CATÓLICA LISBON
BUSINESS & ECONOMICS

Examining factors influencing sales of dedicated car mats using car population and machine learning models

Cyprian Frontczak

Dissertation written under the supervision of Professor Nicolò Bertani

Dissertation submitted in partial fulfilment of requirements for the MSc inBusiness Analytics, at the Universidade Católica Portuguesa, 11.09.2024.

Table of Contents

List of Figures

List of Tables

Abstract

Title: Examining factors influencing sales of dedicated car mats using car population and machine learning models.

Author: Cyprian Frontczak

The goal of this thesis is to examine the relationship between designated car mats sales and car features as primary product. It is based on data from Polish car mat manufacturer with leading position in European market. Their process of selecting future car models for production of new car mats lacks data-driven insight to which this article hopes to lay foundation. Various regression and ensemble models were applied. Results indicate the car population and feature popularity the most influential factors in dedicated car mats sales with premium car mat brand being less favoured. The best model achieved an R2 of 0.596, suggesting that further data expansion and advanced techniques are needed for more accurate predictions. Insights from this study can guide strategic product development and marketing decisions for car mat manufacturers. Future research should incorporate broader datasets and additional variables to capture more comprehensive trends.

Keywords: Dedicated car mats, Coefficient analysis, Machine Learning, Predictive models, Car mat sales, Cross-sectional data

Resumo

Título: Examinando fatores que influenciam as vendas de tapetes automotivos dedicados usando a população de carros e modelos de aprendizado de máquina.

Autor: Cyprian Frontczak

O objetivo desta tese é examinar a relação entre as vendas de tapetes automotivos designados e as características dos carros como produto principal. Baseia-se em dados de um fabricante polonês de tapetes automotivos com posição de liderança no mercado europeu. O processo de seleção de modelos futuros de carros para a produção de novos tapetes automotivos carece de uma visão orientada por dados, para a qual este artigo espera lançar as bases. Foram aplicados vários modelos de regressão e de ensemble. Os resultados indicam que a população de carros e a popularidade das características são os fatores mais influentes nas vendas de tapetes automotivos, sendo que as marcas premium são menos favorecidas. O melhor modelo alcançou um R^2 de 0,596, sugerindo que uma maior expansão de dados e técnicas avançadas são necessárias para previsões mais precisas. Os insights deste estudo podem orientar o desenvolvimento estratégico de produtos e as decisões de marketing para o fabricante de tapetes automotivos. Pesquisas futuras devem incorporar conjuntos de dados mais amplos e variáveis adicionais para capturar tendências mais abrangentes.

Palavras-chave: Tapetes automotivos dedicados, Análise de coeficiente, Aprendizado de Máquina, Modelos preditivos, Vendas de tapetes automotivos, Dados transversais

1. Introduction

Car mats serve both practical and aesthetic purposes, protecting vehicle footwells from dirt and wear while enhancing interior design and personalization. However, when it comes to dedicated car mats, since they are meant to fit a specific car model, creating a range of products poses a challenge for manufacturers since they have to choose car model for dedicated car mat to be produced which should generate sufficient sales.

Despite the market's significance, there is a notable lack of research examining the relationship between car features and car mat sales. Currently, car mat manufacturers (CMM) like the leading Polish company being subject to this thesis, rely on specialist judgement and car popularity for selecting car models, which limits their ability to make fully informed, data-driven decisions.

This research addresses the gap by analysing the factors influencing car mat sales. By leveraging statistical and machine learning models, the study aims to provide a systematic analysis of how car features impact car mat sales, allowing to better understand how different factors like brand popularity or age of the dedicated vehicle influence their sales.

It used data comprised of car mats sales of two brands of the same manufacturer, representation of the vehicles to which they are designated and sales data of car brands over according time period in two variations. This dataset was used with common regression models as well as more prominent in machine learning field ensemble models. Both offer differing insight into the relationship between dependent and independent variables.

Regression models showed that influence strength of each coefficient group representing cars features is paired with the population of represented vehicle brand or the feature itself in the data i.e. most popular body types having the strongest coefficient values. The best model being gradient boosting with hyperparameter tuning achieved an R2 below 0.6, indicating room for capturing more patterns in the data. It also showed through feature importance that the population of the vehicle brand has far greater influence over remaining coefficients. The study highlights the need for a larger dataset and potentially more advanced modelling techniques to improve predictive accuracy and provide deeper insights.

In the following chapter I will outline the context of the CM market. Succeeding chapter will focus on the methodology used, regression models, ensemble models as well as performance evaluation techniques and parameters. Followed by the chapter on data creation, and how cars have been represented in the dataset and data cleaning. The 5th chapter presents the results of the research. The final section will offer conclusion on the findings, their practical potential, recommendation for future improvements as well as explanation of the research limitation.

2. Context

Car mats (CM) are an accessory with global market valued at USD 12.5 billion in 2023. It's also estimated for the market to have CAGR of 4.1% between 2024 and 2032 with target of USD 18 billion by 2032[1]. Their practical purpose is to protect the vehicle footwell from elements. As ancillary products, except for functionality they are also a subject to aesthetical consideration, as they contribute to the overall interior design and personalization of the

vehicle, offering a variety of colours, materials, and custom designs to match the owner's style and preferences. Consequently, the market for car mats is driven not only by the need for protection but also by consumer demand for appeal and luxury. This dual demand stimulates innovation and competition among manufacturers, leading to the development of advanced materials and designs that cater to diverse tastes and enhance the driving experience. This adds to complexity of product development relevant to following research which I will explain in this chapter.

This research is based on a single company in Poland with whom I have personal ties. They are a leading manufacturer of dedicated car mats in European market with almost 40 years of experience in the industry. They will be referred as car mat manufacturer (CMM) on the course of this article. For this study I will focus on two CM brands: premium and basic which are very similar from their practical perspective. The premium product is made from thicker material, it has higher edges and contains cosmetic insert with logo. It is also advertised as superior and priced higher. The basic product is made from the same material, however marginally thinner, it has different design, lower edge and no cosmetic inserts. Nonetheless they have the same practical use and both utilise fixing system and cover dead pedal. They share portfolio of cars they are designated to.

In industry of designated car mats, the main challenge lays in repeated development of product for each distinct car model. In case of CMM each CM model requires a car to be scanned, the measurements must be processed in a software by a specialist and then, form needs to be manufactured. On top of that there are several lesser production constraints associated with manufacturing logistics which contribute to a fixed cost of a single CM model manufacturing. This generates a fixed cost for developing a single model for each brand and therefore becomes a prompt for higher deliberation when it comes to choosing a next most profitable car model.

So far there are no available research articles examining relationship between car features and car mats sales, and the company sharing the data uses specialists' expertise and sales numbers of car to be selected. This prompts interest into more deliberate analysis of the relationship between car mats and their primary products in order to better understand the market but also allow for more precise selection of future models. It could also contribute to marketing efforts of the company.

Szukits and Móricz (2022) describe data-driven decision making as crucial and strategic to modern management approaches. Marcy Farell wrote for Harvard Business Publishing (Jan 6 2023) about benefits of data-driven decision making which allows among others for better decision quality, understanding and quantifying complex issues, recognising patterns and more balanced decision making.

