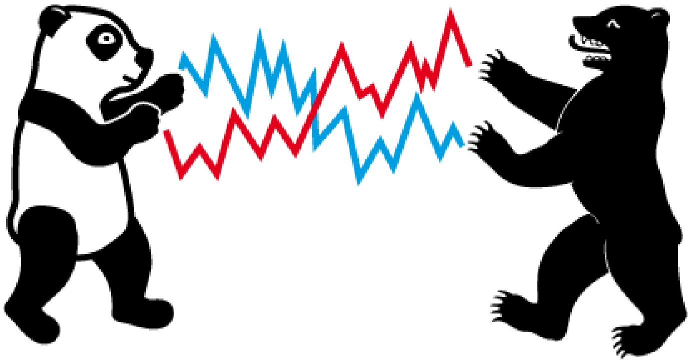
IRTG 1792 Discussion Paper 2018-065

Price Management in the Used-Car Market: An Evaluation of Survival Analysis

Alexander Born [*]
Nikoleta Kovachka [*]
Stefan Lessmann [*]
Hsin-Vonn Seow [*2]



[*] Humboldt-Universität Berlin
[*2] University of Nottingham-Malaysia Campus

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Price Management in the Used-Car Market: An Evaluation of Survival Analysis

Alexander Born[a], Nikoleta Kovachka[a], Stefan Lessmann[a], Hsin-Vonn Seow[b]

[a]*School of Business and Economics, Humboldt-University of Berlin, Unter-den-Linden 6, 10099 Berlin, Germany*

[b]*Nottingham University Business School,University of Nottingham-Malaysia Campus,Jalan Broga,43500 Semenyih, Selangor Darul Ehsan, Malaysia*

Abstract

Second-hand car markets contribute to billions of Euro turnover each year but hardly generate profit for used car dealers. The paper examines the potential of sophisticated data-driven pricing systems to enhance supplier-side decision-making and escape the zero-profit-trap. Profit maximization requires an accurate understanding of demand. The paper identifies factors that characterize consumer demand and proposes a framework to estimate demand functions using survival analysis. Empirical analysis of a large data set of daily used car sales between 2008 to 2012 confirm the merit of the new factors. Observed results also show the value of survival analysis to explain and predict demand. Random survival forest emerges as the most suitable vehicle to develop price response functions as input for a dynamic pricing system.

Keywords: Automotive Industry, Price Optimization, Survival Analysis, Dynamic Pricing

1 Introduction

The paper focuses on management processes in the second-hand car market and develops analytical models to support decision-making in marketing and sales. The automotive sector is a multi-billion dollar industry and a guarantor of growth and wealth in many economies. Operating in increasingly saturated consumer markets, car makers must actively manage and continuously improve business processes concerning the handling of used cars. The strategic importance of the second hand car segment follows from its direct connection to the new car business (Prado, 2010). Selling a new car typically involves the (re-)purchase of a used vehicle. In the leasing business, vendors even face a legal obligation to repossess a vehicle after contract expiration. A large number of take-back obligations also arises from deals with car rental companies, which routinely refurbish their fleets (Desai and Purohit, 1998). Strong interdependence between the new and used car business is also reflected in sales figures. Considering the case of Germany, for example, sales revenues in the car market amounts 186.6 bn Euro, 44 percent of which are attributed to the second-hand market (DAT, 2017).

Low profit margins of about 1 percent (c.f. 8 percent for new cars) contrast the strategic importance of the used car business and represent a key management challenge for car makers and other automotive companies(DAT, 2017). Factors explaining the lack of profit are manifold and include manufacturing overcapacity, excessive supply, increasing discount levels and fierce competition (Jerenz, 2008). In terms of business optimization, specific challenges arise in the used car market. While big marketing campaigns together with a rich set of configuration options and individualization possibilities benefit the selling of new cars, no such measures are available in the used car segment. This constraints the set of management controls to raise profits. Given that the supply of used cars is driven by the new car business, due to retail trade-ins, repossessions, etc., and largely fixed, price is often the only steering mechanism available to improve margins (Du et al., 2009).

Much research has investigated the antecedents of price formation in the used car market; often using prices as cues to shed light on market structure and informational efficiency (Genesove, 1993; Emons and Sheldon, 2009), and how these are affected by the advent of digital channels such as online auctions (Adams et al., 2011). From a microeconomics perspective, price discrimination is a suitable strategy to extract consumer surplus and increase 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).

The overarching objective of the paper is to develop an approach to estimate PRFs using survival analysis. Survival analysis models event time distributions and is part of a larger family of statistical methods to analyze count data (Zhu et al., 2017). Having its origin in medical data analysis, survival analysis has gained popularity in management decision support to predict the probabilities of critical events in a customer relationship such as attrition or credit default (Tang et al., 2014; Dirick et al., 2017). Other management application include predictive maintenance and the modeling a supply-chain risks such as stockouts (Xishu et al., 2016). A common denominator in these applications is that an analyst is interested in the probability of event occurrence, how this probability evolves over time, and is affected by subject characteristics such as behavioral customer data. Survival analysis provides answers to these questions.

In this paper, we define the event of interest to be the sale of a used car. The prize that a seller offers enters the statistical model as an independent variable; other variables including, for example, car age and mileage, special equipments, etc. With this setup, survival analysis allows a decisionmaker to examine how the probability of selling a car evolves with price changes. In other words, the seller obtains a model-estimated PRF, which facilitates price - and eventually profit - optimization. For example, Jerenz (2008) proposes a dynamic programming formulation to identify the optimal amount and frequency of price updates for a used car dealer, and, by means of simulation, estimates his optimal pricing strategy to increase profits by 4.6 percent.

In this paper, we further elaborate on the use of survival analysis for PRF estimation and price optimization as introduced by Jerenz (2008). In this course, the paper contributes to literature in three ways. First, using a large real-world data set form an online marketplace, we show how classical parametric survival methods fail to capture the effect of car characteristics, and how this impedes the accuracy of model-estimated purchase probabilities. In a prize optimization context, inaccurate model predictions lead to suboptimal decisions and eventually diminish sales profit. We further show how a violation of model assumptions explains the inappropriateness of parametric survival analysis. Second, we introduce nonparametric survival methods to the field of PRF estimation. By design, these methods operate in a purely data-driven manner and do not depend on distributional assumptions. We elaborate on the mechanics of corresponding techniques and show how they estimate purchase probabilities more accurately than techniques previously used in price optimization. Third, the empirical results gained throughout predictive modeling of purchase probabilities using survival analysis also provide original explanatory insights concerning the factors that govern used car sales. To that end, we extend the set of features previously employed to model used car sales and, referring to dealership size as a proxy, identify the effect of marketing ability on sales probability. These contributions have important implications for managers in that they provide evidence for the effectiveness of the survival analysis framework in car reselling operations and concrete guidance how to devise a corresponding decision support model.

