Faculty of Engineering, Environment and Computing

School of Computing, Mathematics and Data Science

MSc. Data Science and Computational Intelligence

7151CEM Computing Individual Research Project

Predictive modelling and Analysis: Machine Learning in the UK automotive sector.

Supervisor: Prof. Vasile Palade

Name: Aditya Ashok Gujar

SID: 13366670

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science and Computational Intelligence

Academic Year: 2023/24

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking).

Signed: Aditya Ashok Gujar
Date: 08 Dec 2023

Please complete all fields.

First Name:	Aditya
Last Name:	Gujar
Student ID number	13366670
Ethics Application Number	P165625
1st Supervisor Name	Prof. Vasile Palade
2nd Supervisor Name	Prof. Xiaorui Jiang

This form must be completed, scanned, and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

This research project, titled "Predictive Modelling and Analysis: Machine Learning in the UK Automotive Sector," under the supervision of Prof. Vasile Palade, explores the application of machine learning (ML) algorithms to identify pricing patterns, sales trends, and market dynamics within the UK automotive industry. The study aims to address challenges faced by various stakeholders, including automakers, consumers, sellers, policymakers, and analysts, by providing practical insights and tools for decision-making. The project's objectives include developing ML models for forecasting sales patterns, determining used car prices, analyzing the impact of trim levels on consumer decisions, and enhancing decision-making processes within the UK automotive industry. The intended users of this project encompass a broad spectrum of stakeholders in the UK vehicle industry. These include automakers, consumers, sellers, policymakers, market analysts, researchers, and business intelligence professionals. The project aims to equip these users with valuable resources, such as tools and insights, to navigate the dynamic and diverse automotive landscape efficiently. The research methodology includes getting data from reliable sources, preprocessing it to make sure it is complete and of good quality, exploratory data analysis to get early insights, building ML models with algorithms like linear regression and random forest regression, and measuring success with metrics like mean absolute error and R-squared. The literature review provides insights from recent studies on ML applications in predicting used car prices, emphasising the importance of sophisticated ML strategies for accurate estimations. In summary, this research project endeavours to contribute to the enhancement of decision-making processes in the UK automotive sector through the application of machine learning techniques, providing valuable insights for various industry stakeholders.

Table of Contents

1. Introduction 6
   1.1. Problem Statement 6
   1.2. Intended Users 7
   1.3. Research Questions 8
   1.4. System Requirement 9
2. Literature Review 10
3. Methodology 15
   3.1. Dataset 16
   3.2. Data Cleaning and Preprocessing 16
   3.3. Exploratory Data Analysis 18
      3.3.2. Correlation Heatmap 21
4. Model Architecture 22
   4.1. Linear Regression 22
   4.2. Random Forest 24
   4.3. k-Nearest Neighbors(k-NN): 26
5. Training and Evaluation 28
   5.1. Feature and Target Variable 28
   5.2. Train-Test Split 28
   5.3. Evaluation 30
6. Results and Analysis 33
7. Project Management 34
   7.1. Meeting with Supervisor 34
   7.2. Legal Social Ethical and Professional Considerations 34
8. Critical Appraisal 35
9. Conclusion 35
10. Future Works 36
11. Self-Reflection 37

References 38

Appendix – A – GitHub Link 40

Appendix - B - Presentation 41

Appendix - C – Ethics Approval Certificate 50

ACKNOWLEGEMENT

I would like to extend my sincere thanks to Prof. Vasile Palade, my supervisor, for consistently offering assistance and guidance throughout the full duration of my research project. He regularly provided advice throughout the project and swiftly reviewed my progress. The remarks he supplied have expedited additional improvements in my work.

In addition, I would like to extend my appreciation to my family and friends for their steadfast support and motivation during my academic pursuit, which has continually motivated me to pursue excellence.

1. Introduction

The automobile sector holds a vital global position in the domain of machine learning applications. The utilisation of machine learning has gained significant prominence in automotive research and applications, highlighting the need for collaboration among marketing researchers, automotive designers, and computer scientists to tackle challenges in automotive exterior design, consumer analytics, and sales forecasting. The relationship between machine learning and data is inherent, but difficulties occur because of the confidential nature of important commercial information, such as product sales, which hampers accessibility for researchers. The provided research samples illustrate the dataset's ability to meet the data requirements for business analysis, product design, and market forecasting studies. The dataset aims to be a valuable resource for researchers and practitioners in various fields involved in economic or business-related automotive research. Our suggested dataset, which includes extensive product information gathered from many internet sources, is intended to facilitate the development of additional product datasets for research purposes in various fields such as business, design, and computer science. This work makes major contributions in various aspects. Within the sphere of application, the dataset specifically caters to the need for a thorough automobile dataset for the purpose of economic and business research. This dataset proves advantageous to manufacturers, customers, and marketers. Furthermore, from a machine learning standpoint, the dataset is notable for being the initial extensive automotive dataset, containing approximately 0.3 million specificities from 899 different car models, along with matching specifications and sales data that covers a period of over ten years in the UK market. It meets the criteria of big data, making it appropriate for analytical and predictive work in several fields. Furthermore, with regards to data organisation, the strategy addresses the difficulties identified through a survey study including researchers in the fields of business and computer science, with the goal of making a significant influence in both research and practical applications. Finally, the approach helps to tackle issues associated with integrating varied data formats from various sources, enabling effortless scalability with fresh data and widespread adoption by academics.

1.1. Problem Statement

Attempting to traverse the complex realm of the UK automotive sector can be a challenging endeavour. There are numerous components in motion - deciphering sales patterns, refining price tactics, and analysing model alternatives. Determining precise valuations for pre-owned vehicles can be a complex task. Given the multitude of obstacles, it is arduous for enterprises, customers, and regulators to make sound decisions. Hence, I aspire to develop machine learning model to effectively navigate through complexity. I aim to enhance the influence and authority of significant stakeholders, such as automobile manufacturers, purchasers, and policymakers. I aim to

equip them with valuable insights and user-friendly resources, enabling them to effectively address prevalent challenges in the field. I am excited to construct models that can predict future sales patterns. Thus, automakers can make more intelligent planning judgements at present. My goal is to provide clarity in the complex realm of used car pricing. Customers are entitled to transparent and equitable information while they are evaluating different choices. Examining the influence of trim upgrades on consumer preferences is crucial as it provides valuable insights to automakers on the development of innovative and luxurious features. My main objective is to improve decision-making for everybody involved in the UK automotive industry. In the end, my goal is not only to identify challenging problems in the sector. I aim to make a significant impact by offering cutting-edge machine learning solutions. Envision a scenario in which prominent individuals possess immediate access to forecasts, lucidity, and suggestions. They possess the ability to respond more quickly, construct superior products, formulate strong policies, and make transactions with assurance. This is the significant and positive change that I foresee my work facilitating in the automotive industry in the United Kingdom.

1.2. Intended Users

- Our automotive dataset and machine learning research aim to benefit many stakeholders involved in the UK car industry. Firstly, we have the automakers, ranging from prominent brands such as Jaguar Land Rover to specialised manufacturers like Lotus. We anticipate that our findings will assist in streamlining manufacturing processes to align with consumer preferences and provide guidance for focused marketing efforts.

- Furthermore, both prospective buyers of new and pre-owned vehicles can utilise model data to improve their purchasing selections by considering pricing estimates and comparing values. Sellers of pre-owned autos may find opportune timing for marketing their cars. Our findings has the potential to impact policymakers who are assessing incentives and investments in electric vehicles based on demand predictions.

- From a broader perspective, we anticipate that academics will disseminate our data across various studies to uncover fresh insights into consumer behaviours and other related aspects. Business analysts have the ability to utilise predictive capabilities to implement data-driven initiatives.

- Our primary objective is to create a highly valued asset that encompasses a wide range of tools, models, visualisations, and insights. This resource will be beneficial for anyone involved in the ever-changing UK automotive industry. By equipping important stakeholders with a clear understanding of the present and future dynamics of this intricate and evolving sector, we anticipate the emergence of

significant positive impacts on a large scale. It is imperative to address the current ambiguity around matters in the automotive industry.

1.3. Research Questions

1. Can machine learning models effectively track and predict pricing patterns and strategies employed by different automakers?

    - How accurately and granularly can algorithms discern and analyze the pricing patterns utilized by individual automakers within the dynamic and competitive landscape of the automotive industry? Can meaningful pricing insights be extracted?
    - To what extent can sophisticated machine learning models uncover hidden nuances and strategic subtleties in the pricing trends existing within distinct automotive market segments? Could revelations emerge that contribute to a more comprehensive understanding of segment-specific pricing dynamics?
    - Which specific methodologies among the diverse range within machine learning have shown the greatest proficiency at unveiling even subtle variations in the pricing strategies and approaches implemented by various automakers? Would hybrid models combining algorithms like regression and neural networks yield optimal performance?

2. What are the primary factors, both individual and combined, that prove most instrumental in driving significant shifts - both increments and declines - in the annual and quarterly sales volumes and figures associated with individual car models? Additionally, how do the influences of those key determinants evolve when analyzed over extended multi-year periods?

    - Which tangible factors, such as pricing, styling, performance, brand reputation, etc., as well standalone intangible elements like consumer sentiment, reviews, and market conditions, play the most pivotal roles in influencing the sales figures, volumes and growth trajectories of specific car models year over year? Moreover, how do those major contributors interact in a dynamic manner to collectively bring about the rises and falls observed in market demand?
    - How does the relative isolated impact of tangible external variables like economic cycles, gas prices, infrastructure, and technology innovations compare and contrast to inherent model-specific factors over evolving long term periods? Are there complex cause-effect relationships between external and internal variables that machine learning algorithms could uncover to produce insightful sales predictions?
    - Could imaginative applications of computer vision, natural language processing and other sophisticated machine learning techniques surface

non-intuitive revelations into the key elements catalyzing shifts in sales figures historically? Beyond that identification capacity, can algorithms predict how the coalescing influence of those factors might evolve in the 3-5 year horizon?

3. Stepping back categorically, do data analyses and machine learning modeling point to any overarching patterns or macro trends that might characterize the dynamically changing landscape of diversity and popularity among automotive manufacturers over modern decades and even further back historically?

    - Can innovative machine learning pipelines tapping structured databases as well as unstructured data in the form of articles, ads and social media threads reveal insightful and unexpected patterns in the rising, falling, shifting and merging trajectories of automakers? Could competitive dynamics be illuminated?
    - To what extent do thoughtfully compiled and organized historical sales data combined with specialized predictive modeling contribute to a more informed understanding of the complex factors, consumer behaviors and competitive forces catalyzing ongoing changes in the popularity, relevance and diversity mix of players within the broadly defined automotive industry?
    - How can creative statistical analyses and finely tuned machine learning engines be applied to shed new light on the market forces, consumer preference evolutions, and manufacturer innovations/missteps that have driven seismic shifts in the popularity, reputation, and diversity mix of automakers over modern decades? What future fluctuations seem likely based on those retrospective revelations?

1.4. System Requirement

For our project, we'll need to set up a properly equipped programming environment. The core will be Python 3.x - the latest and greatest version. On top of that, we plan to leverage handy Python libraries, including Matplotlib for graphics, Pandas for data analysis, NumPy for numerical computing, and SciPy for technical computing routines. To tie everything together, we want an integrated development environment (IDE) that streamlines building out machine learning models. Some top options are PyCharm, VSCode, Jupyter Notebooks, or the Anaconda suite. These IDEs make coding so much easier with built-in tools, debugging, and more. On the hardware side, we'll be working on our trusty Windows 10 laptop. It packs substantial power with an Intel Core i7-8550U CPU, 16GB of RAM, and other solid specs that can handle intensive programming and data workloads.

2. Literature Review

Constructing a comprehensive literature review is essential for effectively contextualizing our research. The approach involves doing a comprehensive survey of pertinent studies in various sources such as books, journals, conferences, and other relevant materials to gain a deep understanding of the academic context related to our automobile dataset project. By consolidating existing scholarly findings on machine learning applications and prediction modelling in the automotive industry, we can strategically position our own contributions. Utilising a broad perspective is crucial - we thoroughly investigate several areas of research, ranging from projecting carbon emissions and the transition to self-driving vehicles, to optimising manufacturing processes and anticipating vehicle trade-in values. Step one involves showcasing mastery of the intellectual domain. Subsequently, we systematically categorise recurring patterns related to problems such as data insufficiencies and real-world accuracy obstacles encountered in prediction models. And compare contrasting assumptions amongst different academic fields. By mapping the conceptual battlefield, we can identify significant challenges that require attention in order to enhance our common understanding. Furthermore, this comprehensive overview of past research highlights significant deficiencies in methodology and specific areas of study that require immediate further investigation.

- Precisely forecasting the costs of pre-owned vehicles is a significant and intricate real-life issue that has significant economic ramifications. This study by (Pudaruth, 2014) examines supervised machine learning methods for predicting the prices of secondhand cars in Mauritius, utilising historical data from ads. Nevertheless, numerous research exhibit certain limitations. The majority of individuals depend on exclusive data or limited sets of features, which undermines the ability to reproduce results. Elaborate models are built without thorough validation. Comparative benchmarking is hindered due to the absence of disclosed standard accuracy criteria. Research gaps exist in the development of clear, robust, and extensible machine learning formulations using easily accessible data. This paper is the initial endeavour to develop a model for estimating the prices of pre-owned cars specifically designed for the Mauritian market. It utilises straightforward and interpretable algorithms such as linear regression, kNN, decision trees, and Naive Bayes on publicly available data. Data pre-processing, including normalisation and cleaning, is performed. Evaluating performance is done methodically by utilising criteria such as mean absolute error and test accuracy. The constraints include the limited dataset size and the restricted complexity of the model. However, it establishes a valuable starting point for more sophisticated Mauritian used automobile value models. To summarize, although the paper's effort may be considered modest in comparison to the most advanced methods, it is noteworthy for being the first to apply machine learning techniques to anticipate used automobile prices in Mauritius. It establishes the groundwork for the implementation of larger-scale models that may be utilised by both dealers and customers to their advantage.

