UNIVERSITÀ DI BOLOGNA

School of Engineering

Master Degree in Automation Engineering

Optimization and Machine Learning

Exam Project

Professors:

Andrea Lodi
Antonio Punzo

Students:

Andrea Alboni
Emanuele Monsellato

Academic year 2024/2025

Contents

Chapter 1

Introduction

The objective of this project is to develop a machine learning model capable of predicting the price of a used car based on a set of technical features such as mileage, year of registration, fuel type, gearbox, and others.

The overall methodology adopted in this project follows a structured machine learning pipeline, which consists of the following stages:

- Data exploration and preprocessing: inspection of the dataset, cleaning of inappropriate values, handling of duplicates and outliers, and visualization of key distributions and relationships between variables.
- Feature selection and transformation: identification of the most informative features; transformation and encoding of categorical variables; normalization of numeric attributes to avoid scale imbalances.
- Model development, evaluation, and fine-tuning: implementation of different regression algorithms such as Decision Tree, Random Forest and Gradient Boosting. Each model is trained and validated using appropriate metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score (R^2). Hyperparameter optimization is performed using Grid Search.
- Prediction and interpretation of results: the best performing model is used to predict car prices on unseen data. The results are then interpreted through quantitative analysis and visual inspection.

Chapter 2

Tasks

In this chapter, we explore the practical steps we took to get our machine learning project off the ground. We focused on refining the dataset to ensure it's high-quality and consistent, while also taking measures to prevent data leakage. Plus, we prepared the features to ensure they work effectively with our machine learning algorithms. Finally, we proceeded with model training and evaluation.

2.1 Data Exploration and Preprocessing

The initial steps in addressing any machine learning problem involve thorough data exploration and preprocessing. Examining the cars dataset, it was evident that were presents many duplicates, moreover several rows contained missing, misplaced, or clearly erroneous values. These rows have been removed from the main data frame and the ones having misplaced values have been corrected. Once the correction process was done, they were added to the cleaned data set by concatenation. This allows the model to learn from a greater amount of data leading to a better performance. Out of the 3935 removed rows, 2930 have been corrected and added. In table 2.1 is shown an example of rows that were initially removed and in table 2.2 the corrected version (city and voivodeship columns aren't shown in the tables for a better readability).

Table 2.1: Original incorrect rows to be corrected or removed.

Brand	Model	Price	Mileage	Gearbox	Engine Capacity	Fuel Type	Year
Audi	Audi RS Q3	365500.00	Elektryczny	Automatic	2023	5 km	5 km
Audi	Audi RS e-tron GT	564600.00	Benzyna	Automatic	2023	5 km	3 996 cm3

Once all features were correctly formatted and cleaned, we proceeded to plot the numeric features' distributions, figures 2.1 and 2.2, as well as the scatter plots of the cars' price as a function of the mileage, figure 2.3, and as a function of the year, figure 2.4. The scatter plots confirmed some in-

Table 2.2: Corrected row suitable for training.

Brand	Model	Price	Mileage	Gearbox	Engine Capacity	Fuel Type	Year
Audi	Audi RS e-tron GT	564600.00	5	Automatic	3996	Benzyna	2023

tuitive assumptions: cars with lower mileage or newer manufacturing years generally had higher prices. However, these plots also revealed some outliers, vehicles priced significantly higher than expected given their age and mileage. Additionally, the distribution plots, especially the one displaying vehicle prices, further confirmed the presence of outliers in the dataset.