The remainder of the paper is organized as follows. Section 2 reviews related literature and derives research questions. Section 3 elaborates on survival analysis and nonparametric survival methods in particular. Section 4 introduces the data used in the empirical study, results of which are presented in Section 5. Section 6 concludes the paper with a discussion of the results and their implications.

2 Related literature and research questions

The focus of the paper implies that related works comes largely from two streams of literature, that on the second-hand car market and that on survival analysis. Given the scarcity of prior work on survival analysis for PRF estimation and applications in the automotive sector, Chapter 3 gives a comprehensive overview of survival analysis.

Recent comparative results from another domain are available in Dirick et al. (2017). To identify the research gap concerning decision support in car reselling, the review of related literature focuses on prior work on the used car market.

Since publication of seminal work by Akerlof (1970), this market has received much attention in economics. Market prices are an important cue in corresponding research as they signal the degree to which the market is informational efficient (Levin, 2001), exhibits information asymmetry (Belleflamme and Peitz, 2014), or shows signs of discrimination, (Ayres and Siegelman, 1995), amongst others. Given that digital channels such as online auctions significantly impact market ecology (Bapna et al., 2008), several studies have examined the effect of digital innovation through the lens of the second hand car market (Chen et al., 2013).

The large volume of the second-hand car market implies that it is also relevant from a managerial point of view. Olivares and Cachon (2009) offer valuable insights concerning the competitive disadvantage of large inventories, which they attribute to suboptimal pricing. Close connections to the new car business via retail trade-ins, lease returns, and repossessions from car rental companies (Desai and Purohit, 1998) further contribute to the criticality of pricing decisions (Ratchford and Srinivasan, 1993). To the best of our knowledge, only two studies have examined price optimization in the used car business. Considering the context of online auctions, Du et al. (2009) propose a decision support system to maximize the net auction profit of distributing vehicles on the basis of their estimated auction prices, asset carrying costs and business constraints. Jerenz (2008) embeds the pricing problem in a comprehensive revenue management system for used cars. The system encompasses three components consisting of a forecasting model to estimate residual values, a survival model to construct a PRF, which receives estimated residual values as input, and a dynamic program for determining the optimal pricing policy. The study of Jerenz (2008) is particularly relevant for this paper because it was the first and only application of survival analysis for used car price optimization.

Du et al. (2009) and Jerenz (2008) estimate vehicle prices and residual values, respectively, using ordinary least-squares regression. Subsequent work has shown that data-driven forecasting methods such as neural networks or regression tree ensembles provide significantly more accurate price predictions (Lessmann and Voss, 2017). This finding, through evidencing the potential of advanced data-driven models within price optimization frameworks, motivates the focal study. Available decision support systems entail a temporal modeling of market dynamics. Du et al. (2009) consider an auto-regressive time series model whereas Jerenz (2008) employs parametric survival models. The ramifications of revising the temporal modeling component in a used car price optimization framework and the degree to which the use of a powerful analytic model improves decision support has eluded research. Striving to close this research gap, we focus on the framework of

Jerenz (2008), and thus survival analysis, because it is not specific to auction-based used car sales and generally applicable. In this scope, we aim at understanding the underlying mechanisms of the second-hand car market to deliver insights for managers how to improve pricing decisions and eventually profits. We pursue this objective through proposing the following research questions, which we answer in the empirical part of the paper:

1. **Which factors influence the time a used car spends on the market before it gets sold?**
2. **What is the most accurate statistical method to predict the time on market?**

3 Survival Analysis and Methodology

Survival analysis focuses on an event of interest, such as machine's failure, and the time until this event occurs. The main purposes of survival analysis are estimation of survival and/or risk functions, comparisons of survival functions for different groups at risk and estimation of effects between survival time and external factors (Kleinbaum and Klein, 2006).

Survival analysis can deal with censored data. These are individuals who do not experience the event in the time frame of the underlying analysis. Right censoring occurs for individuals who are still alive at the end of the observed time frame. Left censoring occurs for individuals who have experienced the event before the beginning of the study.

For this study, the event of interest is a sale of a car at a specific point in time. Right censoring is equivalent to a car not being sold at the end of the observation time. Left censoring is not considered in this study.

3.1 Definitions

The survival function $S(t)$ is defined as the probability of an individual to survive past time *t*, where T is a continuous random variable

$$S(t) = Pr(T > t) \tag{1}$$

The lifetime distribution function is defined as the probability for an event to occur latest to time *t*

$$F(t) = Pr(T \leq t) = 1 - S(t) \tag{2}$$

For a differentiable F the event density is defined as

$$f(t) = F'(t) = \frac{d}{dt}F(t), \tag{3}$$

which transforms equation 1 and 2 to

$$S(t) = Pr(T \leq t) = \int_t^\infty f(u)du \tag{4}$$

$$F(t) = Pr(T < t) = \int_0^t f(u)du. \tag{5}$$

The instantaneous risk rate or hazard function is the probability for an event at specific time given $T \geq t$

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t * S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{6}$$

The cumulative hazard function is defined as

$$H(t) = \int_0^t h(u)du = \int_0^t -\frac{S'(u)}{S(u)}du = -\ln S(t) \tag{7}$$

and is interpreted as the number of expected events for each individual by time t in case of a repeatable process. For a continuous random variable T the interchangeability between the survival function and the cumulative hazard function is easily derived

$$S(t) = \exp[-H(t)] = \exp[-\int_0^t h(u)du] \tag{8}$$

3.2 Estimators for survival and hazard functions

We can build and estimator for $S(t)$ as a proportion of all survivors past time t and the total number of survivors. Due to truncation and censoring not all events happen in the observation period. This fact makes the intuitive estimation approach rather troublesome. The Kaplan-Meier estimator examines ascending, ordered event times t_i, the number of events d_i at time t_i and the total number of survivors n_i.

$$\hat{S}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} \tag{9}$$

The Kaplan-Meier estimator (Kaplan and Meier, 1958) is a non-parametric model and provides a step-wise non-increasing function. With equation (7) a Kaplan-Meier estimator for the hazard function $h(t)$ can be derived directly as

$$\hat{H}(t) = -\ln \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} \tag{10}$$

More common the Nelson-Aalen (Nelson, 1969, 2000; Aalen, 1978) estimator is used for the hazard function, defined as

$$\hat{H}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{11}$$

3.3 Cox Proportional Hazards Model

Our general introduction to survival analysis states that one of the goals is to describe effects of variables on the survival or hazard functions. Based on proportionality assumption (Cox, 1992)

$$h(t|X_i) = h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = h_0(t)\exp(X_i * \beta) \tag{12}$$

for the realized values of the variables $X_i = \{X_{i1}, \ldots, X_{ip}\}$ with coefficients $\beta = \beta_1, \ldots, \beta_p$ the effects of different variables X_i, X_j can be put in comparison

$$\frac{h(t|X_i)}{h(t|X_j)} = \frac{h_0(t)\exp(X_i * \beta)}{h_0(t)\exp(X_j * \beta)} = \frac{\exp(X_i * \beta)}{\exp(X_j * \beta)} \tag{13}$$