- Determining the price of pre-owned vehicles is a multifaceted procedure that relies on several aspects such as the brand, model, mileage, and condition. Given the swift expansion of the worldwide used automobile marketplaces, precise price prediction has gained significance in ensuring equitable transactions between buyers and sellers. In this study conducted by (Bao, 2023), the author aims to forecast the pricing of pre-owned vehicles by employing advanced machine learning methods such as ISOMAP for reducing the complexity of the data and Support Vector Machines (SVMs) for training the predictive model. Many analyse performance based on limited private datasets, which limits its applicability to other cases. Deep learning models with high complexity tend to overfit when trained on little data. There is a lack of thorough optimisation of hyperparameters and analysis of the model. This highlights the deficiencies in creating large-scale models that are both straightforward, easily understood, and resistant to errors. Bao (2023) aims to fill these deficiencies by utilising a dataset including 400,000 publicly accessible records. The use of ISOMAP for dimensionality reduction is based on its ability to reveal underlying manifold structures. The Support Vector Machine (SVM) is favoured due to its capacity to generalise effectively even with limited training data. There is a valid rationale for the author's decisions. Nevertheless, ISOMAP renders the model intricate and obscure in contrast to more straightforward alternatives such as PCA. The abysmal 14.3% test accuracy suggests a significant problem of overfitting, most likely caused by ISOMAP disregarding features that are important for predicting prices. There is undoubtedly potential for enhancement through meticulous feature engineering, hyperparameter optimisation, and thorough analysis. To summarise, the work examines effective methods for a crucial issue, but significant improvements in the precision, transparency, accuracy, and applicability of the model are necessary to further enhance research and practical applications. The acquired knowledge facilitates the creation of advanced and reliable data-based models for valuing pre-owned cars.

- Precise forecasting of used automobile prices has become crucial in light of the flourishing used car marketplaces worldwide. This article by (Bukvi´c , Škrinjar, Fratrovi´c, & Abramovi´c, 2022) (Pudaruth, 2014)examines the use of data-driven models that utilise vehicle variables such as age and mileage to predict pricing in Croatia, utilising data from online automobile marketplaces. This study represents one of the initial endeavours to customise the modelling of used automobile prices specifically for the Croatian context, utilising available web data. Performing suitable pre-processing techniques such as eliminating outliers and conducting graphical analysis helps to reveal subtle details within the dataset. Correlation analysis is used to maintain important predictive attributes such as price, mileage, and manufacture year, which helps in developing transparent models. Comparing standard linear regression with advanced classifiers demonstrates that simpler

models can be enough when high-quality data is available. Subsequent dataset analysis confirms that there is a correlation between price inflation and mileage loss, which can be attributed to macroeconomic disturbances. To summarise, although the work is not as advanced as the latest research, it introduces a preliminary attempt to study used automobile valuation in Croatia by utilising open online data and structured machine learning techniques. There is potential for the development of more comprehensive models that incorporate additional variables such as vehicle condition and other aspects that influence pricing. The primary constraints are the limited capacity for model intricacy and the brief duration of the time frame. However, it provides a cost-effective, easily understandable, and adaptable framework for determining the prices of secondhand cars.

- The passage provides an overview of a machine learning endeavour conducted by (Grelle, et al., 2021) that focuses on forecasting the prices of pre-owned vehicles in the United Kingdom. Utilising predictive analysis to determine the value of pre-owned vehicles facilitates equitable pricing for both buyers and sellers in the face of rapidly expanding worldwide markets. Many previous studies rely on exclusive datasets, intricate opaque models, and lack thorough evaluation, which hampers reproducibility, insights, and reliability. This project offers various advantages. Performing suitable data preprocessing techniques such as data cleansing and graphical analysis can reveal intricate details and characteristics of the dataset. Comparing interpretable algorithms such as decision trees and gradient boosting machines on publicly available data fosters confidence and clarity. The reported accuracy range of 90-97% on the held-out test data demonstrates impressive performance, provided it is confirmed by thorough cross-validation. Categorising models based on car manufacturers allows for tailored customisation to specific industry intricacies. Nevertheless, there are certain constraints. The information appears to be limited to 100,000 automobiles, which hinders the possibility of implementing it on a bigger scale. Enhancements to model governance, such as tuning and statistical significance assessment, could increase confidence levels. An analysis of feature importance would yield explanatory insights into prediction patterns. The impact of macroeconomic conditions on price has not been resolved. However, the systematic process of experimentation is promising for the team's understanding and use of machine learning.

- Research by (Narayana, Likhitha, Bademiya, & Kusumanjali, 2021) The technique, encompassing data collection to model evaluation, will be comprehensively explained. A dataset of more than 4,000 records of used automobile sales was collected, with information such as the car's make, model, year, mileage, fuel type, gearbox type, seller type, and sale price. The data was analysed and presented visually to uncover valuable insights about the pre-owned automobile industry. Subsequently, the data underwent meticulous preprocessing and feature

engineering in order to make it suitable for machine learning. Ultimately, the models of linear regression, decision tree regression, and random forest regression were both trained and assessed. The random forest model exhibited superior performance, boasting an accuracy of 85% and minimal error rates. This model has been deployed with an interactive online interface to demonstrate its practical utility. In summary, this project showcases the complete implementation of predictive analytics and machine learning to address a practical issue. The ultimate random forest model can consistently provide accurate guidance for determining the appropriate pricing of pre-owned vehicles. A significant amount of knowledge was acquired regarding the process of gathering and organising data, training models, and assessing their effectiveness. Enhancing accuracy could be further advanced by including more intricate automotive features and comprehensive sales data in the future. The author anticipates that the reader will find this report and the underlying approach to be beneficial, as it provides explanatory insights into predicting patterns. The impact of macroeconomic conditions on price has not been resolved. However, the systematic process of experimentation indicates that the team has a strong understanding of applied machine learning. Ultimately, this innovative endeavour establishes the basic foundation for developing a specialised model to predict the prices of pre-owned vehicles in the UK market. Enhancing the complexity of the model, expanding the range of the dataset, improving interpretability, and conducting further analysis on pricing factors might solidify the practical significance of the results. Commercial teams and policy strategists can utilise these insights to mitigate discrepancies in the high-stakes used vehicle market. Additional scholarly contributions have the potential to enhance financial inclusion in the face of unpredictable macroeconomic conditions.

- The study authored by (Mahfouz, Mosaad, & Belal, 2023) centres on the utilisation of machine learning techniques, specifically employing a federated learning strategy, to predict automobile pricing. The authors employ a combination of numerous machine learning models and federated learning to forecast the values of pre-owned vehicles. This is done by considering several factors, including mileage, fuel type, gearbox type, and number of previous owners. The work is positioned in the broader realm of research that use machine learning methods for sales and demand prediction in many industries. More precisely, it pertains to the current body of work on using data-driven models to anticipate prices and simulate customer purchasing behaviour in the automotive industry. In recent times, the rise of machine learning has led to an increased use of tree ensembles, neural networks, and support vector machines for prediction tasks. However, only a small number of research have utilised sophisticated machine learning techniques to create detailed models for predicting car prices. The federated learning approach developed by Mahfouz et al. fills a gap in the literature. This research is innovative because it adapts advanced federated learning approaches to leverage the

predictive capabilities of machine learning using scattered localised data sources in the automobile field. Empirical evaluations of models such as Fed-KNN, Fed-SVM, and Fed-Random Forest have shown that they achieve higher accuracy compared to traditional alternatives. The authors additionally create suitable evaluation metrics, data preprocessing techniques, feature selection methods, and model validation protocols for the purpose of predicting automobile prices. This research significantly contributes to the advancement of machine learning applications in anticipating consumer-facing pricing and demand in the automotive industry. The combination of the federated learning paradigm and thorough empirical research offers a framework for domain-specific predictive modelling, while also addressing the growing concerns regarding data privacy, security, and accessibility. Subsequent investigations can expand upon these methods to improve linked vehicle technology, personalised pricing engines, and inventory planning systems

3. Methodology

Engaging in this research activity necessitates a comprehensive and methodical methodological approach. Commencing with the data gathering phase, it is imperative to get relevant datasets from trustworthy sources, encompassing the specifics of the "Sales Table," "Price Table," "Trim Table," and "Ad Table." An unwavering emphasis on the precision and comprehensiveness of data serves as the basis for subsequent investigations. The data pretreatment stage seamlessly shifts its focus towards the proactive duties of cleaning and preprocessing. Addressing missing values and outliers, as well as providing uniform data formats, are essential tasks that facilitate the smooth integration and structuring of datasets, laying the groundwork for effective analysis.

The next step entails the use of exploratory data analysis (EDA), which is akin to initiating a conversation with the data. It is essential to have a comprehensive understanding of the dataset's fundamental characteristics by visualising patterns, correlations, and distributions. As we delve further into the realm of machine learning model creation, our focus shifts towards constructing predictive models that address various objectives, such as projecting sales patterns, estimating pricing, evaluating used automobile worth, and examining trim levels. The meticulous selection of algorithms, such as the widely-used linear regression and the dependable random forest regression, emphasises the significance of this critical phase. Validating these models is essential to assure their efficacy and dependability in the analytical domain.

Data visualisation functions as the aesthetic portrayal of important findings and results in the artistic domain. The researcher aims to effectively convey findings to stakeholders by generating visual depictions, thus streamlining intricate data into a cohesive narrative. Within the realm of evaluation metrics, like to a coach assessing team performance, the researcher constructs and employs measurements such as mean absolute error (MAE), root mean square error (RMSE), R-squared, accuracy, precision, recall, or F1-score. These metrics provide a thorough assessment of the models' performance in both regression and classification contexts.

The research narrative culminates in its most significant stage during the reporting and documentation phases. Elaborate reports are meticulously composed, encompassing all facets, ranging from the methodology employed to the discoveries and suggestions. The research highlights the ethical dimensions, showcasing a firm commitment to maintaining ethical standards, particularly when handling sensitive data or developing pricing predictions. The researcher's dedication to comprehending the intricacies of the UK automotive industry is evident through their meticulous and comprehensive methodology, which include employing predictive modelling and analysis.

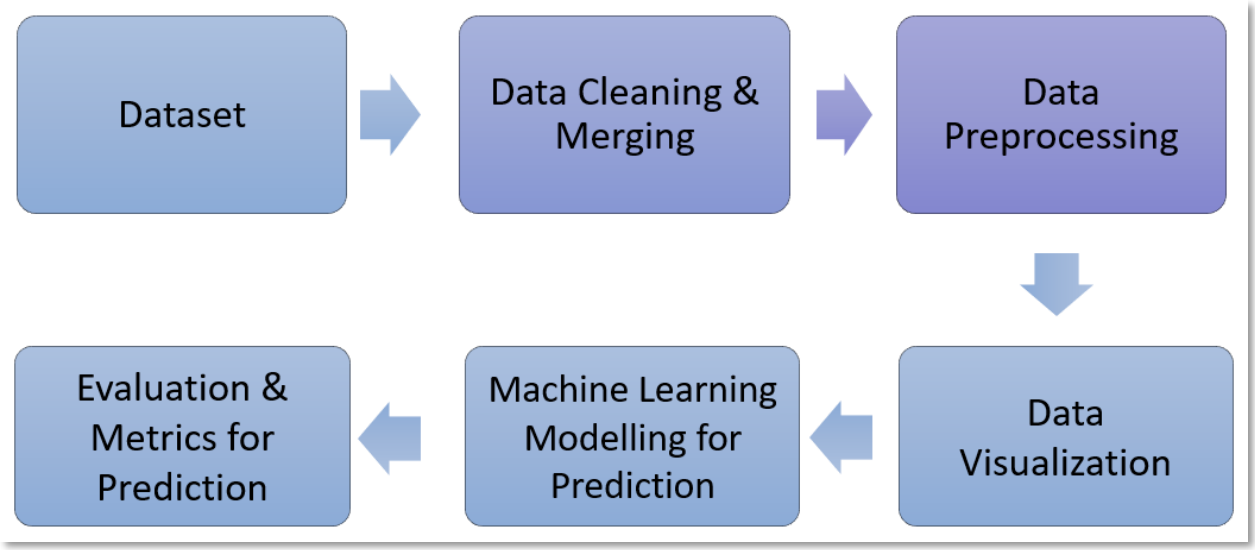


Figure 1-Project Workflow

3.1. Dataset

The dataset used for the project is sourced from Deep Visual Marketing, also known as DVM-Car 2023, which can be accessed at https://deepvisualmarketing.github.io/. A survey was done using Qualtrics to examine common data difficulties in research and practice. The study collected responses from 54 researchers. The group consisted of 26 individuals with a computer science background and 28 individuals with a business studies background, including areas such as economics, marketing, and management.

The study results categorise difficulties into three overarching domains: coverage, accessibility, and quality. Coverage issues happen when a dataset is lacking essential information, either because it is not comprehensive enough or because it has an insufficient number of records or samples. The employed dataset in this context is a comprehensive resource that has been rigorously constructed to include crucial automotive specs and sales variables that are significant to marketing research.

The collection includes a vast number of registered cars in the UK market over several decades. It serves as a valuable collection, intertwining the stories of numerous automobiles, each adding a distinct narrative to the overall fabric of the automotive industry.

