ESTIMATING PRICE-RESPONSE FUNCTIONS AND PREDICTING TIME-ON-MARKET FOR USED CARS IN UKRAINE

by

Vasyl Dyba

A thesis submitted in partial fulfillment of the requirements for the degree of

MA in Business and Financial Economics

Kyiv School of Economics

2025

Thesis Supervisor: _____ Professor Oleh Nivievskyi

Approved by _____
Head of the KSE Defense Committee, Professor [Type surname, name]

_____

_____

_____

Date _____

ACKNOWLEDGMENTS

The author wishes to express gratitude to his thesis advisor Professor Oleh Nivievskyi for his ongoing support and insightful questions during the research process. Without it, most of the research would remain on the surface, lacking academic depth.

Special appreciation is extended to the author's comrades-in-arms from the National Guard of Ukraine, whose patience and understanding make the completion of this study possible despite the challenges of military service.

The author expresses his deepest appreciation to his family and loved ones who provided him with continuous support and love throughout his master's studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFT Accelerated Failure Time

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

C-index Concordance Index

CV Cross-Validation

DOP Degree-of-Overpricing

HR Hazard Rathousandntegrated Brier Score

PH Proportional Hazards

PRF Price Response Function

RSF Random Survival Forest

TOM Time-on-Market

VIMP Variable Importance

CHAPTER 1. INTRODUCTION

The automotive industry is a multi-billion-dollar sector and a significant contributor to wealth in many economies. Ukraine is not an exception, with an estimated revenue of over $7 billion in vehicle sales in 2024. The majority of those sales are coming from the used cars segment, with 1.109 million transactions officially registered in 2024 (Automotive Market Research Institute, 2025). However, despite the enormous revenue and numbers of transactions, profitability remains a significant problem, with profit margins of about 1% (8% for new cars). This doesn't meet the strategic relevance of the used car business and thus is highlighted as one of the most essential management challenges for carmakers and car dealers (DAT, 2017). Amongst the factors contributing to this lack of profit are over-capacity manufacturing, an increase in level discounting, excessive supply, and high competition (Jerenz, 2008).

There are some unique challenges for business optimization in the used car market. Whereas huge marketing campaigns, together with a rich set of configuration options and individualization possibilities, keep the new car business profitable, no corresponding measures are available for the used car business. Therefore, the set of management controls that could boost revenue is restricted. The supply side of used cars is driven by the new car business, which is managed by retail trade-ins, repossessions, etc., so it's largely fixed, and thus price is normally the only steering mechanism available to increase margins (Du et al., 2009). Much research has examined the main variables in the formation of prices in the used car market; often using prices as informative cues to shed light on market structure and informational efficiency (Genesove, 1993; Emons and Sheldon, 2009) From a microeconomic perspective, price discrimination is a suitable strategy for extracting consumer surplus and increasing margins (Avi, 2018). To implement this strategy, sellers require an accurate estimate of consumers' willingness-to-pay or, put differently, the price-response function (PRF).

A common practice in economics is the use of the measure of the consumer's willingness to pay in the estimation of the demand (Skiera & Albers, 2008). However, in the construction of pricing algorithms, one of the factors that is commonly not taken into consideration is the intricate process of specifying the response of customers to a pricing strategy. Often, a functional form of a demand function is either known in advance or chosen based mostly on convenience of analysis. In contrast, in practice, the successful application of a pricing algorithm is primarily contingent upon data acuity as well as the selection of the appropriate functional form of a demand function. Even though survival analysis methods in time-to-event data analysis reference dating back to biostatistic analysis as well as econometrics (Kaplan & Meier, 1958; Cox, 1972), time-on-market analysis in the used-car trade is not yet well-represented in literature, including in the particular environment of the Ukraine.

A second-hand car seller is always caught in the dilemma of spending too much time in selling the car or earning no margin. Pricing the vehicle too high reduces the pool of interested buyers and increases the chances of an excessively long time-on-market. On the other hand, pricing too low may lead to a fast sale but will come at the expense of foregone profit that could have been made with an effective pricing strategy. In standard search theory, there is a positive relationship between asking price and time-on-market because of the effect of listing price on the arrival rates of buyers inspecting a vehicle. In this context, we are interested in the underlying mechanisms of the second-hand car market to provide managers with insights on how to improve pricing decisions and ultimately increase profits.

The main purpose of this research is (i) to find the most reliable and accurate statistical method to predict the TOM in the Ukrainian used car sector, (ii) to estimate price-response functions by modelling a used car's TOM (time-on-market), and (iii) to analyze the influence of the degree-of-overpricing on the probability of a sale.

In our study, we used a sample of used car listings published for May-August 2025, obtained from the largest online marketplace for used car ads in Ukraine – Auto.ria; The final sample contains data from over 90,000 unique vehicle listings. Our dataset includes all the general information of a vehicle and the listing-specific attributes: number of photos,

promotion level, and transaction types offered by the seller. With around 350,000 active used cars for sale and 3.5 million unique visits every month (Similarweb, 2025), Auto.ria.com is a rich data environment to apply statistical techniques for predictions. More specifically, this will involve testing the key assumption from literature in academia on survival analysis within the used car market scenario; thus, testing the hypothesis concerning the asking price of a durable good and its time on the market along with the probability of sale (Jerenz, 2008).

The study would benefit stakeholders in the used car market. These benefits may include cost efficiencies, optimized used car sales pricing strategies, as well as increased transparency for consumers. The thesis would address current critical gaps in the publication of automotive economics. Although there has been much research to price prediction and comparing the performance of machine learning models against linear models (Shymanskyi & Liaskovets, 2023; Kovpak & Orlov, 2019), a gap remains in directly modeling the probability of sale, known as a price response function. As such, contributions to practical knowledge and the generation of new knowledge will enhance the decision-making ability of stakeholders in Ukraine's automobile market and ultimately increase its efficiency.

The results are consistent with economic theory and reveal a clear hierarchy of factors influencing the time-on-market of a used vehicle. The analysis shows that the pricing strategy-in particular, the degree-of-overpricing-is unambiguously the most powerful determinant of sale duration. A secondary but highly influential tier of variables consists of variables that relate to listing quality and information availability, including the length of the description, the number of photos, and the presence of an independent tech report. Core vehicle attributes, such as age, mileage, and engine volume, were found to be significant but only third in their power of impact. Ceteris paribus, expected time on the market increases substantially with a higher asking price. For example, according to the estimates from the AFT model, a 10% increase in the DOP can dramatically lengthen the expected sale duration, underscoring the high price elasticity in the used car market. Moreover, listing quality is a very important factor; greater detail in the description and

more photos posted consistently reduce the time it takes to sell a vehicle, showing the role of information in decreasing buyer uncertainty and increasing transaction speed.

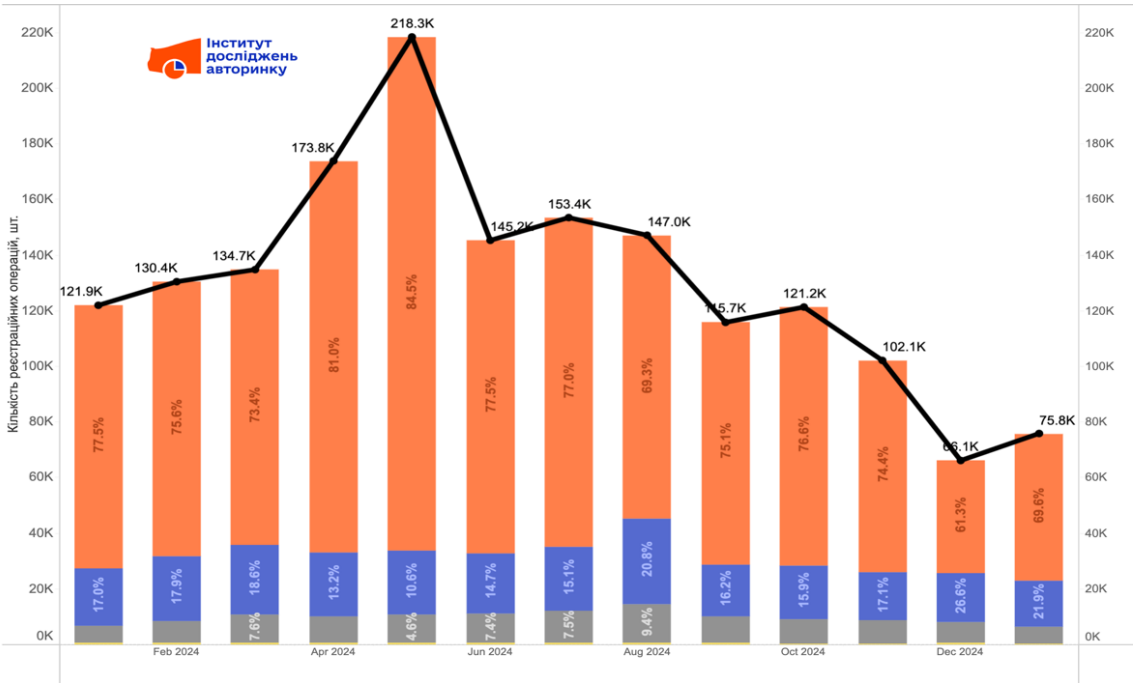
The rest of the paper is organized as follows. Chapter two gives an overview of the used car market in Ukraine and prior findings on the subject. The third chapter discusses empirical methodology, while chapter four presents the data. Chapter five presents and discusses the results. The final chapter summarizes the study, connecting it with real-world applications and suggesting areas for further research.

CHAPTER 2. INDUSTRY OVERVIEW AND RELATED STUDIES

By the definition of Ukraine's Automotive Market Research Institute the car market is divided into three broader categories: new cars (imported and produced domesticaly), used cars imported abroad, and used cars purchased in the domestic market. Last year, the Ministry of Internal Affairs' services centers registered on 15.9% more transactions of used car new owners than in 2023 (Automotive Market Research Institute, 2025). But there is another important point: not all operations reflected in the statistics were truly sales. The change in the rules for mobilizing citizens' transport, according to which the second and other (if more than one) cars can be confiscated for defense needs, caused an immediate reaction from car owners – they responded to this with a wave of mass re-registrations of cars in order to bring the ownership formula to the format of one person – one car. Therefore, the summer peak of "trading" is actually a surge in re-registrations of vehicles to loved ones (as well as donations, rejections – more on this in our other reviews), which has left its mark on the general statistics. IDA experts suggest that the total number of volumes could include approximately 150,000 "idle" re-registrations of cars to family members or relatives; the real number of resales may be approximately the same as in 2023, near 956,000 (Automotive Market Research Institute, 2025).

