National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Aditya Pandey
Student ID:	x23275286.....
Programme:	Master of Science in Data Analytics **Year:** 2024-25
Module:	Statistics and Optimisation (MSCDAD_B)
Lecturer:	Noel Cosgrave
Submission Due Date:	04 JAN 2025
Project Title:	TERMINAL ASSIGNMENT
Word Count:	1307

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Aditya Pandey
Date:	04 JAN 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[MSCDAD_B)]

[Statistics and Optimisation]

Your Name/Student Number	Course	Date
x23275286/Aditya	MSCDAD_B	04 JAN 2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
Grammarly	To correct the grammar mistake during writing the report	https://www.grammarly.com/

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

1. (A)
    a. Decision Variables:
        i. x_i: Binary variable (1 if city i is visited, 0 otherwise), for $i \in \{1,2,...,n\}$.
        ii. t_{ij}: Binary variable (1 if travel from city i to city j is selected, 0 otherwise), for $i \neq j$
        iii. a_i: Arrival time at city i, continuous, $a_i \geq 0$, for $i \in \{1,2,...,n\}$.
    b. Parameters:
        i. v_i: Delivery value for city i.
        ii. c_{ij}: Distance between cities i and j.
        iii. $w_{i,start}, w_{i,end}$: Time window for city i.
        iv. M: Large constant for big-M constraints.
    c. Objectives:
        i. Maximize Delivery Value: We aim to maximize the total value of the cities visited.
        ii. Minimize Travel Time: We aim to minimize the total travel distance between the cities visited.
    d. Constraints:
        i. City Visits: The total number of cities visited should be between 3 and 5.
        ii. Time Windows: Each city must be visited within its specific time window.
        iii. Sequential Travel: The travel sequence should respect the time it takes to travel between cities.
        iv. Dependencies: Some cities must be visited if others are visited (e.g., city 5 must be visited if city 1 is visited).
        v. Mutual Exclusivity: Only one of a pair of cities can be visited (e.g., either city 2 or city 3, but not both).
    e. Variable Types:
        i. x_i and t_{ij} are binary variables (0 or 1).
        ii. a_i is a continuous variable representing the arrival time at city i.
2. (B) - Failed to obtain a solution that is feasible. The problem has contradictory constraints or have constrains that are too stringent. I have attempted to use different approaches (such as cutting) but it could not address the issue. The cities A, D, E, and H was included together with their arrival times, but no feasible solution was obtained.
3. (C) - The solver solves the problem in two steps:

1. Profit Maximization: It aims to maximize the delivery profit while respecting the constraints. The result will show which cities were visited to achieve the highest profit and the associated arrival times.
2. Distance Minimization: It minimizes the total travel distance while respecting the same constraints. The result will show the visited cities that minimize the travel distance and their arrival times.

Both objectives are solved independently, so you get a trade-off between profit and distance, allowing you to compare the two solutions and analyze which cities are visited in each scenario.

4. (D) – NSGA-II Algorithm Steps for Multi-objective Optimization:
    a. Initialization: Randomly generate the first population of solutions and each solution is a sequence of cities to visit.
    b. Non-dominated Sorting: Rank solutions based on dominance
    c. Crowding Distance: Use crowding distance factor to measure diversity each front
    d. Selection: In selecting parents for crossover, tournament selection should be employed in relation to rank and crowding distance.
    e. Crossover: Crossover to exchange two points of parent solutions to create new solutions.
    f. Mutation: Mutate to add diversity and go for new solution.
    g. Replacement: Form a new population by selecting the best solutions from the parent and offspring populations, using dominance rank and crowding distance.
    h. Termination: Stop after a predefined number of generations and provide a Pareto front or set of non-dominated solutions.

Result Analysis:

- Delivery Value: 2800
- Travel Time: 1175.73 hours
- Visited Cities: [1, 1, 1, 1, 1, 1, 1, 1] (All cities visited)

This solution represents the best trade-off between maximizing profit and minimizing travel time. The Pareto front shows that Solution 1 cannot be improved in either objective without compromising the other. If we prioritize minimizing travel time, fewer cities might be visited, but with a lower delivery value.

Thus, NSGA-II effectively provides a set of Pareto-optimal solutions offering different trade-offs, helping the decision-maker choose a solution based on their specific priorities.

