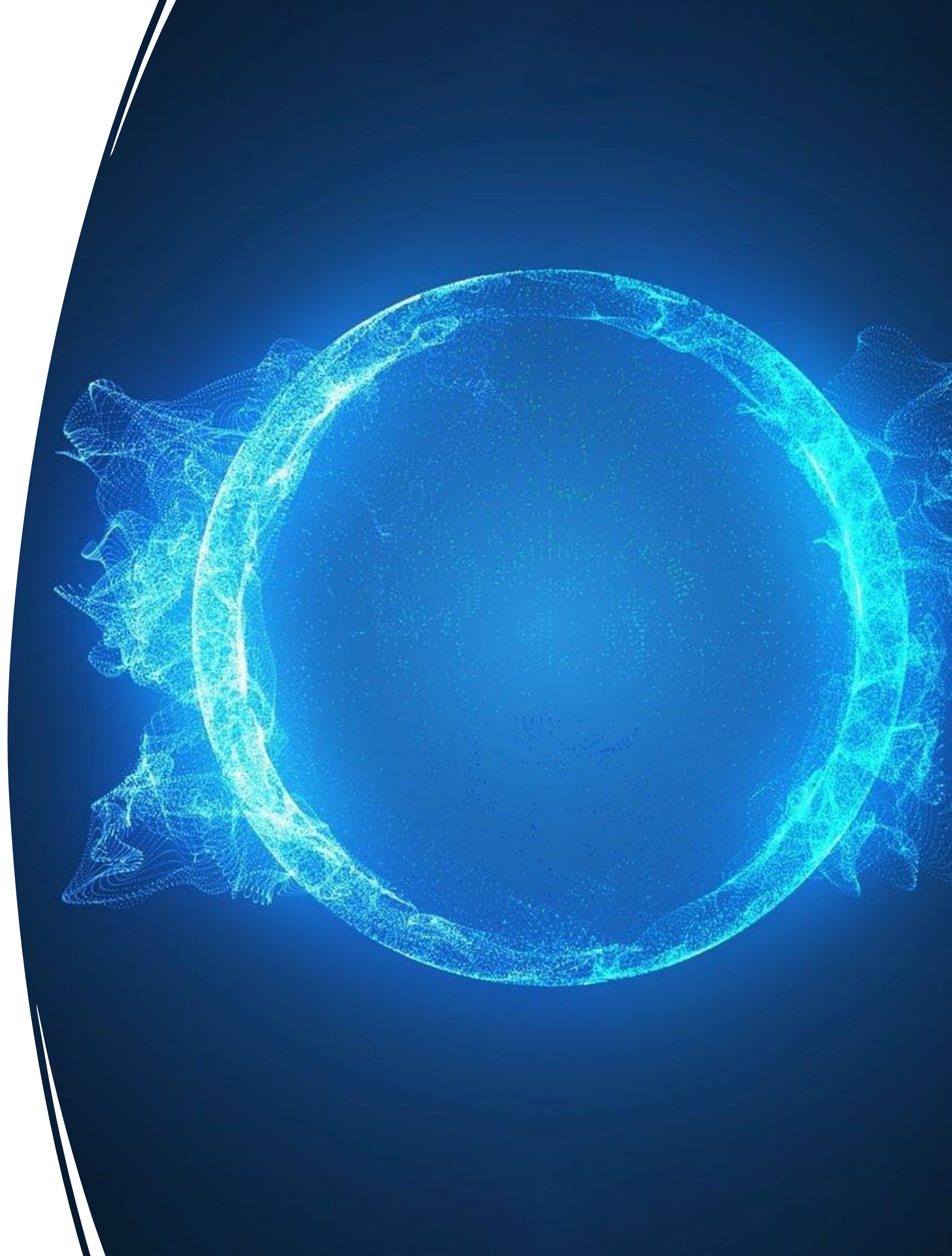
AI-Based Ballpark Quotation Tool

Checkpoint 2: Develop

Team:
Nida Ejaz
Temur Kuchkorov
Javokhirbek Parpikhodjaev



Politecnico di Torino

IVECO

Table of Contents

Checkpoint 2

Politecnico di Torino

1859

IVECO

Project Objective

- Develop an AI-based tool for fast ballpark R&D estimations
- Automatically extract key information from PR Excel files
- Use historical PR–Offer pairs as training data
- Predict **function-level** R&D effort breakdown
- Allow Customer Managers and Function Owners to refine results

Politecnico di Torino

1859

IVECO

Value Proposition

- **For the Customer Manager:**
  - **Speed:** Instant preliminary estimations replace days of manual analysis.
  - **Accuracy:** Data-driven predictions improve consistency across all quotations.
  - **Efficiency:** Drastic reduction in manual workload, allowing focus on high-value tasks