Used cars sold in the domestic market represents the largest transaction category in the car market, accounting for 78.9% of all market transactions within all car categories in 2024 (see Figure 1). In this figure, the breakdown of monthly transactions is shown, with the orange representing these domestic used car sales. The blue segment corresponds to the import of used vehicles. The sale of new cars is divided into two smaller categories: imported new cars, shown in gray, and domestically produced new cars, which are represented by the very thin, almost invisible yellow segment at the base of each bar.

Figure 1. Car Market Transactions Registrations in 2024

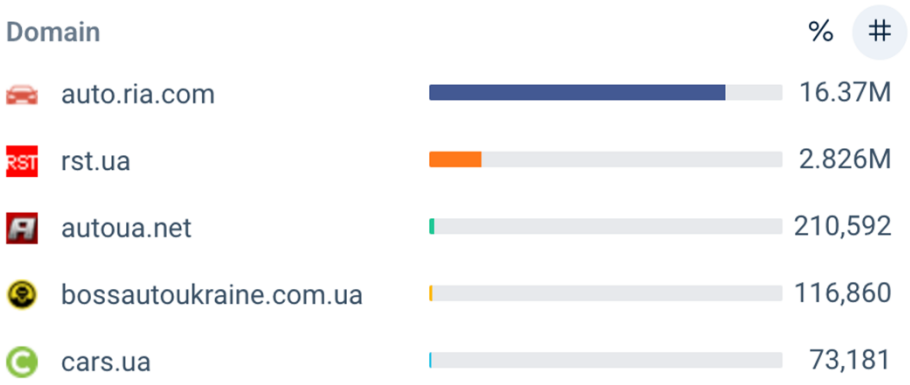


2.1 Online marketplaces for used cars in Ukraine

The trade of used cars is increasingly performed over online platforms where individuals can buy vehicles from Ukraine and unregistered vehicles from the United States and Europe. Online platforms are expansive marketplaces where sellers can advertise their vehicles for others to view, compare, and consider, as well as review the conditions of the cars. The increase in the trade of used cars over online platforms has taken the buying and selling of used cars to another level, with convenience and an extended reach for buyers and sellers. The main online platforms used more and more frequently to trade used cars in Ukraine are: Auto.ria.com, OLX Auto, RST.ua, autoua.net, bossautoukraine.com.ua, cars.ua. For most websites, it is possible to see an estimate of the monthly visits as a proxy for the number of sales taking place on each (see Figure A.2). More traffic to the website yields higher engagement among users and hence, more transactions. However, there is an issue with OLX Auto since it does not have its own website; it exists as part of a giant online classified listing platform where people can come together to buy, sell, or exchange

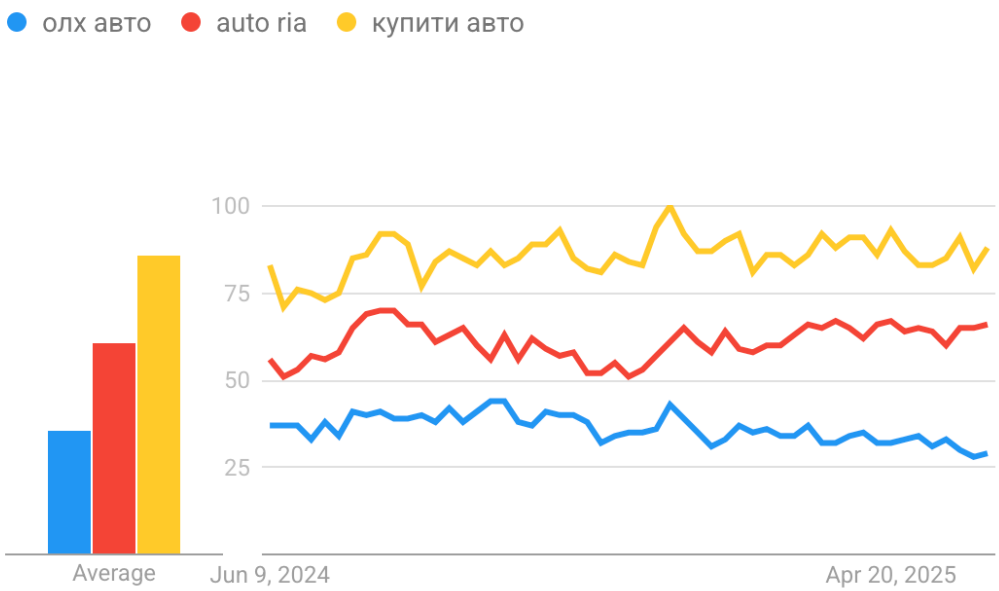
different items or services. The absence of a website complicates the use of traditional web analytics platforms in appearances as it relates to meaningful traffic. However, I could use another metric to determine the level of engagement - Google Search term index.

Figure 2. Websites Total Visits in May 2025



The idea is simple - we used search traffic for keywords, like the name of the classified listing platform, as a proxy to actual traffic on platforms. This calls upon the idea that customers search specific platforms when considering buying or selling. For this purpose, I used the most popular keywords related to each of the platforms - "auto ria" and "олх авто". I also added that general keyword - "купити авто" as a baseline. This word will provide a baseline understanding of the interest in the online car-buying market. Google's Search term indexing is a number from 0 - 100, which represents the amount of search interest relative to the highest search interest in the chart for that region and timeframe of data. This standardization provides ease of use when comparing search interest to terms. As we can see (Figure 3), Auto.ria.com has been the dominant resource in the used cars marketplace, with just under two times higher interest in search traffic than for OLX Auto for the prior 12 months. This illustrates that Auto.ria.com held a greater position within the online marketplace regarding interest or online engagement through used cars listed on each site.

Figure 3. Google Search Index for Last 12 Months



2.2 Main trends in 2024

Which auto brands received more attention and choice from all Ukrainians regarding the secondary market? According to the data from the preceding periods, Volkswagen, with a quota was 12.5%, was twice the rate of those of the brands that received attention after it (see Appendix A). This affirms the brand's visibility and quality within the Ukrainian market. The Renault brand followed with barely a quantitative lead and was ranked second on the list. The presence of various VAZs still does not lose its relevance and significance due to low pricing and the quantity being enough that the descendants of Fiat can still cover a considerable share of the "secondary" market. The next car makers that European motorists prefer are Skoda, Toyota, and BMW.

This prioritization reflects the general preference for European cars in the secondary Ukrainian market. Brands' preferences rely on quality, finishes, and availability of spare parts (Automotive Market Research Institute, 2025). In the list of models, the leader was the Volkswagen Passat. Given that every month of the last calendar year, the Passat held the lead above all the other participants in the rating. The popularity of the Volkswagen

Passat has indirectly confirmed its reputation with consumers in the country for years. The demand and popularity of two more models have been retained: the Skoda Octavia and the Volkswagen Golf. Thus, three successful models affirm the leadership of the Volkswagen and Skoda brands within the market. In 2024, the average age of the cars was 16.3 years in the secondary market, which means there are still cars considered older models, with low prices at the sales point. This might denote customer preference and factor pricing strategies.

How about premium segment cars? Thus, during the 12 months of 2024, a total of 3,372 Porsches, 204 Maseratis, 147 Bentleys, 33 Rolls-Royces, 17 Aston Martins, 12 Ferraris, 11 Lamborghinis, 6 Maybachs, 3 Lotuses, and 3 McLarens were resold on the domestic market (Automotive Market Research Institute, 2025). This highlights the existence of luxury and premium cars within the liking of the used secondary market, and is therefore consumed by a niche of buyers. The most respectable cars were three Mercedes-Benz - one model 130 and two 170V. Referring to the fragility of all their documents, the year of manufacture was also stated as 1936.

2.3 Previous Studies

From the focus of the thesis, it appears that related works come essentially from two main areas: the second-hand car market and survival analysis. The origins of survival analysis can be traced back to the early work on mortality by John Graunt, who introduced the concept of 'life tables' in his book, Natural and Political Observations on the Bills of Mortality (Graunt 1662; Sutherland 1963). The contemporary are of survival analysis began early in the 20th century with studies on industrial device durability. During World War II, the reliability of military equipment became a primary concern, and industrial reliability engineers began to use the term 'lifetime' analysis. In the post-war period, aspects of reliability were applied to studying survival time for cancer patients, and cancer researchers coined the term 'survival analysis. Around this time, two articles were published that laid the groundwork for modern survival analysis. Kaplan and Meier (1958) formalized the

product-limit estimator, and Cox (1972) introduced the proportional hazards model. In the last fifty years, survival analysis has become a leading method for analyzing data in fields such as the survival times of patients in clinical trials (biomedical sciences), the lifetime of a machine component (industrial engineering), and the length of time between unemployment or the length of a strike (economics).

The used car market received significant attention from economists starting with Akerlof's work on asymmetric information (1970). Market prices become a key leading indicator of either the informational efficiency of the market in relevant studies (Levin, 2001), information asymmetry (Belleflamme and Peitz, 2014), and discrimination (Ayres and Siegelman, 1995), among others. The influence of digital channels, especially online auctions, has been well-documented to shape the ecology of this particular market (Bapna et al., 2008). Numerous studies have also evaluated the role of digital innovation in the automotive market (Chen et al., 2013).

The large volume of the second-hand car market also signifies its relevance from a managerial perspective. Olivares and Cachon (2009) provide useful evidence concerning the competitive disadvantage of large inventories, potentially due to less-than-optimal pricing. According to their work, large inventories have a competitive disadvantage, which they attribute to suboptimal pricing. Due to close connections with the new car business, namely, via trade-ins, lease returns, and repossessions from car rental businesses (Desai and Purohit, 1998), the importance of pricing issues is increased, as also suggested by Ratchford and Srinivisan (1993). To the best of our knowledge, there are only two studies that examine survival analysis in terms of price response modelling of the used car market.

