COMP3010 Machine Learning - Assignment S1 2024
@ Computing, Curtin University

Last updated: 26th March, 2024

Weighting:
This assignment is of 100 points, which weighs 40% of the final mark.

Submission:
You need to submit your prediction to Kaggle (see Section 5.1). You also need to submit everything in **a single ZIP** file to Blackboard. Name the file as `<studentID>_<name>_assignment.zip`. The due date is **5 May 2024 11:59 PM**.

Academic Integrity:
This is an **individual** assignment so any form of collaboration is not permitted. This is an **open-book** assignment so you are allowed to use external materials, but make sure you properly **cite the references**. It is your responsibility to understand Curtin's Academic Misconduct Rules, for example, post assessment questions online and ask for answers is considered contract cheating and not permitted.

1 Introduction

Around the globe, a significant volume of oil and gas products, which include hazardous materials, are transported daily through various means. Often, this transportation occurs through densely populated areas, elevating the risk to nearby structures and inhabitants. The road transportation of Liquefied Petroleum Gas (LPG) is particularly prevalent in industrialized nations, raising public safety concerns. One of the critical risks associated with this is the occurrence of Boiling Liquid Expanding Vapour Explosions (BLEVEs). These intense explosions result from complex nonlinear physical processes and can cause significant harm to both structures and people. Predicting the intensity of these explosions remains a challenge with conventional methods. For more details, please refer to research papers on this matter [1, 2].

In response, this project aims to leverage data-driven machine learning techniques to predict the pressure generated by the blast waves from these explosions.

2 Problem Description

For this assignment, your task is to conduct a predictive analysis focusing on the peak pressure caused by BLEVEs. Imagine a scenario where a BLEVE occurs inside a rectangular tank situated within a 3D environment Figure 2. Nearby, a rigid wall mimicking a building structure is positioned at a certain distance from the BLEVE source. This wall alters the trajectory of the blast wave, causing energy reflections and deflections, thereby adding complexity to the situation.

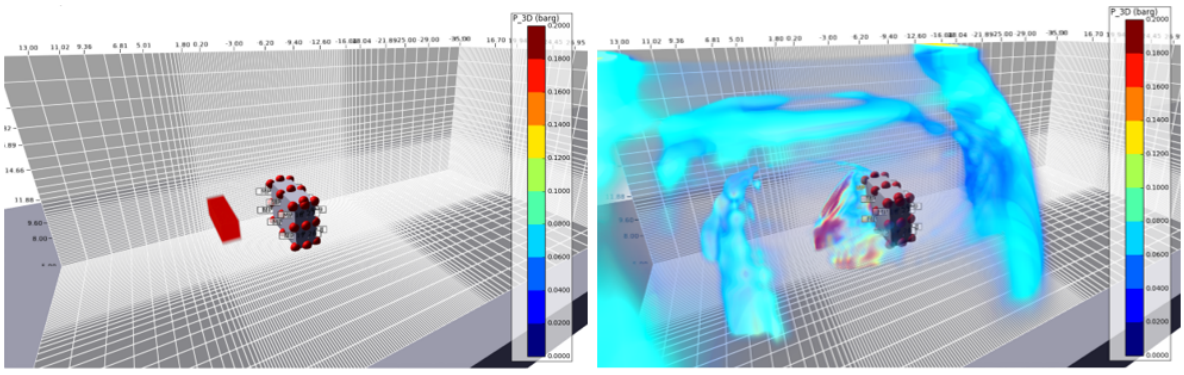

Figure 1: BLEVE blast wave propagation in an obstacle environment.

Your goal is to accurately predict the peak pressure in the vicinity of this obstacle. To facilitate this, 27 sensors are strategically positioned around the obstacle's walls (9 each on the front, back, and side) to gather training data, as shown in Figure 2. These sensor points will also serve as the basis for testing.

