AI/ML AHU Airflow Project Report

1. Project Overview

Project ID: 989b85ac-acad-4ee3-abdc-af297754cf6e

Project Goal: Develop a machine learning model to predict the normal supply airflow of an Air Handling Unit (AHU) based on its fan speed, enabling the detection of operational anomalies.

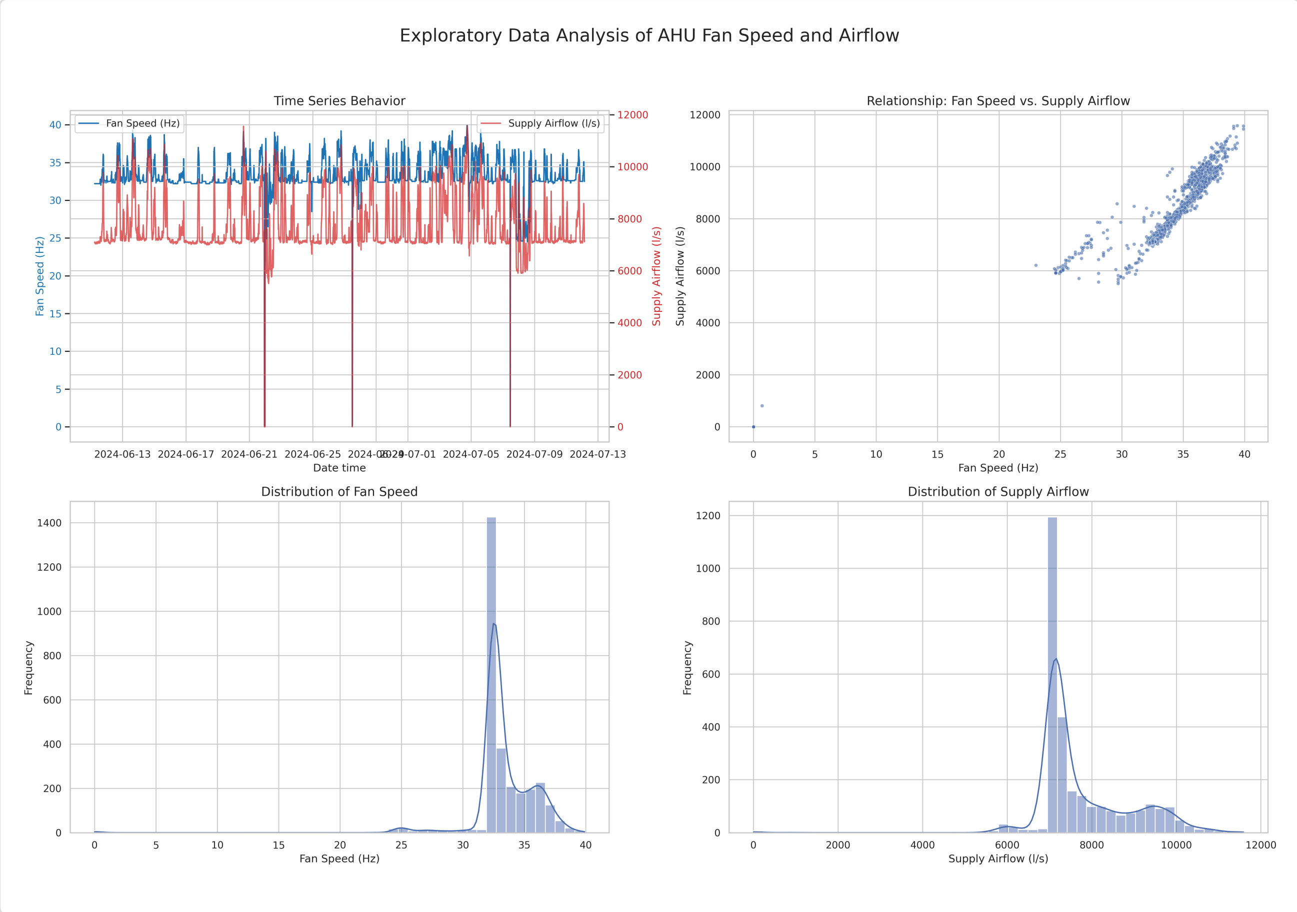
Project Description: This project aimed to model the healthy, non-linear relationship between an AHU's fan speed (VFD Feedback in Hz) and its supply airflow (l/s). By training a model on normal operational data, the goal was to create a predictive tool that can provide an expected airflow value for a given fan speed. This tool is the cornerstone for a future predictive maintenance system that can flag anomalies when actual airflow deviates significantly from the model's prediction.