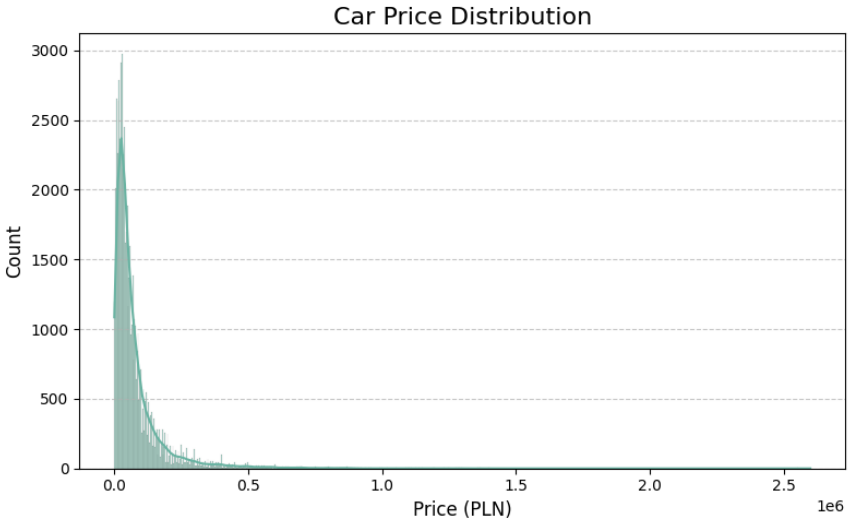


Figure 2.1: Price distribution of used cars.

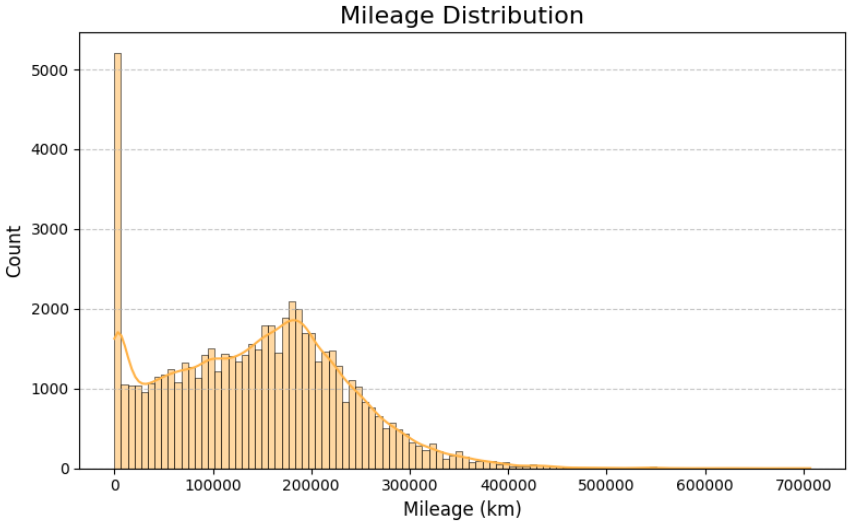


Figure 2.2: Mileage distribution of used cars.

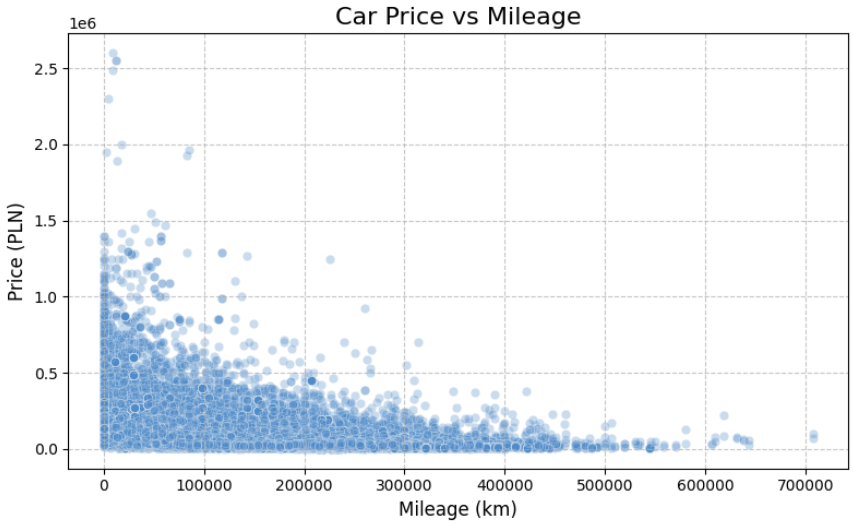


Figure 2.3: Scatter plot of price as a function of the mileage.

