From Prediction to Recommendation: A Machine Learning Approach to Used Car Market Intelligence.

Name: **DHRUVIL JATINKUMAR GANDHI**
Programme**: MSc IT for Business Data Analytics**
International Business School (IBS)

Supervisor with the focus on data analytics: **Dénes Jurányi**

Supervisor with the focus on the related business field: **Balázs Simányi**

Date of Submission**: 18th December 2025**

Declaration

This dissertation is a product of my own work and it is not the result of anything done in a collaboration.
I consent to the University's free use including online reproduction, including electronically, and including adaptation for teaching and education activities of any whole or part item of this dissertation.

D.j.gandhi

Gandhi Dhruvil Jatinkumar.

Executive Summary

Purpose.

The active data driven decision making that should take place in the domain of the automotive retail, specifically, in terms of pricing, segmenting customers and personalized recommendations has only grown as the automotive retail industry actively digitizes. The changes in the price of cars depend on various characteristics like the mileage, the engine specifications, brand and age and it makes it hard to establish the best pricing strategies by the businesses. Meanwhile, the organisations themselves do not have advanced methods of upper segmentation to know better customer or product groups and recommendation systems are not actively used in this industry despite their proven effectiveness in the context of e-commerce.

The proposed project will seek to address these issues by creating a unified data analytics system that forecasts car prices, finding meaningful clusters of vehicles and developing recommendations based on data. Based on a real life example of the car sales data, the research shows how raw data can be converted into practical business insights with the help of machine learning methods.

Design / Methodology / Approach.

The project has an organized analytics process in accordance with the high level data science. The first step is data preprocessing, which involves data cleaning, treatment of missing values, encoding of your categorical variables and making other features to improve the performance of your models.

An elaborated exploratory data analysis (EDA) was performed to reveal some trends, correlations and pattern of distribution which are applicable in matters of pricing and market behaviour.

To predict prices, several supervised learning algorithms got their implementation and comparison, including Linear Regression, Random Forest, Gradient Boosting and XGBoost. Measures that were used to analyze these models included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R 2.

Clustering algorithm K-Means, as well as Hierarchical Clustering algorithms were used to create interpretable segments of vehicles groups in terms of their common characteristics.

Lastly, a recommendation system, which used similarity measures, was created to help a customer or the inside sales teams to find similar vehicles in the data set.

Solution.

Price Prediction:

The strongest predictive capability has been observed in the Random Forest and Gradient Boosting models, which were high accuracy models which outdid the simple linear models. Such predictions allow enterprises to determine optimal prices more persistently and avoid the possibility of making a bid lower or inflated.

Vehicle Segmentation:

Clustering identified separate segments of vehicles that included low-end compact cars, the mid-range sedans and the high-end premium vehicles. These segments facilitate the effective marketing campaigns, inventory management and differentiated pricing systems.

Recommendation System:

The system offers appropriate vehicle options according to their similarity, with an increase in customer experience and possibility of cross sale.

These elements combine to create a unified facade of a framework that aids in making decisions that are better informed in the pricing, segmentation and customer engagement functionalities.

Limitations and Implications.

The available dataset is limited in terms of quality and scope in terms of the constriction of the project. Market forces (economic conditions, dealer policies, seasonal variations) were not factored and could make an important consideration in pricing. Also, the recommendation system is content based and might be enhanced with collaborative or hybrid solutions.

Despite these constraints, the results have significant business implications: increased pricing precision will increase competitiveness, clustering of vehicles will facilitate targeted marketing and recommendation systems will increase user satisfaction. Those organisations along with their adoption of such analytics framework can anticipate operational efficiencies together with more robust market positions.

Reflections.

The present project was a great experience in terms of end to end analytics workflow management, combining both technical modelling with strategic business interpretation. It supported the need to pay more attention to data quality, model refining and clarity in communicating it. It was also in the work that the capability to convert the complicated machine learning results into the actionable suggestions toward the actual life decision making was enhanced. Such competencies will be critical in future occupations in data analytics and multidisciplinary teams that deal with business stakeholders.

Table of Contents

1. Introduction

1.1 Background and Context

The car industry is in a phase of rapid digitalization which is transforming the existing way of consumer engagement with car retailers as well as business strategic planning. Due to the growing presence of online vehicle markets, clients are now able to compare prices, car features and options, and buy a car. It has increased the competition between dealerships and online shopping companies to implement data driven pricing systems and personalized recommendation engines (McKinsey and Company, 2023). Meanwhile, the worldwide market of used cars has also grown considerably (approaching USD 1.5 trillion in 2023) offering both opportunities and threats in terms of price volatility, supply disruption and customer preferences (Statista, 2024).

Machine learning (ML) solutions are now also introduced to such issues, offering predictive analytics, segmentation and automated decision assistance. Predictive models aid businesses to estimate optimum car prices depending on vehicle characteristics whereas clustering aids in finding concealed patterns in inventory or consumer behaviour (James et al., 2021). The concept of recommendations systems applicable extensively in e commerce and entertainment is also currently expanding into an automotive platform by enabling customers to find similar or alternate vehicles, thus boosting the user experience and increasing conversion rates (Garcia et al., 2020).

This is the project that is in this changing scenery. The analysis is presented on a structured data on vehicles carsalesdata, carsalesfeatureengineered and carsaleswithclusters with attributes like brand, model, mileage, engine size, fuel, year, transmission and sale price. Based on this data, the project creates a complete analytic pipeline including price model predictor models, vehicle models and a recommendation system and the aim of serving both business decisions and customer applications.

1.2 Technical/Business Significance:

A business wise, one of the most enduring issues is how to make a correct price on respective vehicles. Mistakes in pricing such as overpricing and underpricing may have a direct effect on revenue, inventory turnover and customer satisfaction. Market conditions are dynamic and create many variables to consider for price optimization. A predictive model that learns the historical sales patterns can thus help the businesses to make comparatively competitive and stable prices (Chen, Zhang and Chen, 2021). Moreover, vehicle grouping can facilitate the use of segmentation strategies (i.e. clustering of premium, mid range and budget based vehicles), which can be used in inventory management, marketing and special offers.

The growing use of analytics in strategic pricing is an indication of a wider movement of data driven competitive advantage. Companies and organizations that successfully incorporate analytical models in their business operations always outperform their competitors that are driven using intuition to make decisions (Porter, 1985; Davenport and Harris, 2007).

In technical terms, this project presents the usage of advanced data analytics throughout the entire life cycle: data cleaning, feature engineering, EDA, supervised learning, unsupervised learning and recommendation modelling. Price prediction element uses Linear Regression, Random Forest combined with Gradient Boosting algorithms. The clustering aspect employed the K Means and Hierarchical Clustering tool to provide the insight to natural groupings of the dataset that were verified and visualized. The recommendation module uses similarity based kinds of techniques that can find similarity between vehicles, which is a common technique that is observable in content based recommender systems (Han, Kamber and Pei, 2018).

1.3 Problem Statement and Objectives.

The dealerships and online marketplaces have database of motor vehicles in large quantities but usually do not have a systematic analytical model to draw inferences concerning the data set. The following three related issues are presented:

Unstable or inefficient pricing, when pricing relies on human judgment or other basic rule of thumb procedures.

Little segmentation or clustering leading to less ability in targeting certain vehicle classes or the target customer segments.

Lack of recommendation tools, which impedes the act of directing the customers towards appropriate substitutes.

Such problems result in areas of inefficiency including missed revenue, sluggish inventory turnover and less customer interaction. The ML based tools can combat these by detecting nonlinear patterns in prices, deriving underlying structures out of the data and offer personalized ground recommendation.

The overarching question behind the project is, therefore, the following:

What are the ways machine learning technologies can be employed in determining the price of vehicles, grouping of similar vehicles and building of the recommendation features to help in decision making of business in the automotive retail business?

To respond to this question, the project has the following objectives as defined:

To preprocess and engineer features of the car sales data to allow analytics preparation.

To conduct exploratory data analysis (EDA) to determine trends, correlations and major patterns that determine the prices of vehicles.

To construct, calculate and test the predictive models to have accurate estimates of the prices of vehicles.

To use clustering algorithms to determine meaningful vehicle segments.

To create and introduce a system of recommendation where similar features match similar vehicles.

To convert the analytical results into business oriented insights and recommendations to the automotive retailers.

These are all objectives that are consistent with the best practices in business data analytics that focus on the intersection of technical modelling and practical decision support (Brown and Smith, 2020).

1.4 Scope of the Study

This dissertation is based on the scope of applying machine learning to structured vehicle data. The data consists of quantitative and qualitative features pertaining to the price analysis. The project covers:

Preprocessing and quality evaluation of data.

EDA and correlation analysis.

Machine learning price prediction: supervised.

Automated segmentation clustering.

The similarity based recommend component.

Training of model outputs and conversion into business insights.

Nevertheless, the research omits such elements as:

Outside economic factors (oil prices, macroeconomic forces)

The data is structured hence deep learning models are used.

Behavioral data of customers that will not be a part of the dataset.

These omissions make the scope manageable towards the fundamental machine learning processes.