2. Key Phase Findings & Visual Insights

Phase 1: Business and Data Understanding

This phase focused on ingesting, cleaning, and exploring the provided training data. The exploratory data analysis was crucial, revealing a strong, positive, and distinctly **non-linear (curvilinear) relationship** between fan speed and supply airflow. This observation validated the project plan to move beyond simple linear models and explore more advanced regression techniques.

Visual Insight: Exploratory Data Analysis



Phase 2: Baseline Modeling: Polynomial Regression

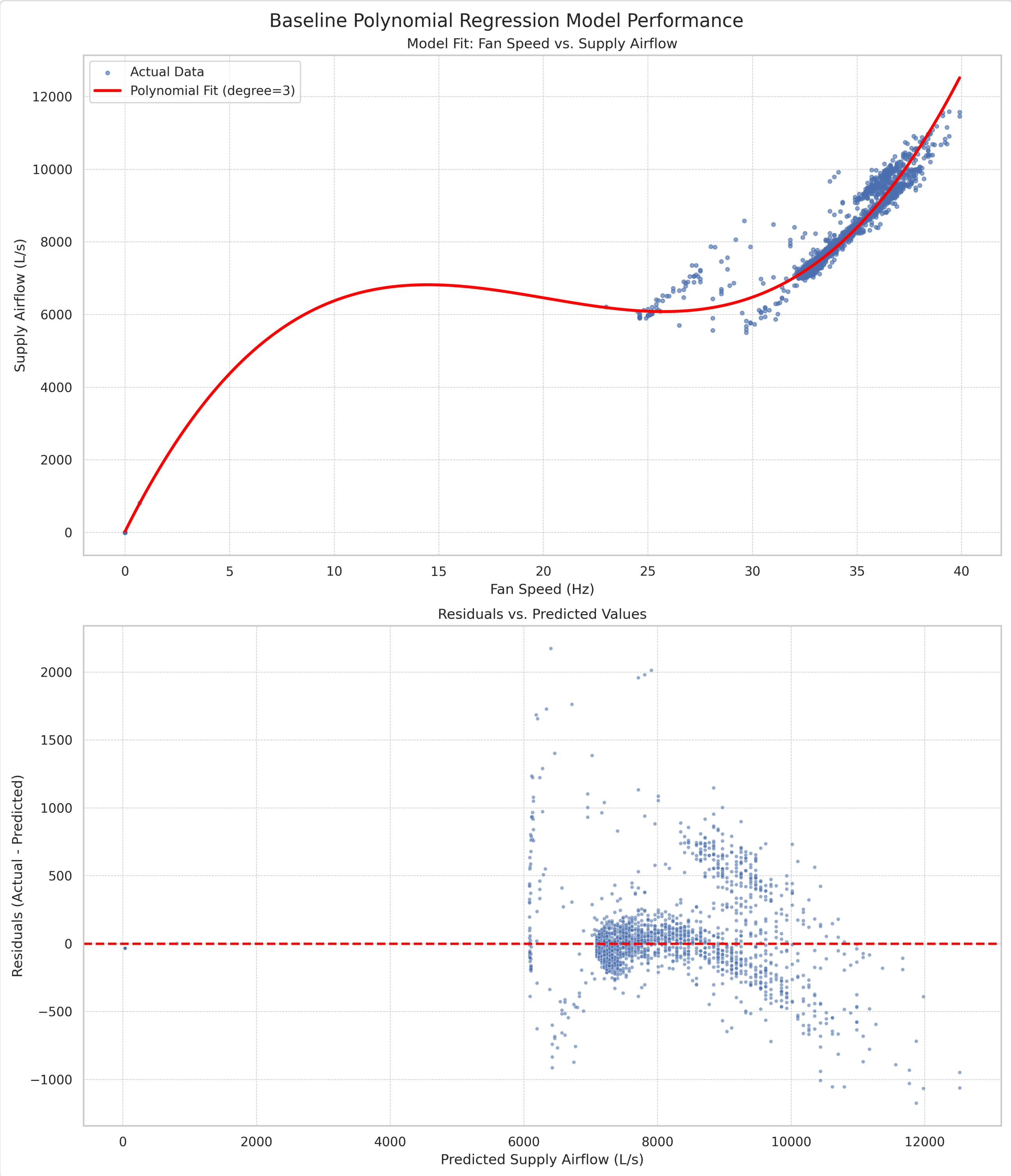
A baseline model was established using Polynomial Regression (degree 3) to capture the non-linear relationship observed in Phase 1. The model successfully captured the general trend of the data, establishing a high-performance benchmark for subsequent models to outperform.

