Course title: **Applied Machine Learning and Data Engineering in Business Context**
[KAN-CDSCV1006U]

Semester: Fall 2021

Course responsible: Raghava Rao Mukkamala

Teacher(s):
Samay Mir
Somnath Mazumdar
Raghava Rao Mukkamala

Course description:
This course aims at making some of the complex mathematical, cloud and data science concepts tangible to the business audience. The primary focus of the course content is to explain data mining, machine learning theories and cloud & data engineering concepts in an animated and business-friendly fashion where students are motivated to understand deeply the assumptions of the models and evaluate their applicability for a given business context.

The course provides hands-on experience on how to use machine-learning methods for solving real-world problems in an organizational context using suitable cloud technologies and business communication practices. It simulates the real-world processes that are experienced by the data scientists in the companies to provide a hands-on experience to the students on data-driven decision-making in an organizational setting. This course is ideal for the students who have a strong technical background in machine-learning and looking forward to enriching their skills on data engineering, cloud technologies, and business communication to have a smooth transition into the data scientist/data engineer careers at a later point of time. This is course is offered in collaboration with Capgemini and other Danish companies and therefore, the students will have an opportunity to interact with domain experts from various industry sectors such as Finance, Marketing, Supply chain and Energy.

The course is structured in three parts, providing the students with a full overview of methods, techniques, and practices that are currently used in the industry with respect to data-driven decision-making in an organization setting as follows.

1. Machine Learning in Business
- Applied linear regression
- Applied PCA
- Applied predictive and classification algorithms such as SVM, Random Forrest and XGBoost
- Applied deep learning algorithms in particular to LSTM

2.Data Engineering and Cloud Technologies
(using Amazon Web Services and Microsoft Azure cloud platforms)
- Data management including compliance
- Ingestion process
- Storage
- Extract, Transform, Load (ETL) end-to-end processes
- ML Computation
- Visualization

3.Business Communication Practices

- How to introduce complex business questions
- How to design approach, assumption and apply machine learning concepts in a business-friendly manner
- How to design PowerPoint slides
- How to clearly demonstrate the business benefits

Learning activities:

The learning objectives can be found on the course homepage:
https://kursuskatalog.cbs.dk/2021-2022/KAN-CDSCV1006U.aspx

Course plan – Fall 2021

Teacher:

Samay Mir (SMI)
Somnath Mazumdar (SMA)
Raghava Rao Mukkamala (RRM)

Week	Lecture	Exercises	Learning objectives	Readings
Week 36 07/10/2021 14:25-16:05 (Room: ks54) 16:15 – 17:55 (Room: SP113)	Course Introduction and Practicalities Intro to DevOps/GitHub + Guest Lecture	Databricks setup and GitLHub	#6	Slides, Articles , Book Chapter 3,4,5 of [Book#1].
Week 37 14/10/2021 14:25-17:55 Room: FC_C1	Module: ML in Business Context Foundations of Regression - Intuition - Optimization using Gradient Decent and Linear Equations - Time Complexity and Big *O* - how do you communicate liner regression in business context?	Databricks setup and examples Linear regression in Databricks	#1, #3, #6	Book Chapter 4 of [Book#2].
Week 38 21/09/2021 14:25-16:05 (Room: ks54) 16:15 – 17:55 (Room: SP113)	Module: ML in Business Context Dimensionality Reduction - Intuition - Optimization - Time Complexity and Big *O*	Dimensionality reduction exercises using SVD/PCA in Databricks	#1, #3, #6	Book Chapter 8 of [Book#2].