Jerenz (2008) developed a comprehensive revenue management system to support used cars price optimization. The system consists of three components. The first is a forecasting model used to estimate residual values. The second is a survival model that constructs a PRF with the residual estimated values as an input. The third is a dynamic program that provides the optimal pricing policy. The study developed by Jerenz (2008) is particularly relevant to this paper because it serves as the first and only example of applying survival analysis to optimize used-car pricing. Born et al. (2018) extend Jerenz's framework

and carefully evaluate survival analysis as a tool for price optimization in the used-car market. Their study filled gaps between classical parametric survival models and modern data-driven techniques. They compared the performance of Cox PH model and data-driven methods (survival tree, random survival forest, conditional inference tree and conditional inference forest) and came to the conclusion that: "We find classical survival analysis methods inappropriate for the real-world data employed in the study, due to strict assumptions of linearity and proportionality. We introduce for the first time data-driven survival methods to the context of the used car market and show superior predictive performance of random survival forest and conditional inference tree." While we don't disagree with their conclusions, we believe that the comparison they made was biased. First of all, unlike Jerenz (2008), who used a highly homogeneous sample of data for performing survival analysis, specifically for one model from one generation with the same engine type and capacity, Born et al. (2018) used a sample of different product lines from a premium car maker, which implied different car types, models, and engines. They essentially neglected one of the most crucial steps before any parametric modeling is conducted. As Jerenz(2008) stated, in the context of the used car sector, it is of interest to identify model types that exhibit similar characteristics in terms of survival and hazard rates. Then, types with similar survival functions can be analyzed as one sample, providing higher explanatory power. Although Du et al. (2009) and Jerenz (2008) estimated vehicle prices and, subsequently, residuals using ordinary least-squares regression, follow-up studies have shown that data-driven forecasting approaches-most notably, neural networks or regression tree ensembles-offer superior estimates of the price (Lessmann and Voss, 2017). For example, Amik et al. (2021) utilized several models, such as a linear regression model, LASSO regression model, decision tree, random forest, and extreme gradient boosting, and chose the best model, XGBoost, outperforming all others in their comparative performance (Amik et al. 2021). It is thus capable of correctly predicting prices over 91% of the time. The analysis showed that the decision tree was overfitting, the random forest provided better replication with gains in generalization, while XGBoost performed

marginally better by incorporating boosting. This suggests the potential of advanced data-driven models for price optimization frameworks.

Concerning the used car market in Ukraine, the topic of modelling PRF remains significantly under-researched in academic literature. To the best of our knowledge, only models predicting the price of the used car and attributes influencing it have been investigated. Most studies have been conducted by graduates of economics at KSE. Hrechanyk (2019) researched the Ukrainian electric vehicle (EV) market to assess its state of readiness for production. Industry data and cost modeling were used, finding that Ukraine had one of the fastest-growing EV markets in the world, and it is almost exclusively populated by used imports, with the Nissan Leaf at the pinnacle of the hierarchy of import vehicles. Tax-free import regimes and updated legislation specifically related to EVs improved the environment, but were insufficient to attract large-scale producers. Summing up, Hrechanyk points out that Ukraine is not yet ready to start full-scale production of EVs but recommends investment in infrastructure and financial incentives that will support future growth of the industry.

Matviichuk (2021) looked at the prices of used cars in Ukraine, considering more than 100,000 transactions within a period covering 2018-2021.Matviichuk (2021) looked at the prices of used cars in Ukraine, making use of more than 100,000 transactions during the period of 2018-2021. Brand was the most influential variable that influenced pricing, while some features had significant effects. The price responses of fuel type in Ukraine were similar to those in France, and showed that diesel cars were about 15% more expensive than their petrol counterparts. In the meantime, there were also price premiums for hybrids and electric cars attached to them, depending on the type-hybrids were between 12-27% more, and electric cars were, on average, about 11% more. These results only further cement the idea that technical and fuel economy specifications are the two most prominent driving forces behind prices. From these findings, it is suggested that hybrid technology may be a pre-specified feature favored in the Ukrainian market due to considerations of either fuel cost or popularity of the cars being imported.

Proshchyna (2020) examined the factors influencing the diffusion of electric vehicles in Ukraine, such as consumer preferences, economic conditions, and policy incentives. By applying both survey-based data and secondary statistics, the study identified motives for switching to EVs, including saving long-term operating costs, fuel independence, and environmental awareness, while the most important barriers include limited charging infrastructure, high purchase prices, and battery replacement issues. The study has also identified that a young and relatively well-off consumer will consider electric vehicle options, particularly in a city with better infrastructure. More importantly, the importance of government incentives and an exemption from import tax has been given much prominence, although these incentives remain somewhat insufficient for mass adoption. Generally, Proshchyna indicated that electric vehicle markets in Ukraine are still at an early stage of development and may need additional policy support for faster transition toward greener mobility.

Among all Ukrainian research on modelling and pricing of used vehicle markets, there were two papers published outside the Kyiv School of Economics: Kovpak and Orlov (2019) and Shymanskyi and Liaskovets (2023). In turn, first compared regression and machine learning models for prediction of used car prices based on more than 200,000 listings from Auto.ria.ua. Authors tried a number of algorithms like linear and polynomial regression, k-nearest neighbors, decision trees, random forest, gradient boosting, and neural networks. They found that the best forecasts in terms of an average relative forecast error of 14.3% were obtained for nonlinear ensemble models which include: (1) random forest, (2) neural media, and (3) gradient boosting. Shymanskyi and Liaskovets (2023) of Lviv Polytechnic National University proposed a cascade machine learning model for forecasting vehicle prices and time-to-sales together. For over 5.3 million records from Auto.ria.ua, the best total accuracy of vehicle price prediction ($R^2 = 0.94$) was demonstrated by XGBoost among the ten tested algorithms. It was found that the deviations of the predicted market prices are helpful for predicting the duration of sales.

CHAPTER 3. METHODOLOGY

Price is the number of monetary units the customer must sacrifice to receive one unit of product or service. In the context of durable goods, this is one element of the marketing mix that plays a defining role in the purchase decision. In addition, price is distinguished from other marketing instruments by the force and speed it has on sales, as well as the brevity of time it takes to change it. This is why knowledge about customers' reservation prices is the key information used to develop rational and optimal pricing strategies. Having information about customers' reservation prices (or their willingness to pay), one can derive the price response functions necessary for developing optimal pricing strategies (Jerenz, 2008).

A used car retailer is subjected to the trade-off between spending too much time selling the vehicle versus receiving no profits. Setting the asking price too high reduces the number of possible buyers, which in turn raises the prospect that the vehicle will be on the market for an excessively long period. On the other hand, by setting the asking price too low, a sale may happen quickly, but the potential profit with a better pricing strategy will be sacrificed. Standard search theory assumes a direct relationship between the asking price and time-on-market because the listing price may influence the rate at which buyers inspect a vehicle.

Survival analysis is a statistical approach that can be used to model the time until an event of interest happens. Here, the event is a sale of a car, and the target is to model time-on-market of a car. This analysis allows us to estimate the survival functions, compare survival functions between groups of data, as well as quantify the relationship of various variables on survival time (Kleinbaum & Klein, 2006). This will help us understand which factors lead to an accelerated sale. The primary benefit of survival analysis over logistic regression is that survival analysis allows us to model censored data in a correct way. For this analysis, a record is considered as a right-censored record if data is collected until the last date of sale if a car is not sold yet. Left-censored data (car sales that happened before

this analysis went into consideration) is not there in this data set as the Auto.ria API does not store any data prior to a sale.

Before applying parametric survival models, it is essential to first verify that the data being analyzed consists of homogeneous subgroups that display common characteristics in terms of their survival behaviour. In an empirical setting, such as the used car market, the entire sample is extremely heterogeneous, and different brands, types of bodies, and price segments exhibit distinct selling behaviours and different hazard patterns. The aggregation of these observations into a single model could not only obscure important facts but also violate a key assumption of the Cox proportional hazards model – the hazard ratio between any two individuals must remain constant over time. To reduce these problems, the data should first be grouped or clustered based on similarity in the estimated survival functions. The survival curves in a non-parametric estimator (e.g., the Kaplan-Meier method) can be compared across different types of vehicles or market segments, for example, using tests such as the log-rank statistic or the Wilcoxon statistic. The subsamples, whose survival curves do not differ significantly, can then be grouped, thereby creating internally consistent clusters that demonstrate common selling behaviors. The step of grouping data, prior to modelling, will enhance the statistical power, thereby ensuring greater robustness of models in the parametric or semi-parametric field and enhancing the interpretability of models forecasting behaviour by adhering to the economic intuition e.g. "mass market saloons" vs. "premium SUVs" etc., implying that they will behave similarly in terms of their common selling pattern, which is empirically homogenous.

Let's start with definitions. The survival function S(t) is the probability of an individual surviving past time t, where T is a continuous random variable. It is a non-increasing function with the value of 1 at the origin and 0 at infinity (Born, Kovachka, Lessmann and Seow, 2018):

$$S(t) = P(T > t) = \int_t^{\infty} f(u)\ du = 1 - F(t). \qquad (3.1)$$

Another representation of survival time distribution is the hazard function (also known as the conditional failure rate and the hazard rate), which is the instantaneous risk of failure at time t, given individual survived to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)} \tag{3.2}$$

We can analyze survival data without any parametric assumption about the form of the underlying distribution. The standard non-parametric estimator of the survival function from observed survival times was proposed by Kaplan and Meier (1958). One of the most important strengths of the Kaplan-Meier (KM) method is that the estimate can be stretched out to include censored data. The estimator considers an increasing ordered set of event times t_i, the number of events d_i at time t_i , and the total number of survivors n_i:

$$\hat{S}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} \tag{3.3}$$

$$\widehat{H}(t) = -\ln \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} \tag{3.4}$$

It should be added that, for the hazard function, Nelson-Aalen would be a more common estimator (Nelson, 1969, 2000; Aalen, 1978). But, as Jarenz (2008) emphasized, the difference between the two estimates can hardly be observed, so we'd keep to KM in our analysis.

Approaches to modelling covariate effects on our outcome of interest can be essentially divided into two classes: proportional-hazards models which include both semi-parametric specifications such as the Cox model and fully parametric variants, for example exponential or Weibull models; and accelerated failure time (AFT) models, also known as location-scale models for log-transformed survival time. The Cox model is, in essence, a

multiple linear regression of the logarithm of the hazard on the covariates X; the intercept term varies with time and constitutes the baseline. The covariates are multiplied on the hazard with regard to the key assumption of the proportional hazards model (Bradburn, Clark, Love, and Altman 2003).

$$h(t|\boldsymbol{X}_i) = h_0(t) * exp\left(\sum (j=1)^p \beta_j x_{(ij)}\right) \tag{3.5}$$

Where:

- $h(t|X_i)$ is the hazard function for car *i* at time *t*, and represents the instantaneous probability of the car being sold at time *t*, given that it has not been sold yet.
- $h_0(t)$ is the baseline hazard, representing the hazard function when all predictor variables are set equal to zero. The Cox model is semi-parametric because this baseline hazard is not assumed to follow any particular probability distribution.
- X_i represents the vector of the *p* predictor variables for car *i*. This includes quantitative attributes such as raceInt and photoCount, among others, plus qualitative attributes using dummy variables, for example, damage and gearboxName.
- β_j is the regression coefficient for the attribute *j*. The exponentiated coefficient, $exp(\sum \beta_j)$, is the Hazard Ratio (HR). Its magnitude corresponds to the multiplicative effect on the hazard for a one-unit change in the attribute x_j. An HR greater than 1 indicates that an increase in the variable speeds up the sale, while an HR less than 1 indicates it slows down the sale.