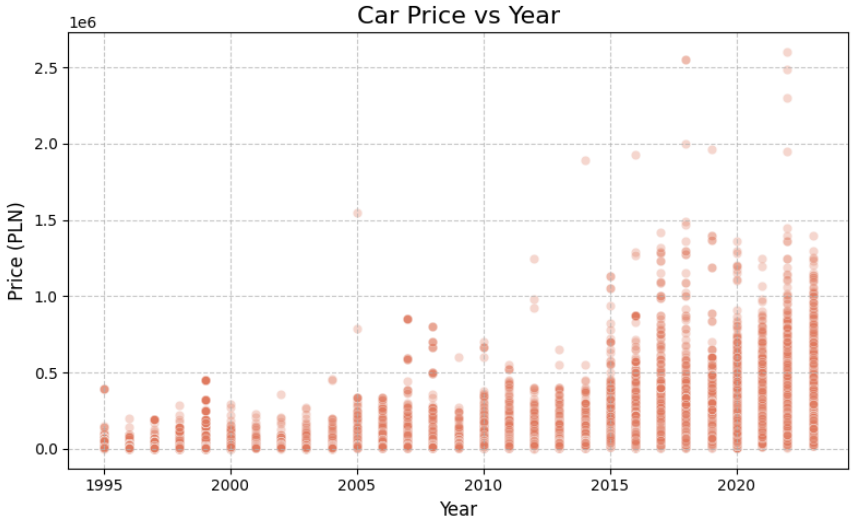


Figure 2.4: Scatter plot of price as a function of the registration year.

In order to address the outlier problem, a box plot was generated for each numeric feature to visually identify them. As shown in Figure 2.5, several cars exhibited extreme values, which can largely be attributed to luxury vehicles with significantly higher prices and engine capacities compared to the most of the dataset, which consisted of standard vehicles.

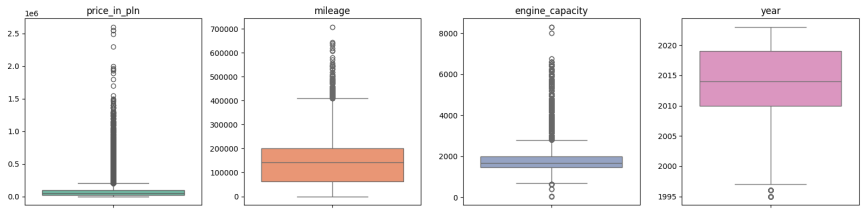


Figure 2.5: Box plots of the numeric features before outlier removal.

To mitigate the impact of these outliers and improve model robustness, we applied upper and lower bounds to each numeric variable and removed the rows containing at least one value outside the defined thresholds. This procedure was preferred to the IQR, Interquartile Range, method, which has been implemented and tested, given the knowledge of the dataset and the specific characteristics of the features. At this point, we plotted the correlation matrix to have a visual representation of the relationships between the numerical features in the dataset, figure 2.6. The elements of the correlation matrix shows that year and mileage are highly correlated, as expected, meanwhile non of the numeric features is particularly correlated to the price, target feature.