The 3D environment encompasses a range of physical variables related to BLEVE, including but not limited to temperature, pressure, the ratio of gas to liquid, and the dimensions of both the tank and the obstacle. The complete list of features is as follows:

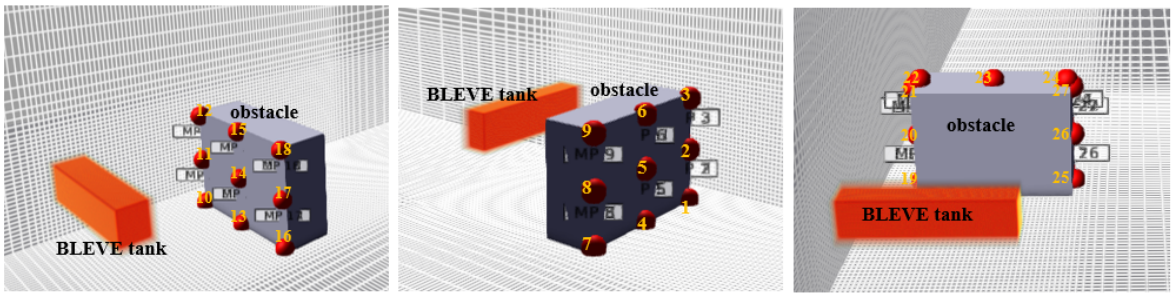

Figure 2: Sensor points on the obstacle. Left to right: font wall, back wall, and side wall.

- `Tank Failure Pressure`: the pressure within the tank when BLEVE happens (in bar)
- `Liquid Ratio`: the ratio of liquid in the tank (liquid and vapour coexist)
- `Tank Width`: the width of tank (in meter)
- `Tank Length`: the length of tank (in meter)
- `Tank Height`: the height of tank (in meter)
- `Vapour Height`: the height of vapour in tank (in meter)
- `BLEVE Height`: the distance of tank to the ground (in meter)
- `Vapour Temperature`: the temperature of vapour (in K)
- `Liquid Temperature`: the temperature of liquid (in K)
- `Obstacle Distance to BLEVE`: the distance of obstacle to BLEVE (in meter)
- `Obstacle Width`: the width of obstacle (in meter)
- `Obstacle Height`: the height of obstacle (in meter)
- `Obstacle Thickness`: the thickness of obstacle (in meter)
- `Obstacle Angle`: the angle between the line connecting obstacle centers and BLEVE centers and the horizontal line
- `Status`: the status of liquid, either subcooled or superheated
- `Substance Critical Pressure`: the pressure required to liquefy a vapour of the substance at its critical temperature (in bar)
- `Substance Boiling Temperature`: the temperature above which liquid of the substance turns into vapour at atmosphere pressure (in K)

- `Substance Critical Temperature`: the temperature above which vapour of the substance cannot be liquefied, no matter how much pressure is applied (in K)
- `Sensor ID`: the ID of the sensor ranging from 1 to 27
- `Sensor Position Side`: the side of the wall where the sensor locates
- `Sensor Position x`: the x coordinate of the sensor
- `Sensor Position y`: the y coordinate of the sensor
- `Sensor Position z`: the z coordinate of the sensor
- `Target Pressure`: the target peak pressure to be predicted (in bar)

3 The Tasks

In the assignment, along with this documentation, you are provided with train.csv and test.csv, which contain the training set and testing set of the data. The target pressure (ground truth) of the training set is given but not for the testing set. You are going to train a machine learning model with the training set and predict the pressure for the testing set.

You will have three main tasks, including data preprocessing, model development, and report.

3.1 Data Preprocessing

Data preprocessing is vital for machine learning. This not only encompasses the selection and potential removal of features with limited or no relevance to the target variable but also invites you to engage in the creation of new features. Additionally, you are encouraged to adapt data types according to the requirements of your chosen model. A significant focus should be placed on data cleaning, a fundamental aspect of preprocessing. This could include:

- **Identifying and Handling Missing Values**: Scrutinize the dataset for any missing or incomplete data entries. You must decide on strategies like imputation or removal based on the context and impact on the dataset.
- **Outlier Detection and Treatment**: Investigate for any anomalies or outliers in the data that could potentially skew the results. Employ appropriate statistical methods to either correct or exclude these outliers.
- **Checking for Duplicates**: Ensure the integrity of the dataset by identifying and removing any duplicate records, as they can lead to biased or inaccurate model training.
- **Rectifying Incorrect Entries**: Be vigilant for any inaccuracies or incorrect entries within the dataset, correcting them as needed to maintain the quality and reliability of your data.