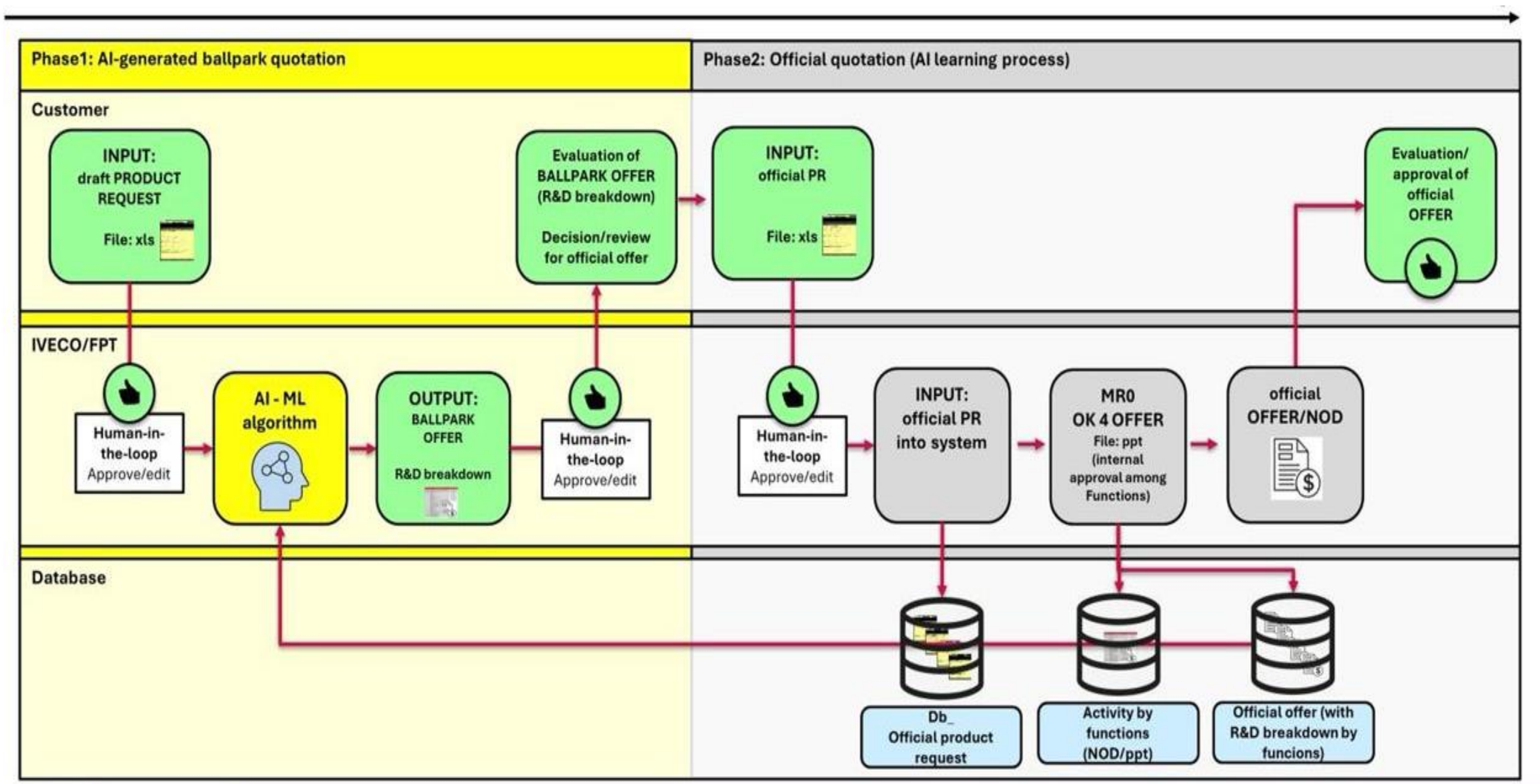
- **For the Business:**
  - **Agility:** Faster "Go/No-Go" decisions accelerate the sales cycle.
  - **Standardization:** A unified process across all functions and departments

Sustainable Development Goal

- **SDG 9: Industry, Innovation and Infrastructure**
  - *Enhancing industrial capability through digitalization and AI-driven automation.*

- **SDG 12: Responsible Consumption and Production**
  - *Reducing "Engineering Waste" by minimizing duplicated efforts and optimizing resource planning*

Functional Diagram



Politecnico di Torino
1859

IVECO

Data Overview

1/2

PRODUCT REQUEST	PR 21031 Rev G
Title: CWL New Model 100 hp	
Platform: CWL	Plant: PLANT-LE
Engine: E3F6	Tier: STAGE V

Vehicle Models:
Mac CE C.WhL

Description:
Mac CE C.WhL: E3F6, STAGE V, Boosted Curve : Rated Power : 71.9 kw @ 2200 RPM, Peak Power : 83.8 kw @ 2000 RPM, Max Torque : 453 Nm @ 1400 RPM
Mac CE C.WhL with E3F6 3.6L SCR-T 72kW Stage V engine with related compliant ATS (same concept of Mac CE T.LB in term of rating, ATS as Mac AG SPE vehicle) for Europe.