The basic issues that could affect the Cox Proportional Hazards model are problems common for complicated observational data, such as that from the car market. First of all, and most importantly, there is a possible violation of the PH assumption. That is a core assumption: the effect of a given predictor (its Hazard Ratio) is constant over time. However, some attributes may have a strong initial impact that fades later-on, for example, a "new listing" effect-or a delayed impact. If the assumption holds, then the estimated

regression coefficients will represent the misleading "average" effect over time, and the hypothesis testing can be incorrect. In order to tackle this problem, the PH assumption will be formally tested using a test based on Schoenfeld residuals (Schoenfeld, 1982). If violations are found, considerations such as stratification or inclusion of time-interaction terms will be made.

The second problem-also common in regression analysis-is multicollinearity, or high correlation between two or more predictors. The VIF will be used to diagnose the multicollinearity among the continuous variables. One way of dealing with multicollinearity is to exclude variables from the model; this approach may, however, introduce omitted variable bias. A more robust approach, used in this research, is to build models of different specifications in order to ensure the stability of the core findings.

Its main competitor is the accelerated failure time (AFT) model, which relates covariates to the logarithm of the failure time in a linear fashion. Under the assumption of time-constant covariates, the mathematical form of the AFT model is obtained by an ordinary regression model where the survival time, T, is transformed by the natural logarithm as follows:

$$Y = \ln T = -x'\beta + \sigma W, \quad W \overset{\text{iid}}{\sim} S_0(\cdot), \tag{3.6}$$

with β as the vector of regression coefficients, σ as a scale parameter, $S_0(\cdot)$ as a known baseline survivor function, and W as an error term with a suitable distribution independent of x. AFT models are based on the concept that the covariates accelerate or decelerate the expected survival time. These usually come in parametric forms; one needs to assume a particular form of the underlying distribution of time-to-event data, such as Weibull or log-normal. However, this is also largely limited in its generality because the selected distribution might not adequately model the data; furthermore, the impact of every predictor in such a model is assumed to be constant and multiplicative on the time scale, which contrasts with the Cox model in which the effect is on the hazard rate.

Estimating a price response function from survival data requires the identification of a significant relationship between price and time-on-market. The raw asking price is an unsuitable predictor because used cars are heterogeneous-the value of any given price is meaningful only in relation to the specific attributes of the vehicle. We address this challenge by defining a standardized metric, the 'degree-of-overpricing' (DOP),

$$\text{DOP} = \frac{\text{Price}}{\exp(\text{Estimated Value})} \tag{3.7}$$

which represents a vehicle's asking price as a ratio of its estimated market value. Once included as a significant covariate within the survival model, it is possible to isolate the independent effect of pricing strategy on the probability of sale and thus derive a price response function, which quantifies the relationship.

To apply the DOP metric, a robust estimate of the underlying market value for a vehicle was required. We used a hedonic price model which presumes the value of a good is a function of its constituent characteristics. In estimating the value for a used vehicle, several alternative functional relationships can be assumed with the linear and the semi-logarithmic form as two basic models whereas the translog functional form and the Box-Cox form represent the foundation for more complex models. Within the context of this study, market value for a used vehicle is estimated assuming the semi-logarithmic form, given a large set of vehicle attributes including age, raceInt, markName, modelName, engineVolume_l, fuelType, gearboxName, and damage. The exponentiated prediction of this model for any given car constitutes its estimated 'Hedonic price,' which forms the denominator in the final DOP calculation used in our survival analysis.

The dependent variable for modelling the TOM is defined by two components: the duration in days from the date the ad was listed to the date it was marked as sold and a event indicator (a dummy variable) which is equal to 1 if the car was sold during the observation period and 0 if the observation was right-censored (i.e., the ad expired or was still active at the date of data collection). This framework is particularly appropriate as it

correctly handles the censored nature of the data, providing unbiased estimates of the influence of each car and listing attribute on the speed of sale.

.

CHAPTER 4. DATA

We extracted a dataset from Auto.ria.com through its private API, which, compared to web scraping, gives better data quality. Each observation represents a used car listing (a vehicle offered for sale). If a car is listed for sale multiple times, only the latest listing information is included in the final dataset. It includes data available on the Auto.ria.com platform during the extraction stage, which took place from June 9th to August 8th. The raw data consisted of 81 variables and 100,750 rows. It has identification, numerical, categorical variables, and dates. Each entry is associated with a unique car and its last published date. We described each variable in the Appendix (Table B.1).

4.1 Data preparation

The data extraction and basic data preparation consisted of building a multi-step pipeline that extracted relevant variables from the Auto.ria.com API's layered JSON object. The following steps were performed to prepare the data for furter analysis:

1. A Python script was created to collect all listing IDs based on particular criteria, gather information from those IDs one at a time, and then compile everything into a CSV for final storage.

2. Another Python script was produced for parsing all of the nested JSON objects and realigning the dataset in a flat structure.

3. An R script for cleaning and grouping the data for subsequent stages of feature engineering and exploratory analysis. In this stage, all duplicated vehicles, invalid rows, and irrelevant columns were eliminated. Additionally, cars with a year of production before 1996 and after 2023 have been deleted because their price dynamics differ from those of "usual" used cars. After the data preparation pipeline, the dataset contained 45 variables and 86,447 rows.

4.2 Feature engineering

We created seven variables from the raw data: time on market, degree of overpricing, quantile, age, length of description, whether VIN is shown, and whether the car brand is considered "premium". The first four variables are based on Jerenz (2008). We propose the last three mentioned as new variables to capture marketing effort, trust in specific listings, and brands considered high-value. As in subsequent studies, we define the target variable as time on market (TOM). TOM is the difference in days between the first and last dates of online presence for the same car. Degree of overpricing (DOP) is the proportion between the price reported in the data and the market value estimates for the specific car. Jerenz (2008) and Born et al. (2018) employed log-linear regression to estimate car market value:

$$\ln V = x'\beta, \tag{4.1}$$

Because the hedonic price represents the intrinsic value of a car (Lessmann and Voss, 2017). However, as mentioned in the literature review, the last studies suggest that XGBoost outperforms other methods in terms of price prediction. The quantile represents the percentile of car prices within the fuel type category. Age represents the age of a car in years. The novel variable description length is the text length for each listing and can be used as a proxy for seller marketing efforts. A boolean variable, is_valid_VIN, was created to represent whether a listing contains a proper Vehicle Identification Number and buyers could see it, which can lend more credibility to the listing. Additionally, the is_premium_brand dummy variable was created, indicating whether cars of a brand are considered a luxury. List of premium car brands is following: "Maserati", "Cadillac", "Jaguar", "Porsche", "Infiniti", "Land Rover", "Lexus", "Tesla", "Mercedes-Benz", "BMW", "Audi".

Here, it would be necessary to mention the shortcomings of our data. In this study, we collected a sample of cross-sectional market data, capturing the sale status of the last listing of a specific vehicle. This captures only price variations and corresponding sale status

for different cars of the same model, but not across a single car. This method is outlined in Jerenz (2008): "One approach would be to extend the method of market data by incorporating the quote history of a product. When the price offer history is combined with sales success data and price adjustment information, it is possible to estimate the probability of a sale being likely. Thus, this approach is a specific form of estimating the price response function and in the following section will be named time duration market data according to the application within time duration models." This means that if the price of a specific vehicle changes, the previous observation is censored, and a new observation begins. In this way, it is possible to incorporate the price variations of specific vehicles into the model, thereby significantly enriching the dataset. However, even without time duration it was still possible to capture price response dynamics.

In addition to technical shortcomings, it is essential to discuss potential value discrepancies. Unfortunately, Auto.ria wasn't open to discuss their data, so some questions remained unanswered relative to listing attributes:

- If the price of a listing changes while it is active, does the added date column update or the update date column?
- When can the listing be marked as sold? When does the buyer confirm the purchase? Or, when he deletes the listing and, as an option, chooses to have it sold on Auto.ria?
- When exactly did the update date column change?

Without this information, it is impossible to address value discrepancies correctly. Therefore, part of the analysis would involve validating the results against previous studies. To correctly identify invalid examples, we deliberately defined our main variable of interest TOM, and filtered all negative values: If the listing was sold, it was equal the difference between soldDate and addDate; If the listing was not sold and collectDate was before expireDate – the difference between collectDate and addDate; If the listing was not sold and collectDate was after expireDate – the difference between expireDate and addDate;

Also, the related fuelType variable was grouped into larger categories. For example, Hybrid (HEV), Hybrid(MHEV), Hybrid(PHEV) have become Hybrid. This would reduce

the dataset's dimensionality after encoding and help eliminate categories with a small number of examples.

A final and critical step in data preparation was filtering outliers and missing values. Based on data exploration, outliers were excluded from the following important numeric predictors: USD, raceInt, engineVolume_l, and age. Here, filtering avoids extreme and possibly erroneous values from having a disproportionate influence on the regression models, resulting in more stable and trustworthy results. As we discussed earlier, the heterogeneity of data for survival analysis modeling can significantly influence the model. So, we also excluded cars that were underrepresented: models(Audi, BMW, etc) with less than 100 examples and categories(sedans, minivans, etc) with less 200 examples.

4.3 Exploratory analysis

The final dataset analyzed consists of 42,763 distinct car listings, of which 11,546 (27%) had been observed sold by the date when collection happened. The remaining 73% of the listings were right-censored, in that they were either still live, or had expired unsold by the date of collection. Table 2 displays a summary of the descriptive statistics of the numerical and nominal variables used in the analysis. The main dependent variable TOM (timeToSale_d) represents the number of days that elapsed after the initial listing date until the car was either sold, or right censored. The mean time-on-market for the full sample is 12.7 days, median 12 days, and the observation interval for any single listing is limited to about 32 days.

