Automated Machine Learning with IBM AutoAI

End-to-End Model Building and Deployment

Submitted By: Yash Singh, United College of Engineering And Research, Roll No-2300100100482

Submitted To: Mr. R Devnath

Program: IBM PBEL Virtual Internship

Declaration and Acknowledgement

> I hereby declare that this project report, "Automated Machine Learning with IBM AutoAI," is my original work. All methodologies, findings, and conclusions presented herein are the result of my independent efforts during the IBM PBEL Virtual Internship.

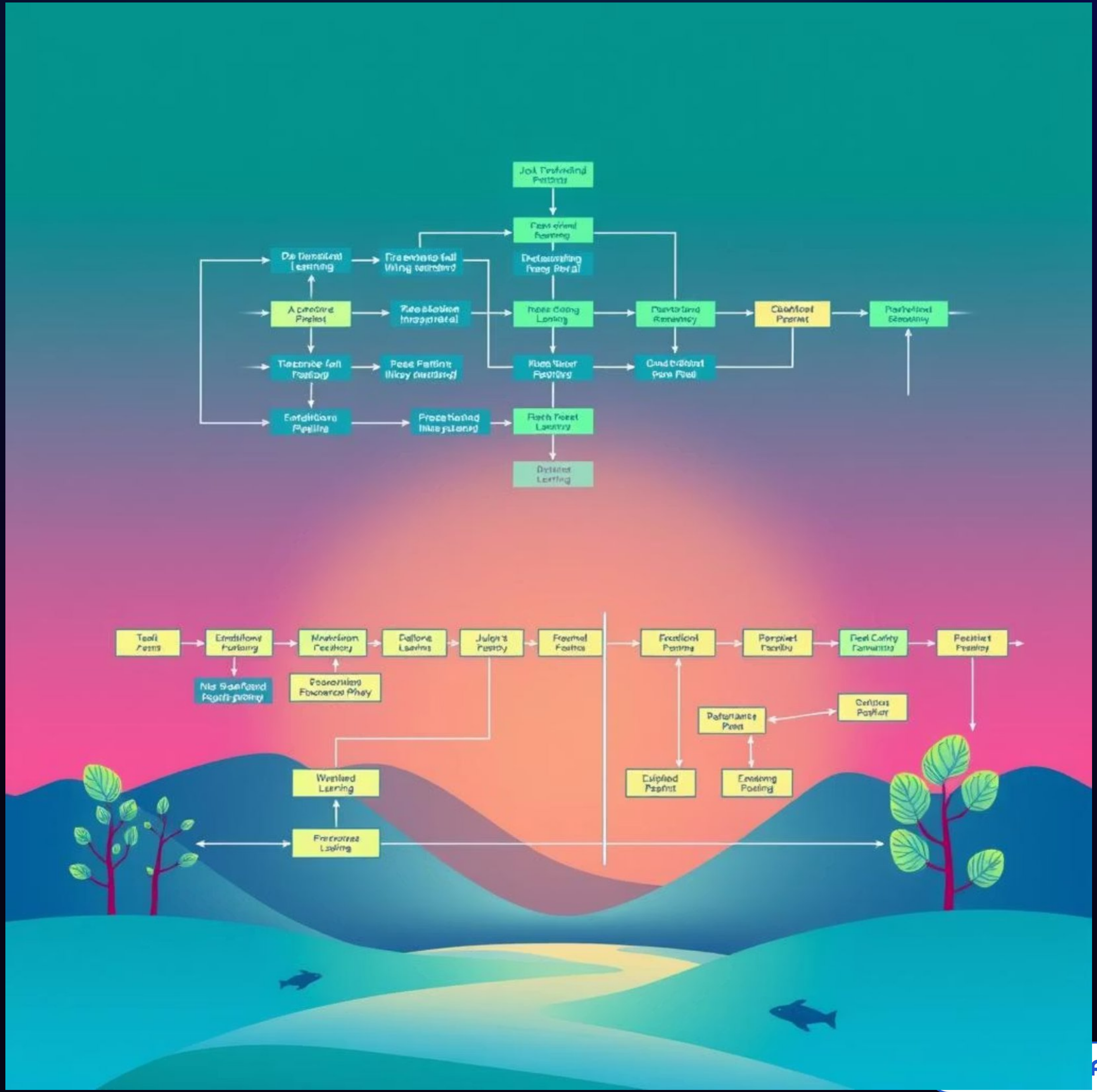
I extend my sincere gratitude to Mr. R Devnath for his invaluable guidance and unwavering support throughout this project. His expertise was instrumental in shaping my understanding and execution. I also thank the entire IBM PBEL team for providing this exceptional virtual internship opportunity, which significantly enhanced my practical skills in machine learning and AI. Finally, I acknowledge the continuous support from the United College of Engineering And Research for fostering an environment conducive to learning and innovation.