Mac CE C.WhL E3F6 3.6L SCR-T 72kW Tier4b engine with related compliant ATS (same concept of Mac CE T.LB in term of rating, ATS as Mac AG SPE vehicle) for NAFTA.

Engine controller will be ECU1. Dataset configured at 500kbps (no auto-baud rate) both for Stage V and T4B.

Replace oil Sump from Mac CE T.LB version to current CWL

Because of the necessities to have a dedicated PTO for emergency steering we need to adopt a

solution similar to SPE1 Tractors trought engine gearbox.

The change require to remove current oil fill tube and relocate it as per SSL solution with:
tube; Plug

The current oil fill tube will be plugged by COMPANY . CUSTOMER will fitup on EU units the Adaptor and pump for emergency steering

The change is required for stage V engine but the same modification can be extended to tier4B in order to manage one engine hardware.

Other change required is to relocate the relief valve currently installed on head cover where will be placed oil fill tube.

CUSTOMER officiallized investments for metal RACKs to support engines logistic transportation. Engine price should be revise consequently (no wooden pallet)

Updated Volumes + Added 1 DU

Evaluate the possibility to have one engine + ATS only, Stage V version, for EPA and ECE homologation on both EU and NA market. The volumes will remain the same

COUNTRIES SOLD TO: Europe, North America, ANZ (Australia, New Zeland)

Politecnico di Torino
1859

IVECO

Data Overview
Proposed quote
2/2

Engineering Activity Summary and R&D expenses forecast (PE.02)

PE Function		Program Main Activities Description	Effort [hrs]			K€
			Manpower	Bench	Vehicle	
Project Management		• Activity tracking and deliverable readiness	1170			100
Design	Base Design	• Release 4 new p/n of engine in PRP new drawings of engine (eVGT and WG), new flywheel, new exhaust flap orientation, new wiring harness (2), new fan pulley (1,4) ratio • Basic tech for ATS CFD, vibration and torsional analysis, verification for front PTO and calculation for 2 front end	2500 950			83
	ATS	• Installation checks of ATS and sensors ; assessment of full exhaust line, Kit ATS release, fluids analysis	840			
	EMS	• Analysis of E/E system architecture and customer requirements according to Functional Safety approach • Verification of vehicle Interface functions	1600			20
	OBD	• OBD verification according to Stage V / Tier4B, support for OBD verification on bench, support for field test	960			48
Bench	Dev & Rel	• Calibration development for top power specific rating with specific base combustion with 1 HW of engine and 1 Kit ATS OBD verification	1080	1080		460
		• E15x1 overload (gamma)	160	500		66
		• E2 (thermal shock)	600	1800		239
		• E39 test (gamma)	1100	3300		438
		• E46 test (gamma)	250	750		99
		• E75 test (gamma)	500	1500		199
		• DF test	2300	4000		812
		• Homologation tests for USA	160	160		50
		• Homologation tests for EU	240	240		80
Application		• QG readiness and dataset release, (4 rating) calibration optimization on 2 vehicles, installation and functional checks (application sign off) / field test support →mid light classification top rating, 1 light and 2 super light • Dataset release	8800			194 18
Supplier R&D		• SupplierB related functions calibration, installation and functional verification on machine for DeNox and FIE • ATS Canning skin temperature for different layout, shaker test for new top rating, release drawing and validation for new DOC and SCROF				150 160
Technical Certification		• Managing official test for Homologation activities for EU/USA (1 parent)	400			30
Materials & Travels		• Materials • travels				40 10
					TOTAL	3.296