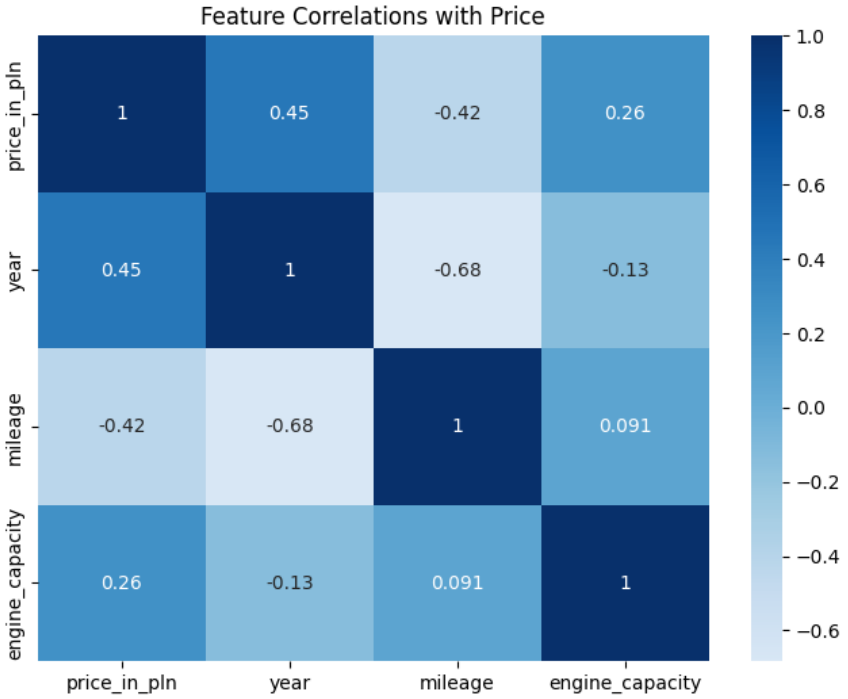


Figure 2.6: Initial correlation matrix.

2.2 Feature Selection and Engineering

The categorical features representing the city and voivodeship were excluded from the analysis, as their influence on the final car price was found to be marginal compared to more relevant attributes such as brand, model, year and mileage. Additionally, these location-based variables had high cardinality, which would have introduced unnecessary noise and complexity without improving model performance.

While working with the dataset, we noticed that the *model* feature, e.g. *Alfa Romeo Spider 2.0-16 TSpark*, actually contained a lot of information about the car such as the brand, the model, the configuration and sometimes the engine capacity and fuel type. Therefore, we engineered, from that features, two new categorical features: the actual car's model and the configuration. At this point, all the categorical variables in the dataset have been encoded using appropriate encoding techniques. In particular:

- One-Hot encoding was used for the fuel type. This transformation creates binary indicators for each category. To avoid multicollinearity, one reference category was dropped.
- Ordinal encoding was applied to both the gearbox and configuration features, as these variables carried inherent hierarchical meaning. An automatic or a car having a specific configuration will most probably cost more than the same car with a manual or with the default configuration.
- Target encoding was applied to the high-cardinality categorical features *brand* and *model*. We decided to aggregate these two features to maximize their expressiveness. For each unique combination *(brand, model)*, the mean of the target variable (car price) was calculated using only the training data. These average values were then used as encoded features. This method allows the model to capture the relationship between specific car types and their typical market values. To prevent data leakage, a fallback value, the global average price over the training data, was used when unseen combinations appeared in the validation or test sets.

Additionally, we decided to construct another feature, the age, which was calculated based on the most recent registration year in the training data. The age variable, along with mileage and engine capacity, were then normalized to ensure consistent feature scaling. In particular:

- Min-Max scaling was applied to the age and engine capacity using the training data range, ensuring that the validation and test sets remained independent.

- Log transformation was applied to the mileage to reduce the skewness and compress the range of values, making the distribution more amenable to regression modeling.

Finally, redundant variables were removed after their transformed representations were created. Then, to gain further insight into the relationships between the engineered features and the target variable, the correlation matrix in figure 2.7 was plotted. As can be seen, the features that are the most correlated to the price are in order: the engineered aggregation of brand and model, the age and the mileage.

