Candidate number: 260

PREDICTIVE ANALYSIS OF CAR PRICES IN THE NORWEGIAN MARKET: A COMPREHENSIVE MACHINE LEARNING APPROACH

Figure 1 - Sourced from DALL-E 3 by OpenAI

CONTENTS

TABLE OF FIGURES/TABLES

1.INTRODUCTION

In this report, I explore the influence of car features like make, model, and mileage on market prices in Norway's unique automotive landscape, known for its substantial adoption of electric vehicles and unique taxation policies. The study is rooted in the use of various machine learning techniques, highlighting their potential in predicting car prices with accuracy and reliability, which is crucial for buyers, sellers, dealerships, and policymakers. Utilizing a dataset from Norwegian used car markets, a Norwegian online marketplace, the analysis involves extensive data cleaning, exploratory data analysis (EDA), and the application of advanced machine learning models such as GBM, XGBM, LGBM, and Neural Networks. The performance of these models is critically assessed using metrics like RMSE, MAE and R^2 to ensure real-world relevance, especially in the context of the Norwegian car market.

The report demonstrates the practical application of data science in the automotive industry, providing insights into market dynamics and consumer decision-making processes. It serves as a bridge between data science, machine learning, and market analysis, showcasing how data-driven methods can effectively address complex challenges in specific markets like Norway. This comprehensive analysis not only contributes to the field of statistical learning but also offers valuable insights for market analysts and policymakers in understanding and predicting car prices.

LITERATURE REVIEW

Impact of Data Preprocessing and Feature Engineering:

A critical yet often understated part in the development of predictive models is data preprocessing and feature engineering. Techniques such as categorical encoding, feature scaling, and handling missing data are essential in optimizing the model's performance. This is confirmed by (Felix, 2019), who emphasized the significance of accurate model inputs and data quality in predictive analytics.

CHALLENGES AND OPPORTUNITIES IN PREDICTION MODELS:

Car price prediction models, especially when using machine learning (ML), face the challenge of model overfitting. This occurs when a model performs well on training data but poorly on unseen data. Regularization and hyperparameter tuning are common strategies to minimize this, as highlighted in recent studies (Shaprapawad, 2023). These challenges open possibilities for further research, especially in improving model generalizability and accuracy.

MACHINE LEARNING TECHNIQUES IN CAR PRICE PREDICTION:

Various ML techniques such as Lasso, decision trees, XGBM (XGBoost), and SVR have been applied in car price prediction, each with distinct merits and limitations. The decision to not utilize SVR in this study stems from its computational complexity and potential overfitting issues in a data-rich environment like car price prediction. The most effective models identified in literature include Random Forest, Gradient Boosting Machines (GBM) variants, which helped in the model selection for this study (Huang, 2022) (Thai, 2019)

SUSTAINABILITY AND ECONOMIC IMPACTS:

The sustainability aspect, particularly in the used car market, is increasingly gaining attention. Predictive models aid in detecting market trends, aligning decisions with economic and environmental sustainability. This aligns with the observations by (Yang, 2023) regarding the shift in consumer awareness towards ecological footprints.

2. METHODS AND DATA

DATA

The selected dataset from Norwegian used car markets offers a thorough view of the Norwegian used car market, encompassing key details such as price, make, model, mileage, and technical specifications. This broad spectrum is vital for an in-depth analysis, reflecting the complexities of car pricing. Its focus on Norway is especially relevant, capturing the market's unique aspects like environmental policy impacts and the prevalence of electric vehicles. The high quality of the dataset is apparent in its detailed and diverse features, ranging from model year to body type and wheel drive, this ensures an insightful analysis. In addition, the practicality of the dataset stands out, with results aimed at benefiting consumers, dealerships, and industry analysts in the Norwegian car market. This balance of academic depth and real-

world applications makes the dataset an invaluable tool for understanding and navigating Norway's automotive landscape.

Table 1 - Description of dataset

Number of Datapoints: 33400		
Feature	Translation (from Nor to Eng)	Type
Pris	Price	Integer
Merke	Brand	Categorical
Merke ID	Brand ID	Categorical
Model	Model	Categorical
Model ID	Model ID	Categorical
Postnummer	Postal number	Categorical
Farge	Color	Categorical
Hjuldrift	Wheel Drive	Categorical
Effekt	Power	Integer
Sylindervolum	Cylinder Volume	Float
CO2-utslipp	Co2 emissions	Integer
Antall seter	Number of seats	Categorical
Karosseri	Body Type	Categorical
Antall dører (Antall dorer)	Number of doors	Categorical
Rekkevidde (WLTP)	Range	Integer
Salgsform	Sales Form	Categorical
Avgiftsklasse	Tax Class	Categorical
Kilometer	Mileage	Integer
Modellår (Modellaar)	Model Year	Categorical
Girkasse	Gearbox	Categorical
Drivstoff	Fuel	Categorical

DATA PREPROCESSING TECHNIQUES

Label Encoding: For handling categorical data in my dataset, such as 'Merke' (Brand) and 'Model ID', label encoding was the chosen method. This data preprocessing technique converts categorical values into a numerical format, enabling machine learning algorithms to interpret them effectively. It assigns a unique integer to each category within a feature.

Standard Scaling (StandardScaler): Different features in my dataset come with various scales, which can skew the performance of certain algorithms, especially those sensitive to feature scaling like neural networks. StandardScaler standardizes features by removing the mean and scaling to unit variance. This scaling, which assumes that features follow a normal distribution, ensures that all features contribute equally to the model's predictions and helps in speeding up the convergence of gradient-based optimizers.

Handling Missing Values: My dataset, like many real-world datasets, contained missing values. I addressed this through strategic filling in and deletion. Where data was missing at random and the feature was crucial to my analysis, I used distribution-based data augmentation methods. This involved replacing missing values with estimations based on the distribution of the available data, ensuring a realistic and statistical approach. For features with excessive missing values or those less critical to my analysis, I chose to remove them to maintain the overall integrity and quality of the dataset.

Train, test, validation split: An essential step in preparing my dataset for the machine learning models involved dividing the data into distinct subsets: 60% for training, 20% for validation, and 20% for testing. This division was chosen to ensure a robust training process while providing enough data for validating and testing the models' performance. The training set was used to build and tune the models, the validation set to fine-tune the models' parameters and prevent overfitting, and the test set to evaluate the models' effectiveness on unseen data.

Feature Selection and Removal

An essential step in my preprocessing phase was the careful selection and removal of certain features from the dataset. This process was guided by the principles of reducing redundancy and ensuring data quality, both of which are essential for effective modeling. Here we discuss the rationale behind the removal of specific features:

Model and Merke (Brand): Removed due to redundancy with 'Model ID' and 'Merke ID'. Eliminating these string format duplicates avoids having multiple features telling us the same thing, thus enhancing model clarity and efficiency.

Sylindervolum (Engine Volume): Removed because of inconsistent data quality and a high number of missing values. Removing it prevents inaccuracies in model predictions and maintains dataset robustness.

CO2-Utslipp (CO2 Emissions): like 'Sylindervolum', it was removed due to excessive missing values. Removing it ensures data integrity and reliability for predictive modeling.

This approach focuses on the most essential aspects of feature removal, emphasizing data quality, reduction of redundancy, and the importance of accurate model inputs.