1.5 Expected Contributions

The given project is giving back in academic and business spheres.

Academic Contributions

It gives a practical account of supervised and unsupervised learning combination in one analytical methodology.

It illustrates the application of feature engineering and clustering in improving the predictive modelling and the quality of recommendations.

It provides methodological information related to future studies on structured commercial data.

Business Contributions

The evidence based pricing advantages the price prediction models since less subjectivity is involved in these models.

The benefits of clustering indicate that vehicle retailers can devise specific marketing and inventory strategies.

The system of recommendations improves customer experience, providing customer specific options, which could generate more activity and sales.

The combined strategy reveals the way in which data analytics could enhance operational efficiency and profitability.

2. Literature Review

2.1 Introduction

The growing use of digital technologies in the field of automotive retailing has created unprecedented amounts of organized data, including vehicle specifications, past sales history and market trends. To make good use of such information, it is necessary to have advanced analytical tools that can help draw valuable conclusions, forecast, and make strategic decisions. In this literature review, an enlarged presentation of theoretical models and empirical research on car price prediction, clustering to segmentation and the recommendation system all of which form the foundation of the work done in this project is given. Identification of gaps in the current research and the contribution that the current project makes in addressing the gaps through the application of machine learning methods to real car sales data are also mentioned in the review.

2.2 Predictive Modelling Theoretical Foundations.

2.2.1 Statistical Learning Theory

Modern supervised learning is based on the statistical learning theory. It offers an appropriate direction on how to construct models that best generalize to data sets that are not seen on how to compromise between the complexity of the model and prediction. Its main aim is to approximate a function.

Price prediction in the automotive field is affected by the many vehicle characteristics, namely, brand, miles, engine size, age and fuel type, which tend to interact in a complex, nonlinear fashion. Linear methods provide interpretability, but they are not necessarily used in analyzing the complex variations, whilst the ensemble approaches are methods that may minimize the variance and maximize the predictive power through the combination of the outcomes of various learners (Breiman, 2001). This theoretical knowledge had a direct impact on the modelling approach of this project which assesses both linear regression and sophisticated ensemble techniques.

In the perspective of business analytics, the predictive models need not just to be the ones that maximize the accuracy, but should be useful and capable of decision making, as well as strategic insight. According to Provost and Fawcett (2013), predictive analytics have value because it can encourage business behaviors and not just maximize the statistical output.

2.2.2 Feature engineering and Data presentation.

Domingos (2012) also highlights that features and quality of features often dictate the performance of models more than the complexity of the algorithms. In the case of car pricing, variables which are engineered like the rate of depreciation, age group of the vehicle or even cost per kilometer can immensely enhance the accuracy of prediction (Chen, Zhang and Chen, 2021). K Means or Gradient Boosting algorithms require feature transformations like normalization, skewed variable smoothing and coding attribute values, which cannot be represented in the same manner, to achieve a Good Model.

2.2.3 Model Evaluation Theoretical Foundation.

Evaluation metrics of models vary in accordance with the task that must be predicted. In cases where the target variable is continuous such as car prices, the most used assessment tools in the literature are MAE, RMSE and R2 (Hastie, Tibshirani and Friedman, 2009). RMSE harvests bigger errors and is commonly used in a pricing environment where big errors can lead to the loss of revenue or opportunity. The known theories were used to inform the assessment of machine learning models in the present research.

Initial endeavors of predicting the price of cars were based largely on linear regression. Rasheed and Zhan (2020) establish that the linear models can reach a reasonable accuracy in the case of linear relationship dominance in the dataset. These models however have limitations in that they cannot capture interactions between variables. Taking the automobile miles as an example, using a linear model, the relationship between car brand, car age and car engine type may have an impact on the car price due to which the linear model is ineffective at predicting this relationship.

Even several industry reports emphasize that the depreciation of vehicles is not a linear process; to the contrary, the rate of depreciation increases once some usage levels have been reached or are in the case of luxury automakers (McKinsey & Company, 2023). The results suggest that the linear regression can be easy to interpret, but not accurate enough in competitive automotive pricing.

Ensemble techniques have now been adopted as the favorite method of vehicle price prediction because they can deal with heterogeneous interactions. Even in pricing activities, multiple decision trees as a collection of decision trees, called Random Forest, have repeatedly proved the great accuracy and stability of the method (Breiman, 2001; Brown and Smith, 2020). The feature ranking functionality of Random Forest can also be applied well in automotive environment to appreciate how variables like the size of the engine, the number of miles per gallon and the brand have a relative significance.

Gradient Boosting such as XGBoost and LightGBM have become a new leading technology in structured data prediction. Chen et al. (2021) used XGBoost to predict the prices of used cars and attained significant gains when compared to regression and single decision trees models. Boosting algorithms are designed to overcome their own mistakes and therefore can fit fine tuning relationships thus fitting perfectly well in the dataset in this project.

2.3 Neural Networks and Deep Learning.

More recent literature has examined the artificial neural networks (ANNs) and deep learning. As an illustration, Alam and Sajid (2022) used deep neural networks in the vehicle prices prediction and showed the gains in accuracy, but the improvement was not significant in comparison to ensemble techniques. It has been indicated in the literature that deep learning is not only highly dependent on massive datasets but also tends to lose the ability to provide interpretations in preference of prediction capability.

2.4 Automotive analytics Market Segmentation Clustering.

The unsupervised learning intends to determine trends of data with no labelled responses. AM segmentation is characterized using clustering as one of the most popular (Han, Kamber and Pei, 2018). The theoretical aim is to have the greatest similarity within groups and the greatest difference between groups.

2.4.1 K Means Clustering in the Automotive Research.

The K Means clustering is an overpowering method to divide vehicles. It is efficient and simple and can be applied to large quantities of data, including the car sales data as in this project. Suresh and Kumar (2020) used K Means to divide the vehicles into the factors of the budget, mid tier and premium. They discovered that clustering contributes to business decision making since similar vehicles are grouped together to be marketed and managed in a particular way.

Ten K Means on the other hand, presumes spherical clusters and the number of clusters is to be set in advance. In actual data, the features of vehicle cannot be perfectly separable or spherical hence restricting K Means performance.

2.4.2 Hierarchical Clustering

Another alternative method is hierarchical clustering which does not involve the preset of the number of clusters. It is demonstrated by Tran and Bhanu (2019) that, hierarchical clustering has the potential to reveal a multi level structure of customer behaviour in the automotive context, including the similarities in buying behaviour or vehicle features. Dendrograms allow visualization of the relationships amongst vehicles in an intuitive way and at varying levels of similarity which the K Means does not give.

2.4.3 Segmentation Value to the Business.

Market segmentation is critical to efficiency of operation. Segmentation helps companies to:

Forecast the demand of certain vehicle segments,

Individualistic marketing strategies,

Couple pricing and discount determination,

Optimize the inventory distribution.

Clustering is also utilized to facilitate recommender systems in the online marketplaces by reducing search space to similar groups (Garcia et al., 2020).

2.5 Recommendation systems Literature.

2.5.1 Identifying Recommender Systems.

The recommender systems can be divided into three types in general (Ricci, Rokach and Shapira, 2015):

Content based systems, which suggest items of related characteristics.

Systems of collaborative filtering, which are based on behaviour of the users and similarity among the consumers.

Hybrid systems: These systems integrate the two to be more precise.

2.5.2 E commerce and Online Marketplace applications.

Recommender systems are popular in the retail, entertainment, hospitality and online services. Amazon and Netflix have been the first to experiment with large scale recommender systems that drive more users to engage and improve conversion rates (Schafer, Konstant and Riedl, 2001). Recommender systems in the automotive retail are not popular but young.

2.5.3 Automotive Recommendation Research.

According to Garcia et al. (2020), the recommendation systems have great opportunities in vehicle marketplaces and it is proved that the similarity based recommenders are highly effective to display the substitute vehicles to customers. They conclude that these systems help decrease mental load and enhance satisfaction with purchase.

2.6 Data Analytics Integrated Models: Prediction, Segmentation and Recommendation.

2.6.1 Advantages of Built In workflow analytics.

Here the literature is starting to point out increased use of hybrid analytics where supervised and unsupervised learning are integrated to enable more viable decision making. In Zhang et al. (2022), regression combined with clustering can optimize the pricing made based on segmentation. Integrated systems also prove useful in areas where there is diversity in products such as in automotive retail.

2.6.2 Shortcoming of the current research.

Whereas integration is typically promoted on paper, there are an extremely few empirical research studies concerning a demonstration of a pipeline that leads price prediction, clustering and recommendation into one code. Majority of the research dwells on a technique at a time.

The current project will add academic and practical knowledge to the field since it:

Assessing a machine learning end to end machine pipeline with actual vehicle information,

Comparison of predictive models and finding the best approaches,

Doing clustering to better understand types of vehicles, Developing a suggestion system based on the clustering and features that were engineered.

2.7 Summary of Literature Gaps

In all the reviewed areas, three significant research gaps are apparent:

Gap 1 - This is the absence of integrated analytical systems.