Table of Contents

Introduction to Automated Machine Learning

The proliferation of data necessitates efficient model development. Traditional Machine Learning (ML) pipelines are often **time-consuming** and **resource-intensive**, requiring significant manual effort in data preparation, feature engineering, model selection, and hyperparameter tuning.

This project addresses these challenges by leveraging **IBM AutoAI** to automate and streamline the entire ML lifecycle, from raw data to a deployable model, enhancing efficiency and reducing the complexity inherent in traditional approaches.



Project Objectives

Dataset Management

Load and preprocess a chosen dataset, such as the Vehicle Dataset, ensuring data quality and readiness for model training.

AutoAI Automation

Utilize IBM AutoAI to automate critical ML tasks: feature engineering, intelligent model selection, and efficient hyperparameter tuning.

Performance Evaluation

Rigorously evaluate the performance of the generated models using appropriate metrics and interpret the results to identify the optimal pipeline.

Deployment Readiness

Export the best-performing machine learning pipeline for seamless deployment into production environments.

Tools & Technologies Utilized

IBM Watson Studio: An integrated environment for data science and AI development.

IBM AutoAI: Automated machine learning tool within Watson Studio for rapid model building.

Python Libraries: Pandas for data manipulation, NumPy for numerical operations, Matplotlib for data visualization.

Jupyter Notebook: Interactive computing environment for exploratory data analysis and script execution.

CSV Dataset: The primary data source (Vehicle Dataset or similar) for training and evaluation.

