Automatic price prediction from CAD files

Robin Sutter[1] , Yannik Holenstein[1], Bella Fernández[2] and Bernhard Bermeitinger[2]

[1]School of Computer Science, University of St. Gallen (HSG), St. Gallen (CH)

[2]Institute of Computer Science Vorarlberg, University of St. Gallen (HSG), Dornbirn (AT)

IC3K 2026 · KDIR · Angers, France · 28-30 October 2026

University of St.Gallen
Institute for Computer Science in Vorarlberg

ROBOTUNITS®
DER AUTOMATISIERUNGSBAUKASTEN

Problem introduction and related work

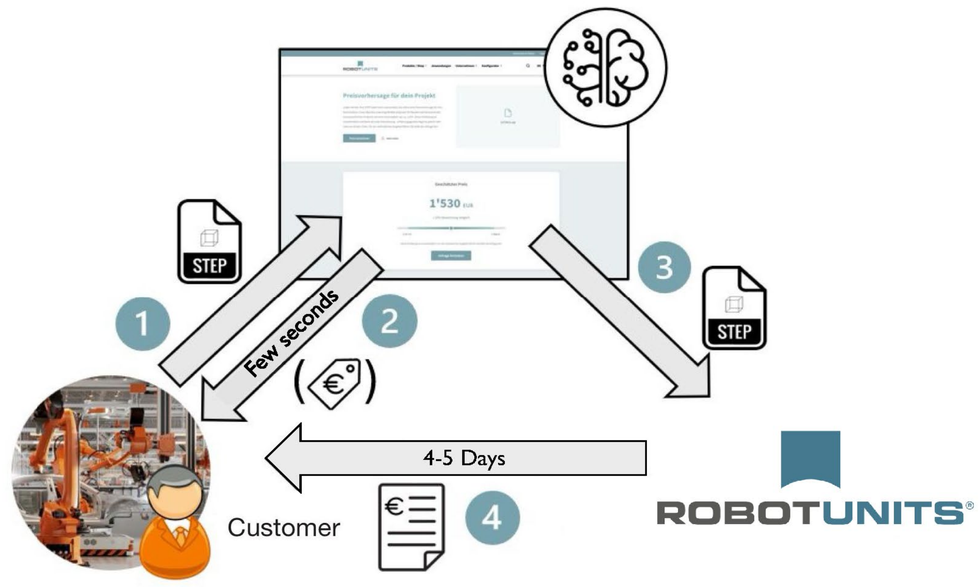
Problem: Quotation is largely a manual process that requires substantial time and professional resources.

The effort is potentially unproductive and reduces the capacity for inquiries with a higher probability of closing.

Related work:

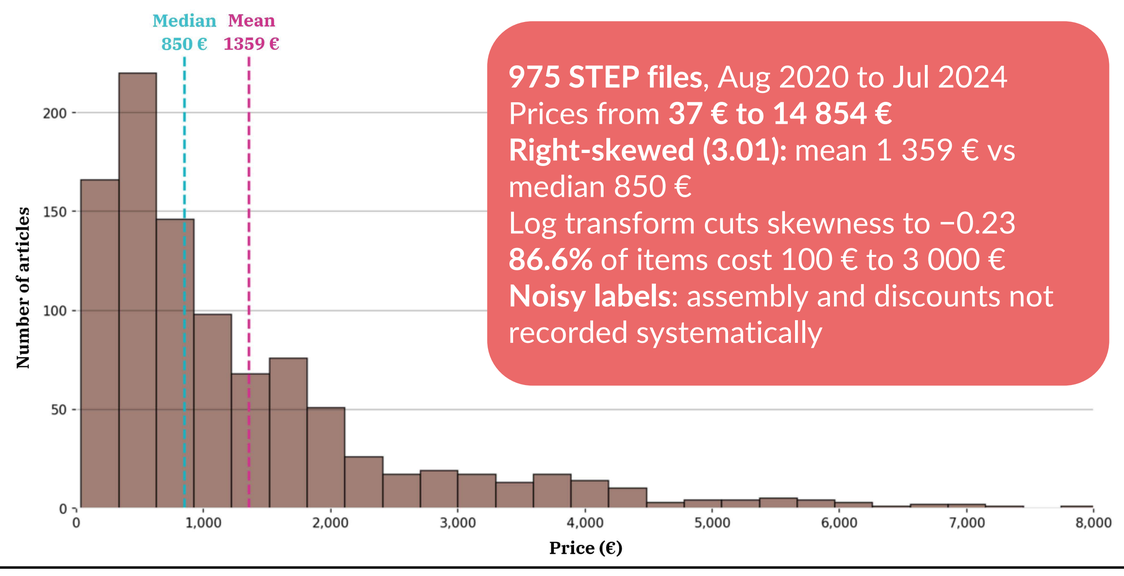
- STEP files feed **automated quotation,** but via **intermediate representations**, not direct price prediction
- Price prediction uses **regression, random forests, gradient boosting** and TabNet.
- STEP geometry needs **compact, learnable representations** before models can use it

