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

<u>Prepared by:</u>

Ryan Kho Yuen Thian (22WMR04097)

Bachelor of Computer Science in Data Science

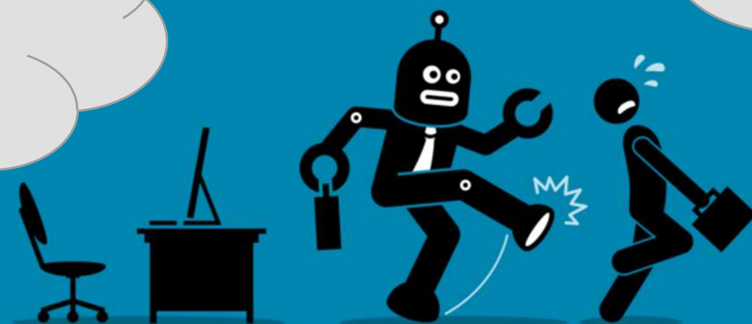
Topic to be Covered:

1. Problem Statement & Proposed Solution
2. Methodology & System Design
3. Implementation & Deployment
4. Discussion of Results
5. Live Demo of the 3 Applications

1a) Problem Statement

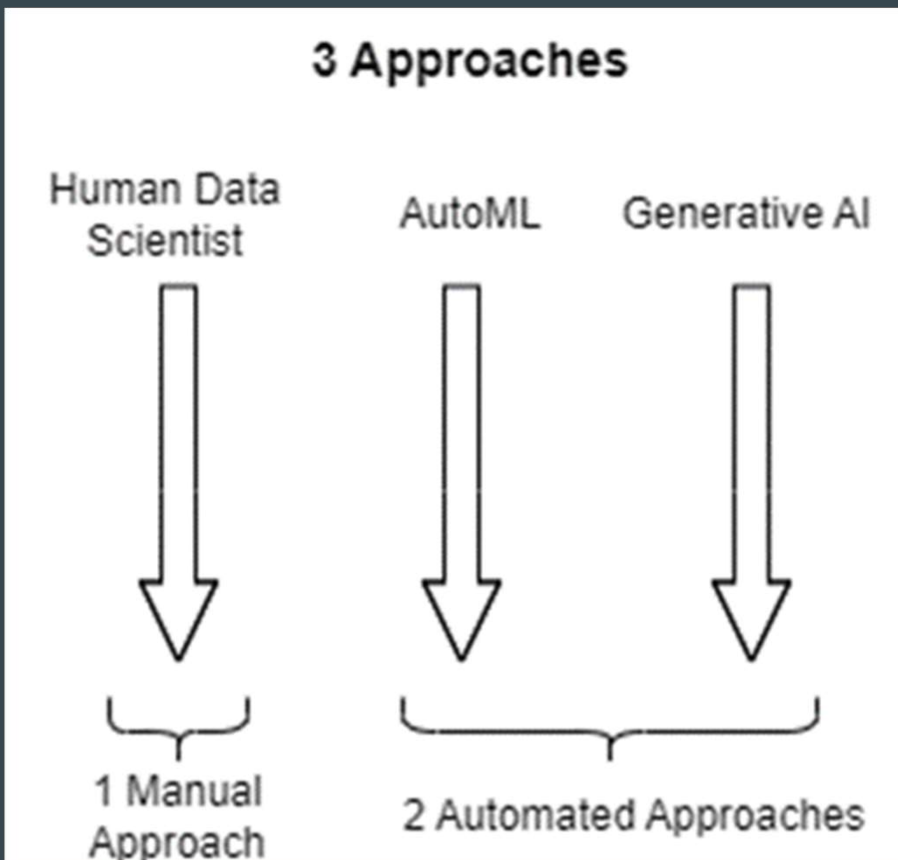
Will ChatGPT take Data Science Jobs?

Can AutoML replace Data Scientists?



Will ChatGPT Put Data Analysts Out Of Work?

1b) Objectives & Proposed Solution



- To explore the 3 approaches of Machine Learning (ML) model development, evaluation and deployment.
- To compare & evaluate the approaches in terms of how well they perform against established ML Best Practices using a realistic case study (Credit Risk Assessment).
- To determine whether or not Generative AI or AutoML can replace Human Data Scientists.

