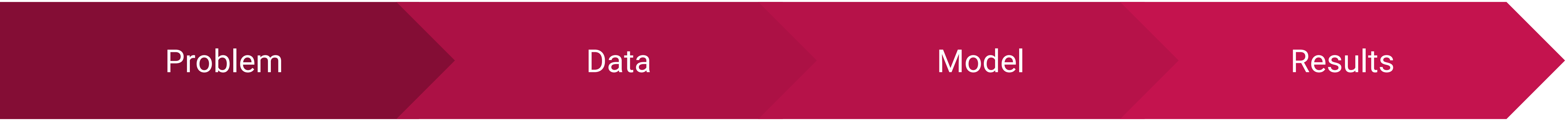
Caregivers & Machine Learning 2023

Capstone Project Outline

April 10, 2023

VECTOR
INSTITUTE

Capstone Project Goal



Problem

- Select a dataset
- Define the problem
- Determine goal of ML model

Data

- Explore data
- Visualizations
- Data cleaning and organizing
- Features
- Label

Model

- Select ML algorithms (2 or more)
- Train models
- Compare results
- Select best model
- Choose appropriate performance metrics

Results

- Interpret results
- Practical implications of your findings
- Discuss limitations and future work

Capstone Project

Dataset options

- We are going to provide some dataset options
- Individuals with “interesting” datasets personal/offices/workplaces are encouraged to use them, please contact Flora for approval

Teams

- You may choose to work individually or in teams of 2

Image credit- Unsplash.com

Capstone Project Structure

- Executive summary
- Introduction
- Problem definition
    - *Brief on dataset*
    - *Problem you are trying to solve*
    - *Proposed model and approach*
- Data exploration and description
- Model
    - *Describe your model and algorithm*
- Results and findings
- Conclusions and future work
- References

Capstone Project Structure

Section	Description
Executive Summary	Brief overview of the capstone project in not more than 250 words.
Introduction	Discuss why you are carrying out the project and talk about the type of problem you are trying to solve. Briefly discuss previous work in this area (if applicable). What is your hypothesis?
Problem definition	Tell us more about the problem you are trying to solve and discuss why it is a challenge. Discuss what type of data you are using. Discuss your approach, model options, the model you chose, and why you selected it.
Model	Describe the details of your modeling and algorithm.
Results and findings	Present your results and interpret the outcome of your modeling. You should have answers to the questions you had in your introduction here (if possible). Talk about practical applications and implications of your findings.
Conclusions and future work	Make inference from your project and discuss limitations and strengths of your work. Discuss future opportunities.
References	Remember to provide references

Capstone Report

- 4-6 pages long (excluding references), any format you choose (LaTex, Word, etc.)
- If you are providing your data – please do not include sensitive content or trade secrets
- Please submit **Capstone Report** and **Python Notebook** to course portal by **Friday, May 5 @ 11:59pm**

Capstone Grading Scheme

Code	
Clean and readable code (modular, comments)	10%
Reproducibility and documentation	10%
Performance (i.e., accuracy, error, etc.)	10%
Total	**30%**
Report	
Formatting and structure	10%
Problem description and exploratory data analysis	15%
Methodology	20%
Results and conclusions	15%
Appropriate use of figures, graphs, tables, and references	10%
Total	**70%**

Capstone Datasets

	Dataset	Industry	Possible prediction model	Problem type	Link to dataset
1.	Credit Card Fraud Detection	Financial	Predict if a credit card transaction is fraudulent	Binary classification on imbalanced data	https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
2.	Auto Insurance	Insurance	Predict the auto insurance claim amount	Regression	https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data
3.	Diabetic Patients' Re-admission	Medical	Predict if a diabetic patient was readmitted to the hospital	Multi-class classification	https://www.kaggle.com/datasets/saurabhtayal/diabetic-patients-readmission-prediction
4.	Telecom customer churn	Telecommunications	Predict if a customer will churn	Binary classification or clustering	https://www.kaggle.com/datasets/abhinav89/telecom-customer
5.	Cyberbullying	Social Media	Predict the type of cyberbullying in a tweet	Multi-class text classification	https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification
6.	Electric Motor Temperature	Manufacturing	Predict electric motor components' temperature	Timeseries regression	https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature

Credit Card Fraud Detection

- The dataset contains transactions made by credit cards in September 2013 by European cardholders.
- This dataset presents transactions that occurred in two days, containing 492 frauds out of 284,807 transactions.
- The dataset is imbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numerical input variables which are the result of a PCA transformation.