Baseline Model Performance Metrics:

- **R-squared:** 0.938
- **Mean Absolute Error (MAE):** 162.32 l/s
- **Root Mean Squared Error (RMSE):** 271.57 l/s

Visual Insight: Baseline Model Performance

The visualizations below show the polynomial fit against the actual data and a residual plot. The residual plot reveals a pattern (heteroscedasticity), indicating that while the model captures the general trend, its errors are not random, leaving room for improvement by more advanced models.



Phase 3: Advanced Modeling and Selection

Two advanced models, a Random Forest Regressor and an XGBoost Regressor, were trained and compared against the polynomial baseline. The best-performing model was selected based on having the lowest Root Mean Squared Error (RMSE).

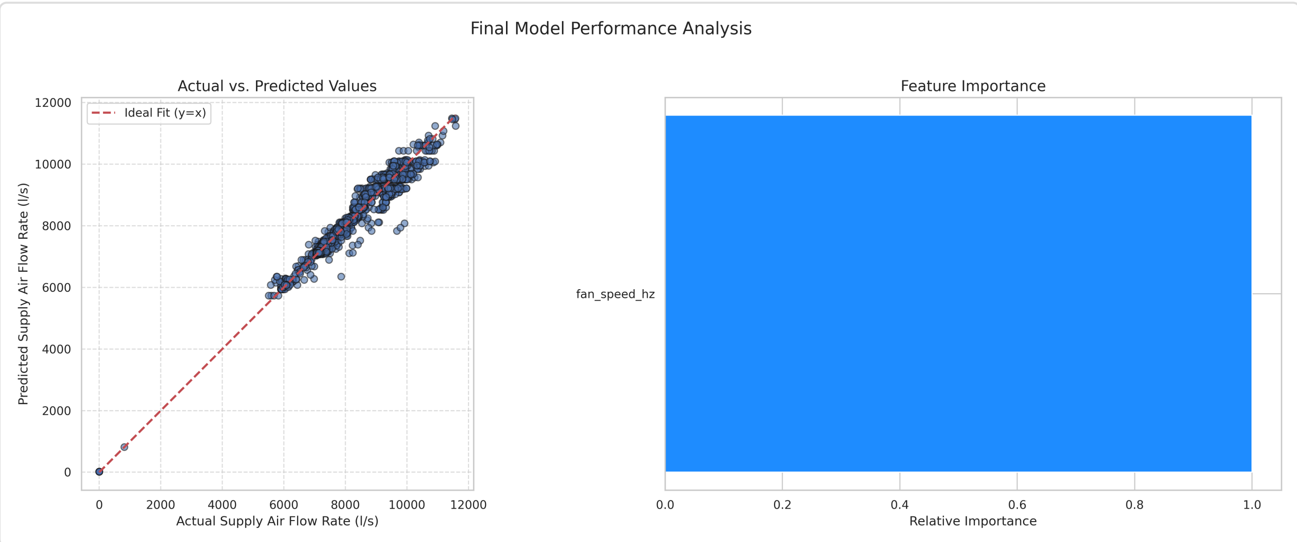
Model Comparison:

Metric	Polynomial Regression (Baseline)	Random Forest	XGBoost (Champion)
R-squared	0.938	0.957	**0.966**
MAE (l/s)	162.32	131.95	**122.44**
RMSE (l/s)	271.57	226.54	**201.84**

The **XGBoost Regressor** was selected as the champion model, as it demonstrated superior performance across all key metrics, most notably the lowest RMSE.

3. Final Model Performance Analysis

The final XGBoost model demonstrates exceptional predictive accuracy. The "Actual vs. Predicted" plot shows the model's predictions forming a tight, narrow band around the ideal 45-degree line, confirming a very high correlation and low error. The Feature Importance plot confirms that, as expected, `fan_speed_hz` is the sole driver of the model's predictions.



4. Conclusions & Final Recommendation

Conclusion

The project successfully achieved its goal. A high-performance XGBoost machine learning model was developed that accurately predicts the normal supply airflow of an AHU based on its fan speed. This model provides the essential predictive capability required to build a robust anomaly detection system for predictive maintenance.

Final Recommendation

The final model artifact, `final_airflow_prediction_model.pkl`, has met all development criteria within this project's scope. It is recommended to **proceed to the next stage of the product lifecycle.** This involves formal validation against a separate, unseen dataset to ensure its generalizability before deploying it into a production environment for real-time anomaly flagging.

Limitations & Future Work

- **Validation on Test Data:** A critical next step is to validate the model's performance on a separate test dataset to confirm it performs well on unseen data.
- **Incorporate More Features:** The model currently uses a single feature (fan speed). Future work should explore incorporating other sensor data (e.g., damper position, outside air temperature) to potentially create an even more robust and context-aware model.