Politecnico di Torino
1859

IVECO

Data Collection and Preparation – CSV File

pr_id	scenario_id	revision	hardware_code	market	application
21086_C	21086_C_S0	C	Engine_E9C0	EMEA;NAFTA;Korea;Japan	Mod_AGST_XL360, Mod_AGST_XL390, Mod_AGST_XL435
22111_A	22111_A_S0	A	V8	EMEA;NAFTA	Mod_AG_N11, Next Gen Customer_N11
21138_A	21138_A_S0	A	V8	EMEA;NAFTA;APAC	FRH1000, FRH1200
21086_B	21086_B_S0	B	Engine_E9C0	EMEA;NAFTA;Korea;Japan	Mod_AGST_XL360, Mod_AGST_XL390, Mod_AGST_XL435
22043_E	22043_E_S0	E	Engine_E6N0	EMEA	Mod_A.PH_T6175, Mod_A.PH_T6175, Mod_A.PH_T6175, Mod_A.PH_T6175, Mod_A.PH_T6175, CMSB_T7190, Mod_A.PH_T6175
24078_A	24078_A_S0	A	Engine_E5FC	NAFTA	Machine_CESSL_SV240
21086_A	21086_A_S0	A	Engine_E9C0	EMEA;NAFTA;Korea;Japan	Mod_AGST_C.CMXL_39, Mod_AGST_C.CMXL_40, Mod_AGSV_XL
21090_A	21090_A_S0	A	E6N0	EMEA;NAFTA;Korea;Japan	CMHD_7T340
21026_B	21026_B_S0	B	Engine_E5FC	EMEA;Turkey;Korea	Mod_CEWI_SK1, Mod_CEWI_SK2, Mod_CEWI_SK3, Mod_CEWI_SK4, SVB, TRB, Mod_CEWI_SK7
24027_B	24027_B_S0	B		NAFTA; ANZ	
23033_A	23033_A_S0	A	Engine_ES80	APAC	Rocket Edition, Mod_M.DL_5510, Mod_M.DL Rocket edition, Mod_N.DL 47 hp with Engine_ES80, Mod_N.DL with 49.5 hp
21026_A	21026_A_S0	A			
23130_A	23130_A_S0	A	Engine_E5FC	EMEA;NAFTA;Israel;Turkey;Not Regulated Countries	13ton SR
22158_A	22158_A_S0	A	Engine_E9C0	EMEA;NAFTA	Mod_CE_WhL_1, Mod_CE_WhL_2
24033_A	24033_A_S0	A		EMEA;NAFTA	
22099_A	22099_A_S0	A	Engine_E2F8	NAFTA;India	Mod_CE_T.LB_1
21110_B	21110_B_S0	B	E1C3	EMEA;NAFTA	Mod_AGFS_A825, Mod_AGFS_C890, Mod_AGFS_C990
22039_A	22039_A_S0	A	Engine_E0N0	EMEA;NAFTA;APAC;ANZ	Mod_CMSB_PM15, Mod_CMSB_T6rcpc
21062_A	21062_A_S0	A	E6N0	LATAM	Model_PM14, Model_PM15, Model_PM17, Model_PM18, Model_PM19, Model_PM20, Model_PM21, Model_PM23, Model_T7175, Model_T7190,
18094_D	18094_D_S0	D		APAC	
21026_B	21026_B_S1	B	Engine_E5FC	EMEA;Turkey;Korea	Mod_CEWI_SK1, Mod_CEWI_SK2, Mod_CEWI_SK3, Mod_CEWI_SK4, SVB, TRB, Mod_CEWI_SK7
21031_C	21031_C_S0	C	Engine_E5FC	EMEA;NAFTA;ANZ	
23131_A	23131_A_S0	A	Engine_E5FC	NAFTA;EMEA;Israel;Turkey	13ton SR
24019_A	24019_A_S0	A	Engine_E9C0	EMEA;NAFTA	Mod_CE_WhL_1, Mod_CE_WhL_2
22027_A	22027_A_S0	A	Engine_E5FC	NAFTA; ANZ; Puerto Rico; Japan	Mod_CEWI_C.234, Mod_CEWI_C.255, Mod_CEWI_C.332, Mod_CEWI_SK2, Mod_CEWI_D.L550, Mod_CEWI_L.223, Mod_CEWI_L.321, Mod_CEWI_L.328, Mod_CEWI_SK3, Mod_CEWI_S.R210B, M
22122_A	22122_A_S0	A	E6N0	APAC	Model_33, Model_34
19111_C	19111_C_S0	C	Engine_E6N0	APAC	Mod_AG_C.CMHD_O27, Mod_AG_C.CMHD_O30
22100_A	22100_A_S0	A	Engine_E2F8	APAC	Mod_CEV.C952
21031_G	21031_G_S0	G	E3F6	EMEA;NAFTA;ANZ	Mac_CE_C.WhL
23074_A	23074_A_S0	A	Engine_E6N0	GLOBAL	Mod_CE_Grad_83, Mod_CE_Grad_85
21021_B	21021_B_S0	B	Engine_E5FC	NAFTA	Mod_CE_C.WhL__5 TIER4, Mod_CE_C.WhL__6 TIER4, Mod_CE_C.WhL__7 TIER4, Mod_CE_C.WhL__8 TIER4
24084_A	24084_A_S0	A	Engine_E6N0	EMEA	Program_WLSIL
22145_A	22145_A_S0	A	NEF	NAFTA	TH6.32, TH7.32 classic, TH7.42 ELITE
24002_C	24002_C_S0	C		EMEA;NAFTA	
21132_B	21132_B_S0	B	Engine_E4N0	EMEA	Mod_CELE_TH51_6, Mod_CELE_TH51_7c, TH7.37 Plus, Mod_CELE_TH51_7e
18094_E	18094_E_S0	E			
21028_A	21028_A_S0	A	E8S0	GLOBAL	Mod_INDTRA_56S TREM 4