3.2. Data Cleaning and Preprocessing

In the field of data science, raw data is sometimes compared to unprocessed crude oil extracted from deep underground. While it possesses energy and potential, it requires

refinement before it can effectively fuel our systems. Prior to utilising raw data for analytics, modelling, and visualisation, it usually necessitates refining and modification, which is sometimes referred to as data pre-processing. View data pre-processing as a comparable operation to an intake and sorting facility located at the beginning of a power plant. Data streams from multiple sources are disorganised and contaminated with dust, containing discrepancies in format, errors, redundant observations, and missing information. Our data pre-processing facility subjects each data load to a rigorous quality assessment and conditioning procedure. This process aims to identify and remove anomalies, resolve any discrepancies, enhance the context of the data where necessary, and condense everything into standardised and well-organized data packages that are ready for analysis. Particular pre-processing jobs encompass data cleansing, data integration, data minimization, and data transformation, among other duties. Data cleaning is the process of identifying and rectifying any inaccurate, redundant, or unfinished items that necessitate adjustments or removal. Data integration is the process of merging different data sources into cohesive and consistent datasets. Data reduction is a process that reduces the amount of data by removing unnecessary repetition and concentrating on important information. Data transformation is the reorganisation of raw data types into objects that are suitable for analysis and can be used with our software tools, such as matrices or data frames (Coventry University, n.d.).

```
#Dropping unwanted columns
df_trim = df_trim.drop(columns=['Maker','Genmodel'])
df_sales = df_sales.drop(columns=['Maker','Genmodel'])

# Merge the tables
merged_data_left = pd.merge(df_basic,df_trim, on='Genmodel_ID', how='outer')
merged_data_left.head()

# Merge the tables
merged_data_left = pd.merge(merged_data_left,df_sales, on='Genmodel_ID',how='outer')

# Droping NaN Values
merged_data_left = merged_data_left.drop(merged_data_left.index[0])
merged_data_left = merged_data_left.dropna()
```

Figure 2 Data Cleaning

This code snippet is specifically created for the purpose of cleaning and prepping data in the context of merging and managing missing values in a dataset that is represented by Pandas DataFrames. The columns 'Maker' and 'Genmodel' are eliminated from the DataFrames df_trim and df_sales since they are deemed useless for the upcoming study. The code subsequently merges two DataFrames, df_basic and df_trim, by performing an outer join operation based on the 'Genmodel_ID' column. The resulting DataFrame, named merged_data_left, is further merged with df_sales using the 'Genmodel_ID' column and an outer join. After performing these procedures, the initial row of merged_data_left is excluded to eliminate any possible NaN values that may have occurred due to the outer join. Afterwards, any residual rows that have NaN values are

methodically eliminated, guaranteeing the overall integrity and comprehensiveness of the dataset. The purpose of this methodical data preprocessing strategy is to optimise the dataset for future analyses, modelling, or visualisation. It achieves this by resolving problems associated with unnecessary columns and missing values through a sequence of focused procedures.

Label encoding is a straightforward and efficient method for transforming category information into numerical representation. The LabelEncoder class from scikit-learn allows for simple encoding of categorical data, facilitating its preparation for subsequent analysis or input into machine learning algorithms (Great Learning Team, 2023).

```
46  from sklearn.preprocessing import LabelEncoder
47
48  label_encoder = LabelEncoder()
49  merged_data_left['Automaker'] = label_encoder.fit_transform(merged_data_left['Automaker'])
50  merged_data_left['Genmodel'] = label_encoder.fit_transform(merged_data_left['Genmodel'])
51  merged_data_left['Trim'] = label_encoder.fit_transform(merged_data_left['Trim'])
52  merged_data_left['Fuel_type'] = label_encoder.fit_transform(merged_data_left['Fuel_type'])
```

Figure 3 Label Encoder

In the above code, the LabelEncoder class from the scikit-learn library to preprocess categorical data. The goal is to convert categorical variables into numerical representations, making it easier to use machine learning methods that specifically demand numerical inputs. The transformation is applied to four specific columns, namely 'Automaker,' 'Genmodel,' 'Trim,' and 'Fuel_type,' in the DataFrame called 'merged_data_left'. The label_encoder object is created, and the fit_transform method is used to encode the categorical values in each column. By performing this procedure separately for each column, the categorical labels are substituted with numerical equivalents, resulting in a more appropriate format for subsequent analysis or model training. This method is very useful when working with machine learning models that require numerical input features, improving the overall compatibility and efficacy of the dataset for predictive modelling and analytical tasks.

3.3. Exploratory Data Analysis

Exploratory data analysis (EDA) is an essential initial step in analysing a new data set. Prior to engaging in formal modelling or statistical testing, exploratory data analysis (EDA) enables us to gain a more profound understanding of the subtleties and complexities inside our data using a variety of quantitative and visual inspection tools (Patil, 2018). In essence, Exploratory Data Analysis (EDA) encompasses the essential tasks one must perform to gain a deep understanding of the available data. Before attempting to model or draw inferences from our data observations, it is important to thoroughly examine and understand the profiles, contours, personality quirks, and backstories that exist within them, just as you would want to know someone well before making significant judgements or predictions about their future behaviour. When conducting exploratory data analysis

(EDA), our specific objectives include identifying abnormal data points that deviate from anticipated patterns, uncovering the existence of missing values or outliers that could distort results, examining hypotheses regarding relationships between variables, and assessing whether crucial modelling assumptions, such as linearity, normality, homoscedasticity, or independence, are satisfied. In addition, we utilise Exploratory Data Analysis (EDA) to identify inherent patterns, clusters, or latent structures within the data that were not initially expected.

3.3.1. Data Visualization

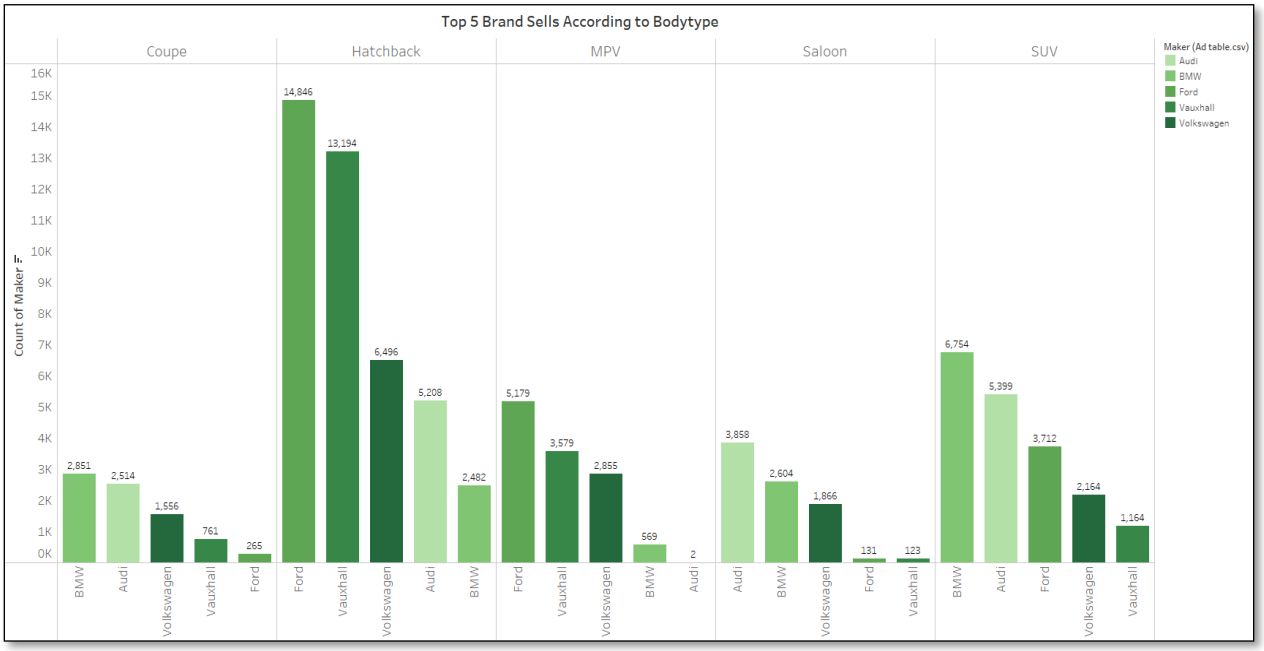


Figure 4 Top 5 Brand Sells according to Bodytype

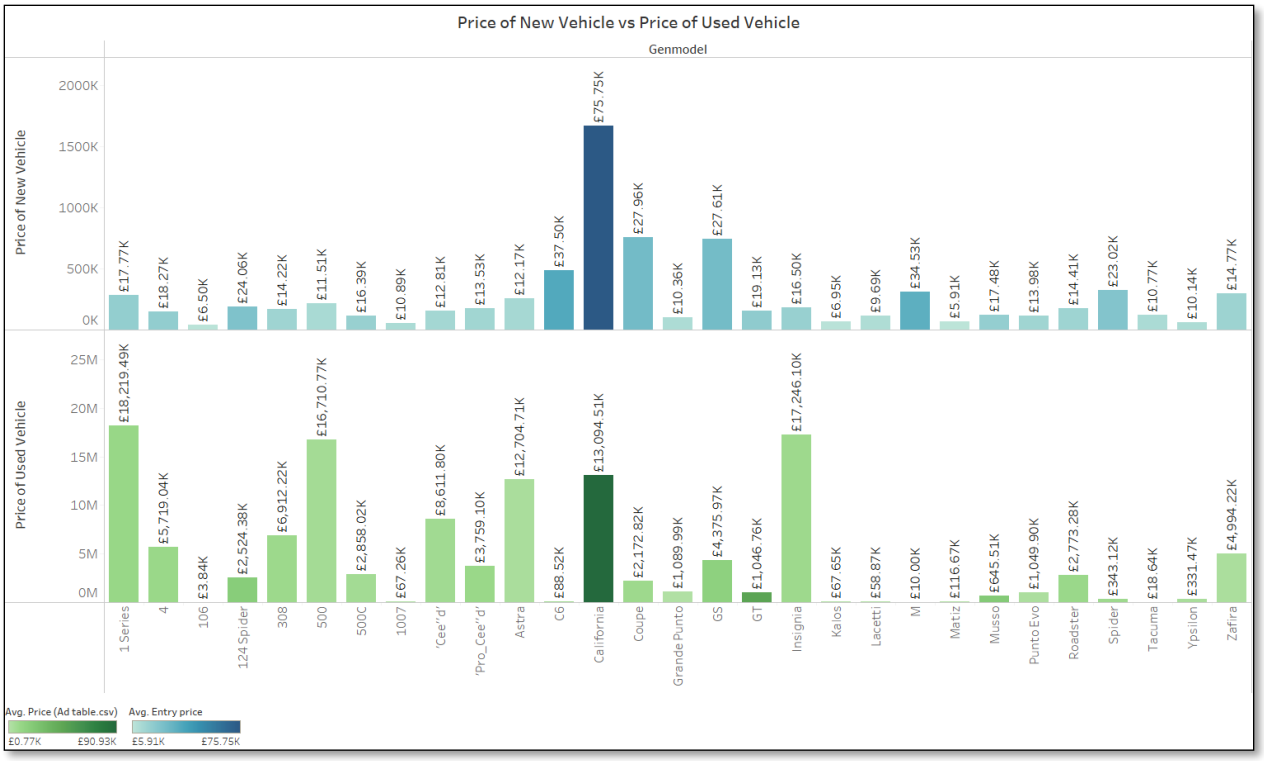


Figure 5 Price of New Vehicle vs Used Vehicle

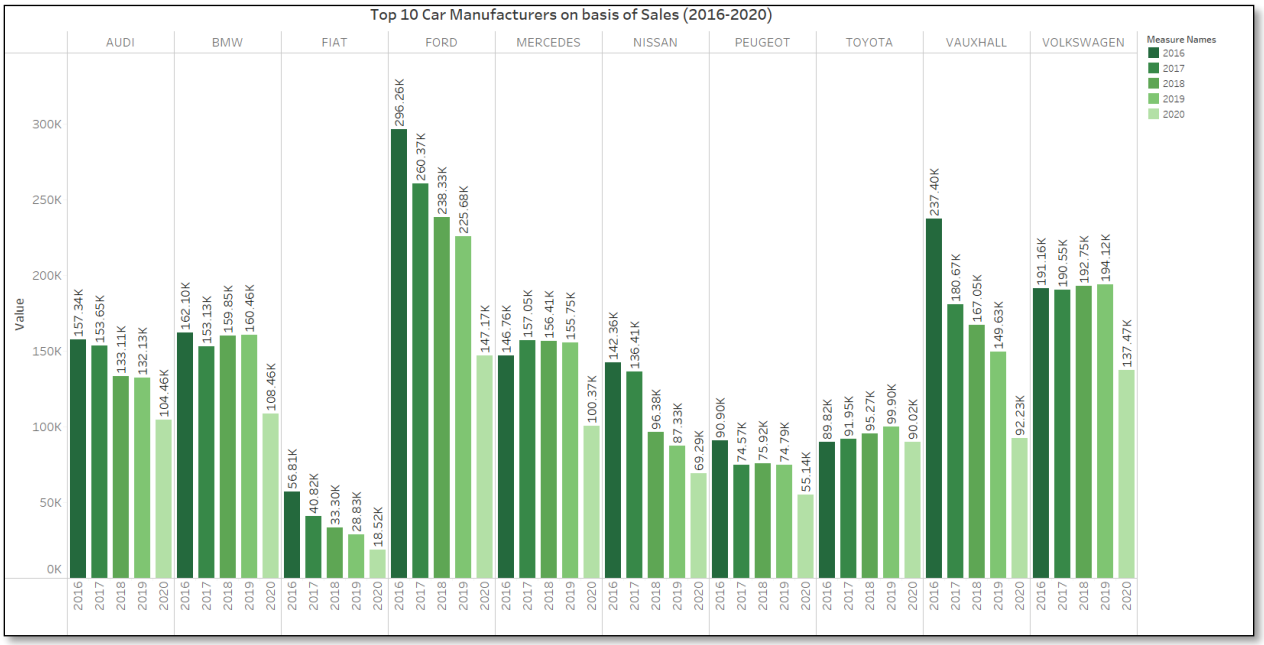


Figure 6 Top 10 Car Brans on basis of Sales(2016-2020)

3.3.2. Correlation Heatmap

Heatmaps are highly useful visual aids for identifying trends and patterns in data across time. Heatmaps utilise colour intensities to encode values, enabling immediate identification of anomalies, clusters, and cycles as they progress over time. Nevertheless, traditional static heatmaps have obstacles when dealing with large time series or datasets with a high number of distinct values. Attempting to create a single massive heatmap becomes counterproductive when dealing with numerous data points over multiple time intervals. The pertinent signals become overwhelmed by an abundance of extraneous noise. In order to address this challenging situation of finding a certain item among many others, we rely on interactive heatmap solutions. Interactive heatmaps allow users to easily filter datasets along both axes using customizable inclusion-exclusion rules. Instead of confronting a large, unified system, one now uses sidebar widgets to selectively choose certain sections of interest. This may entail identifying significant monthly or hourly patterns within yearly cycles or focusing on a select few top performers among numerous product lines (Kumatani, Itoh, Motohashe, Umezu, & Takatsuka, 2016).

