Estimating Liquidity in Online Marketplaces via Machine Learning and Survival Analysis

Marcell P. Granát, (ORCID: 0000-0002-4036-1500, https://scholar.google.com/citations?user=BCrw9NsAAAAJ)[1,2]
and Péter Vékás, PhD (ORCID: 0000-0002-2626-814X, https://scholar.google.com/citations?user=m-CY0cMAAAAJ)[3*]

[1]Central Bank of Hungary.
[2]Eötvös Loránd University, 1 Egyetem tér, Budapest, 1053, Hungary.
[3*]Institute of Operations and Decision Sciences, Corvinus University of Budapest, 8 Fővám tér, Budapest, 1093, Hungary.

*Corresponding author(s). E-mail(s): peter.vekas@uni-corvinus.hu;
Contributing authors: marcigranat@student.elte.hu;

Abstract

Estimating market liquidity is critical for buyers and sellers navigating markets for heterogeneous goods. This paper introduces an integrated workflow for analyzing liquidity across diverse online marketplaces using machine learning and survival analysis. Our approach estimates the expected time to sell a product in three steps: (1) data collection and preparation, (2) dimensionality reduction and price prediction using machine learning, and (3) survival analysis.

We demonstrate the effectiveness of the proposed workflow through a case study of a popular online used car marketplace. Our results reveal that after estimating market-consistent asking prices from observed listing characteristics via machine learning, the ratio of actual and predicted offer prices is a strong predictor of listing duration and improves survival model discrimination relative to the raw asking price. Our main findings remain robust under several plausible alternative model specifications.

Our analysis demonstrates the broad applicability of the proposed framework to online marketplaces with diverse goods and significant liquidity considerations. Additionally, the workflow can be used by market analysts and integrated into online marketplace interfaces to inform sellers of expected selling times based on their offer prices and product characteristics.

Keywords: empirical economics, market liquidity, online marketplaces, machine learning, survival analysis.

1 Introduction

Online marketplaces have become integral to today's economy and everyday life, serving as platforms for the exchange of goods and services with unprecedented accessibility and efficiency. These marketplaces cater to a diverse range of products, from everyday essentials to specialized items like used cars and real estate, making them vital components of the modern digital economy. Understanding the dynamics of such platforms – particularly pricing and liquidity – offers critical insights for buyers, sellers, and policymakers.

To analyze these marketplaces, machine learning and survival analysis provide a complementary toolbox. Machine learning excels at uncovering patterns and relationships in large datasets, offering predictive power for pricing and market trends. Meanwhile, survival analysis models time-to-event outcomes, making it particularly suitable for studying the liquidity of goods in online marketplaces, such as the time it takes for a product to sell.

For example, cars are among the most expensive consumer durables owned by households, so buyers carefully consider their decisions. These vehicles are commonly resold in online marketplaces after use, and it is important to know how much owners can sell their cars for and how long it will take. The value of each vehicle varies, so to estimate how quickly it will sell at a given price, one must first assess its market-consistent asking price given its characteristics. Intuitively, a very important determinant of the time to sell is the price-to-value ratio, defined as the ratio of the advertised asking price to the estimated market-consistent asking price. Additionally, car characteristics are also associated with listing duration. Beyond advancing economic knowledge, understanding these mechanisms can provide valuable information to all stakeholders in online marketplaces. In particular, dealers can improve their planning by knowing how quickly they will be able to dispose of the assets they currently hold, which may have storage costs and capacity constraints.

Our paper contributes to the growing literature on online marketplaces by demonstrating how machine learning-based valuation and survival analysis can be integrated into a coherent empirical workflow for estimating liquidity in markets for heterogeneous goods. Rather than introducing a fundamentally new methodological approach, we combine established machine learning techniques for price prediction with Cox proportional hazards modeling and random survival forests (RSFs) to estimate time-to-sale as a function of the price-to-value ratio and product characteristics. This applied methodological integration provides a practical and transferable framework for analysing liquidity in heterogeneous online marketplaces.

We present a literature review and the proposed methodological workflow in Sections 2 and 3, followed by demonstrating its effectiveness and potential applicability to a wide range of online markets through a case study on the Hungarian used car market in Section 4 and the associated numerical results in Section 5. We state the limitations of our study in Section 6, and finally, we conclude with Section 7.

2 Literature review

Rosen (1974) introduced the popular framework of hedonic pricing models to estimate the impact of attributes on prices of heterogeneous goods. These models decompose the price into constituent characteristics, allowing for an assessment of how each feature contributes to the overall value. Although widely used, these traditional econometric approaches rely on restrictive functional forms and face difficulties handling the high-dimensional feature spaces typical of modern online marketplaces.

Machine learning techniques such as random forests, support vector machines, and artificial neural networks have also been employed in this domain to overcome these limitations and improve predictive accuracy. For example, Pudaruth (2014) applies machine learning techniques to predict used car prices, achieving varying degrees of accuracy across different models. Pal et al. (2017) propose random forests to predict prices based on various features, achieving substantial accuracy in used car price prediction. Pangallo and Loberto (2018) investigate if online interest can be used as a proxy of housing demand using supervised machine learning techniques on data from online home advertisements. Jin (2021) develops a pricing model for used cars in online marketplaces based on factors like mileage, manufacturing year, fuel consumption, and engine size. Bukvić et al. (2022) utilize supervised learning methods to predict used car prices in the Croatian market, incorporating features such as production year and mileage. Akyüz et al. (2023) introduce a hybrid model combining decision trees, support vector machines, and artificial neural networks to enhance house price estimation. Saini and Rani (2023) assess the performance of popular machine learning models, finding that eXtreme Gradient Boosting (XGBoost) achieved the best predictive performance among the evaluated methods. Dutulescu et al. (2023) demonstrate on car datasets that recent advancements in deep learning have further expanded the capabilities of these tools, capturing complex feature relationships and providing robust price estimates. Çılgın and Gökçen (2024) develop a machine learning architecture that integrates clustering and stacking ensemble techniques, achieving high accuracy in real estate price prediction. Madhusudhanan et al. (2024) introduce Prob-SAINT, a probabilistic tabular regression model that not only provides accurate point predictions but also quantifies uncertainty.

While these studies achieve strong predictive performance, they focus almost exclusively on price estimation, leaving aside the temporal aspect of how long it takes for a product to sell. In contrast, survival analysis provides a complementary framework that models time-to-event outcomes and is particularly suited for analyzing liquidity in online marketplaces. Demiriz (2018) employs survival analysis to investigate the lifespan of used car advertisements on online platforms like Craigslist. By treating the delisting of an advertisement as a proxy for sale, he explores how factors like asking price and vehicle attributes influence listing duration, yielding actionable insights into market liquidity. A broader methodological review by Wiegrebe et al. (2024) highlights the convergence of deep learning and survival analysis, emphasizing the potential of neural networks to model time-to-event outcomes and to integrate diverse data types, such as numerical features and unstructured data like images or text.

Regarding the economic background of these marketplaces, Akerlof (1970) examines the concept of information asymmetry in markets of heterogeneous goods, as illustrated by the used car market, highlighting how quality uncertainty can lead to inefficiencies—a phenomenon termed the "market for lemons". This theory suggests that sellers possess more information about a product's quality than buyers, potentially resulting in adverse selection and market failure. Such information asymmetries make liquidity analysis particularly relevant for heterogeneous goods.

Liquidity is a multidimensional concept whose appropriate measurement depends on the institutional characteristics of the market under consideration. In continuously traded financial markets, empirical indicators include bid–ask spreads, turnover measures, and measures of the price impact of trading. For example, Amihud (2002) defines an illiquidity measure as the ratio of the absolute daily stock return to dollar trading volume, averaged over a specified period, and interprets it as a rough measure of the price response associated with one dollar of trading volume. Such measures are particularly suited to markets in which transaction prices and trading volumes are observed.

Online marketplaces for durable and heterogeneous goods differ substantially from continuously traded financial markets. Transactions are relatively infrequent, individual items are imperfect substitutes, and information on trading volume, order flow, bid–ask spreads, and transaction-level price impact is generally unavailable. A time-based operational definition is therefore more applicable in this setting. Lippman and McCall (1986) define an asset's liquidity as the expected time until it is sold while an optimal selling policy is pursued. Consistent with this conception and with the literature on markets for heterogeneous durable assets, the present study operationalizes liquidity through listing duration. Because the available data identify the disappearance of an advertisement rather than a verified transaction, the empirical measure is more precisely interpreted as time-to-delisting used as a proxy for time-to-sale.

Despite the cited advances, prior research on price prediction and liquidity modeling has largely evolved in isolation. Machine learning–based studies excel at predicting prices but typically ignore how pricing relates to selling time, while survival analysis–based studies effectively model time-to-sale but treat offer prices as exogenous and fixed, without accounting for whether they represent fair market value. This separation limits the ability to jointly analyze pricing and liquidity, especially in markets characterized by heterogeneity and changing conditions.