Spiller and Belogolova (2016) indicate in their study that if two brands have nonuniform pricing and the premium product is positioned on belief of higher quality, there is evidence that customers would likely choose the more expensive alternative. It is because consumers perceive quality as an objective attribute which are believed to reflect the actual state of the product. This, in the eyes of the consumers justifies the higher price of the product. And such behaviour is expected in the results.

In the next chapter I will present methodology used in this study: what models have been used, performance metrics and hyperparameter tuning for regression forest and gradient boosting.

3. Methodology

3.1. Models

This section is dedicated to presenting a methodology utilised in the research. Five regression and two ensemble models are used, each with specific quality allowing for unique insight into the data. Regression models allow for coefficient analysis and ensemble models offer superior predictive performance.

3.1.1. Linear Regression

Linear regression is a statistical technique used to model and analyse the relationship between a dependent variable and one or more independent variables. The primary goal of linear regression is to establish a linear relationship between the dependent variable and the independent variable. There are two types of linear regression: univariate and multivariate. The difference is the prior uses one independent variable to explain the dependent variable whereas the latter utilises multiple independent variables for this purpose. As data in this research contains multiple variables, we will be using multivariate regression following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Here *y* represents the dependent variable, β_n values are the regression coefficients, x_n are the independent variables and ε represents the error.

The purpose of linear regression is to model the relationship between a dependent variable and multiple independent variables which can serve multiple purposes like predicting the dependent variable based on the regressors values, understanding how they influence the dependent variable individually or quantifying the relationship between dependent variable and independent variables and modelling those complex relationships.

3.1.2. Stepwise regression

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration. The availability of statistical software packages makes stepwise regression possible, even in models with hundreds of variables. By eliminating some variables, it creates a model that is simpler and possibly more accurate than if one containing all variables. It is one of the most basic regressions. Due to its parsimonious approach, it allows for balanced complexity and explanatory power of the model resulting in good interpretability, Kuhn, Johnson (2013).

There are three types of stepwise regressions:

Forward Selection starts without any variables, it tests each variable as it is added to the model and keeps the one which are statistically significant. The process is repeated until optimal model is built.

Backwards Elimination is a reverse version of Forward selection. The model starts with all variables included which then are removed one at a time. Then model is tested for statistical significance and the process is repeated until combination of the most optimal variables in the model is refined.

Bidirectional elimination is the combination of both previous methods which tests at each step which variables should be included or excluded.

The main benefit of stepwise regression is that it helps create a parsimonious model that balances complexity and predictive power. However, according to Whittingham et al. (2006), stepwise regression can lead to biased parameter estimates, overfitting, and incorrect significance tests due to model selection based on parameter significance. The outcome can vary depending on the type of stepwise selection used and since it identifies a single best model it ignores the potential of multiple models fitting the data well which can lead to overconfidence and misleading conclusions.

3.1.3. Lasso Regression

Lasso stands for Least Absolute Shrinkage and Selection Operator. Used in machine learning to handle high dimensional due to its automatic feature selection functionality. Also known as L1 regularization it does this by adding a penalty term to the residual sum of squares (RSS), which is then multiplied by the regularization parameter (lambda). As a result, some coefficients can be shrunk to zero, essentially removing them from the equation. The penalty is controlled by lambda where higher value results in more coefficients being shrunk to zero. Lasso regression follows such equation:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Where 뭇i is the dependent variable, β_0 is the intercept, β_j are the coefficients of the predictors

叭ij, λ is the regularization parameter that controls the strength of the penalty, it is alsoconstrained:

$\lambda \geq 0$, and the number of predictors is represented by 支.

Such penalization of coefficients helps to avoid multicollinearity and overfitting of the model. It promotes model sparsity which should be helpful in dataset with many variables (Kuhn, Johnson, 2013).

3.1.4. Ridge Regression

Ridge regression, also known as L2 regularization works similarly to Lasso regression, however it does not shrink the coefficients to zero. It helps to stabilize the model and works in favour of smaller datasets by balancing bias-variance trade-off, improving generalization without performing feature selection. It's represented by equation:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Where 웃$_i$ is the dependent variable, 鄕$_0$ is the intercept, 鄕$_j$ are the coefficients of the predictors 叭$_{ij}$, 劃 is the regularization parameter that controls the strength of the penalty, it is alsoconstrained: 劃 ≥ 0, and the number of predictors is represented by 支.

It handles multicollinearity well, however highly corelated estimators can become unstable and have large variances. This is mitigated by penalization of the size of the coefficients. Since it is not performing feature selection it has potential use for cases when dataset contains predictors with multicollinearity but all hold relevance. (Kuhn, Johnson. 2013)

3.1.5. Elastic Net Regression

Elastic Net regression combines both L1 and L2 regularization from Lasso and Ridge regressions allowing it for both shrinking or eliminating the coefficients. It is largely beneficial for complex datasets with multicollinearity problem where model allows for feature selection and balancing coefficients. It handles correlated features better than Lasso which tends to randomly select one feature from a group of highly correlated features in contrast, EN tends to include groups of correlated features in or out of the model together. It is represented by equation:

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1 \right)$$

Where: 웃 is the response factor, 昼 is the matrix of predictor variables, β is the vector of coefficients, 잊 is the number of samples, $||$잊 − 昼β$||^2$ is the residual sum of squares, $||\beta||_1$ is the L1 norm, $||\beta||^2$ is the squared L2 norm, α is the parameter mixing L1 and L2 penalties, and λ is the regularization strength.

The Elastic Net is positioned as a superior method for regression shrinkage and selection, particularly in high-dimensional data scenarios like microarrays. It effectively balances the strengths of lasso and ridge regression, making it a valuable tool for statistical modelling (Zou, Hastie. 2003).

3.1.6. Random Forest

Introduced by Leo Breiman in 2001, Random Forest is a machine learning algorithm that can be used in classification and regression tasks by combining multiple regression trees. it starts by generating multiple bootstrap samples from the original dataset. They are created randomly, and each data point can be chosen multiple times. For each sample a tree model is trained independently and in case of regression, the results are being averaged for most accurate prediction. Additionally, each bootstrap sample has one third of the data set aside as out-of-bag sample which is later used for cross-validation.

As for tree models, they work by splitting the data into a tree-like structure mapping their decisions which are made based on feature values. The structure of regression tree made of three elements: Root node, which is the starting point of the model – it represents the entire dataset. This is later split into subsets which are decision nodes – they split the data based on specific features until reaching Leaf nodes – terminal nodes representing the final output of class label. In case of regression the splitting is done by the algorithm based on variance reduction or with goal of choosing most effective separation into distinct classes. The splitting is stopped until a criterion is met which could be maximum tree depth, minimum number of samples per node or minimum improvement threshold. The main disadvantages of a decision tree are that they are prone to overfitting and are unstable since small changes to data can greatly influence the tree structure. They are also biased in feature selection where they prefer features with more levels. This is where Random Forest becomes a more reliable model. By utilising mentioned before mechanisms it allows for more reliable results with proper tuning, minimising the risk of overfitting. With such complexity main drawback of this method is it's computationally expensive (Genuer et.al. 2008).

3.1.7. Gradient Boosting

The principle of Gradient boosting method is to combine multiple simple models to create a stronger model with superior performance. It works by building an ensemble of models, typically decision trees, in a step-by-step fashion. Each new model is trained to predict the residual errors of the combined ensemble of all previous models. This iterative process continues until a specified number of models have been added or model performance reached plateau.