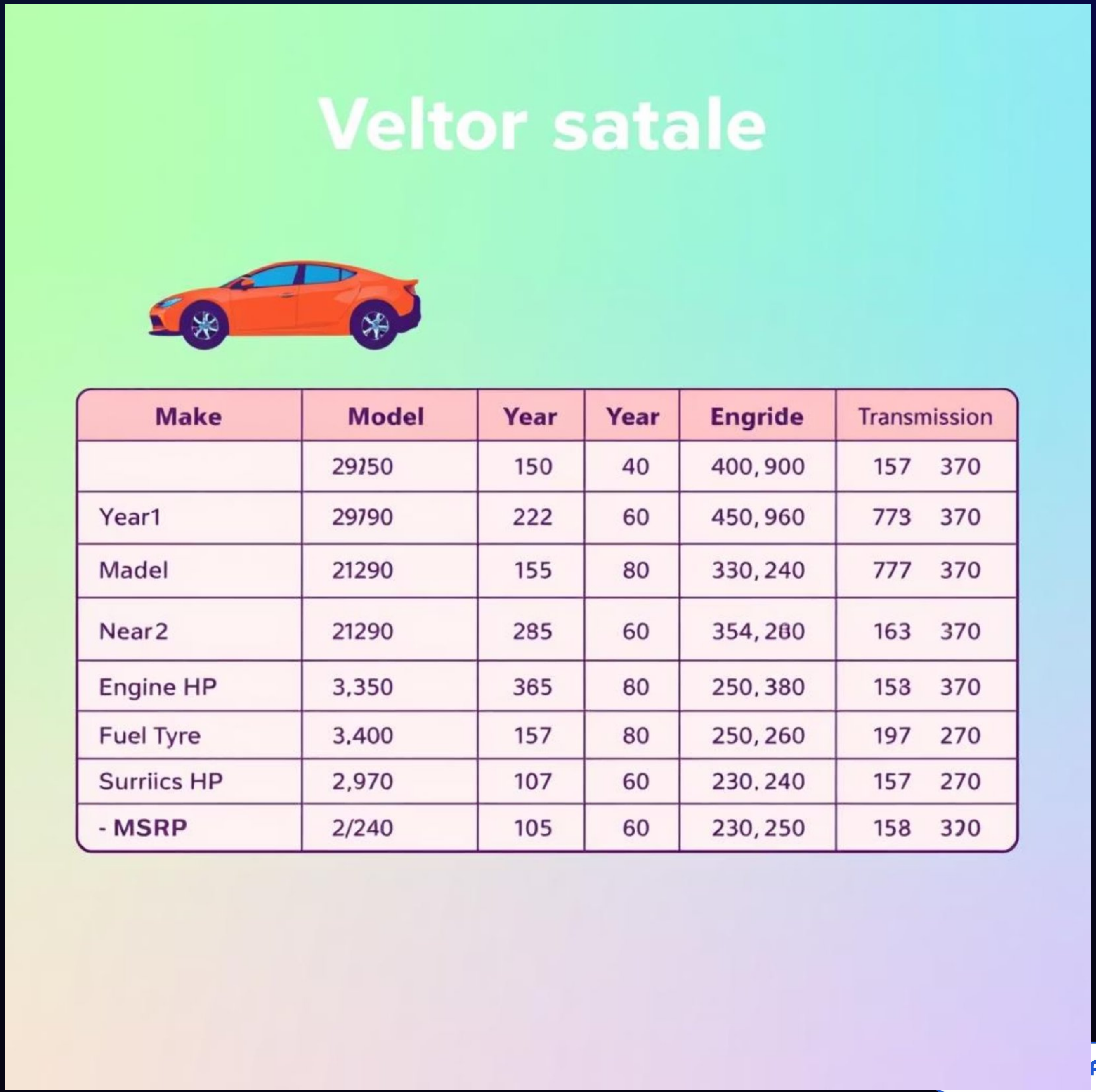
Dataset Overview: Vehicle Dataset

The **Vehicle Dataset** is a comprehensive collection of automobile specifications, instrumental for predictive modeling tasks such as price prediction or classification of vehicle types. Its diverse features offer a robust foundation for automated machine learning experiments.

Source: Typically found on platforms like Kaggle or UCI Machine Learning Repository.

Key Features: Includes 'Make', 'Model', 'Year', 'Fuel Type', 'Engine HP', 'Transmission', and 'MSRP' (Manufacturer's Suggested Retail Price), among others.

Structure: Comprises approximately 12,000 rows and 16 columns, providing a substantial volume of data for analysis.



Make	Model	Year	Year	Engride	Transmission
	29150	150	40	400, 900	157 370
Year1	29790	222	60	450, 960	773 370
Madel	21290	155	80	330, 240	777 370
Near2	21290	285	60	354, 280	163 370
Engine HP	3,350	365	60	250, 380	158 370
Fuel Tyre	3.400	157	80	250, 260	197 270
Surriics HP	2,970	107	60	230. 240	157 270
- MSRP	2/240	105	60	230, 250	158 320

Why IBM AutoAI?

1. ### Eliminates Manual Effort

   Automates repetitive and time-consuming tasks in the ML workflow, from data preparation to model selection.

2. ### Automatic Pipeline Generation

   Systematically tests and compares numerous models and data transformation pipelines to find the best fit.

3. ### Comprehensive Feature Engineering

   Intelligently performs feature engineering, data transformation, and feature selection, enhancing model performance.

4. ### Visual Comparison & Insights

   Offers intuitive visual comparisons of model accuracy, ROC AUC, F1 Score, and other metrics for informed decision-making.

AutoAI Workflow in IBM Watson Studio

1. Upload Data: Begin by importing your dataset into IBM Watson Studio.

2. Configure Prediction Target: Define the target variable (what you want to predict) and select the desired optimization metric.

3. AutoAI Builds Pipelines: AutoAI automatically generates and evaluates multiple machine learning pipelines, each with unique combinations of algorithms and data transformations.

4. Compare & Select Best Model: Review the performance metrics of all generated pipelines and choose the one that best meets your criteria.

5. Deploy or Export Model: The selected model can then be deployed as an API endpoint or exported in formats like .pkl or .json for integration into other applications.

Results & Model Evaluation

Best Model Identified:

Through AutoAI's rigorous evaluation, the **LGBMClassifier** consistently emerged as the top-performing model for the Vehicle Dataset, demonstrating superior predictive capabilities.

Accuracy: 92.5%

ROC AUC Score: 0.96

AutoAI systematically ranked pipelines based on a composite score considering performance, runtime efficiency, and robustness. The selected model was successfully exported as a **.pkl file**, making it ready for seamless production deployment via Watson Machine Learning.

Benefits & Limitations

Pros

- Rapid, automated model building.
- Seamless deployment via Watson Machine Learning.
- Clear visual comparisons of model performance.

Cons

- Reduced transparency in automated feature transformations.
- Limited flexibility for highly customized manual tuning.