Oxford Machine Learning Summer School 2023

Report for the Master's Program

Author: Jauroyon Maxime

Abstract

This report provides an overview of my participation in the Oxford Machine Learning Summer School 2023, organized by AI for Global Goals in collaboration with CIFAR and the University of Oxford's Deep Medicine Program. The summer school focused on the advancements and practical applications of machine learning, offering modules that covered both fundamental principles and real-world case studies. In this report, I will present my experiences in the ML x Fundamentals and ML x Cases modules, highlighting the knowledge gained and insights obtained during the program.

Acknowledgments

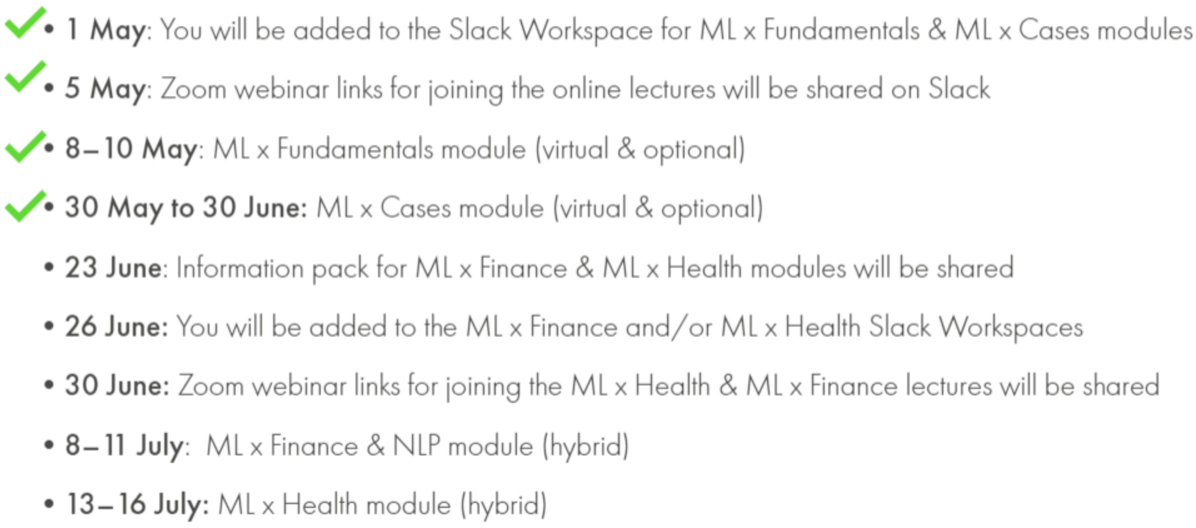
I would like to express my gratitude to the organizers of the Oxford Machine Learning Summer School 2023 for providing me with the opportunity to be part of this exceptional educational event. I am grateful to the speakers and experts who shared their knowledge and expertise, as well as to my fellow participants for the engaging discussions and collaborations.

Contents

1 Introduction

The Oxford Machine Learning Summer School 2023 was a transformative educational experience, offering participants the opportunity to delve into the cutting-edge advancements and practical applications of machine learning. This report aims to provide an overview of my participation in the summer school, highlighting the ML x Fundamentals and ML x Cases modules that I attended during the May-June session.

Selected among a competitive pool of applicants from 106 countries, I was thrilled to be part of this prestigious event organized by AI for Global Goals in collaboration with CIFAR and the University of Oxford's Deep Medicine Program. The summer school brought together a diverse group of participants, including students, postdocs, lecturers, and professionals, with a shared passion for advancing the field of artificial intelligence.



- ✓ **1 May**: You will be added to the Slack Workspace for ML x Fundamentals & ML x Cases modules
- ✓ **5 May**: Zoom webinar links for joining the online lectures will be shared on Slack
- ✓ **8–10 May**: ML x Fundamentals module (virtual & optional)
- ✓ **30 May to 30 June:** ML x Cases module (virtual & optional)
- **23 June**: Information pack for ML x Finance & ML x Health modules will be shared
- **26 June:** You will be added to the ML x Finance and/or ML x Health Slack Workspaces
- **30 June:** Zoom webinar links for joining the ML x Health & ML x Finance lectures will be shared
- **8–11 July**: ML x Finance & NLP module (hybrid)
- **13–16 July:** ML x Health module (hybrid)

The ML x Fundamentals module, held from 8th to 10th May, served as an essential foundation for the subsequent modules. Designed to equip participants with the necessary theoretical background, this module covered key areas such as optimization, statistical/probabilistic machine learning, deep learning, and more. As a participant, I had the opportunity to deepen my understanding of these fundamental concepts and develop a solid theoretical framework for modern machine learning.

Following the ML x Fundamentals module, I engaged in the ML x Cases track from 30th May to 30th June. This module focused on real-world machine learning problems and the practical aspects of ML development and implementation. Led by experienced ML and data scientists, the sessions provided insights into efficient data collection, enrichment, cleaning, labeling, and building ML models tailored to domain-specific tasks. Through hands-on exercises and case studies, I acquired valuable skills and knowledge that will contribute to my future endeavors in the field.