ADDITIONAL PREPROCESSING STEPS

Before the application of machine learning models, two key preprocessing steps were undertaken to enhance the quality and consistency of the dataset:

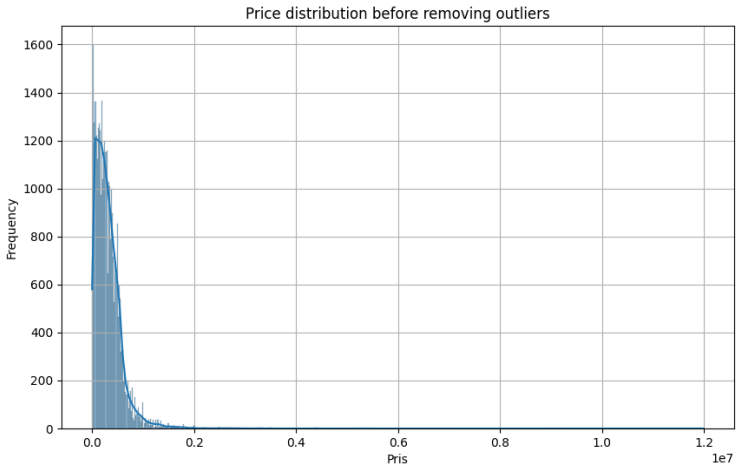

Figure 2 - Price distribution in dataset, before removing outliers

Filtering Data Within 3 Standard Deviations: The dataset was filtered to include only data points within three standard deviations from the mean, reducing the impact of outliers. This step, which reduced the dataset from 33,423 to 33,145 data points, enhances the robustness and reliability of the analysis by minimizing extreme values negatively affecting the performance.

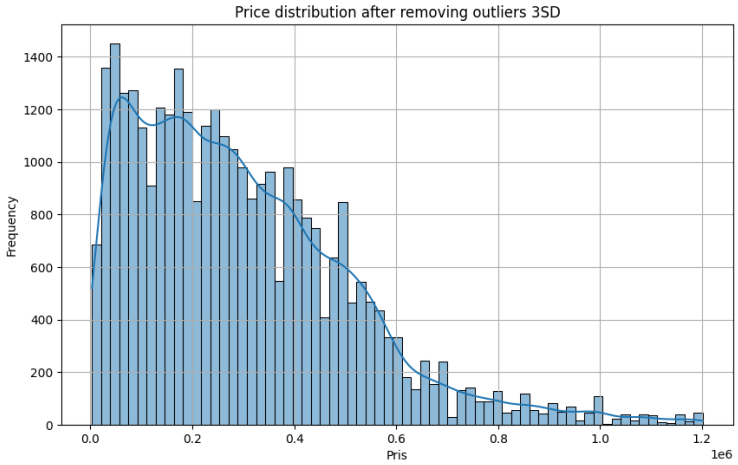

Figure 3 - Price distribution in dataset, after removing outliers.

Logarithmic Transformation (Log(x)): I used this to deal with unevenness in the data, helping to even out the distribution. It was particularly helpful for my dataset, which had lower-priced cars. This transformation helps to make the data's variation more consistent, which in turn makes it fit better with the basic rules used in statistical modeling, ultimately leading to better performance of the model.

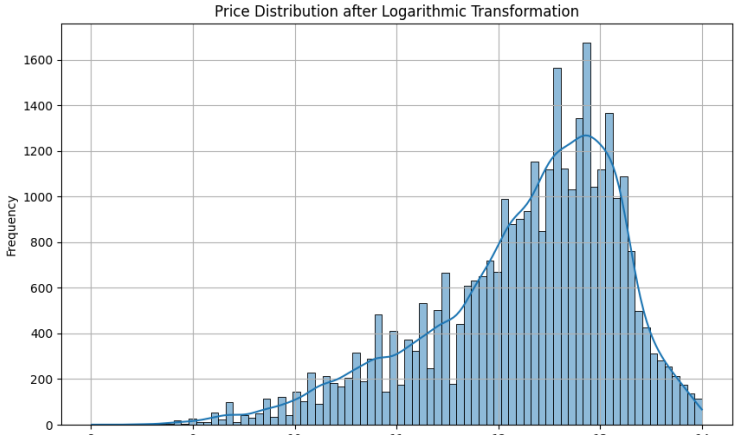

Figure 4 - Price distribution in dataset after logarithmic transformation

These preprocessing steps were vital in preparing the dataset for exploratory data analysis and machine learning modeling, ensuring data quality and fostering more accurate and generalizable insights.

CHOICE OF MACHINE LEARNING MODELS

Gradient Boosting Machine (GBM): GBM specializes in sequential error correction, making it highly precise. It is particularly effective in scenarios where the relationship between features and the target variable is complex. While it stands out in accuracy, the need for careful parameter tuning and potential overfitting have to be considered when applying it.

Extreme Gradient Boosting (XGBM): XGBM is renowned for its speed and performance, even on large datasets. It's a go-to choice for winning solutions in machine learning competitions. Its ability to handle sparse data and its features for regularizing models set it apart, making it less prone to overfitting than standard GBM.

Light Gradient Boosting Machine (LGBM): LGBM is extremely efficient in handling large datasets quickly, making it ideal for applications where time is a constraint. Its innovative sampling and bundling techniques ensure that it does not compromise on accuracy while dealing with a large volumes of data.

Lasso Regression: Lasso Regression is particularly effective in situations with more features than observations, efficiently reducing the dimensionality of the data. It not only helps in feature selection but also enhances the prediction accuracy by eliminating irrelevant features that might otherwise introduce noise into the model.

Random Forest (RF): RF is a strong contender for dealing with imbalanced datasets, as it inherently manages class imbalance. It also handles missing values effectively, reducing the need for extensive pre-processing. Its ensemble nature makes it less sensitive to outliers, enhancing its robustness.

Neural Networks (NN): NNs excel in capturing intricate patterns in high-dimensional data, making them ideal for complex regression tasks like car price prediction. They are particularly powerful in capturing interactions between variables that are not immediately apparent, offering deep insights into the underlying data structure and requiring less feature engineering the most ML models.

In conclusion, the selection of these diverse machine learning models - GBM, XGBM, LGBM, Lasso Regression, Random Forest, and Neural Networks - reflects a thorough approach to the unique requirements and characteristics of my dataset. Each model brings its own set of strengths and advantages, making them collectively adept at tackling the complex challenge of predicting car prices in the Norwegian market.

Each model has been meticulously chosen not only for its individual merits but also for how it complements the others, creating a well-rounded, robust predictive framework. This strategic selection ensures that we are equipped to tackle the challenges of predicting car prices with accuracy and depth, leveraging the strengths of each model to gain a comprehensive understanding of the factors influencing car prices in the Norwegian market.

3. Results

In this section, we delve into the heart of our analysis, presenting the findings taken from a careful exploration of the dataset and the application of advanced machine learning models. My journey began with an exploratory data analysis (EDA), a critical step to gain initial insights and understand the underlying structure of the Norwegian car market data. Following this, I transitioned into the realm of predictive modeling, where various advanced machine learning algorithms were employed to forecast car prices.

EXPLORATORY DATA ANALYSIS (EDA)

In this section, I will present the findings from My exploratory data analysis, which is divided into 4 key areas: linear correlation analysis, correlation matrix analysis, data distribution examination, and permutation feature Importance. Each of these areas provided unique insights, paving the way for informed model selection and predictive analysis.