Motivation



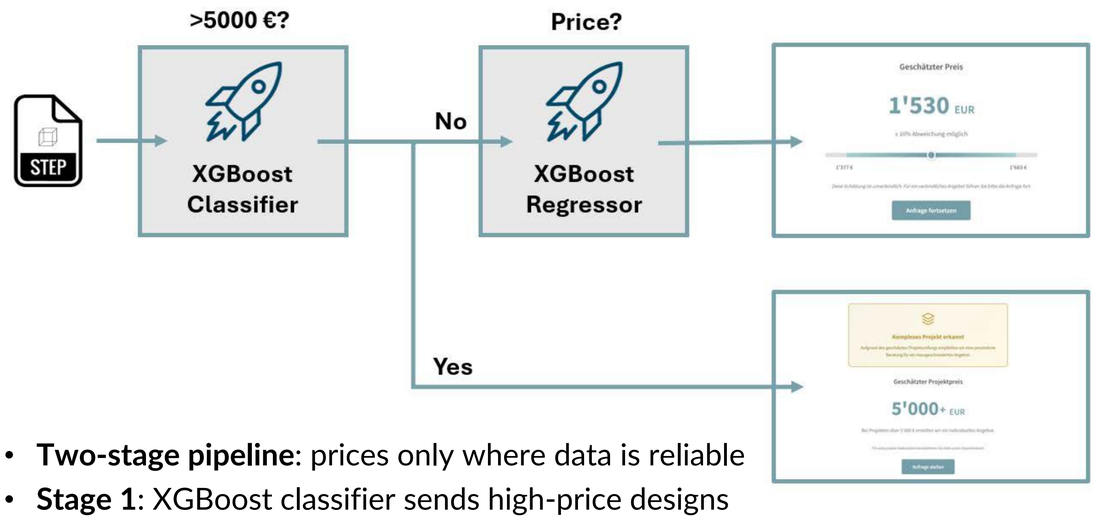
Main Goal: Users can receive an accurate price quote within minutes

Dataset



975 STEP files, Aug 2020 to Jul 2024
Prices from **37 € to 14 854 €**
Right-skewed (3.01): mean 1 359 € vs median 850 €
Log transform cuts skewness to −0.23
86.6% of items cost 100 € to 3 000 €
Noisy labels: assembly and discounts not recorded systematically

Model Architecture



- **Two-stage pipeline**: prices only where data is reliable
- **Stage 1**: XGBoost classifier sends high-price designs to sales for manual review
- **Stage 2**: XGBoost regressor prices supported designs
- **Extendable**: new designs and prices retrain the model

Modelling approach

Regression model	MAE
LSTM (First Tokens)	500 €
LSTM (Deterministic)	550 €
LSTM (Random)	600 €
Linear Regression (Mass)	407 €
Linear Regression (Multi)	559 €
Random Forest (12f)	216 €
Random Forest (Top6)	217 €
XGBoost (Top6)	215 €
XGBoost (21f)	207 €
XGBoost (Tuned)	204 €
XGBoost (Filtered)	252 €
TabNet (Baseline)	312 €
TabNet (Extended)	425 €
TabNet (Final)	207 €
Final Pipeline Regressor	**163 €**