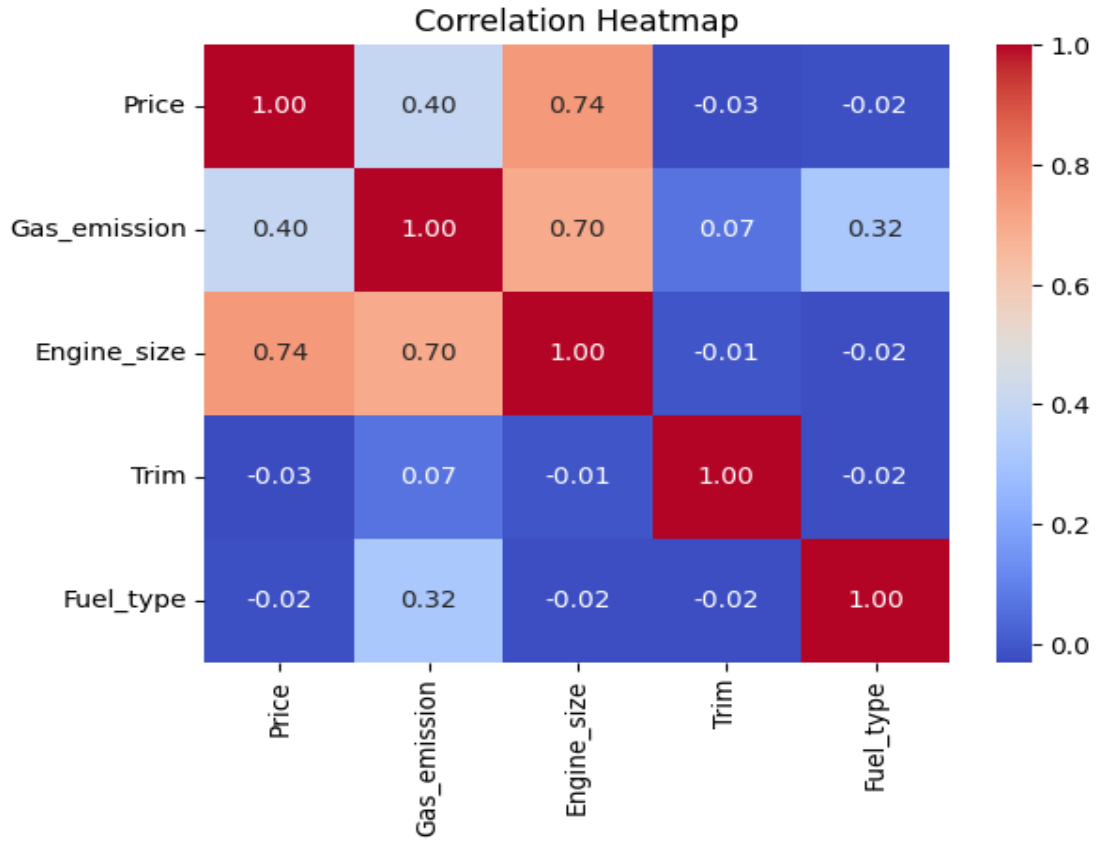


Figure 7 Correlation Heatmap of Target Variables

```
# Correlation heatmap
correlation_matrix = merged_data_left[selected_features].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

Figure 8 Correlation Heatmap

The Seaborn heatmap displayed above illustrates the correlation matrix as a grid. The grid includes price, emissions, engine size, trim, and fuel type as the axes. A colour gradient is used to indicate the strength of the statistical relationship between two attributes, whether it is strongly positive, strongly negative, or negligible. Additionally, the correlation coefficients are printed within each cell. In order to provide immediate clarity regarding the content of this visual representation, we provide it a descriptive term, namely 'Correlation Heatmap'. Lastly, we utilise the pyplot.show() method from matplotlib to display the correlation heatmap visualisation. By constructing and examining such a plot, we promptly identify which characteristics of the car display intriguing correlations that may necessitate additional predictive modelling or research to provide valuable insights. The result provides valuable insights that inform subsequent actions for more sophisticated analytics, helping both data science aims and business goals.

4. Model Architecture

4.1. Linear Regression

Linear regression is a type of supervised learning technique that analyses labelled data by comparing input variables (X) and output variables (Y). It is utilised to determine the correlation between two variables and forecast future outcomes based on previous associations (What is Linear Regression?, n.d.).

The objective is to determine the most accurate linear regression model that represents the associations within the vehicle dataset. Optimally, the line of best fit minimises the overall prediction error among the data points, visually represented by the line that is closest to all plotted points. In mathematical terms, it is desirable for each data point to exhibit minimal deviation from the values predicted by the regression line. This line effectively achieves a balance between representing the general patterns in the data and acknowledging the natural fluctuations within the data. The process of modifying the line of best fit involves utilising statistical analysis and visualisation tools to accurately represent patterns within a certain margin of error. Plotting prediction errors aids in fine-tuning the accuracy of the line in summarising the primary trend of the data while avoiding exaggeration of random fluctuations (What is Linear Regression?, n.d.).

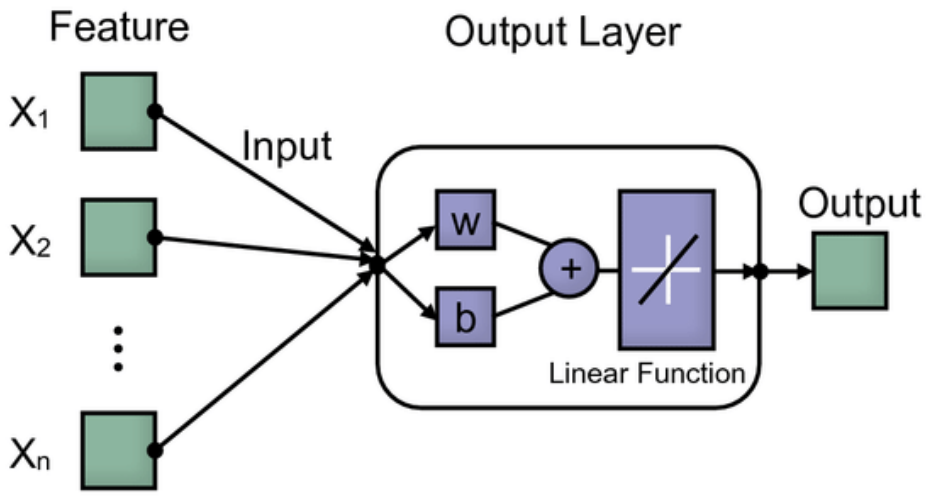


Figure 9 Linear Regression Architecture (Hong Min & KooYoon, 2021)

The structure of linear regression is quite straightforward. The model comprises a solitary layer of input nodes that represent the predictor variables, along with a lone output node. The fundamental equation for a straightforward linear regression is represented as (Y' = a + bX), where (Y') denotes the anticipated score, (b) signifies the gradient of the line, and (a) represents the y-intercept. The objective of the linear equation is to determine the line that optimally corresponds to the data, by minimising the overall prediction error, represented as the smallest distance between each data point and the regression line (Hong Min & KooYoon, 2021) (Statistics How To, n.d.) (What is Linear Regression?, n.d.).

The equation of the line follows the format (Y = a + bX), where (X) represents the independent variable and (Y) represents the dependent variable. The line has a slope of (b), and the intercept is (a) which represents the value of (Y) when (X = 0) (Hong Min & KooYoon, 2021) (Statistics How To, n.d.) (What is Linear Regression?, n.d.).

Advantages

- Simple Implementation: One of the main advantages of linear regression is its ease of implementation and interpretation, which are useful qualities when first exploring predictive modelling. By estimating a single line, one can easily evaluate initial concepts before exploring more complex alternatives (Rout, n.d.) (Pranavnath, 2023).
- Efficient to train: Model training is highly efficient, as it only requires optimizing the slope and intercept of a line. When there are strong linear tendencies, the process of finding the line that best fits the data is accomplished at a remarkable pace. This enables the rapid iteration of various modelling experiments (Waseem, 2023).
- Handles Overfitting: Compressing intricate data into a solitary linear correlation inherently combats overfitting, offering a form of regularisation. The inclusion of

supplementary mathematical restrictions and the use of cross-validation further strengthens the emphasis on fundamental patterns rather than idiosyncrasies. Therefore, linear regression exhibits a reasonable level of generalization (Waseem, 2023).

- Extrapolation: The linear equation allows for carefully expanding projections slightly beyond the training distribution for new inputs. Although not universally applicable, this cautious extrapolation occasionally provides additional usefulness (Waseem, 2023).
- Interpretability: The model's transparent linear structure allows for a direct examination of the relationships between input and output variables in the data. The ability to interpret the results is extremely valuable during the initial stages of exploration, prior to achieving optimised predictions (Rout, n.d.).

Disadvantages

- Outlier Sensitivity: Linear regression is particularly susceptible to deviations such as anomalies and noise, which can significantly impact its accuracy. Outliers have a detrimental effect on the accuracy and reliability of a model by pulling the line of best fit away from the typical data. Prior to fitting, it is crucial to thoroughly inspect and clean the data (Pant, n.d.).
- Risk of underfitting: Limiting the analysis to a solitary linear line increases the likelihood of inadequately capturing intricate data patterns. When relationships deviate from a linear pattern, a single line is often insufficient to capture complex phenomena, leading to inadequate adaptation and performance. This requires careful consideration prior to implementation (Pant, n.d.).
- Assumes Linearity: Linear regression inherently assumes that there is a linear relationship between the input and output variables. However, the actual dependencies may deviate or exhibit a non-linear pattern in reality. As a result, the rigid linear structure may fail to accurately describe more complex relationships within the data (Pranavnath, 2023) (Rout, n.d.).
- Lack of capacity to manage non-linear situations: If exploration uncovers non-linear relationships, applying linear regression to the data would result in imprecise modelling and projections. More adaptable methodologies such as polynomial regression or non-linear neural networks are required to accurately characterise the relationships (Pranavnath, 2023).

4.2. Random Forest

Random forest is a type of supervised learning algorithm that utilises ensemble modelling. It combines multiple decision trees to carry out predictive tasks. It has become widely adopted for predicting results in areas such as finance and e-commerce, using behavioral data patterns (Mbaabu, 2020) (IBM, n.d.).

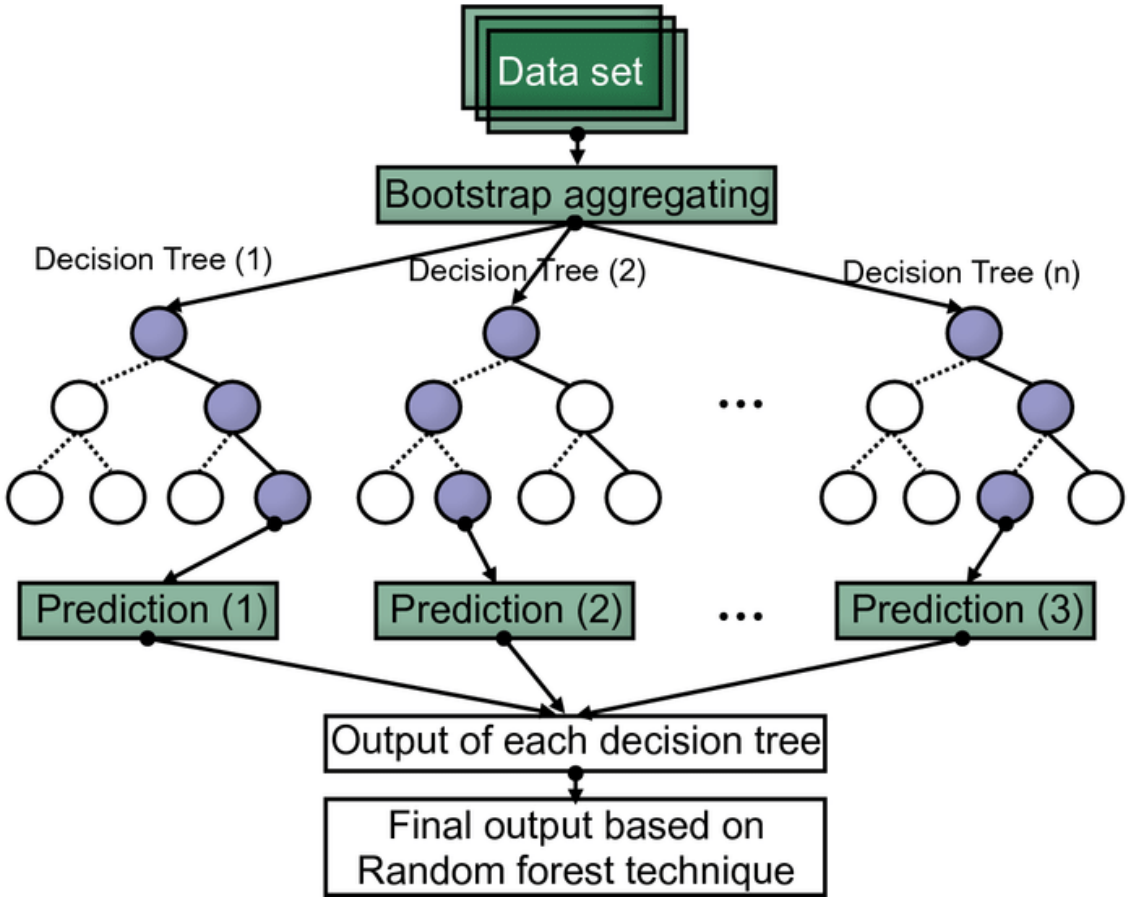


Figure 10 Random Forest Architecture (Hong Min & KooYoon, 2021)

Architecturally, a random forest consists of numerous individual decision trees, each trained on a randomly selected subset of the complete dataset. This ensemble methodology utilises distinct signals derived from various decision tree viewpoints, combining them to form consolidated predictions (IBM, n.d.). The algorithm employs bagging during the process of constructing trees, which involves randomly selecting training data points. Bagging reduces the risk of overfitting compared to individual trees by mitigating irregularities. Consequently, the accuracy of the ensemble is enhanced. Random forest combines the outputs of multiple trees in a classification task by using a majority vote across all the trees. Meanwhile, in regression tasks, the estimates of each individual tree are combined by taking their average to produce overall forecasts. Therefore, the forest achieves harmony among the trees by utilising the collective intelligence of the crowd (Mbaabu, 2020) (IBM, n.d.).

Advantages

- Reduction of overfitting: Using a collection of decision trees helps reduce the problem of overfitting in individual trees, while also enhancing the overall accuracy of predictions and the efficiency of model training (Great Learning Team, 2023).

- Versatility: The random forest algorithm is versatile, as it can be used for both classification and regression tasks. It is particularly effective in handling data that includes both categorical and continuous variables (Great Learning Team, 2023) (Donges, n.d.).
- Robustness: Random forest demonstrates resilience to noise and outliers by aggregating predictions from a variety of decision trees, making it a dependable method across a wide range of real-world datasets (What are the Advantages and Disadvantages of Random Forest, 2023).
- Automated handling of missing data: Another advantage is the automated management of missing data values without requiring imputation. Decision trees inherently enable predictions to be made even when inputs are missing (Great Learning Team, 2023) (Vadapalli, 2021).
- No need for feature normalization: Due to their rule partitioning approach, decision trees do not rely on mathematical optimization. As a result, random forest does not necessitate extensive data normalization (Great Learning Team, 2023).

Disadvantages

- Computational Intensity: The process of creating multiple decision trees significantly increases the computational and memory requirements for training random forests, which hinders the ability to scale up (Great Learning Team, 2023) (Donges, n.d.) (Vadapalli, 2021).
- Slow Prediction: The presence of a large number of trees hinders the ability to make real-time predictions, which limits the practicality of using them in applications that require low latency (Donges, n.d.) (Vadapalli, 2021).
- Lack of Interpretability: The lack of transparency in the internal mechanisms hinders the intuitive understanding of data relationships within the model (Donges, n.d.).
- Increased Accuracy Requires More Trees: Engaging in the pursuit of small improvements in accuracy carries the potential of significantly amplifying the complexity of the model and the associated costs in terms of running time (Donges, n.d.) (Vadapalli, 2021).

4.3. k-Nearest Neighbors(k-NN):

The K-Nearest Neighbors algorithm operates by assessing similarity. It utilises the labels or values of the K most similar points in the training data to make predictions for new data points (Geeks for Geeks Team, n.d.).

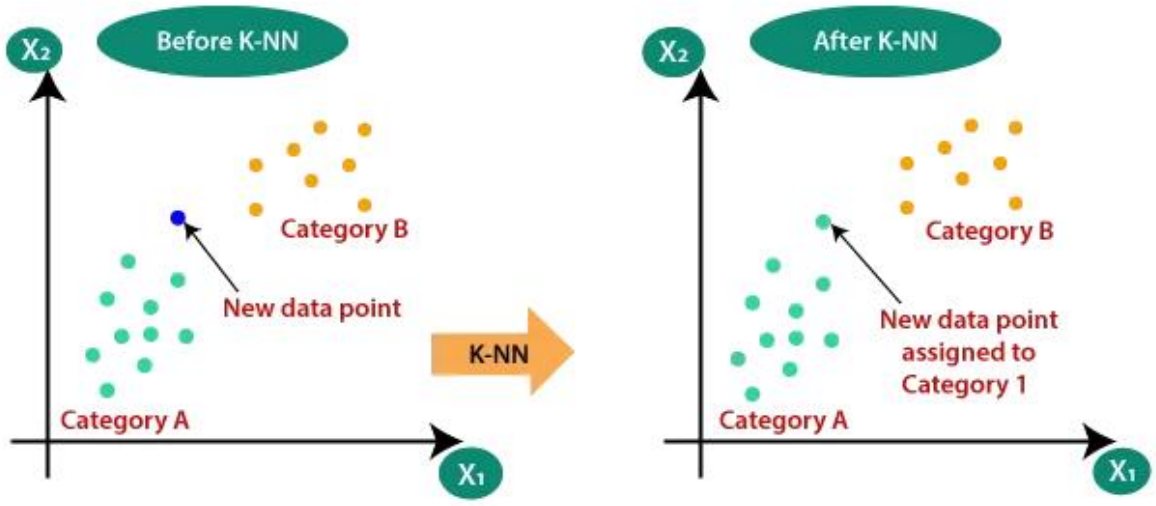


Figure 11 KNN Architecture (Castogna, n.d.)

Upon the arrival of a new data point, the K-Nearest Neighbors (KNN) algorithm examines the K training points that are most proximate to it, utilising a distance metric such as Euclidean distance (Geeks for Geeks Team, n.d.). The new point is assigned the value that is most common among its K nearest Neighbors (Korstanje, n.d.). K-Nearest Neighbors (KNN) algorithm stores all the training data and makes predictions by calculating similarity when a prediction is required. The fundamental concept is that points in close proximity are likely to possess similar labels or values (Korstanje, n.d.).

Advantages

- No Training Period: Being a "lazy learner," KNN does not necessitate any formal training period prior to making predictions. This results in a higher speed compared to algorithms such as SVM that require training (Kumar, 2019).
- Seamless Addition of New Data: Additional data points can be incorporated smoothly without affecting the precision of the model. Retraining is unnecessary as KNN solely relies on similarity during prediction (Kumar, 2019).
- Ease of Implementation: The KNN algorithm is easily configured, necessitating only the selection of the number of Neighbors (K) and the distance function. The significant advantage lies in the simplicity of implementation (Kumar, 2019).
- Capturing Complex Interactions: It is capable of capturing intricate relationships among input variables without the need for explicitly defining a statistical model. KNN is solely dependent on proximity (MyEducator Team, n.d.).
- No Assumptions: It is unnecessary to make conjectures regarding the distribution of the data. KNN exclusively emphasizes the examination of nearby regions (MyEducator Team, n.d.).

Disadvantages

- Performance with Large Datasets: The performance deteriorates when dealing with extremely large datasets due to the significant computational cost of comparing each new data point with every existing data point (Kumar, 2019).
- Challenges with High Dimensions: High-dimensional datasets present challenges as distance calculations become intricate across all dimensions (Kumar, 2019).
- Need for Feature Scalling: Feature scaling is necessary prior to applying the K-nearest Neighbors (KNN) algorithm to ensure that variations in different ranges do not affect the distance measures (Kumar, 2019).
- Sensitivity to Noisy Data and Outliers: The presence of noise, missing values, and outliers can have a substantial negative impact on the accuracy of the model, as KNN places great emphasis on the nearest data points. Data cleaning is essential (Kumar, 2019).
- Computational Intensity: The computational intensity of KNN increases at prediction time due to the distance calculations, resulting in slower speed and higher memory requirements as the dataset size increases (MyEducator Team, n.d.).

5. Training and Evaluation

5.1. Feature and Target Variable

To train a machine learning model, it is necessary to provide a substantial number of examples that illustrate the connections between input features and the target variables that are to be predicted. Initially, the data at hand is partitioned into distinct features, which will serve as inputs for the model, and targets, which represent the anticipated predictions for specific inputs. The model is subsequently trained by processing this dataset, establishing associations between input features and their corresponding targets. By being exposed to consistent inputs and outputs multiple times, the algorithm discovers patterns that enable it to make accurate predictions for unfamiliar features. Efficiently training the model to precisely capture the relationships between inputs and outputs is crucial for guaranteeing high-quality performance. Appropriate dataset partitioning, feature curation, and dimension determination aid in the model's ability to generalise the input-target associations. The ultimate objective is to create an algorithm that can effectively anticipate the corresponding target variables by utilising learned mappings from the training data when provided with new input features.

5.2. Train-Test Split

The train-test split is a fundamental machine learning technique used to assess the performance of a model. The process entails partitioning your initial dataset into two distinct sets - one designated for training purposes and the other for testing purposes.

The objective is to train your model using a specific dataset and subsequently evaluate its performance on a separate "novel" dataset to replicate its ability to handle unfamiliar real-world data. This provides an indication of the extent to which your model is likely to apply to new, unseen data.

Usually, the division of data into training and testing sets is carried out in Python using the ‘sklearn’ library. This function enables the random division of your dataset, giving you the ability to specify the sizes of both the training and testing sets.

Implementing a train-test split is crucial for obtaining a precise assessment of model efficacy and circumventing issues such as overfitting. It facilitates the prediction of future performance on new data by leveraging the current proficiency in modelling the training data.

The train-test split procedure is frequently utilised in machine learning guides and tutorials. Visual diagrams illustrate the concept by demonstrating how the initial dataset is divided into two separate subsets that are utilised at different stages of the modelling process.

Essentially, the train-test split involves dividing a single dataset into separate training and test datasets. This method is used to evaluate models and ensure their ability to generalise to new data.