The process begins with an initial model, often a simple decision tree, that makes predictions on the training data. Then the residuals are calculated. They represent the errors the model aims to correct in the subsequent steps. A new decision tree is trained to predict the residuals. This new model focuses specifically on the errors made by the initial model, thereby learning to correct these mistakes. The predictions from the new model are combined with the existing model predictions. This combination is controlled by a parameter called the learning rate, which dictates the contribution of the new model to the overall predictions.

Gradient Boosting offers flexibility with several parameters available for tuning:

- Number of iterations,
- Learning rate,
- Subsample fraction,
- Loss function,

- Early stopping,
- Hyperparameters of the chosen base learner.

It is also known to have high performance, winning numerous competitions, it's also versatile for different types of data, provides insight into feature importance allowing for determining which variables carry the biggest impact for the model. It is also robust against overfitting including datasets with noise; however, it highly depends on proper hyperparameter configuration. This model can also be expensive computationally (Natekin, Knoll. 2013).

3.2. Performance metrics

In all cases models used train and test split 80-20% where 20% is used for testing and remaining 80% for training in a random manner using basic R function 'sample()'. For assessing model's performance three measures were used as proposed by Khun and Johnson (2013):

- **R Squared (R^2)** is a statistical measure indicating proportion of the variance in the dependent variable that is predictable from the independent variables. It can be interpreted as how much data is explained by the model. It fits on the scale 0-1 where 1 means that 100% of the variance in the dependent variable is explained by the model. It can be explained by formula:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals (SSR)Total}}{\text{Sum of Squares (SST)}}$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  Where: SSR is the sum of squared differences between the observed values and the predicted values. SST is the sum of squared differences between the observed values and the mean of the observed values. y_i is the observed value of the dependent variable, $\hat{y}_i$ is the predicted value, and $\bar{y}$ represents the mean of the observed values.

- **Root-mean-square-error (RMSE)** is a performance indicator measuring the average difference between values predicted by a model and the actual values. It allows for estimating models' accuracy by measuring how well is it able to predict the target value.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

  y_i is the observed value of the dependent variable, $\hat{y}_i$ is the predicted value, and n isthe number of observations.

- **Mean absolute error (MAE)** calculates the average magnitude of errors between predicted values and actual values without considering the direction of the errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  y_i is the observed value of the dependent variable, $\hat{y}_i$ is the predicted value, and n isthe number of observations.

Both RMSE and MAE can be interpreted as either how far (on average) the residuals are from zero or as the average distance between the observed values and the model predictions. It's important to note that they are interpreted within the scale of the dependent variable.

3.3. Hyperparameters

Both Regression Forest and Gradient Boosting models were subjected to hyperparameters tuning. Hyperparameters are external configuration parameters of the model used to manage its training. In this research at first trainControl and expand.grid functions from the caret package are used to perform a grid search for the more optimal parameters which later are used as reference for manual tuning with goal to obtain most optimal performance measures. Another reference of model's performance was scatterplot of actual and predicted values with perfect prediction slope and loess line.

- Locally Estimated Scatterplot Smoothing (LOESS) represents a smoothed curve that captures the relationship between the actual and predicted values. It works by fitting multiple regressions in local neighbourhoods of the data and gives more weight to the points closer to the line.

The next chapter will explain the data used for the research as well as the steps taken to create a final dataset to be used for the analysis. It will also present a reasoning behind chosen representation of the vehicles in the data.

4. Data and transformation

This section describes the process of building the dataset, composed of three separate sources. It explains choices and logic behind form of representation of cars and how the population of car brands has been displayed. Lastly, coefficient multicollinearity will be examined resulting in creation of sibling dataset which will be used in the analysis parallel to the original dataset as it shows promise of better performance.

4.1. Representation of cars in the model

For this research, it's important to define two main categorising logics for the following analysis: definition of the car and submodels. Due to complexity of automotive market many terms and definitions have room left for the interpretation. Therefore, throughout this thesis I have adhered to the interpretations of the CMM.

The main challenge of defining a car for this research was posed by multiple features describing a vehicle on the market and behaviours connected to cars and automotive societies. Personal vehicles serve not only a variety of practical uses but are beacons of status, passion and form of expression as well. Those are tightly tied to psychological patterns of their owners and therefore are outside of scope of this research. As an example, some car models are strongly favoured as subjects of tuning or become automotive icons, therefore their market price is based not only on utilitarian capabilities but become biased to be chosen by more passionate owners. This can be specific to brands or single models. People who are passionate about their vehicles are also more prone to purchase accessories i.e. car mats. Unfortunately, there is no research or list of such car models group which is purely subjective and this behaviour cannot be documented in this research. In order to create a reliable, general representation of cars for this model I chose three main features that are dependable and define well each car's position on the market: Car Brand, Body Type and Segment. On top of that I have included production years to record age of the vehicle. Car brand represents manufacturer and is indispensable for defining a vehicle. It also carries specific associations on the market. Body type aside from defining the vehicle as well, allows for representation of utilitarian values a vehicle offers: SUV and kombi offers much more trunk space; therefore, it can be assumed that they will be subjected to more practical exploitation which could require additional protection of the interior in form of CM. Both define the vehicle well; however, their representation of vehicle's status is subjective to the age of the vehicle and therefore position on the market as it is well-known that cars depreciate over time and brand's statuses shift alike. Hence introduction of segments which are impervious by time. Segments categorise cars by their size and prestige and are represented by first letters of the alphabet with A being the smallest and cheapest cars and F being the largest and most expensive. I chose to adhere to CMM's information system which is Polish Wikipedia pages on every car model. Those websites are renowned for their meticulous and detailed vehicle description, providing uniform and clear car classification by segment. Its reliability is described by Voss (2005) as result from open-source nature of the website, with multiple markers and mechanisms, minimising error inducing factors. This was particularly important as vehicle classification is not centrally regulated and can vary throughout sources and institutions. Therefore, provided trustful source, used by the company itself, with detailed and precise vehicle segment classification it allowed for data reliability and uniformity. Few car models lacking necessary documentation in the chosen source were categorised accordingly to company's expert considerations.

The issue of submodels is more specific to vehicle accessories and applies directly to CM research. Each car model can be divided into what I will be referring to as submodels. In this research submodel is understood as a variant of a vehicle of the same brand, model and generation which equipment specification results in different shape of the vehicle's floor. As an example: Alfa Romeo Giulia is a sports sedan manufactured under "Alfa Romeo" brand, which model is "Giulia" and as of currently there are two existing generations of this vehicle with current one being manufactured since 2016. In this case we can divide it into two submodels: with all-wheel-drive (AWD) which supplies power to all four wheels and rear-wheel-drive (RWD) which powers only the rear wheels of the vehicle. This is because added complexity of the AWD requires more space in the transmission tunnel which in turn alters the shape of the front footwell. Because of that company offers two distinct car mat sets within the same CM brand for the same car model. Other examples could be: longer variant of

the vehicle, hybrid drivetrain, automatic/manual gearbox, pre/post-lift, etc. Despite some features following a rule of thumb they cannot be defined within this analysis as some vehicles do not fall subject to change of floor's shape from varying equipment as others do, hence they do not possess different CM models and submodels are not registered within company's datasets. Because of that it's impossible to split it into submodels as distribution of CM sales would be unknown. This issue will be addressed in later part of this chapter.