All of the predictor variables are considered time-invariable, determined at the time of the observation. A variable of great interest is the `degree-of-overpricing' (DOP_hedonic), which has a mean value of 1.01 and a standard deviation of 0.50. This suggests that the market is not very efficient, in that most asking prices are clustered in a wide band about ± 50% of their estimated market value. The dataset also has a rich listing of quality factors, which are a prime focus of this study; as, for instance, the average car is

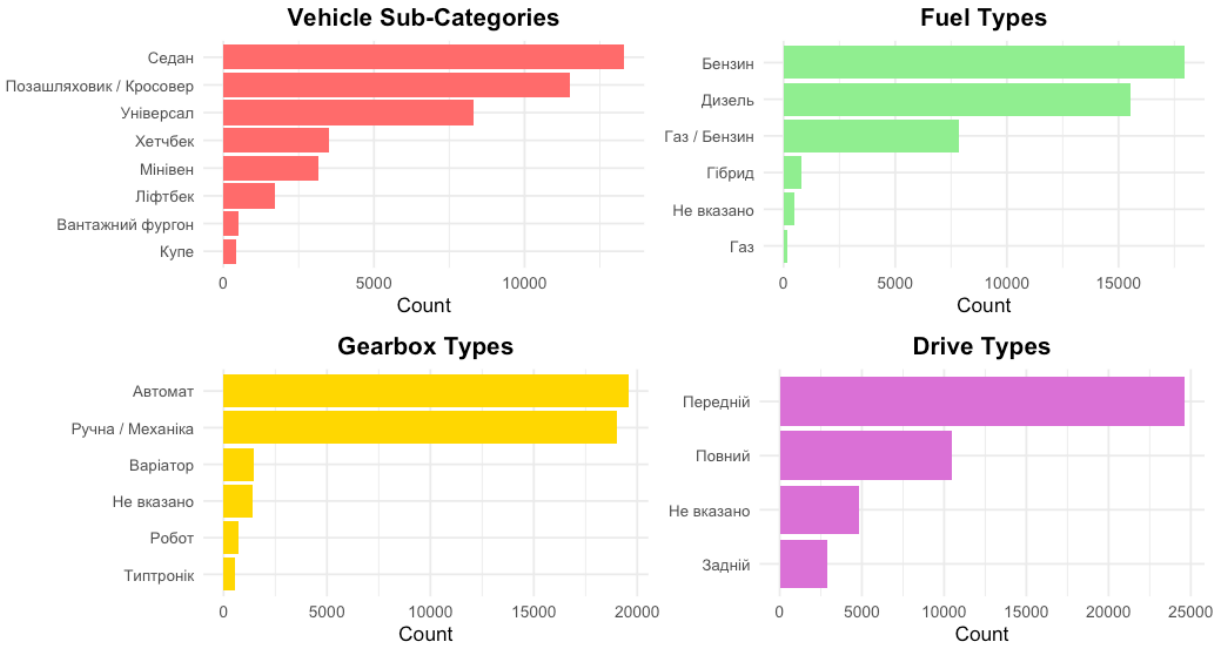
14.4 years of age and presented with about 22 illustrations and a description containing 345 characters.

Table 1. Descriptive statistics of numeric and boolean variables

	count	mean	std	min	25%	50%	75%	max	range
USD	42763.00	11484.13	10267.89	239.00	5300.00	8700.00	14500.00	250000.00	249761.00
raceInt	42763.00	219.12	95.86	1.00	155.00	218.00	275.00	999.00	998.00
firstTime	42763.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00	1.00
dealer_IsReliable	42763.00	0.05	0.21	0.00	0.00	0.00	0.00	1.00	1.00
dealerVerified	42763.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00	1.00
haveInfotechReport	42763.00	0.91	0.29	0.00	1.00	1.00	1.00	1.00	1.00
damage	42763.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00	1.00
level	42763.00	0.35	2.12	0.00	0.00	0.00	0.00	60.00	60.00
period	42763.00	0.37	1.92	0.00	0.00	0.00	0.00	31.00	31.00
photoCount	42763.00	21.66	17.86	0.00	11.00	17.00	27.00	242.00	242.00
withVideo	42763.00	0.03	0.17	0.00	0.00	0.00	0.00	1.00	1.00
exchangePossible	42763.00	0.24	0.43	0.00	0.00	0.00	0.00	1.00	1.00
auctionPossible	42763.00	0.63	0.48	0.00	0.00	1.00	1.00	1.00	1.00
onModeration	42763.00	0.29	0.45	0.00	0.00	0.00	1.00	1.00	1.00
is_valid_VIN	42763.00	0.99	0.10	0.00	1.00	1.00	1.00	1.00	1.00
age	42763.00	14.35	5.75	2.00	10.00	14.00	18.00	29.00	27.00
descriptionLen	42763.00	344.97	349.80	0.00	103.00	236.00	473.00	2015.00	2015.00
engineVolume_l	42763.00	2.03	0.61	0.12	1.60	1.98	2.23	6.60	6.48
timeToSale_d	42763.00	12.73	8.20	1.00	6.00	12.00	19.00	32.00	31.00
sale_status	42763.00	0.27	0.44	0.00	0.00	0.00	1.00	1.00	1.00
log_price	42763.00	9.06	0.77	5.48	8.58	9.07	9.58	12.43	6.95
is_premium_brand	42763.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00	1.00
price_quantile	42763.00	0.51	0.27	0.00	0.28	0.51	0.73	1.00	1.00
USD_hat	42763.00	11150.15	8510.03	275.02	5332.42	8788.39	14716.25	90427.05	90152.03
DOP_hedonic	42763.00	1.02	0.50	0.12	0.90	1.01	1.12	94.28	94.16
USD_boost	42763.00	11317.22	9910.89	276.71	5267.53	8570.99	14389.99	237422.80	237146.10
DOP_boost	42763.00	1.01	0.09	0.11	0.99	1.00	1.03	4.54	4.44

Figure 2 provides a summary of the distributions for the main categorical variables in the data set. Analysis of the vehicle sub-categories discloses that the Ukrainian used car market represented in this sample is characterised by three major categories: Sedans,

Figure 4. Distribution of vehicles by categories

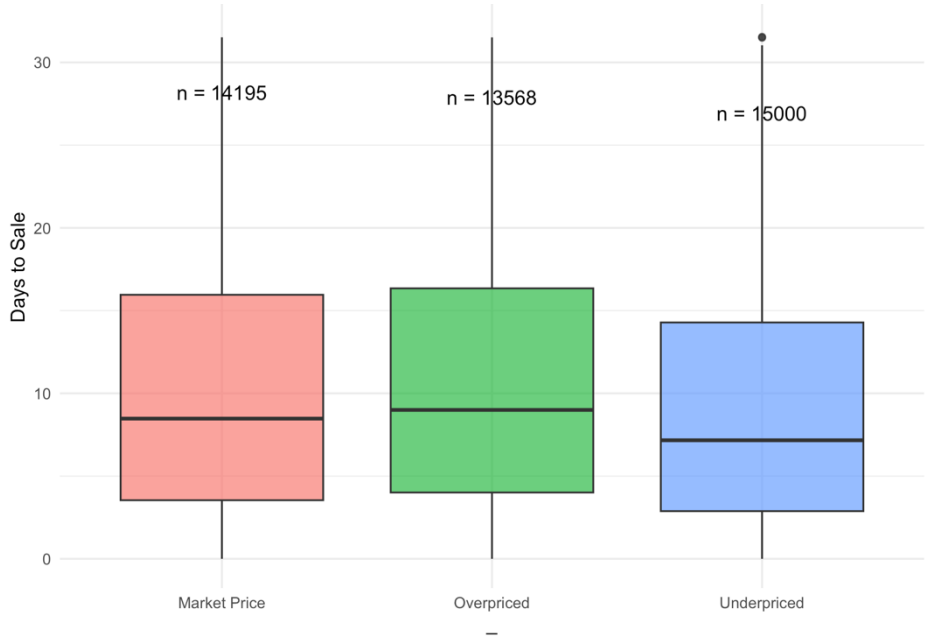


Crossover / SUVs and Wagons which together make up a large part of the all of the listings. Minivans and hatchbacks represent considerable, but smaller, parts of the market. In terms of technical specifications, the fuel types figure shows a noticeable bias towards traditional powertrains, in that petroleum and gas oil are the two most popular. Many cars also run on a dual gas/gasoline basis. The distribution in the gearbox types is much more equal, with a reasonable split between automatic and manual gearboxes, which together account for nearly the whole dataset. The drive types are mostly those where the drive is to the front, with all-wheel drive being the second most common.

In order to investigate one of the most important variable for estimating PRF, degree-of-overpricing, we build Box plots of TOM for three categories: “Underpriced”, “Market Price”, and “Overpriced”. These categories are an approximate representation of

the market estimate. We calculated the average for each model within the same year of production and then compared it to the vehicle's price. If it was different than average by more than $500, we gave it the class "Overpriced". A similar procedure was used for the "Underpriced" category. Even with such a simple rule of thumb, we already see differences in time-on-market between different categories. This suggests a strong cue for the significant influence of the degree of overpricing of TOM.

Figure 5. Box-plots of time on market by price category



CHAPTER 5. RESULTS

This chapter presents the results of the multi-stage modeling analysis, establishing a robust methodology for estimating the Price Response Function (PRF), which serves as the basis for profit optimization (Jerenz, 2008). This methodology aims to optimize model and feature selection for any given dataset in the context of survival analysis for the used car market. First, we present the results of our preliminary analysis, which was conducted to find the most promising model from classical survival methods and compare it with the best data-driven survival method for the used cars market: Random Survival Forest (Born at el. 2018). This process identified an extended Weibull AFT model with regression splines as the best fit for our dataset. Second, we compare the best models from two classes using 10-fold cross-validation on the sub-sample, which we used for feature selection, and the full dataset. And finally, we use the best-performing model to estimate the individual price response function and analyze the influence of the degree-of-overpricing on the probability of sale.

In the first stage, we determined which of the classical models best described our data. The methodology we applied to identify the classical survival model that best describes our data and is reliable in its assumptions was largely described by Jerenz (2008). Here, we would give a short summary and clarify where it was extended with data-driven methods:

1. The first stage involved identifying car models with similar survival characteristics by comparing estimated survival functions prior to any parametric modeling. We grouped data into separate car types, such as sedans, hatchbacks, etc. Then, within each car type, clusters of models with the same characteristics in terms of survival rates are identified using pairwise comparison of log-rank tests and a weight function of $W(t) = 1$. Models are clustered into a sub-sample only if the null hypothesis of identical survival curves could not be rejected for every possible pair of models within that group, thus ensuring statistical interchangeability ($p > 0.05$).

