A04 Colab\_ITAI 1371 ML

Contributors: Martin Demel, Jiri Musil, Jasmine Campos, Munachimso Benson

Department of Science, Technology, Engineering & Math, Houston Community College

ITAI - 1371 Introduction to Machine Learning

Professor Vishwa Rao

September 20th, 2025.

# Introduction

# Each team member ran the entire notebook from start to finish, then we met for a working session to compare what we observed, discuss questions, and finalize our answers together. The task was to predict book prices, and AutoGluon guided us from raw files to a trained model. By executing every cell by ourselves, we learned to read logs, interpret metrics, and understand the defaults set by the library.

# Body

We learned to start small and move quickly. Working with a 1,000-row sample gave us quick feedback while maintaining the problem's core, which helped us discuss trade-offs. AutoGluon identified a regression task and created a feature pipeline that handled categorical data and free text. Seeing nine original columns expand into 441 processed features demonstrated how much value automated feature generation can provide, especially when text is converted into usable signals.

We also learned to interpret evaluations correctly. AutoGluon prints mean squared error with a flipped sign, so higher values are better, which prompted us to convert the reported score back to the true metric and consider what that magnitude means for price predictions. The training trace made AutoML feel transparent. We observed LightGBM, CatBoost, neural nets, random forests, XGBoost, and others train sequentially, then a second-level weighted ensemble combined the top performers. The best validation result came from the ensemble at a mean squared error of 3780.2518, with learned weights of about 0.533 for NeuralNetTorch and 0.467 for CatBoost. That allowed us to explain why diverse models capture different structures in the same data and why blending them can lower error.

# Total training time was approximately 24 seconds, and the logs estimated prediction throughput at around 10,138 rows per second, which is sufficient for batch use. AutoGluon automatically created an internal validation split and saved artifacts in a timestamped folder with a single-line loader, reinforcing good practices for reproducibility and handoff.

# Conclusion

As a team, we moved beyond just running cells and gained practical judgment. We can explain why this is a regression problem, what the feature generator does, how to interpret AutoGluon’s metric convention, and why the ensemble won. We validated the value of small, quick iterations and saw how clear logs transform AutoML from a black box into a guided tour. Together, we built confidence in the workflow and in the trained model, which we can now reload and use.

# References

GeeksforGeeks. (2025, July 17). AutoGluon: an open source AutoML library. GeeksforGeeks. <https://www.geeksforgeeks.org/machine-learning/autogluon-an-open-source-automl-library/>