customer cohort.

THE PROBLEM

The problem was defined as a customer-level purchase prediction problem, that is, to predict which customers were most likely to buy each of the five top brands in the 30-day period following the last date in the dataset. The output of our model would ultimately be a prediction on the scale of 0 to 1 for each customer, for each of the five top brands.

BASELINE MODEL

A baseline model was created by Rubikloud Chief Data Scientist Brian Keng and made available to the class. On November 9, 2017, Rubikloud Chief Engineer Adrian Petrescu made his way down the 401 to stop in at our class, provide a demo of the RubiOne platform and walk us through the baseline model. The model was based on the popular recency, frequency, monetary (RFM) method of assessing customer value and engagement. Customers were scored in these three areas by aggregating their purchasing habits over the two-year period, and were binned into groups from one to ten on each of these three characteristics. Predictions were then made by summing an individual customer's scores and dividing by 30.

EVALUATION

Due to the class imbalance present in the dataset, classification accuracy would not have been a suitable metric for the evaluation of predictions, as the naive example of predicting everyone to purchase nothing would result in accuracy in excess of 98%! Therefore, the area under the receiver operator characteristic curve (ROC) was selected by Rubikloud as the metric to compare models, the relative measure of true positives rate to false positives rate. While the baseline RFM model was very simple in nature, it provided a reasonably high mean ROC of 0.7240 across the five brands. Rubikloud's challenge for our class was to use ML techniques to construct a model which would provide a better mean ROC than their baseline model… which was no easy task, as the relative simplicity of their model exceeded the expectations of their Chief Data Scientist! Clearly, we would need a plan.

MAKING A PLAN

After some preliminary research into the field of marketing, it was apparent that the RFM features extracted in the baseline model are widely used by retailers to segment their customers. While the technique of binning customers, making predictions and using these predictions to make business decisions is common and effective in the marketing world, it does not have any inherent learning aspect to it, and thus was ripe for improvement with ML. With this in mind, our plan of action became clear – we would apply machine learning techniques to the RFM model, add a couple simply engineered features from the data, and create a retail prediction model that could be easily implemented at retailers with limited overhead in further data processing.

OUR MODELS

We compared three different ML models in our quest for the best: logistic regression (LR), a deep neural network (DNN) and a long-short term memory (LSTM) network. LR is the workhorse of classification algorithms in ML, relatively simple and very fast to implement. In our review of related literature, the algorithm provided strong results in similar problems, so we were interested if the KISS (keep it simple, stupid) methodology would be the way to go. DNNs extend LR to perform nonlinear calculations through a hierarchy of progressively rich layers of features. Having a model that could model nonlinear patterns was important, since we had the intuition that purchasing habits can be both predictable and sporadic. The LSTM network was our sexiest model, also the most specific to the problem at hand. An LSTM network is a type of Recurrent Neural Network (RNN) that uses outputs from previous time steps as inputs to future predictions – very prudent to our prediction of future purchase habits based on historical data.

FEATURE EXTRACTION

Four features were extracted from the transaction data for each of the five brands, for each customer in the time-period January 1st, 2015 to December 31st, 2016:

- Recency: number of days from the final day in the time-period to the last purchase of the customer
- Frequency: count of unique transactions a customer made in the time-period
- Monetary: sum of products purchased in the time-period
- Amount: count of total products a customer purchased in the time-period

One additional feature was also extracted for each customer – the number of days since registration, which we call Duration. Each customer therefore contained four features for each of the five brands and one additional feature for a total of 21. We refer to these features as RFMAD.

EXPERIMENTS AND RESULTS

The hyperparameters of our models were tuned using five-fold cross-validation. For the LR model, regularization strength was chosen via a grid search. For the DNN and LSTM, the learning rate, number of hidden units and number of hidden layers were selected by an initial coarse random search, followed by a finer Bayesian optimization using the SigOpt API. Many hours were spent tuning these hyperparameters. As the progress bar slowly inched forward, we contemplated the big questions of life… “Who am I”, “What is my purpose”, and most notably, “Can I afford a new GPU next semester?”.

In the LR and DNN models, features were calculated annually, for the LSTM, monthly. Features were fed into each model as inputs, with the target output being a multi-label prediction of the five brand-level, binary labels of purchases in the subsequent month.

The benchmark RFM model provided by Rubikloud achieved a mean ROC of 0.7240 across the five brands. Our RFMAD features, combined with the ML algorithms, were able to significantly improve on this performance. The optimized LR achieved a mean ROC of 0.7522, the DNN achieved a mean ROC of 0.7528, and the LSTM was the top performer, achieving a mean ROC of 0.7563. Assuming the distribution of ROC scores from the cross-validation procedure were approximately normal for each individual model, we conducted t-tests to determine if differences in mean ROC across models were significant. All models were found to be significantly better performers than the baseline RFM model, with the LSTM model being significantly better than the LR and NN at the 10% significance level.

Although the evidence was not striking, the LSTM model was the best classifier of the three. This was not surprising, given the temporal nature of the dataset and the proven success of LSTM models in many applications with temporal components. Retailers who are already tracking customer segmentation variables such as RFM can make demonstrable improvements in their predictive power by including machine learning models in their retail purchase forecasts. For optimal results, an LSTM network is recommended. However, given the large differences in total training time between the LSTM network and the LR classifier and only marginal improvements in performance, it would be very reasonable to incorporate a logistic regression classifier.

PARTING THOUGHTS

Having the opportunity to work on a real-world problem and dataset was an invaluable experience. In much of our earlier coursework, datasets were small, assignments more defined, and guidance was provided all along the way. The path taken on this project was much more of our own vision – we were provided a large dataset and a problem to approach, but the freedom to attack it however we saw fit. We experienced much more of the data science pipeline than we had in the classroom. We wrestled with Pandas, spent hours whiteboarding our ideas, tinkered with code, read man page after man page, and after hours of work thought “oh! But what if we included xyz” … and we did it all over again!

We are grateful to Rubikloud for access to their platform and their data, and allowing our class to have this experience. We hope it is just the first of many projects for us in the world of Data Science!