```
433  # Trim Level Prediction
434  features_trim_level = merged_data_left.drop(['Trim'], axis=1)
435  target_trim_level = merged_data_left['Trim']
436  X_train_trim_level, X_test_trim_level, y_train_trim_level, y_test_trim_level =
437  train_test_split(features_trim_level, target_trim_level, test_size=0.2, random_state=42)
438
```

Figure 12 Train-Test Split

Figure no is related to the task of predicting the trim level of automobiles. The dataset is preprocessed for this prediction task by partitioning it into features and the target variable. The features, referred to as features_trim_level, encompass all columns in the DataFrame 'merged_data_left' with the exception of the 'Trim' column, which serves as the target variable. The target variable, referred to as target_trim_level, encompasses the 'Trim' column.

Afterwards, the dataset is divided into training and testing sets using the train_test_split function from the scikit-learn library. The feature and target variables are divided into training and testing sets. X_train_trim_level and X_test_trim_level store the feature values for the training and testing sets, respectively. Similarly, y_train_trim_level and y_test_trim_level store the corresponding target variable values for the training and testing sets. The dataset is divided into training and testing sets using a test size of 20% (test_size=0.2). To ensure reproducibility of results, a random state of 42 is set. Prior to training a machine learning model for trim level prediction, it is customary to perform this

preparation, wherein the training set is utilised to train the model and the testing set is utilised to assess its performance.

5.3. Evaluation

```
# Linear Regression for Trim Level Prediction
lr_trim_level_model = LinearRegression()
lr_trim_level_model.fit(X_train_trim_level, y_train_trim_level)
lr_trim_level_predictions = lr_trim_level_model.predict(X_test_trim_level)

# Random Forest for Trim Level Prediction
rf_trim_level_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_trim_level_model.fit(X_train_trim_level, y_train_trim_level)
rf_trim_level_predictions = rf_trim_level_model.predict(X_test_trim_level)

# KNN for Trim Level Prediction
knn_trim_level_model = KNeighborsRegressor(n_neighbors=5)
knn_trim_level_model.fit(X_train_trim_level, y_train_trim_level)
knn_trim_level_predictions = knn_trim_level_model.predict(X_test_trim_level)
```

Figure 13 Model Building

The given code snippet utilises three separate machine learning models to forecast automobile trim levels. The initial step involves initialising the first model, Linear Regression. It is then trained using the given training data, specifically `X_train_trim_level` and `y_train_trim_level`. Finally, the model is used to predict trim levels on the test data, `X_test_trim_level`. The final forecasts are saved in the variable `lr_trim_level_predictions`.

Transitioning to the second model, we utilise a Random Forest Regressor to predict the trim level. The model is trained using the training data, employing 100 decision trees and a fixed random state for reproducibility. Subsequently, predictions are made on the test data and stored in the variable `rf_trim_level_predictions`.

Finally, the third model utilises a K-Nearest Neighbours (KNN) Regressor. The KNN model is trained on the given data and then used to predict trim levels on the test set, with an initial set of 5 neighbours. The predictions are stored in the variable `knn_trim_level_predictions`. These three models collectively provide various methods for predicting trim levels, each with its distinct features and potential benefits.