Building on these two strands of literature, our study demonstrates a unified empirical workflow. Specifically, we employ established machine learning techniques for dimensionality reduction and price prediction and subsequently apply Cox regression and RSFs to model expected time-to-sale as a function of the price-to-value ratio and product characteristics. The contribution is therefore an applied methodological integration rather than the development of a new prediction or survival-analysis method. The use of an interpretable Cox model in the final stage preserves interpretability, while the RSF aims to maximize predictive accuracy.

In summary, this paper contributes to the literature in two ways: (1) it demonstrates how machine learning-based valuation and survival analysis can be integrated into a single empirical workflow, and (2) it illustrates the applicability of this framework to analyze the relationship between pricing and liquidity of heterogeneous goods in dynamic online markets through a case study of the Hungarian used car market.

3 Methodological workflow

The objective of this study is to present an integrated empirical workflow for jointly predicting prices and assessing liquidity in online marketplaces for heterogeneous goods. The workflow consists of three principal steps:

1. data collection and preparation,
2. dimensionality reduction and price prediction, and
3. survival analysis.

3.1 Data collection and preparation

Estimating liquidity in online marketplaces requires granular and temporally consistent data that capture both item characteristics and listing dynamics. Two main data acquisition strategies are suitable for this purpose: webscraping and Application Programming Interfaces (APIs). Both approaches allow the systematic collection of advertisements at regular intervals, ensuring that each listing's lifetime, pricing history, and observable attributes are recorded.

Webscraping retrieves publicly available information from marketplace websites, while APIs—where accessible—offer a structured and legally compliant data source that often includes metadata not available on front-end interfaces. Data should be collected frequently enough to observe the full life cycle of listings, including price changes and delisting dates. The resulting dataset must then be cleaned, filtered, and validated to ensure internal consistency and external reliability. Although implementation details (e.g., scraping tools or programming languages) may vary, the key design principle is the temporal completeness and accuracy of listing-level data.

We assume that for every item $i = 1, 2, \dots, n$, its observable characteristics are given by $x_i = (p_i, u_i, v_i)$, where p_i denotes the offer price, u_i represents quantitative attributes, and v_i encodes qualitative attributes as binary variables. Sales outcomes are represented by $y_i = (s_i, t_i)$, where $s_i = 1$ if the item has been delisted and 0 otherwise, and t_i measures its time on the market. Following Demiriz (2018), we treat delisting as a proxy for successful sale, an assumption supported by negligible listing costs, industry practices such as the Cox Dealer Delisted Price Index (Cox Automotive Market Insights, 2023), and the rarity of identical relistings. This proxy enables the approximation of selling times even in the absence of direct transaction data.

3.2 Dimensionality reduction and price prediction

Online marketplaces feature highly heterogeneous products with a large number of categorical attributes and relatively few continuous ones. This leads to multicollinearity and high-dimensional design matrices that challenge both interpretability and computational feasibility. To mitigate these issues, we apply dimensionality reduction to the one-hot encoded qualitative characteristics v_i using Multiple Correspondence Analysis (MCA), a method particularly suited for categorical variables (Venables and Ripley, 1999; Abdi and Valentin, 2007). Factors with eigenvalues exceeding the average are retained (Kaiser, 1960), yielding reduced feature representations $v_i^* = v_i R$ and updated characteristics $x_i^* = (p_i, u_i, v_i^*)$.

This step balances two competing objectives: dimensional parsimony and preservation of interpretability. While techniques such as autoencoders could achieve more efficient compression, MCA was chosen because it allows straightforward mapping of the derived factors back to interpretable categorical dimensions, aligning with our aim to maintain transparency throughout the pipeline.

In the next stage, supervised machine learning models are trained to predict the observed advertised asking price π_i of each vehicle from its observable characteristics:

$$\pi_i = f(u_i, v_i^*, \omega),$$

where f denotes the trained model with hyperparameters ω. Consequently, the model output should be interpreted as an estimated market-consistent asking price, rather than as a realized transaction price or an intrinsic economic value of the vehicle. We employ machine learning rather than conventional econometric models to enhance predictive accuracy and capture nonlinear feature interactions, as demonstrated in related literature (James et al., 2013). Models such as tree-based ensembles (e.g., Gradient Boosting Machines, XGBoost, LightGBM) and neural networks can flexibly approximate complex relationships between attributes and prices.

Our framework does not pursue causal inference but rather accurate price prediction as an input to the subsequent survival model. Hence, model selection focuses on predictive performance rather than parameter interpretability. Optimal algorithms and hyperparameters should be selected based on a chosen performance metric (e.g., MSE or MAE) using robust validation schemes such as repeated k-fold cross-validation or bootstrapping. After training, the ratio of the actual and predicted offer prices defines the price-to-value ratio:

$$r_i = \frac{p_i}{\pi_i}.$$

This ratio operationalizes the concept of overpricing and underpricing, linking valuation to liquidity in the final stage.

3.3 Survival analysis

The final step models time-to-sale as a function of the price-to-value ratio and item characteristics. To distinguish between explanatory and predictive objectives, we propose two complementary survival analysis approaches. We employ the Cox proportional hazards model (Cox, 1972):

$$H_i(t, r) = e^{\beta u_i + \gamma v_i + \delta r} H_0(t), \tag{1}$$

where $H_i(t, r)$ is the cumulative hazard and $H_0(t)$ the baseline hazard. The choice of the Cox model reflects both theoretical and practical considerations. Unlike fully parametric survival models, the Cox model does not assume a specific distribution of selling times, making it suitable for heterogeneous goods with unknown hazard shapes. Moreover, it provides interpretable coefficient estimates that directly quantify the marginal associations of pricing decisions and product characteristics with the selling hazard. This interpretability is central to our research question, which aims to examine the association between deviations from the estimated market-consistent asking price and liquidity.

In parallel, we estimate an RSF model (Ishwaran et al., 2008). As a non-parametric ensemble learning method, RSF naturally accommodates nonlinear relationships and higher-order interactions while avoiding the proportional hazards assumption required by the Cox model. Consequently, RSF serves as a flexible predictive model that complements the interpretable Cox regression framework.

The proportional hazards assumption—that covariate effects remain constant over time—can be evaluated using scaled Schoenfeld residuals (Grambsch and Therneau, 1994). If violations arise, remedies include stratified Cox models or time-dependent covariates (Fisher and Lin, 1999; Zhang et al., 2018), or alternatively, the RSF model remains applicable without requiring proportional hazards. To avoid overfitting given the potentially large number of covariates, we propose to estimate the Cox model with LASSO regularization (Simon et al., 2011), selecting the penalty parameter λ via cross-validation and measuring fit with Harrell's concordance index (Schmid et al., 2016). Variables retained by LASSO should then be re-estimated in an unregularized model to reduce the shrinkage bias introduced by penalization.

The survival function, derived as

$$G_i(t, r) = \exp[-H_i(t, r)],$$

captures the probability that an item remains unsold beyond time t. For the RSF model, survival probabilities are estimated non-parametrically from the ensemble of survival trees. Averaging these functions across selected subsets or the whole set of listings allows for the empirical assessment of liquidity patterns across categories, price segments, or over time.

This design ensures coherence between methodological choices and research objectives. Dimensionality reduction addresses the high cardinality of categorical attributes while preserving interpretability; machine learning provides flexible estimation of market-consistent asking prices under nonlinear relationships; the Cox model provides interpretable coefficient estimates, while the RSF model offers a flexible alternative focused on predictive accuracy without requiring proportional hazards. Comparing their out-of-sample performance enables an empirical evaluation of the trade-off between interpretability and predictive performance. These components bridge the gap between valuation and time-to-sale modeling within a single empirical workflow. By contrast, most earlier approaches in the literature have typically focused on only one of these tasks.

Alternative methodological choices are possible for both the valuation and liquidity components of the proposed workflow. For price estimation, traditional hedonic regression models remain widely used because of their interpretability, while a broad range of machine learning methods, including random forests, gradient boosting, and neural networks, have been applied when predictive accuracy is the primary objective. Likewise, selling times can be analysed using linear or generalized regression models when censoring is ignored, classical Cox proportional hazards models when interpretability is desired, or more flexible machine learning survival methods such as RSFs. The objective of the present study is not to provide a comprehensive benchmark comparison among these alternatives. Rather, we propose an integrated workflow that combines market-consistent asking-price estimation with survival analysis, thereby enabling the investigation of how relative pricing is associated with listing duration within a unified empirical framework.

4 Case study: the online used car market in Hungary

To illustrate our proposed approach on real-world data, we turn to the website `www.hasznaltauto.hu`, widely recognized as the largest online marketplace for used cars in Hungary, significantly surpassing any competitor in terms of listings, web traffic, and overall market reach (Hungary Today, 2024; Portfolio.hu, 2024). It hosts on average over 110,000 active listings at any given time, totaling roughly one million vehicle advertisements annually, making it Hungary's largest database for used vehicles. Recent web analytics confirm that it receives around 11 million visits monthly, far exceeding the closest competitors, whose monthly traffic remains significantly lower (SimilarWeb, 2025).

