Machine Learning Pipeline Report: Secondhand Machinery Price Prediction

Cristian Perlog

July 16, 2025

1 Introduction

This report summarizes the results of the machine learning pipeline developed to predict secondhand machinery sale prices. The pipeline includes data preprocessing, feature engineering, and evaluation of multiple regression models. The best achieved R^2 score was 0.87 using XGBoost.

2 Data Exploration

2.1 Data Overview

The dataset contains information about secondhand machinery sales including model description, sale date, base model, and other details. Key preprocessing steps included:

- Handling missing values (dropping columns with more than 70% missing)
- Encoding categorical variables (Target Encoding for high cardinality, Label Encoding for low)
- Feature engineering (e.g., calculating machine age from year made and sale date or splitting sales date into year, month and day)

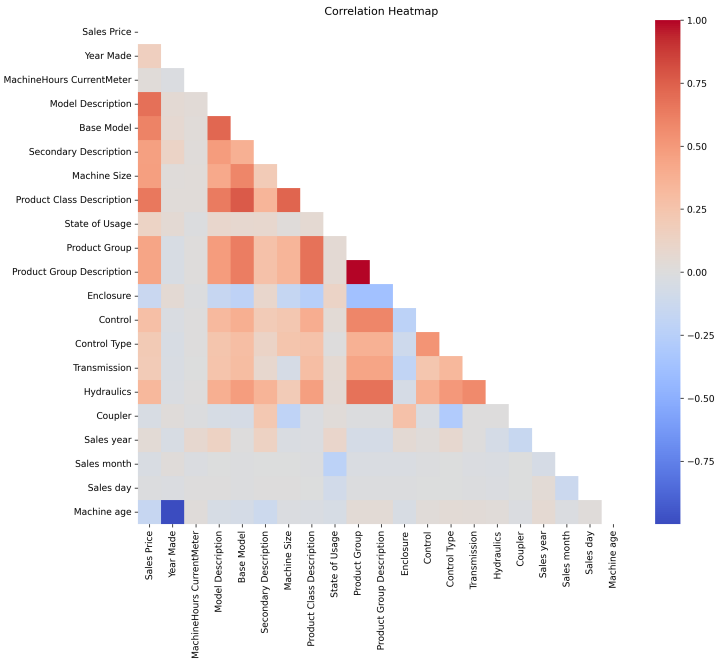

Figure 1: Correlation Between Features

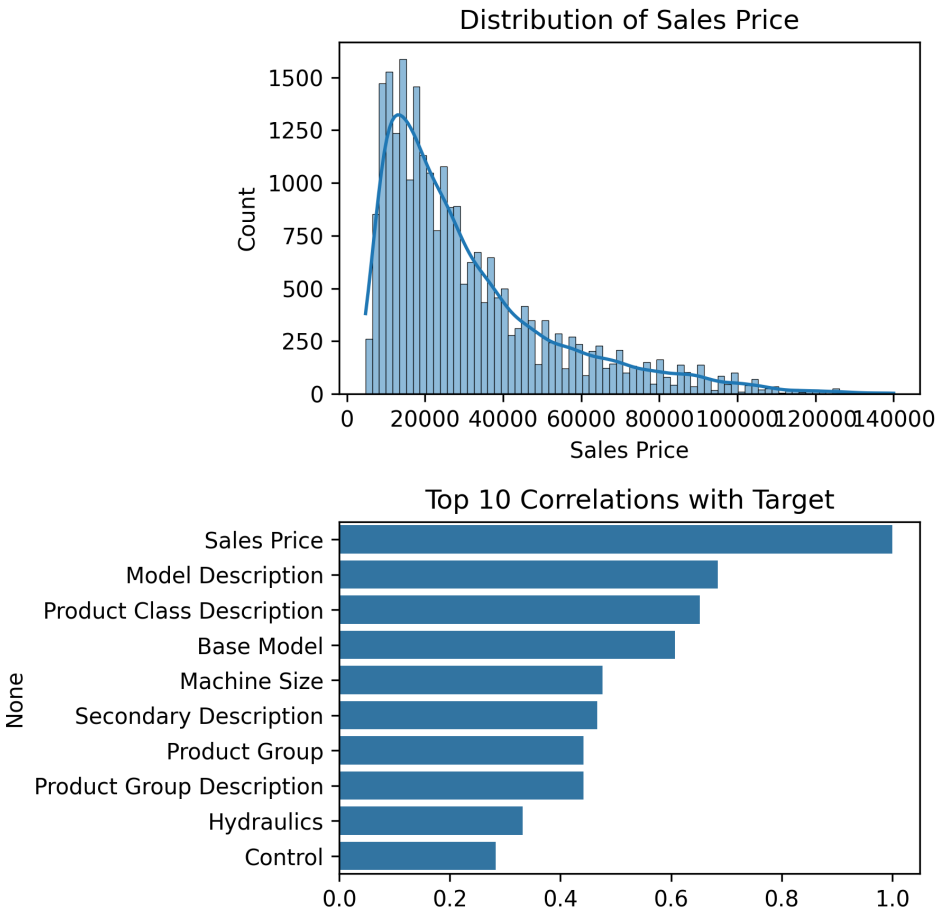

Figure 2: Exploratory Data Analysis: Target distribution and top correlations

