**AI5213/EC4213: Machine Learning and Deep Learning** **Student ID:**
**Homework 3 (Kaggle Competition)** **Name:**

**Due: Sunday, Nov 24, 11:59PM**

Welcome to HW3! This assignment features a Kaggle competition on flight price prediction where you'll apply tree-based models and practice feature engineering. The homework consists of a prediction competition and a technical report.

**Essential Resources:**

- Tree Models in ISLP: https://islp.readthedocs.io/en/latest/labs/Ch08-baggboost-lab.html
- Kaggle Competition Guide: https://www.kaggle.com/learn/guide/kaggle-competitions
- Feature Engineering: https://www.kaggle.com/learn/feature-engineering

**Recommended Workflow:**

- Start with EDA and basic data cleaning
- Implement a simple baseline model (e.g., decision tree)
- Gradually explore advanced models (Random Forest, XGBoost) and ensemble them
- Use cross-validation to prevent overfitting
- Document your progress and insights

**Deliverables:**

- Submit a single PDF via Gradescope ("HW3 Write-Up")
- Include your competition results and technical report
- Attach your code in the appendix
- Use LaTeX template provided (or neat handwriting if necessary)

**Important Guidelines:**

1. Sign the honor code statement on the next page
2. Document any collaboration or help received
3. Write all responses in English
4. Properly mark question sections in Gradescope

**For staff use only**

| Honor Code | Q1 (Part 1) | Q2 (Part 2) | Total |
|---|---|---|---|
| / 5 | / 45 | / 50 | / 100 |

## Honor Code [5 pts]

**Declare and sign the following statement:**

*"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."*

*Signature*:

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

(a) Did you receive any help whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g., "Alex shared insights on optimizing hyperparameters for XGBoost during a group discussion.")

(b) Did you give any help whatsoever to anyone in solving this assignment? If your answer is 'yes', give full details (e.g., "I advised Chris to check out the Kaggle feature engineering guide for handling missing values.")

(c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

## Q1. Mini Competition: Flight Ticket Price Prediction [95 pts]

Welcome to a data science challenge! In this mini-competition, you'll work with real-world data to build and improve your predictive models. Your journey will be evaluated on two aspects: your model's performance on the Kaggle leaderboard (45 pts) and a technical report documenting your approach (50 pts). This is your chance to apply classroom knowledge to a practical problem while competing with your peers.

### Part 1: Competition Performance [45 pts]

Your task is to predict flight ticket prices. Strive to surpass the target score and submit your results. You're free to use any libraries available - such as scikit-learn, pandas, numpy, statsmodel, ISLP, xgboost, polars and others. Your grade for this part will be determined entirely by your final performance on the private leaderboard. Follow these simple rules: Use your student ID, play fair, give it your all, and, most importantly, enjoy the mini-competition: https://www.kaggle.com/t/d51884752a1d46aba5b1048bd895f15f

#### Score guideline

- The leaderboard uses Mean Absolute Percentage Error (MAPE) as the evaluation metric, where smaller scores indicate better models.
- While you'll see and capture your public leaderboard score during development, final grading will be based on the private leaderboard score to ensure your model generalizes well to unseen data.
- You can submit up to five times per day. We encourage you to divide the dataset into train/validation sets and use your internal validation set to select the best model for submission.
- Final points will be determined by your private leaderboard score $s$ with the target score $t$.
- Who achieved better the target score $(s \leq t)$ will receive the full points (45 pts).
- Those who didn't beat the target score $(t < s)$ will receive $\max(45(1 - (s - t)), 0)$ points.
- You're free to use the Ed & Kaggle discussion tab for sharing your thoughts or your code snippets.
- However, those who break the honor code will be handled accordingly. Since we have access to last year's leaderboard and code submissions, referencing or reusing code from previous students will be treated as an honor code violation. (Honor code at https://sundong.kim/courses/mldl24f)