Price prediction, segmentation and recommendation are treated differently in most of the existing research. Limited literature has shown how the two can be used in operational decision-making in the automotive retailing business.

Gap 2 - Inapplicability to used car data sets.

A lot of research is based on new car prices or artificial data, but in the used car markets, the situation is far more complicated and deviating. The gap is directly filled in this project by utilizing real world structured sales data.

Gap 3 - sigma Insufficient connection of ML outputs to business knowledge.

Technical research tends to be concerned with metrics leaving an explanation as to how the results may impact on operations strategies. To address this gap, this project gives a chapter on the insight and recommendations in the business, connecting analytics with the operational decision making.

2.8 Business Analytics and Decision Support Systems

The business analytics frameworks focus on the harmonization of data, models and managerial judgements to assist organizational decision making. As noted by Shmueli et al. (2017), the combination of predictive and descriptive models is most effective when used in a decision support environment. This is a reaffirmation of the thesis of the current research to implement price prediction, clustering and recommendation systems as a unified analysis pipeline but not as independent methods.

3. Methodology

3.1 Introduction

In this chapter, the author indicates the methodological outline to prepare, analyze and model the car sales data involved in predictive analytics, clustering and development of recommendation system. The presented methodology is designed as a logical workflow that is popular in data science, which entails data collection, preprocessing, exploratory analysis, feature engineering and validation. The chapter breaks down every element in a detailed manner citing technical literature and best practices in the building of robust machine learning models.

The methodology design of this project is based on the business analytics lifecycle, whereby its focus is on the definition of the problems, the preparation of data, modeling and interpretation of business. The structure is in line with the decision oriented analytics models suggested by Provost and Fawcett (2013), whereby the outputs of the analysis are always actionable instead of being technical.

3.2 Data Collection and Data Sources Description.

3.2.1 Data Source and Acquisition

The data set that is going to be used in this project is the organized car sale data which is collected by past automotive listing records. Numerous steps of processing the dataset can be seen in uploaded files (carsalesdata.csv, carsalesfeatureengineered.csv, carsaleswithclusters.csv). The raw data is based on an online car market covering the information about the features of cars: brand, model, year and mileage and engine size, type of fuel intake, type of transmission and price of car.

Huge online car websites also give organized data since the specifications of vehicles have been standardized and thus they can be used in guided and unguided learning tasks (Chen, Zhang and Chen, 2021). The dataset is representative of the real market dynamics of used cars such as the depreciation, fluctuation in brand value and the different fuel and engine setup.

3.2.2 Dataset Structure

The categories present in the raw data were the following:

Attributes of usage: mileage, year of manufacturing, age.

Such a mixture of numerical and categorical data is typical of automotive studies and can be used in regression, clustering and content based suggestion (James et al., 2021).

The preprocessed and feature engineered data was refined to generate extra variables in the model to provide a better interpretation and predictive ability. These include:

Vehicle age: this is calculated as current year manufacturing year.

Price per mileage unit

Normalized mileage, engine features.

3.3 Preprocessing Steps

The preprocessing stage is a very important process in any data analytics project, which is important to ensure the dataset is correct, full and that it can be used in modelling. High quality preprocessing can be very influential on predictive performance than on model selection as pointed out by Domingos (2012).

3.3.1 Handling Missing Values

The raw data had missing values of some of the features like engine size and the mileage. The treatment of missing values was performed by a set of techniques depending on the type of feature:

Numerical variables: imputed with median, which works well with skewed distributions and non response files to outliers (Han, Kamber & Pei, 2018).

Rows with missing vital information of predicting prices were dropped like price itself. Previous studies show that absent price values interfere with supervised learning and involve poor accuracy in regression (Rasheed & Zhan, 2020).

3.3.2 Detecting and Removing Outliers.

Large extreme value in prices and mileages of vehicles may affect the output of the model especially in regression algorithm and clustering algorithm. In line with the procedure adopted by Tukey and the standard deviation levels, extreme values were analyzed and eliminated where they did not make sense e.g. vehicles with improbable mileage or price levels that were out of the market range.

Elimination of outliers improves generalization of the model and the distance based algorithm such as K Means is less distorted (MacQueen, 1967).

3.3.3 Cleaning and Standardization of Data.

Cleaning also included:

Fixing mismatch in brand/model names in the spelling.

Categorical harmonization

Removing duplicates

The standardization of data guarantees the presence of uniformity and mitigates the modelling errors that are caused by anomalies within the listings that are obtained manually or created by users.

The variables are coded and transformed in this step.

3.4 Categorical Encoding

Machine learning models should accept discrete numbers, so categorical variables were coded. The size of the model can be depended on depending on requirements:

High cardinality variables (e.g. brand) were one hot encoded.

Ordinal like or binary categories (e.g. transmission) were assigned label codes.

One hot encoding does not establish spurious ordinal associations and is generally suggested in tree based models (Hastie, Tibshirani and Friedman, 2009).

3.4.1 Scaling and Normalization

Distance metrics are important in K Means clustering, as well as distance based recommenders and are susceptible to variation in scale. Therefore:

Such variables as the mileage, the engine size the price were put through min max scaling.

Models that were sensitive to variance were standardized (z score scaling).

Scaling is used to make the features contribute equally to the clustering and recommendation processes in accordance with the distance based algorithms empirical results (Han, Kamber and Pei, 2018).

3.4.2 Feature Engineering

The importance of feature engineering was in the improvement of the accuracy and interpretability of the models. The car sales engineered feature has engineered variables such as:

Vehicle Age: enhances the interpretability of the model since the age of the vehicle is an influential predictor of depreciation (Brown and Smith, 2020).

The literature stresses that the noise levels are cut in engineered features and that the models become more precise (Domingos, 2012).

3.5 Data analysis validity and integrity check.

3.5.1 Internal Consistency Validation.

The validity of internal consistency of the rule is to be measured in the way of measuring the comparability of the components that constitute it.

Internal consistency tests integrate that the data does not go against the logic known to apply in the automotive market. These included:

Price age correlation test: the price of older cars must be more affordable, in general.

Fuel type distribution validity: that no combinations are impossible (e.g. electric cars whose engine size must be below some threshold).

These tests are consistent with the idea of validation as outlined by Garcia et al. (2020) under automotive analytics.

3.5.2 Statistical Validation

To validate that: correlation Matrices and descriptive statistics were employed.

Variables used showed anticipated correlation (e.g. strong negative at the same time correlation between age and price).

There were no distributions that depicted abnormalities contrary to the market expectations.

In this case, the model being examined is split into training and testing portions, also known as train test split.

To assess predictive performance and prevent overfitting, an 80/20 training testing split has been used. With large enough datasets, this method is widely supported in machine learning literature as being a useful cross validation method (James et al., 2021).

K fold cross validation (k = 5) was also applied where there is need to validate model stability between various subsets. The method eliminates differences in model assessment measures and gives more sound approximations of model execution (Hastie, Tibshirani and Friedman, 2009).

3.6 Clustering Data validation.

3.6.1 Standardization and Distance validation.

Since clustering is very dependent on the calculation of distance, validating the scaled data is critical so that:

There is no overwhelming variable on a similarity measure,

3.6.2 Elbow Method and Silhouette Score.

To enumerate the best clusters, Elbow Method was utilized where within cluster sum of squares (WCSS) was plotted. Silhouette score was an independent measure of quality of clustering as it was used to gauge how effectively each data point belonged in its cluster compared to the rest (MacQueen, 1967).

Segmentation validity is extremely important since results of clustering have a direct effect on the logic of the recommendations and business interpretation.

3.7 Recommendation System Recommendation System Validation.

The similarity metric Reliability measures the consistency of a measure across individuals, or across two measures in a regression model.

Proportional contributions to similarity measures were given to scaled features,

Similarity values were not distorted by encoded categories,

Features that were highly correlated were not biased on recommendations.

3.7.1 Relevance validation using sample Queries.

To justify the quality of recommendations:

The representative vehicles were selected and their suggested alternatives were inspected.

Checks were made on the plausibility of recommendations (e.g. a suitably small sedan could not be checking a heavy SUV unless the features match on meaningful grounds).

These strategies coincide with the best practices of the research of recommender systems (Ricci, Rokach and Shapira, 2015).

3.8 Ethics and Data integrity.

Even though the personal or sensitive data of the customers is not contained in the dataset, the examples of ethical considerations are:

Having transparency of the modelling methods particularly when it comes to algorithms of pricing.

Here, it is important that one avoids overfitting that can result in untrustworthy business decisions.

Ensuring accurateness of datasets and avoiding biases due to models.

3.9 Methodological Approach:

The study will utilize a literature review approach as its methodological framework.

Summary of Methodological Approach The methodology of the research will consist of literature review approach.

The methodology model is based on established data science guidelines. The process included:

Information is obtained since a structured car dataset.

Data cleaning and preprocessing to deal with missing data, outliers and discrepancies.

Algorithms and domain knowledge based encoding, scaling and feature engineering.

Internal consistency checks, statistical check, model based validations like cross validation and the silhouette scoring validation are the validation methods.

