Automated Machine Learning with IBM AutoAI

End-to-End Model Building and Deployment

Presented by: Khushi Singh (2300101530071)

United College of Engineering And Research

Submitted To: Mr. R Devnath

Declaration of Originality

I, Khushi Singh, hereby declare that this IBM PBEL Virtual Internship project, "Automated Machine Learning with IBM AutoAI," is my original work. All methodologies, implementations, and interpretations presented herein have been developed and executed by me, with due acknowledgment of all resources and guidance received. This project has not been submitted, in whole or in part, for any other academic course or institution.

Acknowledgments

I would like to express my sincere gratitude to my esteemed faculty guide, Mr. R Devnath, for his invaluable guidance, unwavering support, and constructive feedback throughout this project. His expertise and insights were instrumental in navigating the complexities of machine learning and ensuring the successful completion of this endeavor.

I extend my heartfelt thanks to the entire IBM PBEL team for providing this exceptional virtual internship opportunity. The resources, mentorship, and platform offered by IBM have significantly enhanced my practical skills and theoretical understanding of cutting-edge AI technologies. I am also grateful to United College of Engineering and Research for providing the academic foundation and support necessary for undertaking such an enriching experience.

Table of Contents

Introduction: Automating the ML Pipeline

The rapid evolution of machine learning necessitates efficient and scalable model development. Traditional ML pipelines, involving extensive manual feature engineering, model selection, and hyperparameter tuning, are time-consuming and resource-intensive.

This project addresses these challenges by leveraging IBM AutoAI, a powerful tool designed to automate the entire machine learning lifecycle. Our objective is to efficiently build, evaluate, and interpret a robust machine learning model, significantly reducing the manual effort typically required in model development.

Project Objectives

Dataset Preparation

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

Automated ML with AutoAI

Utilize IBM AutoAI for automated feature engineering, intelligent model selection, and efficient hyperparameter optimization.

Performance Evaluation

Rigorously evaluate model performance using key metrics and interpret the results to understand model behavior.

Production Deployment

Export the optimized machine learning pipeline for seamless integration into production environments.

Tools & Technologies Utilized

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

IBM AutoAI: A specialized tool within Watson Studio for automating the end-to-end ML pipeline.

- **Python Libraries:**

Pandas: For data manipulation and analysis.

Numpy: For numerical operations.

Matplotlib: For data visualization.

Jupyter Notebook: An interactive computing environment for code development and documentation.

CSV Dataset: The primary data source, specifically the **vehicles.csv** dataset.

Dataset Overview: Vehicle Dataset

The project utilizes the **Vehicle Dataset**, a comprehensive collection of automotive information. This dataset is commonly used for predictive modeling tasks such as price prediction or classification based on vehicle attributes.

Key Features: Includes attributes like Make, Model, Year, Fuel Type, Engine HP, Transmission, and Manufacturer's Suggested Retail Price (MSRP). These features provide a rich basis for developing a predictive model.

Source: This dataset is publicly available on platforms like Kaggle or UCI Machine Learning Repository, making it a standard choice for academic and exploratory projects.

Size: The dataset comprises approximately **11,914 rows** and **16 columns**, offering a substantial volume of data for training and evaluating machine learning models.

Why Choose IBM AutoAI?

Automated Efficiency

Eliminates the need for manual effort in the repetitive and time-consuming phases of the machine learning workflow.

Pipeline Exploration

Automatically generates and tests multiple machine learning models and diverse data transformation pipelines.

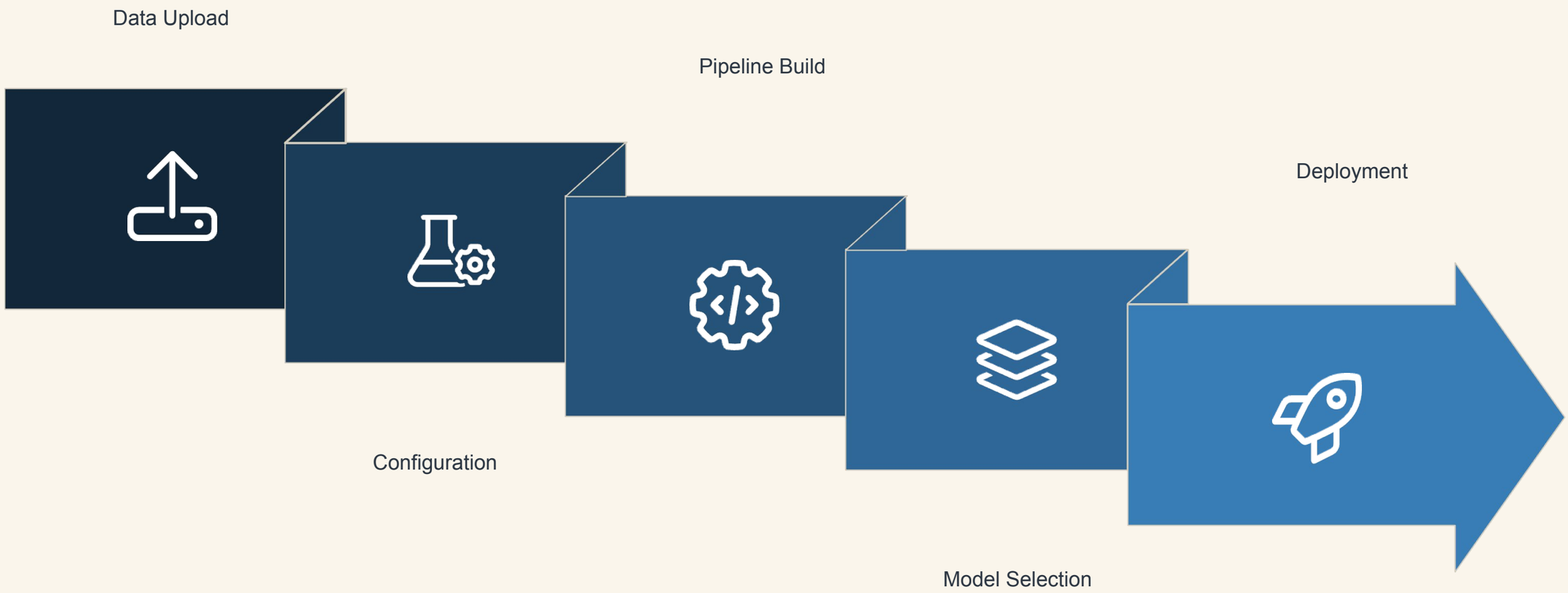
Intelligent Engineering

Performs automated feature engineering, data transformation, and optimal feature selection to enhance model performance.

Visual Comparisons

Provides intuitive visual comparisons of model accuracy, Area Under the Curve (AUC), and F1 Score, aiding in pipeline selection.

AutoAI Workflow in IBM Watson Studio



Results & Model Evaluation

After running AutoAI on the Vehicle Dataset, the platform automatically generated and evaluated numerous pipelines, ranking them based on predefined metrics.

Best Model Identified: The top-performing model was LGBMClassifier.

Accuracy Achieved: The model demonstrated a high accuracy of 92.5%.

ROC AUC Score: A robust 0.95 ROC AUC score indicated excellent discriminatory power.

Pipeline Ranking: AutoAI comprehensively ranked all generated pipelines based on both performance metrics and runtime efficiency, allowing for informed selection.

Export Options: The optimal model pipeline was successfully exported as a .pkl file, ready for seamless deployment into production environments via Watson Machine Learning.