A Study on whether Automated Machine Learning or Generative AI can replace the Human Data Scientist

Ryan Kho Yuen Thian

BCS(Hons) in Data Science

Supervisor: Ms Fatin Izzati Binti Ramli

Abstract

The study examines whether Generative AI and AutoML can replace human data scientists in performing Machine Learning (ML) tasks. It compares the performances of the human Data Scientist, H2O AutoML and ChatGPT Data Analyst (CDA) by implementing a Credit Risk Assessment Model using the CRISP-DM methodology and a set of ML Best Practices. The strengths and weaknesses of each approach are analyzed focussing on accuracy, efficiency, and adaptability. It explores the potential for AI and AutoML to complement or replace human expertise and considers whether a hybrid approach combining both might be the most effective solution.

Problem Statement

Generative AI and AutoML, which automate code generation and machine learning tasks, are debated as potential replacements for data scientists. Articles like “Can AutoML Replace Data Scientists?”, “Will ChatGPT take Data Science Jobs?” and “Will ChatGPT Put Data Analysts Out Of Work?” explore this. While full replacement seems unlikely, these technologies could transform the data scientist role. Leveraging them effectively is key for productivity and staying competitive, as their capabilities continue to evolve.

Objectives

- To explore the 3 approaches (human Data Scientist, AutoML, Generative AI) of ML model development, evaluation and deployment.
- To compare & evaluate the 3 approaches using a realistic case study (Credit Risk Assessment) and implementing established ML Best Practices.
- To come to a conclusion as to whether or not Generative AI or AutoML can replace human Data Scientists.

Design and Methodology

Methodology: The case study was developed using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology and a realistic dataset (the German Credit Dataset). All 3 approaches are required to implement a set of established ML Best Practices. ML Algorithms used include: Logistic Regression, Random Forest, Gradient Boosting Machines and Neural Networks.

Design: The design of the Credit Risk Assessment model covers the major phases of CRISP-DM including: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. For each phase, the appropriate ML Best Practices are identified. A user interface design is proposed for data input and outcome prediction. A deployment setup is outlined, which involves the user interface passing user input to a backend API which in turn accesses the ML Model object that provides the prediction.

Main Software tools used: H2O AutoML, ChatGPT Data Analyst, Python, Anaconda, Jupyter Notebook, Spyder

Frameworks: Scikit-Learn, Tensorflow-Keras, Pandas, MatPlotLib, Seaborn, Imblearn, Shap, Streamlit, FastAPI, MLflow, etc.

Stakeholders: Data Scientists/Analysts, IT Management, Business Executives and Decision Makers, Data Science Educators and Trainers, Business End Users and the Community

Construction and Testing

Construction: *ChatGPT Data Analyst (CDA)* was used to implement the case study by issuing prompts to it. To ensure it performs the ML tasks in the correct order and that it does not omit any important tasks, it is guided via explicit prompts. CDA can make mistakes so its work must be checked carefully. *H2O AutoML* streamlines the workflow by running multiple algorithms and configurations, selecting the best-performing model based on predefined metrics. However, during Data Preprocessing and EDA, manual coding is required. A *human Data Scientist* implements a ML model by gathering and preprocessing data, selecting algorithms, training the model, evaluating its performance using suitable metrics, and tuning and testing the model to ensure it generalizes well to new data.

Deployment: Each of the 3 approaches uses a frontend Streamlit user interface and a FastAPI backend which interacts with the ML Model and Scaler objects. Since local deployment is used, a single laptop or desktop will suffice. To execute each approach, two Anaconda Prompt windows are needed: one to run the Streamlit server and open the app in a browser, and the other to start the FastAPI app with Uvicorn.

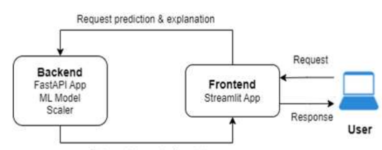

Figure 1.1: Local Deployment

User Interface & API: The user interface includes multiple input fields for entering loan application details. Once submitted, the system will display a prediction of whether the applicant is likely to default, along with a SHAP plot to explain the prediction. For the AutoML and Human Data Scientist approaches, the user interface and API were primarily developed manually. For the Generative AI approach, most of the work was handled by the ChatGPT Data Analyst (CDA), though human intervention was still required for checking and debugging.

Contribution

The study evaluates AutoML and Generative AI's potential to replace human data scientists, exploring their strengths and limitations. It provides insights for organizations on these technologies' capabilities and their role in automating data science tasks, informing both academic and industry perspectives on human versus machine expertise.

Conclusion

All 3 approaches ((human Data Scientist, AutoML, Generative AI) were able to successfully implement the Credit Risk Assessment case study, including the set of ML best practices. Through this exercise, their strengths and weaknesses were identified which led to the conclusion that the hybrid approach is the best solution, leveraging the strengths of both human and AI/automation.