2. In the second stage, the parametric regression modelling is conducted to identify the best set of features for a given cluster of data and model (Cox PH or AFT). While in Jerenz(2008), data with similar car attributes and pre-defined features (degree-of-overpicing, market size, price quantile for same cars, age, and number of previous owners) were used, our data was limited in terms of number of examples for each car model, but rich on listing attributes (level of promotion, number of phtos, length of text description). Therefore, we used the largest cluster of models with similar survival characteristics for rigorous feature selection. We applied multiple model-selection criteria (forward, backward, stepwise AIC, and stepwise BIC) and retained only variables selected by at least three of the four procedures. This ensemble-based consensus mitigates bias associated with any single feature selection method and provides a data-driven foundation for the final modelling. For the AFT models, we also performed an average AIC comparison between groups of distributions (exponential, Weibull, log-normal, and log-logistic) and found that our data resemble the Weibull distribution.
3. Residuals analysis revealed that both models violated their key assumptions: the Cox model failed to meet its proportional hazards assumption for several key covariates, while the AFT model exhibited significant non-linear relationships with its predictors. In the case of the Cox model, we included time-varying effects for each variable that violated the proportional hazards assumption. In the case of AFT, we extended it with restricted cubic splines to account for any non-linearity. This technique helps model often unknown, nonlinear functional relationships between the continuous covariates and the log-hazard ratio, thereby fulfilling the linearity assumption and enhancing the model's accuracy. Using an iterative data-driven process, the number of knots for each continuous variable was chosen optimally, with AIC serving as the selection criterion. After addressing the main assumption violations in the models, we conducted a 10-fold cross-validation comparison of the extended models (see

Table 5). Both achieved remarkably high discrimination scores. However, the Weibull AFT model showed superior goodness-of-fit, with an AIC approximately 1.6 times lower than the Cox PH model. As a result, the extended Weibull AFT with cubic regression splines was selected as the champion model among classical survival analysis methods

Table 2. Discrimination and goodness-of-fit of classical models

	C-index		AIC		BIC	
Model	**C_index_mean**	**C_index_sd**	**AIC_mean**	**AIC_sd**	**BIC_mean**	**BIC_sd**
AFT_Spline	0.695	0.027	11442.11	96.681	11513.96	96.691
Cox_PH_Extended	0.695	0.028	18076.38	191.449	18117.19	191.532

In the second stage, we applied internal validation using both a subsample from the dataset used for classical model tuning and the full dataset to assess the robustness of our model specification. We also tested the Random Survival Forest model as the best-performing model to have a benchmark.

The 10-fold cross-validation results (see Table 3 and 4) provide a very consistent and nuanced story of the performance of the two champion models. On the smaller, more homogeneous Wagons sub-sample where extensive feature engineering and model tuning were performed, both models' performances are remarkably close, as shown in Table 6. The RSF has a marginal advantage in overall accuracy with a mean IBS of 0.143 versus 0.147 for the AFT Spline model. This slight edge in accuracy comes at the cost of a slightly less stable C-index, as shown by its higher standard deviation.

Table 3. Model's performance on a sub-sample of data (Wagons)

	Integrated Brier Score (IBS)		Concordance Index (C-Index)	
Model	**Mean IBS**	**SD IBS**	**Mean C-Index**	**SD C-Index**
AFT_Spline	0.147	0.009	0.658	0.063
RSF	0.143	0.007	0.663	0.074

Table 4. Model's performance on the full dataset

	Integrated Brier Score (IBS)		Concordance Index (C-Index)	
Model	**Mean IBS**	**SD IBS**	**Mean C-Index**	**SD C-Index**
AFT_Spline	0.155	0.003	0.636	0.027
RSF	0.148	0.003	0.664	0.022

However, the real difference in the capabilities between the models becomes apparent when they are applied to the full, more heterogeneous, dataset shown in Table 7. Here, the Random Survival Forest shows a clear and robust advantage. Its mean IBS of 0.148 is significantly lower than the 0.155 obtained by the AFT Spline model, indicating a greater improvement in overall predictive accuracy. Of course, more importantly, the RSF's discriminative ability-as measured by a mean C-index of 0.664 is substantially higher than the AFT model's 0.636. The RSF also tends to be more stable on the full dataset, with lower standard deviations in its IBS and C-index scores across the 10 folds.

These results lead to a clear answer to our first research question: while a carefully specified classical model, such as the Spline AFT, degrades much more substantially on the full dataset's complexity and heterogeneity, it achieved a very competitive performance on a clean, homogeneous sub-sample. In contrast, the data-driven Random Survival Forest proves more robust and accurate overall. Due to its ensemble nature and the possibility of internal handling of complex interactions, it generalizes better to a diversity of vehicles and outperforms all other methods with respect to both the prediction accuracy (IBS) and discrimination (C-index). Thus, in view of TOM prediction in the Ukrainian used car sector, the Random Survival Forest is considered the most reliable statistical method.

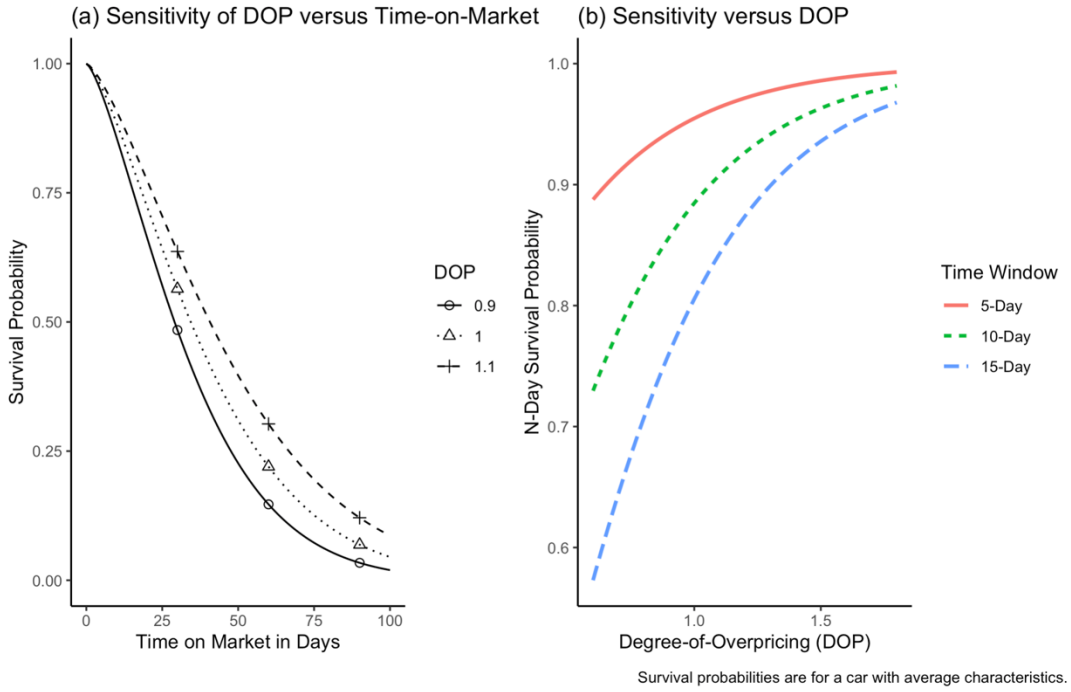
However, classical methods, if properly tuned and validated, exhibit remarkably close performance while offering significantly greater interpretability. This trade-off becomes particularly evident when addressing the primary goal of this research: the estimation of a Price Response Function for profit optimization. By its very nature, the AFT model yields an explicit, parametric equation relating the asking price to the expected probability of sale. It is thus possible to directly translate its coefficients into time ratios, thereby providing a clear and quantifiable answer to the question, "By what percentage will my expected selling time increase if I raise my price by 10%?" This functional form is precisely what is needed for the type of profit optimization frameworks described by Jerenz (2008).

On the other hand, while the RSF proved to be the more accurate predictive tool, however, it doesn't offer a simple, functional relationship between price and the probability of a sale. While this is a powerful tool for understanding the direction and relative magnitude of the price effect, it does not provide the explicit β coefficients needed for direct input into a mathematical profit maximization formula. Thus, while the RSF is superior for forecasting which cars will sell and when, the Spline AFT model remains invaluable for explaining why and providing the specific, interpretable parameters required for strategic, model-based pricing decisions.

This fundamental difference in interpretability versus raw predictive power vividly comes into view when comparing the Price Response Functions estimated by each model in Figure X and Figure Y. Both models successfully answer the research question and confirm the core economic principle: a higher Degree-of-Overpricing provides a higher survival probability,. However, they provide different insightful nuances about the nature of this relationship.

The Spline AFT model (Figure Y) assumes a smooth, monotonic PRF. Its survival curves are well separated, with a classic, clean, parametric shape, offering an interpretable representation of the price effect. It indicates a predictable decrease in probability of sale with each incremental increase in price. This smooth, functional form is best suited for theoretical analysis and direct inclusion in a profit optimization formula.

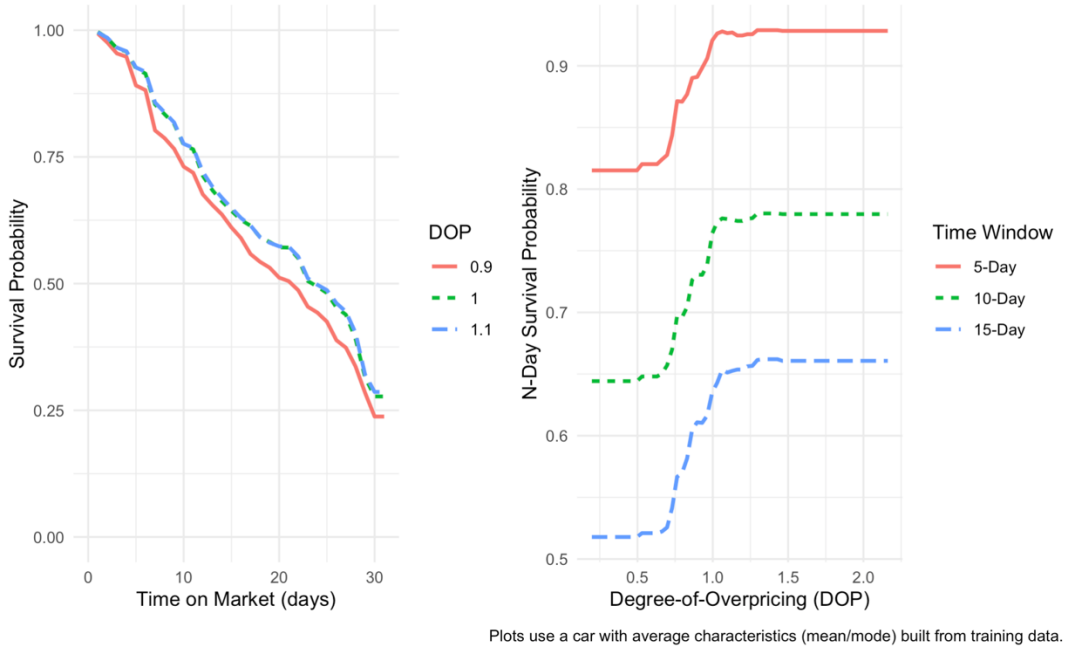
Figure 6. Price-response function from optimized Weibull AFT model



By contrast, the RSF model (Figure X) provides a more complex-and likely more realistic-non-parametric relationship. The PRF it estimates is not smooth but "steppy," with clear plateaus and thresholds. For example, the N-Day Survival plot shows that the model has learned from the data that the largest penalty comes from crossing the threshold from underpriced (DOP < 1.0) to overpriced. It also reveals us that once a car is significantly overpriced, further price increases have a minimal negative impact on the probability of a sale. What is more, the RSF suggests that the practical difference in sales velocity between a fairly priced car and a slightly overpriced one is marginal, as their survival curves are nearly overlapping. This is another data-driven insight that cannot be captured by the smooth curves of the AFT model, but it does suggest that the market reacts to pricing in distinct tiers rather than as a smooth continuum.