LINEAR CORRELATION ANALYSIS

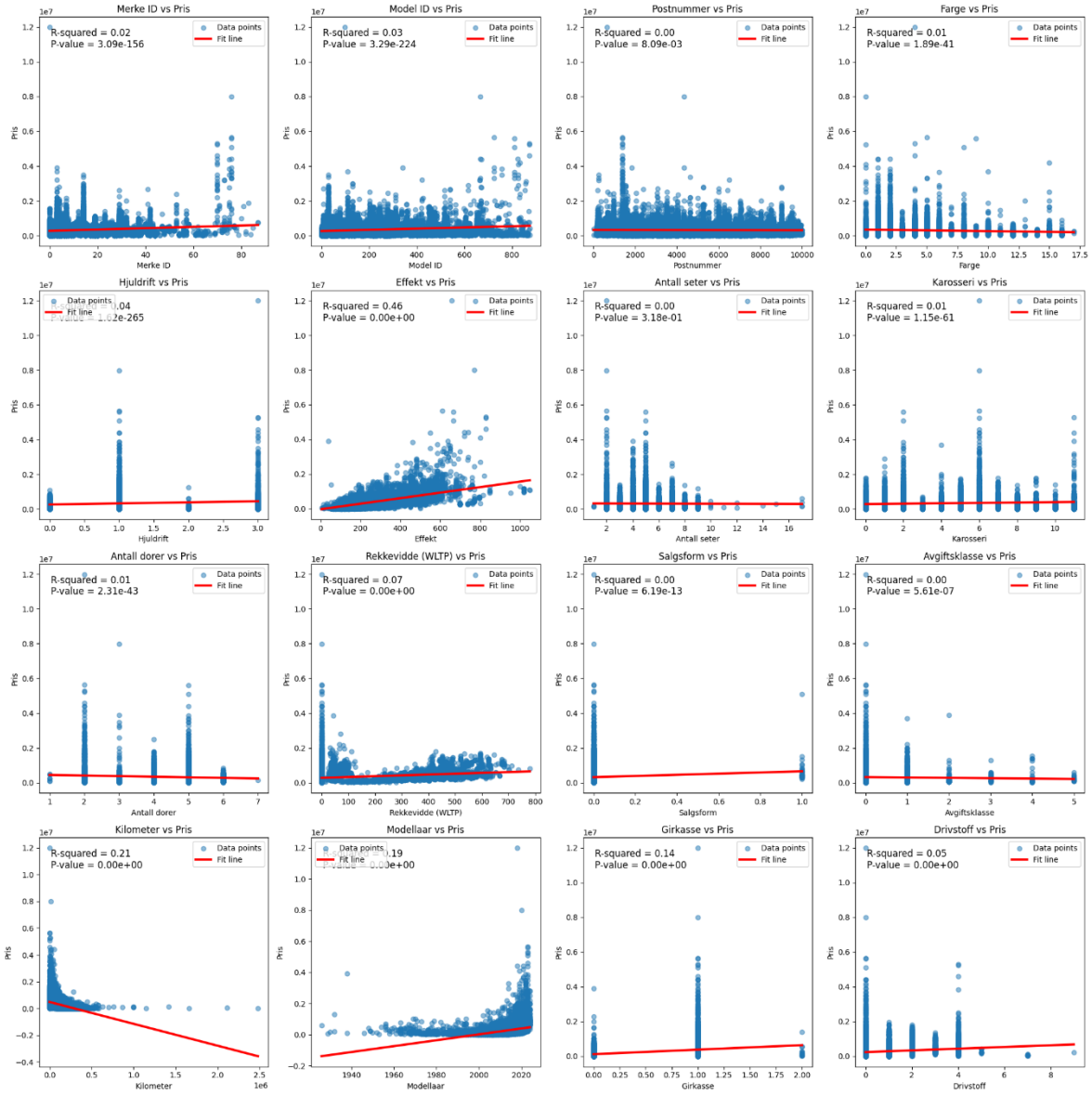

Figure 5 - Combined plots of linear correlation between Price and every other feature

The first phase of my EDA involved conducting a linear correlation analysis between each feature and the target variable, 'Pris' (Price). This step was very helpful in identifying potential key drivers influencing car prices. By measuring the strength and direction of the relationship between each feature and the car price, I gained valuable insights into which variables might be the most Influential. This analysis served as a preliminary filter, highlighting the features that needed closer examination and possibly influencing our feature selection for the machine learning models.

CORRELATION MATRIX ANALYSIS

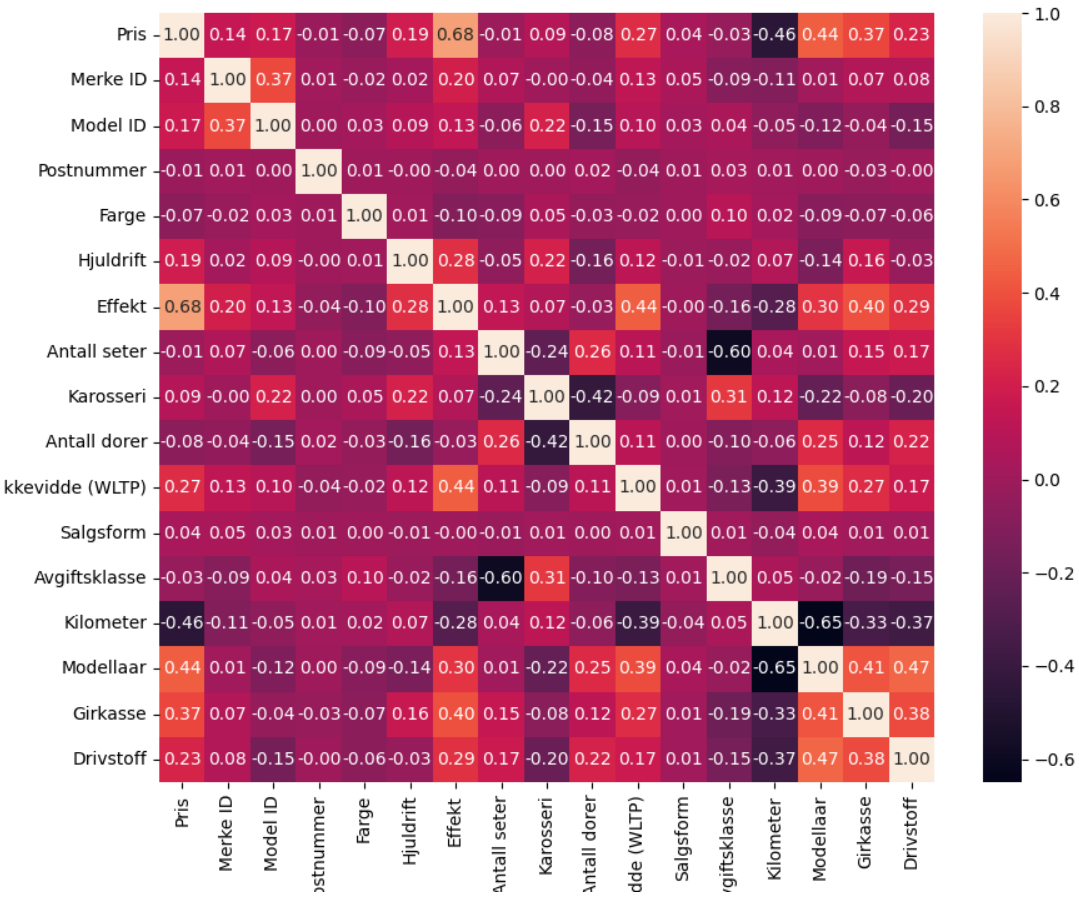

Figure 6 - Correlation matrix between every feature

Next, I constructed a correlation matrix to gain a full view of the interrelationships between variables. This step was important for spotting instances where variables were overlapping or duplicating information, which can make models less dependable. The correlation matrix helped in pinpointing which features might be redundant and which features were more correlated with the target feature than others.

DATA DISTRIBUTION ANALYSIS

The final aspect of my EDA focused on examining the data distribution across various features This part of the analysis was necessary to reveal underlying patterns, anomalies, or trends within the dataset. Understanding the distribution of data helped in identifying outliers, and determining whether any transformations were necessary for better model performance. This examination ensured that the data fed into the predictive models was reliable and aligned with the assumptions of the algorithms used.