We shall focus on two research questions: (Q1) **the relationship between the selling time and the price-to-value ratio**, and (Q2) **how other vehicle characteristics are associated with selling time**.

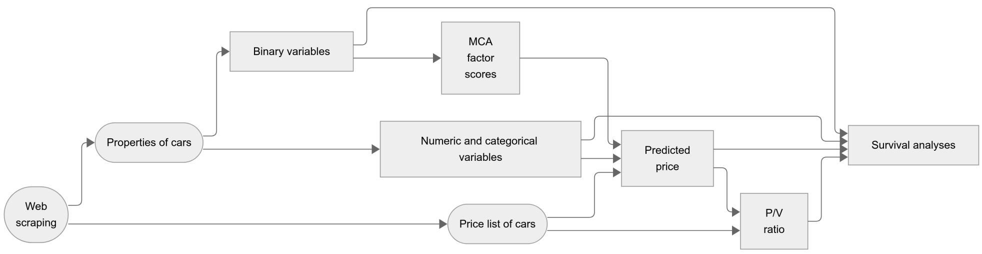

Fig. 1 Overview of the study design

We illustrate the study design in Figure 1. We shall present the practical implementation of the phases of the analysis described in Section 3 in separate subsections.

4.1 Data collection and preparation

We used webscraping to regularly collect data including 250 observed characteristics from all car advertisements posted on `www.hasznaltauto.hu` between May 2021 and March 2022, a period spanning 10 months. Due to technical difficulties, it was not possible to download all data daily without omission.[1] Appendix A shows the days on which data were downloaded. We scraped data from 613,604 advertisements throughout the period.

We collected the IDs and prices of all cars available on each day, as well as all vehicle characteristics that were first available on that day. In other words, we assumed that all technical details of a given car remained constant over the lifetime of the advertisement in order to reduce the computational need for regular data downloads and the memory size of the data collected. By default, we use the last observed offer price of each listing, and alternatively, as a sensitivity analysis, we use the time-weighted average of its advertised prices over the entire observed listing period, as well as the initial offer price and the simple mean of the initial and final asking prices. Following academic and industry standards, we treat delisting as a proxy for sale, while recognizing that some listings may disappear for other reasons.

In our baseline scenario, cars that remained listed when data collection ended on March 14, 2022, were treated as right-censored observations in the survival analysis.[2] Due to the regular downloading, it is known when a particular car advertisement started and when it disappeared from the website or, for censored listings, how long it remained observable until the end of the data-collection period. This information allows us to calculate the observed listing duration[3] and to distinguish between listings that ended during the observation period and those that were still active at its conclusion. These durations and censoring indicators are used as the outcome variables in the survival models.

An important technical detail for the analysis is that three types of data are linked to each car during data collection:

1. *Identifiers and offer prices*, which are downloaded daily.
2. *General characteristics* of cars, which are available for all cars (sometimes refused to be filled in). These 23 variables include the number of kilometers driven, the year of construction, engine power, etc., and are categorical or numeric (see the full list in Figure 2).
3. *Accessories* that only some cars have, such as cruise control, bluetooth, leather seats, etc. These are listed in the descriptions of cars. We created binary variables (whether the car has them or not) for all the accessories that occur. This gave us a total of 235 dummy variables, from which we later created factor scores using MCA to reduce the number of dimensions.

We imputed missing general characteristics by the k Nearest Neighbors (kNN) procedure. The proportion of missing data points is shown in Figure 2 for each variable. We applied $k = 5$ and Gower's distance (Liao et al., 2014, p. 5), which is well-suited

[1] But with a personal computer dedicated to this purpose, the process will continue without interruption in the future.

[2] Additionally, as a sensitivity analysis, we evaluated three alternative assumptions: (i) right censoring as before but excluding repeatedly reposted listings, (ii) the removal of listings active at the end of the period, and (iii) the removal of these listings and all others whose prices changed during the observation period.

[3] Alternatively, we used the total number of days the listing was observed as active, excluding temporary removals, for sensitivity analysis.

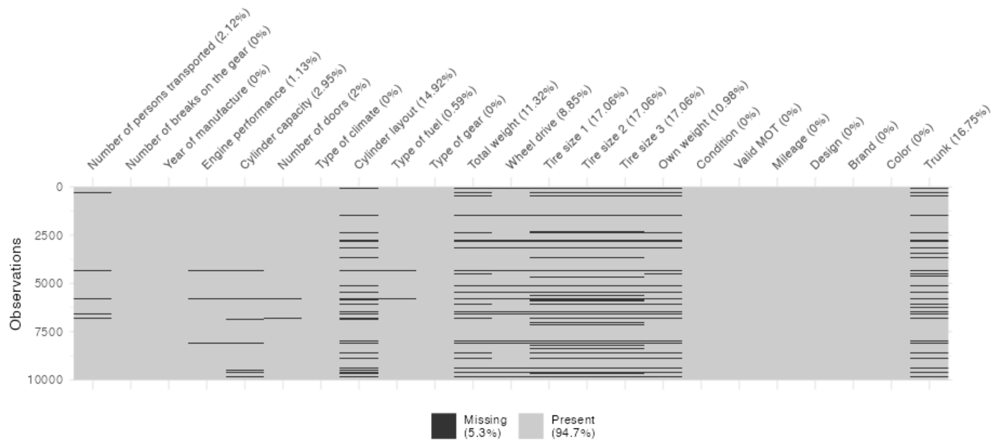

Fig. 2 Missingness among the general characteristics of cars. Values in parentheses indicate the proportions of missing data. This figure was created on a simple random sample of 10,000 advertisements (the proportions remained stable after regenerating the sample).

for datasets containing both categorical and numeric variables.[4] Including all possible predictor variables in the kNN algorithm involves considerable computational effort, so we used theoretical considerations to determine the most important predictors for each variable and whether to treat them as categorical or numeric.[5] Table 1 lists the imputed variables with their predictors that we used in the kNN imputation procedure.

Although kNN imputation is a widely used approach for handling missing data, it has the potential to distort the multivariate dependence structure and other statistical properties of the data because the imputed values are derived from the observed multivariate structure, which may lead to optimistic model fit or biased coefficient estimates in subsequent analyses (Sterne et al., 2009; Beretta and Santaniello, 2016; van Buuren, 2018). The proportion of missing data was generally low across variables (Figure 2), with the highest level of missingness observed for trunk volume (approximately 16%), which reduces the likelihood that the choice of imputation method materially affects the study conclusions. Additionally, we evaluated two alternative imputation strategies as sensitivity analyses: median imputation for numerical variables combined with mode imputation for categorical variables, and Multiple Imputation by Chained Equations (MICE) using the `mice` package in R (van Buuren and Groothuis-Oudshoorn, 2011) with default settings.

The submitted listing forms may contain intentional or unintentional data entry errors. Errors in the reported offer price are particularly important for our analysis. We omitted advertisements with offer prices below HUF 300,000 or above HUF 30,000,000,

[4] Gower's distance is calculated as the average of the variable-wise distances between observations. For categorical variables, the distance is 0 if the values are identical and 1 otherwise. For numeric variables, the distance is the absolute difference between values, normalized by the range of the variable.

[5] For example, we considered car tire size with 3 possible values to be categorical, as it is usually a characteristic of a car type, strongly related to the price of the car, and its impact on the price is not linear, i.e., there are sizes that are significantly more expensive for premium models. This assumption was motivated by the advice of tire manufacturer and distributor Schnell Service 2000 Kft.

Table 1 Variables used for imputation.

Imputed variable	Predictors
cylinder capacity	brand, year of manufacture, number of persons transported
cylinder layout	brand, year of manufacture, cylinder capacity, engine performance
number of doors	brand, year of manufacture, own weight, number of persons transported
number of persons transported	brand, year of manufacture, own weight
own weight	brand, year of manufacture, engine performance, number of persons transported, number of doors
tire size 1	brand, year of manufacture, own weight
tire size 2	brand, year of manufacture, own weight
tire size 3	brand, year of manufacture, own weight
total weight	brand, year of manufacture, engine performance, number of persons transported, number of doors
trunk	brand, year of manufacture, number of persons transported, own weight, total weight
type of fuel	brand, year of manufacture, engine performance, cylinder capacity
wheel drive	brand, year of manufacture, engine performance, type of fuel

which was the case for about 10,000 vehicles. In most of these cases, the advertising price was HUF 1, which was presumably entered to display the car in front of the ascending order by price. In a few cases, we found car advertisements with prices above HUF 30,000,000 that were very unrealistic based on their technical details, and their inclusion could have severely distorted the analysis. We determined the lower and upper limits by browsing the advertisements on the website, setting a realistic range of values for a used car. We also removed advertisements of cars whose mileages were not reported, as this is likely to have a strong influence on prices, and imputing it would lead to very noisy estimates. After removing all outliers and handling missing data, we performed the modeling on 566,998 complete observations.