3 Models Used

Four types of models with hyperparameter tuning were evaluated:

- **Linear Models**: Baseline (Linear, Ridge, Lasso)
- **Random Forest**: Robust to outliers with feature importance
- **Gradient Boosting**: Sequential error correction
- **XGBoost**: Advanced boosting with regularization

4 Results

4.1 Model Performance Comparison

Table 1: Model Comparison

Model	Test R^2	Test RMSE	Test MAE
XGBoost	0.871	8,133.74	5,393.81
Gradient Boosting	0.870	8,178.66	5,358.42
Random Forest	0.852	8,727.84	5,678.27
Linear Regression	0.575	14,771.56	10,362.33

Key observations from the model comparison:

- Tree-based models significantly outperformed linear models (XGBoost R^2 0.871 vs Linear Regression 0.575)
- XGBoost and Gradient Boosting showed nearly identical performance
- The best parameters show preference for deeper trees (maximum depth=5-7) and moderate learning rates (0.1)

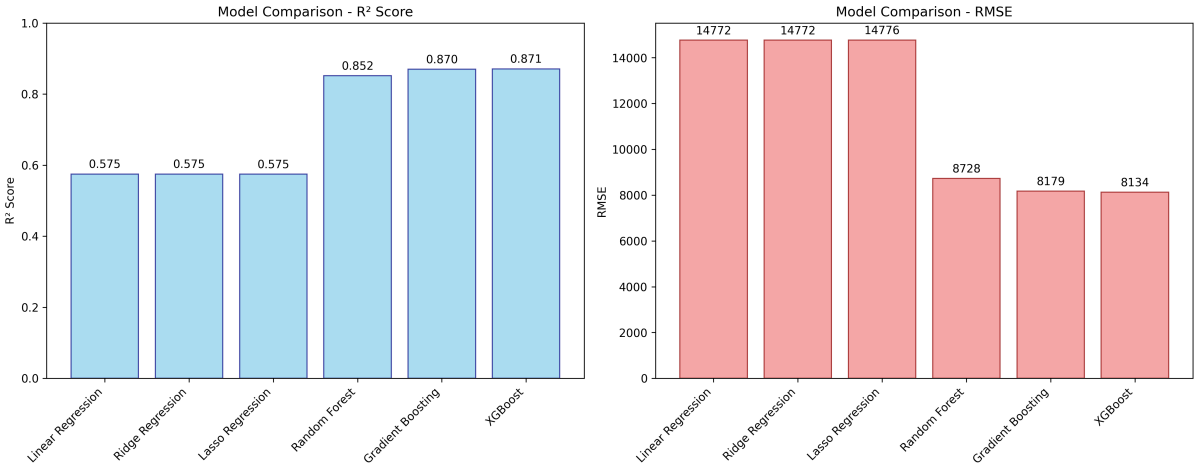

Figure 3: Model performance comparison on test set

4.2 Best Model Performance

The XGBoost model achieved the best performance with the following metrics:

Table 2: Best Model (XGBoost) Evaluation Metrics

Metric	Test Set
R^2 Score	0.8711
RMSE	8133.74
MAE	5393.81

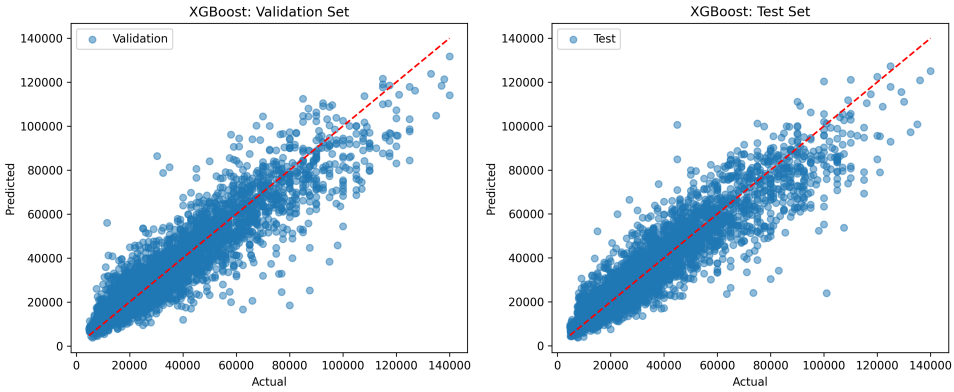

Figure 4: Visualization of the XGBoost Predictions

4.3 Feature Importance

The Random Forest feature importance analysis revealed the most predictive features:

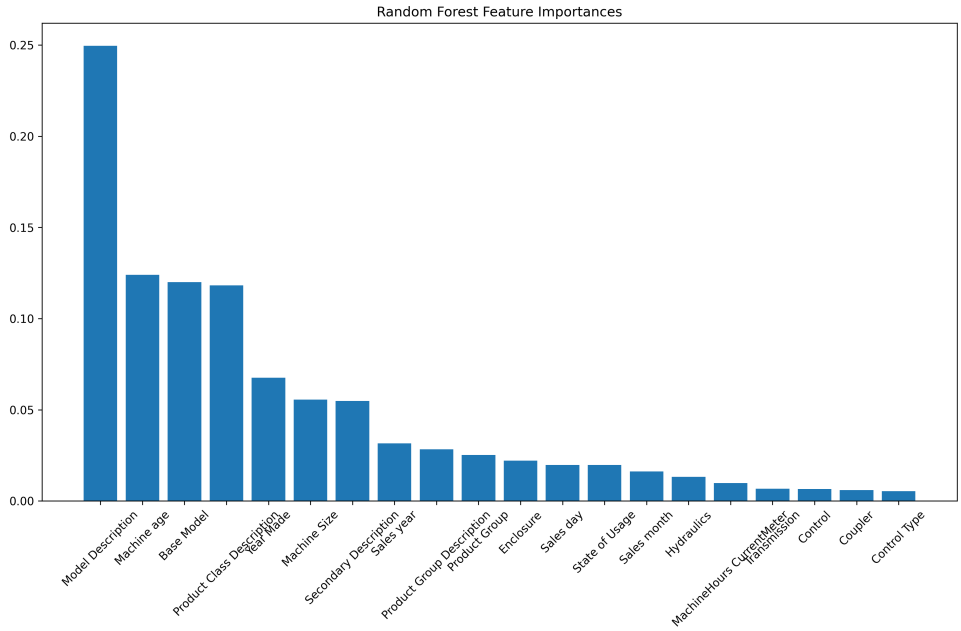

Figure 5: Top 20 important features from Random Forest

5 Conclusion

The XGBoost model demonstrated the best performance in predicting secondhand machinery prices, achieving an R^2 of 0.8711. Key findings include:

- Machine age and model description were among the most important features
- Tree-based models outperformed linear models significantly
- The model shows good generalization with similar validation and test performance