Degree-of-overpricing is the single most powerful and significant determinant of a used car time-on-market. To rank the predictive power of all covariates, a robust variable importance (VIMP) analysis averaged over 50 bootstrap runs of the final RSF model was

Figure 7. Price-response function from RSF model

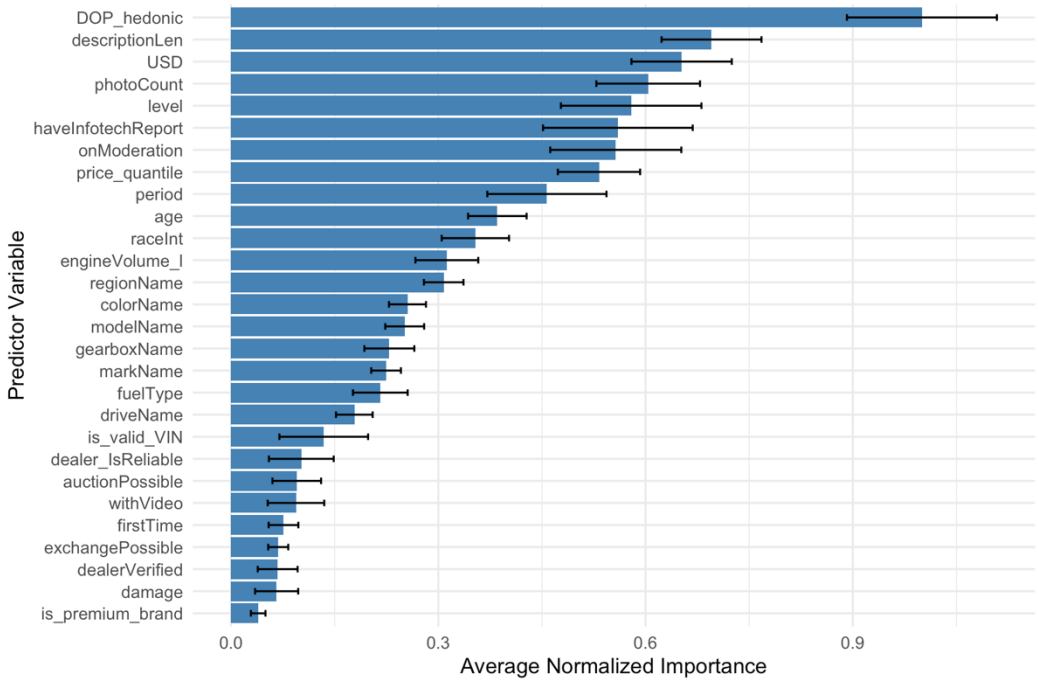


performed (Figure Z). We find that DOP is unambiguously the most influential variable with a normalized importance of 1.0, whose influence is substantially greater than that of the next most important variables, such as descriptionLen (0.69) and USD (0.65). These data-driven findings confirm that a vehicle's price relative to its estimated market value is the most critical piece of information for predicting its sale duration.

The final extended AFT model provides a parametric verification of this. The coefficient for the linear DOP term is 1.68 with a p-value of < 0.0001, making it by far the most statistically significant predictor in the model. In an AFT framework, a positive coefficient indicates an increase in the expected survival time. The Time Ratio of $\exp(1.68) \approx 5.37$ provides a dramatic quantification of this effect: every one-unit increase in DOP-the differential of pricing a car at 200% above market value versus 100%-increases the expected time-on-market by a factor of more than five, all else being held constant. Partial Dependence Plots from the final RSF model visualize this powerful influence (Figure Y). The left-hand plot, "Sensitivity of DOP versus Time-on-Market," clearly illustrates that the

survival curve for an underpriced car (DOP = 0.9) is constantly steeper and lower than the curves for fairly priced or overpriced cars, showing evidence of a faster time to sale. The right-hand plot, "Sensitivity versus DOP," quantifies this relationship at specific time windows. It reveals a non-linear, threshold-based effect: the probability of a car remaining unsold increases dramatically as the DOP crosses the 1.0 "fair price" threshold. For example, for a 5-day window, the survival probability leaps from ~80% to over 90% as the DOP goes from 0.9 to 1.1, confirming that overpricing has a massive and immediate negative impact on the probability of a quick sale.

Figure 8. RSF Variable Importance (averaged over 50 bootstrap runs)



CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

This research sought to answer three fundamental questions about the dynamics of the Ukrainian used car market: finding the best predictive model for time-on-market, quantifying the price impact, and estimating a price response function. The multistage modeling framework, culminating in a head-to-head comparison of a highly specified classical model and a tuned data-driven model, has yielded clear answers and, more importantly, a set of actionable strategic recommendations for market participants. The empirical analysis showed that while a very carefully optimized Spline AFT model can produce outstanding predictive performance, the RSF proved to be marginally more accurate and robust in general, particularly when applied to a large and heterogeneous dataset. This confirms the suspicion that modern data-driven techniques enjoy a slight edge in terms of pure predictive power.

The single most powerful determinant of a vehicle's TOM across all models was the degree-of-overpricing. The second tier of variables involved listing quality and information transparency, such as description length, photo count, and tech reports, which were more influential than many core vehicle attributes, including age and mileage. Finally, the estimated PRF shows a nonlinear, threshold-based relationship between price and probability of sale, with critical inflection points in pricing strategy.

This hierarchy of influence refines the approach put forward by Jerenz (2008), who centered his work largely on the DOP as the main input to a profit optimization module. While our findings confirm that the DOP is indeed the most critical factor, they also show that substantial parts of a vehicle's sales velocity come from actionable "listing quality" metrics, which should not be disregarded. This means that a complete profit optimization strategy must consider not just the price invested in the ad, but also the investment in its presentation. More importantly, the unique strengths of our two champion models illuminate two separate, valuable business applications. Predicting time-on-market is not an end in itself but rather a means to two different strategic ends.

A first application, following the framework of Jerenz 2008, is strategic profit maximization. This requires a clear, functional relationship between price and sale probability to model revenue over time. Our Optimized Spline AFT model is perfectly suited for this. While marginally less accurate in raw prediction, it returns an explicit mathematical formula $\log(T) = \beta_0 + \beta_1 x_1$. directly integrable into a profit optimization algorithm. Business can now easily simulate the financial impact of different pricing policies, e.g., "what is the expected profit if we price all our cars at a DOP of 1.05 vs. 1.10?"; it provides the interpretable parameters needed for high-level, model-based strategic decisions. The second application is tactical opportunity identification, such as finding the best deals or predicting which specific cars will sell quickly. For this purpose, raw predictive accuracy is paramount. Our RSF model, as the cross-validated winner in predictive power, is the ideal tool. It can be deployed to screen thousands of listings and generate a "risk score" for each, allowing a dealer to quickly identify undervalued assets (DOP < 1.0 and a low predicted TOM) for acquisition.

One promising direction for future research would be to try to improve the RSF performance by incorporating the time-dependent variables first set out by Jerenz 2008. While this study relies on cross-sectional data, an RSF model trained on time duration market could learn from dynamic market predictors, such as fluctuating levels of similar models of cars available (market size), shifts in regional demand (cars bought in the last month). This could potentially lead to big improvements in predicting time on market with data-drive methos. Also, a direct next step would be to operationalize the profit optimization framework proposed by Jerenz (2008) in the Ukraine used car market, using explicit coefficients from our AFT model as the core input. Such a module could allow for the simulation and testing of dynamic pricing strategies, going from prediction to active revenue management. An important extension of this framework would be the inclusion of competition-that is, modeling the impact of competitor pricing on the optimal price path. And finally, a very interesting area of research would be to enrich the RSF model with unstructured and behavioral data. This could include variables derived from Natural Language Processing (NLP) to score the quality and sentiment of the description text, or

features from Computer Vision models to assess the photographic quality and "curb appeal" of the listing images. Even more, adding user engagement metrics such as the ratio of views-to-favorites could provide a powerful, real-time proxy for a vehicle's desirability. The acid test for such an improved model would of course be to evaluate its predictive performance against that of human experts in a live forecasting competition. By pitting the algorithm against seasoned dealership managers in the task of predicting which cars are likely to sell within the next 30 days, we could put a true value on data-driven insights. The most compelling result would be the creation of a hybrid decision support system: the provision of a strong data-driven signal to augment the intuition of human experts, leading to a new frontier in optimizing inventory management, refining pricing strategies, and accelerating sales velocity in the used car market.

REFERENCES

Aalen O. 1978. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 6 (July): 701–726. http://www.jstor.org/stable/2958850.

Akerlof G. A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84 (March): 488-500. https://doi.org/10.2307/1879431

Amik F. R., Lanard A., Ismat A. & Momen S. 2021. Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh. *Information* 12: 514.

Avi H., 2018. Optimal Pricing and Replenishment of an Expiring Inventoried Product under Heterogeneous Consumer Sensitivities. *Decision Sciences* 49 (June): 522-552. https://doi.org/10.1111/deci.12276

Ayres I., Siegelman P. 1995. Race and Gender Discrimination in Bargaining for a New Car. *The American Economic Review* 85 (June): 304-321. http://www.jstor.org/stable/2118176

Bapna R., Jank W., Shmueli G. 2008. Consumer Surplus in Online Auctions. *Information Systems Research* 19 (4): 400-416. https://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0173

Bergmann S. & Feuerriegel S. 2024. Machine Learning for Predicting Used Car Resale Prices using Granular Vehicle Equipment Information. *Expert Systems with Applications* 263 (March).