The coefficient β is estimated via the partial log-likelihood method. First consider the probability of a unique event at time t such that a time-event indicator $C_i = 1$ for the event occurred and $C_i = 0$ for censoring time is $C_i = 1$ and $Y_i = t$

$$L_i(\beta) = \frac{\exp(X_i * \beta)}{\sum_{j:Y_j \geq Y_i} \exp(X_j * \beta)} \tag{14}$$

Assuming statistical independence the joint distribution for all realized events is given by

$$L(\beta) = \prod_{i:C_i=1} \frac{\exp(X_i * \beta)}{\sum_{j:Y_j \geq Y_i} \exp(X_j * \beta)} \tag{15}$$

The partial log-likelihood is derived as

$$l(\beta) = \sum_{i:C_i=1} \Big(X_i * \beta - \log \sum_{j:Y_j \geq Y_i} \exp(X_j * \beta) \Big) \tag{16}$$

While giving an appealing framework, Cox proportional hazards model is based on several assumptions. The most obvious is the proportionality assumption of $h_0(t)$ for all observations. With regards to the different nature of different car models, like technical specifications, geographical location, size of the dealership and its marketing and negotiations abilities, the proportionality to the baseline hazard function is a very restrictive

assumption.
The derivation of partial log-likelihood reveals the log-linearity assumption in variables. This assumption might be violated for continuous variables but does not have the same emphasis on the model as proportionality.
Additionally, the assumption of non-informative censoring must hold (Ranganathan et al., 2012). The mechanism behind the censoring of individual observations should not be related to the probability of the event's occurrence. Inability to sell a vehicle for a specific price and resulting price reduction is considered as right-censoring in this study. Thus, non-informative censoring assumption might not always hold.

3.4 Survival trees

The assumptions of proportionality, log linearity and non-informative censoring, which are indispensable for the mathematical model, propose strict restrictions. To overcome these restrictions a modeling approach using survival trees may be used here. The concept of Classification And Regression Trees (CART) developed by Breiman (Breiman et al., 1984) sets the guidelines for the development of the survival tree framework (Bou-Hamad et al., 2011).
The two main parts of a tree algorithm are node splitting and stopping rule or pruning criteria. The first part is necessary for partitioning the variables space in smaller sub-partitions. The second part is necessary for the reduction of fully grown trees to prevent overfitting. For censored data there is no natural measure of node homogeneity. Thus, the impurity reduction splitting rule from CART algorithm is not directly applicable. Similar problems arise with regards to the definition of a natural loss function for censored data and the pruning part of a tree algorithm.
The first step is to construct a proper metric for the node splitting process (Crowley et al., 1995). Consider two random variable $X_1 \sim F_1$ and $X_2 \sim F_2$. The Wasserstein distance for two distributions can be described as

$$\left[\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p du \right]^{\frac{1}{p}} \tag{17}$$

For censored observations the integral has to be adjusted. Consider estimates $\hat{F}_1$ and $\hat{F}_2$ for F_1 and F_2 respectively.
Define

$$\lim_{u \to \infty} \hat{F}_1(u) = m_1 \leq 1$$
$$\lim_{u \to \infty} \hat{F}_2(u) = m_2 \leq 1$$

and assume further without loss of generality

$$m_1 \leq m_2$$

Set

$$m_3 = F_2^{-1}(m_1)$$

and define

$$\tilde{F}_2(u) = \begin{cases} \hat{F}_2(u) & \text{if } u < m_3 \\ m_1 & \text{if } u \geq m_3 \end{cases}$$

The L_p Wasserstein distance for censored data is then

$$\Big[\int_0^{m_3} |\hat{F}_1^{-1}(u) - \tilde{F}_2^{-1}(u)|^p du\Big]^{\frac{1}{p}} \tag{18}$$

In case of ordinary L_p metric given as

$$\Big[\int_0^{\infty} |F_1(u) - F_2(u)|^p du\Big]^{\frac{1}{p}}$$

define

$$m_4 = \min\Big(\hat{F}_1^{-1}(m_1), \hat{F}_2^{-1}(m_2)\Big)$$

and get the ordinary L^p metric for censored data

$$\Big[\int_0^{m_4} |\hat{F}_1^{-1}(u) - \hat{F}_2^{-1}(u)|^p du\Big]^{\frac{1}{p}} \tag{19}$$

With regards to a precise formulation and deep understanding of the survival tree methods a formal framework for the development of survival trees is given here.
Let U be the true survival time and C the true censoring time. Define $z = \min(U, C)$ as either event or censoring. Further define an indicator $\delta = I(U \leq C)$ with $\delta = 1$ corresponding to an event and $\delta = 0$ corresponding to a censoring. Define $x = (x_1, \ldots, x_p)$ a vector of variables. With n independent subjects the learning sample is defined as

$$\mathcal{L}_n = (z_i, \delta_i, x_i)$$

Further define $\hat{S}_t$ as an estimator of the survival function and $\hat{\delta}_{\hat{S}_t}$ as a step function. The reduction in impurity at a node t based on the learning sample$\mathcal{L}_n$ is given by

$$G(t) = p(t)d(\hat{S}_t, \hat{\delta}_{\hat{S}_t}) - [p(l(t))d(\hat{S}_{l(t)}, \hat{\delta}_{\hat{S}_{l(t)}}) + p(r(t))d(\hat{S}_{r(t)}, \hat{\delta}_{\hat{S}_{r(t)}})] \tag{20}$$

where

- $p(t)$ is the proportion of observations at node t
- $d(\hat{S}_t, \hat{\delta}_{\hat{S}_t})$ is the L^p Wasserstein distance
- $r(t)$ is the right child of the node t at split
- $l(t)$ is the left child of the node t at split

Based on the learning sample $\mathcal{L}_n$ and the splitting rule from equation (20) a tree $T(\mathcal{L}_n)$ can be constructed.
Practitioners in R have to resort to **rpart** package with the implementation of the splitting rule proposed by LeBlanc and Crowley thoroughly discussed above. Multiple further approaches on survival trees are introduced for within-node homogeneity as well as for between-node heterogeneity. Nevertheless, only a few have been implemented as widely used statistical routines.

3.5 Conditional inference trees

Based on the CART approach survival trees inherit its intrinsic drawbacks, overfitting and selection bias towards nodes with multiple possible splits or missing values. Application of pruning overcomes the first issue, but the selection bias still remains. The approach of growing trees derived from conditional inference is designed to overcome both issues. A recursive binary partitioning with a generic algorithm is performed in 3 steps (Hothorn et al., 2004, 2006a,b)

1. Test for conditional independence between the variables $X = X_1, \ldots, X_m$ and the response variable Y.
2. Perform a split based on predefined selection criteria.
3. Recursively repeat 1 and 2 until the hypothesis of conditional independence cannot be rejected.