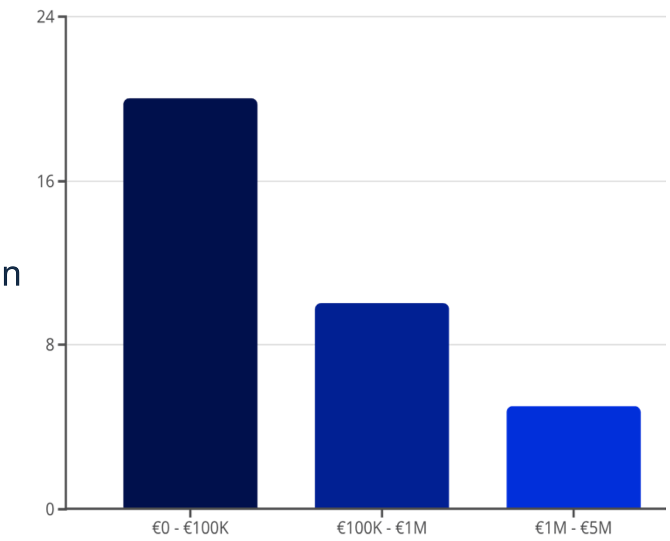
1859 Politecnico di Torino

IVECO

The Data Challenge

Key Statistics:

- Dataset: very small, 37 Cleaned Historical Projects
- Cost Range: €7K to €5M.
- Challenge: Extreme variance requires Log-Transformation



Histogram of Project Cost Distribution

Politecnico di Torino

IVECO

Research Questions

- **RQ1:** Which technical PR features have the highest correlation with R&D effort per function?
- **RQ2:** Which activities are affecting the cost of each function?
- **RQ3:** Can synthetic data improve model performance given the very limited number of historical PR–Offer pairs?
- **RQ4:** Which ML model (RF, XGBoost, Linear Models, Transformers?) gives the most interpretable & accurate predictions?
- **RQ5:** How consistent are AI-generated ballpark estimations compared to expert-generated historical offers?

Feature Engineering Strategy

Log-scale applied to Hours & Costs

Transformation

Design Ratio & Calibration Ratio

Ratios Created

One-Hot Encoding for Project Sizing

Encoding

12 Engineered Features ready for ML

Result

Engineering Features for Better Prediction

- **Handling Variance:** Applied Log-Transformation to Total Hours and Cost to stabilize the extreme range (€7k vs €5M).
- **Capturing Complexity**: Created Ratios (e.g., Calibration Ratio) because the proportion of work often dictates complexity more than just raw hours.
- **Categorical Context:** Used One-Hot Encoding for Sizing Level (Small, Mid, Full) to help the model distinguish between project scales

Politecnico di Torino
1859

IVECO

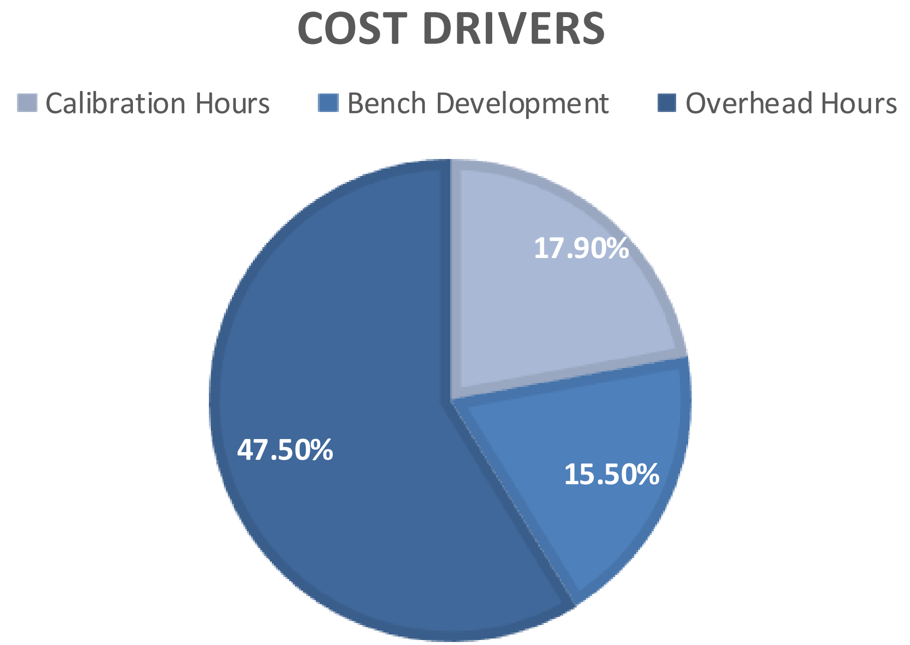
Feature Importance (Drivers of Cost)

What Drives the Cost?