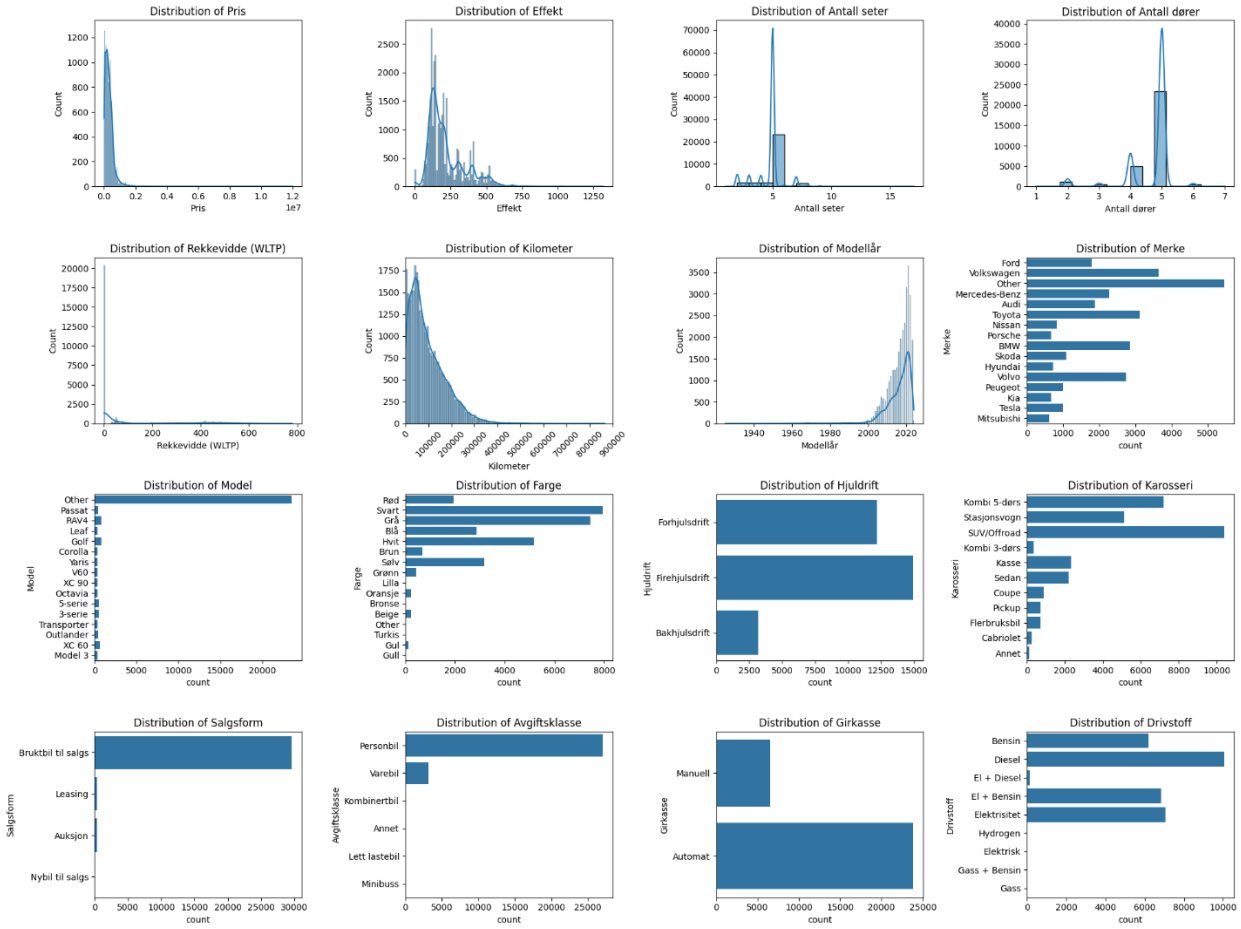

Figure 7 - Data distribution of all data before processing the data

PERMUTATION FEATURE IMPORTANCE ANALYSIS

Permutation Feature Importance is a technique used to measure the importance of each feature in the predictive model. This method involves randomly shuffling the values of each feature and observing the impact on model performance. A significant decrease in the model's accuracy when a feature's values are shuffled shows high importance. This approach is

particularly beneficial as it is independent of the model's internal structure, making it applicable to a variety of models, including complex ones like gradient boosting machines.

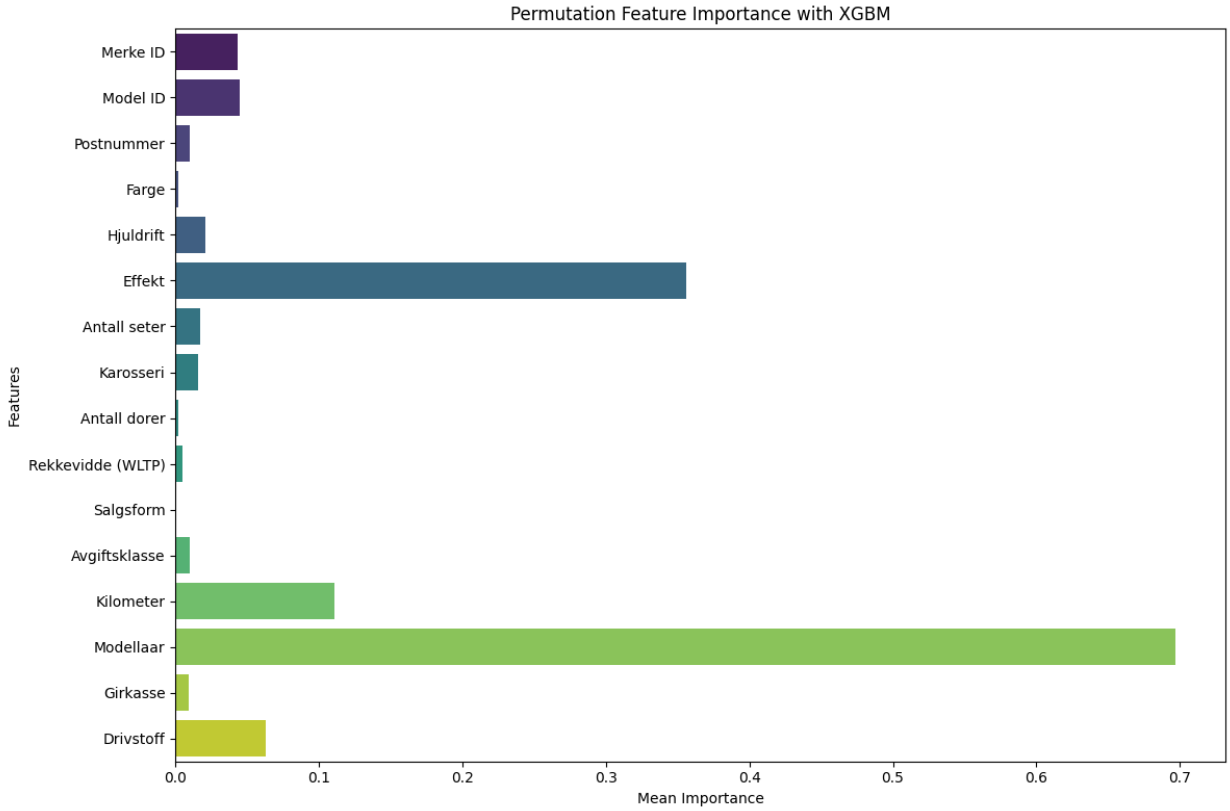

Figure 8 - Permutation feature importance using XGBM

In my analysis, I visualized the permutation feature importance results using a bar chart, where each bar represents a feature, and its length represents the mean importance calculated over multiple shuffles. This visualization offers a clear and concise representation of the contribution of each feature to the model's predictions, aiding in identifying key factors influencing car prices in the dataset.