Lastly, some cars share platforms. Those are basic structures upon which vehicles are built and it became a common practice for car manufacturers to share platforms for production since it contributes to economies of scale and in turns lowers the production, as well as development costs. As a result, different car models from different brands can share CM set. An example would be Audi Q8 being twinned by Lamborghini Urus with the same engine and built on the same platform, who both share car mats set. This behaviour has been described by Zac Palmer for Autoblog (June 2 2022)[2]. It's also important to note that different generations of the same vehicle model in general differ largely, therefore they do not share a CM set and virtually in all cases are separate in the data.

4.2. Building the dataset and transforming the dependent variable

The main dataset is company's sales data comprised of individual orders for every offered product in format of time period of 9 months starting from January 2023. It has been shared directly by the representative of the company for the purpose of this research. The chosen time period also covers both high and low periods of the company's sales cycle which is from January to September. Choice of this data was also limited by integration of the new data management system, which had no accessible records over the past beyond January of 2023. It contains all orders for European contractors in B2B department. Allowing for direct representation of the demand representative of the whole continent. After filtering the dataset from the customers details and irrelevant information what was left was: data on product's ID, quantities sold per order and product description, which also contained vehicles description to which each product was dedicated to. For the analysis this data has been compacted into cross-sectional format by combining sales of the same products into one position, ultimately resulting in quantities of every unique car mat set sold over 9 months period.

As the next stage data required annotation in order to represent vehicles. For each unique product ID, a fitting vehicle description was extracted. This list has been sent to professional annotation company. The goal was to extract each vehicle's brand, body type and class in form of dummy variables. One of the reasons for choosing dummy variables was to address previously explained behaviour of single CM fitting multiple vehicles. This resulted with each car mat description contents to be converted into dummy variables of three groups: brand, body type and segment. After receiving annotated data, it was bound by ID to the original dataset.

Next, I chose to address the submodel issue. Because it's impossible to determine distribution of CM sales among different submodels for the car models that do not undergo such division, I chose to combine submodels into a single observation. Referring to previous example of Alfa Romeo Giulia, instead of appearing as two ID's, it would now figure as a single register of Alfa Romeo Giulia. Within this process I have created two additional variables

representing number of combined submodels and flag for combined positions. The process of eliminating submodel level resulted in uniformity among all data points.

The last stage of creating the dataset was to include the total number of sales for each vehicle's brand per year in Europe. This would allow for representation of each vehicle's model estimated population which was presumed to have direct correlation in number of cars sold as the CMM itself uses quantities sold for model selection. Unfortunately, I was unable to reliably acquire sales of each car model with chosen criteria for all observation, therefore upon finding consistent brand sales data I chose to utilise it. The dataset was downloaded directly from European Automotive Manufacturers' Association website[3] and it covered all brands sold in Europe from 1990 to year 2023. In order to implement it into the main dataset, using Excel, I removed all brands sales that were not included in the main dataset. In this form each year was represented in separate sheet. Using VBA code, I compiled data from all sheets into long format table, which then using Pivot Table tool in Excel was converted to wide format with columns for years and rows for brand names. Only one car brand sales, accounting for 4 observations, were not available and it has been removed from the dataset.

The main challenge of implementing this data into the main dataset was addressing single CM ID fitting multiple brands. This would mean that it was necessary to represent sales of combined brands over multiple years. To solve this problem, several matrices were created based on main dataset. Each row was represented by manufacturer car brand corresponding to product's ID, and columns were representing each year from 1990 to 2023 with dummy variables corresponding to each year a car has been produced. The number of matrices was determined by number of distinct car brands sharing a row. Conveniently, despite CM fitting distinct models, they all shared the same production years in Europe. I assume this was a result of technical/logistical/financial convenience of platform sharing among manufacturers. For the next step each matrix has been multiplied by EU sales data frame and then summed together.

Additionally, three variables were created:

prod - Flag for cars that were still in production in 2023,

prod_lenght - Length of production of a car in years,

v_age - Age of the vehicle since the first production year until 2023.

At this stage dataset consists of 975 observations. The dependent variable *quant*, representing CM sold per observations ranges from 1 to 1721 units. Mean of quant variable is 154.6 units and median: 101.

Figure 1 displays distribution of quantity variable. Here y-axis shows number frequency of quantities from x-axis. It's immediately apparent that the distribution is heavily right-skewed. There are also several possible outliers.

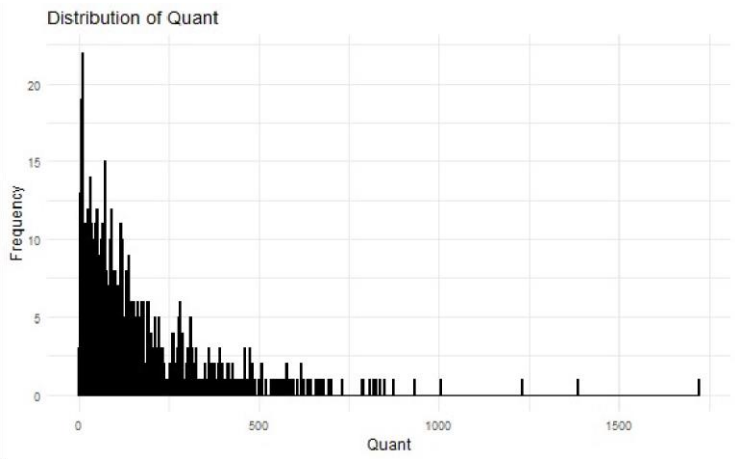

Figure 1 Distribution of the dependent variable quant

To address this, I applied logarithmic transformation with goal of bringing the distribution to the shape of normal distribution.

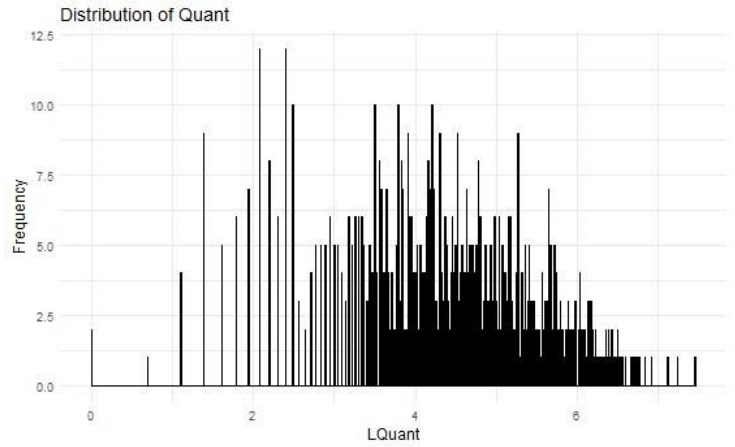

Figure 2 Distribution of the dependent variable quant after logarithmic transformation (now log_quant)

Despite displaying features of Gaussian distribution, in current state log_quant contains features of bimodal distribution. It's a type of multimodal distribution and its main characteristics are two peaks, in this case the major mode is around 2 on x-axis and the minor mode, around 4. This would suggest that current dataset is a sum of two separate groups with different modes, therefore they should be split and analysed separately. This, in the light of a fact that the main dataset is a combination of two distinct products implies that different brands follow separate normal distributions. However, upon graphing each brand separately it proves not to be the case.