2a) Methodology

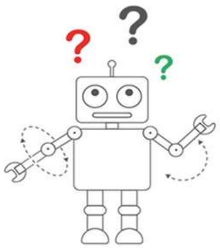
AutoML Framework

H_2O AutoML

Generative AI

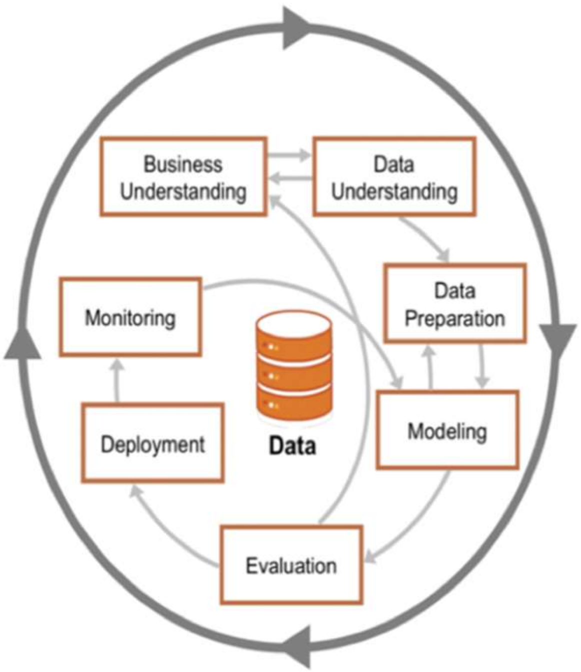
ML Algorithms

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting Machine
- Neural Networks



2a) Methodology (continued)

CRISP-DM Methodology



Assessment

Machine Learning Best Practices

- Outlier Analysis
- Exploratory Data Analysis
- Feature Engineering
- Feature Selection
- Cross-validation
- Hyperparameter Tuning
- etc.

Case Study

Credit Risk Assessment (Loan Default)

2a) Methodology (continued)

Major Software Tools

- H2O.ai AutoML
- ChatGPT Data Analyst
- Anaconda
- Jupyter Notebook
- Spyder

Frameworks

- Scikit-Learn
- Tensorflow-Keras
- Imblearn
- Shap
- Streamlit
- FastAPI
- MLflow
- etc.

2a) Methodology (continued)

German Credit Dataset

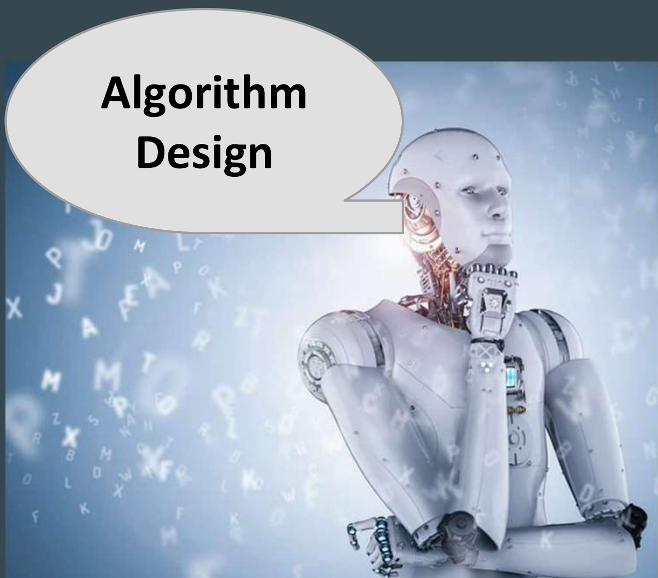
Attribute	Description	Type
1	Status of existing checking account	Categorical
2	Duration in month	Numerical
3	Credit history	Categorical
4	Purpose	Categorical
5	Credit account	Numerical
6	Savings account/bonds	Categorical
7	Present employment since	Categorical
8	Installment rate in percentage of disposable income	Numerical
9	Personal status and sex	Categorical
10	Other debtors/guarantors	Categorical
11	Present residence since	Numerical
12	Property	Categorical
13	Age	Numerical
14	Other installment plans	Categorical
15	Housing	Categorical
16	Number of existing credits at this bank	Numerical
17	Job	Categorical
18	Number of people being liable to provide maintenance for	Numerical
19	Telephone (yes/no)	Categorical
20	Foreign worker	Categorical