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Credit Card Fraud Detection

- **Input**: time, amount, 28 features which are the result of a PCA transformation
- **Output**: Fraud class (0 or 1).
- **Suggested Task**: Train the model using the input features to predict the fraud class (0 or 1).
- **Possible models**: Standard classification models (e.g., logistic regression, decision trees, neural networks).

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Auto Insurance dataset

- Dataset contains 9134 claims which include the information about the customer, the vehicle, the insurance policy, and the total claim amount.

- Data contains both categorical and numerical variables.

https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data

Auto Insurance dataset

- **Input**: 23 attributes such as State, Coverage, Education, Gender, Employment, Vehicle class, etc.
- **Output**: Total claim amount.
- **Suggested Task**: Train the model using the input features to predict the total claim amount.
- **Possible models**: Standard regression models (e.g., linear regression, decision trees, neural networks).

https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data

Telecom customer churn

- The dataset consists of 100 variables (attributes) and approximately 100,000 records.
- This dataset contains both numerical and categorical variables.
- Various factors are considered important while dealing with customers of telecom industry (e.g., monthly minutes of use and recurring charges).

https://www.kaggle.com/datasets/abhinav89/telecom-customer

Telecom customer churn

- **Input**: various information about customers of telecom industry.
- **Output**: whether the customer will churn or not.
- **Suggested Task:** develop a model to predict the customers who would churn.
- **Possible Models:** Standard classification models (e.g., logistic regression, decision trees, neural networks).

https://www.kaggle.com/datasets/abhinav89/telecom-customer

Diabetic Patients' Re-admission

- The dataset consists of 49 variables (attributes) and approximately 100,000 records.
- The dataset contains information about diabetic patients in the US from 1999 to 2008.
- This dataset contains both numerical and categorical variables.

https://www.kaggle.com/datasets/saurabhtayal/diabetic-patients-readmission-prediction

Diabetic Patients' Re-admission

- **Attributes**: Various demographic and clinical information of a patient.
- **Output**: whether the patient is readmitted in less than 30 days, more than 30 days, or not readmitted.
- **Suggested Task:** Develop a model to predict the high-risk diabetic-patients who are most likely to get readmitted within 30 days.
- **Models that can be used:** standard classification models (e.g., logistic regression).

https://www.kaggle.com/datasets/saurabhtayal/diabetic-patients-readmission-prediction