Throughout the summer school, I had the privilege of learning from esteemed speakers and experts in the machine learning domain. The program committee, comprised of accomplished individuals from academia and industry, curated an impressive lineup of speakers who shared their expertise and research findings. These interactions enriched my understanding of the latest trends, challenges, and breakthroughs in machine learning, inspiring me to explore new avenues and push the boundaries of my own capabilities.

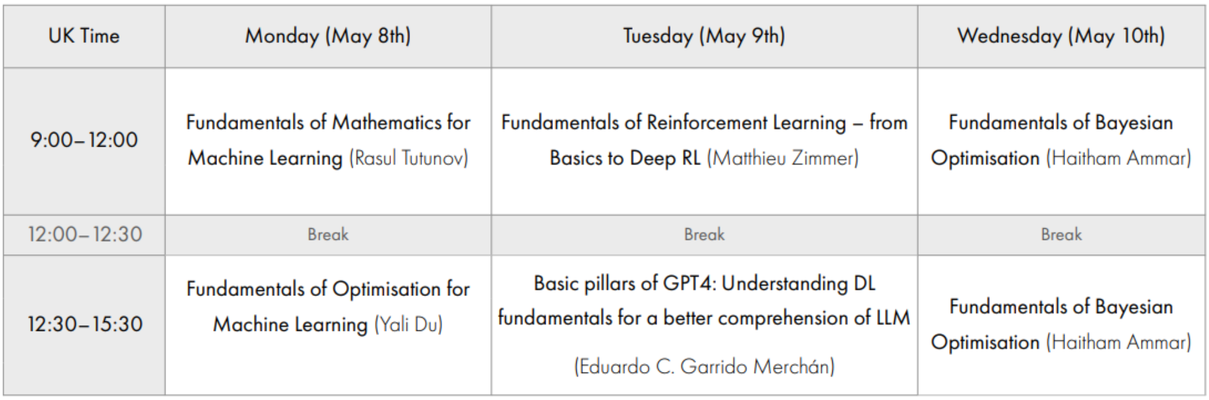
In this report, I will summarize the key takeaways from the Oxford Machine Learning Summer School 2023, reflecting on the knowledge gained, practical skills acquired, and the impact of this educational experience on my personal and professional growth. By examining the modules attended and discussing notable case studies, I aim to showcase the breadth and depth of the summer school's curriculum and its significance in the rapidly evolving field of machine learning.

2 ML x Fundamentals

2.1 Module Overview

The first module of the OxML summer school focused on providing an introduction to machine learning. This module served as the foundation for the subsequent topics covered throughout the program, ensuring participants had a solid understanding of the fundamental principles and techniques in machine learning.

2.1.1 Topics Covered



UK Time	Monday (May 8th)	Tuesday (May 9th)	Wednesday (May 10th)
9:00–12:00	Fundamentals of Mathematics for Machine Learning (Rasul Tutunov)	Fundamentals of Reinforcement Learning – from Basics to Deep RL (Matthieu Zimmer)	Fundamentals of Bayesian Optimisation (Haitham Ammar)
12:00–12:30	Break	Break	Break
12:30–15:30	Fundamentals of Optimisation for Machine Learning (Yali Du)	Basic pillars of GPT4: Understanding DL fundamentals for a better comprehension of LLM (Eduardo C. Garrido Merchán)	Fundamentals of Bayesian Optimisation (Haitham Ammar)

In this section of the report, we focus on a collection of courses that cover fundamental topics in machine learning and deep learning. These courses serve as building blocks for understanding the concepts and techniques discussed in the subsequent sections.

- The first course, 'Fundamentals of Mathematics for Machine Learning,' provides a comprehensive introduction to the mathematical foundations of machine learning. It covers essential topics such as linear algebra, calculus, and probability theory, which are crucial for understanding the underlying principles of various machine learning algorithms.

- The second course, ’Fundamentals of Optimization for Machine Learning,’ explores the optimization techniques used to train machine learning models effectively. It covers algorithms like gradient descent, stochastic gradient descent, and advanced optimization methods, equipping learners with the tools to optimize model performance.
- Moving on, the course ’Fundamentals of Reinforcement Learning from Basic to Deep RL’ delves into the exciting field of reinforcement learning. It introduces learners to the fundamental concepts, including Markov decision processes, value iteration, and policy gradients, ultimately leading to an understanding of deep reinforcement learning.
- The next course, ’Basic Pillars of GPT-4: Understanding DL Fundamentals for a Better Comprehension of LLM,’ focuses on deep learning fundamentals and their relevance to large language models (LLMs) like GPT-4. This course provides insights into the key concepts of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which form the foundation of LLMs.
- Lastly, the course ’Fundamentals of Bayesian Optimization’ covers the principles and applications of Bayesian optimization, a powerful technique for optimizing hyperparameters in machine learning models. Learners gain an understanding of Bayesian optimization algorithms and their practical use in improving model performance.

By summarizing the key topics covered in these courses, we aim to provide an overview of the foundational knowledge necessary for comprehending advanced concepts in machine learning and deep learning. The following summary section highlights the core principles and techniques discussed throughout these courses, providing a solid basis for the subsequent sections of this report.