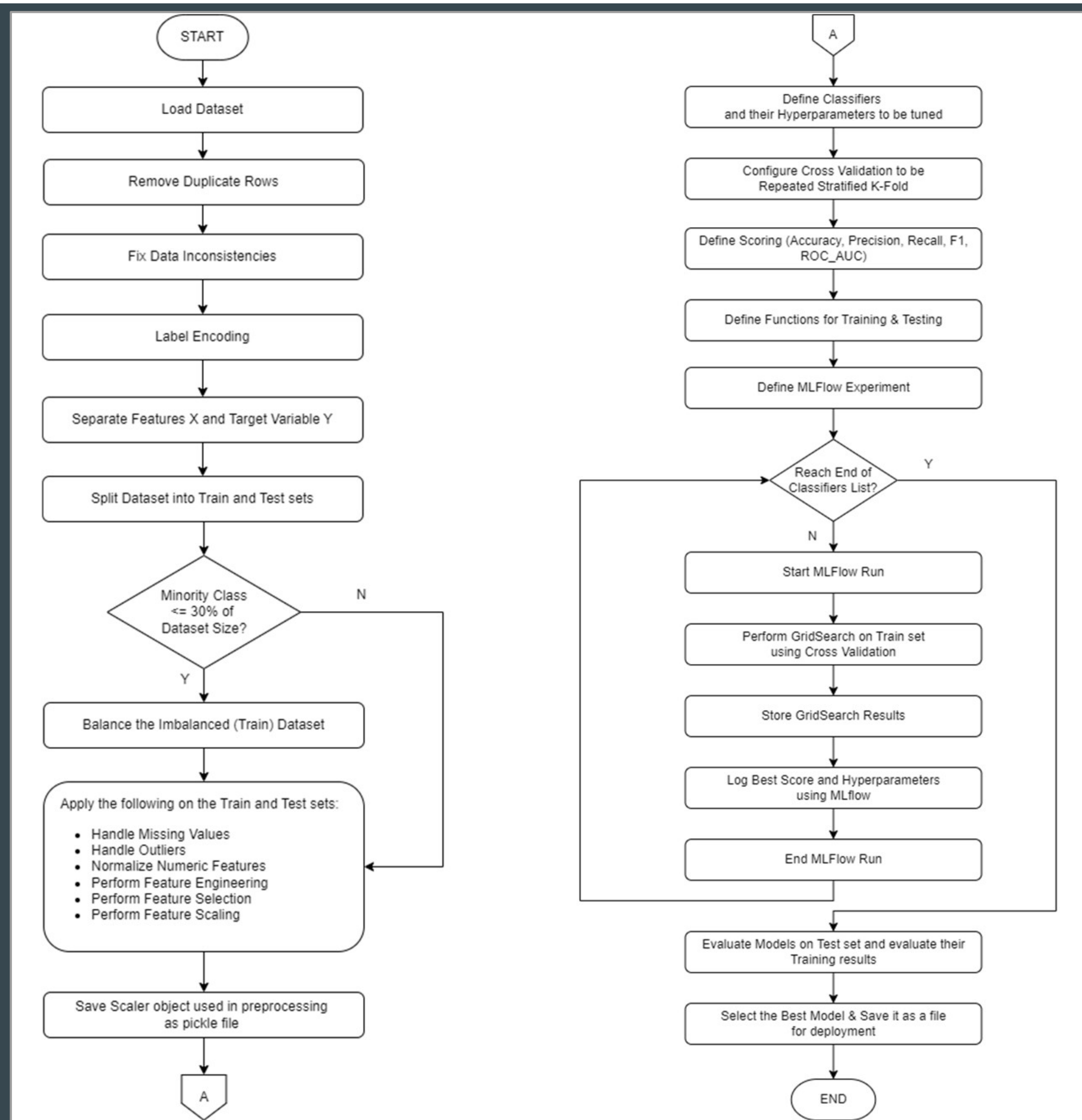
Attribute 21 Credit Risk Numerical

- 1: Good credit
- 2: Bad credit

Row Count: 1,000

2b) System Design





2b) System Design (continued)

Machine Learning Best Practices:

- Understand the Business Problem and Define Clear Objectives
- Preliminary Data Understanding
- Identify Data Quality issues
- Remove Duplicate Data
- Fix Data Inconsistencies
- Split the Dataset
- Balance the Imbalanced Dataset
- Handle Missing Values
- Handle Outliers
- Exploratory Data Analysis (EDA)
- Perform Feature Engineering

2b) System Design (continued)

Machine Learning Best Practices (continued):

- Encode Categorical Features into Numerical Values
- Perform Feature Scaling on Numeric Features
- Perform Feature Selection
- Track the Model Development Process (MLOps)
- Select Model Evaluation Metrics
- Perform Cross Validation
- Choose the Right ML Algorithm
- Perform Regularization
- Perform Hyperparameter Tuning
- Perform Ensemble Learning
- Model Interpretability and Explainability

2b) System Design (continued)

