Preprocessing Automation with Qwen2.5

F. Messina, F. Nocella, N. Cherchi

Academic Year 2024/2025

The Challenge of Data Preprocessing

- **Data Growth**: Massive data from sources like customer interactions and IoT sensors requires streamlined management for competitive insights.
- **Importance of Data in Business**: High-quality data is a key asset, empowering decision-making, enhancing customer experience, and driving business innovation.
- **Data Quality Issues**: Raw data is often incomplete or inconsistent, demanding extensive cleaning and preparation before analysis.
- **Manual Preprocessing Challenges**: Current processes are time-consuming, requiring specialized skills to manage complex datasets effectively.

Project Goals

Automating Data Preprocessing with Qwen2.5

- **Objective**: Evaluate Qwen2.5's potential to automate preprocessing, minimizing manual intervention.
- **Expected Benefits**:
  - Reduce time and costs by automating repetitive tasks.
  - Improve data quality and consistency, enabling scalable, reliable analysis.

Designing an Application for Business Optimization

- **Model Selection**: Qwen2.5 1.5B, chosen for its speed and efficiency.
- **Two-Level Architecture**:
  - **Logical Level**: Creating a framework for scalable, reusable data preprocessing workflows.
  - **Code Generation Level**: Automating the generation of executable code for rapid deployment in business environments.
- **Validation**: Testing on datasets to ensure accuracy and reliability before business integration.

Technologies Used

Core Technologies

- **Qwen2.5**: Open-source language model for natural language processing, providing secure, scalable, and cost-effective automation solutions for enterprises.
- **Python**: With libraries such as PyTorch, enabling quick deployment and flexibility.

Mockup of the application - 1/2

Qwen2.5 *Automated preprocessing*

Speed Up Your machine learning workflow with Qwen 2.5!

Process data in just a few clicks

Planning to train machine learning models? Need to handle large volumes of data without losing hours in tedious work? Our automatic preprocessing app makes it easier than ever. With Qwen 2.5, every step is simplified, accelerating your workflow and allowing you to focus on insights and innovation rather than endless coding.

Key Benefits:

Smart Automation: Qwen 2.5 is powered by advanced algorithms to automatically transform your dataset, from basic cleaning to feature engineering.

Flexible Outcome: Take the preprocessing transformations best suited for your project: no coding required.

Time Saver: Why spend valuable time on repetitive tasks? Let Qwen 2.5 handle it for you.

Get started now and experience the power of simplicity with **Qwen 2.5**!

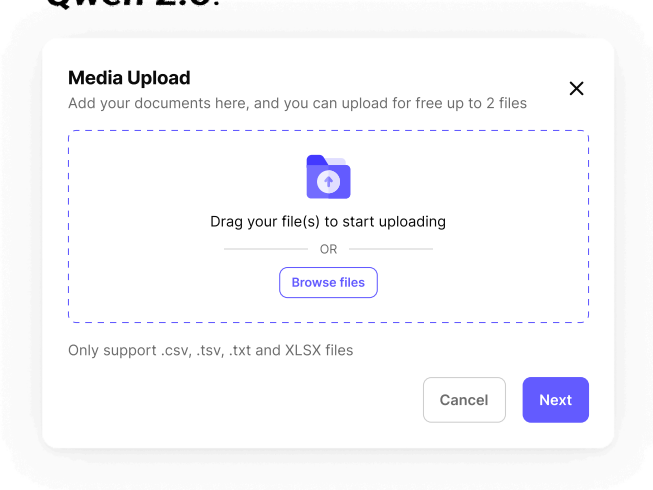


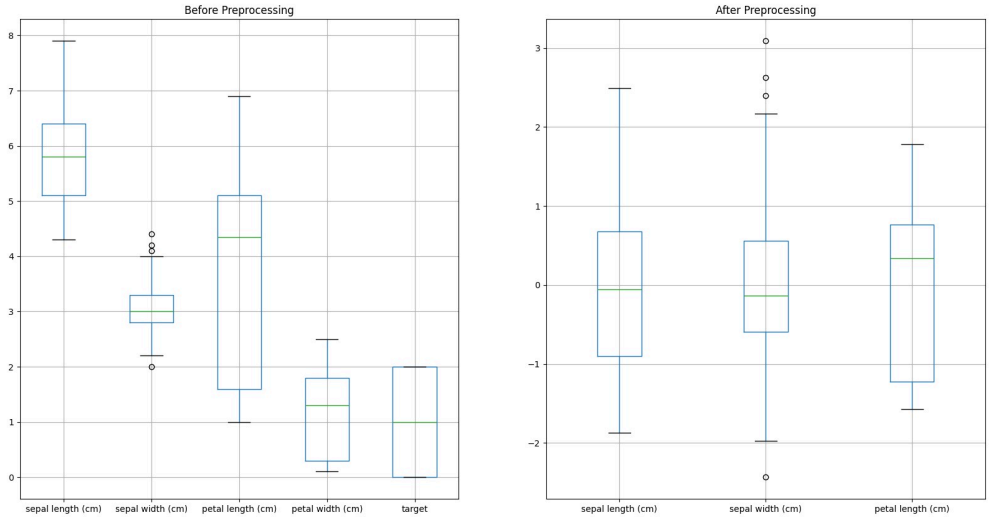
Figure 1: Main page of the Qwen2.5 application

Mockup of the application - 2/2

Qwen2.5 *Automated preprocessing*

Preprocessing Complete! Your Data is Ready!