Table 2 provides an overview of the offer prices. As can be seen in the table, prices on the second-hand car market increased during the observation period, just like prices in the total vehicle market, as Hungarian Central Statistical Office (2022) documented a 4.5% price increase in vehicle purchases in 2021. This illustrates that the market is not stable over time, which may be due to changes in the economic environment.[6] The descriptive statistics also reveal the strongly positively skewed and leptokurtic distribution of the prices. Hence, we used the logarithm of the offer price in our models.

To visualize the distribution of the times to sell, we computed the Kaplan-Meier estimate of the survival function, and plotted it in Figure 3. The median survival time is 11 days, while the average survival time is 21.1 days. Not surprisingly, this reveals that the distribution of selling times is also positively skewed.[7]

[6]Price fluctuations may reflect changes in the exchange rate, as numerous used automobiles are imported into Hungary. Consumers' car purchases may also be affected by the economic cycle. For instance, only 61,324 passenger cars were placed on the Hungarian market after the financial crisis, in 2010, compared with 149,137 in 2000 and 295,431 in 2018 (Hungarian Central Statistical Office, 2020).

[7]Selling times have a lower bound of zero, but no upper bound in theory (the length of the observation window in practice).

Table 2 Descriptive statistics of offer prices, considering only the cars used in the modeling, without outliers. Prices are denominated in HUF.

Quarter	Mean	Median	Standard deviation	Skewness	Kurtosis
2021-04/2021-07	3,710,100	1,949,000	4,821,653	2.795249	11.63271
2021-07/2021-10	3,774,025	1,950,000	4,883,133	2.731550	11.19031
2021-10/2022-01	3,904,783	1,990,000	5,057,759	2.626280	10.32879
2022-01/2022-04	4,068,776	2,200,000	4,873,486	2.524668	10.10797
Total	3,871,298	1,999,000	4,931,980	2.659425	10.71338

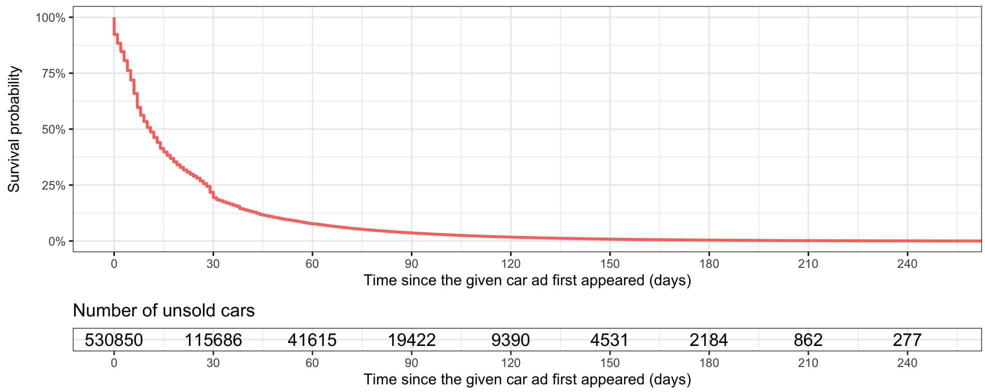

Fig. 3 Kaplan-Meier estimate of the survival function.

4.2 Dimensionality reduction

Following data preparation, we reduced the dimensionality of the numerous qualitative features of used cars by means of MCA. We used the brand and all accessories variables for this purpose.[8]

4.3 Supervised learning

Following dimensionality reduction, we applied supervised learning to predict the observed advertised asking price of each vehicle from its observable characteristics, using the following techniques:

- Linear Regression,
- CART Regression Trees,
- Random Forests,
- eXtreme Gradient Boosting (XGBoost),
- k Nearest Neighbors (kNN),
- Artificial Neural Networks (ANN),
- Support Vector Machines (SVM).

[8]In reality, the application of MCA was indispensable here, as some supervised machine learning methods such as random forest failed to handle the multitude of dummy variables representing car brands.

Table 3 Numbers of hyperparameters, combinations of tested parameters, and total runtimes of hyperparameter tuning.

Model	# hyperparameters	# combinations	Runtime (sec)
EXtreme Gradient Boosting	7	128	66,280
Ordinary Least Squares	0	1	4
Single-hidden-layer neural network	3	46	37,353
k Nearest Neighbors	3	32	213,280
Random Forest	2	32	60,923
RBF Support Vector Machine	3	128	184,857
Regression tree	3	64	814

Additionally, we built a simple average ensemble of all the models in the above list to reduce model uncertainty, as previous research has shown that such unweighted combinations often perform competitively relative to more complex weighting schemes, partly because they avoid the additional estimation error associated with learning ensemble weights (Smith and Wallis, 2009; Claeskens et al., 2016).

The supervised learning models were trained on log-transformed offer prices, and the predicted prices were obtained by applying the exponential transformation to the predicted log prices. As a result, the estimated market-consistent prices correspond to conditional medians rather than conditional means. We deliberately adopted this approach because median predictions are more robust to the right-skewed distribution of used car prices and to the influence of extreme observations.

We tuned the hyperparameters of the models in a stratified k-fold cross-validation framework: we selected a random sample of 10,000 (this was necessary due to the high computational demands of the models) observations sold until July, and split it 10 times into training (9,000 observations) and validation (1,000 observations) sets, stratified by quartiles of the offer price. We chose the hyperparameter combination with the highest average R^2 in the 10 validation samples.

After tuning, we created an independent test sample of 10,000 cars sold before July 2021 in order to select the best out of seven candidate models and their ensemble.

Table 3 displays the number of hyperparameters, the number of parameter combinations, and computation times[9] per model. See Appendix B for a full list of tuned hyperparameters.

As supply prices generally increased in the market (see Table 2), we ordered sold listings by their calendar delisting dates, divided them into consecutive groups of 10,000, and built models to estimate market prices on each subsample. Although this approach partly accounts for inflation, its primary purpose is to allow the relationship between vehicle characteristics and market prices to evolve over time rather than assuming a constant pricing function. Changes in consumer preferences, technological developments, regulatory policies, and supply conditions may alter the implicit value of attributes such as fuel type, engine size, transmission, or mileage. Rolling time windows therefore allow the pricing function itself to adapt over time. As a sensitivity analysis, we additionally estimated two single global models including quarterly and monthly dummy variables, respectively, instead of using rolling time windows. As there

[9]CPU: Apple M1 Pro (8 cores in parallel, the combination numbers of the tested hyperparameters are therefore multiples of 8), RAM: 17,2 GB.

are also cars that were already available before March 2022 but had not yet been sold by the end of the data collection period, we estimated the price of unsold cars using a model that was trained on the last subsample of sold cars.

4.4 Survival analysis

After predicting the prices of cars by supervised learning, we computed the price-to-value ratios, which—along with several car characteristics—can be used to estimate selling times by the Cox proportional hazards model with LASSO regularization. As tuning the penalty hyperparameter λ by k-fold cross-validation is a computationally demanding process, we only used a random sample of 10,000 observations for this purpose, but later estimated the best Cox model on all observations. After variable selection, we fitted the unregularized Cox model with the terms selected by LASSO. As an alternative, we also trained a more flexible RSF model using the `ranger` package in R with its default hyperparameter settings and 50 decision trees, as preliminary experiments showed that predictive performance had already stabilized at this value, while larger forests substantially increased the computational burden with negligible improvements in predictive accuracy. Additionally, as a sensitivity analysis, we reestimated both models after removing the bottom and top 5–5% of price-to-value ratios (i.e., heavily under- or overpriced cars), and using the raw offer price instead of the price-to-value ratio.

The predictive performance of the survival models was evaluated independently of the train–validation–test procedure used for the supervised learning models. For the Cox specifications, LASSO variable selection was first performed on a random subsample of 10,000 observations using cross-validation with Harrell's concordance index as the selection criterion. The selected covariates were then re-estimated with an unregularized Cox proportional hazards model on the full relevant sample. Out-of-sample concordance was assessed by five-fold cross-validation: in each fold, the Cox model was fitted on four fifths of the data and evaluated on the held-out fifth, after which Harrell's concordance was computed from the pooled out-of-fold linear predictors. For the random survival forest specifications, we used the out-of-bag prediction error from the implementation in the `ranger` R package; the reported concordance is therefore calculated as one minus the out-of-bag prediction error.

5 Results

5.1 Unsupervised learning

We kept all factors with eigenvalues greater than the average of the eigenvalues. This yielded 236 factors out of more than 600 binary variables.

To facilitate the interpretation of the latent dimensions arising from MCA, we examined the category loadings on the first two principal components. Figure 4 displays the dummy variables with the largest absolute loadings in the two-dimensional loading space, allowing the dominant contributors to each principal component to be identified and interpreted.