Driver 1: Overhead Hours (47.5% importance) — administrative complexity is the biggest cost factor.

Driver 2: Calibration Hours (17.9%) — technical tuning is the second most critical predictor.

Driver 3: Bench Development (15.5%) — hardware testing phases significantly impact the budget.



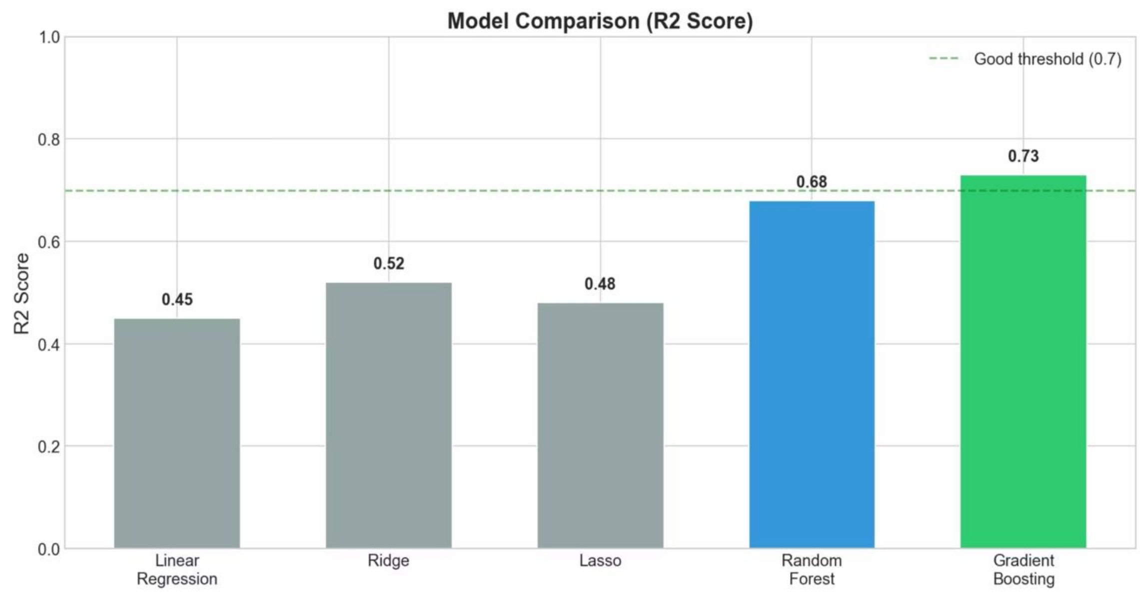
Politecnico di Torino

IVECO

Model Selection Strategy

Linear Regression Too simple for complex R&D costs	Neural Networks Overfitting risk on small data	Gradient Boosting Chosen: Robust to outliers, best for tabular data

Gradient Boosting achieves the highest R^2 Score (0.73), exceeding the 0.7 threshold for good model performance. It outperforms simpler models (Linear, Ridge, Lasso) and ensemble alternatives (Random Forest), making it the optimal choice for this complex R&D cost estimation task.



Politecnico di Torino 1859

IVECO

Validation Strategy & Model Performance

Strategy: Leave-One-Out Cross-Validation (LOO)

- **The Challenge:** With only **37 samples**, a standard 80/20 split leaves too few examples for testing, leading to unstable results.
- **The Solution:** We implemented **LOO Cross-Validation**:
  1. Train the model on **36 projects**.
  2. Test on the **1 remaining project**.
  3. Repeat **37 times** (once for each project).
- **Benefit:** This maximizes the training set while ensuring every single data point is tested effectively.

- **R^2 Score: 0.75**
  - *High Explanatory Power:* The model explains 75% of the cost variance (up from 0.08 in baseline).
- **Median Error: 34%**
  - *Acceptable Baseline:* For an early-stage "Ballpark" tool, this error margin is viable for decision support.
- **Success Rate: ~50%**
  - *Consistency:* Nearly half of all predictions fall within a tight $\pm 30\%$ error margin.

Politecnico di Torino

1859

IVECO

Approach with Synthetic Data Generation and LLM Usage

The Data Scarcity Problem

Only 37 Real Projects

We possess historical data, but the volume is drastically insufficient for robust Machine Learning training.

The Consequence

- Severe overfitting risk
- Model memorizes data instead of learning patterns
- Unreliable predictions on new quotes

Target: We need 500+ samples to generalize effectively.

Politecnico di Torino
1859

IVECO

Two Approaches Attempted In Synthetic data generation phase

APPROACH 1: CTGAN

Industry Standard (Failed)

We initially deployed a Conditional Tabular GAN (Generative Adversarial Network), a state-of-the-art deep learning model for tabular data synthesis.