	- SVD, PCA - Interpretation of Dimensionality Reduction in Business Context			
Week 39 28/09/2021 14:25-16:05 (Room: ks54) 16:15 – 17:55 (Room: SP113)	Module: ML in Business Context Random Forests, XGBoost and Neural Networks - Intuition - Optimization - Time Complexity and Big O Intuition behind Error Back Propagation in Neural Networks	Exercises in Databricks on Random Forests, XGBoost Tensor Flow /PyTorch in Google Colab	#1, #3, #6	Book Chapter from 6,7 and 10 of [Book#2]
Week 40 05/10/2021 14:25-16:05 (Room: ks54) 16:15 – 17:55 (Room: SP113)	Module: Data Engineering and Cloud Data Management Strategies, Data Architecture Data Validation and Compliance How to build end-to-end data processes?	Guest talk from Industry expert on Data Compliance and GDPR	#2, #3, #5	[J-1]
Week 40 08/10/2021 12:35 – 17:00 (Room: PH_Ovnhallen - mobil auditoriet)	Workshop – 1 Workshop on Data validation and deployment of Production ML pipelines using TensorFlow Extended (TFX) pipelines		#2, #5	
Week 41 12/10/2021 14:25-17:55 Room: FC_C1	Module: Data Engineering and Cloud Extract Transform Load (ETL) Processes -ETL Concepts -Migration to cloud -Build from source to target -Rewrite the logic in Databricks	Demo of migration logic from on-premises to cloud using Azure Data Factory and Data Bricks	#2, #3, #5	[Book#3, J-2]
Week 43 28/10/2021 13:30 – 17:00 Room: DH.Ø.0.41	Module: Data Engineering and Cloud DevOps -DevOps Concepts -Unit testing -Continuous Integration and Continuous Delivery (CI/CD) -Branching Strategies	Exercise on Azure DevOps / GitLab and Databricks	#2, #3, #5	[J-3]
Week 43 – Workshop 29/10/2021 12:35 – 17:00 (Room: PH_Ovnhallen - mobil auditoriet)	Workshop – 2 Workshop on ETL processes and pipelines in cloud environment		#2, #5	

Week 44 02/11/2021 08:00 – 11:30 Room: DH.Ø.0.41	Module: Business Communication Concept of management consulting -why, how & what Business Communication Frameworks	How to conduct a workshop with business communication? How to design PowerPoint slides for business communication?	#3, #4, #6	
Week 44 – Workshop 05/11/2021 12:35 – 17:00 (Room: PH_Ovnhallen - mobil auditoriet)	Workshop – 3 Workshop om sliding as a management consultant - Building Storyline - Framework - Presenting Data Architecture		#2, #3, #4	
Week 45 09/11/2021 08:30-11:30 Room: DH.Ø.0.41	Module: Business Communication How to create an end-to-end proposal? Recap with a use case	Presentation of end-to-end ML pipeline using a use case	#3, #4, #6	
Week 45 – Workshop 12/11/2021 12:35 – 17:00 (Room: PH_Ovnhallen - mobil auditoriet)	Workshop – 4 Plan in a group on how to conduct a brain-storming session/workshop to find relevant use cases for a given company Develop a 10-slide executive PowerPoint presentation indicating: - the company's current situation/complication and key question - Your approach on how to solve the key question - ML approach - Data engineering architecture Your deliverables i.e., what will the outcome of your approach be in terms of revenue or optimization A roll-out plan of your approach		#2, #3, #4, #5,	
Week 46 16/11/2021 08:30-11:30 Room: DH.Ø.0.41	Exam discussion and Q & A session - Students Presentations		#3, #4, #6	

Compulsory assignments - Applied Machine Learning and Data Engineering in Business Context				
	Release Date	**Submission date**	**Teacher assigned**	**Platform – Canvas or DE?**
Assignment 1	2021-09-29	2021-10-10	SMA	Canvas
Assignment 2	2021-10-29	2021-11-07	SMI	Canvas
Assignment 3	2021-11-10	17-11-2021	SMI	Canvas

Literature

	Authors(s)	Title	Publisher/ ISBN/ DOI
[Book 1]	Robert Ilijason	Beginning Apache Spark Using Azure Databricks: Unleashing Large Cluster Analytics in the Cloud	Springer Nature Switzerland AG. https://link.springer.com/book/10.1007/978-1-4842-5781-4
[Book 2]	Aurélien Géron	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems	O'Reilly
[Book 3]	Ralph Kimball and Joe Caserta	The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data	Wiley Publishing, Inc.

Journal papers	
[J-1]	Mazumdar, S., Seybold, D., Kritikos, K., & Verginadis, Y. (2019). A survey on data storage and placement methodologies for cloud-big data ecosystem. *Journal of Big Data*, *6*(1), 1-37.
[J-2]	Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, *5*(3), 1-27.
[J-3]	Leite, L., Rocha, C., Kon, F., Milojicic, D., & Meirelles, P. (2019). A survey of DevOps concepts and challenges. *ACM Computing Surveys (CSUR)*, *52*(6), 1-35.

Slides will be uploaded before the lecture.