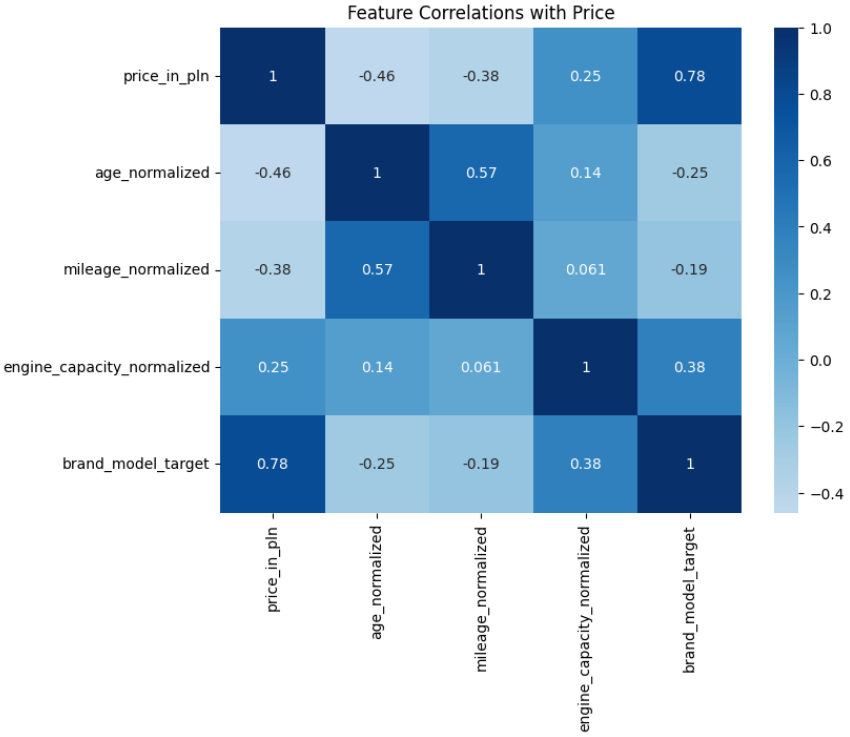


Figure 2.7: Correlation matrix.

2.3 Model Development and Evaluation

At the end of the feature selection and engineering process, we end up with eleven meaningful features, therefore, no further feature selection techniques have been applied. So, we proceeded with model definition and training.

To predict the price of used cars, we evaluated a range of models, with a focus on tree-based regression algorithms. The selected models include Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor and XGBoost Regressor. Each model is associated with a specific

hyperparameter grid, which will be used in combination with a grid search with 3-fold cross-validation. This ensures the optimization of each model's hyperparameters on the training set. The tuning and the training processes were guided by the coefficient of determination (R^2).

After fitting the model, predictions were generated on the whole unseen test set. Each model performance was evaluated using three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). These metrics provide complementary insights: MAE reflects the average prediction error, MSE penalizes larger deviations more heavily, and R^2 indicates the proportion of variance in the target variable explained by the model. The results obtained over both the training and test sets are shown in the table 2.3 and figure 2.8 shows the models predictions over 7 randomly chosen samples from the test set.

Table 2.3: Model performance on training and test sets.

Training Set			
Model	**MAE**	**MSE**	**R²**
Random Forest	18,337.69	1.22×10^9	0.911
Gradient Boosting	20,627.79	1.61×10^9	0.883
Decision Tree	20,532.68	1.67×10^9	0.878
XGBoost Regressor	21,617.46	1.90×10^9	0.862
Test Set			
Model	**MAE**	**MSE**	**R²**
Random Forest	21,650.87	2.10×10^9	0.869
Gradient Boosting	22,287.70	2.19×10^9	0.864
Decision Tree	22,518.47	2.21×10^9	0.863
XGBoost Regressor	23,269.08	2.62×10^9	0.837

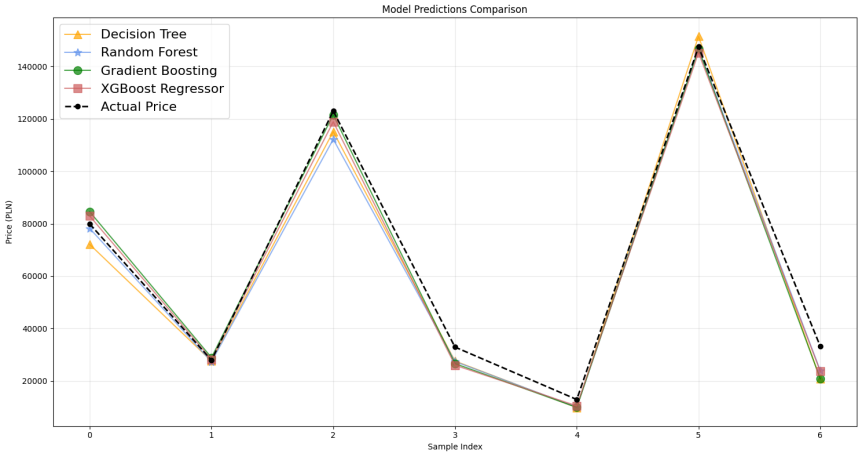


Figure 2.8: Trained models predictions on 7 randomly chosen samples.