2.2 Topics Overview

2.2.1 Fundamentals of Mathematics for Machine Learning

The course ”Fundamentals of Mathematics for Machine Learning” provides a comprehensive foundation in mathematical concepts and techniques essential for understanding and implementing machine learning algorithms. In this course, you will explore key topics such as linear algebra, probability theory, vector calculus, and their applications in machine learning. By gaining a solid understanding of these mathematical fundamentals, you will be equipped with the necessary tools to analyze data, develop models, and make informed decisions in the field of machine learning. The following is a summary of the main topics covered in the course, highlighting their relevance and significance in the context of machine learning.

- Linear algebra:

  Covers the basics of linear algebra, including vectors, matrices, linear independence, inner products, norms, and eigenvalues. These concepts are fundamental in machine learning for understanding data representation and transformation.

- Probability Theory:

  Introduces the fundamental concepts of probability theory, including probability distributions, random variables, and basic probability rules. Probability theory is essential for modeling uncertainty and making informed decisions in machine learning.

- Vector Calculus:

  Covers the basics of vector calculus, including derivatives, differentiation rules, partial derivatives, gradients, and the chain rule. Vector calculus is important for understanding the optimization algorithms used in machine learning, such as gradient descent.

- Loss functions in ML:

  Discusses loss functions used in machine learning, which quantify the discrepancy between predicted and actual values. Different types of loss functions are explored, and their properties and applications in various machine learning algorithms are discussed.

- Euclidean Spaces:

  Introduces Euclidean spaces, which are spaces where points are defined by coordinates, and distances between points are calculated using a distance formula. Covers coordinate systems, distance formulas, and the geometric interpretation of vectors in Euclidean spaces.

- Matrices:

  Explores matrices, including their definition, properties, transpose operation, linear operations, and matrix multiplication. Matrices play a crucial role in representing and transforming data in machine learning models.

- Linear Independence:

  Discusses linear independence of vectors and its importance in linear algebra. Covers the concept of linear combinations and how linear independence relates to the existence of unique solutions in systems of linear equations.

- Eigenvectors/Eigenvalues:

  Introduces eigenvectors and eigenvalues of matrices and their significance in linear transformations. Covers eigendecomposition and the diagonalization of matrices. Eigenvalues and eigenvectors have various applications in machine learning, such as dimensionality reduction techniques like Principal Component Analysis (PCA).

- Matrix Norms:

  Explores different matrix norms, including the Frobenius norm, maximum norm, infinity norm, and spectral norm. Matrix norms quantify the size or magnitude of matrices and are important in analyzing the convergence of iterative algorithms and measuring the error of approximations.

- Vector Calculus (continued):

  Covers further topics in vector calculus, including differentiation rules, Jacobian matrices, partial derivatives, gradients, and Taylor series expansions. These concepts are crucial for understanding optimization algorithms and the backpropagation algorithm used in neural networks.

These topics provide a solid foundation in the mathematical concepts necessary for understanding and applying machine learning algorithms.

2.2.2 Fundamentals of Optimisation for Machine Learning

The course on optimization methods for machine learning and deep learning models offers a comprehensive exploration of various techniques and algorithms aimed at minimizing loss functions and achieving optimal solutions. Throughout this course, participants are introduced to fundamental principles and strategies that enable efficient optimization. From traditional gradient descent approaches to cutting-edge adaptive methods, the curriculum covers a wide range of topics essential for practitioners in the field. Whether you're new to machine learning or an experienced professional, this course equips you with the knowledge and tools to effectively train and optimize your models, enabling you to unleash their true potential.The following is a summary of the main topics covered in "Optimization for Machine Learning":

- Gradient Descent:

  The course starts with the introduction of gradient descent, a fundamental optimization algorithm for finding the minimum of a function. It explains both the batch gradient descent and stochastic gradient descent methods and their differences.

- Stochastic Gradient Descent (SGD):

  The course dives deeper into stochastic gradient descent, which is widely used in machine learning. It explains how SGD works by randomly selecting subsets of data for each iteration and updates the parameters based on the gradients computed on those subsets.

- Convergence Rates:

  The course discusses convergence rates of optimization algorithms. It explores how the convergence rates differ for convex and non-convex functions and introduces concepts like Lipschitz convexity and smoothness.

- Non-Convex Optimization:

  The course delves into non-convex optimization problems, which are prevalent in deep learning and neural networks. It explores the challenges associated with non-convex optimization, such as the presence of local minima, saddle points, and the possibility of getting stuck in suboptimal solutions.

- Smooth Functions:

  The course introduces the concept of smooth functions and their properties. It explains how smoothness can be utilized to analyze the convergence behavior of optimization algorithms and provides theoretical bounds on the number of iterations required to reach a certain level of accuracy.

- Adaptive Methods:

  The course explores adaptive methods for optimization, focusing on the Adam optimizer. It discusses how adaptive methods adjust the learning rate dynamically during the optimization process and provides a theorem with convergence guarantees for Adam.