MACHINE LEARNING MODEL TRAINING AND RESULTS

In this crucial phase of my analysis, I evaluated the performance of a diverse range of machine learning models on the task of predicting car prices in the Norwegian market. The models included Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBM), Light Gradient Boosting Machine (LGBM), Random Forest (RF), Neural Networks (NN), and Lasso Regression. Their effectiveness was evaluated using metrics such as Root Mean Square Error

(RMSE), Mean Absolute Error (MAE) and R-squared (R^2), across train, validation, and test datasets.

Table 2 - Metrics, showing results off all the models used in testing

Model	Training			Validation			Test		
Metric	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
GBM	32720.98	19921.63	0.99	31861.32	19646.45	0.99	82808.27	31685.4	0.93
XGBM	46140.41	23948.71	0.98	39474.46	23317.72	0.98	65777.04	31211.71	0.95
LGBM	36268.47	17467.66	0.98	31314.27	17267.24	0.99	67911.32	30536.06	0.95
RF	51982.54	12119.98	0.97	41268.17	11804.88	0.98	77540.45	32144.95	0.93
Lasso	178492.38	82404.26	0.63	170205.61	82628.1	0.58	177305.4	83667.52	0.66
NN	57826.15	34475.47	0.96	88843.65	40266.15	0.89	90013.56	42597.52	0.91

Each model's performance metrics across different datasets are laid out above for a clear comparison of their effectiveness in predicting car prices.

Comparative Analysis and Insights

In the assessment of machine learning models for predicting car prices in Norway, the following observations were made:

- The Gradient Boosting Machine (GBM) model excelled in accuracy on the training and validation sets but experienced a notable decrease in performance on the test set.
- Extreme Gradient Boosting (XGBM) and Light Gradient Boosting Machine (LGBM) models were closely aligned in performance, with both demonstrating a lesser reduction in accuracy from training and validation sets to the test set, suggesting better generalization than GBM.
- Random Forest (RF) showed consistent performance across all datasets without significant fluctuations.
- Neural Networks (NN) underperformed relative to other models in this study.
- Lasso Regression was the least accurate in predicting car prices, as indicated by the metrics.

These findings capture the core performance aspects of each model regarding their use in car price prediction for the Norwegian market.

RESULTS OF FEATURE REDUCTION ON MACHINE LEARNING MODEL PERFORMANCE

This section outlines the outcomes of a systematic feature reduction process on the performance of four machine learning models: Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBM), Light Gradient Boosting Machine (LGBM), and Random Forest (RF). The method involved the iterative removal of the least impactful features as determined by permutation feature selection. Features were removed one at a time, with the models retested after each removal to measure the impact on performance metrics.

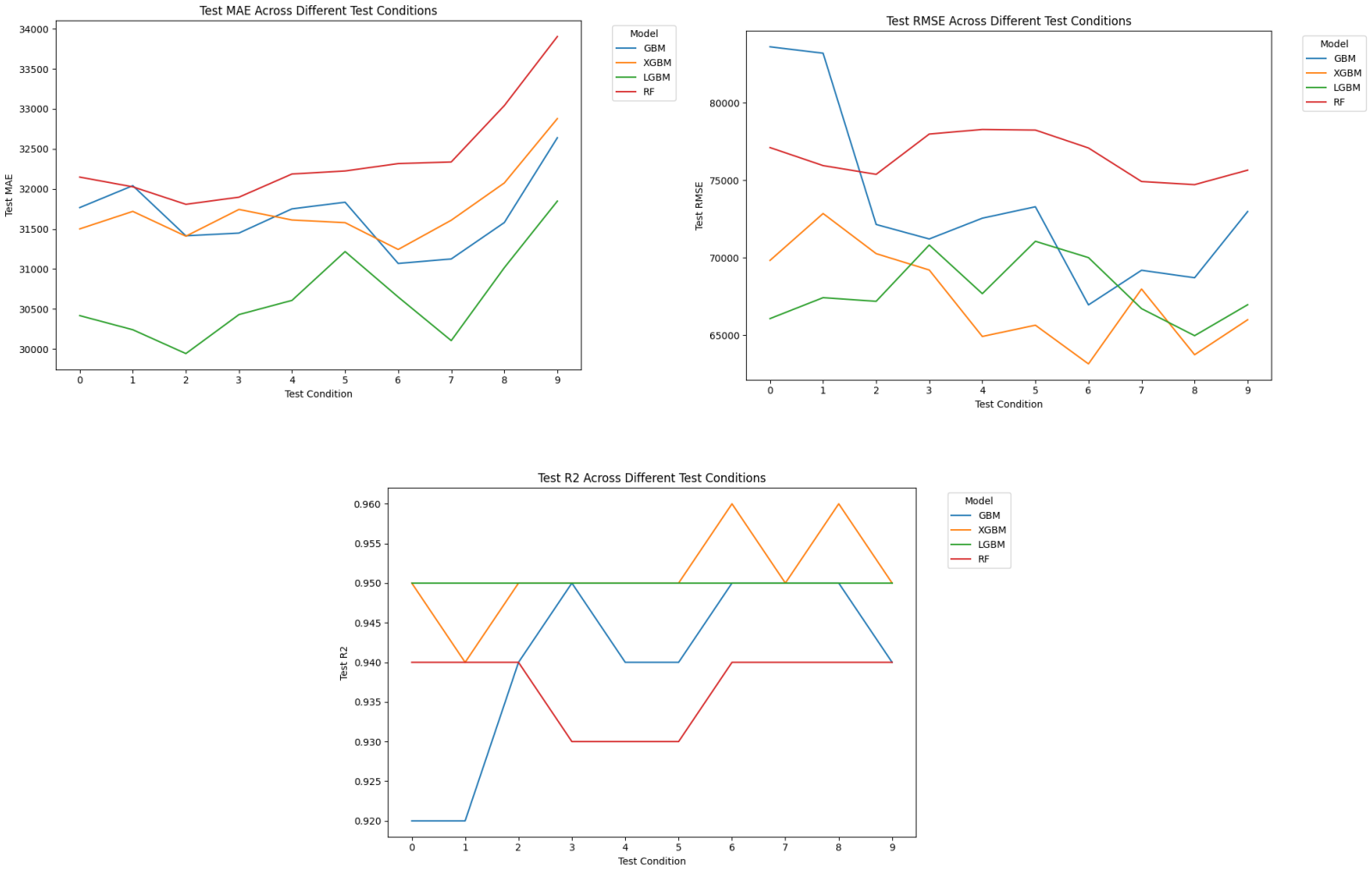

Figure 9 - Results of feature reduction on model performance

The testing cycle went through ten conditions, labeled 0 through 9. Each increment on the x-axis of the resulting graphs corresponds to one round of feature removal. The sequence of removal began with "Salgsform," identified as the least impactful feature (Figure 8), and

proceeded through "Antall dører," "Farge," "Rekkevidde (WLTP)," "Girkasse," "Avgiftsklasse," "Postnummer," "Karosseri," "Antall seter," and finally "Hjuldrift."

By the ninth test condition, the dataset was reduced to the following features:

- Merke ID
- Model ID
- Effekt
- Kilometer
- Modellår
- Drivstoff

The graphs plot the performance of each model across the test conditions. Neural Networks (NN) and Lasso Regression were removed from this analysis due to high variability in performance across test conditions, this inconsistency made it hard to clearly see how the primary models of interest were affected.

The initial results show that the performance of the model gets better as we gradually take away some of the features. These metrics give us some insights into the relative importance of each feature and the models' ability to adapt to the reduced feature set.

