STATS 607 Project 1Reflection Document

Name: Judy Wu

Project url: https://github.com/judywu4800/spaceship_titanic.git

Original State of the Analysis

The original analysis was conducted as part of **STAT GU4241 (Spring 2022)** using a single Jupyter notebook. The notebook contained exploratory data analysis, preprocessing, and model fitting steps in a linear fashion, but the workflow was not modularized. There was no clear directory structure, no environment management, and no testing framework, making it difficult for others to reproduce or extend the work. The original state has been preserved in the repository and can be restored with:

```
git checkout original
```

Challenges in the Transformation

The biggest challenges were creating a clean virtual environment and curating a requirements file that would work consistently across machines since I am new to this. Another difficulty was separating data processing, model fitting, and evaluation from a single Jupyter notebook into modular Python scripts, and then constructing a pipeline that could utilize each module. It was challenging to design meaningful tests that could validate that the codes run and gives correct results.

Improvements with Most Impact

- **Project Structure**: Implementing a standardized directory layout (`data/`, `src/`, `artifacts/`, `results/`, `tests/`) gives a better clarity and reproducibility.
- **One-Command Execution**: Adding `run_analysis.py` allowed the entire pipeline to be executed with a single command, ensuring others can reproduce results without manual intervention.
- **Documentation**: Writing a professional README with setup and usage instructions made the project accessible to new users.

Future Improvements

In future projects, I would try to set up clear directory structure and modularize the workflow from the beginning. I will also try to set up environment management and testing framework so that it will be more readable and easier for others to reproduce the results.

Time Allocation

- Finding a suitable project and setting up GitHub repository and structure: ~1 hours
- Refactoring notebook into modular scripts, implementing testing strategy, debugging and polishing pipeline: ~5 hours
- Documentation (README + docstrings + reflection): ~1 hours
  Total: ~7 hours