- Visualization of Loss Landscapes:

  The course briefly mentions a study on visualizing the loss landscapes of neural networks. It highlights the importance of network architecture and training parameters in shaping the loss landscape and influencing the generalization performance.

Overall, the course provides a comprehensive overview of optimization methods for machine learning, covering both theoretical aspects and practical techniques. It emphasizes the challenges and considerations specific to non-convex optimization, which is crucial in the context of deep learning and neural networks.

2.2.3 Fundamentals of Reinforcement Learning from Basic to Deep RL

The course on Markov Decision Processes (MDPs) and Reinforcement Learning provides a comprehensive study of decision-making in uncertain environments. This course explores the theoretical foundations and practical techniques for solving MDPs using reinforcement learning algorithms.Throughout the course, participants delve into the concepts and methodologies behind MDPs, which serve as a mathematical framework for modeling decision problems with uncertain outcomes. The course covers various solution methods for MDPs, including both model-based and model-free approaches.Participants gain hands-on experience by coding and experimenting with reinforcement learning algorithms using popular libraries.By summarizing the essential elements of the course, this report aims to provide a concise overview of the knowledge gained and the skills developed throughout the study of Markov Decision Processes and Reinforcement Learning.

- The course covers various aspects of solving Markov Decision Processes (MDPs), which are mathematical frameworks used to model decision-making in uncertain environments. The course explores different methods for solving MDPs, including both model-based and model-free approaches.

- Model-based methods involve having knowledge of the transition function (T) and reward function (R) of the environment. Value iteration and policy iteration are discussed as model-based algorithms for finding an optimal policy that maximizes a given criterion. These algorithms aim to find the optimal value function first and then derive the optimal policy from it.

- In contrast, model-free methods do not rely on having explicit knowledge of the transition and reward functions. Instead, they learn from experience through trial and error. Actor-only methods, critic-only methods, and actor-critic methods are covered as model-free approaches.

- Actor-only methods focus on learning a policy directly, while critic-only methods estimate value functions. Actor-critic methods combine the advantages of both actor-only and critic-only methods to learn both policy and value functions simultaneously.

- The course also covers exploration versus exploitation strategies in reinforcement learning, such as pure exploration, pure exploitation, and trade-off strategies. It discusses the use of neural networks to represent policies, including deterministic and stochastic policies.

- Various algorithms are presented for model-free learning, including Q-Learning, SARSA, and Deep Q-Networks (DQN). These algorithms estimate value functions or action-value functions to find an optimal policy.

  Additionally, the course introduces reinforcement learning techniques like REINFORCE and Proximal Policy Optimization (PPO). These algorithms utilize policy gradients to optimize the policy and reduce variance using baselines.

- The coding exercises in the course involve implementing and interacting with RL algorithms using libraries such as Gym and RLlib. There are also sections on training, monitoring, saving, loading policies, and evaluating their performance.

Overall, the course provides an overview of the fundamental concepts, algorithms, and practical aspects of solving MDPs using different approaches, with a focus on reinforcement learning and its applications.

2.2.4 Basic pillars of gpt4 : Understanding DL fundamentals for a better comprehension of LLM

The course begins by providing a solid foundation in deep learning and variational autoencoders, which form the basis of generative models. We then explore diffusion models, which allow us to recover meaningful data from randomness and offer improved training stability. Additionally, we delve into generative pretrained transformers, such as GPT-2 and GPT-3, which have achieved remarkable results in generating high-quality and contextually relevant outputs.By summarizing the key concepts and techniques covered in this course, we aim to provide a concise overview of the latest advancements in generative AI.

- Deep Learning:

  Deep learning is a field of machine learning that focuses on training neural networks with multiple layers to learn hierarchical representations of data. The course references the book ”Deep Learning” by Bengio, Goodfellow, and Courville as a foundational resource in this area.

- Variational Autoencoders:

  Variational autoencoders (VAEs) are generative models that combine an encoder network and a decoder network to learn latent representations of data. The course mentions a tutorial by Doersch as a resource for understanding VAEs in detail.

- Diffusion Models:

  Diffusion models involve a process of adding noise to data and progressively denoising it to recover the original data. Denoising score matching is used to learn the score function and estimate the gradient of the data distribution. The reverse diffusion process aims to reconstruct the original data from corrupted samples. Markov Chain Monte Carlo (MCMC) is used to sample from the data distribution. Diffusion models have applications in image synthesis, text generation, inpainting, and generative tasks.

- Generative Pretrained Transformers:

  Generative pretrained transformers are deep neural networks that have achieved significant scalability in terms of model size and capacity. They are based on the transformer architecture, which incorporates self-attention mechanisms, layer normalization, and positional encoding. Stacked transformer layers increase the model's capacity and depth. Masked causal attention enables autoregressive generation by attending only to previous tokens. These models are typically pretrained on large corpora and fine-tuned for specific tasks with smaller datasets.