In step 1 m partial hypotheses of independence for each variable X_j are postulated and combined to a global hypothesis.

$$H_0^j : D(Y|X_j) = D(Y)$$
$$H_0 = \cap_{j=1}^m H_0^j$$

with $D(Y|X)$ being the conditional distribution of the response variable Y given the variable X. With the learning sample $\mathcal{L}_n$ and the case weights w the association between the response variable and the j-th variable is measured by a linear statistic of the form

$$T_j(\mathcal{L}_n, w) = \text{vec}\Big(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, \ldots, Y_n))^T\Big) \in \mathbb{R}^{p,q} \tag{21}$$

where

- $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is a non-random transformation of the variable X_j
- $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ is an influence function depending on the responses $Y_1, \ldots, Y_n$ in a permutation symmetric way
- vec operator converts $p_j \times q$ matrix into a $p_j q$ column vector by column wise combination

The distribution of $T_j(\mathcal{L}_n, w)$ is usually unknown and permutation tests under H_0^j are used to reject the dependency on the joint distribution of Y and X_j. The splitting rule of the CART framework cannot be used in the step 2 with the original metric on censored data. Thus, the following splitting criteria is used

$$T_{j^*}^A(\mathcal{L}_n, w) = \text{vec}\Big(\sum_{i=1}^{n} w_i I(X_{j^*} \in A) h(Y_i, (Y_1, \ldots, Y_n))^T\Big) \in \mathbb{R}^q \tag{22}$$

where

- A is a possible split subset
- j^* denotes the variable index for the variable with the strongest association to Y deducted in step 1
- $I(\cdot)$ is the indicator function

Maximization of a test statistic over all possible subsets A leads to the optimal split

$$A^* = \underset{A}{\text{argmax}}\, c(t_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A) \tag{23}$$

where

- $\mu_{j^*}^A$ is the conditional expectation derived under the use of permutation tests
- $\Sigma_{j^*}^A)$ is the conditional covariance derived under the use of permutation tests

- $c : t \in \mathbb{R}^{pq} \to \mathbb{R}$ is a univariate test statistic mapping an observed multivariate linear statistic to $\mathbb{R}$

A pruning routine is obsolete as overfitting and excessive tree sizes are prevented through conditional independence test before each split. Conditional inference trees can be used on censored survival data without further considerations. Their predictive power is shown to be on par with trees from CART framework. For R practitioners the conditional inference trees are implemented in the **party** package.

3.6 Random Survival Forest

Random forest is one of the most popular machine learning algorithm with numerous applications for regression and classification. While the basic idea of random forest is assumed to be known by an interested reader several adjustments to random survival forest are given here (Ishwaran et al., 2008).

As previously discussed, the splitting rule during the tree growing phase has to be adjusted to explicitly involve survival time and censoring information. Random survival forest incorporates four splitting rules

- log-rank splitting rule (Segal, 1988; LeBlanc and Crowley, 1993)
- conservation-of-events splitting rule (Naftel et al., 1985)
- log-rank score rule (Hothorn and Lausen, 2003)
- random log-rank splitting rule (Hothorn and Lausen, 2003)

For a terminal node h and the $N(h)$ distinct event times $t_{1,h}, \ldots, t_{N(h),h}$ a cumulative hazard function (CHF) is defined as Nelson-Aalen estimator

$$H(t|x_i) = \hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}, \text{if } x_i \in h \tag{24}$$

where

- $d_{l,h}$ is the number of events at time $t_{l,h}$
- $Y_{l,h}$ is the number of individuals at risk at time $t_{l,h}$

With $I_{i,b}$ an indicator function for out-of-bag cases and $H_b^*(t|x_i)$ the CHF for a tree from b-th out of total B bootstrap sample the out-of-bag ensemble CHF is given by

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|, x_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{25}$$

and the bootstrap ensemble CHF for i is given by

$$H_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*(t|, x_i) \tag{26}$$

The prediction error is calculated on the out-of-bag data using the concordance index, where for pre-specified unique times $t_1^0, \ldots, t_m^0$ the case i has a worse predicted outcome than case j if

$$\sum_{l=1}^{m} H_e^{**}(t_l^0 t|x_i) > \sum_{l=1}^{m} H_e^{**}(t_l^0 t|x_j) \tag{27}$$

The concordance index (C-index) is constructed in such a way that the derived prediction error lies in a rage $[0, 1]$ with value of 0.5 equal to a random guess. The five-step, random survival forest algorithm is implemented in R in the **randomSurvivalForest** package.

3.7 Conditional Inference Forest

The methodology chapter is completed with a short remark on the ensembling technique for conditional inference trees (Hothorn et al., 2004, 2006a). The conditional inference forest grows trees in the way described in the subsection 3.6. The predictor for a new individual with variable X_{new} is the Kaplan-Meier estimator based on all observation from the learning sample $\mathcal{L}_n$ from the same leaf as X_{new}

$$\hat{S}_{\mathcal{L}_n}(\cdot|X_{new}) = \hat{S}_{\mathcal{L}_{n(X_{new})}}(\cdot)$$

The averaging technique differs from random survival forest as the weights are not equally distributed. Conditional inference forest assigns weights based on the total number of subjects at risk $Y_{l,h}$ for a given terminal node h. For R practitioners the conditional inference forest is implemented in the **party** package.

4 Data and feature engineering

4.1 Raw data

The raw data set comprises daily car prices from the 18th September 2008 till the 18th December 2012. Each price observation corresponds to an observed car with various attributes concerning make, model, performance characteristics and configuration. Each car observation also corresponds to a vendor, who is described through its location and a binary flag indicating whether the vendor is a professional dealer. The raw data contains 5,915,774 unique car IDs, 747,102 unique vendor IDs and 190,323,612 price observations.

All cars and price observations belong to different product lines of a premium car maker. The data has been gathered from *mobile.de*, a major online marketplace for used cars. An agreement between mobile.de and the focal car maker ensured compliance of data gathering. This agreement has also facilitated data collection via API programming, which benefits data quality compared to web scraping. To reduce the dimensionality of the data, we focus on seven main types of cars. Oldtimers and young cars have been deleted because their price dynamics differ from that of ”ordinary” used cars. In addition, we exclude private vendors from the data. On the one hand, this decision accounts for the ongoing trend of professional car dealerships dominating the market. On the other hand, professional dealers are the recipients of our research. We do not expect private dealers to use data-driven models to support pricing decisions. Concentrating on professional dealers also creates a more homogeneous and more comparable group of vendors.

As the data represents daily observations, each day a car spends on the market contributes one price observation. To obtain the standard data input format for survival analysis, we merge the multiple observations for a unique car (i.e., for different time points) into a single data point that represents the time in days a used car has been observed on the market. Restricting the data to seven car types and professional vendors, and merging over time for identical car IDs reduces the initial 190 millions of price observations to 4,875,850 observations.

4.2 Variables for survival analysis

We generate six variables from the raw data: time on market, market size, degree of overpricing, quantile, age, and size of dealership. The first five variables are based on Jerenz (2008) and used with minor adjustments. We propose size of dealership as a new variable to capture the marketing ability and image of a specific dealer. First, this study defines the target variable time on market (TOM). TOM is the difference in days between the first and the last date of online presence for the same car with the same price.