Figure 4 indicates that the first dimension primarily captures a modern-equipment profile. Positive values on this dimension are associated with comfort and driver-assistance features such as adaptive cruise control, reversing camera, keyless start, parking radar, traffic sign recognition, and electrically folding exterior mirrors. Hence, this dimension separates better-equipped, more technologically advanced vehicles from cars without this equipment profile. The second dimension captures a different source of heterogeneity: the absence of basic safety and convenience features. Higher values on this axis are associated with missing airbags, ABS, adjustable steering, central locking, power steering, front electric windows, and seat-height adjustment.

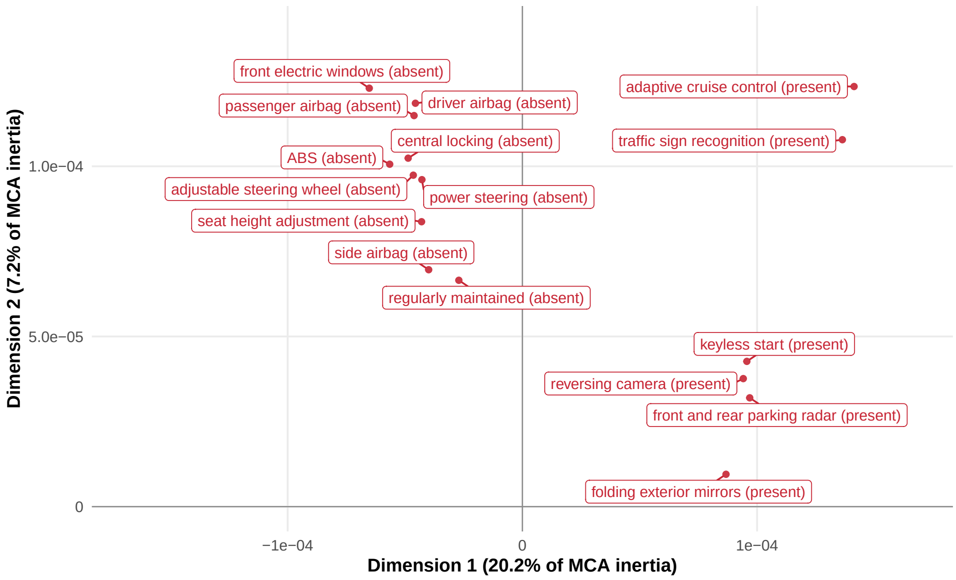

Fig. 4 Loading plot for the first two retained dimensions. Red points and labels mark the highest-contributing category levels for the first two MCA dimensions.

5.2 Supervised learning

The predictive performance of the models used to estimate market-consistent asking prices is visualized in Figure 5, while the predictive performance on the validation set for each combination of hyperparameters is shown in more detail in Appendix B.

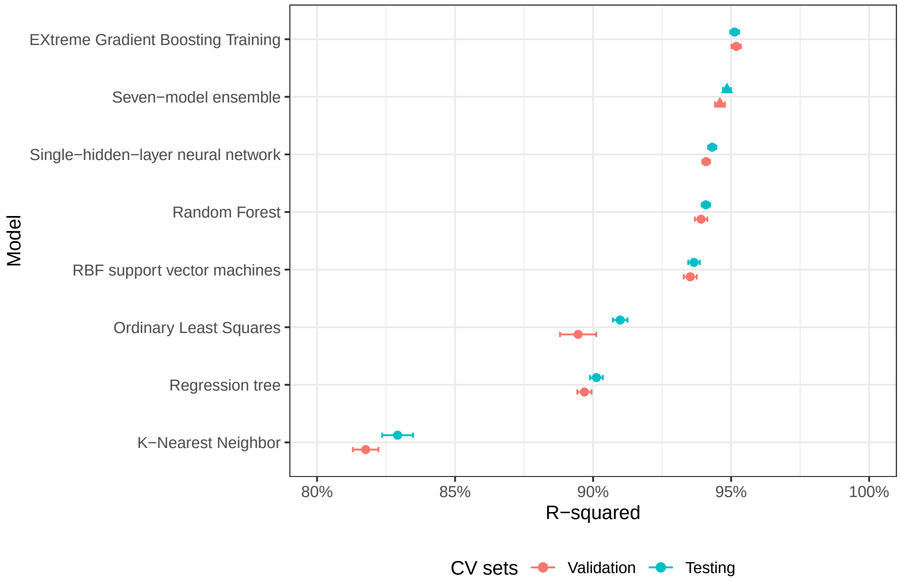

Fig. 5 Comparing the estimation accuracy of the investigated models, including the seven-model ensemble. Points report mean fold-level R^2 values and horizontal intervals report one standard error across folds.

Figure 5 demonstrates that XGBoost has the best predictive capacity on the held-out testing folds. The seven-model ensemble performs very similarly and ranks second on the testing folds, but it does not improve on the best individual model. This suggests that the additional models do not add enough independent predictive information to offset the weaker performance of some ensemble members. Consequently, we retain XGBoost as the benchmark price prediction model. Its strong performance is consistent with previous evidence on gradient boosting methods (Nguyen-Sy et al., 2020; Sheridan et al., 2016).

Using the predictions of the XGBoost model, we calculated price-to-value ratios. Table 4 reports the minimum, quartiles, median, and maximum of these. Half of the offer prices do not differ from the predicted values by more than 3%. However, the highest observed offer price exceeds the corresponding predicted market-consistent asking price by more than a factor of three based on the estimation. This value is still plausible, but it is a reasonable precaution to perform a robustness check on the survival model by omitting these outlier values.

5.3 Survival analysis

Tuning the Cox model with LASSO regularization using the one-standard-error criterion yielded the optimal regularization hyperparameter $\lambda = 0.0178$. Using this value, we obtained a model with a total of 55 predictors. We visualize the tuning process in Figure 6.

Table 4 Descriptive statistics of price-to-value ratios, computed as the proportions of offer prices to predicted values from XGBoost. To report in this table we subtracted 1 from the values.

Quarter	Min (%)	L. quartile (%)	Median (%)	H. quartile (%)	Max (%)
2021/04-2021/07	-48.13%	-2.83%	-0.08%	2.66%	61.95%
2021/07-2021/10	-51.17%	-2.69%	0.06%	2.87%	208.45%
2021/10-2022/01	-50.43%	-2.75%	0.07%	2.96%	198.91%
2022/01-2022/04	-47.48%	-2.63%	0.10%	2.89%	124.62%
Total	-51.17%	-2.72%	0.06%	2.89%	208.45%

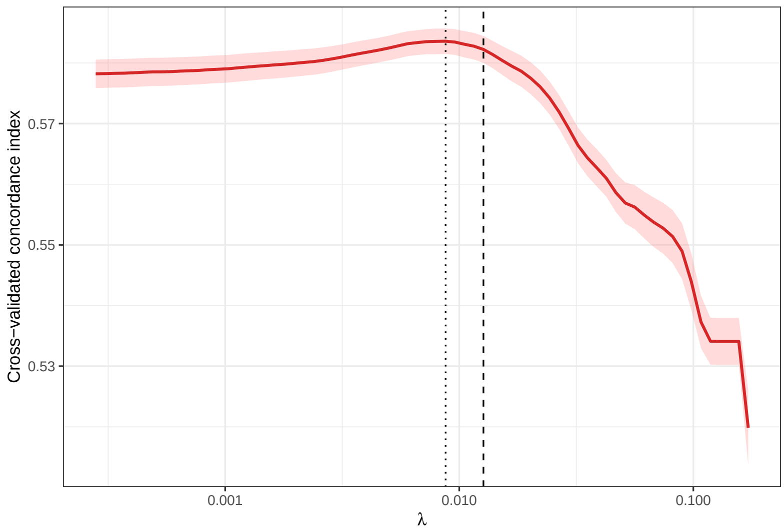

Fig. 6 10-fold cross-validation results for finding the optimal Cox LASSO hyperparameter λ. The dotted line corresponds to maximum cross-validated concordance, while the dashed line represents the popular one-standard-error heuristic.

To ensure robustness, we ran the feature selection algorithm and estimated the Cox model with and without dropping the cars in the bottom and top 5% of the price-to-value distribution. However, the results were almost identical, so the model is robust in this aspect. Figure 7 visualizes the estimated coefficients for both models. The figure shows only minor differences between them; while the same terms were selected by the regularized LASSO model, the confidence intervals of the estimated coefficients do not overlap for seven variables (year of manufacture, car set off, VAT refund status, price-to-value ratio, leather steering wheel, and lacking or having 3-year MOT validity). Furthermore, two minor covariates (Apple CarPlay and 2-year MOT validity) exhibit a sign switch, though their estimates remain close to zero. Overall, the coefficients remain highly consistent for the vast majority of predictors.

Some examples of how car characteristics are related to selling times, whose interpretation may require expert knowledge:

- American models remain associated with significantly longer expected selling times (exhibiting one of the largest negative coefficients in the model), assuming all else

Fig. 7 Estimated coefficients of the right-censored Cox model and the restricted price-to-value-range robustness model, with confidence intervals.

is unchanged. The likely reason is that American cars are produced with different standards, and it is harder to find mechanical parts in Hungary. Thus they sell slower because of the smaller pool of customers.