PRICE PREDICTION RESULTS

Figure 10 comprising of six scatterplots provides a visual comparison of actual vs. predicted car prices for each of the machine learning models tested. These plots serve as a valuable tool for visually evaluating the performance and accuracy of the models. On these plots, the actual prices (x-axis) are plotted against the predicted prices (y-axis), offering a clear representation of how closely each model's predictions align with the actual values. Notably, the R^2 scores depicted in these scatterplots are the same as those presented in Table 2, reinforcing the analysis of the machine learning models with a visual perspective. These plots are particularly insightful as they visually demonstrate the degree of variance in the predictions, further improving our understanding of each model's effectiveness in the context of the Norwegian used car market.

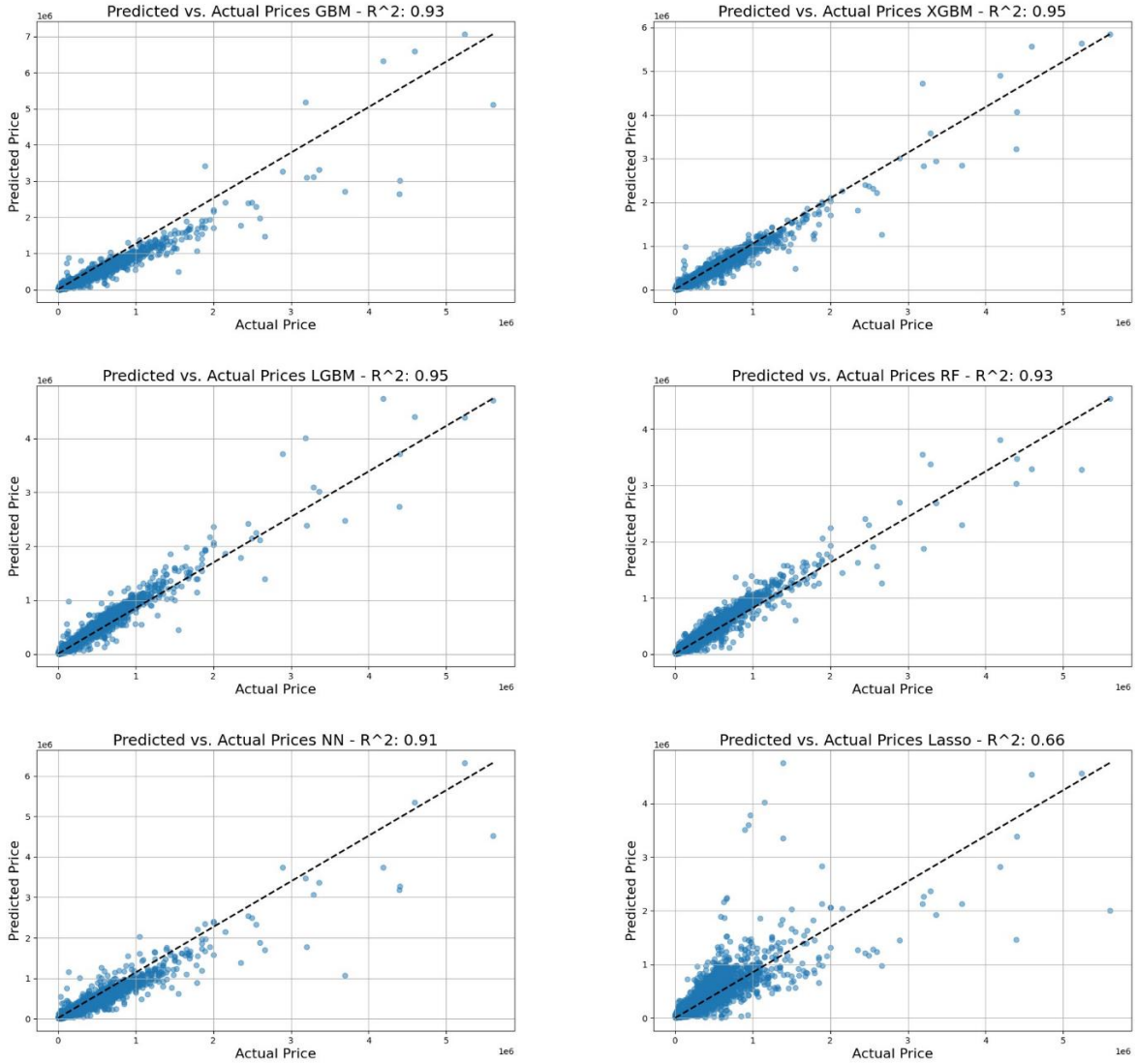

Figure 10 - Scatterplots of actual price vs predicted price by the models.

Table 3 - 10 random Actual Prices vs Predictions comparisons

XBGM				Lasso			
Actual (NOK)	Predicted (NOK)	Error (%)	Difference (NOK)	Actual (NOK)	Predicted (NOK)	Error (%)	Difference (NOK)
86777	114500.5	31.95	27723.51	899000	3516626	291.17	2617626
334147	282758.4	15.38	51388.6	126777	36934.27	70.87	89842.73
1029000	934850.8	9.15	94149.25	26777	35573.76	32.85	8796.757
1079000	989220.2	8.32	89779.8	3199000	2263794	29.23	935206.1
54961	51239.65	6.77	3721.355	249000	205206.8	17.59	43793.17
549900	519603.8	5.51	30296.2	195000	161164.8	17.35	33835.15
549000	526900.8	4.03	22099.25	184147	152797.8	17.02	31349.22
319000	326455.1	2.34	7455.12	329000	291386	11.43	37613.98
559000	569215.4	1.83	10215.4	569000	514072.6	9.65	54927.41
463147	458980.2	0.9	4166.84	424900	403950.1	4.93	20949.91

Following the scatterplots, Table 3 is shown to further highlight the performance of the models, specifically comparing XGBM (my best-performing model) and Lasso (the least effective in this context). This table showcases ten randomly selected prediction results from each model, organized into four columns: Actual Price, Predicted Price, Error (%), and Difference (Actual - Predicted). The data clearly shows the superior accuracy of XGBM, with its predictions more accurately reflecting the actual prices, as demonstrated by lower error percentages and smaller differences. In contrast, the Lasso model's predictions differ significantly from the actual values, resulting in higher error rates and larger differences. This table effectively captures the contrast in performance between the two models, providing solid examples of XGBM's precision and Lassos shortcomings when it comes to predicting car prices in the Norwegian market.

4. Discussion

In the discussion chapter, we'll take a closer look at what the results from our machine learning models mean. We'll go beyond just the numbers to see what these findings tell us about Norway's car market, how different data analysis methods work, and what it all means for people involved in this market. We'll also talk about what worked well and what didn't in our approach, linking our real-world results back to the theories we talked about earlier in the report.

EXPLORATORY DATA ANALYSIS (EDA)

LINEAR CORRELATION ANALYSIS (LCA)

My EDA, specifically the Linear Correlation Analysis (LCA) (Figure 5), yielded insightful observations regarding the relationship between various features and car prices. This analysis showed that certain features demonstrated a more obvious linear correlation with price, while others showed little to no correlation. Such findings are in line with logical expectations, as not all features would influence the price of a vehicle.

CORRELATION OBSERVATIONS

Features with Minimal Correlation: The LCA revealed that features like Postal number, Color, Number of seats, Body type, and Tax class had negligible or no obvious linear correlation with the car price. This lack of correlation is visually evident from the fit line of these variables.

Features with Mild Correlation: A subset of features displayed a modest correlation with price. These include Brand ID, Model ID, Wheel drive, Number of doors, Range, Sales Form, Gearbox, and Fuel. While these correlations were statistically noticeable, they were not very strong.