Your data has been transformed and optimized explore the results and download with one click.



The charts on the left showcase the transformations applied, giving you a visual snapshot of your freshly prepared data. Spot trends, patterns, and insights instantly, making it easier than ever to see the value in your dataset, already processed and ready for analysis.

Download

Figure 2: Results page of the Qwen2.5 application

Results and Comparative Analysis

Key Findings and Comparative Insights

- **Performance Comparison**: Qwen2.5 vs. manual approaches and LLama.
- **Preprocessing Time Reduction**: Qwen2.5 automates tasks, achieving up to faster dataset cleaning avoiding the time spent on reasoning and decision-making.
- **Consistency and Scalability**:
  - Qwen2.5 demonstrated robust performance on standardized preprocessing tasks.
  - Limitations emerged with complex datasets, showing need for iterative improvements.
- **Future Potential**: Highlights of Qwen2.5's scalable applications and integration in larger workflows.

Preprocessing Comparison

Preprocessing Approaches

- **Comparison of Approaches**:
  - Manual preprocessing (our team, Kaggle) vs. automated approaches (Qwen2.5 and LLama3).
- **Qwen2.5 Performance**:
  - Outperforms LLama3 with higher accuracy and consistency.
  - Requires less human intervention but still trails manual methods on complex datasets.
  - Shows strong performance on small datasets (e.g., Iris), indicating potential for lightweight applications.
  - Poor performance respect to manual preprocessing on complex datasets,.