Machine Learning Best Practices (continued):

- Ensure Scalability
- Model Deployment
- Production Model Monitoring

3a) Implementation

i) Human Data Scientist Approach

- Followed the Proposed Algorithm Design in System Design
- Explored the list of ML algorithms mentioned in Methodology
- Best Model: KerasClassifier (Neural Network)

Performance Metrics	Training	Testing
Accuracy	80.15%	70.5%
Precision	79.15%	50.57%
Recall	83.56%	73.33%
F1-Score	81.07%	59.86%
ROC-AUC Score	80.08%	71.31%

3a) Implementation

ii) AutoML Approach

- Mostly followed the Proposed Algorithm Design in System Design
- Adopted programmatic approach instead of using Flow GUI
- Explored the list of AutoML algorithms available in H2O.ai (except XGBoost)
- Best Model: H2O Gradient Boosting Machine Grid

Performance Metrics	Training	Testing
Accuracy	88.05% (at 0.535711 threshold)	71.5% (at 0.5 threshold)
Precision	100% (at 0.98678 threshold)	52% (at 0.5 threshold)
Recall	100% (at 0.083427 threshold)	78.33% (at 0.3 threshold)
F1-Score	88.26% (at 0.439246 threshold)	57.78% (at 0.5 threshold)
ROC-AUC Score	89.03%	74.93%

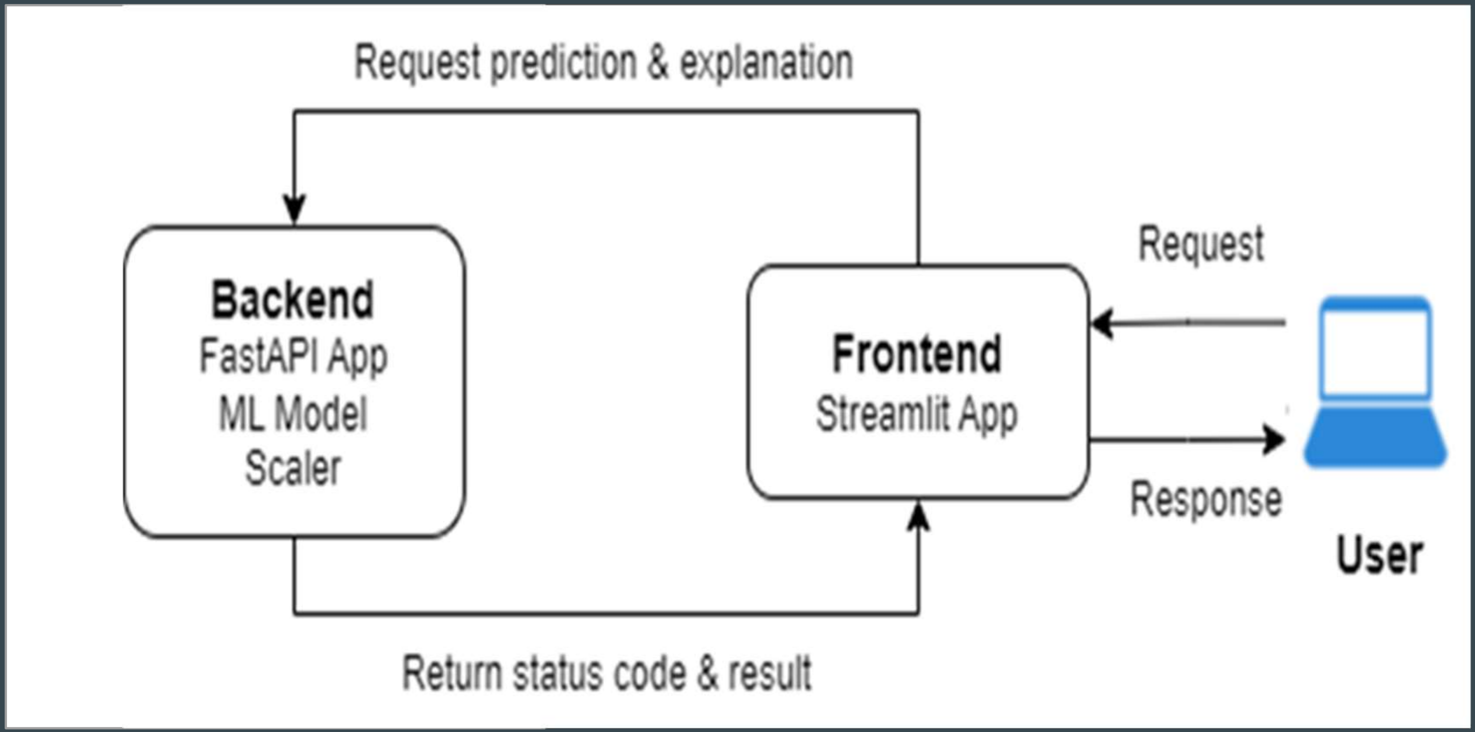
3a) Implementation

iii) Generative AI Approach

- Mostly followed the Proposed Algorithm Design in System Design
- Experimented with a few versions before deciding on the final script version
- Explored Logistic Regression, Random Forest and Gradient Boosting Machine
- Best Model: Gradient Boosting Machine