Belleflamme P., Peitz M. 2014. Asymmetric Snformation and Overinvestment in Quality. *European Economic Review* 66: 127-143.

Born A., Kovachka N., Lessmann S., Seow H. 2018. Price Management in the Used-car Market: an Evaluation of Survival Analysis. *IRTG 1792 Discussion Papers*. https://www.econstor.eu/handle/10419/230775.

Bradburn M. J., Clark T. G., Love S. B. and Altman D. G. 2003. Survival Analysis Part II: Multivariate Data Analysis - An Introduction to Concepts and Methods. *British Journal of Cancer* 89: 431-436. https://doi.org/10.1038/sj.bjc.6601119

Cox D. R. 1972. Regression Models and Life Time Tables. Journal of the Royal Statistical Society 34: 187-220.

Chen J., Esteban S., Shum M. 2013. When do Secondary Markets Harm Firms? *The American Economic Review* 103: 2911–2934. https://doi.org/10.1257/aer.103.7.2911

Desai P., Purohit D. 1998. Leasing and Selling: Optimal marketing strategies for a durable goods firm. *Management Science* 44 (November): 19-34. https://doi.org/10.1287/MNSC.44.11.S19

Du J., Xie L., Schroeder S. 2009. Practice Prize Paper: Pin optimal distribution of auction vehicles system: Applying price forecasting, elasticity estimation, and genetic algorithms to used-vehicle distribution. *Marketing Science* 28 (July-August): 637–644. http://www.jstor.org/stable/23884237

Emons W., Sheldon G. 2009. The market for used cars: New evidence of the lemons phenomenon. *Applied Economics* 41: 2867-2885. https://doi.org/10.1080/00036840802277332

Genesove D. 1993. Adverse selection in the wholesale used car market. *Journal of Political Economy* 101 (August): 644-665. https://www.jstor.org/stable/2138742

Graunt J. 1662. Natural and Political Observations on the Bills of Mortality. *John Martyn and James Allestry*. https://name.umdl.umich.edu/A41827.0001.001

Hrechanyk N. 2019. Electric Vehicle Market in Ukraine. *Kyiv School of Economics*. https://kse.ua/wp-content/uploads/2019/04/BFE_Thesis-final_Hrechanyk_Nazarii.pdf

Jerenz A. 2008. Revenue Management and Survival Analysis in the Automobile Industry. Springer. *Gabler Verlag Wiesbaden*. https://doi.org/10.1007/978-3-8349-9840-8

Kaplan E. L., Meier P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53 (June): 457-481. https://doi.org/10.2307/2281868

Kleinbaum D. G., Klein M. 2006. Survival analysis: a self-learning text. *Springer Science and Business Media* 3. https://doi.org/10.1007/978-1-4419-6646-9

Kovpak E. & Orlov F. 2019. Comparative analysis of machine learning models and regressions for car price prediction. *Bulletin of V N Karazin Kharkiv National University Economic Series*: 31-40. https://doi.org/10.26565/2311-2379-2019-97-04

Lessmann S. & Voß S. 2017. Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting* 33: 864-877. https://doi.org/10.1016/j.ijforecast.2017.04.003

Levin J. 2001. Information and the market for lemons. *The RAND Journal of Economics* 32: 657-666. https://doi.org/10.2307/2696386

Matviichuk R. 2021. Pricing of used cars in Ukraine: looking into more than 100,000 deals in the automotive aftermarket. *Kyiv School of Economics.* https://kse.ua/wp-content/uploads/2021/12/Roman-Matviichuk.pdf

Nelson W. 1969. Hazard plotting for incomplete failure data. *Journal of Quality Technology* 1 (February): 27-52. https://doi.org/10.1080/00224065.1969.11980344

Olivares M., Cachon G. P. 2009. Competing retailers and inventory: An empirical investigation of general motors' dealerships in isolated us markets. *Management Science* 55 (September): 1586-1604. https://doi.org/10.1287/mnsc.1090.1050

Prieto M., Caemmerer B. & Baltas G. 2014. Using a hedonic price model to test prospect theory assertions: The asymmetrical and nonlinear effect of reliability on used car prices. *Journal of Retailing and Consumer Services* 22 (January): 206-212. https://doi.org/10.1016/j.jretconser.2014.08.013

Proshchyna T. 2020. Estimation of hedonic pricing model for light vehicles: the case of Ukrainian market for new cars. *Kyiv School of Economics.* https://kse.ua/wp-content/uploads/2021/04/Master-thesis_Proshchyna.pdf

Ratchford B. T., Srinivasan N. 1993. An empirical investigation of returns to search. *Marketing science* 12 (February): 73-87. https://doi.org/10.1287/mksc.12.1.73

Schoenfeld D. 1982. Partial residuals for the proportional hazards regression model. *Biometrika* 69 (April): 239-241. https://doi.org/10.2307/2335876

Shymanskyi V. & Liaskovets V. 2023. Cascade model for price and time of car sales prediction. *In ProfIT AI:* 152-167. https://ceur-ws.org/Vol-3641/paper14.pdf

Sutherland I. 1963. John Graunt: a tercentenary tribute. *Journal of the Royal Statistical Society* 126: 537-556. https://doi.org/10.2307/2982578

Wertenbroch K. and Skiera B. 2002. Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research* 39 (May): 228-241. https://doi.org/10.1509/jmkr.39.2.228.19086

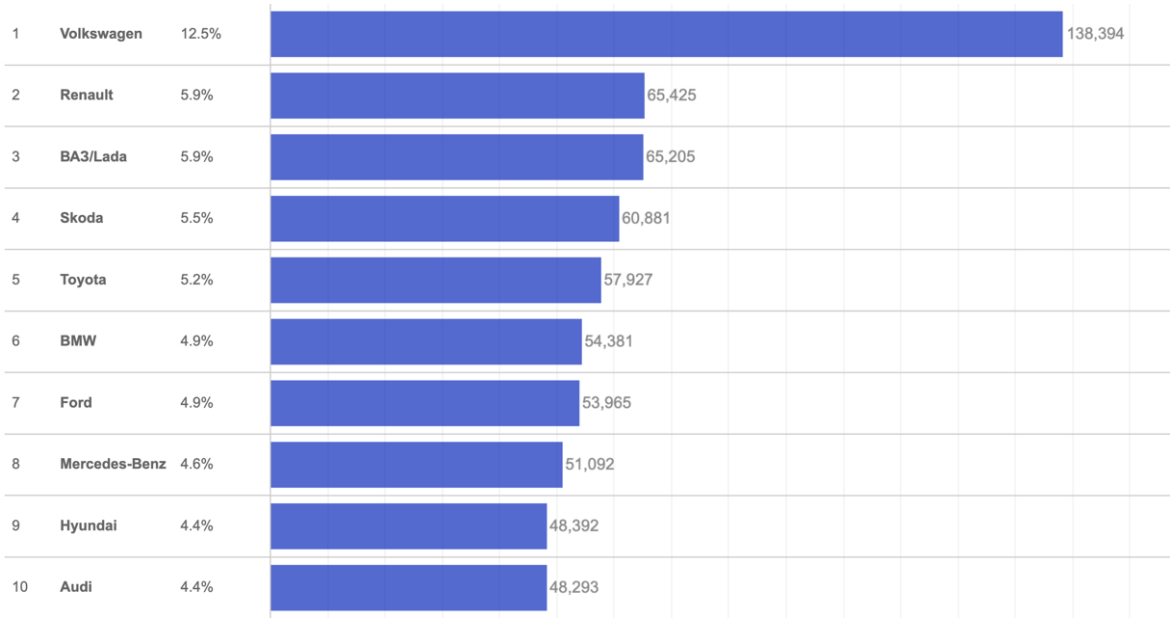
SimilarWeb. Estimated website statistics. https://www.similarweb.com.

Automotive Market Research Institute. Private Analytical Centre. https://eauto.org.ua.

Dat report 2017.
https://www.dat.de/fileadmin/de/images/produkte/DAT_Barometer/Archiv/GESAMT2017.pdf

APPENDIX A
FIGURES

Figure A.1 Top 10 passenger car brands, domestic market, 2024



APPENDIX B

DATA TABLE

Table B.1 Dataset variables with Descriptions

Variable	**Description**
autoId	Listing ID
title	Title
linkToView	Link to the listing
VIN	Vehicle identification number
markName	Car brand
modelName	Car model
categoryName	Vehicle category name
bodyName	Body style (e.g., sedan, hatchback
UAH	Price in Ukrainian hryvnias
EUR	Price in euros
USD	Price in US dollars
minMonthLeasingBuPay	Minimum lease amount (uah)
age	Car age (years)
raceInt	Mileage (000's km)
fuelType	Fuel type
engineVolume_l	Engine volume (litres)
gearboxName	Gearbox type (transmission)
driveName	Drive type
generationName	Car model generation
colorName	Exterior color name
colorHex	Exterior color name (hex code)
locationCityName	City
regionName	Region
isSold	Indicates whether the car is sold (yes/no)
statusId	Status of listing (0 – active, 1 – sold/deleted)
photoCount	Number of photos
withVideo	Video availability (yes/no)
auctionPossible	Negotiable (auction possible)

mainCurrency	Main currency
isShow	Display VIN (true/false)
agreeShowVIN	If user agreed to show VIN
seoLink	Link to the listing photo used for SEO
description	Description of listing
descriptionLen	Number of characters in the description
level	Premium listing level
addDate	Date and time added
updateDate	Date and time updated
soldDate	Date when the listing was marked sold
expireDate	Date and time expired (listing expiration)
level_expireDate	Expiration of Premium listing status
firstTime	First-time listing indicator (yes/no)
dealerName	Dealer or seller name
dealer_IsReliable	Dealer reliability flag (yes/no)
dealerVerified	Dealer verification status (yes/no)
wasCrashed	Was in traffic accident (yes/no)
haveInfotechReport	Infotech report available (yes/no)
techCondition_annotation	Technical condition (1 – Fully undamaged; 2 – Professionally repaired damage; 3 – Unrepaired damage; 4 – Not drivable / for parts)
techCondition_title	More details on technical condition
damage	Is damaged (yes/no)
vat	Value-added tax included (yes/no)
onRepairParts	Listing is for repair/parts only (yes/no)
underCredit	Under financing/credit arrangement (yes/no)
exchangePossible	Is exchange possible (yes/no)
exchangeType	Exchange type