```
lr_trim_level_predictions_class = np.round(lr_trim_level_predictions).astype(int)

# Compute metrics for Trim Level Prediction using Linear Regression
accuracy_trim_level = accuracy_score(y_test_trim_level, lr_trim_level_predictions_class)
precision_trim_level = precision_score(y_test_trim_level, lr_trim_level_predictions_class, average='weighted')
recall_trim_level = recall_score(y_test_trim_level, lr_trim_level_predictions_class, average='weighted')
f1_trim_level = f1_score(y_test_trim_level, lr_trim_level_predictions_class, average='weighted')

print("\nTrim Level Prediction Metrics:")
print(f"Accuracy: {accuracy_trim_level}")
print(f"Precision: {precision_trim_level}")
print(f"Recall: {recall_trim_level}")
print(f"F1 Score: {f1_trim_level}")

rf_trim_level_predictions_class = np.round(rf_trim_level_predictions).astype(int)

# Compute metrics for Trim Level Prediction using Random Forest
accuracy_trim_level = accuracy_score(y_test_trim_level, rf_trim_level_predictions_class)
precision_trim_level = precision_score(y_test_trim_level, rf_trim_level_predictions_class, average='weighted')
recall_trim_level = recall_score(y_test_trim_level, rf_trim_level_predictions_class, average='weighted')
f1_trim_level = f1_score(y_test_trim_level, rf_trim_level_predictions_class, average='weighted')

print("\nTrim Level Prediction Metrics:")
print(f"Accuracy: {accuracy_trim_level}")
print(f"Precision: {precision_trim_level}")
print(f"Recall: {recall_trim_level}")
print(f"F1 Score: {f1_trim_level}")

knn_trim_level_predictions_class = np.round(knn_trim_level_predictions).astype(int)

# Compute metrics for Trim Level Prediction using KNN
accuracy_trim_level = accuracy_score(y_test_trim_level, knn_trim_level_predictions_class)
precision_trim_level = precision_score(y_test_trim_level, knn_trim_level_predictions_class, average='weighted')
recall_trim_level = recall_score(y_test_trim_level, knn_trim_level_predictions_class, average='weighted')
f1_trim_level = f1_score(y_test_trim_level, knn_trim_level_predictions_class, average='weighted')

print("\nTrim Level Prediction Metrics:")
print(f"Accuracy: {accuracy_trim_level}")
print(f"Precision: {precision_trim_level}")
print(f"Recall: {recall_trim_level}")
print(f"F1 Score: {f1_trim_level}")
```

Figure 14 Computing Metrics

The figure assesses the performance metrics for trim level prediction by employing three distinct machine learning models: Linear Regression, Random Forest, and K-Nearest Neighbours (KNN).

The predicted trim levels for each model are rounded to the nearest integer using the np.round function and then converted to integers. Following that, essential classification metrics are calculated, such as accuracy, precision, recall, and F1 score. These metrics offer a thorough evaluation of the performance of each model in predicting trim levels using the test data.

The initial set of metrics relates to the Linear Regression model. The accuracy, precision, recall, and F1 score are computed and displayed, providing valuable information about the overall performance of the model. The identical procedure is iterated for the Random Forest and KNN models, with each model offering a distinct viewpoint on the efficacy of the corresponding algorithms in predicting trim levels.

```
# Evaluate models
def evaluate_model(predictions, true_values):
    mse = mean_squared_error(true_values, predictions)
    r2 = r2_score(true_values, predictions)
    return mse, r2

```

Figure 15 Model Evaluate Function

The given code presents a function called evaluate_model, which is intended to evaluate the effectiveness of machine learning models. The function accepts two parameters: predictions and true_values. The predictions parameter represents the predicted values, while the true_values parameter represents the actual (true) values. The function computes two evaluation metrics, namely Mean Squared Error (MSE) and R-squared (R2). These metrics provide insights into the accuracy and adequacy of the models by quantifying the discrepancy between predicted and actual values. The function subsequently returns the calculated Mean Squared Error (MSE) and R-squared (R2) as a tuple, furnishing a succinct yet informative assessment summary for the performance of the machine learning model.

```
# Evaluate Trim Level Prediction models
lr_trim_level_mse, lr_trim_level_r2 = evaluate_model(lr_trim_level_predictions, y_test_trim_level)
rf_trim_level_mse, rf_trim_level_r2 = evaluate_model(rf_trim_level_predictions, y_test_trim_level)
knn_trim_level_mse, knn_trim_level_r2 = evaluate_model(knn_trim_level_predictions, y_test_trim_level)