- Similarly, exhibited cars (demonstration vehicles) tend to sell much slower, ceteris paribus, again reflected by a strongly negative coefficient.

- Amplifier output correlates with significantly shorter expected selling times, having the largest positive coefficient in the model. However, the causal effect of this variable is questionable, as it is likely correlated with other characteristics attractive to young customers (stickers, lighting, etc.).
- While variables like 'Cylinder type: Other' appear relevant in unregularized explorations, the LASSO algorithm drops them entirely, indicating that their variation is spurious noise or already captured by other characteristics. Conversely, the model reveals that cars where the VAT is refundable tend to sell significantly slower, likely because they target a narrower, more rigid corporate market.

Based on the absolute values of the standardised coefficients, the price-to-value ratio emerges as a strong predictor of selling time. Its sign aligns with our prior expectations: holding all else constant, a higher price-to-value ratio is consistently associated with longer selling times. Notably, this effect is stronger in the full sample than within the restricted inter-ventile range of common selling prices. This difference is economically plausible: extreme price deviations (e.g., heavily overpriced vehicles) may belong to professional sellers who apply larger margins and are less motivated by immediate liquidity. Their resulting exceptionally long selling times amplify the aggregate effect, explaining why the coefficient becomes smaller in absolute terms once these extremes are excluded.

The predicted responses associated with changes in the price-to-value ratio are visualized in Figure 8. For prices below or equal to the predicted market value, both the Cox and RSF specifications exhibit a similar pattern, with median selling times typically falling between 12 and 13 days. However, for higher prices, the RSF captures a strong nonlinearity that the Cox model smooths over due to its rigid parametric nature. When a vehicle is significantly overpriced, its median selling time increases steeply with the price-to-value ratio. This reflects a sound economic reality: professional dealerships often incorporate an extra commission margin of 10% or more into their prices, and they do not necessarily want or need to sell quickly if it means sacrificing that premium.

In summary, the results of this case study reveal that: (1) XGBoost outperforms other machine learning methods in estimating the price-to-value ratio, (2) which is a strong predictor of the time to sell, and (3) some other features of cars are associated with selling times, such as whether they were American brands, exhibit cars, or had amplifier outputs. The model enables dealers and other market stakeholders to estimate how long it will take to sell a given car for a given price.

5.4 Sensitivity analysis

To assess the robustness of the proposed workflow to alternative data processing and modeling choices, we performed a sensitivity analysis considering the following specifications:

- Censoring:
  - right-censored: listings that remained active at the end of the observation period are treated as right-censored observations.

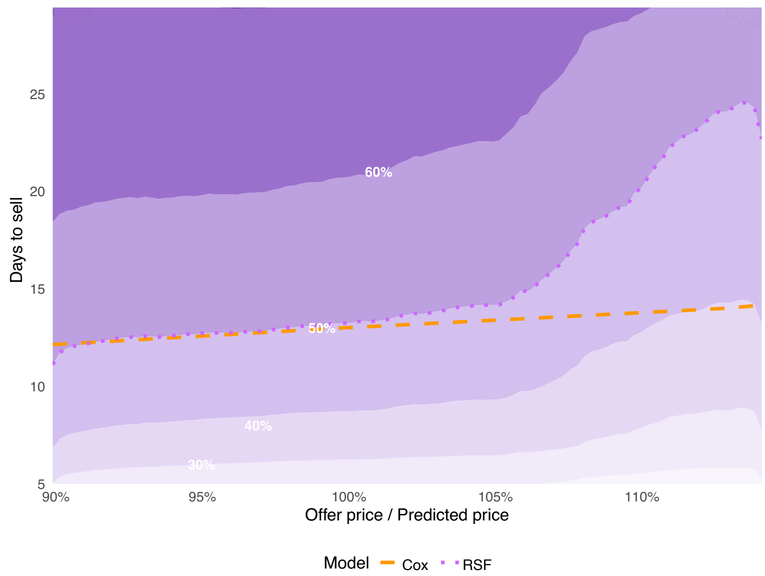

Fig. 8 Partial dependence of selling times on the price-to-value ratio in the Cox and RSF models using a 200-vehicle random reference sample. Solid contours correspond to percentiles of time-to-sell in the more accurate RSF model, while the dashed and dotted lines denote the median selling times under the two specifications.

  - sold only: only listings that disappeared during the observation period are retained.
  - right-censored (excl. reposts): right-censored after removing repeatedly reposted listings.
  - sold only (fixed price): sold only, after excluding listings whose advertised asking prices changed during the observation period.

- Selling time:
  - end–start: elapsed time between the first and last observed listing dates.
  - days on site: total number of days the listing is observed as active, excluding temporary removals.
- Offer price:
  - final
  - initial
  - time-weighted average
  - average: average of final and initial offer prices
- Imputation:
  - kNN
  - median/mode
  - MICE

- Price prediction model:
  - XGBoost
  - ensemble
- Time adjustment in price prediction:
  - rolling window
  - quarterly dummies
  - monthly dummies
- Price variable:
  - price-to-value ratio
  - price-to-value ratio (5–95%): excluding the bottom and top 5% of observations based on the price-to-value ratio before estimating the survival models.
  - offer price: advertised asking price used directly, without the price prediction stage.
- Survival model:
  - Cox: Cox proportional hazards model
  - RSF: Random Survival Forest

Unless otherwise stated, the baseline specification uses right-censoring, the end–start measure of listing duration, the final observed offer price, kNN imputation, XGBoost estimated on rolling time windows, and the price-to-value ratio as a predictor.

The numerical results in Table 5 show that the main findings are robust to alternative assumptions concerning censoring, the measurement of selling time and offer prices, imputation, temporal adjustment, price prediction, and survival modeling. Across all Cox specifications, the hazard ratio of the standardized price covariate remains below one. Thus, whether relative pricing is measured using the baseline specification or one of the alternative data-processing choices, higher prices relative to the estimated market-consistent asking price are consistently associated with a lower delisting hazard and therefore longer expected listing durations. The price prediction models are similarly stable: where reported, their R^2 values range only from 0.952 to 0.958.

The RSF specifications achieve higher Harrell's concordance indices than their directly comparable Cox counterparts. Using the price-to-value ratio, the concordance index increases from 0.593 in the baseline Cox model to 0.626 in the RSF model. The same pattern is observed when the raw offer price is used directly, for which concordance rises from 0.584 under Cox to 0.613 under RSF, and after restricting the price-to-value ratio to its 5th–95th percentile range, for which it rises from 0.583 to 0.599. This consistent advantage is plausibly attributable to the greater flexibility of RSF, which can accommodate nonlinear effects, as observed in Figure 8, without imposing proportional hazards. This is particularly relevant because the proportional hazards assumption is rejected in every Cox specification: the tests based on Schoenfeld

Table 5 Results of the sensitivity analysis, including deviations from the baseline specification, sample sizes in the price prediction and survival models, numbers of delisting events, R^2 in the price prediction models, e^{β} hazard ratios of the standardized price covariate and Harrell's concordance indices in the survival models.

Model	**Specification change**	**Prediction** N	**Survival** N	**Events**	R^2	**HR**	**Conc.**
Baseline	–	563,479	512,965	471,169	0.952	0.836	0.593
Censoring							
C1	sold only	563,479	460,187	460,187	0.952	0.941	0.576
C2	right-censored (excl. reposts)	563,479	500,750	461,581	0.952	0.837	0.592
C3	sold only (fixed price)	563,479	422,857	393,466	0.952	0.820	0.588
Selling time							
D1	days on site	563,479	512,965	471,169	0.952	0.837	0.595
Offer price							
P1	initial	563,479	510,724	469,277	0.952	0.794	0.604
P2	average	563,479	512,965	471,169	0.952	0.809	0.596
P3	time-weighted average	561,338	535,429	494,029	0.955	0.811	0.599
Imputation							
I1	median/mode	563,479	531,880	490,084	0.956	0.890	0.589
I2	MICE	563,479	531,880	490,084	0.958	0.892	0.589
Time adjustment in price prediction							
T1	quarterly dummies	563,479	537,941	496,145	0.955	0.891	0.586
T2	monthly dummies	563,479	537,941	496,145	0.955	0.890	0.586
Price prediction model							
E1	ensemble	563,479	537,941	496,145	0.949	0.891	0.586
Price variable							
V1	offer price	563,479	512,965	471,169	–	0.884	0.584
V2	price-to-value ratio (5–95%)	563,479	461,667	442,664	0.952	0.926	0.583
Survival model							
S1	RSF	563,479	512,965	471,169	0.952	–	0.626
Mixed specifications							
M1	time-weighted average + sold only	561,338	483,109	483,109	0.955	0.877	0.579
M2	offer price + RSF	563,479	512,965	471,169	–	–	0.613
M3	price-to-value ratio (5–95%) + RSF	563,479	461,667	442,664	0.952	–	0.599

residuals return numerically $p = 0$ both for all covariates jointly and for the price-to-value ratio separately. The Cox estimates therefore remain useful for summarizing the direction and approximate magnitude of the association in an interpretable form,[10] but the RSF results provide more reliable predictive discrimination when the objective is to estimate expected listing durations.