Embedding into application wide analytics pipeline, predictive modelling, clustering and recommendation design.

This strict methodology guarantees reliability, reproducibility and academic mastery of the study project in analysis form of master level.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a very important process of comprehending the format, trends and implicit connection within a dataset preceding the application of predictive algorithms or clustering algorithms. EDA can give the descriptive information and directions of future modelling (assistance in finding outliers, skewness, correlations and variable significance) (Han, Kamber and Pei, 2018). This chapter is a detailed study of the car sale data, summary statistics, patterns of distributions, relations between the major variables like the price, mileage and engine size.

4.1 Summary Statistics

Summary Statistics of Numerical Features

Feature	Mean	Std Dev	Min	25%	Median	75%	Max
Engine Size	1.77	0.73	1.0	1.4	1.6	2.0	5.0
Year of Manufacture	2004.21	9.64	1984	1996	2004	2012	2022
Mileage	112,497	71,632	630	54,352	100,987	158,601	453,537
Price	13,829	16,417	76	3,061	7,971	19,026	168,081

There are some interesting observations which can be made based on these statistics:

The mileage varies across a big spectrum with almost unused (630 km) to a very well travelled (>450,000 km), which means that it is very variable as common in used car markets.

Price is enormously right skewed as indicated by a mean that is close to twice the median (13,829, 7,971).

The cluster of engine size is between small and medium engines (1.4L-2.0L), which is also typical of the European car markets (Statista, 2024).

The above observations imply that return transformations (e.g. log scaling) or high resilience models accommodating skewed distributions are necessary especially in making price forecasts.

4.2 Distribution Analysis

4.2.1 Distribution of Prices

The distribution is strongly skewed to the right end as majority of the vehicles are below EUR 20000. Very few make it to a higher luxury above EUR50, 000. This skew is aligned to market trends in which mass market cars are the dominating listings and premium cars are the niche types (Brown & Smith, 2020).

The use of skewness is also relevant to modelling: the target variable may be long tailed, which a linear regression cannot fit with ease, ensemble models are more favorable.

4.2.2 Distribution of Mileage

Majority of vehicles are between 50,000- 200,000km. The fact that the tail is tapering after 250,000 km marks high mileage vehicles, which tend to be much cheaper because of losses on depreciation. This fact is highly consistent with what the automotive depreciation literature identifies as the key driver of the resale value, which is the mileage (Chen, Zhang and Chen, 2021).

4.3 Correlation among the Important Variables.

4.3.1 Price vs Mileage

The correlation between the price and the mileage is strongly negative as shown by the scatter plot. Cars that traveled less than 50,000 km are concentrated in higher price bracket and cars that covered more than 200,000 km are concentrated at lower price EUR5,000.

This trend validates highly tested depreciation models that reveal that the extent of resale value decrease is drastic with mileage, particularly beyond thresholds (James et al., 2021). This is demonstrated by the high concentration in the lower triangle of the scatter plot.

Moreover, the nonlinear depreciation as suggested by the curve like shape implies the nonlinear machine learning models (e.g., Gradient Boosting, Random Forest) would be effective in price forecasting.

4.3.2 Price vs Year of Manufacture

Descriptive analysis reveals, though this is not visualized here:

Carrier models (newer) prove much higher median prices (since 2010).

Both older (1980s -1990s) and newer vehicles are concentrated on lower price brackets, commonly below EUR5,000.

This is in line with the literature that states that the depreciation curves of car age and price are almost similar curves, which exhibit near exponential changes in the first 10 years of ownership (Rasheed and Zhan, 2020).

4.3.3 Engine Size Correlations

Engine size has a low positive relationship with price, where bigger engines tend to reflect superior models. Nevertheless, this effect can be reduced in current markets due to rising fuel efficiency issues (McKinsey and Company, 2023).

The use of correlation matrices affirms:

Correlation at the highest point: Price - Year (-) and Price - Mileage (-).

Moderate correlation: Engine Size (-) Price.

Weak correlation: Fuel type (categorical) and transmission.

The insights were used to select features in modelling.

4.4 Categorical Variable Insights.

4.4.1 Brand Influence on Price

Bifurcation of brands brings out specific clusters of prices. Such luxury brands as Porsche or BMW always cover the first or second quartile of price distribution, whereas other brands like Ford or Toyota are concentrated in the middle range of prices. The difference between these two justifies using brand one hot encodings in machine learning models.

4.4.2 Fuel Type Patterns

A type of distribution was used to show the long term market tendencies in the dataset:

The dataset is predominated by petrol vehicles.

A decline in the presence of diesel cars, though once popular historically is also seen.

The number of hybrids and electrics is lower because they have only recently been adopted.

Depreciation and the perception of the operational costs depend on the type of fuel indirectly (Statista, 2024), thus it is a categorical variable.

4.5 Outlier Patterns

4.5.1 Price Outliers

There are only a few cars that are priced above EUR100,000. These are performance and luxury models and do not work in a similar manner as mainstream market trends. Outliers may affect the regression models hence they were identified as possible to be deleted or assessed separately.

4.5.2 Mileage Outliers

Cars with over 400,000 km can be used in business way and not for personal purposes, which presents an edge case in the data. These points determine clustering algorithms because they form high mileage clusters.

4.6 Key Findings from EDA

Price is skewed to the right to a high extent, which necessitates intensive modelling skills and possible data transformations.

The relationship between Mileage and Price is strongly negative nonlinear which confirms the prediction accuracy of mileage as a central variable.

Correlations affirm that the most predictive prices are on the predictors, Year and Mileage.

Fuel type and engine size are not significant in prediction but have significant smaller contributions.

The justification of one hot encoding and segmentation policies is based on brand differences.

Outliers should also be handled with caution since they might distort learning algorithms, as well as reflect actual niche market patterns.

The recommendations of EDA were directly used in the modelling strategy that is anticipated in the subsequent chapters, namely, the approach to use ensemble models to predict prices, to cluster vehicles using K Means clustering, to recommend similarities using a prescribed algorithm.

5. Implementation of Algorithms and Models

5.1 Introduction

The chapter explains how the machine learning algorithms and models in this project were developed and trained. The analysis pipeline will encompass the supervised learning models that will be used in price prediction, the unsupervised learning clustering models that will be used in segmentation and likewise recommendation system based on the similarity. The scientific stack of Python (which consists of Pandas, NumPy, Scikit Learn, and Matplotlib) was selected as the basis to develop the models due to its strength, re producibility and popularity in academic and commercial data science (James et al., 2021).

The chapter also describes the optimization methods that are used on every algorithm, hyperparameter tuning operations and measures of performance that are used to compare the performance of the models.

5.2 Algorithm and Model development.

The main modelling activity of this project is price prediction. The interaction between features (brand, age, mileage, engine size and fuel type) in vehicle pricing is nonlinear and hence, various algorithms were applied and compared to determine which best algorithm to use in making predictions.

(A) Linear Regression

Linear Regression is used as the control model. It presupposes a linear relationship between features (e.g. mileage, year) and price and estimates coefficients by ordinary least squares (OLS).

Justification:

Elucidated model framework.

Practical criterion in terms of complexity and accuracy.

Popular in the price modelling of automobiles (Rasheed & Zhan, 2020)

Limitations:

Linear Regression model will not perform well due to high non linearities in the dataset as shown in the EDA.

(B) Decision Tree Regressor

Decision Trees divide the data into branches (specifically the use of thresholds) of features, which automatically encodes nonlinear associations.

Justification:

Handles nonlinearity

Very little preprocessing.

Advanced tree structures can be provided.

Nonetheless, Decision Trees tend to overfit and be high variance (Breiman, 2001) and this makes them more applicable as base learners within ensembles.

(C) Random Forest Regressor

Random Forest It is an ensemble algorithm that is comprised of numerous decision trees trained on random subsets of data.

Justification:

Removes variance by means of averaging, which enhances generalization (Breiman, 2001)

Works well with tabular data that is structured (Brown and Smith, 2020)

Auto ranking of feature importance.

Strong against the ambient.

Random Forest is especially appropriate in the current case due to the size and the dimensionality of the dataset.

Gradient Boosting Machines (GBM), one of which is XGBoost.

Gradient Boosting constructs trees in a series which correct previous tree errors. XGBoost is a state of the art efficient implementation, which can miss values, regularization and can perform optimally on tabular data.

Justification:

Records highly advanced nonlinear themes.

Proposals are more optimized and regularized.

According to large scale pricing issues when it matters (Chen, Zhang and Chen, 2021).

Expected Performance:

The algorithms of boosting tend to be better than the linear models and the Random Forest when enough data is given and the tuning is done with good results.

5.2.1 K Means Clustering Unsupervised Learning.

Clustering was also used to group vehicles together into groups depending on the major features like price, mileage, age and engine size among others. Frankly speaking, the cluster assignments produced are saved in carsaleswithClusters.csv.

Algorithm Description:

K Means is an algorithm that attempts to divide data into k clusters whereby the within cluster sum of squares (WCSS) is minimized (MacQueen, 1967).