print(f"Trim Level Prediction - Linear Regression MSE: {lr_trim_level_mse}, R2 Score: {lr_trim_level_r2}")
print(f"Trim Level Prediction - Random Forest MSE: {rf_trim_level_mse}, R2 Score: {rf_trim_level_r2}")
print(f"Trim Level Prediction - KNN MSE: {knn_trim_level_mse}, R2 Score: {knn_trim_level_r2}")
```

Figure 16 Preditction Evaluation

The given code assesses the efficacy of machine learning models specifically for the purpose of predicting Trim Level. The evaluate_model function is used to compute two important metrics, namely Mean Squared Error (MSE) and R-squared (R2), for three distinct models: Linear Regression, Random Forest, and K-Nearest Neighbours (KNN). Subsequently, the outcomes are displayed, presenting a succinct overview of the assessment metrics for every model. The Mean Squared Error (MSE) values represent the average of the squared differences between the predicted and actual Trim Level values. On the other hand, the R2 scores quantify the proportion of variability in the target variable that can be accounted for by the model. The evaluation phase provides valuable insights into the precision and forecasting capabilities of the Trim Level Prediction models.

A consistent and thorough methodology is employed for other forecasts, including Price, Engine Size, Gas Emission, and Fuel Type. This indicates a methodical and all-encompassing approach to assessing the effectiveness of these machine learning models in different prediction tasks.

6. Results and Analysis

Random forest Results

	Sales	Trim	Fuel Type
Accuracy	**0.8445384093223215**	**0.17356318588075575**	**0.9980201380246634**
Precision	**0.8778885713099969**	**0.17466115948138872**	**0.9987584681617188**
Recall	**0.8848125316092347**	**0.17356318588075575**	**0.9980201380246634**
F1 Score	**0.8714743041503068**	**0.1627496613036963**	**0.9983742320830735**
R2 Score	**0.9359011234582523**	**0.9979975056547717**	**0.9985552889005931**

k-Nearest Neighbor Results

	Sales	Trim	Fuel Type
Accuracy	**0.7940348743476969**	**0.0384658898065392**	**0.7138816608213598**
Precision	**0.8753232885400432**	**0.03597739626505151**	**0.949829303595221**
Recall	**0.8439220424506715**	**0.0384658898065392**	**0.7138816608213598**
F1 Score	**0.8495153101138857**	**0.033367592567638**	**0.8050522553954704**
R2 Score	**0.9507895036950341**	**0.9921008031413852**	**0.6947091033275099**

7. Project Management

Effective project management involves initiating, planning, and controlling a range of tasks to deliver a new or altered product or service over a finite timeline. Formally managing projects that produce tangible or intangible deliverables, are likely to be complex, require managing change and risk, and have a defined start and end date leads to substantial benefits. Careful project management provides a greater probability of achieving the desired outcome, ensures efficient utilisation of resources, and satisfies the needs of different stakeholders. By investing in robust initiation, planning, and control, organisations can deliver successful projects that bring value to sponsors while making the best use of team members' time and talents. Students examining project management case studies can compare notes on which techniques work well versus areas needing improvement to become thoughtful managers capable of guiding projects to maximise results and stakeholder satisfaction (Association for Project Management Team, 2023).

7.1. Meeting with Supervisor

During my research project, I scheduled bi-monthly supervisory sessions with Professor Vasile Palade using Microsoft Teams. Each session lasted for 30 minutes. These meetings were crucial for reviewing progress and obtaining essential guidance. They provided a priceless opportunity to synchronize project objectives and tackle any arising obstacles. Prior to each meeting, thorough preparations were conducted, which included gathering relevant information and creating a list of discussion topics encompassing completed activities since the last session, upcoming priorities, and specific areas requiring direction or feedback. Professor Palade diligently scrutinized the preliminary studies and findings during our discussions, furnishing valuable critiques and practical suggestions for enhancement. His ethical mentorship and provision of resources played a pivotal role in both my personal growth and the successful attainment of project objectives. I extend my sincere appreciation for the extensive mentorship provided throughout this journey. Highlights the importance of this crucial stage. It is crucial to validate these models to ensure that they are effective and reliable in the analytical field.

7.2. Legal Social Ethical and Professional Considerations

As the author, I consider data to be the crucial element in machine learning research. Therefore, it is crucial to thoroughly consider the origins and characteristics of the data utilized. Upon careful examination from an ethical perspective, the data set in this specific study demonstrates strong integrity. Primarily, the data was acquired without any financial expenditure from openly accessible web sources, so circumventing concerns related to restrictive licensing or costly paywalls. This granted the author unrestricted access to crucial data necessary for the goals of investigation. Furthermore, the author made a deliberate effort to ensure that the content does not contain any sensitive personal information or identifiers that could have a negative impact on persons if used in this context. The author shown conscientiousness by appropriately acknowledging the legal ownership and permissions granted for this dataset. In summary, measures were

implemented to obtain non-controversial, morally sound data in a cost-effective manner through trustworthy channels. The complete documentation demonstrating compliance with ethical data usage can be found in the Appendix appended to this report. The author has given priority to ethical considerations and set a high standard for responsible conduct when doing machine learning research in order to access and utilize these data resources.

8. Critical Appraisal

During our research on predictive modelling in the automobile sector, I acquired significant hands-on expertise in utilising machine learning techniques such as Linear Regression, KNN, and Random Forests for the purpose of forecasting. The primary objective was to utilise data-driven insights to enhance technical knowledge and outcomes for vehicle manufacturers and partners.

One notable accomplishment was the effective application of widely-used machine learning techniques to the field of car forecasting. The shown accuracy in other domains, such as sales predictions, confirms the adaptability of regularly used methodologies in fields like finance and healthcare. This highlights a broader range of opportunities to improve data-driven decision making in automotive technology using predictive analytics.

Personally, the practical application resulted in significant knowledge acquisition. Despite having prior knowledge in machine learning foundations, this project enhanced my expertise in the specialised field of time-series data analysis for predicting applications. Consistent use of python libraries such as pandas, matplotlib, and scikit-learn enhanced coding aptitude as well. Utilising Tableau enhanced my proficiency in data reporting through the creation of visual representations.

Nevertheless, certain constraints emerged that indicate promising avenues for further expansion. Specifically, conducting more thorough statistical benchmarking could enhance the evaluations of the Linear Regression, KNN, and Random Forest methods that were utilised. Assessing their ability to forecast outcomes in relation to established benchmarks would provide a more comprehensive understanding of the advantages and disadvantages. Further investigation into established methodologies for predicting car trends would facilitate more insightful evaluations.

Furthermore, increased collaboration with automotive industry specialists at every stage of the project may have boosted the contextual significance. Specialised industry perspectives are essential for effectively turning data insights into tangible technological or manufacturing advancements. Collaboration across different disciplines is still necessary in order to achieve the most impact on company.

9. Conclusion

- When assessing machine learning models for prediction tasks, we evaluate their performance using various metrics to determine the most suitable strategy

depending on planned use cases and priorities. Both the random forest and K-nearest neighbours (KNN) algorithms yield similar R-squared values when used for sales volume forecasting. Nevertheless, KNN has a reduced mean squared error (MSE), indicating a more effective decrease of forecast deviations from the actual results.

- If the primary goal is to minimise forecast error, KNN would be the recommended choice. However, if our objective is to optimise forecast accuracy or accurately identify unusual events, random forest may be more effective in achieving these goals, even if it results in some errors. The assessment of acceptable trade-offs among precision, interpretability, uncertainty levels, and calculation durations relies on the business objectives and the degrees of risk that may be tolerated.

- Comparable trade-offs arise when estimating both automotive trim levels and fuel kinds. Random forest frequently surpasses KNN in terms of accuracy, precision, recall, and F1 score, which are model evaluation measures. Additionally, it reduces the mean squared error (MSE) for both prediction tasks. Random forests are effective in minimising errors when categorising vehicles into luxury, mid-range, or basic trim bands, as well as accurately discriminating between petrol, diesel, hybrid, or electric cars.

- The consistent improvement in performance across assessment metrics for both classification objectives indicates that optimising random forest may have broader advantages. Nevertheless, it is crucial to thoroughly investigate predictive parity in all vehicle sectors, as biases may arise unexpectedly. However, according to the latest empirical findings, random forest models have shown distinct benefits compared to KNN alternatives when it comes to sales regression, as well as predictions for trim level and powertrain using this particular car dataset.

10. Future Works

- Building upon this first investigation, I am of the opinion that there is potential in enhancing prediction abilities through the integration of deep learning and neural network approaches. Neural network topologies, despite their complexity, have the potential to describe the numerous linkages and uncertainties in automobile data more effectively than my current approach. The increased flexibility has the potential to boost the precision of our forecasts.

- In addition, my analyses have only depended on organised data tables. A compelling next endeavour is to investigate the large amounts of unorganised visual data in the automotive industry, encompassing images of vehicle parts, assembly processes, and so on.

- Advanced image recognition algorithms provide systematic techniques for extracting visual insights from such photographs. By incorporating these computer vision-derived patterns into our current textual record databases, we can enhance our understanding and gain a more comprehensive viewpoint. There is a potential to greatly enhance the accuracy of predictions beyond what present methodologies can achieve.

- Specifically, I believe it is worth exploring the use of convolutional neural networks to classify car images based on the forms of their components, the production process, defect identification, and other relevant aspects of manufacturing. Long-short term memory networks can be utilised to integrate computer vision outputs with streams of timeseries data from cars and suppliers. This has the potential to stimulate more efficient monitoring of problems, flexible improvement of quality control, and potentially more adaptable, robust coordination of the supply chain.

- The integration of deep learning, imaging, and timeseries data analysis presents engineering obstacles but holds the potential to enhance automotive design and production to unparalleled levels. The raw data is currently available, awaiting future advancements to fully use its collective potential.

11. Self-Reflection

- When I started working on this vehicle data modelling project, I didn't fully realise the important influence that the origin and history of the dataset have on the results of our analysis. In my previous machine learning efforts, I used carefully selected and customised data frameworks created by Python pipelines, specifically designed for training algorithms. I erroneously presumed that all datasets readily adhere to machine learning standards.

- Nevertheless, the endeavour to manage this particular car collection designed only for computer vision assignments proved very burdensome. The configuration of the records was specifically optimised for the purpose of detecting imagery in Javascript, rather than for conducting exploratory, descriptive, or predictive modelling in Python. Statistical rules regarding missingness encoding, categorical variables, feature engineering, and target leakage were not taken into account, and no concessions were made for them.

- The process of converting the data into a suitable format for effective training and assessment required extensive searching and cleaning. Despite being cleaned, the persistent remnants of its picture categorization origins continued to hinder the performance of the modelling. Each step obstinately resisted, as if offended by my

efforts to reduce it to a single variable. My well-organized models struggled to perform smoothly instead of executing gracefully as they usually do.

- This event made me realise that real-world data is often difficult to work with and sometimes even directly opposes my modelling goals. I cannot assume the presence of clean and compliant data, nor can I immediately enforce any framework only based on the convenience provided by Python tooling. To generate reliable insights, one must possess a profound comprehension of the origins of the data, including the intentions, decisions, and processes that have influenced its current state.

- This understanding is crucial in order to effectively convert the data into analytical tools. Data wrangling is not merely a superficial need, but rather an artistic process that demands the same level of creative skill as subsequent modelling stages. Acknowledging my lack of knowledge in this area will surely enhance my skills as a flexible, perceptive, and skilled data scientist.

