Ames Housing Project: A Journey Through Data Science Roles

Kristin Henderson

This project, centered on the Ames Housing dataset, served as an exploration of various interconnected data science roles. The journey highlighted several overarching themes in practical data science work. Selecting the most effective AI model for collaboration presented its own learning curve, while working with AI also demonstrated that prolonged, continuous interactions can sometimes reduce efficiency; breaking tasks into smaller, focused conversations might optimize prompting and interaction speed. Furthermore, the distinct roles often revealed significant overlap, underscoring that real-world project execution is rarely linear. Instead, it demands an iterative approach, revisiting and refining work as new insights emerge from different "perspectives" or roles. As the project expanded, maintaining meticulous organization and ensuring consistent updates across all documentation became paramount. Finally, this phase primarily involved a focused exploration of individual roles, often with a necessarily narrow scope. Consequently, significant opportunities remain for future work, particularly in fully integrating these roles and tasks, such as incorporating newly acquired datasets (e.g., school data) into the modeling pipeline and subsequent, more holistic analyses.

What follows is a concise summary of the key challenges, learnings, and accomplishments experienced within each defined project role.

1. Project Manager

The Project Manager role underscored the importance of aligning objectives with stakeholder needs early on, a challenge met by prioritizing stakeholder identification *before* defining objectives. This user-centered approach and the value of a clear roadmap were key learnings. The primary accomplishment was developing the `project_roadmap.md`.

2. Stakeholder Liaison

This role highlighted the demand for a broad skill set (communication, technical insight, organization) and the centrality of communication in bridging technical and business teams. The liaison's value, especially in complex projects, and their role in ongoing user education and feedback loops were key insights. Developing `stakeholder_liaison.md` was a core contribution.

3. Data Engineer

The Data Engineer tackled challenges such as correctly encoding diverse categorical features and managing the sequence of engineered features. A key iterative loop involved adapting the ETL pipeline (`src/etl/etl_main.py`) for evolving modeling needs, like providing integer-coded nominals for deep learning. Refining AI prompts for specific tasks and ensuring creation of data dictionaries/schemas were also part of the journey. Learnings emphasized communication, iterative refinement, and robust documentation. Accomplishments include the core ETL script, related documentation, and the foundational `housing_cleaned.csv`.

4. Data Acquisition Specialist

Challenges included manual school mapping due to dataset limitations (no direct property IDs) and failed scraping attempts for school metrics, requiring manual collection. This led to a key learning: the need for robust data sources or advanced scraping for scalability. Contributions involved acquiring and processing school data and authoring a `future_acquisition_plan.md`.

5. Data Steward / Governance Officer

A unique challenge was implementing governance for *potential* sensitive data, as the initial dataset was a simple CSV. This clarified the Steward's role in policy and documentation, distinct from data processing. Learnings included designing scalable access controls and logging, even for simple storage. Accomplishments include governance scripts, access logs, and guides.

6. Data Scientist

Overcoming the deep learning model's initial poor performance was a major focus, resolved by log-transforming the target and, crucially, using Embedding layers. This breakthrough required close collaboration with the Data Engineer to modify the ETL pipeline. Debugging hyperparameter tuning efforts also presented challenges. Key learnings included confirming Random Forest's strength as a baseline and the significant impact of Embedding layers. Developing baseline and advanced models, and spearheading necessary ETL changes, were key accomplishments.

7. Machine Learning Engineer

This role focused on the technical implementation of models. Challenges included setting up a consistent deep learning environment (managing Python versions, package compatibility, and a Conda to `venv` transition) and iterative model debugging that influenced universal data loading strategies. Early deep learning predictions were also consistently low. Learnings included initial exposure to the Keras Functional API, model serialization (`.h5`, `.joblib`), and the criticality of saving all artifacts. Successfully training and saving both models and their artifacts was a key output.

8. Front-End Engineer

The primary challenge was ensuring the Gradio app inputs correctly mapped to model expectations, requiring close ML Engineer collaboration. The ongoing consideration of user-friendly features and iterative app design (potentially adding visualizations) was also key. Learnings included how model choice impacts UX and how small UI tweaks improve usability. Developing two Gradio apps and a user guide were the main contributions.

9. Visualization Expert / Data Storyteller

Challenges included deploying the Gradio app to Hugging Face (troubleshooting dependencies) and making metrics like MAPE intuitive for stakeholders via clear explanations. Learnings involved web app deployment, the importance of consistent visual styling, and accessible metric explanations. A core insight was that data storytelling aims to produce clear, tangible reports (like `docs/model_comparison_report.md`) documenting findings and utility. Developing the visualization script, generating figures for the report, and supporting app deployment were key accomplishments.

10. Explainability Engineer

Communicating SHAP value concepts to a non-technical audience was the main challenge, addressed by creating a dedicated explanation document. Learnings involved gaining initial experience with SHAP for interpreting model behavior, the value of combining visual and narrative explanations, and seeing how SHAP reveals non-linear impacts. Performing SHAP analysis, generating the summary plot, and authoring the SHAP explanation guide were key contributions.