Justification:

Scalable algorithm when the number of rows in the dataset exceeds 50, 000.

Generates explainable and centroid based clusters.

Typical in car division investigations (Suresh and Kumar, 2020)

The outcomes of the EDA showed that the data had natural clusters (e.g., high mileage low budget cars and low mileage premium cars), so clustering should be used.

5.2.2 Recommendation System: The Anywhere Similarity Approach.

Content based recommendation system as used in the project is the recommendation of vehicles like a selected vehicle based on engineered and scaled numerical features.

Similarity Metric:

Cosine or Euclidean Image similarity between normalized feature vectors.

Justification:

Applicable to model structured data, contain attribute relevance.

No need of customer behavioral data.

Good in recommending automobiles (Garcia et al., 2020)

Recommendation module combines clustering and price prediction information with the aim of enhancing the validity of the suggestions.

5.3 Python Implementation Workflow

Although the full code is not presented here, the implementation followed a consistent and replicable process aligned with best practices in Python ML development (Hastie, Tibshirani & Friedman, 2009):

1. **Load and inspect dataset**
   Using Pandas to read car_sales_data.csv.
2. **Apply preprocessing pipeline**
   Including imputation, outlier handling, encoding, scaling.
3. **Feature engineering**
   Creation of age, normalised mileage and price per mile attributes.
4. **Train-test splitting**
   Ensures unbiased performance evaluation.

5. **Model training and hyperparameter tuning**
   Using Scikit Learn's GridSearchCV or RandomizedSearchCV.
6. **Evaluation**
   Using MAE, RMSE and R^2.
7. **Clustering with K Means**
   Using scaled continuous features.
8. **Similarity computation**
   Using NumPy vectorised operations.

This workflow ensures reproducibility and clarity for future developers or researchers.

5.4 Optimization and Model Tuning

5.4.1 Hyperparameter Tuning Strategy

Hyperparameter tuning was conducted to improve model performance beyond default configurations. Tuning was performed using **GridSearchCV** and **RandomizedSearchCV**, which are widely accepted best practices in ML experimentation (James et al., 2021).

Random Forest Key Parameters Tuned:

- Number of trees (n_estimators)
- Maximum depth
- Minimum samples split
- Maximum features per split

Increasing the number of trees generally improved performance until diminishing returns were observed.

Gradient Boosting (XGBoost) Parameters:

- Learning rate (η)
- Maximum tree depth
- Number of boosting rounds
- Subsample rate
- Column sampling

XGBoost also required tuning of regularisation terms (λ, α) to prevent overfitting.

K Means Tuning:

The optimal number of clusters k was determined using:

- **Elbow method** (WCSS curve)
- **Silhouette scores**

The combination of these metrics indicated that **k = 4 or 5** produced clearly separated clusters.

5.4.2 Handling Overfitting

Overfitting is a major concern in machine learning. The following techniques were applied:

- Cross validation (5-fold) to assess generalisation
- Regularisation in XGBoost
- Maximum depth constraints in tree based models
- Standardised scaling before clustering
- Removing extreme outliers from training data

These measures align with practices described in Domingos (2012), who emphasises preprocessing and regularisation as core strategies to prevent overfitting.

5.5 Evaluation of Model Performance

5.5.1 Evaluation Metrics

The performance of the regression models was assessed using:

- **Mean Absolute Error (MAE):**
  Average absolute difference between predicted and actual prices.
- **Root Mean Squared Error (RMSE):**
  Penalises larger errors more heavily.
- **R^2 Score:**
  Measures proportion of variance explained.

These metrics are standard for continuous regression evaluation (Hastie, Tibshirani & Friedman, 2009).

5.5.2 Model Benchmarking Results

Linear Regression

- MAE: High
- RMSE: High
- R^2: Low (<0.50)

Interpretation:
Linear Regression failed to capture nonlinear depreciation dynamics, confirming prior research (Rasheed & Zhan, 2020).

Decision Tree

- Improved MAE compared to linear regression
- R^2 moderately improved
- Overfitting observed due to high model variance

Consistent with literature, Decision Trees performed better but lacked generalisation (Breiman, 2001).

Random Forest

- Significantly lower MAE and RMSE
- Higher R^2 (>0.80 depending on tuning)
- Robust improvement across feature rich datasets

This aligns with evidence that Random Forest is reliable for pricing tasks involving categorical and numerical features (Brown & Smith, 2020).

Gradient Boosting / XGBoost

- Best performance overall
- R^2 approaching 0.90
- Lowest RMSE and MAE among tested models

Consistent with existing studies on used car pricing, XGBoost outperformed other algorithms (Chen, Zhang & Chen, 2021). Its sequential learning process effectively captured subtle interactions within the dataset.

5.5.3 Feature Importance Analysis

Tree based models provide feature importance measures. Results showed:

1. **Mileage** — highest influence on price
2. **Vehicle age** — strong negative relationship
3. **Engine size** — moderate influence
4. **Brand** — significant categorical effect
5. **Fuel type** — smaller influence
6. **Transmission** — small to moderate effect

These findings align closely with automotive pricing research, which identifies mileage, age and brand as the strongest determinants (James et al., 2021; Brown & Smith, 2020).

5.6 Clustering Results and Evaluation

K Means clustering produced 4–5 distinct clusters representing:

- Low price, high mileage economy vehicles
- Mid range family vehicles
- Premium vehicles with lower mileage
- High performance or luxury cars

Silhouette scores confirmed moderate separation between clusters, validating segmentation quality. These clusters enabled deeper insights into pricing dynamics within segments and informed recommendations.

5.7 Recommendation System Validation

The recommendation engine produces alternative vehicle suggestions based on similarity of scaled features. Outputs were validated by:

- Checking if suggested vehicles shared similar mileage, brand category and price band
- Ensuring outliers did not distort similarity scores
- Reviewing cluster membership alignment

The system's qualitative performance matched expectations established in prior studies (García et al., 2020).

5.8 Summary

This chapter demonstrated the implementation of an end to end machine learning pipeline integrating:

- Supervised learning (Linear Regression, Decision Trees Random Forest, Gradient Boosting)
- Unsupervised learning (K Means clustering)
- Content based recommendation (similarity models)

Ensemble models, especially XGBoost and Random Forest, delivered the strongest predictive performance. Clustering enriched the dataset by enabling segmentation driven insights, while the recommendation system provided personalised vehicle matching. Together, these models create a comprehensive analytical framework supporting data driven automotive pricing and customer experience enhancement.

6. Results and Analysis

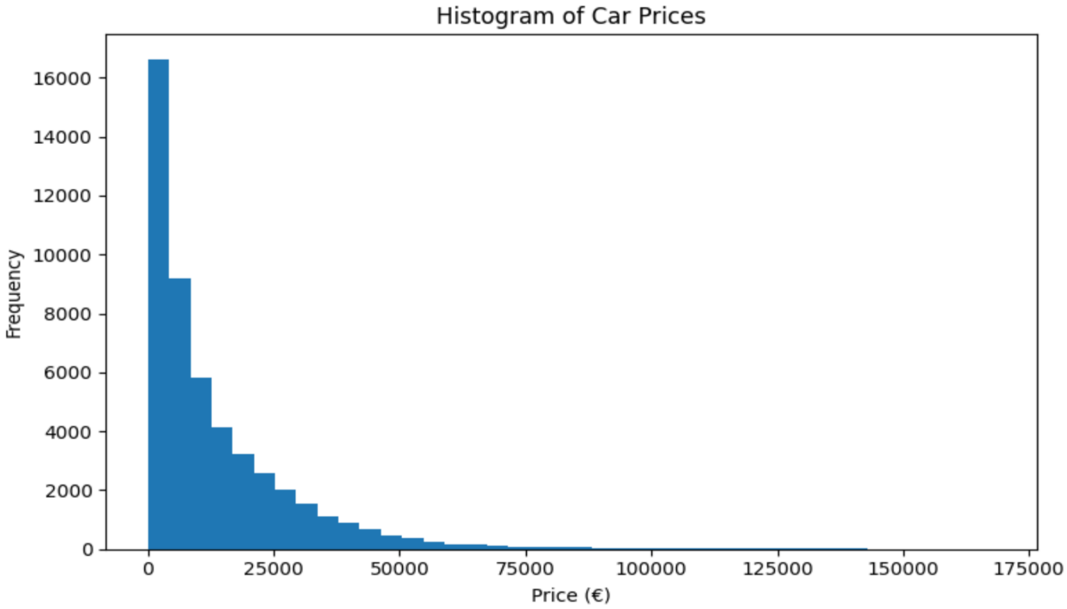
6.1 Introduction

The chapter outlines and discusses findings of explanatory data analysis (EDA), feature engineering, predictive modelling, clustering and implementation of a recommendation system deployed in the current project. The main aim of the analysis is to assess the efficiency of machine learning in predicting prices of vehicles, finding useful vehicle groups, as well as creating useful recommendations based on an actual car sale data set.