Market size (MS) represents the number of equivalent cars being offered simultaneously to the car of interest, where equivalent cars are from the same model and the same car category. Additional restrictions on age and mileage within the group ensure the same behavior for the cars under consideration. The age for equivalent cars lies within an interval of plus/minus three months. The mileage is in the interval plus/minus 10 000 km. Market size is intrinsically related to TOM, as vehicles with larger TOM tend to show higher market size. A more advanced definition of market size is part of future work.

Degree of overpricing (DOP) is the proportion between the price reported in the data and the hedonic price for the specific car, which we estimate using a log-linear regression on registration date, car performance, fuel type, age, type and category of the car.

$$\begin{aligned}\log(\text{Hedonic price}) = \beta_1\text{Registration} + \beta_2\text{Performance} + \beta_3\text{Fuel}+ \\ \beta_4\text{Age} + \beta_5\text{Type} + \beta_6\text{Category} + \epsilon\end{aligned} \tag{28}$$

The hedonic price represents the intrinsic value of a car defined by a set of variables (Lessmann and Voss, 2017). We then define DOP as

$$\text{DOP} = \frac{\text{Price in dataset}}{\exp\left(\text{Hedonic price}\right)} \tag{29}$$

Quantile represents the percentile of the car price for prices in the same car category. Age represents the age of a car in months. The novel variable size of dealership (SOD) counts the number of unique cars under the same vendor ID.

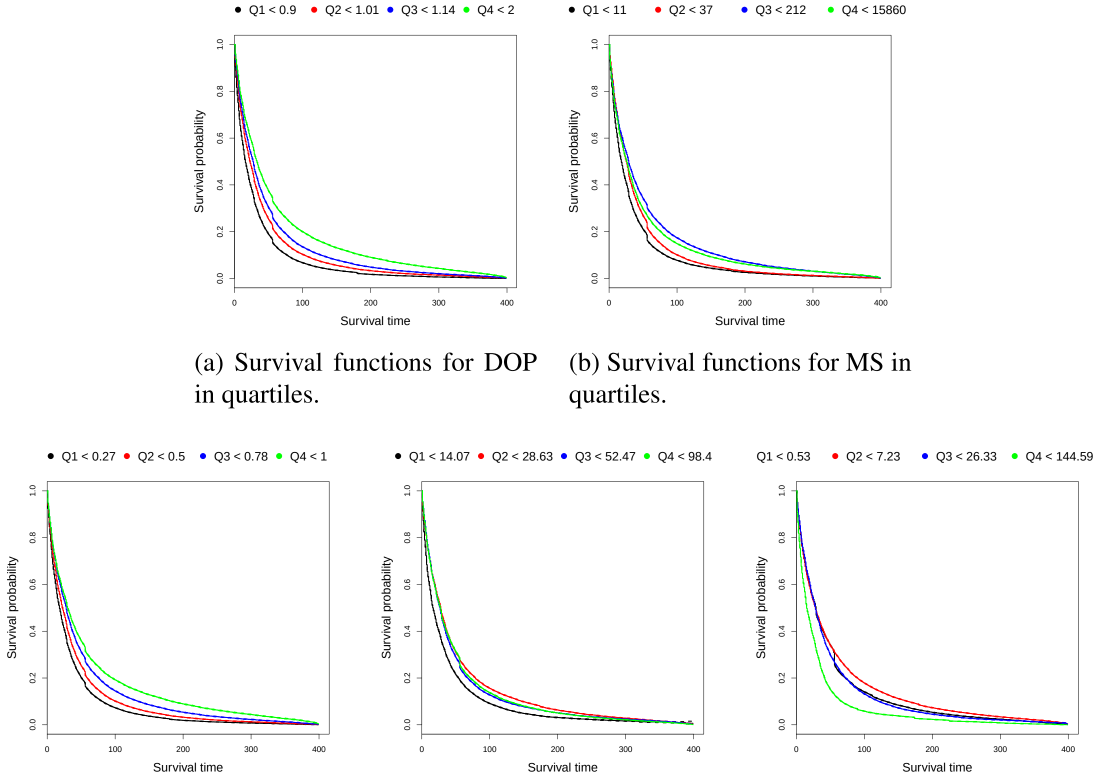
5 Results

This section reports the results observed during the application of survival analysis methods for PRF estimation. We first present results for traditional and data-driven survival models individually, and then compare their performance using the Brier score and the C-index for assessing calibration and discrimination performance, respectively. A discussion how the observed results answer our research questions follows in Section 6

5.1 Classical survival analysis methods

Kaplan-Meier estimators and the Cox proportional hazards model represent classical survival analysis methods. These methods are pragmatic, easy to implement and allow for a quick assessment of variable dynamics by visual analysis and statistical measures. Their results also allow for an interpretation of market dynamics.

Figure 1 depicts Kaplan-Meier curves for each of the independent variables. The visual analysis provides a first indication how variables (e.g., car features) affect the probability of selling a used car. To account for the fact that variable effects might differ across value ranges, we group observations according to the quartiles of variables' values and estimate the survival curves from the grouped data. We support results of Figure 1 with statistical analysis using the Cox proportional hazards model. Corresponding results in form of exponentiated model coefficients and significance are presented in Table 5.1, where we estimate individual survival models for individual car categories. Stratification by car category accounts for possible heterogeneity across car types and is feasible because every category is well represented in the large data set. Note that an exponentiated coefficient less than (above) one implies that the probability to sell a car decreases (increases) with larger values of the variable.



(a) Survival functions for DOP in quartiles.

(b) Survival functions for MS in quartiles.

(c) Survival functions for Quantile in quartiles.

(d) Survival functions for Age in quartiles.

(e) Survival functions for SOD in quartiles.

Figure 1: Assessing the effect of variables with Kaplan-Meier curves. Plot of Kaplan-Meier estimated survival probabilities for several variables. Groups are divided in quartiles. Survival time is defined in days.