  GPT-2, 3 and 4 are specific models within the generative pretrained transformers framework. GPT-2 has a large number of parameters and shows improved performance compared to its predecessor. GPT-3 further increases model size and capacity, introduces sparse attention mechanisms, and leverages in-context learning. GPT-3 demonstrates strong performance on a wide range of tasks without fine-tuning. GPT-4, a model with even more parameters, acknowledges the continuous evolution and growth of generative pretrained transformers.

- Generative AI Models and Applications:

The course provides a comprehensive review of generative AI models and their applications across various domains such as text, images, video, speech, music, code, business, video games, and the brain. It references a state-of-the-art review by Gozalo-Brizuela and Garrido-Merchan as a resource for further exploration.

2.2.5 Fundamentals of Bayesian Optimisation

The lecture on the topic of ”Bayesian Optimization: Gaussian Processes & Beyond” focuses on the concept of Bayesian Optimization (BO), which is a data-driven optimization approach used in various applications such as molecule design, material design, power plant tuning, data center cooling, hyper-parameter tuning, NAS (Neural Architecture Search), and AutoML (Automated Machine Learning). Bayesian Optimization aims to maximize a black-box function that is not known analytically and is expensive to evaluate. It involves a sequential data-driven scenario where an agent interacts with a black-box function and uses the acquired knowledge to improve the guess of the optimum while being data-efficient.

- The presentation discusses the two main steps of Bayesian Optimization: Step I - Learning Bayesian Models, and Step II - Maximizing acquisition.

- In Step I, Gaussian Processes (GPs) are used as a model to represent the unknown function. The presentation explains the basics of Gaussian Processes and their properties.

- Step II involves finding new probes or inputs to evaluate the black-box function. Various acquisition functions are discussed, such as Expected Improvement, Probability of Improvement, Simple Regret, Upper Confidence Bound, Knowledge Gradients, and Entropy Search. These acquisition functions trade-off exploration and exploitation to find the best next input to evaluate.

- The presentation also introduces a Bayesian Optimization library called HEBO (Heteroscedastic and Evolutionary Bayesian Optimization), developed by Huawei Noah’s Ark. HEBO is described as a flexible and easy-to-use library that supports parallel and high-dimensional Bayesian Optimization, mixed-variable optimization, multi-objective optimization, and interfaces with other libraries such as GPyTorch, GPy, GPFlow, and Pymoo.

- Furthermore, the presentation discusses the challenges of applying Bayesian Optimization to problems with discrete or mixed variable types. It introduces specific kernels, such as overlap kernels and substring kernels, that can be used for modeling similarity in such cases. It also mentions the challenge of acquisition function maximization in combinatorial optimization problems and suggests leveraging local search algorithms within trust regions.

Overall, the presentation provides an overview of Bayesian Optimization, Gaussian Processes, and their applications in various real-world scenarios. It highlights the HEBO library and its capabilities for efficient and flexible Bayesian Optimization.

2.3 Personal Reflections

The first module of the OxML summer school was an enriching experience that not only laid the foundation for our journey into the world of machine learning but also consolidated and expanded upon the knowledge acquired during my M1 studies.

Through this module, we were introduced to a comprehensive range of key concepts, techniques, and applications in machine learning. The interactive workshops, guest lectures, and collaborative learning activities created an immersive learning environment that allowed us to actively engage with the material and gain practical insights.

Moreover, the module went beyond the scope of my M1 studies by including courses on reinforcement learning and Bayesian optimization, which were not covered in my previous coursework. These courses provided valuable new perspectives and expanded my understanding of these advanced topics, equipping me with additional tools and techniques for addressing complex machine learning challenges.

In the next section, we will explore the second module of the summer school, focusing on hands-on challenges.

3 ML x Cases

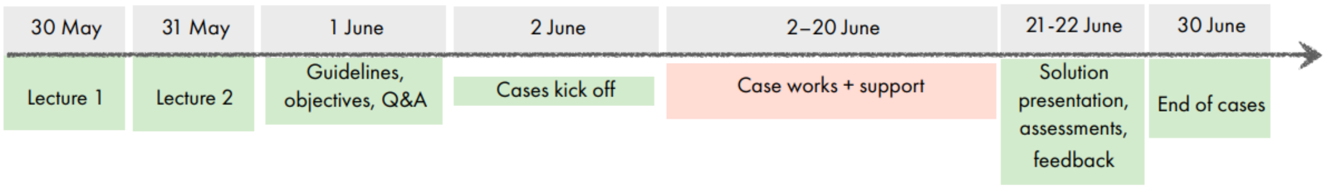
3.1 Module Overview

The second module of the OxML summer school focused on providing participants with a hands-on learning experience through two challenging projects in the domains of finance and health. This module aimed to enhance our practical skills in applying machine learning techniques to real-world problems.

During this module, we were presented with two distinct challenges, one in the finance domain and the other in the health domain. These challenges were designed to simulate real-world scenarios and required us to analyze and solve complex problems using machine learning methodologies.

I personally chose to tackle the finance challenge, as it was aligned with my choice for the Finance x NLP module. The finance challenge was a two-task challenge and i will delve into the details of the tasks and discuss the approaches and techniques employed to overcome the challenges presented in the next section.