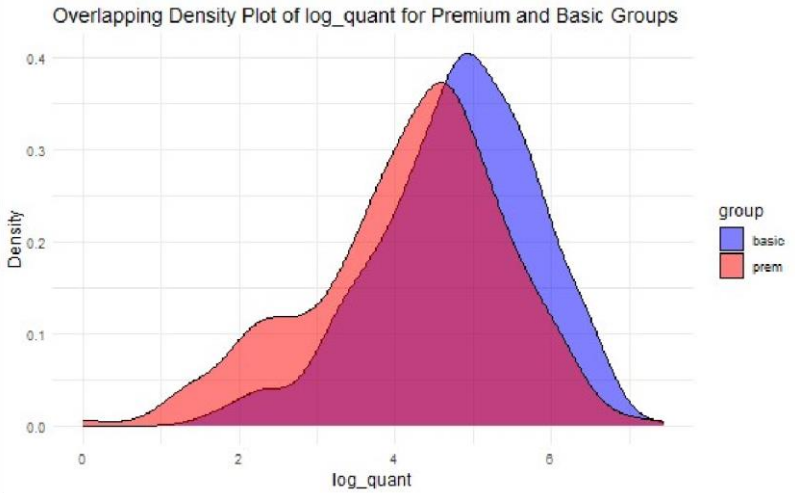

Figure 3 Overlapping Density Plot of log_quant for premium and Basic CM brands

By plotting an overlapping density graph for both groups, we can see that they share similar distribution patterns. This is further supported by summary statistics, where median and mean is 4.43 and 4.29 for premium group and 4.89 and 4.83 for basic group respectively. It's also apparent that both groups distribution share a small peak between log_quant = 2 and 3 on the x-axis which proves that bimodal distribution in the main dataset is not a result of merging two groups of brands. Because the secondary peaks are not sufficiently defined for both premium and basic brands datasets to allow splitting, I chose to address that by removing potential outliers. Several groups of data were removed. First was determined by z-scores, second was compiled by car brands that were either lacking EU sales data or, in case of Saab, have been long gone from the market what makes this brand irrelevant for this study. Lastly, some CM models that have been introduced by CMM during the analysed time period were removed.

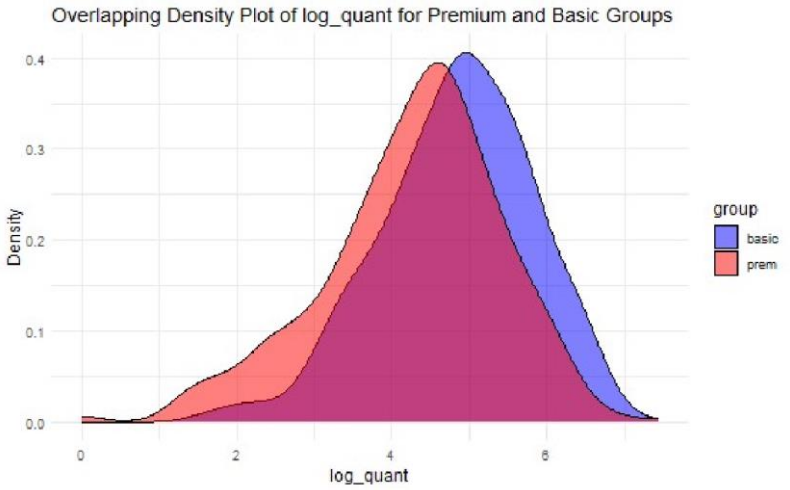

Figure 4 Overlapping Density Plot of log_quant for Premium and Basic CM brands after outlier removal

As a result, the smaller peaks shared by both distributions have been reduced. In this state data is fit to be used in models. Table 1 displays description of all variables. In total data contains 62 variables and 926 observations.

Name of variable	Type of variable	Description
log_quant	Continuous	Dependent variable quant representing quantity of CM sets sold. After logistic transformation
A	Dummy	Group: Segment. Vehicle segments. 0: car to which CM is fitted does not belong in the segment, 1: car to which CM is fitted does belong in the segment
B		
C		
D		
E		
F		
SUV	Dummy	Group: Body Shape. Body style of the car. 0: car to which CM is fitted is not made in this body style, 1: car to which CM is fitted is made in this body style
Coupe		
Fastback		
Hatchback		
Kabriolet		
Kombi		
Kombivan		
Liftback Minivan		
Pick.up		
Sedan		
Crossover		
VAN		
Alfa.Romeo	Dummy	Group: Car Brand. Car brand to which CM is fitted. 0: car to which CM is fitted is not made under this brand, 1: car to which CM is fitted is made under this brand
Audi		
BMW		
Chevrolet		
Citroen		
Cupra		
Dacia		

Dodge		
DS		
Fiat		
Ford		
Honda		
Hyundai		
Jaguar		
Jeep		
Kia		
Land.Rover		
Lexus		
Mazda		
Mercedes		
Mini		
Mitsubishi		
Nissan		
Opel		
Peugeot		
Porsche		
Renault		
SEAT		
Skoda		
Smart		
SsangYong		
Subaru		
Suzuki		
Toyota		
Volkswagen		
Volvo		
1990 - 2023	Discrete	Group: Yearly Car Sales (YCS). Sum of total yearly sales for each car brand listed in the observation for the specified year.
is_comb	Dummy	Denotes whether observation is combined of several CM sets
n_comb	Discrete	Denotes number of CM sets that were combined into single observation
prod	Dummy	0: the car is no longer in production, 1: the car is in production as of 2023
prod_lenght	Discrete	Number of years the car was produced
v_age	Discrete	Age of the vehicle counted in years since the first year of production until 2023
prem	Dummy	Marker for CM brand. 0: denotes basic brand, 1: denotes premium brand
pop	Discrete	Sum of total yearly sales for each car brand listed in the observation over the specified years. Represents a total population of the brands over specified years on the European market (sum of 1990 – 2023 variables)

Table 1 Detailed list of all variables contained in the dataset.

4.3. Multicollinearity and sibling dataset

In order for the data to generate reliable results for regressive models as well as allow for their interpretability, it's important to handle the issue of multicollinearity. This happens when multiple predictors within the data are highly correlated to each other. The reason for its importance is that when data contains variables with high collinearity their individual effect on dependent variable cannot be easily distinguished, therefore making the model hard if not unfeasible for interpretation since correlated predictors share information. Moreover, multicollinearity heavily contributes to overfitting, especially for out-of-sample testing. There is also a risk that in more severe cases, the model can become sensitive, this means that minor changes can result in big fluctuations of the results which in turn renders the model unreliable. Therefore, in order to create a reliable dataset for the research it is an indisputable step.

The graph below illustrates heat map Correlation Matrix for all variables used in the model.