Table 1: Cox proportional hazards. Comparing buyer preferences for major classes of a premium OEM

	Dependent variable							
	TOM (in days)							
	Compact	Van	Mid-size	Executive	SUV	Luxury	Luxus Cabrio	Sports Cabrio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MS	1.000***	0.999***	0.999***	0.999***	0.997***	0.998***	0.999***	0.998***
DOP	0.300***	0.398***	0.321***	0.379***	0.317***	0.257***	0.283***	0.513***
Quantile	0.841***	0.710***	0.835***	0.755***	0.746***	1.037**	0.889***	0.578***
Age	0.994***	0.995***	0.994***	0.996***	0.994***	0.988***	1.002***	0.996***
SOD	1.009***	1.008***	1.007***	1.006***	1.005***	1.003***	1.005***	1.006***
Observations	466,048	279,333	681,465	513,884	126,469	81,529	24,955	91,901
R^2	0.116	0.097	0.083	0.067	0.091	0.080	0.065	0.093
Max. Possible R^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Log Likelihood	-4,111,137.000	-2,267,910.000	-5,710,268.000	-4,073,737.000	-915,396.800	-536,513.300	-150,094.000	-651,078.700
Wald Test (df = 5)	57,624.660***	27,713.880***	56,778.420***	34,138.110***	11,176.410***	6,638.950***	1,694.180***	8,598.460***
LR Test (df = 5)	57,669.120***	28,400.600***	58,967.620***	35,511.800***	12,005.200***	6,779.878***	1,676.916***	8,938.150***
Score (Logrank) Test (df = 5)	57,228.840***	27,198.330***	55,774.100***	33,865.190***	10,946.070***	6,550.558***	1,675.045***	8,439.512***

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

DOP shows a monotonic increase in survival probability across all classes in Figure 1(a), with cars with the highest prices in relation to their hedonic value spending the most time unsold. Results of Cox proportional hazard models in Table 1 support this fact for each car class. The distribution of the coefficient clusters at around 0.3 for most car types; sports cars being an exception with a DOP coefficient of 0.513. A relatively large DOP coefficient for sports cars implies that offering a price above the (model estimated) market value decreases the sales probability of sports cars less than the sales probability of other cars. In other words, estimated DOP coefficients evidence how the price elasticity of second-hand sports cars is less than that of other cars.

Figure 1(b) suggests a non-monotonic increase in sales probabilities with larger market size. Curves with lowest sales probabilities represent the first and the second quartiles of market size. Corresponding cars face low competition from similar cars in price, age and mileage, which explains a low chance of selling. Survival curves of the third and the fourth quartiles do not show the same trend and overlap. We attribute this pattern to heterogeneity originating from the variable MS capturing both demand- and supply-side effects. Large MS values may be a consequence of the new car business in that popular car models exhibit large production and sales volume, which eventually lead to a large supply of used cars in the second-hand market. On the other hand, low demand may be the dominating determinant of large market size. Corresponding cars may be regarded as the ”lemon” of the second-hand market. With coefficient estimates close to one for all car classes, results of the Cox proportional hazards model suggest a small effect of MS on the sales probability. To correctly interpret the exponentiated coefficient near one and the high significance of the feature we recall the definition of MS, which is the number of equivalent cars being offered simultaneously. The high range in values of MS is responsible for the exponentiated coefficient very close to one.

The variable quantile shows a similar monotonic behavior as DOP in Figure 1(c). We explain this result with the direct relationship between Quantile and the vehicle’s price. Our analysis by car classes supports this result across all classes but Luxury. For most classes, the coefficients lie within the range of 0.7 - 0.9. Notably, the lowest coefficient of 0.578 is present in the category sports cabriolet. This reveals that the willingness of a customer to pay a surplus for specific technical characteristics, resulting in a lower DOP effect, is compensated by a higher sensitivity regarding the price positioning of the vehicle within its own class.

The variable age displays a non-monotonic behavior in Figure 1(d). The lowest quartile of the variable Age shows expected behavior as the survival probability is the lowest among the four curves. Rather surprisingly is the behavior of the second quartile (up to 28 months old), as the survival probability is higher as for the third and fourth quartiles. This result implies that cars up to 28 months of age stay longer in the market as older cars. An explanation might be the price of these vehicles. Older cars tend to have lower prices, which highly increase the probability of being sold, as shown by the survival curves for DOP and Quantile. The German used car market has two highly frequent age segments. These are cars around 12 and 48 months of age. The former come from the automobile OEMs and their employees while the latter come from leasing returns. These segments are also known to customers (Johnson and Waldman, 2003). It may be plausible that dealers set higher prices for cars up to 28 months age relatively to their actual market value as they position them to the category of OEM employees' cars rather than to the leasing returns. Car class-wise results of the Cox model also provide results with very small impact on the survival probability. All classes have coefficients close to one, with Luxury Cabrio being the only class with a coefficient slightly above one. This fact is similar to the interpretation of the MS results, as Age has a high variance in its values.

The variable SOD shows a non-monotonic increase in survival probability in Figure 1(e). SOD represents the size of the dealership and displays lowest survival probability for cars from large vendors. This might be due to more efficient sales processes, more effective marketing and higher know-how as the competitive edge. For the third and second quartile observations, the survival probability gradually rises. The survival curve representing the first quartile crosses the curves of the second and the third quartiles. An explanation might be dealerships with less than 10 cars on the market, where chance and individual skill might be more important than the effects arising from the size of the dealership alone. In the class by class analysis, SOD is the only variable with consistent positive effect on the sales probability across all car types. This fact undermines the finding from Kaplan-Meier estimations and validates SOD as a proxy for efficient processes and better market knowledge. Notably, the coefficients are close to one in the Cox model implying a low impact of SOD in the Cox proportional hazards model. Again, the variance in SOD values contributes to the near-one coefficient of the feature.

Overall, the results from Cox proportional hazards model largely agree with results from Kaplan-Meier curves. Variables with low impact according to the Cox model show non-monotonic behavior for the probability of a car being sold in the Kaplan-Meier curves. We have provided possible explanations for non-monotonic behavior for the variables MS, Age and SOD. However, low impact of these variables according to the Cox proportional hazards model does not mirror our interpretation of Kaplan-Meier results. Therefore, we proceed with checking model assumptions to analyze whether the results may be attributable to our data failing to fulfill the linearity and proportionality assumptions of the Cox proportional hazards model.

We analyze the linearity assumption using visual analysis of the martingale residuals, which are defined as

$$r_{M_i} = \delta_i - r_{C_i}, \tag{30}$$

where δ_i is the event status for the i-th observation and

$$r_{C_i} = \hat{H}(T_i, x_i) = \exp(\beta x_i)\hat{H}_0(T_i), i = 1, \cdots, n \tag{31}$$

is the Cox-Snell residuals. We plot the residuals against the single features of the model to test for a non-zero slope, which is an indication for the violation of the linearity assumption (Therneau et al., 1990). Figure 2 provides corresponding results and reveals that each variable displays a non-zero slope with the highest deviations for MS, SOD and Age. This suggests that the initial result of low to zero impact from Cox proportional hazards model (see Table 1) for these variables comes from the limitations of the model. The graphical analysis also provides some evidence for the existence of non-linear relationships between the variables and the target variable TOM.

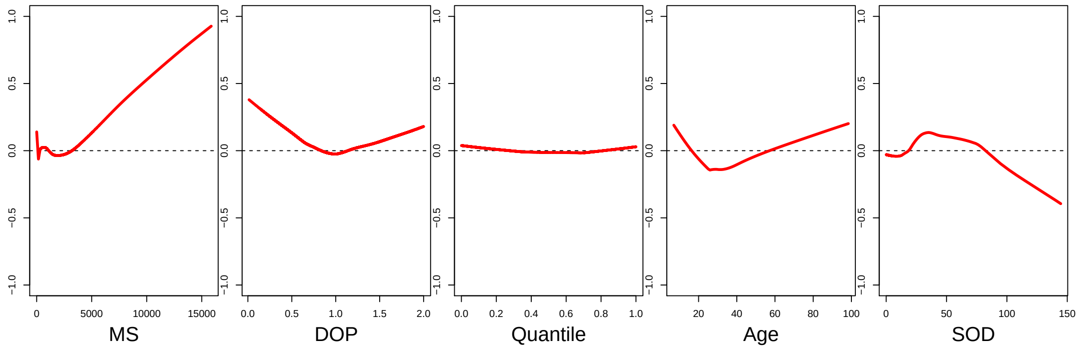


Figure 2: Assumptions check - martingale residuals for single features. Variables are plotted against martingale residuals for the linearity check.