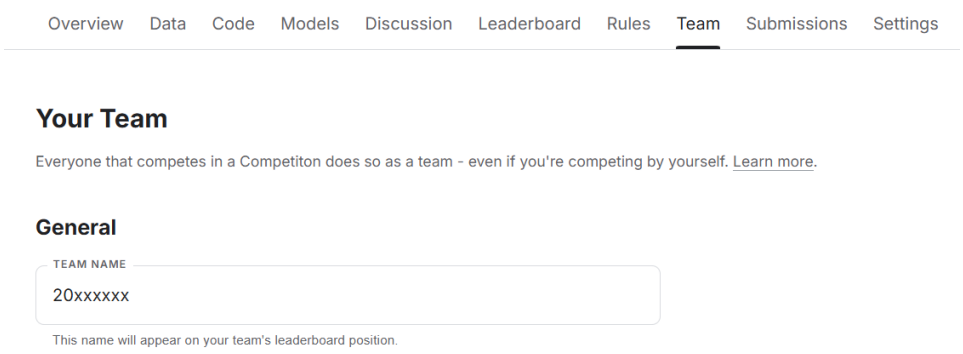



Figure 1: Use your student ID as a team name

**Part 2: Technical Report [50 pts]**

Write a self-contained report (at most 4 pages) documenting your approach and findings. For reference on how to structure your report, you can check example write-ups at https://wsdm-cup-2018.kkbox.events/.

Your report should include:

- **Problem Description and Data**
  - Clear statement of the task and evaluation metric
  - Dataset characteristics and initial insights
  - Train/validation split strategy
- **Data Preprocessing & Feature Engineering**
  - Data cleaning and handling of missing values/outliers
  - Feature creation rationale and importance analysis
  - Feature selection process
- **Modeling & Results**
  - Model architecture and validation methodology
  - Hyperparameter tuning strategy
  - Performance analysis and key findings
- **Technical Implementation**
  - Key libraries/frameworks used
  - Challenges faced and solutions

**Format requirements:**

- Please limit the main document to 4 pages in the current single-column LaTeX format, with unlimited references allowed beyond the page limit.
- Include a screenshot of your public leaderboard score as evidence
- Include representative code snippets for important steps
  (Data preprocessing, feature engineering, model training)
- Include relevant figures/tables to support your discussion
- Proper citations if you reference any external resources
- Clear structure with appropriate section headers
- Use clear and concise technical writing

Write your technical report here.

[Note: Please remove the following text and figures when writing your report.] Participating in this Kaggle competition [1] was an amazing experience! It was my first time applying LightGBM [2] and XGBoost [3] to real-world datasets in a competitive programming setting. See Figure 2 for the overall architecture. After 12 attempts, I was able to achieve MAPE $= 0.0705$ (Figure 4). I've included a LaTeX code of the figure and table from previous work [4, 5] to demonstrate how to incorporate them in your report.