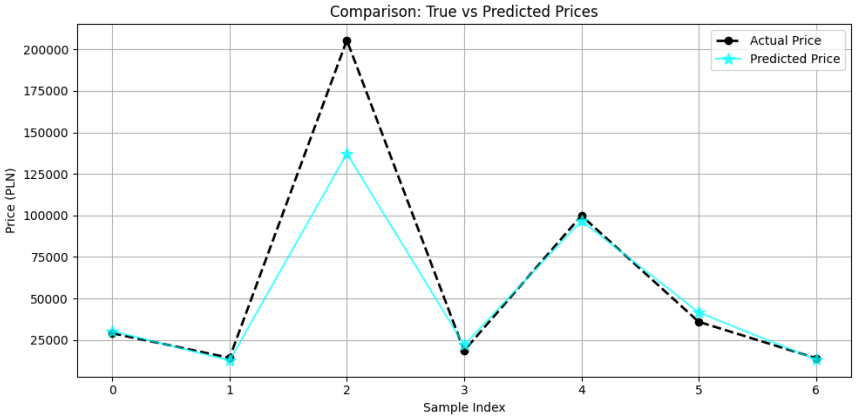
In conclusion, the best performing model is Random Forest which reaches a R^2 score of 0.869 over the test set. The Gradient Boosting followed closely obtaining a similar performance. While the overall results are not exceptionally high, the models demonstrated a consistent ability to provide reliable price predictions in the majority of cases.

2.4 Prediction and Interpretation

Once the best performing model has been obtained, we tested it on 7 randomly chosen test samples. The results are shown both numerically and visually in 2.9. The model demonstrates reasonable predictive ability across most cases, particularly for vehicles within low to mid-range price brackets, where the absolute and percentage errors remain relatively low (e.g., Samples 0, 4, and 6 with errors below 6.5%).

However, both overestimations (e.g., Samples 3 and 5) and underestimations (e.g., Samples 1 and 2) are present, indicating a lack of systematic bias but a sensitivity to variability in the unseen data. In some cases, identical or similar values for common features like mileage, engine size, and model year result in significantly different prices, which the model fails to fully capture. This likely reflects either unmodeled real-world pricing factors, such as interior/exterior condition or ownership history, or inconsistencies and noise within the dataset itself.

In conclusion, while the model exhibits some limitations, it still performs reasonably well across the majority of typical cases. The predictions for average-priced vehicles are accurate, and the model demonstrates an ability to learn meaningful patterns from the available data.



Sample ID	Real Price	Predicted Price	Absolute Error	Error (%)
0	29,000.00	30,314.34	1,314.34	4.5
1	14,200.00	12,549.90	1,650.10	11.6
2	205,600.00	137,274.84	68,325.16	33.2
3	18,500.00	21,935.88	3,435.88	18.6
4	99,900.00	96,777.64	3,122.36	3.1
5	35,900.00	41,586.56	5,686.56	15.8
6	13,999.00	13,145.78	853.22	6.1

Figure 2.9: Best model predictions on 7 randomly chosen samples, showing both numerical and visual results.

Chapter 3

Conclusions and Recomendations

The analysis highlights several limitations in both the current modeling approach and the underlying dataset, which collectively constrain the model's predictive accuracy. While the model is able to capture certain patterns, its overall performance remains limited.

Notably, the dataset doesn't include several important features that are known to significantly affect used car prices, such as visible exterior defects (e.g., cracks or dents), interior condition, the number of previous owners, and vehicle color. The absence of such factors restricts the model's ability to account for real-world variations in market value.

In addition, data quality issues further hinder model reliability. For example, many vehicles share identical values for attributes like mileage, engine size, and year of manufacture, yet are associated with different prices. These inconsistencies may reflect either input errors or unobserved factors, addressing them through more rigorous data cleaning and feature engineering is essential for improving model performance.

Moreover, incorporating more detailed information about car brands and models could enhance predictive accuracy. However, inconsistencies in how this information is recorded pose a challenge to effective integration.

In conclusion, while the current model captures some patterns in the data, its overall predictive performance remains limited. This outcome reflects the inherent complexity of used car price prediction, which depends on a wide range of interacting factors. Future improvements should focus on enriching the dataset with more relevant attributes, addressing data inconsistencies, and refining feature representations to better capture the underlying dynamics of the used car market.