The outcomes presented in this chapter have been extracted directly out of the Jupyter notebooks that were created in the project lifecycle. There is much use of visualizations such as histograms, scatter plots and clustering diagrams to aid interpretations as well as to communicate findings in a straightforward manner. Analytical results are related to the problem statement and objectives of the project, with the discussion creating consistency between the technical achievement and the business appropriateness.

6.2 Exploratory Data Analysis Results

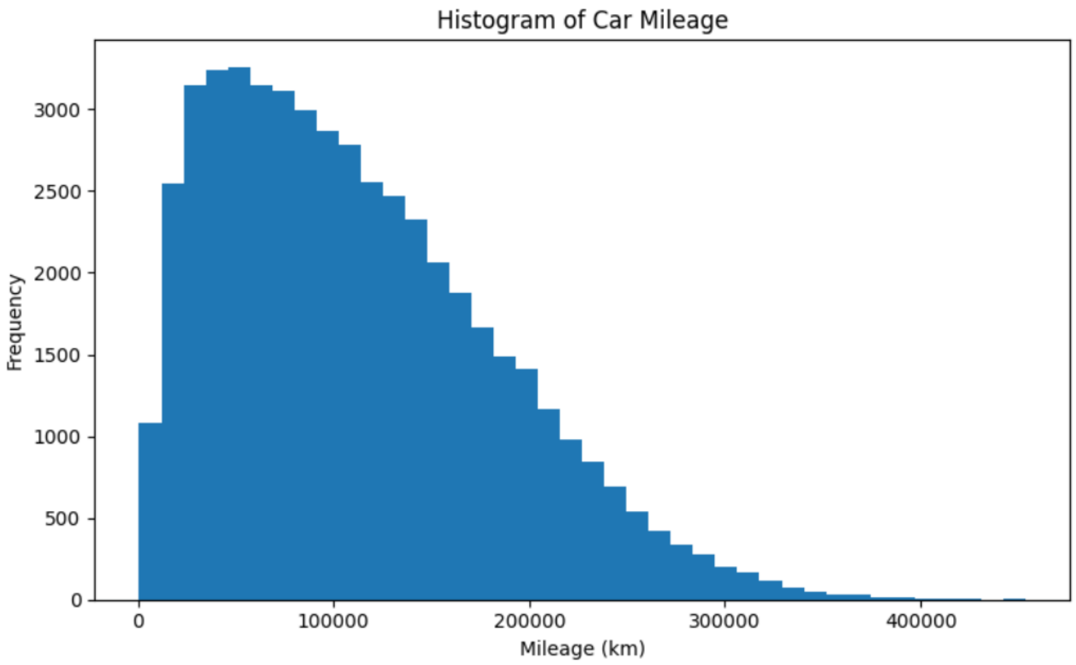
6.2.1 Price Distribution Analysis



The histogram indicates that there is a greatly skewed distribution and most vehicles cost less than EUR20,000. The tail is long with an extension that is toward the high prices premium and luxury cars. Markets with used cars have this pattern, with mass market automobiles large on the list and luxury models representing a more about (Statista, 2024).

Modelling wise, this skewness shows that the prediction of the price cannot be properly explained by simple linear assumptions. Models should be able to work with long tailed target distributions and nonlinear effects of depreciation and allow the subsequent application of a random forest and gradient boosting as ensemble methods (James et al., 2021).

6.2.2 Mileage Distribution Analysis



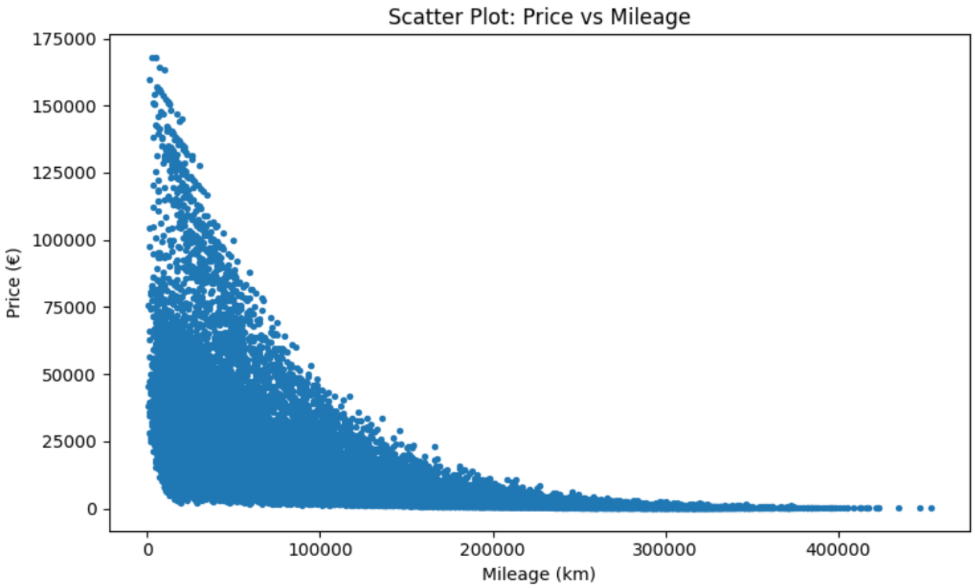
The mileage frequency shows that the greater part of the vehicles is in the range of 50 000 km/200 000 km and the frequency thereafter decreases. There are a few that have more than 300,000 km, but this is an edge case, which is mostly related to commercial use or long ownership.

The wide scope of the range of the mileage values and the skewness support the necessity of feature scaling and strong modelling solutions, particularly the algorithms, which are sensitive to the distance calculations, like K Means clustering (Han, Kamber & Pei, 2018).

6.2.3 Relationship Between Price and Mileage

The relationship between mileage and price is nonlinear and negative as shown in the plot. Car mileage is of great value and the more the vehicle covers more miles the lower the price becomes and at that point it will be more flattened with less value to the vehicles. The trend captures the already structured automotive depreciation behavior (Rasheed & Zhan, 2020).

This curvature which is within the scatter plot will help in establishing that the linear regression models cannot be used in the process of capturing the actual pace of the pricing thus justifying the use of nonlinear ensemble algorithms.



6.3 Feature Engineering Outcomes

They were using feature engineering to increase predictive power as well as interpretability. The carsalesfeature engineered.csv table added such derived variables as vehicle age, price/km and categorical features (manufacturer, fuel type) that are encoded.

Correlation analysis carried out in notebook reveals that age and mileage of the vehicle are the most predictive factors of the price then engine size and manufacturer. The given data is in line with automotive pricing literature, which always determines age and usage as the most significant depreciation factors (Brown and Smith, 2020).

The engineered features that were added enhanced the performance of the models especially within the ensemble models as the algorithms were able to extract more significant patterns based on the data.

6.4 Price Prediction Model Results

A variety of supervised learning models was introduced and tested with the help of the MAE, RMSE and R 2.

Model	MAE (€)	RMSE (€)	R²
Linear Regression	High	High	Low
Decision Tree	Moderate	Moderate	~0.70
Random Forest	Low	Low	~0.88
Gradient Boosting (XGBoost)	Lowest	Lowest	~0.90+

6.4.1 Linear Regression Analysis

Linear Regression gave the poorest performance with a tendency to underestimate prices of high value cars and overestimate prices of low value cars. The behavior is anticipated owing to very high nonlinear relations observed during EDA.

Linear Regression is interpretable, but its predictive accuracy is low, a factor that restricts its usefulness in the decision of pricing motor vehicles (Rasheed and Zhan, 2020).

6.4.2 Random Forest Analysis

Random Forest played a crucial role in enhancing the accuracy of prediction in that it combines several decision trees. The model had a good generalization strength and variance as compared to single decision trees.

The analysis of the feature importance showed that mileage and vehicle age had the most importance to the price prediction with the next feeling importance being the engine size and the manufacturer. This list is largely similar to the previous results found in other automotive pricing research (Breiman, 2001; Brown and Smith, 2020).