We analyze the proportionality assumption using visual and statistical analysis of the Schoenfeld residuals (Schoenfeld, 1982; Andersen, 1982; Aranda-Ordaz, 1983), which we plot against the survival time in Figure 3. A non-zero slope is once again an indication for non-proportionality; that is a violation of model assumption. We further secure the results of a visual inspection of Figure 2 through performing a χ-squared test with the H_0 hypothesis of the hazards being proportional on the data at the significance level of 95%. Note that the significance level of the Schoenfeld residual is dependent on the size of the sample. As our data has several millions of observations, an even more conservative significance level may be necessary to accept H_0 (Lin et al., 2013). Table 2 provides

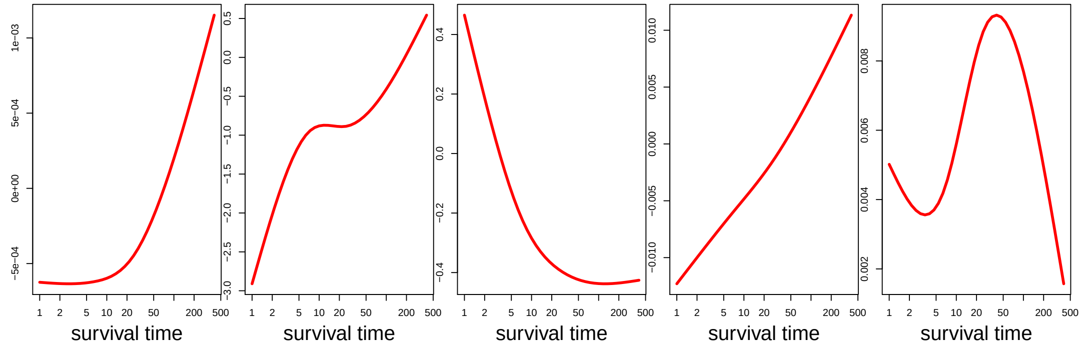


Figure 3: Assumptions check - Schoenfeld residuals for individual variables. Variables' values are plotted against Schoenfeld residuals for the proportionality check.

the results of the χ-squared test statistic and reveals that the proportionality assumption cannot be accepted for any of the variables at the significance level of 95%. In summary, the χ-squared test statistic and the visual examination provide strong evidence for non-proportionality.//

Table 2: Proportionality check. Schoenfeld residuals test.

Variable	ρ	χ^2	p-value
MS	0.0718	8587	0
DOP	0.0747	9723	0
Quantile	-0.0517	4499	0
Age	0.0806	11168	0
SOD	0.0550	3751	0
Full model		27253	0

The analysis of the assumptions of the Cox model suggests that neither the linearity nor the proportionality assumption can be accepted. This result restrains further use of the Cox proportional hazards framework. It also calls for the consideration of methods with less rigid assumptions and inbuilt ability to capture non-linear and non-proportional relationships in the data. We argue that this finding is managerially meaningful. On the one hand, survival analysis in general provides a powerful statistical framework to aid pricing decisions in the used car market (Jerenz, 2008). On the other hand, the only application of this framework for car reselling by Jerenz (2008) restricts itself to classical methods like the Cox model, which our analysis shows to be inadequate.

5.2 Data driven survival methods

Previous results on the limitations of the Cox model, as representative for classic parametric survival models, motivate an analysis of data-driven survival models for PRF estimation. We select four different methods, all of which ground on the concept of decision tree learning (Breiman et al., 1984): survival tree, random survival forest, conditional inference tree and conditional inference forest. We discuss our modeling strategy and corresponding results using the example of the random survival forest (RSF). Detailed results for other methods are available in the Appendix. The analysis of model results displays much similarity across different data-driven survival methods. Focusing on one method, therefore avoids repetitive discussions. We discuss RSF in the main part of the paper because it performs best among the four data-driven techniques; as shown later in the paper.

The performance of RSF, as well as other data-driven survival methods, depends on meta-parameter settings to be selected by the analyst. RSF meta-parameters include the number of trees in the ensemble and the number of variables to be used at random in each split (Breiman, 2001). To tune meta-parameters, we consider candidate values of each meta-parameter and empirically determine the best combination using grid-search (Lessmann and Voss, 2017). The out-of-bag prediction error, a measure naturally generated in a random forest framework Breiman (2001), facilitates assessing the predictive accuracy of a specific combination of meta-parameter values.

Figure 4 provides the results of RSF parameter-tuning on a random sample of 20,000 observations. We draw a random sample because the estimation and assessment of several candidate RSF models on the full data set would be computationally intractable. Considering candidate settings of 1, 2, and 3, we find the best number of variables selected at random per split to be 2. Figure 4 also shows how the out-of-bag error decreases when adding additional survival trees. This beneficial effect, however, diminishes with forest size, while adding trees always increases run times and memory requirements. In the light of Figure 4, we select the forest size for subsequent analysis to be 100 trees. We suggest this value to provide close to minimal prediction error while avoiding excessive run times on the full data set.

The first research question we strive to answer concerns the mechanisms that govern car resale in the second-hand market. Results of the Cox model (Table 5.1) have given a preliminary answer to this question through estimating the direction, magnitude, and significance of variable coefficients. Violation of model assumptions draws these results into perspective. To revisit results of the Cox model and obtain a clearer view on the way in which features affect resale probabilities, we examine the variable importance of the RSF. Random forest based variable importance ranks are a popular approach to appraise how much a variable impacts model predictions and to inform feature selection (Ishwaran et al., 2011). Roughly speaking, the magnitude of an importance score captures the degree

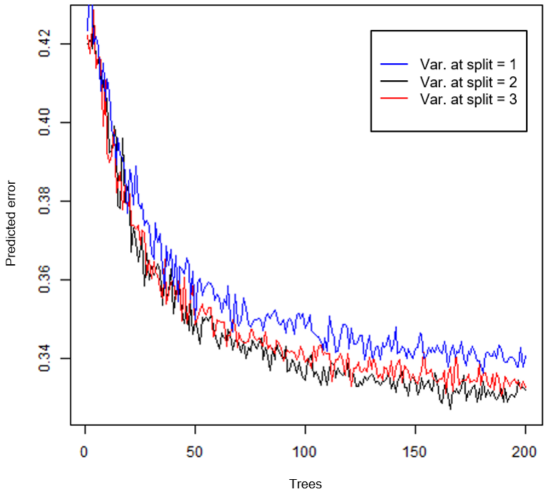


Figure 4: Tuning tree parameters - check predicted error stability across changing number of trees and variables at each split. Number of observations - 20.000.

to which predictive accuracy decreases, if the information within the variable were not available to the model (Breiman, 2001). We develop 100 RSF models, run them on disjoint data samples and average the normalized variable importance scores across all models. Figure 5 provides corresponding results.