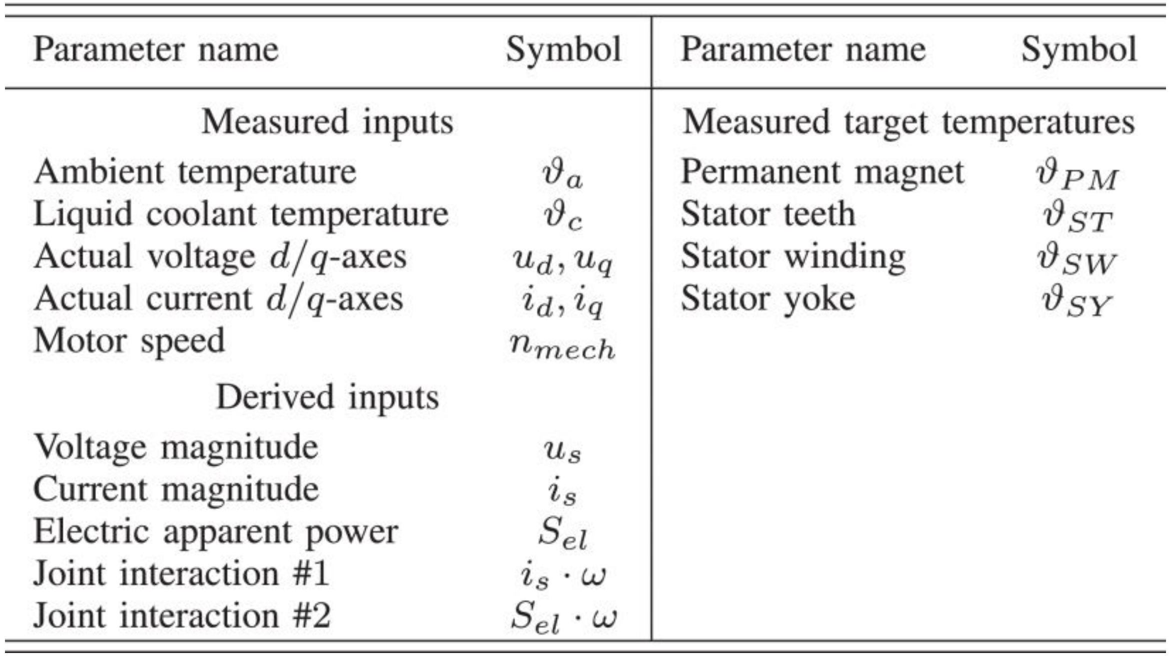
Electric Motor Temperature

- The dataset comprises several sensor data collected from a permanent magnet synchronous motor (PMSM) deployed on a test bench.

- Dataset is in csv format and each row represents all measurement sessions and features sampled at 2 Hz frequency.

- This dataset contains timeseries data and requires special handling of the inputs to insure proper model learning.

- This task is more challenging and only recommended for groups who want to explore timeseries methods.



Parameter name	Symbol	Parameter name	Symbol
Measured inputs		Measured target temperatures	
Ambient temperature	ϑ_a	Permanent magnet	ϑ_{PM}
Liquid coolant temperature	ϑ_c	Stator teeth	ϑ_{ST}
Actual voltage d/q-axes	u_d, u_q	Stator winding	ϑ_{SW}
Actual current d/q-axes	i_d, i_q	Stator yoke	ϑ_{SY}
Motor speed	n_{mech}		
Derived inputs			
Voltage magnitude	u_s		
Current magnitude	i_s		
Electric apparent power	S_{el}		
Joint interaction #1	$i_s \cdot \omega$		
Joint interaction #2	$S_{el} \cdot \omega$		

Input and target variables

https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature
https://ieeexplore.ieee.org/abstract/document/9296842

Electric Motor Temperature

- **Input**: Sensor measurements at 2Hz frequency.
- **Output**: stator yoke temperature, stator winding temperature, stator tooth temperature, permanent magnet temperature.
- **Suggested Task**: Develop a model to predict temperatures based on current and previous sensor readings.
- **Possible Models**: Dense networks, Convolutional neural networks, Recurrent neural networks .

https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature
https://ieeexplore.ieee.org/abstract/document/9296842

Cyberbullying

- The dataset contains more than 46000 tweets labelled according to the class of cyberbullying: Age; Ethnicity; Gender; Religion; Other type of cyberbullying; Not cyberbullying

- The data has been balanced in order to contain around 8000 of each class.

- This dataset contains text input which needs to be converted into numerical values before it can be used with machine learning models. This process is called text embedding and is a required step for this task.

- This task is more challenging and only recommended for groups who want to explore text processing methods.

https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification
J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.

Cyberbullying

- **Input**: Tweet text.
- **Output**: Class of cyberbullying.
- **Suggested Task**: Develop a model to predict the class of cyberbullying.
- **Possible models:** Neural networks, Recurrent neural networks, Language models.

https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification
J. Wang, K. Fu, C.T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.