- **Feature Selection**: Some features may have little or no correlation with the target and can probably be removed. More in general, a "sparse" model can be trained with the most important features
- **Feature Engineering**: You do not have to restrict yourself in the set of features provided. You can create new features on your own! E.g., the ratio $\frac{\texttt{Tank Width}}{\texttt{Tank Length}}$ can be considered as another feature to your model
- **Data Type Conversion**: Depending on the model you use, you may need to convert features to the suitable data type. E.g., you may want to convert categorical features to numeric ones with certain encoding methods
- **Feature Scaling**: Consider normalization or standardization techniques to scale your data, especially if your model is sensitive to the magnitude of input values.
- **Data Augmentation**: To enhance the robustness of your model, you may consider increasing data instances using augmentation techniques.
- **Others**: Any other data preprocessing techniques

Note that some preprocessing may have effects on each other. Pay attention to the order of preprocessing steps.

3.2 Model Development

In the model development phase of this assignment, your objective is to explore and compare a diverse range of machine learning models to identify the one most suited for predicting peak pressure in BLEVE scenarios. A key requirement is to examine **at least three fundamentally different types of machine learning models**. These models need to be evaluated and compared with at least two different metrics. Things may include:

- **Model selection**: you need to examine at least three different machine learning models, such as linear models, support vector regression, random forest, xgboost, neural networks, or any other models you think it is suitable. It is important to note that choosing multiple variations of similar model types, such as Random Forest and Gradient Boosting Decision Trees (GBDT), does not fulfil the requirement of exploring three distinct model types.
- **Hyparameter tuning**: Remember, each machine learning model comes with its set of hyperparameters, which are crucial for optimizing its performance. Invest time in fine-tuning these hyperparameters, using techniques such as hold-out validation or cross-validation, to achieve the best results.
- **Evaluation metrics**: Your models should be evaluated using compulsory metrics like MAPE (Mean Absolute Percentage Error) and R^2, both available in the Scikit-learn library. You may also consider additional metrics like RMSE (Root Mean Square Error) or MAE (Mean Absolute Error) to gain a more comprehensive understanding of your models' performance.

- **Model Ensemble**: After developing multiple models, consider leveraging the power of model ensembling. This technique involves combining different models to improve the final prediction performance.

Once you are happy with the model you trained, you can apply it to the test set to get predictions. A Kaggle competition is available for you to evaluate the performance (see Section 5.1).

3.3 Report

As part of this assignment, you are required to compile a comprehensive report documenting each step undertaken throughout the project. This report should serve as a detailed record of your methodologies, choices, and insights gained during the assignment. The following checklist provides guidance on the key components that your report should cover:

- Data cleaning: Detail the types of data issues identified (missing values, outliers, duplicates, incorrect entries, etc.) and the specific actions you took to address them.
- Data processing: Describe the preprocessing steps executed, such as normalization, feature engineering, and data type conversion. Explain the rationale behind these steps.
- Model selection: Discuss the various models you investigated and the reasons for selecting or rejecting each. Emphasize the diversity in the types of models chosen and their relevance to the problem at hand.
- Hyperparameter tuning: For each algorithm, outline the hyperparameters that were tuned, the strategy employed for searching (grid search, random search, etc.), the range of values considered, and the optimal values determined.
- Prediction: Explain how the final predictions on the test set were derived. This might include the use of a single model, a combination of models, or an ensemble method.
- Self-reflection: This section is more open-ended and personal. Reflect on your overall understanding and feelings about the project. Discuss any difficulties encountered, lessons learned, and how you might approach the project differently if given another chance. Feel free to include any other relevant thoughts or insights.

Your report should be concise and informative. It must **NOT exceed 10 A4 pages** in length.

4 Python Environment

You will use `Python` for this assignment and you can use any library you like. You can conduct experiments with your local python environment but the final submission has to be a **Jupyter Notebook that can be run on Google Colab**. Note that the notebook you

submitted should contain necessary comments or markdown cells to briefly explain what you are doing.

When saving the notebook for submission, make sure it contains the cell output. Colab does save cell outputs by default. If you are not sure, double-check the notebook setting and make sure the "Omit code cell output when saving this notebook" is disabled.

5 Submission

5.1 Kaggle submission

A private Kaggle competition has been created specifically for this assignment. This platform allows you to monitor your model's performance on the test set. The test set is divided into two equal parts: one half is used for the public leaderboard, and the other half for the private leaderboard.