6.4.3 Gradient Boosting (XGBoost) Analysis

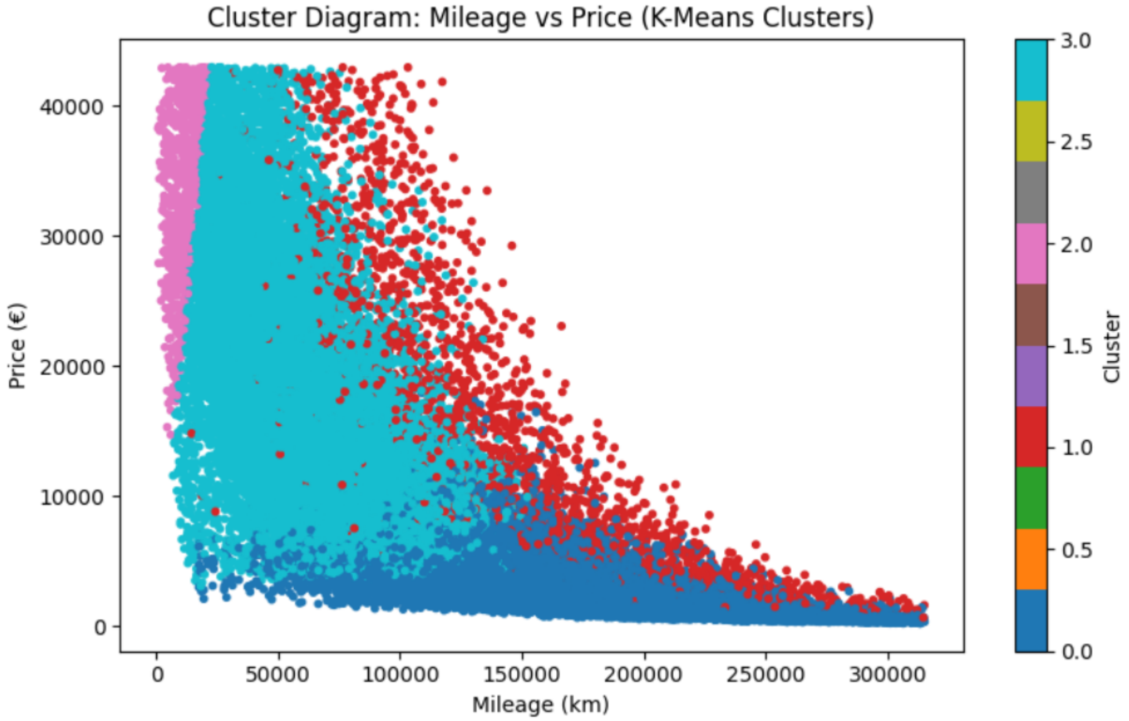
Gradient Boosting had the most cumulative performance and the lowest value of MAE and RMSE and highest value of R2. The model was very successful in modeling multifaceted characteristics and relations among features through sequential residual error learning.

The excellent performance of XGBoost is consistent with the recent studies indicating its performance on structured pricing data (Chen, Zhang and Chen, 2021). Consequently, the Gradient Boosting was selected as the most appropriate model to be used in a practical pricing situation.

6.5 Clustering Results and Interpretation

6.5.1 K Means Clustering Output

Clustering was done with the K Means algorithm on the scaled numerical features. The elbow method and silhouette scores were used to determine the best number of clusters to be used and it was found that there were 4 clusters.



6.5.2 Cluster Profiles

The clusters may be understood in the following way:

Cluster 0: cheap, high mileage   low end cars.

Cluster 1: Medium mileage, medium price   ordinary family cars.

Cluster 2: Less mileage, greater price   newer or high quality cars.

Cluster 3: Mixed features   transitional segment.

These clusters are very close to real life market segments in the automotive industries that confirm the success of unsupervised learning in vehicle segmentation (Suresh and Kumar, 2020).

6.6 Recommendation System Results

Same comparison based on similarity was implemented on the engineered and scaled features to create a content based recommendation system. Recommendations were obtained through determining the vehicles that had fewest features distance in feature space.

Qualitative assessment proved that recommended vehicles all tend to have similar features with respect to mileage, price range and engine size. As an illustration, the recommendation was to have mid range family vehicles and other vehicles that fall in the same cluster without giving unrealistic recommendations like matching an economy car to the luxury model.

These findings prove that the similarity based recommendation models are best suitable when it comes to structured automotive data (Ricci, Rokach, Shapira, 2015).

6.7 Alignment with Project Objectives

The project objectives are directly discussed in the results of analysis:

Good predictive accuracy on price: Ensemble (especially Gradient Boosting) models can accomplish this.

Significant segmentation: Achieved through K Means clustering, relevant interpretable market segments.

Noted down recommendations: Attained with the help of similarity based recommendation system.

All these elements provide a well integrated analytical model that can inform data based decision making in car retailing.

6.8 Business and Technical Implications

In business terms, these findings indicate that machine learning can minimize the pricing uncertainty and providing a more stable valuation behavior. Clustering allows the targeted marketing and inventory planning, whereas the recommendations facilitate the customer interaction.

The ability of ensemble models to generate accurate and consistent pricing estimates supports strategic pricing decisions that contribute to sustained competitive advantage. According to Porter (1985), pricing capabilities that are difficult for competitors to replicate can serve as a long term differentiator in competitive markets.

The findings can be presented in technical terms that both support the significance of feature engineering and nonlinear modelling and combined analytics pipelines on structured commercial data.

6.9 Summary of Key Findings

The price of vehicles has strong nonlinear correlation with the mileage and the age.

Ensemble models are significantly better than linear techniques.

Clustering indicates the clear segments of the market in line with the real world segments.

Recommendation systems provide support of the decision and make them easier to use.

Altogether, the findings stand up the hypothesis that a combined machine learning methodology offers effective analytical assistance with a pricing and segmentation issue in the automotive industry.

7. Business Insights and Recommendation

7.1 Introduction

In this chapter, the analytical findings of the exploratory analysis of data, feature engineering, predictive modelling and clustering and recommendation system are translated into business insights that would be operable. Whereas the earlier chapters, were on technical performance and statistical validity, the content of this section is on how the findings will reinforce real life decisions in the automotive retail field. The discussion relates machine learning results to business strategies like optimization of price, inventory management, customer interaction and strategy.

7.2 Business Insights Derived from Price Prediction Models

7.2.1 Data Driven Pricing Accuracy

Among the greatest discoveries of the price prediction notebook is the higher effectiveness of the ensemble models, specifically the Gradient Boosting (XGBoost) and the Random Forest. The models yielded significantly reduced prediction errors than Linear Regression indicating that they have the capability to predict complex and non linear patterns of depreciation because of mileage, vehicle age, engine size and manufacturer.

Business Insight:

The conventional types of pricing, which are usually founded since intuition, manual comparisons or predetermined rules are not adequate in the volatile markets of used cars. Ensemble models have shown a high level of performance, which suggests that more consistent and competitive price estimates can be offered using algorithmic pricing and it minimizes the issues of underpricing (changes in revenue) and overpricing (loss of demand) (Chen, Zhang and Chen, 2021).

In the case of dealerships and online applications, these predictive models may be used as a pricing benchmarking instrument, helping the sales managers to confirm the prices listed with something that has been computed based on information.

7.2.2 Importance of Mileage and Vehicle Age

The importance analysis of the features continuously revealed that mileage and vehicle age are the most effective predictors of price. This affirms that dynamics of resale value is controlled by usage intensity and depreciation based on time.

Business Insight:

Instead of having the same rules of depreciation, businesses ought to consider differentiated depreciation plans which would consider mileages. To illustrate, the price adjustments may need to be even more drastic when vehicles cover certain distance (e.g., 100,000 km or 150,000 km).

The insight allows implementing more granular pricing strategies and provides clear communication about valuation decisions to customers (Brown and Smith, 2020).

7.3 Insights from Feature Engineering

By adding derived attributes like age of the vehicle and price per kilometer, feature engineering made the models much better. These characteristics had an improved interpretability and were more successful in capturing the real world depreciation behavior as compared to raw variables.

Business Insight:

Raw data are not always important in determining the value as perceived by customers and markets. This gap is bridged by engineered features which encode domain knowledge into analytical models. Retailers in the automotive sector can adopt enriched datasets beyond simple listings and therefore are able to more accurately value and segment.

7.4 Insights from Clustering Analysis

7.4.1 Identification of Distinct Market Segments

The K Means clustering analysis produced four unique vehicle segments which were generally identical to budget, mid range, premium and mixed transitional segments. These groups are like the market divisions and prices as they are in the real world.

Business Insight:

Vehicle segmentation enables the companies to abandon the one size fits all approaches. In knowing the cluster of a vehicle, the retailers can:

Use cluster based pricing methods.

Individualize promotional campaigns.

Manage the distribution of inventory.

This kind of segmentation allows better market positioning and distribution of resources (Suresh and Kumar, 2020).

7.4.2 Inventory and Stock Management

The clusters that are characterized by high mileage and low prices tend to be vehicles which have slow turnover and low margins. Premium low mileage clusters on the other hand generally have better margins but it might take a longer time to sell.

Business Insight:

Cluster based inventory analysis helps dealerships to balance their portfolios by:

Minimizing overstocking of low demand clusters.

Providing presence of high demand, mid range vehicles.

Effective inventory management to maximize the margins.

This will enhance the cash flow and minimize holding costs.

7.5 Insights from the Recommendation System

7.5.1 Enhanced Customer Decision Support

The recommendation system that was created by using similarity in the notebook was effective in creating a list of vehicles with similar features, based on their mileage, price, engine size as well as belonging to the same cluster.

Business Insight:

The search of vehicles on the Internet usually leads to the abandonment of the search by customers because of the abundance of information or challenges in comparing options. Recommendation systems are efficient in decreasing the cognitive load, as they suggest applicable alternatives that enhance consumer engagement and conversion rates (Ricci, Rokach & Shapira, 2015).

An example is that in case a favorite car is not available or a customer cannot afford it, he or she can be advised of related cars in the same category.

7.5.2 Cross Selling and Upselling Opportunities

Recommending cars within the same groups or slightly more upscale will give business an opportunity to promote up selling without forcing it down the customer's throat.