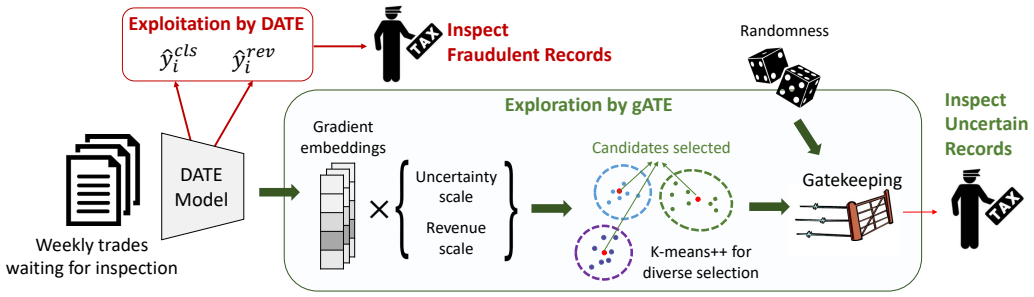
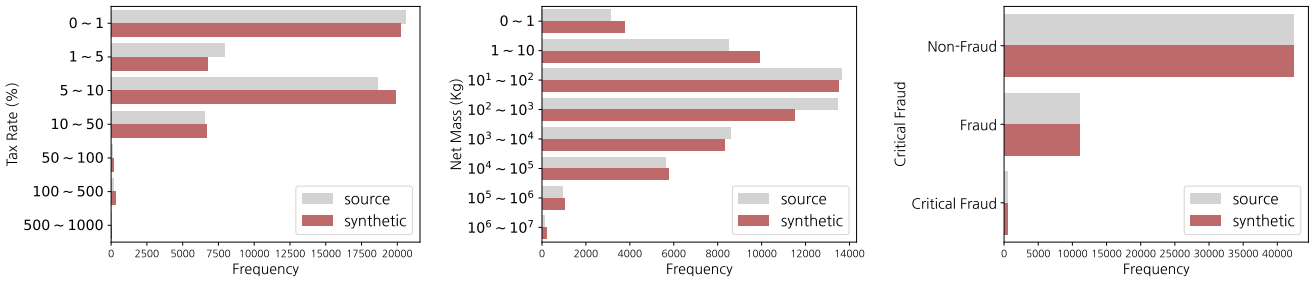



Figure 2: Illustration of the hybrid selection framework.




(a) Tax Rate (b) Net Mass (c) Critical Fraud

Figure 3: Exploratory data analysis on two predictors and target variable

Table 1: Data description

| Attribute | Value | Explanation |
|---|---|---|
| Declaration ID | 97061800 | Primary key of the record |
| Date | 2020-01-01 | Date when the declaration is reported |
| Office ID | 13 | Customs office that receives the declaration (e.g., Seoul regional customs) |
| Importer ID | HQ0W7JA | Consumer who imports the item |
| Seller ID | PBP2MYI | Overseas business partner which supplies goods to Korea |
| Courier ID | MWIDNS | Delivery service provider (e.g., DHL, FedEx) |
| HS6 Code | 090121 | 6-digit product code (e.g., 090121 = Coffee, Roasted, Not Decaffeinated) |
| Country of Departure | JP | Country from which a shipment has or is scheduled to depart |
| Tax Rate | 8.0 | Tax rate of the item (%) |
| Tax Type | A | Tax types (e.g., FTA Preferential rate) |
| Net Mass | 1262.0 | Mass without any packaging (kg) |
| Item Price | 1437418.0 | Assessed value of an item (KRW) |
| Critical Fraud | 1 | Among frauds, critical frauds that can threaten public safety, are marked as 2 (0/1/2 Ternary). |

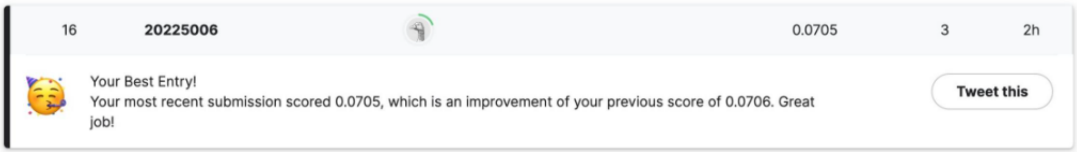



Figure 4: Screenshot of leaderboard score

# References

[1] S. Kim. (2024) GIST-MLDL24f-HW3. [Online]. Available: https://www.kaggle.com/competitions/gist-mldl24f-hw3

[2] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[4] C. Jeong, S. Kim, J. Park, and Y. Choi, "Customs Import Declaration Datasets," *arXiv:2208.02484*, 2022.

[5] S. Kim, T.-D. Mai, S. Han, S. Park, T. Nguyen, J. So, K. Singh, and M. Cha, "Active Learning for Human-in-the-loop Customs Inspection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 039–12 052, 2023.