3.1.1 Topics Covered



30 May	31 May	1 June	2 June	2–20 June	21-22 June	30 June
Lecture 1	Lecture 2	Guidelines, objectives, Q&A	Cases kick off	Case works + support	Solution presentation, assessments, feedback	End of cases

- Introduction to the challenges in finance and health: The module began by providing an overview of the specific challenges in the domains of finance and health, highlighting their relevance and impact.
- Selection of challenge: Participants were given the opportunity to choose one of the two challenges based on their interests and career aspirations. This personalized approach allowed them to focus on a domain that aligned with their professional goals.
- Hands-on experience: The module offered a hands-on learning experience by providing participants with real-world challenges in finance and health. It required applying machine learning techniques to develop practical solutions.

3.2 Work Overview

3.2.1 Task 1

In the realm of natural language processing and text classification, the identification and categorization of Environmental, Social, and Governance (ESG) documents hold great importance. ESG factors are critical indicators of a company's sustainability and ethical practices, making the classification of documents based on these factors a significant challenge.

Our primary objective was to develop an effective and accurate classifier that could discern and categorize ESG documents based on their content. This involved training models to understand and extract key information from textual data, enabling them to make informed predictions about the ESG category to which a document belongs.

To evaluate the performance of our models, we employed the F1 score, a widely used metric in text classification tasks. The F1 score considers both precision and recall, providing a balanced assessment of the model's ability to correctly classify ESG documents across different categories. Through this metric, we could gauge the accuracy and effectiveness of our classifiers and make informed comparisons between different approaches.

Now, let us delve into our journey of tackling this challenge, exploring our experimental setups, methodologies, and key findings in the realm of ESG document classification.

- We embarked on a comprehensive exploration of various machine learning and deep learning models to tackle the ESG classification problem. Our goal was to identify the most effective approach for accurately categorizing documents into the four ESG classes: governance, social, environmental, and other.

- Initially, we experimented with various traditional machine learning models, including logistic regression, support vector machines (SVM), random forests, and naive Bayes classifiers. For the finance challenge in ESG classification, we explored the effectiveness of these models in combination with Tf-idf vectorization and CountVectorizer for text representation. Among these models, logistic regression using Tf-idf vectorization stood out, achieving the highest f1-score of 77.777%. However, the other models also showcased promising results, demonstrating their potential for the ESG classification task.

- To delve deeper into the realm of deep learning, we explored models using Word2Vec and Glove embeddings with recurrent neural networks (RNNs). However, these attempts did not yield satisfactory results, and we encountered challenges in capturing the nuanced patterns present in ESG-related text. Despite careful tuning of model architectures and hyperparameters, the performance of these models fell short, leading us to explore other avenues.

- Motivated by the success of pre-trained language models, we attempted to fine-tune popular models such as BERT (base and large) and DistilBERT. Our initial attempts using the run_glue.py script did not surpass the baseline performance. However, we persevered and adopted the Trainer API from Hugging Face to fine-tune the DistilBERT model, which yielded more promising results. With the Trainer, we achieved an impressive f1-score of 80.392%, highlighting the value of leveraging pre-trained models and the effectiveness of the fine-tuning process in ESG classification tasks.

- To further evaluate the performance of the fine-tuned DistilBERT model, we conducted experiments with a more robust train/test split. Surprisingly, the model's performance decreased very slightly, resulting in an f1-score of 79.392%. While this outcome was unexpected, it emphasized the importance of carefully considering the data partitioning strategy and the impact it can have on model performance.

- In our pursuit of even higher performance, we explored a domain-specific model, namely the ”yiyanghkust/finbert-esg” model from Hugging Face. This model had been fine-tuned on 2,000 manually annotated sentences from firms' ESG reports and annual reports. Initial results without any further fine-tuning showcased its potential, as it achieved an f1-score of 72%. However, limitations arose due to the model's input size, as it was restricted to only 512 characters. Consequently, the model's predictive capabilities might have been hampered, hindering its performance on our dataset.
- In summary, our comprehensive investigation encompassed a wide range of models and techniques for ESG classification. Through careful experimentation and fine-tuning, we identified the most promising approach, which involved leveraging the DistilBERT model using the Hugging Face Trainer. This configuration achieved an impressive f1-score of 80.392%. Our findings highlighted the importance of pre-trained models and the benefits of leveraging their capabilities for text classification tasks. Furthermore, we emphasized the significance of thoughtful data preprocessing and the impact it can have on model performance.

For this first task, I achieved a ranking of 18th out of 36 participating teams on the public leaderboard. However, when considering the final standings based on the private leaderboard, which accounts for approximately 68% of the test data, I secured a ranking of 7th out of 36 participating teams. This demonstrates the significant improvement and competitiveness of our ESG document classifier when evaluated on the larger portion of the test data.

3.2.2 Task 2

In the ever-evolving field of document analysis and understanding, one of the fundamental tasks is table detection. Tables play a crucial role in organizing and presenting structured information, making their accurate detection a significant challenge.