References

Association for Project Management Team. (2023). *What is project management?* Retrieved from Association for Project Management: https://www.apm.org.uk/resources/what-is-project-management/

Bao, Y. (2023, June 09). *EWA Publishing - Applied and Computational Engineering.* Retrieved from Applied and Computational Engineering: https://ace.ewapublishing.org/article.html?pk=e3762c9134974aa79b992dd74b5cb29b

Bukvi´c , L., Škrinjar, J. P., Fratrovi´c, T., & Abramovi´c, B. (2022). Price Prediction and Classification of Used-Vehicles UsingSupervised Machine Learning. *Research Gate*.

Coventry University. (n.d.). *Data cleaning and pre-processing. What do we mean by 'data cleaning' and why is it necessary?* Retrieved from Future Learn : https://www.futurelearn.com/info/courses/applied-data-science/0/steps/169276#:~:text=Data%20cleaning%20(also%20known%20as,with%20incomplete%20or%20missing%20data.

Donges, N. (n.d.). *Random Forest: A Complete Guide for Machine Learning*. Retrieved from Built In: https://builtin.com/data-science/random-forest-algorithm

Geeks for Geeks Team. (n.d.). *K-Nearest Neighbor(KNN) Algorithm*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/k-nearest-neighbours/

Great Learning Team. (2023, November 8). *Lable Encode in Python*. Retrieved from Great Learning: https://www.mygreatlearning.com/blog/label-encoding-in-

python/#:~:text=Label%20encoding%20is%20a%20simple,input%20into%20machine%20learning%20algorithms.

Great Learning Team. (2023, June 3). *Random Forest Algorithm in Machine Learning*. Retrieved from Great Learning: https://www.mygreatlearning.com/blog/random-forest-algorithm/

Grelle, C., Franke, V., Verma, A., Miyanyedi, O., Cere', V., & Devadiga, A. (2021, April 04). *Used car price prediction in the UK market — Determining the price of used cars based on their features.* Retrieved from Medium: https://techlabsdus.medium.com/used-car-price-prediction-in-the-uk-market-determining-the-price-of-used-cars-based-on-their-dbb2185ccafd

Hong Min, D., & KooYoon, H. (2021, March). *Suggestion for a new deterministic model coupled with machine learning techniques for landslide susceptibility mapping.* Retrieved from Research Gate: https://www.researchgate.net/publication/350319555_Suggestion_for_a_new_deterministic_model_coupled_with_machine_learning_techniques_for_landslide_susceptibility_mapping

IBM. (n.d.). *What is Random Forest?* Retrieved from IBM: https://www.ibm.com/topics/random-forest

Jin, C. (2021, June). *IEEE International Conference on Emergency Science and Information Technology (ICESIT).* Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/9696839

Korstanje, J. (n.d.). *The k-Nearest Neighbors (kNN) Algorithm in Python*. Retrieved from Real Python: https://realpython.com/knn-python/

Kumar, N. (2019, February 23). *Advantages and Disadvantages of KNN Algorithm in Machine Learning*. Retrieved from The Professionals Point: https://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html?m=1

Kumatani, S., Itoh, T., Motohashe, Y., Umezu, K., & Takatsuka, M. (2016, September 1). *Time-Varying Data Visualization Using Clustered Heatmap and Dual Scatterplots*. Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/document/7557905

Mahfouz, M. M., Mosaad, S. M., & Belal, M. A. (2023). Forecasting Vehicle Prices using Machine Learning Techniques based on Federated Learning Strategy. *International Journal of Computer Applications (0975 – 8887)*.

Mbaabu, O. (2020, December 11). *Introduction to Random Forest in Machine Learning*. Retrieved from Section.io: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

MyEducator Team. (n.d.). *Advantages and disadvantages of KNN*. Retrieved from MyEducator: https://app.myeducator.com/reader/web/1421a/11/q07a0/

Narayana, C. V., Likhitha, C. L., Bademiya, S., & Kusumanjali, K. (2021). Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business. *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC).* IEEE Xplore.

Pant, M. (n.d.). *Advantage and Disadvantage of Linear Regression*. Retrieved from Kaggle: https://www.kaggle.com/discussions/general/352609

Patil, P. (2018, March 23). *What is Exploratory Data Analysis?* Retrieved from Towards Data Science: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Pranavnath. (2023, July 26). *Advantages and Disadvantages of Linear Regression*. Retrieved from Tutorials Point: https://www.tutorialspoint.com/advantages-and-disadvantages-of-linear-regression

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning.

Rout, A. R. (n.d.). *ML - Advantages and Disadvantages of Linear Regression*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/

Statistics How To. (n.d.). *Linear Regression: Simple Steps, Video. Find Equation, Coefficient, Slope*. Retrieved from Statistics How To: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/

Vadapalli, P. (2021, June 18). *Random Forest Classifier: Overview, How Does it Work, Pros & Cons*. Retrieved from upGrad: https://www.upgrad.com/blog/random-forest-classifier/

Waseem, M. (2023, August 2). *Linear Regression for Machine Learning*. Retrieved from Edureka: https://www.edureka.co/blog/linear-regression-for-machine-learning/

What are the Advantages and Disadvantages of Random Forest. (2023, February 19). Retrieved from Rebellion Research: https://www.rebellionresearch.com/what-are-the-advantages-and-disadvantages-of-random-forest

What is Linear Regression? (n.d.). Retrieved from Master's in Data Science with edx: https://www.mastersindatascience.org/learning/machine-learning-algorithms/linear-regression/

Appendix – A – GitHub Link

https://github.com/adiguj/7151CEM.git

Appendix - B - Presentation

Predictive Modelling and Analysis: Machine Learning in the UK Automotive Sector.

Supervisor: Prof. Vasile Palade
Name: Aditya Ashok Gujar
SID: 13366670

Introduction

The UK automobile sector struggles to understand and optimize sales trends, pricing strategies, trim level preferences, and used car pricing. Goal is to create machine learning prediction and analysis that gives automakers, customers, and policymakers practical tools to solve their problems.

Motivation

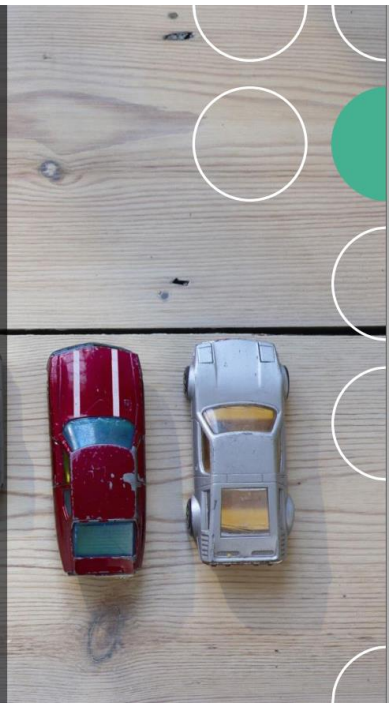
- Predictive Analysis
- Competitive Analysis
- Customer Segmentation and Targeting
- Innovation and Research

Literature Review

- Yan Bao (2023, June 16) examines the subjectivity of appraiser-based used car pricing. With 22.5% training accuracy and 14.3% test accuracy, the deep learning model trained with a classical SVM performed poorly. The study recommends advanced neural computing approaches to estimate used automobile prices more accurately.

- In the 2020/21 TechLabs "Digital Shaper Programme," the Düsseldorf team produced a machine learning model that reliably predicts used vehicle market value, helping customers and sellers (Grelle, et al., 2021). The study examines 100,000 UK pre-owned automobile selling prices (£) to determine cost-affecting auto components.

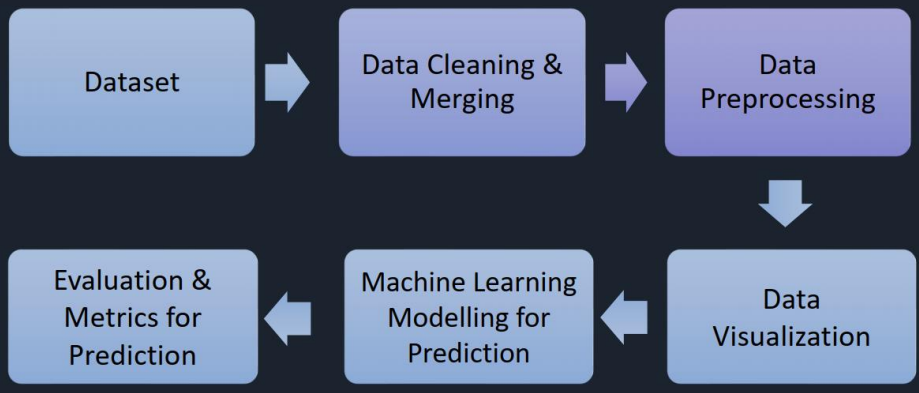
- Jin, C.(2021, June) predicted used car values using machine learning and neural networks. The study recommends using Naive Bayes, LSTM, and Gradient Boosting to improve accuracy and collect more data due to the dataset's 12,988 samples.

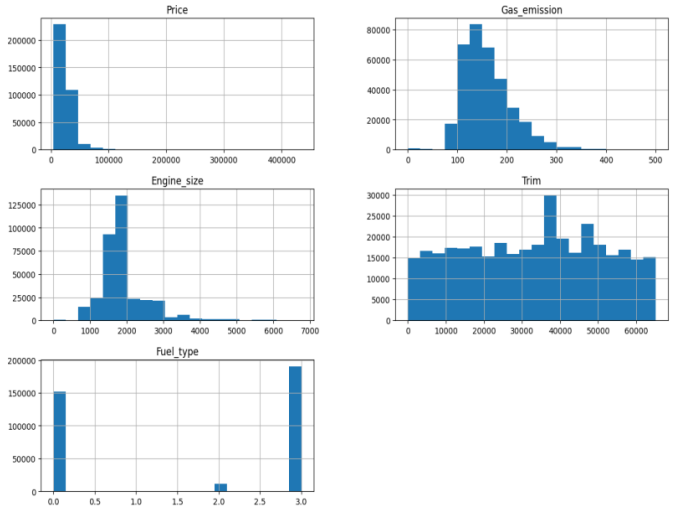


Dataset

File	Description
Basic	Its primary purpose is to index the data in other tables.
Trim	It contains 0.33 million trim-level details, including price, fuel type, and engine size. It is intended for those seeking automobile model prices at specific trim levels.
Sales	It includes information about vehicle sales in the UK market from the year 2001 to 2020.

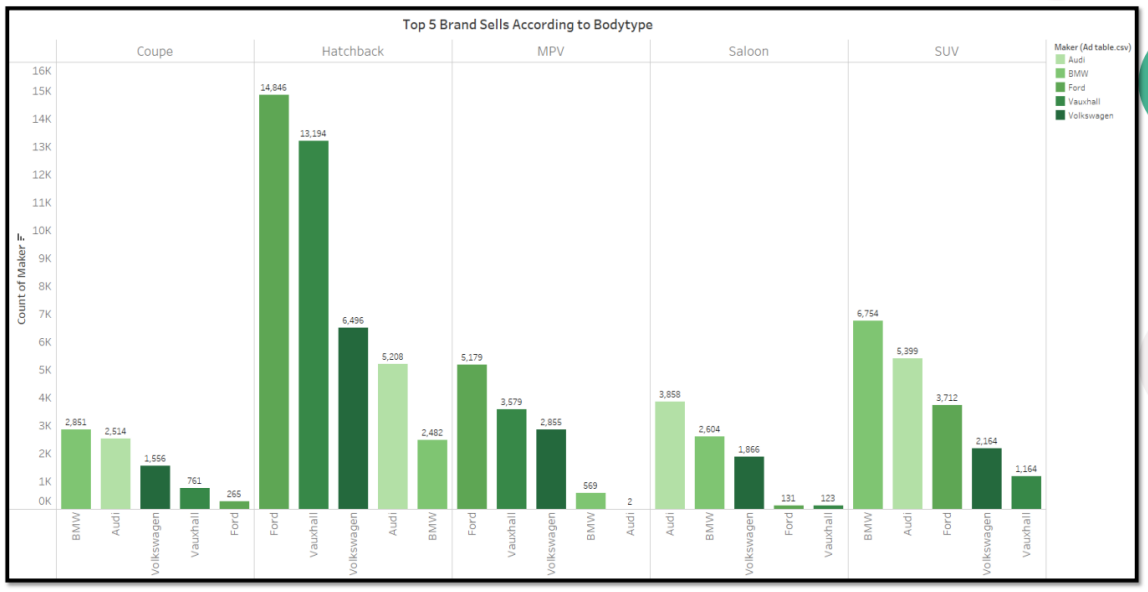
Process Flow

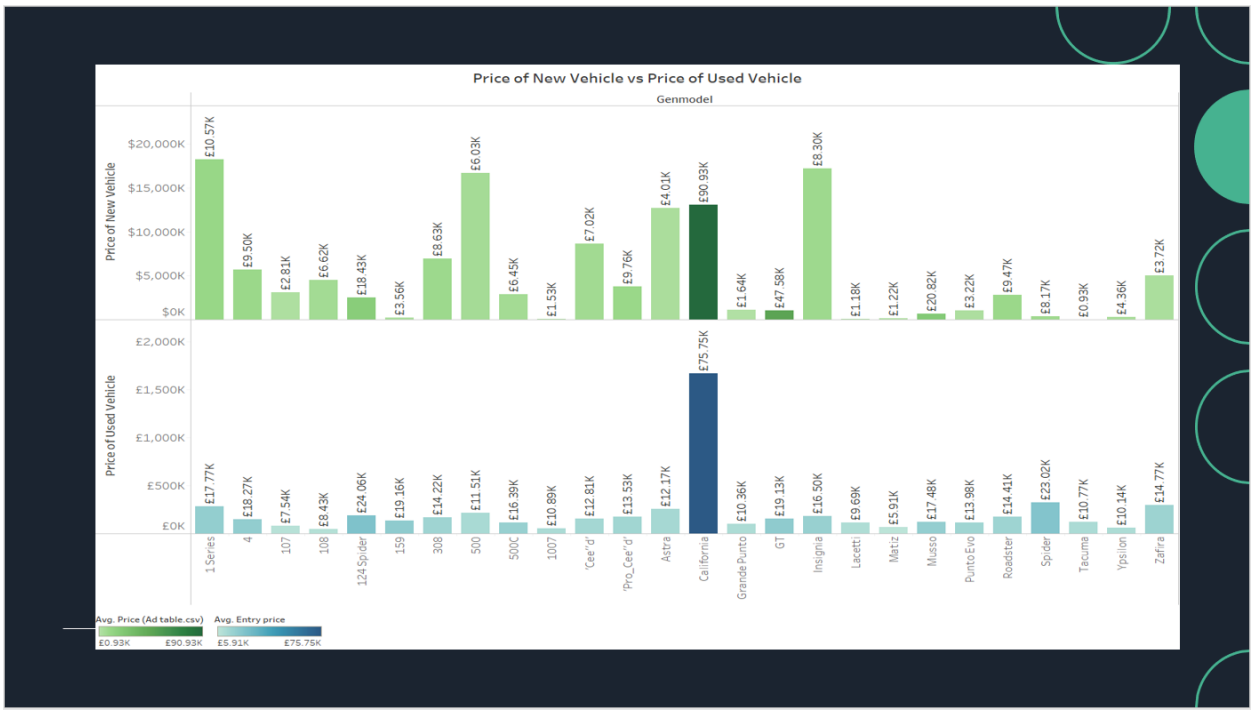


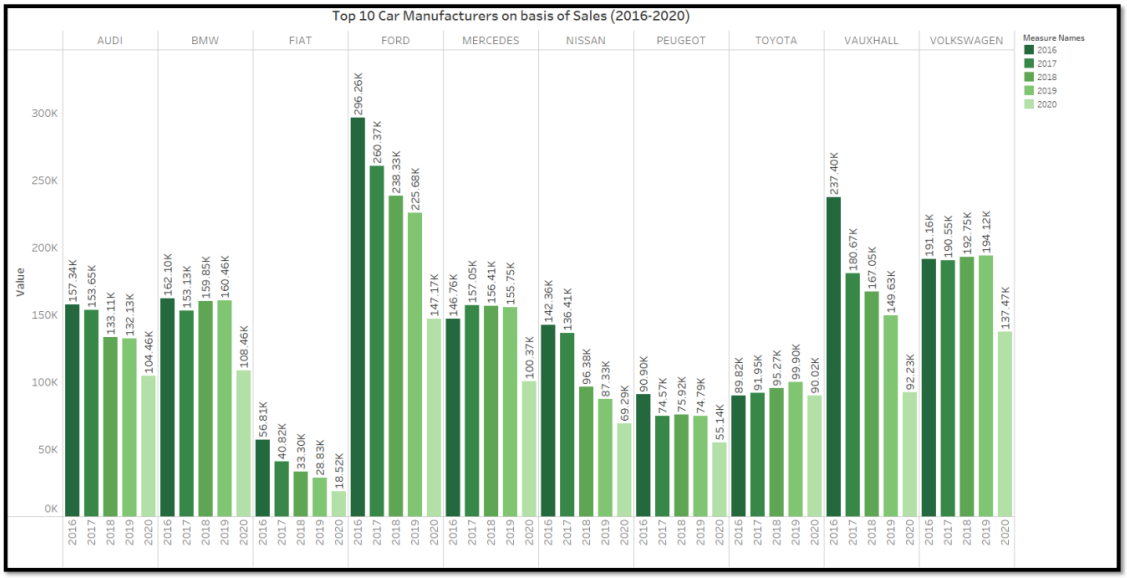


Exploratory Data Analysis

- The distribution of values for each selected feature, such as Price, Sales, Trim, and so on, can be better comprehended with the assistance of this visualization.
- A granular picture of the distribution can be obtained by setting the bins=20 parameter, which divides the range of possible values into 20 intervals.
- The figsize=(15, 10) argument guarantees that the figure that displays all of the histograms has the size that has been chosen.




Price of New Vehicle vs Price of Used Vehicle
Genmodel
Price of New Vehicle
$20,000K
$15,000K
$10,000K
$5,000K
$0K
£10.57K
£9.50K
£2.81K
£6.62K
£18.43K
£3.56K
£8.63K
£16.03K
£6.45K
£1.53K
£7.02K
£9.76K
£4.01K
£90.93K
£1.64K
£47.58K
£8.30K
£1.18K
£1.22K
£20.82K
£3.22K
£9.47K
£8.17K
£0.93K
£4.36K
£3.72K
Price of Used Vehicle
£2,000K
£1,500K
£1,000K
£500K
£0K
£17.77K
£18.27K
£7.54K
£8.43K
£24.06K
£19.16K
£14.22K
£11.51K
£16.39K
£10.89K
£12.81K
£13.53K
£12.17K
£75.75K
£10.36K
£19.13K
£16.50K
£9.69K
£5.91K
£17.48K
£13.98K
£14.41K
£23.02K
£10.77K
£10.14K
£14.77K
1 Series
4
107
108
124 Spider
159
308
500
500C
1007
'Cee'd'
'Pro_Cee'd'
Astra
California
Grande Punto
GT
Insignia
Lacetti
Matiz
Musso
Punto Evo
Roadster
Spider
Tacuma
Ypsilon
Zafira
Avg. Price (Ad table.csv)
£0.93K
£90.93K
Avg. Entry price
£5.91K
£75.75K

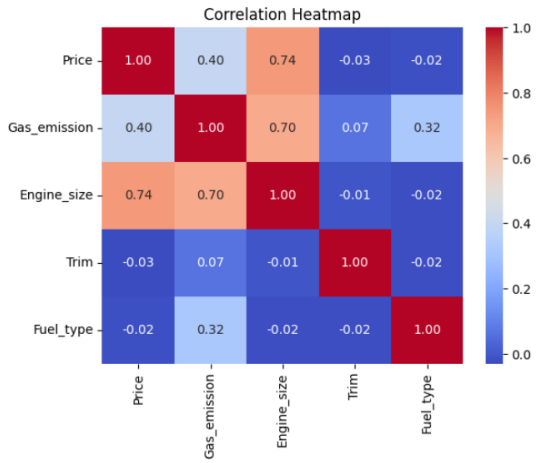

Top 10 Car Manufacturers on basis of Sales (2016-2020)
AUDI
BMW
FIAT
FORD
MERCEDES
NISSAN
PEUGEOT
TOYOTA
VAUXHALL
VOLKSWAGEN
Measure Names
2016
2017
2018
2019
2020
Value
300K
250K
200K
150K
100K
50K
0K
157.34K
153.65K
133.11K
132.13K
104.46K
162.10K
153.13K
159.85K
160.46K
108.46K
56.81K
40.82K
33.30K
28.83K
18.52K
296.26K
260.37K
238.33K
225.68K
147.17K
146.76K
157.05K
156.41K
155.75K
100.37K
142.36K
136.41K
96.38K
87.33K
69.29K
90.90K
74.57K
75.92K
74.79K
55.14K
89.82K
91.95K
95.27K
99.90K
90.02K
237.40K
180.67K
167.05K
149.63K
92.23K
191.16K
190.55K
192.75K
194.12K
137.47K

Correlation Heatmap

The heatmap provides a graphical representation of the association between the many pairs of data in the dataset, including Sales, Engine Size, Trim, and so on.

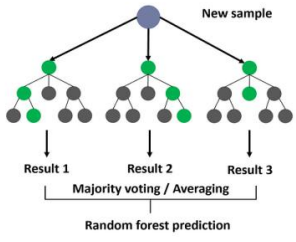
When a correlation is positive, the colour will be warmer (getting closer to 1.0), and when a correlation is negative, the colour will be cooler (getting closer to -1.0).

The intensity and direction of the correlation between each relevant pair of features is indicated by the numeric values that are contained within each cell of the heatmap.



Random Forest

- Random Forest is a flexible ensemble learning method that solves regression issues by creating several decision trees during training and then producing an average forecast based on the trees' data.
- Each variable (Sales, Trim, Fuel Type) has training and testing sets. Each target variable has its own Random Forest model. To predict sales, rf_sales_model is trained. and similarly for Trim, and Fuel Type.
- The n_estimators hyperparameter (100 in the code) controls the forest's decision trees. For reproducibility, more hyperparameters like random_state are set.
- Random Forest models make test set predictions (rf_sales_predictions, etc.) after training.
- MSE and R2 Score are used to evaluate Random Forest models.

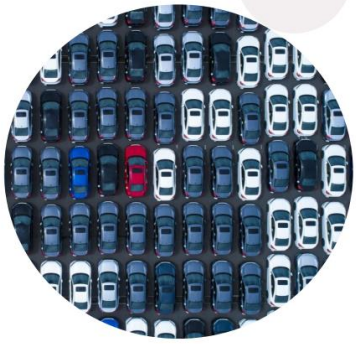


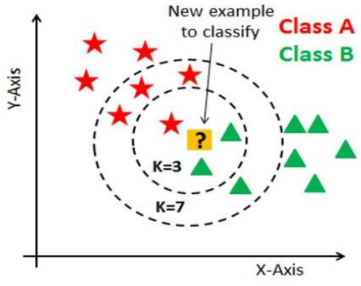
Random Forest - Results

	Sales	Trim	Fuel Type
Accuracy	0.8445384093223215	0.17356318588075575	0.9980201380246634
Precision	0.8778885713099969	0.17466115948138872	0.9987584681617188
Recall	0.8848125316092347	0.17356318588075575	0.9980201380246634
F1 Score	0.8714743041503068	0.1627496613036963	0.9983742320830735
Mean Square Error	14865744.77416272	652153.1019801726	0.00310832531960629
R2 Score	0.9359011234582523	0.9979975056547717	0.9985552889005931

K-Nearest Neighbors

- KNN is an instance-based learning algorithm that is non-parametric and generates predictions based on the similarity of the data points that are entered.
- The KNN models are trained independently for each of the target variables. For example, the knn_sales_model is trained to predict sales, and similarly, the trim and fuel type models are trained as well.
- KNN's important hyperparameter is n_neighbors. The code predicts using the majority class of the 5 nearest neighbours because this hyperparameter is set to 5.
- The distance metric (metric) and other hyperparameters can be modified to meet prediction job requirements.
- After being trained, KNN models are able to produce predictions on the test sets that correspond to them (knn_sales_predictions, etc.).
- To determine how well each KNN model performs, evaluation metrics like Mean Squared Error (MSE) and R2 Score are produced.





K-Nearest Neighbors - Results

	Sales	Trim	Fuel Type
Accuracy	0.7940348743476969	0.0384658898065392	0.7138816608213598
Precision	0.8753232885400432	0.03597739626505151	0.949829303595221
Recall	0.8439220424506715	0.0384658898065392	0.7138816608213598
F1 Score	0.8495153101138857	0.033367592567638	0.8050522553954704
Mean Square Error	11601397.931722593	2572534.472705623	0.6568395746125127
R2 Score	0.9507895036950341	0.9921008031413852	0.6947091033275099

Conclusion

- In terms of Sale, R2 values are similar for Random Forest and KNN models, however KNN has a lower MSE, indicating higher prediction error reduction. Analysis goals and metrics trade-offs determine the decision between Random Forest and KNN. For accuracy and recall, Random Forest may be best. KNN may reduce prediction mistakes.
- In Trim Level Prediction, Random Forest surpasses KNN in accuracy, precision, recall, and F1 score. Random Forest excels at prediction error reduction with a reduced MSE in regression measures. Based on these results, Random Forest is the best Trim Level Prediction model.
- Fuel Type Prediction accuracy, precision, recall, and F1 score are better with Random Forest than KNN. Random Forest has a reduced regression MSE, reducing prediction mistakes. Fuel Type Prediction is best with Random Forest, based on these results.

References

- Bao, Y. (2023, June 09). *EWA Publishing - Applied and Computational Engineering.* Retrieved from Applied and Computational Engineering
- Grelle, C., Franke, V., Verma, A., Miyanyedi, O., Cere', V., & Devadiga, A. (2021, April 04). Used car price prediction in the UK market — Determining the price of used cars based on their features.
- Jin, C. (2021, June). IEEE International Conference on Emergency Science and Information Technology (ICESIT).

Diagram Reference

- Nour Al-Rahman Al-Serw(2021, May 17). *Analytics Vidhya Publishing - K-nearest Neighbor:* The maths behind it, how it works and an example.
- Dr. Roi Yehoshua(March 25). *Published on Medium – Random Forests.*

Thank You

Appendix - C – Ethics Approval Certificate

Certificate of Ethical Approval

Applicant:	Aditya Gujar
Project Title:	Predictive modelling and Analysis: Machine Learning in UK automotive sector.

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	11 Oct 2023
Project Reference Number:	P165625