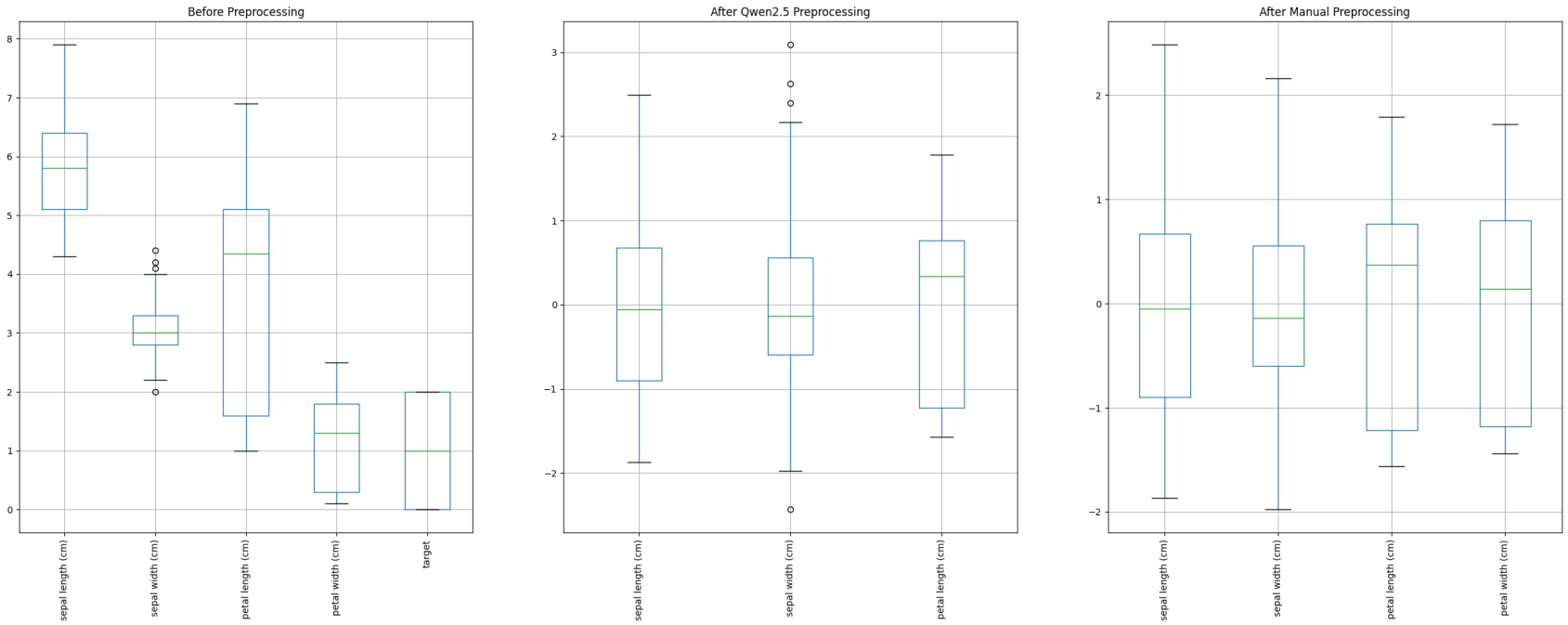


Figure 3: Preprocessing Comparison on Iris Dataset

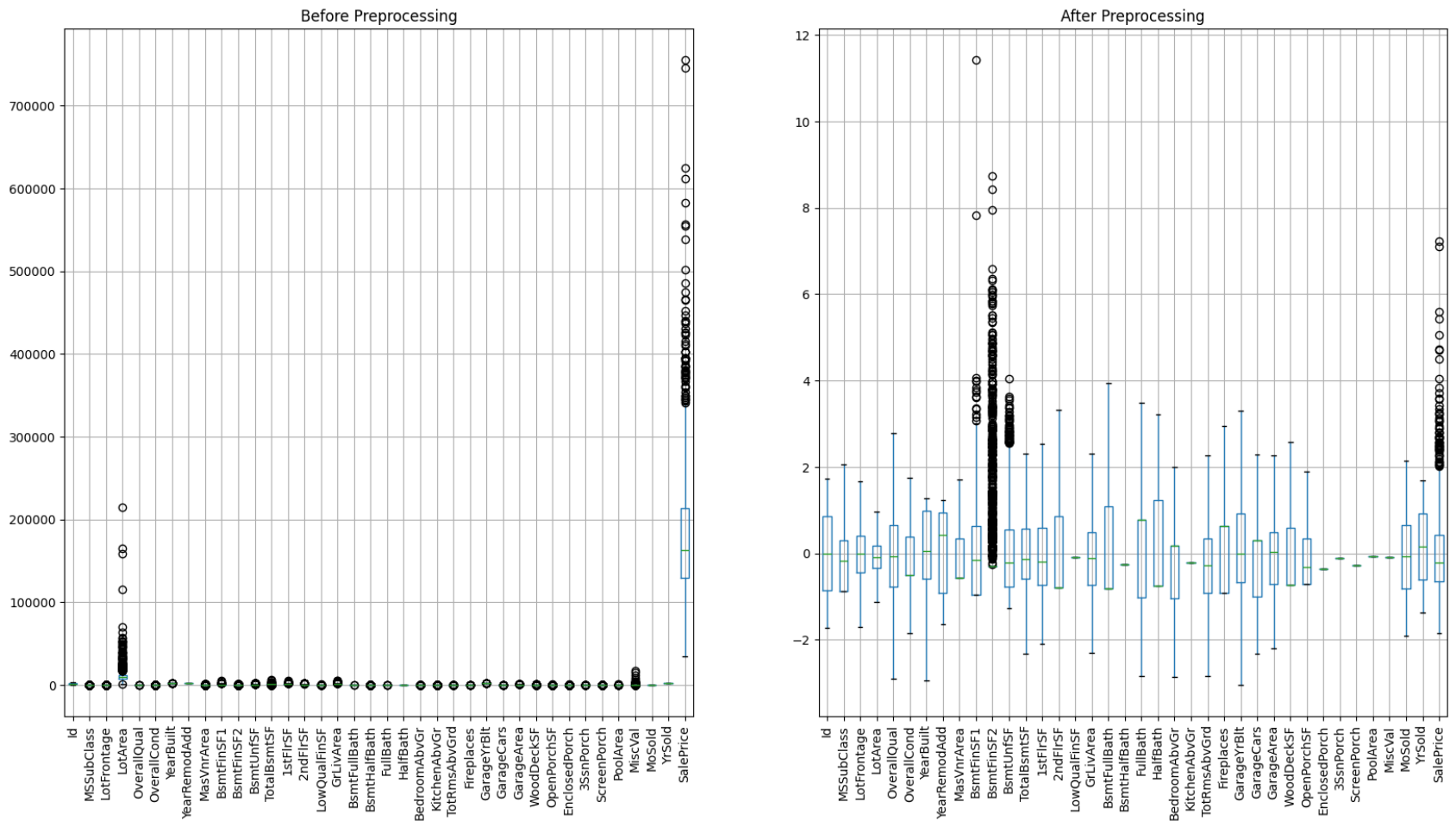


Figure 4: Preprocessing on House Prices Dataset

Conclusions

Outcomes

- The experiment demonstrated that Qwen2.5 can automate preprocessing for simple datasets, saving time and costs. However, improvements are needed for handling more complex data, as results in this area were less promising.
- It also highlighted the potential of using LLMs like Qwen2.5 for automating preprocessing, offering businesses a way to reduce the time spent on data preparation, improve efficiency, and allow data professionals to focus on more strategic tasks.

Conclusions

Future Potential

- **Next Steps for Business Integration**:
  - Fine-tuning the model to support even more complex business use cases.
  - Scaling the solution to handle larger datasets and more diverse business applications.
  - Continuous improvement through feedback loops, optimizing for business efficiency and accuracy.
  - Delevoping an iterative approach to improve the model's performance on complex datasets.
  - Integration of the automated preprocessing system into existing business workflows.

Thank You for your attention!

Feel free to ask any questions or share your feedback.

f.messina16@studenti.unipi.it
f.nocella1@studenti.unipi.it
n.cherchi@studenti.unipi.it