Result: Lost critical business logic.
Correlation: ~0.0 (Random noise).
Outcome: Model R^2 = 32.9% (Poor).

APPROACH 2: CUSTOM ALGORITHM

Domain-Informed (Success)

We developed a custom algorithm that encodes engineering domain knowledge directly into the generation process.

Result: Preserved causal relationships.
Outcome: Model R^2 = 49.3% (Improved).

Politecnico di Torino
1859

IVECO

CTGAN Architecture

How GANs Work

CTGAN uses a game-theoretic approach where two neural networks compete:

- **Generator:** Creates fake data samples from random noise.
- **Discriminator:** Tries to distinguish between real and fake samples.

The system trains until the Discriminator can no longer tell the difference. Ideally, this captures the statistical distribution of the original dataset.

Why CTGAN Failed: Correlation Destruction

GANs focus on matching marginal distributions (the shape of individual columns) but often fail to capture conditional relationships (how columns affect each other) in small datasets.

Feature Relationship	Real Data Correlation	CTGAN Output	Status
Sizing Score → Cost	+0.500	~0.0 (Random)	LOST
Hardware Change → Cost	+0.396	+0.006	LOST
Calibration Change → Cost	+0.363	-0.026	LOST

Impact: The ML model trained on this data performed worse than random guessing on "Hours Prediction" (-3.0% R^2).

Politecnico di Torino 1859

IVECO

Solution: Correlation-Preserving Generation

- Instead of letting a neural network guess the relationships, we explicitly encoded domain knowledge as mathematical rules.

Domain Logic

Hard-coded rules derived from engineering expertise (e.g., "Hardware changes always increase cost").