Features with High Correlation: On the other hand, features like Effect, Mileage, and Model year showed a notably higher correlation with car prices, indicating their greater relevance in the price determination process.

P-VALUES AND CORRELATION INTERPRETATION

Statistical vs. Practical Significance: The analysis, especially the p-values, suggested some correlation for most features. The p-values ranged considerably, with the highest being 3.18e-01 (Number of seats) and the second highest being 8.09e-03. However, it's important to differentiate between statistical significance (low p-values) and practical significance. A low p-value indicates a low probability of the observed data under the null hypothesis but does not necessarily mean the result has practical implications, especially in large datasets.

Strength of Correlation: The presence of a statistically significant correlation does not equate to a strong correlation. The correlation coefficient, such as Pearson's R-squared, is a better

measure of the strength and direction of a linear relationship. There can be cases where the correlation is statistically significant but weak in terms of the correlation coefficient.

Influence of Data Distribution: The distribution of data points also plays a role in the visual interpretation of the linear relationship. Certain distributions may mask the presence of a linear correlation, making it less apparent visually despite being statistically significant.

The Linear Correlation Analysis within our EDA phase provided an interesting understanding of how different features relate to car prices. While some correlations were expected and aligned with assumptions, others provided some new insights, pointing out the complexity and complex nature of factors influencing car prices. This analysis forms a critical foundation for following predictive modeling and offers valuable perspectives for stakeholders in the Norwegian car market.

CORRELATION MATRIX ANALYSIS INSIGHTS

The correlation matrix analysis in Figure 6 provided valuable insights into how different features relate to the target variable, the price of cars. This analysis was particularly revealing in confirming some expected correlations while also bringing to light a few surprising ones.

Expected Correlations: As anticipated, certain features showed a strong and logical correlation with price. Model Year, with a correlation of 0.44, aligns with the general understanding that newer cars are usually more expensive. Mileage showed a negative correlation of -0.46, highlighting the typical decrease in car value with increased usage. Additionally, the Effect (power of the car) showed a significant positive correlation of 0.68 with price, which aligns with the expectation that more powerful cars are more expensive.

Surprising Correlations: Some correlations, however, were unexpected. The correlation between Fuel and Price at 0.23, and Gearbox and Price at 0.37, were higher than I anticipated. These findings hint at a bigger impact of these features on car pricing than presumed. Also, the correlation between Range and Price was lower (0.27) than expected. This reflects a nuance of the dataset, which includes both electric and non-electric vehicles, with many data points with a range of 0, specifically for non-electric cars.

Other Notable Correlations: The analysis also highlighted correlations that provide insights into car features and the car market. The negative correlation of -0.60 between the Number of Seats

and Tax Class indicates how tax regulations influence car design in terms of seating. The correlation of -0.42 between Body Type and Number of Doors is logical, as certain body types inherently have a specific number of doors, like coupes typically having two. Range's correlation with Effect (0.44) suggests a trend towards higher power in electric vehicles, which are often present in newer models. The correlation of -0.65 between Mileage and Model Year is in line with the expectation that older cars generally have higher mileage.

These correlations from the matrix paint a thorough picture of the factors influencing car prices. They not only validate some common assumptions about the car market but also provide nuanced insights, especially regarding the evolving trends and consumer preferences in the Norwegian market. This analysis is helpful in understanding the complex nature of car valuation and the relationship of various features in determining price.

DATA DISTRIBUTION EXAMINATION INSIGHTS

The examination of data distribution from Figure 7 within the Norwegian used car market yielded several interesting insights, specifically highlighting trends and preferences in the automotive sector.

Fuel Type Distribution: A significant discovery was the distribution of 'Drivstoff' (Fuel), which revealed that nearly half of the cars on the market are electric or hybrid. This trend aligns with what is known about new car sales in Norway but seeing this mirrored in the used car market is particularly fascinating. It underscores that Norway is leading in the adoption of eco-friendly vehicles and reflects a shift in consumer preferences towards more sustainable cars. This distribution not only informs us about the current market but also hints at future trends in the car industry, both in Norway and potentially globally.

Brand Popularity and Electric Vehicle (EV) Penetration: The data showed Volkswagen, Toyota, BMW, and Volvo as the most common brands in the used car market, with Tesla also featuring prominently in the top ten. Tesla's strong presence is a notable difference from global trends, demonstrating Norway's unique position in the EV market. The popularity of these brands, particularly Tesla, further reinforces the observation of a significant shift towards electric and hybrid vehicles. (Yang, 2023)

Additionally, the high number of data points with zero in 'Rekkevidde' (Range) is a result of non-electric cars.

These insights from the data distribution examination provide a deeper understanding of the Norwegian used car market, revealing popular trends and consumer preferences that are crucial for both market analysts and potential buyers. They also offer a unique perspective on the rapid evolution of the car industry in the context of environmental sustainability and technological advancement.

PERMUTATION FEATURE IMPORTANCE ANALYSIS

The results from the permutation feature importance analysis (Figure 8) provided valuable confirmation and some surprises regarding the significance of various features in predicting car prices. Consistent with our previous analyses, Model Year, Effect (power), and Mileage were the most crucial features, though their order was somewhat surprising. In contrast to the initial assumption based on linear and matrix correlation analyses that placed Effect at nr.1, Model Year took the lead, followed by Effect and then Mileage. This insight shifts my understanding of which factors were the most influential in determining car prices. On the other hand, features like Sales Form, Number of Doors, and Color showed negligible importance, which was to be expected. Similarly, features such as Range, Tax Class, Gearbox, and Postal Number were found to have minimal impact on price prediction. These findings help refine our model by highlighting the features that warrant more focus and those that contribute less to the accuracy of the models.

MACHINE LEARNING MODEL TRAINING AND RESULTS

In this section, we delve into the heart of my analysis by examining the outcomes of training various machine learning models, including GBM, XGBM, LGBM, Random Forest, Neural Networks, and Lasso Regression. This phase was essential for applying theoretical insights to a real-world problem and testing the effectiveness of these models in predicting car prices. I thoroughly evaluated each model's performance using key metrics like RMSE, MAE and R^2. This discussion will focus on interpreting these results, exploring the strengths and limitations of each model, and understanding their implications within the Norwegian car market. This analysis serves as a crucial link between our data preparation efforts and the practical applications of our findings.

The evaluation of various machine learning models in predicting car prices revealed insightful findings, particularly when examining the performance metrics across the training, validation,

and test datasets as shown in Table 2. A key observation was the variation in model performance, with certain models showing superior results in the test phase, which is critical for assessing real-world applicability.

GBM with 7000 Trees: This model demonstrated good performance in the training and validation phases, but a notable increase in RMSE to 82808.27 during testing indicated potential overfitting issues.

XGBM (4500 Trees): This model stood out with robust performance across all phases, particularly in the test set with an RMSE of 65777.04 and an R^2 of 0.95. XGBM's ability to balance accuracy in training with generalizability in testing makes it a strong candidate for predicting car prices.

LGBM: Echoed the success of XGBM, showing excellent performance in the training and validation stages, and maintaining a strong standing in the test phase with an RMSE of 67911.32 and R^2 of 0.95.

Random Forest: Showed consistent performance with a notable RMSE of 77540.45 in the test set, though it lagged slightly behind XGBM and LGBM.

Neural Networks: Despite their complexity, it did not perform as well as other models, particularly in the test phase with an RMSE of 90013.56 and R^2 of 0.91, suggesting potential issues in model specification or training.

Lasso: Demonstrated significantly higher RMSE values across all phases, with the test set RMSE reaching 177305.40, indicating a less effective model for this particular dataset and task.