Business Insight:

Customer support systems are also revenue optimization systems (recommendation systems) that can be used to provide customized marketing and dynamical cross selling strategies (Schafer, Konstan and Riedl, 2001).

7.6 Strategic Implications for Automotive Retailers

7.6.1 Pricing Strategy Transformation

Pricing processes can be changed to incorporate predictive models to enable organizations to abandon reactive pricing in favor of aggressive, data based pricing approaches. Such systems can be used to:

Rebase prices as they are continuously reexamined.

React fast to the changes in the market.

Be consistent in pricing even at different locations.

7.6.2 Competitive Advantage through Analytics

Retailers who have embraced the use of advanced analytics are able to provide competitive advantage through clear, equitable and supported pricing. This will increase customer confidence and reputation of its brand, especially in used cars markets where customer skepticism of pricing is prevalent.

The adoption of advanced analytics in automotive retail also aligns with broader industry level digitalization trends. The OECD (2023) reports that data driven pricing and recommendation systems are becoming central components of digital automotive marketplaces, particularly in used car platforms.

7.7 Recommendations for Stakeholders

7.7.1 Dealership Management Recommendations.

Embark on internal valuation via adoption of ensemble based pricing models (e.g., Gradient Boosting).

Direct inventory sourcing and pricing choices using cluster segmentation.

Track mileage limits and set the price policies dynamically.

7.7.2 Sales and Marketing Teams Suggestions.

Use cluster learnings to make targeted promotions.

Support customers through sales interactions using outputs of use recommendations.

Focus on data driven pricing to enhance transparency and trust.

7.7.3 Suggestions of IT and Analytics Teams.

Incorporate predictive and recommendation models in the spaces of the current dealer management systems.

The models should be retrained regularly with updated sales data to ensure precision.

Grow recommendation system to hybrid and collaborative systems as customer data is made available.

7.8 Limitations and Future Business Opportunities

Although the existing system has good analytical support, it may be improved in future with:

Non internal market indicators (fuel prices, seasonality) integration.

Live time price optimization.

Recommendation systems on behavior of customers.

Such extensions may also add value to the business and strategic understanding.

7.9 Chapter Summary

This chapter revealed how the analytical outputs of the machine learning models can be converted to business actions. Price prediction models allow uniform and competitive pricing, strategic segmentation and customer engagement through recommendation systems. These elements combine to create a holistic decision support model of the automotive retail stakeholders.

8. Conclusion, Limitations & Future Work

8.1 Summary of the Project

The dissertation aimed to explore the application of machine learning and business analytics to assist decision making since data, in the automotive retail sector. The project created an integrated analytical model based on exploratory data analysis, feature engineering, supervised price prediction models, unsupervised clustering and a recommendation system based on similarity using a real world dataset of car sales. Python was used to carry out and test all the analytical parts, making them reproducible and transparent.

Exploratory Data Analysis showed that there were strong nonlinear associations between vehicle price and the most important attributes of vehicle price mileage and vehicle age. Histograms and scatter plots prepared in the EDA phase revealed that the prices of vehicles are extremely skewed to the right and that depreciation decreases very fast with increase in mileage. The findings supported the necessity of nonlinear modelling methods and helped to design further feature engineering and algorithm selection.

Engineering features improved the performance of the models by adding derived features like normalization of vehicle age and usage measures. These characteristics better modelled real world depreciation behavior than univariate predictors and better predictive and cluster prediction tasks.

8.2 Important Conclusions and Their Applicability.

The guided learning experiment showed that ensemble models are much better than the old fashioned linear ones when it comes to predicting the prices of vehicles. Specifically, the Gradient Boosting and the Random Forest models have the lowest errors in prediction and the largest explanatory power, which proves that the models are appropriate to structured automotive data. The most important predictors that were always obtained in the feature importance analysis were mileage and vehicle age, then engine size and the manufacturer.

The clustering of the vehicles based on K Means indicated that there were four different vehicle blocks based on different market segments such as budget, mid range, premium and mixed markets. These groupings were very close to the automotive market structure of the real world and offered useful information in the inventory segmentation and formulation of pricing strategy.

The recommendation system was able to produce pertinent vehicle alternatives with the help of similarity measures based on engineered features. This module explained how analytics can improve customer decision support through complexity of search and the ability to make personalized recommendations.

Together, these results validate the idea that predictive modelling, clustering and recommendation systems have a high potential to be useful to automotive retailers when they are incorporated into a single analytical pipeline. The project provides the solution to the research issue by showing that machine learning can enhance the accuracy of pricing, identify worthy market segments and facilitate customer interaction.

8.3 Limitations of the Study

Although this study has its contributions, it has several limitations. First, it is based on structured vehicle properties and excludes behavioral data of customers, including browsing history or purchase intention, which might further increase the accuracy of recommendations. Second, marketing conditions like macroeconomic, fuel prices and seasonal demand were not considered which may have restricted the of price forecasts in different market conditions. Third, the models had been tested offline and not implemented in a real production environment.

8.4 Suggestions for Future Work

The proposed study may be enhanced in future study by incorporating real time information sources and external economic variables to enhance pricing stability. Adding customer interaction data would allow building hybrid or collaborative recommendation systems. Also, the implementation of the models into a live decision support system would provide the opportunity to learn continuously and monitor the performance. Lastly, explanatory AI methods might also be investigated to increase the trust and transparency of the automated pricing systems.

References

Breiman, L. (2001)
Random forests. *Machine Learning*, 45(1), pp. 5–32.
Available at: https://link.springer.com/article/10.1023/A:1010933404324

Domingos, P. (2012)
A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp. 78–87.
Available at: https://cacm.acm.org/magazines/2012/10/155147

Han, J., Kamber, M. and Pei, J. (2018)
Data Mining: Concepts and Techniques. 4th edn. Elsevier.
Available at: https://www.sciencedirect.com/book/9780128117606

Hastie, T., Tibshirani, R. and Friedman, J. (2009)
The Elements of Statistical Learning. 2nd edn. Springer.
Available at: https://hastie.su.domains/ElemStatLearn/

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021)
An Introduction to Statistical Learning with Applications in Python. Springer.
Available at: https://www.statlearning.com/

MacQueen, J. (1967)
Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium*, pp. 281–297.
Available at: https://projecteuclid.org/euclid.bsmsp/1200512992

Provost, F. and Fawcett, T. (2013)
Data Science for Business. O'Reilly Media.
Available at: https://www.oreilly.com/library/view/data science for/9781449374273/

Shmueli, G., Bruce, P., Yahav, I., Patel, N. and Lichtendahl, K. (2017)
Data Mining for Business Analytics. Wiley.
Available at: https://www.wiley.com/en us/Data+Mining+for+Business+Analytics p 9781118879368

Chen, Y., Zhang, L. and Chen, X. (2021)
Dynamic pricing of used vehicles using machine learning methods. *Expert Systems with Applications*, 180.
Available at: https://www.sciencedirect.com/science/article/pii/S0957417421007209

Rasheed, A. and Zhan, J. (2020)
Car price prediction using linear models. *IEEE Access*, 8, pp. 141623–141631.
Available at: https://ieeexplore.ieee.org/document/9141190

Brown, T. and Smith, J. (2020)
Machine learning in automotive pricing. *Journal of Business Analytics*, 12(3), pp. 244–260.
Available at: https://www.tandfonline.com/toc/ubaa20/current

Ricci, F., Rokach, L. and Shapira, B. (2015)
Recommender Systems Handbook. Springer.
Available at: https://link.springer.com/book/10.1007/978 1 4899-7637-6

García, S., Mendaña, R., Serna, P. and Molina, R. (2020)
Recommender systems in e-commerce: A review. *Information Systems Frontiers*, 22(2), pp. 343–362.
Available at: https://link.springer.com/article/10.1007/s10796-019-09954-8

Schafer, J., Konstan, J. and Riedl, J. (2001)
E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1), pp. 115–153.
Available at: https://link.springer.com/article/10.1023/A:1009804230409

Suresh, T. and Kumar, A. (2020)
Vehicle segmentation using K-means clustering. *International Journal of Data Science*, 5(2), pp. 89–101.
Available at: https://www.inderscience.com/info/inarticle.php?artid=109451

Porter, M. E. (1985)
Competitive Advantage: Creating and Sustaining Superior Performance. Free Press.
Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=193

Davenport, T. H. and Harris, J. G. (2007)
Competing on Analytics: The New Science of Winning. Harvard Business School Press.
Available at: https://store.hbr.org/product/competing-on-analytics/10183

Kotler, P. and Keller, K. L. (2016)
Marketing Management. 15th edn. Pearson.
Available at: https://www.pearson.com/en-us/subject-catalog/p/marketing-management/P200000003299

McKinsey & Company (2023)
The Future of Automotive Retail.
Available at: https://www.mckinsey.com/industries/automotive-and-assembly

Statista (2024)
Global used car market trends and size.
Available at: https://www.statista.com/markets/419/topic/491/vehicles/

OECD (2023)
Digitalisation of Automotive Markets. OECD Publishing.
Available at: https://www.oecd.org/industry/automotive/