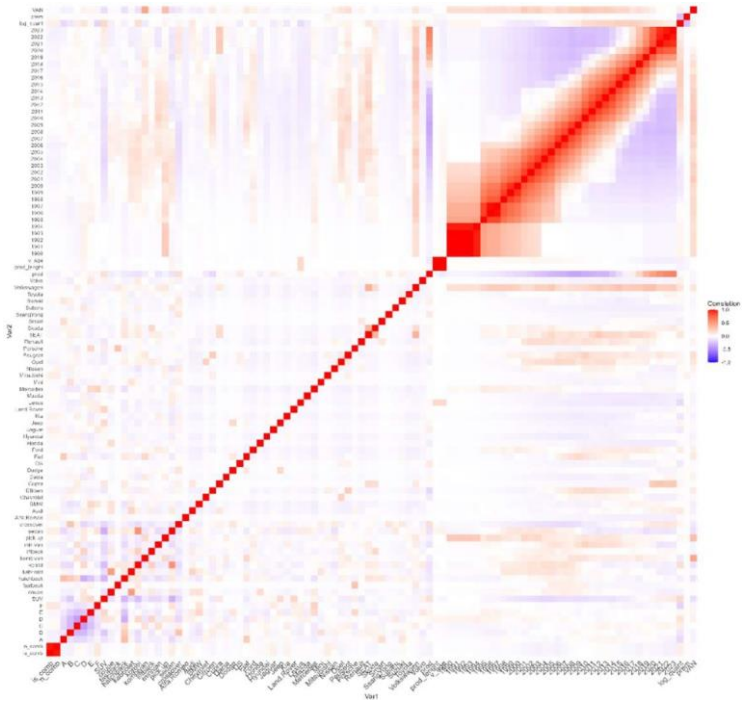

Figure 5 Correlation Matrix of all features including YCS group

The main feature represented by this matrix is the high correlation pattern between yearly population variables from 1990 to 2023 which are defined on the graph by the wide red section. For this reason, these variables will be combined into a single *pop* variable, representing the sum of those yearly population variables. This ultimately describes the population of brands per observation for the production years. Additionally, number of variables will be greatly reduced, simplifying the model which could improve the predictive power since the data has relatively little observations thus reducing its complexity should allow the models to develop better generalizations.

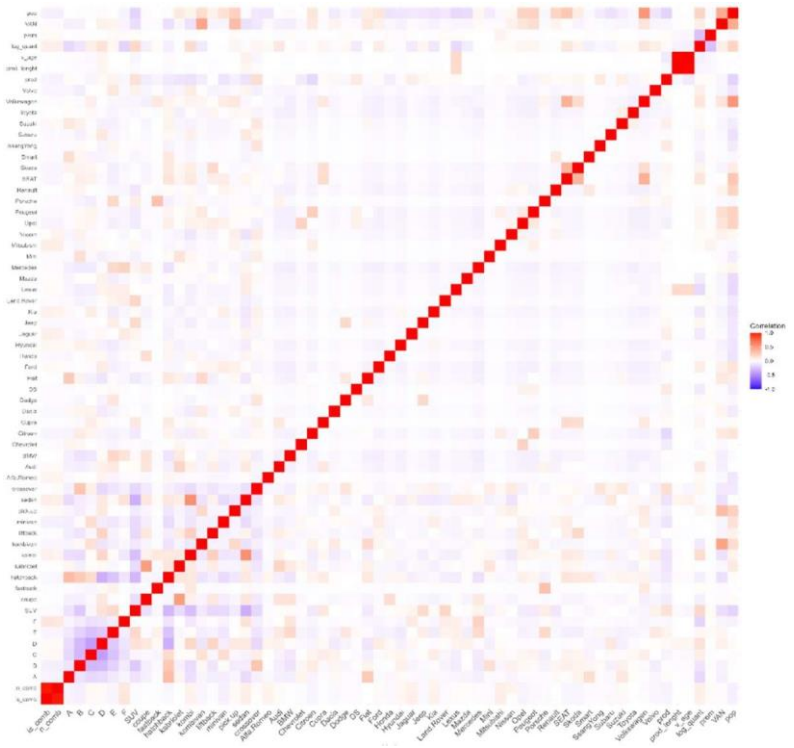

Figure 6 Correlation Matrix of all features including pop variable as an alternative to YCS group

After reduction of car population variables two more independent variables show correlation. Variables is_comb and n_comb display the same pattern, however n_comb carries additional information therefore it will be favoured over is_comb. Variables prod_lenght and v_age also share very high collinearity. In this case v_age will be selected.

The next chapter offers the results of presented methodology on data with both representation of car brand population. It will begin with performance of regressive models and continue into ensemble models before and after hyperparameter tuning. For each group the best performing model will be examined in detail.

5. Results

The aim of this research is to examine relationship between car features and sales of car mats. Following sections will present results for both regressive and ensemble methods with detailed description of best out-of-sample performing models from each group. Due to scarcity of observations the best performing regressive model will be trained on the whole population and its coefficients will be interpreted in context of the data. The best performing ensemble model will be interpreted for feature importance.

5.1. Regressive models

Two variants of data have been used in the models. The first variant represented car brands population using YCS group, and the other subsidized it with *pop* variable. The reasoning for it was to compare whether variable simplification associated with use of *pop* variable would

benefit the model's performance as data with YCS group contained 92 variables whereas its pop counterpart – 61 which could play significant role due to small size and complexity of data.

As such data contained variables: *log_quant*, *segment* group, *body shape* group, *car brand* group, *prem*, *prod*, *v_age* and either *yearly car sales* group or *pop*.

For all models measures for out-of-sample performance were: R^2 score, Root Mean Square Error and Mean Absolute Error. Starting with regression models, 5 models were developed: Linear Regression, Forward Stepwise Regression, Lasso Regression, Ridge Regression, and Elastic Net Regression.

Model	R-squared	RMSE	MAE
LR YCS	0.381	0.866	0.701
LR pop	0.377	0.859	0.694
FS LM YCS	0.352	0.874	0.698
FS LM pop	0.314	0.900	0.712
Lasso YCS	0.381	0.853	0.689
Lasso pop	0.374	0.853	0.682
Ridge YCS	0.392	0.846	0.679
Ridge pop	0.383	0.852	0.686
EN YCS	0.385	0.851	0.682
EN pop	0.376	0.856	0.692

Table 2 Performance metrics for all regression models for both datasets

Starting with Linear Regression, in out of sample performance data with YCS has achieved R^2 of 0.492, RMSE: 0.773 and MAE: 0.630 whereas its *pop* counterpart achieved R^2: 0.455, RMSE: 0.801 and MAE: 0.650. Since both models include high number of variables, next model was Forward Stepwise Selection. Contrary to model utilising *pop* variable, the YCS variant produced improved metrics however minimal. This trend follows through all regressive models.

Figure 7 Summary of performance metrics of regression models

Based on presented metrics we can conclude that Ridge model for YCS data variant has performed best with highest R^2 of 0.392 and lowest RMSE of 0.846 and MAE being 0.979. This implies that regardless of added complexity in pair with scarce observations, larger number of the variables have beneficial influence on models' performance. This is the case for almost all models presented in this section with only Lasso and Linear models having inferior performance in MAE metrics for YCS dataset and Linear RMSE compared to pop variant. Compared to Lasso and Elastic net models which offer more functionality than ridge model it would also indicate that variables in the dataset are indispensable in context of whole since the Ridge as the only model except linear regression does not perform variable selection.

Following these results, because of the small number of observations a new Ridge model was trained on the whole YCS dataset to achieve more accurate results for analysing coefficients.