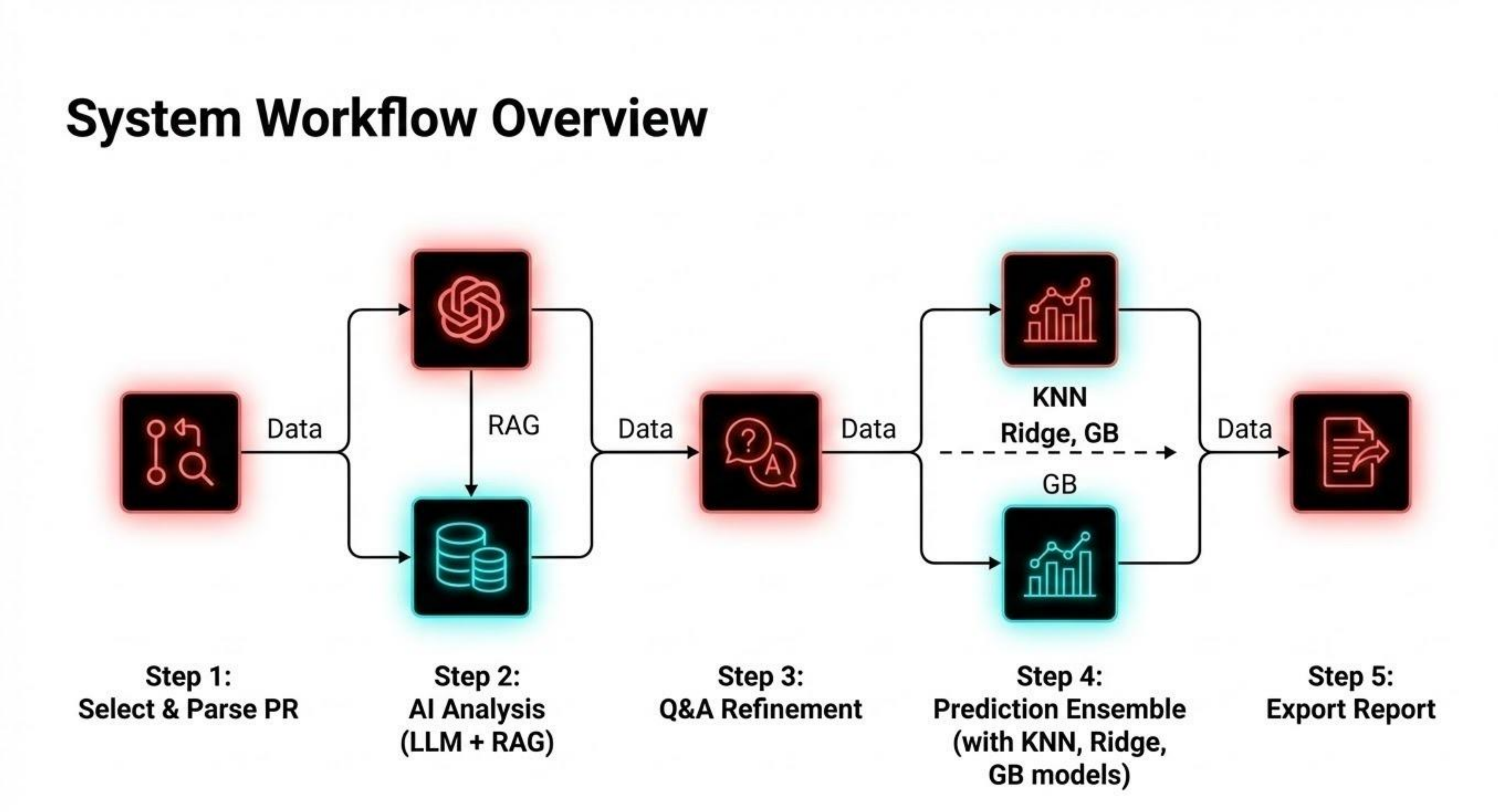
Distribution Sampling

Base costs are sampled from normal distributions specific to project sizing (Small, Medium, Large).

Causal Links

Features are generated first, then multipliers are applied to the base cost, preserving cause-and-effect.

LLM Based Pipeline and Interface for Tool



Politecnico di Torino

IVECO

Step 1: Input & Parsing

PRExcelParser Module

Processes raw **.xls** dataset files using Regex patterns to identify key identifiers.

- **ID Pattern:** PR_(\d+)_rev_([A-Z])
- **Product Families:** NEF, CURSOR, F1, EOCO
- **Target:** Extracts raw text for LLM processing.

PRDocument Dataclass

Structured data container holding extracted information.

```
@dataclass class PRDocument: pr_id: str revision: str title: str product_family: str raw_text: str # For LLM platform: str plant: str
```

Politecnico di Torino

1859

IVECO

Step 2: AI Analysis Engine

LLM Engine

Powered by **DeepSeek V3** via OpenRouter API.

Context window optimized for technical documentation reading and feature extraction.

System Prompt

"You are an expert R&D Cost Estimation Engineer at FPT Industrial..."

Embeds domain knowledge: Sizing levels (X-small to Full) and Cost Drivers.

LLMAnalysis Extraction

- Boolean Flags (Hardware/ATS change)
- Complexity Score (1-10)
- Initial Cost & Sizing Estimate
- Missing Info Identification

Step 2: Vector Store (RAG)

FPT Knowledge Base

Retrieval Augmented Generation injects domain expertise into the LLM context.

- **Database:** ChromaDB (Localhost:8000)
- **Embeddings:** all-MiniLM-L6-v2 (384-dim)
- **Content Types:**
  - Acronyms (ATS, DOC, SCR)
  - Sizing Definitions (Cost/Hours)
  - Historical Projects (N=37)

Step 3: Human-in-the-Loop

Question Generation

The LLM analyzes low-confidence areas in the initial extraction.

- Identify missing scope items.
- Clarify ambiguous requirements.
- *"Is this a new certification or carry-over?"*

Refinement Logic

incorporate_answers()

User inputs via Streamlit UI are fed back into the LLM to update the **LLMAnalysis** dataclass.

- Updates Sizing & Cost estimates.
- Updates Confidence score.
- Allows manual override of Boolean flags.

Step 4: Prediction Ensemble

KNN (40%)

K-Nearest Neighbors

Finds the most similar historical projects based on feature vectors. Best for precedent-based estimation.

Ridge (25%)

Ridge Regression

Captures linear relationships between cost drivers and total cost. Applies L2 regularization to prevent overfitting.

GB (35%)

Gradient Boosting

Handles non-linear patterns and complex interactions between features (e.g., Engine Family + ATS Tech).

Uncertainty: Conformal Prediction guarantees 90% coverage intervals.

Step 5: Excel Generation

PE.02 Forecast Report

A comprehensive 6-sheet Excel file generated using openpyxl, styled with FPT Brand colors.

- **Sheet 1:** Executive Summary (Cost, Hours, Confidence)
- **Sheet 3:** AI Understanding (Scope & Features)
- **Sheet 4:** PE.02 Breakdown (A1-D R&D Functions)
- **Sheet 6:** Similar Projects (Validation Data)

Complete Data Flow

Data travels from raw Excel input through the Parsing Module, is enriched by the Vector Store, processed by the LLM Engine, refined by User Input, predicted by the ML Ensemble, and finally exported.

Key Dataclasses:

- PRDocument
- LLMAnalysis
- FeatureVector
- PredictionResult
- ReportData

Technology Stack

Core Logic

Python 3.9+

Pandas, NumPy, Scikit-Learn

AI Model

DeepSeek V3

via OpenRouter API

Vector DB

ChromaDB

Sentence Transformers

Frontend

Streamlit

Wizard Interface

What’s Next?

- Collect more data and analyse this.
- Finalize the features.
- Test extraction accuracy
- Create first version of synthetic data
- Train baseline ML models
- Prepare internal demo for IVECO manager

THANK YOU

IVECO • GROUP

WE GO BEYOND