Performance Metrics	Testing
Accuracy	73.5%
Precision	77.70%
Recall	87.14%
F1-Score	82.15%
ROC-AUC Score	64.40%

3b) Deployment



4) Discussion of Results

To answer the question:
“Can Generative AI (ChatGPT) or AutoML (H2O) can replace the Human Data Scientist?”

- All 3 approaches produce reasonably good results (metrics) with more emphasis on the Recall scores (since False Negatives are to be minimized)
- All 3 approaches implemented all or most of the ML best practices successfully.
- From my development experience and results obtained, the following slides summarize the strengths and weaknesses of each of the 3 approaches.
- From this analysis, conclusions will be drawn on the question posed above.

ChatGPT Data Analyst

Strengths	Weaknesses
Ability to implement most ML Best Practices (when guided)	CDA can perform ML tasks in wrong order (needs guidance).
Ability to rapidly create new code and make code changes	CDA can omit important ML tasks (need to check code).
Generates documentation, along with its generated code	Code can go missing especially after several iterations of change.
Ability to generate visualizations for EDA and results	When code is updated, CDA may not adjust all other required changes.
Ability to provide insights into dataset, visualizations, and results	Cannot install new packages in CDA runtime environment.
Ability to provide advice on ML Best Practices and other issues	CDA runtime cannot run models except baseline ones (default params).
	CDA runtime cannot run grid search or random grid search.

Human Data Scientist

Strengths	Weaknesses
Better at human-focused activities like stakeholders engagement	Human coding speed is several orders of magnitudes slower than that of ChatGPT
Has organization-specific knowledge (process, people, laws)	Human Data Scientist's coding skills not as extensive as that of ChatGPT
May have deeper domain knowledge	Human Data Scientist's knowledge of ML not as broad as that of ChatGPT
Required to supervise CDA's work which is unreliable if unguided	
Required to organize, make decisions and integrate overall ML project	
Required to intervene when CDA and AutoML cannot proceed (e.g. exceptions)	
Human-written code does not have the same kind of reliability issue as ChatGPT's	

AutoML

Strengths	Weaknesses
Predefined parameters for Flow GUI make it easier for "Citizen Data Scientists"	No support for Business Understanding phase
Model Selection is automated and simplified with Leaderboard	Little support for Exploratory Data Analysis (EDA)
Stacked Best-in-Family Ensembles Models are automatically selected	Little automation for Data Preprocessing (programmatic)
Cross-Validation is automatically applied by default to reduce overfitting	H2O Python API coding is not more productive than normal Python coding
Built-in Model Explainability and Feature Importance	Web Flow GUI offers limited functionality
Scalable platform supporting big data and distributed, parallel, in-memory processing	Little support for automating deployment
MLOps Monitoring platform available but proprietary	

WHY CHATGPT DATA ANALYST CANNOT REPLACE THE HUMAN DATA SCIENTIST

- CDA must be supervised by a Human Data Scientist because it can make mistakes.
- Currently, the CDA runtime can only perform light processing. Anything medium or heavy must be offloaded to another platform, which requires human intervention.
- Currently, the CDA runtime does not allow new packages to be installed. This severe restriction means real-world models cannot be developed and executed in the CDA runtime.
- In matters requiring human collaboration and decision-making such as during the Business Understanding phase, CDA and in general AI cannot play the main role.

WHY AUTOML CANNOT REPLACE THE HUMAN DATA SCIENTIST

- AutoMLs were developed to play a limited role. For example, they do not provide support for the business understanding phase where the Human Data Scientist play a critical role.
- H2O AutoML has weak support for EDA (Exploratory Data Analysis) and needs to be augmented by other Python visualization frameworks (which needs human initiation).
- H2O Web Flow GUI offers limited functionality e.g. data preprocessing.
- Deployment requires significant manual coding and integrations with other frameworks and systems

5) Live Demo of the Applications