(Intercept)	3.520	Porsche	-0.679
n_comb	0.373	Renault	0.354
A	-0.595	SEAT	0.335
B	-0.124	Skoda	0.351
C	0.072	Smart	-0.204
D	0.119	SsangYong	-0.288
E	-0.019	Subaru	-0.483
F	-0.759	Suzuki	0.305
SUV	0.804	Toyota	0.541

coupe	-0.214	Volkswagen	0.492
fastback	0.518	Volvo	0.594
hatchback	0.107	prod	0.160
kabriolet	0.154	v_age	-0.009
kombi	0.575	X1990	0.027
kombivan	-0.872	X1991	0.025
liftback	-0.238	X1994	-0.015
minivan	0.081	X1995	0.004
pick.up	0.389	X1996	0.019
sedan	0.069	X1997	0.023
crossover	0.441	X1998	-0.014
VAN	0.561	X1999	-0.029
Alfa.Romeo	0.455	X2000	0.009
Audi	0.702	X2001	-0.023
BMW	0.475	X2002	-0.022
Chevrolet	-0.030	X2003	0.000
Citroen	0.259	X2004	0.013
Cupra	-0.719	X2005	0.007
Dacia	0.548	X2006	-0.012
Dodge	0.461	X2007	0.010
DS	-0.997	X2008	-0.006
Fiat	0.447	X2009	-0.001
Ford	0.545	X2010	0.014
Honda	0.081	X2011	0.006
Hyundai	0.165	X2012	0.002
Jaguar	-0.608	X2013	0.009
Jeep	0.037	X2014	-0.003
Kia	-0.284	X2015	0.012
Land.Rover	-0.036	X2016	-0.005
Lexus	-0.247	X2017	0.013
Mazda	-0.225	X2018	0.007
Mercedes	0.348	X2019	-0.001
Mini	0.184	X2020	-0.002
Mitsubishi	0.136	X2021	0.020
Nissan	0.165	X2022	-0.001
Opel	0.468	X2023	-0.015
Peugeot	0.377	prem	-0.525

Table 3 Coefficient output of Ridge regression model for YCS dataset

Analysing the coefficient output of YCS Ridge regression we can see that premium factor has negative influence over sales, meaning that contrary to Spiller and Belogolova study, premium products are less likely to be purchased over the basic substitute. In the segment group, first two and last two segments (A, B, E, F) had negative coefficients with segments C and D being only positive ones, this follows vehicle population represented by YCS group where mentioned segments are the most popular, representing 60% of the vehicle population. This leads to another observation where segments' A and F coefficients are far more

impactful than the rest from this group. This is interesting because they are the least popular representing roughly 8% of the population. The reason for this could be that the remaining segments are far more popular and therefore cars represented by them are more diverse, having multiple body types offered and being subject to platform sharing.

Body type group and brand group both seem to follow the population of each, with the most popular variables producing the most significant positive coefficients.

It's important to notice that none of the regressive models reached R^2 higher than 0.5.

5.2. Random Forest and Gradient Boosting

In Random Forest models the performance was expectedly better than regression models. Interestingly, for those models the dataset utilising pop variable has exceeded YCS counterpart performance-wise. In the table below we can see the different performance metrics for all combinations of data and models before hyperparameter tuning. It's also worth noting that without tuning, gradient boosting model does not generate any significant performance metrics.

Model/data	R^2	RMSE	MAE
RF YCS	0.400	0.892	0.687
RF pop	0.431	0.875	0.679
GB YCS	0.005	1.157	0.917
GB pop	0.006	1.157	0.916

Table 4 Performance metrics for random forest and gradient boosting models before hyperparameter tuning

5.3. Hyperparameter tuning

With the goal of improving the model's performance, both Random Forest and Gradient boosting models hyperparameters were tuned. First, a grid search was performed for finding a general direction and later, based on the best results each model was manually tuned for best performance.

Model/data	R^2	RMSE	MAE
RF YCS	0.416	0.886	0.680
RF pop	0.442	0.866	0.672
GB YCS	0.568	0.762	0.600
GB pop	0.596	0.737	0.562

Table 5 Performance metrics for random forest and gradient boosting models after hyperparameter tuning

We can see that Gradient boosting model became competitive after hyperparameter tuning and outperforms random forest. Similarly to models before, data using pop variable had greater performance than its YCS counterpart. The best performing model was GB pop with $R^2 = 0.596$, RMSE = 0.737 and MAE = 0.562. Hyperparameters used for this model were:

n.trees = 90000
interaction.depth = 10
shrinkage = 0.001
cv.folds = 5
n.minobsinnode = 7

Examining the best performing model closer (GB pop), we can see on the graphs below that it is still not capturing the full pattern within the data. As visualised by loess curve which forms a curvy line on a different angle than the ideal relationship red line. This suggests nonlinearities or biases in the prediction.

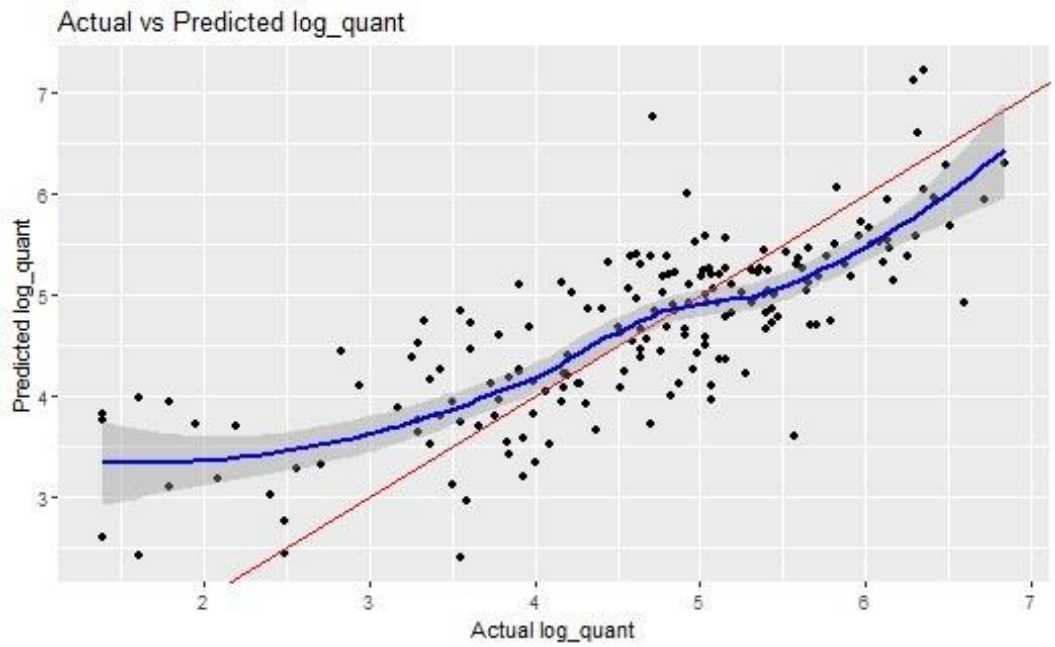

Figure 8 Scatterplot of Actual vs Predicted values of dependent variable log_quant with perfect relationship line (red) and loess curve (blue)

Examining models' feature importance graph we can see that vehicle population and age as well as premium factor of the car mat brand are the most influential in the model.