In my analysis, the Neural Networks model, configured with a 128x64x32 neuron structure across its layers, did not achieve optimal performance in the test phase. This outcome implies a potential need for optimization. Neural Networks are known for their sensitivity to architecture and hyperparameters, indicating that varying the network's depth, neuron count, or exploring different activation functions and dropout rates could significantly improve its predictive capability. Therefore, the moderate results observed initially should be viewed as a preliminary step, encouraging more extensive experimentation and fine-tuning to fully harness the model's potential in predictive analyses.

FEATURE REDUCTION

My analysis of feature reduction's impact on model performance, as presented In Figure 9, offered significant insights, particularly regarding the relevance of certain features in price prediction. The results suggested that removing less relevant features could enhance model performance, a conclusion drawn from observing changes in RMSE and R^2 scores across many test conditions.

IMPACT ON RMSE AND R^2

Reduction in RMSE: The most notable change was observed in the GBM model, where RMSE decreased by 24.9% from test condition 0 to test condition 6. This trend of decreasing RMSE was consistent across all models, but it was particularly pronounced in test condition 6 for GBM and XGBoost (15% decrease from test condition 1) and test condition 8 for LGBM (9% decrease from test condition 5) and RF (4% decrease from test condition 4). These findings indicate that the models became more accurate in predicting car prices as less relevant features were removed.

Slight Increase in R^2: Alongside the reduction in RMSE, there was a marginal increase in R^2, particularly notable in test conditions 6 and 8 with XGBoost, showcasing improved model fit and predictive power.

IDENTIFICATION OF IRRELEVANT FEATURES

Permutation Feature Importance Analysis: The analysis identified features like "Antall dører" (Number of doors), "Farge" (Color), "Rekkevidde (WLTP)" (Range), "Girkasse" (Gearbox), "Avgiftsklasse" (Tax class), "Postnummer" (Postal number), "Karosseri" (Body type), and "Antall seter" (Number of seats) as the least important. Removing these features did not adversely affect, and in many cases, even improved the models' performance.

TREND IN MAE AND OPTIMAL FEATURE REDUCTION

Increase in MAE: However, it's noteworthy that MAE showed an upward trend from test condition 7 and onwards, peaking in test condition 9. This suggests that while feature reduction generally improved model accuracy, excessive removal of features might lead to a decline in performance.

Optimal Number of Features: The results imply that removing 5-6 of the least important features optimizes model performance, particularly for RMSE and R^2, without significantly impacting MAE. This finding is essential for model optimization, as it highlights the balance between feature inclusion and model accuracy.

This analysis of feature reduction highlights the importance of feature selection in machine learning models. It shows that including more features is not always better and that identifying and removing less important features can lead to more efficient and accurate models. This insight is especially relevant for predictive modeling in complex datasets like those in the car industry, where detecting the most influential factors is essential to accurate price prediction.

LIMITATIONS

LACK OF VISUAL AND DESCRIPTIVE INFORMATION

One significant limitation of the models and their performance is the absence of two important aspects that can heavily influence a car's price: the visual condition and description. The dataset lacks complete descriptive information that could provide insights into the car's condition, such as maintenance history, abnormalities, or unique features. Similarly, the absence of visual data, like images, means we cannot evaluate the car's physical condition, which is vital in determining its market value. Factors like scratches, rust, or overall aesthetic appeal, apparent in pictures, can significantly affect a car's price, without considering its technical specifications.

NO ACCOUNT FOR MARKET DYNAMICS

Additionally, the models do not account for fluctuations in car prices due to external economic factors, such as inflation or market trends. The dataset covers a relatively short time frame, which limits my ability to analyze or predict price changes influenced by broader economic conditions. This exclusion could lead to an oversimplified understanding of price factors, as real-world car valuation is often affected by a load of economic factors, including inflation, supply-demand dynamics, and industry trends.

5.CONCLUSION

SUMMARY OF FINDINGS

This report embarked on a broad journey through the complex landscape of predicting car prices in the Norwegian market, using a blend of exploratory data analysis (EDA), machine learning models, and feature reduction techniques. The EDA revealed key insights into linear

correlations between various features and car prices, with certain attributes such as Model Year, Mileage, and Effect showing significant influence, while others like Color and Chassis appeared less impactful. This initial analysis laid a solid foundation for the subsequent predictive modeling.

MACHINE LEARNING MODEL PERFORMANCE

In my thorough machine learning analysis, I rigorously trained and evaluated a range of models, including GBM, XGBM, LGBM, Random Forest, Neural Networks, and Lasso Regression. The test results, a vital measure of real-world performance, were mainly led by XGBM and LGBM, both demonstrating strong predictive capabilities. XGBM slightly outperformed LGBM in the test phase, though this difference was marginal. This phase is particularly crucial as it measures the model's ability to adapt to new, unseen data - a key aspect of practical uses. So XGBM and LGBM emerged as the most effective models for predicting car prices in the Norwegian market. The underperformance of Lasso Regression and the moderate outcomes from the Neural Networks highlighted areas for potential improvement, such as architecture and hyperparameter adjustments. On the other hand, the Random Forest model, despite not surpassing XGBM and LGBM, proved to be a reliable benchmark, indicating its consistent performance and versatility. This analysis highlights the importance of not only achieving high accuracy during training but also ensuring robustness and generalizability in real-world scenarios.

FEATURE REDUCTION AND MODEL OPTIMIZATION

My exploration into feature reduction revealed that excluding less impactful features, as identified by permutation feature importance analysis, could improve model accuracy. This was clear in the decreased RMSE and slightly increased R^2 scores in models like GBM and XGBM. It highlighted the balance necessary in feature selection – too few features might oversimplify the model, while too many could introduce noise.

LIMITATIONS AND FUTURE DIRECTIONS

The study, however, was not without its limitations. The absence of visual and descriptive data about the cars, such as physical condition and detailed descriptions, could have provided more depth to the price prediction models. Additionally, the models did not account for broader economic factors like inflation, which can significantly influence car prices. These areas present possibilities for future research, suggesting the integration of image analysis, natural language processing, and economic modeling to achieve a more full approach to car price prediction.

CONCLUDING THOUGHTS

In conclusion, this report demonstrates the intricate relationship between various features and their impact on car prices. Through thorough data analysis and machine learning modeling, I gained valuable insights into the Norwegian car market. The findings of this study not only contribute to the academic understanding of statistical learning and data science but also offer practical implications for stakeholders in the automotive industry. Future studies should aim to address the identified limitations and explore the incorporation of more nuanced data, thereby improving the predictive models and enhancing their real-world applications.

BIBLIOGRAPHY

AlShared. (2021). Used Cars Price Prediction and Valuation using Data Mining. *RIT*, 37.

Bukvić, Š. F. (2022). Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. *Sustainability*, 18.

Felix, L. (2019). Systematic literature review of preprocessing techniques for imbalanced data. *IET Software*, 479-496.

Huang, N. H. (2022). Used Car Price Prediction Analysis Based on Machine Learning. *ICAID*, 356-364.

Intelligence, M. (2022). *Norway Used Car Market*. Retrieved from Mordor Intelligence: https://www.mordorintelligence.com/industry-reports/norway-used-car-market

Jørgensen, M. P. (2016). Brand loyalty among Norwegian car owners. *ScienceDirect*, 256-264.

Monburinon, C. K. (2018). Prediction of prices for used car by using regression models. *IEEE*, 5.

Shaprapawad, B. K. (2023). Car Price Prediction:An Application of Machine Learning. *IEEE*, 7.

Thai, S. T. (2019). Prediction car prices using quantify qualitative data and knowledge-based system. *IEEE*, 5.

Yang, L. Y. (2023). Electric vehicle adoption in a mature market: A case study of Norway. *ScienceDirect*, 12.