Our primary objective was to investigate the impact of data size on the performance of table detection models. Data size has long been recognized as a critical factor in machine learning, influencing the ability of models to generalize and make accurate predictions. We aimed to unravel the relationship between data size and model performance by experimenting with pre-trained models trained on different quantities of annotated table images. This exploration allowed us to gain valuable insights into the scalability and effectiveness of models as the size of the training data varied.

To evaluate our models and quantify their performance, we employed the Mean Columnwise root mean squared error (MCRMSE) metric, which is a widely used metric for evaluating regression tasks. This metric measures the average root mean squared error across all columns of the predicted table compared to the ground truth table. By leveraging this metric, we could assess the effectiveness of our models and make informed comparisons between different approaches, ensuring the accuracy and reliability of our table detection.

Now, let's delve into our exploration of this challenge and delve into the details of our experiments, methodologies, and key findings.

- We aimed to develop a model that could accurately detect tables in documents. We conducted several experiments using different pre-trained models, evaluating their performance based on a test dataset. The main objective was to investigate the impact of data size on model performance.

- We started by exploring various pre-trained models from Hugging Face's model hub. Among them, the ”Benito/DeTr-TableDetection-5000-images” model showed the best performance, achieving a score of 0.0259. This model had been pre-trained on a large dataset of 5000 images, indicating the positive impact of data size on model effectiveness.

- Next, we attempted to finetune the ”facebook/detr-resnet-50” model with our own dataset of 500 images. Unfortunately, this finetuned model performed poorly, failing to detect any tables in the training data.

- To further investigate the impact of data size, we then directly finetuned the ”Benito/DeTr-TableDetection-5000-images” model with our dataset. However, this approach yielded a lower score of 0.10307 compared to the model without finetuning. This outcome suggests that the finetuning process and the characteristics of the additional data may have affected the model's performance negatively.

- Finally, we experimented with the ”Benito/DeTr-TableDetection-3000-images” model, which was the same as the best-performing model but finetuned on a smaller dataset of 3000 images. This model achieved a score of 0.16729, which was noticeably worse than both the original model and the finetuned version with the larger dataset. This result emphasizes the importance of data size and demonstrates the negative impact of reducing the finetuning dataset.

- In conclusion, our analysis highlights the significance of data size in table detection models. The ”Benito/DeTr-TableDetection-5000-images” model, pretrained on a larger dataset, proved to be the most effective. Finetuning with a smaller dataset or using a different base model led to inferior performance. These findings emphasize the importance of having a sufficient amount of high-quality training data for optimal model performance in table detection tasks.

For this second task, I obtained a remarkable ranking of 2nd out of 12 participating teams on the public leaderboard. When looking at the private leaderboard, we maintained a strong performance and secured a ranking of 3rd out of 12 participating teams. This further emphasizes the reliability and effectiveness of the pre-trained table detection model.

3.3 Personal Reflections

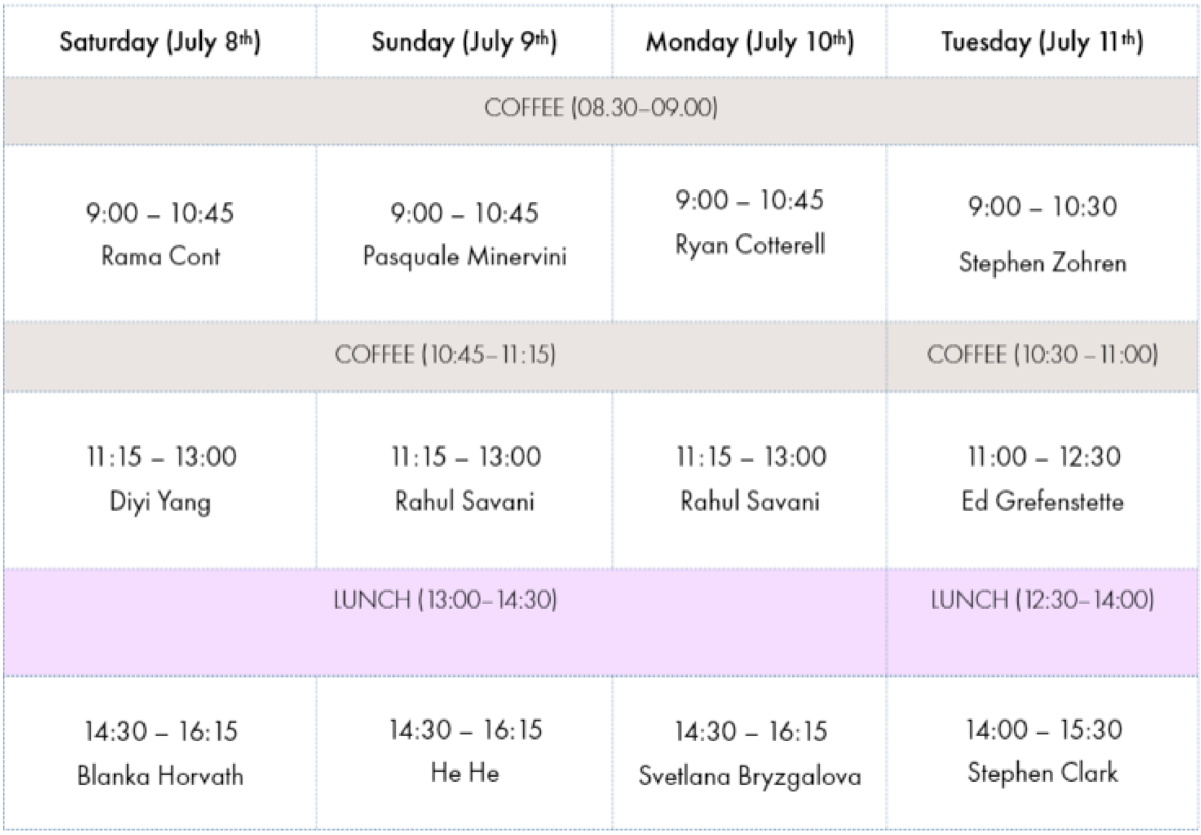
The MLX cases module was an invaluable experience that allowed us to apply our theoretical knowledge to real-world machine learning problems. Through hands-on projects, interactive workshops, and engaging guest lectures, we gained practical insights into the challenges and intricacies of working with MLX cases.