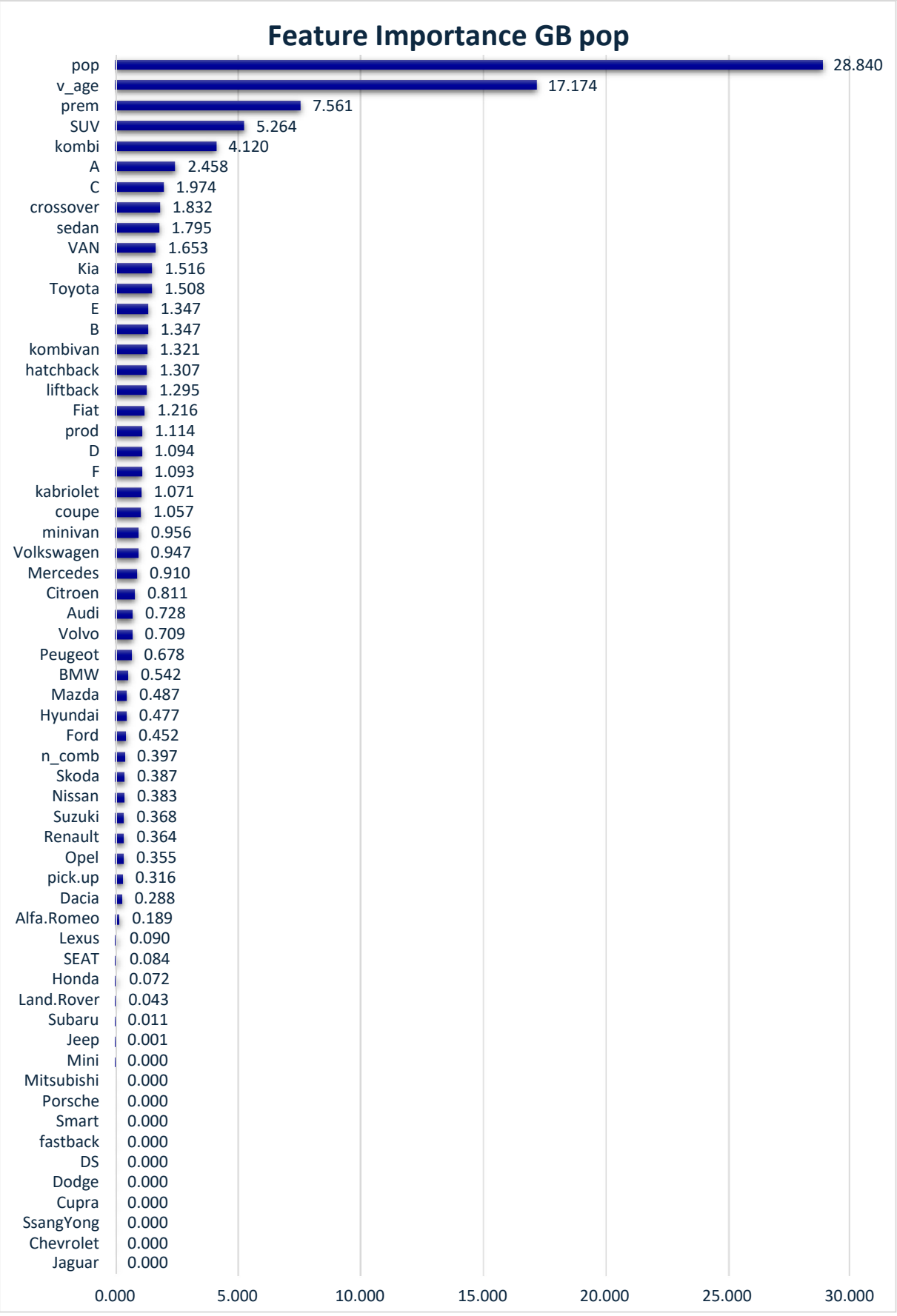

Figure 9 Feature importance for Gradient Boosting model for dataset with pop variable with best hyperparameter tuning

The following, final chapter offers a conclusion based on the findings, practical use of the research, and limitation as well as suggestions for improvements in the future.

6. Conclusions

The point of this research was to explore factors influencing the sales of designated car mats. Several models were tested in out-of-sample performance rendering Ridge regression and Gradient boosting best in their class. From their output we can conclude that the population of primary product i.e. cars is by far the most influential factor in number of sales of dedicated car mats. We can also conclude that contrary to Spiller and Belogolova (2016) study; while sharing the fundamental functionality, the premium variant of CM is less likely to be sold over its basic counterpart. This means that despite premium marketing, additional cosmetic features and added functionality of higher edges and thicker material, consumers are more likely to choose cheaper, more basic alternative of the CM brand.

For the CMM, these insights can inform strategic decisions in product development and marketing. Focusing on car models and segments with higher sales can optimize resource allocation for the production of a new model, while emphasizing popular car brands and body types in marketing efforts can enhance market penetration as well as efficiency. The negative performance of premium car mats indicates a need for reevaluating the marketing and pricing strategies for these products. However, it is important to notice that since the company is under dynamic development, this could be subject to change due to sales department efforts and continuous B2B partners portfolio expansion.

It cannot be ignored that even the best model has achieved R^2 below 0.6 and its loess curve deviated from the perfect line. This indicates that there is quite a lot of patterns in the data to be captured. I'm confident that the primary reason for this is the number of observations being too small for the high complexity of variables therefore not allowing for reliably capturing all patterns within the data. It is also possible that the method of representation of the vehicle was suboptimal, and not representing the cars in the best form for models used in this study as hinted by the inferiority of statistical significance of more popular car segments.

Lastly, it's important to keep in mind the complexity of automotive market. With plenty of factors contributing to the form of the vehicle and complex relationship between the cars and their owners, creating a model reflecting this relationship better could require more detailed data and previous insight which are scarce on this topic at the time of writing.

Future research should consider expanding the dataset to include more observation to capture more comprehensive trends and patterns. The form of representation of the vehicle should also be re-evaluated. Incorporating additional variables, such as economic factors or consumer demographics, could also enhance the analysis. Moreover, exploring more advanced machine learning techniques, including deep learning, could further improve predictive accuracy and provide deeper insights into the factors driving CM sales.

7. References

https://www.autoblog.com/article/platform-sharing-why-build-multiple-models-same-chassis/

https://www.expertmarketresearch.com/reports/automotive-floor-mats-market

Passenger car registrations in Europe 1990-2021, by manufacturer - ACEA - European Automobile Manufacturers' Association

Farell, M. (2023). *"Data and Intuition: Good Decisions Need Both"*. https://www.harvardbusiness.org/data-and-intuition-good-decisions-need-both/

Genuer et al. (2008). *"Random Forests: some methodological insights".* https://www.researchgate.net/publication/23420370_Random_Forests_some_methodological_insights

Kuhn, J., Johnson, K. (2013). "*Applied Predictive Modelling".* *http://appliedpredictivemodeling.com/*

Natekin, A., Alois, K. (2013). *"Gradient boosting machines, a tutorial".* https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021/full

Spiller, S., Belogolova, L. (2016). "*On Consumer Beliefs about Quality and Taste"*. Journal of Consumer Research, Volume 43, Issue 6, April 2017, Pages 970–991. https://doi.org/10.1093/jcr/ucw065

Szukits, Á., Moricz, P. (2022). "*Towards data-driven decision making: the role of analytical culture and centralization efforts"*. Review of Managerial Science. https://doi.org/10.1007/s11846-023-00694-1

Wittingham et al. (2006). *"Why do we still use stepwise modelling in ecology and behaviour?".* Journal of Animal Ecology. https://pubmed.ncbi.nlm.nih.gov/16922854/

Voss, J. (2005). *"Measuring Wikipedia"*. http://eprints.rclis.org/6207/

Zou, H., Hastie, T. (2003). *Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays".* https://www.researchgate.net/publication/228781252_Regression_shrinkage_and_selection_via_the_elastic_net_with_applications_to_microarrays