The comparison between the price-to-value ratio and the raw offer price also supports the usefulness of the initial price prediction stage. In the Cox models, replacing the price-to-value ratio with the raw offer price reduces Harrell's concordance index from 0.593 to 0.584. In the more flexible RSF models, the corresponding reduction is from 0.626 to 0.613. Hence, the predictive advantage of relative pricing is not specific to the Cox functional form. By first estimating a market-consistent asking price from vehicle characteristics, the price-to-value ratio distinguishes an expensive vehicle from a vehicle that is expensive relative to comparable alternatives. The resulting improvement in concordance indicates that this normalization contains information

[10]Because the assumption is rejected, the Cox hazard ratios should be interpreted as average associations over the observed listing duration rather than as effects that remain constant over time.

about liquidity that is not captured by the advertised asking price alone, thereby supporting our central contribution of combining machine-learning-based valuation with survival analysis.

The deviations in the Cox hazard ratios are also economically and methodologically plausible. Excluding repeatedly reposted listings has virtually no effect: the hazard ratio changes only from 0.836 to 0.837, while concordance remains almost unchanged at 0.592. Likewise, replacing elapsed end–start duration with days on site produces an HR of 0.837 and concordance of 0.595, indicating that temporary listing removals do not materially affect the main relationship. By contrast, restricting the analysis to sold listings attenuates the estimated relationship, with the HR moving to 0.941. This is expected because the restriction removes listings that remain active at the end of the observation period, including potentially overpriced vehicles with especially long durations, and therefore removes part of the variation that identifies the association between relative price and liquidity. The associated decline in concordance to 0.576 is also consistent with the loss of informative censored observations.

Alternative definitions of the offer price yield somewhat stronger estimated associations than the final-price baseline. The HRs are 0.794 for the initial offer price, 0.809 for the average of the initial and final prices, and 0.811 for the time-weighted average. This pattern is logical because the final observed price may already reflect price reductions made in response to a listing's failure to sell. Initial and time-averaged prices more closely represent the asking prices prevailing over the listing period and may therefore capture the relationship between the seller's pricing decision and listing duration more clearly. Their concordance indices, ranging from 0.596 to 0.604, are nevertheless close to the baseline value, indicating that their predictive power does not depend on a particular definition of the observed offer price.

The alternative imputation procedures and temporal adjustments produce somewhat more attenuated HRs, between 0.890 and 0.892, but do not alter the direction of the relationship. Median/mode imputation and MICE both yield a concordance index of 0.589, while replacing rolling windows with quarterly or monthly dummies produces concordance indices of 0.586. The latter result is consistent with the purpose of the rolling-window approach: a single global model with period indicators can account for common shifts in price levels, but it is less able to accommodate changes over time in the implicit values of individual vehicle characteristics. Nevertheless, the persistence of HRs substantially below one under all these alternatives indicates that the main finding is not driven by the baseline imputation method or temporal-adjustment procedure.

Excluding the bottom and top 5% of price-to-value ratios attenuates the Cox HR from 0.836 to 0.926 and lowers concordance from 0.593 to 0.583. This seems to be a logical consequence of removing the listings with the most pronounced underpricing or overpricing, where the relationship between relative price and selling time is strongest and most nonlinear, although the hazard ratios cannot be directly compared due to the different standard deviations used for standardization. The association nevertheless remains in the expected direction in the restricted sample, while the corresponding RSF model continues to outperform Cox.

Overall, the sensitivity analysis indicates that the exact numerical magnitude of the estimated association varies in predictable ways with the information retained by each specification, and its direction, economic interpretation, and predictive relevance remain robust.

6 Limitations

We acknowledge several limitations of the present study. First, following academic and industry practice, we treat delisting as a proxy for completed sales, but it does not necessarily imply that the vehicle was successfully sold. Listings may also be removed because the seller withdrew the vehicle, sold it through another channel, or chose not to continue advertising. Consequently, the observed selling times should be interpreted as the duration until delisting rather than confirmed transaction times. The analysis cannot account for repeated listings of the same vehicle or sales completed after switching to another online platform. Such cases may lead to measurement error in the estimated selling times and introduce additional uncertainty regarding the true duration of the selling process.

Second, the analysis relies on advertised asking prices rather than realized transaction prices. Asking prices may differ from final sale prices due to negotiations between buyers and sellers. While asking prices are the most consistently available information in online marketplace data, access to transaction prices would enable a more precise assessment of pricing behaviour and market liquidity.

Third, we do not perform a comprehensive benchmark comparison against all plausible valuation and survival modelling approaches. Although the supervised learning and survival models employed in this study represent strong and widely used methods for their respective tasks, alternative approaches may exhibit different predictive accuracy, computational requirements, or levels of interpretability. Consequently, the present paper should be viewed as demonstrating the feasibility of the proposed integrated workflow rather than establishing its superiority over all competing methodologies. A systematic benchmarking study constitutes an important direction for future research.

Fourth, the empirical analysis is based on data from the Hungarian used car market and one online marketplace. Although the proposed workflow is designed to be transferable to other online marketplaces for heterogeneous goods, the empirical findings may not generalize directly to markets with different institutional settings, consumer behaviour, platform characteristics, or pricing mechanisms. Applying the workflow to additional product categories, countries, and marketplace platforms would provide further evidence regarding its robustness and external validity.

Fifth, the market is not stable over time, and both the overall price level and the relationship between vehicle characteristics and prices may evolve during the observation period. We addressed this by estimating the supervised learning models on rolling time windows, which allows the pricing function to adapt over time. As our dataset covers only ten months, however, it is not possible to separately identify macroeconomic trends from the one-off effects of the COVID-19 pandemic (McKinsey & Company, 2023).

7 Conclusion

In this paper, we have demonstrated an integrated empirical workflow for assessing liquidity in online marketplaces of heterogeneous goods. The workflow combines automatic data collection, unsupervised machine learning for dimensionality reduction, supervised machine learning for price prediction, random survival forests, and Cox proportional hazards models with and without LASSO regularization to analyse how price-to-value ratios and product characteristics are related to selling times.

This integrated workflow combines established approaches to price estimation and survival analysis within a single empirical framework. While prior studies typically examined either price estimation or time-to-sale separately, our approach illustrates how these complementary techniques can be used together to provide a joint view of value formation and market liquidity. Because it relies on data that are commonly available in online marketplaces, the workflow can be adapted to different product categories and platform types.

From an applied perspective, the framework has several tangible use cases. Sellers can utilize it to optimize pricing strategies by first obtaining the estimated market-consistent asking price for a vehicle and then evaluating alternative candidate asking prices through their corresponding price-to-value ratios. The survival model provides the predicted listing duration associated with each candidate ratio, enabling sellers to compare alternative pricing strategies before selecting an asking price. Platform operators can embed the model into their interfaces to display real-time liquidity indicators, design liquidity-aware recommendation systems, or support dynamic pricing mechanisms.

Based on our proposed workflow, it is also possible to aggregate individual predictions to form a market liquidity index—such as the median predicted time-to-sale per month—to track liquidity trends over time and assess how they respond to macroeconomic or platform-specific factors.

In summary, this paper demonstrates the practical value of integrating machine learning-based valuation with survival analysis for studying liquidity in online marketplaces. It shows how established techniques can be combined into a coherent and transferable empirical workflow. By linking price formation with time-to-sale, the proposed framework provides a practical foundation for analyzing market behaviour and supporting decision making by sellers, platform designers, and policymakers alike.

References

Abdi, H. & Valentin, D. (2007). Multiple correspondence analysis. *Encyclopedia of measurement and statistics*, *2*(4), 651–657.

Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Akyüz, S., Eygi Erdoğan, B., Yıldız, O., & Karadayı Ataş, P. (2023). A novel hybrid house price prediction model. *Computational Economics*, *62*, 1215–1232, https://doi.org/10.1007/s10614-022-10298-8.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56, https://doi.org/10.1016/S1386-4181(01)00024-6.

Beretta, L. & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(Suppl 3), 74, https://doi.org/10.1186/s12911-016-0318-z.

Bukvić, L., Škrinjar, J. P., Fratrović, T., & Abramović, B. (2022). Price prediction and classification of used vehicles using supervised machine learning. *Sustainability*, *14*(24), 17034, https://doi.org/10.3390/su142417034.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762, https://doi.org/10.1016/j.ijforecast.2015.12.005.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox Automotive Market Insights (2023). Dealer delisted price index. Accessed: March 2025 https://www.coxautoinc.com.