Moreover, the significance of the private leaderboard rankings cannot be understated. As they reflect the final standings based on a larger portion of the test data, they provide a more accurate assessment of our solutions' performance in real-world scenarios. Consistently achieving a commendable ranking on the private leaderboard further reinforces the robustness and practicality of our approaches.

Overall, the MLX cases module was a pivotal part of the OxML summer school, bridging the gap between theory and practice. It empowered us to apply our machine learning expertise, gain practical insights, and develop the skills necessary to address complex real-world challenges. This module undoubtedly enriched our learning journey and prepared us for future endeavors in the field of machine learning.

4 Conclusion

I am excited to be participating in the MLx Finance and NLP module as part of the summer school organized by AI for Global. This module will provide a comprehensive exploration of the intersection between machine learning, finance, and natural language processing (NLP). Although specific details about the content and topics covered in this module are not available at the moment, I am eagerly looking forward to engaging with leading experts in the field and gaining valuable insights into how machine learning techniques can be applied to finance and NLP applications.



Saturday (July 8th)	Sunday (July 9th)	Monday (July 10th)	Tuesday (July 11th)
COFFEE (08.30–09.00)			
9:00 – 10:45 Rama Cont	9:00 – 10:45 Pasquale Minervini	9:00 – 10:45 Ryan Cotterell	9:00 – 10:30 Stephen Zohren
COFFEE (10:45–11:15)			COFFEE (10:30 – 11:00)
11:15 – 13:00 Diyi Yang	11:15 – 13:00 Rahul Savani	11:15 – 13:00 Rahul Savani	11:00 – 12:30 Ed Grefenstette
LUNCH (13:00–14:30)			LUNCH (12:30–14:00)
14:30 – 16:15 Blanka Horvath	14:30 – 16:15 He He	14:30 – 16:15 Svetlana Bryzgalova	14:00 – 15:30 Stephen Clark

I am confident that this module will provide me with valuable insights and practical knowledge in the field of MLx Finance and NLP. I am eagerly anticipating the opportunity to learn from experts, engage in discussions, and explore the latest advancements in this exciting field.

Until now the OxML summer school has been an enriching and rewarding experience, providing participants with a comprehensive understanding of machine learning and its practical applications. Through a series of modules, workshops, and projects, we have explored fundamental concepts, advanced techniques, and real-world case studies, equipping us with valuable knowledge and skills in this rapidly evolving field.

Throughout the summer school, we had the opportunity to learn from renowned experts and researchers in the machine learning community. Their lectures, presentations, and interactions have broadened our perspectives and deepened our understanding of the diverse applications and challenges in machine learning.

The hands-on projects and collaborative activities have been instrumental in solidifying our theoretical knowledge and sharpening our practical skills. By working on MLX cases, we gained valuable experience in problem formulation, data preprocessing, feature engineering, and model evaluation. The interactive workshops and peer discussions further enhanced our learning process, allowing us to exchange ideas, seek feedback, and explore innovative approaches.

Moreover, the summer school fostered a vibrant and inclusive learning environment. Interacting with fellow participants from diverse backgrounds and cultures enriched our learning experience and broadened our horizons. The discussions, debates, and networking opportunities enabled us to build connections and forge collaborations that may extend beyond the duration of the program.

As we conclude our journey at the OxML summer school, we carry with us a solid foundation in machine learning principles and a passion to continue exploring and contributing to this exciting field. The knowledge, skills, and experiences gained during this program will undoubtedly shape our future endeavors and empower us to make meaningful contributions in academia, industry, and beyond.

We extend our heartfelt gratitude to the organizers, instructors, guest speakers, and fellow participants for their valuable contributions and unwavering support throughout the summer school. Their dedication and commitment have made this experience truly remarkable and unforgettable.

In closing, we look forward to applying our newfound knowledge and skills in machine learning to tackle real-world challenges, pushing the boundaries of innovation and making a positive impact in the world.