Logistic Regression

1. (A) - Logistic Regression Equation - The logistic regression equation is:
   $P(y=1) = 1 / (1 + \exp(-(\beta0 + \beta1x1 + \beta2x2 + ... + \beta30x30)))$
   Explanation:
   y: Dependent variable (binary target: 0 = benign, 1 = malignant).
   x1, x2, ..., x30: Independent variables (30 features in the dataset).
   $\beta0$: Intercept.
   $\beta1, \beta2, ..., \beta30$: Coefficients for each feature.
   Probability threshold: If $P(y=1) >= 0.5$, classify as malignant; otherwise, benign
2. (B)
   Load and Explore Dataset
   - Dataset contains 569 rows and 31 columns (30 features + 1 target).
   - Target variable distribution:
     - Malignant tumors (1): 357 cases (63%).
     - Benign tumors (0): 212 cases (37%).
   Observation: Moderate class imbalance exists.
3. (C) - Model Evaluation Summary:
   a. MSE (0.0633): The model's predictions are fairly accurate, with a small average squared error.
   b. R^2 (0.7304): The model explains 73% of the variance in the target variable, indicating a good fit.
   c. MAE (0.1941): On average, the model's predictions are off by about 0.1941 units.
   d. MedAE (0.1633): For 50% of the predictions, the error is less than 0.1633, showing robustness to outliers.
   e. EVS (0.7344): The model explains about 73.4% of the variance, like R^2.
   Conclusion:
   Finally, this model performs well it is giving reasonably accurate predictions. We can still improve model, by feature engineering, regularization, or maybe trying more complex models.
4. (D) - With **alpha = 0.5**, the model found that **x1**, **x2**, **x3**, **x4**, **x5**, and **x9**. These features have very low p-values, suggesting they play an important role in the model. Other features like **x6**, **x7**, **x8**, and **x10** appear less important and might be excluded for a more efficient model.
   When alpha was set at 0.5 the model showed x1, x2, x3, x4, x5, and x9 as are significant predictors. These all have very low p-values, and thus we can say that they are large importance to them in the model.
   Features such as x6, x7, x8, x10 seem to be of lesser significance and can hence be dropped in an effort to come up with a more effective model.

5.

Variable	Odds Ratio	Interpretation
Mean Texture	0.69	A one-unit increase in mean texture decreases the odds by 46%.
Mean Area	0.78	A one-unit increase in mean area decreases the odds by 29%.
Mean Smoothness	0.89	A one-unit increase in mean smoothness decreases the odds by 12%.
Mean Radius	0.9	A one-unit increase in mean radius decreases the odds by 11%.
Mean Perimeter	0.91	A one-unit increase in mean perimeter decreases the odds by 9%.

Among the independently significant variables, Mean Texture has the greatest impact of dominating the odds of the outcome with decreased odd by 46% for each variation in unit.
Mean Perimeter is the least sensitive variable with a mean decrease of 0.09 in odds for a unit upturn in the perimeter.

5. Predicted Probability of Malignancy: 0.7239 Prediction: Malignant

---

Time Series

1.
    a. **Nature:** Highly volatile with no clear trend or seasonality; contains significant outliers.
    b. **Stationarity:** Likely non-stationary; ACF shows gradual decay.
    c. **Outliers:** Extreme values detected, requiring treatment.
    d. **Model Insights:** PACF suggests potential AR(1) component.
    e. ADF Statistic: -21.611395679774343 p-value: 0.0 The time series is stationary.
2. The performance comparison of both ARIMA and SARIMA models reveal that SARIMA performs better than ARIMA in all aspects. Indeed, we have seen that for SARIMA the value of AIC and BIC is lower and that means that the model has a better fit, but the number of its parameters is smaller as well. It also yields slightly lower estimate of MSE implying better forecasting ability but the difference is very negligible. However, SARIMA has the higher R-squared meaning it accounts for more of the variance in the data than the other model, although both have negative R-squared figures, meaning none of the models are a very good fit to the data. Specifically, in these evaluation metrics for these datasets, model comparison shows that SARIMA is more appropriate than SETAR for these applications but they could still be fine tuned or other methodology employed for better forecasts.
3. Done in code