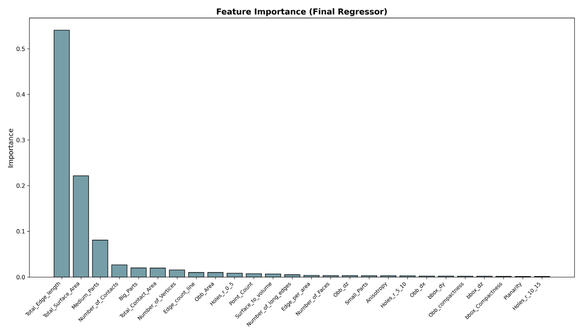
Classification	Accuracy
7 bins (fixed)	74.0 %
3 bins (quantile)	85.9 %
5 bins (quantile)	75.6 %
3 bins + 26 features	88.8 %
Binary (> 5000 €)	98.9 %
Two-stage pipeline	**MAE 218 €**

Tree-based models halve the error: MAE 407 € → 204 €

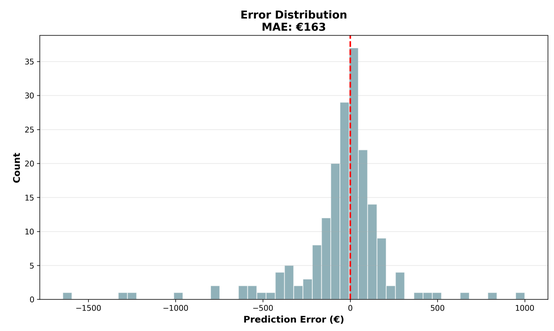
Binary classifier flags items above 5000 €: 98.9 % accuracy

Results

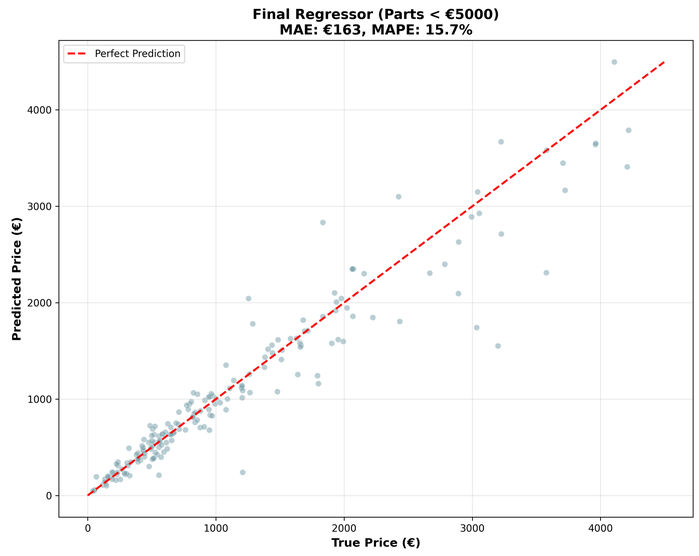
Feature importance, final regressor



Prediction error distribution



True vs predicted price (< 5000 €)



- **Predictions follow the ideal diagonal**: MAE 163 €, MAPE 15.7 %
- **Errors centred on zero**: little bias, but expensive parts tend to be underestimated
- **Edge length and surface area** dominate feature importance; few features matter

Accurate below 5000 €; spread grows for costly parts

Conclusions

Key Takeaways:

- **Final regressor: MAE 163 €**; two-stage pipeline 218 €, slightly worse
- **Tree-based models win**: XGBoost most stable, Random Forest close
- **Engineered geometric features** beat raw STEP text: LSTM ~2.5× worse
- **TabNet** matches XGBoost only after retraining (+4 %), but is less stable
- **Hybrid pipeline**: classifier filters atypical cases, regressor prices the rest

Conclusion:

- **Data quality** limits performance more than model type
- **High-priced designs** cannot be priced reliably, so it's better to do a manual review
- **Next step**: standardise price definitions (kit vs assembled, discounts)

Automated price quotes are feasible as a tool for standard designs; experts handle the rest