The public leaderboard enables you to track your model's performance. You can submit your predictions to Kaggle before the assignment deadline and view your model's MAPE score on the public leaderboard, alongside scores from other participants. It is important to note that this is not a competition in the traditional sense; it is primarily a tool for you to gauge your progress. On Kaggle, you have the option to use a nickname for anonymity. Remember to include your Kaggle name clearly at the beginning of your notebook and report (otherwise 'performance' marks may not be given to you correctly).

You are allowed up to five submissions per day. On the assignment's due date, Kaggle will determine your final score based on your best prediction (as reflected in your public leaderboard scores) against the other half of the test set. This score, displayed on the private leaderboard, will be used to assign marks for 'model performance' (detailed in Section 6).

Please note that it is not mandatory to submit predictions to Kaggle daily. However, you must submit your final prediction before the assignment deadline to receive a score on the private leaderboard and avoid losing the 'performance' marks (which account for 20% of your grade). Regular submissions are strongly recommended as a way to monitor your progress and avoid potential surprises in performance discrepancies between your training and testing results, especially if submissions are made at the last minute.

The Kaggle link for this competition is
`https://www.kaggle.com/t/65bcb11b674fafbe1a7d97c2623d8a35`.
If you encounter any issues participating in this competition, please contact the unit coordinator.

5.2 Blackboard submission

Beside the Kaggle submission, you will also need to make a final submission to BlackBoard. You are required to submit a single zip file that contains all documents, including:

- The source code main.ipynb with your code, comment, output, and Kaggle name
- The report report.pdf with your documentation and Kaggle name

- The csv file that contains your prediction for the test set **prediction.csv** (as in the format of sample_prediction.csv on Kaggle)
- The signed **declaration form**
- (Optional) The **README** file which contains all other information that is not suitable to put into markdown cells of your jupyter notebook

6 Marking

The assignment carries a total of 100 marks, distributed across various components as follows:

- **Satisfactory submission [10 marks]**: This category assesses the overall compliance with submission requirements, including:
  - Inclusion of all required files
  - Proper naming and organization of files
- **Data preprocessing [20 marks]**: Marks in this section are awarded based on the quality of data preprocessing, which may include:
  - Correct identification and treatment of data issues, such as missing values, outliers, duplicates, etc.
  - Appropriate handling of data types and normalization
  - Implementation of advanced preprocessing techniques that enhance model performance, such as effective feature engineering
- **Model development [30 marks]**: This section evaluates the rigour of the model development process, including:
  - Consideration of a diverse range of machine learning models
  - Thorough hyperparameter tuning, utilizing methods like hold-out validation or cross-validation
  - Fair and comprehensive evaluation of the models
- **Prediction [20 marks]**: This component focuses on the performance of the final model, including:
  - The model's performance metrics on the testing set (MAPE on the private leaderboard of Kaggle)
  - **There will be a bonus of 10 marks for the top 10 competitors on the private leaderboard**
  - Analysis of the difference between training and testing performance

Table 1: Marking guide based on testing MAPE

MAPE	Mark(s)
< 0.2	20
0.2 - 0.25	18
0.25 - 0.3	15
0.3 - 0.4	12
0.4 - 0.5	8
0.5 - 0.6	4
> 0.6	0

- **Report [20 marks]**: It should provide comprehensive and insightful documentation of the project, including:
  - Inclusion of all required sections as outlined in the report guidelines
  - Additional insightful interpretation and analysis of the data and model
  - Clear, knowledgeable insights drawn from the project

It is important to note that while some factors might not be directly marked, they can significantly influence the marks awarded in all sections, such as the readability and quality of your notebook (both in terms of code and text). If any part of your submission is unclear or not understandable, it could result in the loss of marks for that section. Always strive for high-quality, well-documented code and a clearly written report to effectively communicate your work.

This is the end of the assignment specification. Have fun!

References

[1] Jingde Li, Qilin Li, Hong Hao, and Ling Li. Prediction of bleve blast loading using cfd and artificial neural network. *Process Safety and Environmental Protection*, 149:711–723, 2021.

[2] Qilin Li, Yang Wang, Yanda Shao, Ling Li, and Hong Hao. A comparative study on the most effective machine learning model for blast loading prediction: From gbdt to transformer. *Engineering Structures*, 276:115310, 2023.