Demiriz, A. (2018). Used car pricing and beyond: A survival analysis framework. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)* (pp. 65–68).

Dutulescu, A., Catruna, A., Ruseti, S., Iorga, D., Ghita, V., Neagu, L.-M., & Dascalu, M. (2023). Car price quotes driven by data: Comprehensive predictions grounded in deep learning techniques. *Electronics*, *12*(14), 3083, https://doi.org/10.3390/electronics12143083.

Fisher, L. D. & Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*, *20*(1), 145–157.

Grambsch, P. M. & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

Hungarian Central Statistical Office (2020). *Statistical Yearbook of Hungary, 2019*. Budapest: Hungarian Central Statistical Office.

Hungarian Central Statistical Office (2022). Consumer Price Indices by COICOP classification. Accessed: March 2025 https://www.ksh.hu/stadat_files/ara/en/ara0005.html.

Hungary Today (2024). German brands dominate the domestic used car market. Accessed: March 2025 https://hungarytoday.hu.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841–860, https://doi.org/10.1214/08-AOAS169.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Jin, C. (2021). Price prediction of used cars using machine learning. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)* (pp. 223–230).

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Liao, S. G., et al. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*, *15*(1), 346, https://doi.org/10.1186/s12859-014-0346-6.

Lippman, S. A. & McCall, J. J. (1986). An operational measure of liquidity. *American Economic Review*, *76*(1), 43–55.

Madhusudhanan, K., Behrens, G., Stubbemann, M., & Schmidt-Thieme, L. (2024). ProbSAINT: Probabilistic Tabular Regression for Used Car Pricing. *arXiv preprint arXiv:2403.03812* https://arxiv.org/abs/2403.03812.

McKinsey & Company (2023). Pumping the brakes on used-car prices https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/charts/pumping-the-brakes-on-used-car-prices.

Nguyen-Sy, T., Wakim, J., To, Q.-D., Vu, M.-N., Nguyen, T.-D., & Nguyen, T.-T. (2020). Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method. *Construction and Building Materials*, *260*, 119757.

Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S. S. (2017). How much is my car worth? a methodology for predicting used cars prices using random forest. *arXiv preprint arXiv:1711.06970*.

Pangallo, M. & Loberto, M. (2018). Home is where the ad is: online interest proxies housing demand. *EPJ Data Science*, *7*(1), 47.

Portfolio.hu (2024). Berobbanhat a használtautó-piac a következő hónapokban. Accessed: March 2025 https://portfolio.hu.

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *International Journal of Information and Computation Technology*, *4*(7), 753–764.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34–55, https://doi.org/10.1086/260169.

Saini, P. S. & Rani, L. (2023). Performance evaluation of popular machine learning models for used car price prediction. In *Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023* (pp. 577–588).: Springer.

Schmid, M., Wright, M. N., & Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. *Expert Systems with Applications*, *63*, 450–459.

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, *56*(12), 2353–2360.

SimilarWeb (2025). Használtautó.hu traffic analytics market share. Accessed: March 2025 https://www.similarweb.com.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5), 1–13 https://www.jstatsoft.org/v39/i05/.

Smith, J. & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355, https://doi.org/10.1111/j.1468-0084.2008.00541.x.

Sterne, J. A. C., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393, https://doi.org/10.1136/bmj.b2393.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC, 2 edition.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67, https://doi.org/10.18637/jss.v045.i03.

Venables, W. N. & Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. New York: Springer, 3 edition.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review*, *57*, 65, https://doi.org/10.1007/s10462-023-10681-3.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine*, *6*(7).

Çılgın, C. & Gökçen, H. (2024). A hybrid machine learning model architecture with clustering analysis and stacking ensemble for real estate price prediction. *Computational Economics*, https://doi.org/10.1007/s10614-024-10703-4.

Statements and declarations

Funding

This research was funded by the János Bolyai Research Scholarship and the National Program titled "Ageing: A New Challenge for the Economy and Society" (project code: NP2025-IDGT-3/2025) of the Hungarian Academy of Sciences.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

The authors contributed equally to this work.

Availability of data and materials

The data and code used in the analysis are available from the corresponding author upon reasonable request.

A Download dates

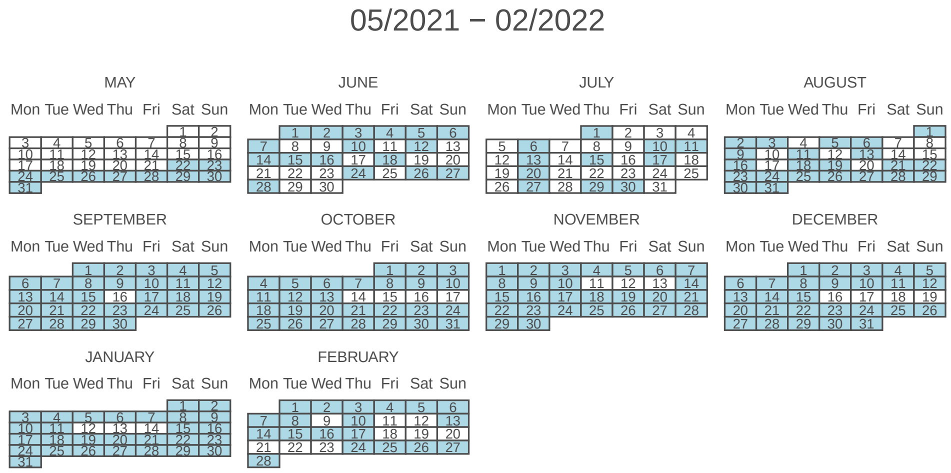

Fig. 9 Download dates (highlighted in blue). Starting in mid-August 2021, we were able to run the data download code on a personal server, so the daily download became more regular from then on.

B Hyperparameter tuning

Table 6 Top 10 hyperparameter combinations for eXtreme Gradient Boosting.

Trees	Min N	Tree Depth	Learn Rate	Loss Reduction	Sample Size	Stop Iter	Average R-Squared
1746	16	6	0.0258	0	0.7102	14	0.9523
1211	4	10	0.0074	0	0.315	20	0.9517
1779	23	13	0.0528	0.0099	0.6565	13	0.9506
1593	8	9	0.0082	0.0026	0.8579	12	0.9505
869	12	14	0.006	0.0003	0.4159	5	0.9504
1040	37	8	0.0113	0	0.4688	17	0.95
1634	15	4	0.0823	0	0.4631	5	0.9485
653	34	9	0.0185	0.0014	0.1692	14	0.9459
595	17	3	0.05	0.0001	0.8412	16	0.9454
1272	6	15	0.0098	0	0.951	9	0.9448

Table 7 Top 10 hyperparameter combinations for single-hidden-layer Neural Networks.

Hidden Units	Penalty	Epochs	Average R-Squared
5	0.8302	644	0.9404
5	0.5594	680	0.9403
4	0.3096	851	0.9401
5	0.1741	471	0.936
4	0.0202	675	0.9359
4	0.017	721	0.9353
5	0.379	222	0.9329
5	0	592	0.9321
6	0.001	510	0.9312
5	0	405	0.9309

Table 8 Top 10 hyperparameter combinations for K-Nearest Neighbors.

Neighbors	Weight Func	Dist Power	Average R-Squared
2	rank	0.1465	0.8243
13	Gaussian	0.2589	0.8222
9	rank	0.3331	0.8205
11	cos	1.6533	0.8194
13	optimal	1.4462	0.817
11	optimal	1.8215	0.8164
8	triangular	1.865	0.8163
12	cos	0.8301	0.8161
14	rectangular	0.1868	0.8132
9	inv	1.5549	0.8132

Table 9 Top 10 hyperparameter combinations for Random Forests.

Mtry	Min N	Average R-Squared
42	6	0.9404
35	3	0.9404
52	8	0.94
75	3	0.9398
47	16	0.9397
65	14	0.9396
28	11	0.9394
72	12	0.9393
58	19	0.9393
41	22	0.9391

Table 10 Top 10 hyperparameter combinations for RBF Support Vector Machines.

Cost	Rbf Sigma	Margin	Average R-Squared
9.583	0.0025	0.1218	0.9347
13.3887	0.0018	0.159	0.9345
3.8105	0.0035	0.1333	0.9316
1.8583	0.0006	0.1903	0.9152
1.6808	0.0088	0.1661	0.9145
3.2048	0.0004	0.0201	0.9129
16.6023	0.0002	0.0103	0.9121
6.1335	0.0094	0.1731	0.9114
8.9552	0.0002	0.0002	0.9103
5.4173	0.0001	0.1546	0.9063

Table 11 Top 10 hyperparameter combinations for Regression Trees.

Cost Complexity	Tree Depth	Min N	Average R-Squared
0	10	39	0.8958
0	8	25	0.8954
0	9	34	0.8954
0	9	28	0.8952
0	8	36	0.8952
0	10	25	0.8949
0	10	21	0.8945
0.0002	9	24	0.8938
0	7	21	0.8925
0	11	27	0.892