AI for Business — Titanic Lab Assignment

EDA + Logistic Regression + Random Forest + Neural Networks

Instructor: Miao Liu

Due: **Mar 1, 2026 (Midnight)**

Overview

In this lab, you will work with the **Kaggle Titanic** dataset, a classic introductory problem in data science and machine learning. The goal is to predict passenger survival using features such as age, sex, class, fare, and embarkation port. You will:

- Perform **exploratory data analysis (EDA)** and visualization.
- Build a baseline **logistic regression** model.
- Train and interpret **random forest** and **neural network** models, including versions with **hyperparameter tuning**.

Kaggle resource (tutorial + competition materials):
https://www.kaggle.com/code/alexisbcook/titanic-tutorial#The-data

Learning objectives

By the end of this lab, you should be able to:

- Load, inspect, and summarize structured data; identify missingness and basic data quality issues.
- Visualize distributions and group-level survival patterns; connect plots to a clear interpretation.
- Build end-to-end ML workflows: preprocessing → training → evaluation → prediction.
- Explain (in plain language) what a model is doing and why its performance changes across model classes.
- Communicate results to a **non-technical audience** (client/boss) with clarity and professionalism.

How to start (Google Colab + notebooks)

1. Go to Google Colab: http://colab.research.google.com/.
2. Click `File` → `Upload notebook`, and upload the provided notebooks:
   - `Titanic_Part_1.ipynb` (EDA + logistic regression),
   - `Titanic2_RH.ipynb` (random forest),
   - `Titanic2_NN.ipynb` (neural network).
3. Download the Kaggle Titanic datasets (`gender_submission.csv`, `train.csv`, and `test.csv`) from the Kaggle page above, and save them in a folder in your **Google Drive**.
4. **Important: update file paths in the notebook.** My code uses my own Drive folder. You must change the portion of each path *after* `My Drive/` to match where you saved the files in *your* Drive. For example, my path is:
   `'/content/drive/My Drive/Teaching/AI for Business/AI for Business 2025/Lab Assignment/Lab1 Titanic/train.csv'`
   You need to replace `Teaching/AI for Business/AI for Business 2025/Lab Assignment/Lab1 Titanic/` with your own folder path (e.g., `MyFolder/Titanic/`).

If you are new to Python, you may find this short course helpful: https://www.kaggle.com/learn/intro-to-programming.

Writing standard (very important)

Your target audience is a **client or boss with no AI background**. In your write-up:

- Use plain language (minimize jargon; define terms when needed).
- For every figure/table: include a **1–3 sentence interpretation**.
- When describing code: focus on the **workflow and purpose** of steps, not line-by-line details.

Student Handout

Deliverables

Submit **one PDF report** via Canvas. Your report should include:

- Answers to all required questions below,
- Clean plots and tables (with captions/labels),
- Clear interpretations and model-comparison discussion.

Optional (not required unless instructed otherwise): Include your Kaggle public leaderboard score(s) if you submit predictions.

1 Part A: Exploratory Data Analysis (EDA) and Logistic Regression

A1: Numerical variable histograms

Draw histograms for **four numerical variables** in the dataset (e.g., `Age`, `Fare`, `SibSp`, `Parch`). Provide short descriptions of the distributions you observe. Your description should address at least:

- range and typical values,
- skewness/outliers,
- whether the distribution seems "well-behaved" or suggests transformations.

A2: Correlation analysis with survival

For each numerical variable, compute and plot its correlation with the `Survived` dummy variable. Which numerical variable has the **highest correlation** with `Survived`? Provide a short interpretation of what the sign and magnitude suggest.

A3: Survival rates by categories

Examine three categorical variables (e.g., `Pclass`, `Sex`, `Embarked`). Plot survival rates for each group/category and provide brief descriptions of patterns or insights you observe. Your write-up should include at least one sentence about **why** the pattern might occur.

A4: Fare binning and survival

Since `Fare` often shows a strong relationship with survival, divide `Fare` into bins (quantiles or custom bins) and analyze its relationship with `Survived`. What do you observe? Explain how binning changes (or clarifies) the relationship compared to using raw `Fare`.

A5: Missing data analysis

Identify missing observations in each variable:

- Which variable(s) have many missing values?

- Which missing values can be handled by simple imputation (and why)?
- Which missing values might be more problematic (and why)?

A6: Logistic regression model (baseline)

Train a logistic regression model using the training dataset. In your report:

1. Describe the key steps in the workflow:
   - data preprocessing (encoding, scaling if applicable, handling missing values),
   - model training,
2. Report performance using appropriate metrics (e.g., accuracy; precision; recall; f1-score; ROC). If you are not familiar with some of these metrics, search and learn how they are related to type I and II errors.
3. Interpret the results: What types of passengers does the model predict as more likely to survive (based on coefficients or patterns)?

2 Part B: Random Forest Models

B1: Simple random forest model (no tuning)

Train a basic random forest model **without hyperparameter tuning**. Describe the workflow, including:

- data preprocessing steps,
- model training process,
- model evaluation approach and metrics.

Finally, interpret the outputs: What insights can be gained from predictions and performance? If your notebook reports feature importance, discuss the top features and what they mean in plain language.

B2: Hyperparameter-tuned random forest model

Train a more sophisticated random forest model with hyperparameter tuning. Address:

- workflow differences compared to the simpler model,
- the list of hyperparameters being tuned,
- the **candidate values** chosen for each hyperparameter,
- calculation of the **total number of hyperparameter combinations** explored.

Finally, interpret the outputs: What (if anything) improved? Why might tuning help (or not help) on this dataset?

3 Part C: Neural Network Models

C1: Simple neural network model (no tuning)

Train a basic neural network model **without hyperparameter tuning**. Describe:

- the model workflow (inputs, architecture at a high level, loss/optimizer),
- the training process (epochs, batch size if applicable),
- evaluation approach and metrics.

Interpret the results: What insights can be learned? Comment on whether the neural network seems to overfit or underfit.

C2: Hyperparameter-tuned neural network model

Train a more advanced neural network model with hyperparameter tuning. Describe:

- workflow differences compared to the simpler network (what new features are added and why),
- hyperparameters being tuned,
- candidate values specified for each hyperparameter.

Finally, interpret the outputs: What changed in performance? What tradeoffs (accuracy vs complexity vs interpretability) do you see?

4 Part D: Model comparison and recommendation

Using results from Parts A–C, write a short comparison (roughly 1/2 to 1 page) that answers:

- Which model performed best on your validation (or cross-validation) metric(s)?
- Which model would you recommend to a non-technical client, and **why**?
- If your best-performing model is not the most interpretable, how would you justify using it?

Submission checklist

Before submitting your PDF, make sure you have:

- Included all required plots (with captions/labels) and brief interpretations.
- Clearly stated your evaluation approach (train/validation split or cross-validation).
- Clearly interpreted outputs for each model family (logistic regression, random forest, neural network).
- Provided a model comparison and recommendation (Part D).
- Ensured writing is clean, readable, and professional for a non-technical audience.

Submission guidelines

- Submit your assignment as a **PDF file** via Canvas.
- Late submissions will not be accepted unless prior approval has been obtained.