We find SOD to be the most important variable, closely followed by Age and MS. On average, importance scores of the latter two are similar, whereby MS displays larger variation in that the minimum and maximum variable importance observed across the 100 RSF models show larger spread compared to Age. Interestingly, both price related variables, Quantile and DOP, show relatively lesser impact. Especially DOP, the variable previous work identifies as most relevant predictor of sales probabilities (Jerenz, 2008), comes out as relatively least important variable in Figure 5. This result indicates that consumers' price sensitivity depends on factors such as age and competition (captured in MS). We also find some evidence that deeper market knowledge and more efficient sales processes, which we associate with SOD, facilitate larger dealerships to charge a premium price. Last, the fact that the maximum variable importance of DOP is roughly equal to the minimum importance score of SOD provides strong evidence in favor of our proposition to include SOD as predictor in survival models for PRF estimation.

We secure results from the variable importance analysis through examining the development of the integrated Brier Score when iteratively discarding variables. To that end, we first calculate the Brier score of the full model using a training sample of 13.500 observations and a random sample of 1.500 observations for the prediction. Next, we exclude one variable, reestimate the model, and recalculate the Brier score. The Brier Score mea-

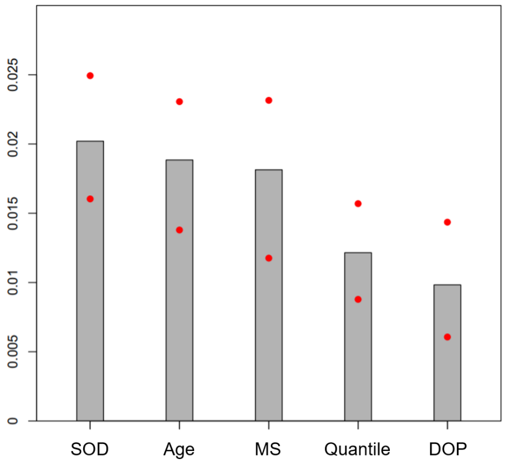


Figure 5: Variable importance for the five variables - mean values from a sample of 100 models. Red points indicate maximal and minimal values

sures the degree to which RSF estimated sales probabilities agree with actual values and is formally defined as the mean of the squared residuals between model forecasts and a zero-one coded binary target variable (Eren, 2014). It is common practice to consider the integrated Brier Score in survival analysis. There, an integration of scores occurs in that Brier Scores can be calculated for each discrete time point. Table 3 provides corresponding results, where an increase in the Brier score indicates a decrease of model performance. One finding of Table 3 is that every variable exclusion decreases the performance of the RSF model. This confirms that each variable is important. The variables SOD and Age belong once more to the group of most important variables. Their exclusion causes the largest decrease in performance. As in Figure 5, exclusion of DOP hurts model performance the least. The relevance of the variables MS and Quantile differs between the Brier Score and RSF variable importance analysis.

Additionally, we perform similar analysis using the C-index. Note that in contrary to the integrated Brier score, a higher C-index corresponds with better model performance. Table 4 shows a drop in the C-index for each model with a variable excluded. Exclusion of SOD results in the biggest drop of the C-index and thus model performance, followed by Age and Quantile. The results of C-index analysis are in line with the integrated Brier score.

Overall, variable importance, Brier score and C-index results emphasize the importance of SOD, Age and, to lesser extent, MS and Quantile. DOP shows the lowest variable importance and the lowest improvement of model calibration and discrimination. These

Table 3: Variables analysis. Integrated Brier score for random survival forest models under exclusion of variables

	Integrated Brier Score
Full model	0.090
w/o SOD	0.096
w/o Age	0.096
w/o Quantile	0.096
w/o DOP	0.094
w/o MS	0.094

Table 4: Variables analysis. C-index for random survival forest models under exclusion of variables

	Concordance index
Full model	0.600
w/o SOD	0.564
w/o Age	0.579
w/o Quantile	0.579
w/o DOP	0.584
w/o MS	0.586

results stand in contrast to the results of the Cox proportional hazards model. Our analysis of its limitations through strict assumptions delivers a first indication for the deviating results. While it is possible that classical methods correctly assign the impact of each variable, we assert that data-driven, non-linear tree methods better fit the underlying data and facilitate a more precise detection of patterns, variable importance, and predictive power. To confirm this, we compare classical models and data- driven survival methods in the next section.

5.3 Survival model performance comparison

We compare the Cox proportional hazards model and the data-driven approaches with regards to calibration and discrimination performance. We assess the calibration performance of the models using the Brier score and use the C-index to assess the discriminative power of each model. We denote the set of models as M with a model $M_i, i = 1, \ldots, 5$ referencing to one specific survival model. We sketch our benchmarking approach in Algorithm 1.

Algorithm 1 Comparison of calibration and discrimination performance of classic and data-driven survival methods.

```
 1: Randomly split observations in training set $T$ and validation set $V$ with a 60/40 ratio.
    Define $\tilde{V}$ as a fixed subset of $V$ with 1500 observations
 2: Split the training set $T$ in disjoint subsets $T_j$ with $T = \bigcup_{i=J}^{100} T_j$
 3: for $j = 1, \dots, 100$
 4:     Fit model $M_i, i = 1, \dots, 5$ on $T_j$ and define it as $M_i^j$
 5:     Make a prediction up to 120 days using $M_i^j$ on $\tilde{V}$
 6:     Calculate Brier score for $M_i^j$ and define it as $Brier_i^j$
 7:     Calculate C-index score for $M_i^j$ and define it as $C_i^j$
 8:     Calculate paired difference $Brier_{i,k}^j = Brier_k^j - Brier_i^j$ for $i, k = 1, \dots, 6$
 9:     Calculate paired difference $C_{i,k}^j = C_k^j - C_i^j$ for $i, k = 1, \dots, 6$
10: Calculate mean and standard error for $Brier_{i,k}^j$ and $C_{i,k}^j$ over all $j$
11: Calculate confidence intervals for means of paired differences based on a $t$-distribution
```

Figure 6 and Figure 7 provide comparative results in terms of $Brier_{i,k}$ and $C_{i,k}$, respectively, together with the respective confidence intervals. Note that according to Step 8 and Step 9 of Algorithm 1, we calculate the difference in model performance as *performance of model in row - performance of model in column*. A red horizontal line represents no difference in model performance.

Figure 6 reports pairwise comparison of Brier scores. In the context of pairwise comparisons, a difference in means above the red line indicates better performance of the column model compared to the row model. The difference is statistically significant if the confidence interval, represented by the slashed lines, does not include zero.

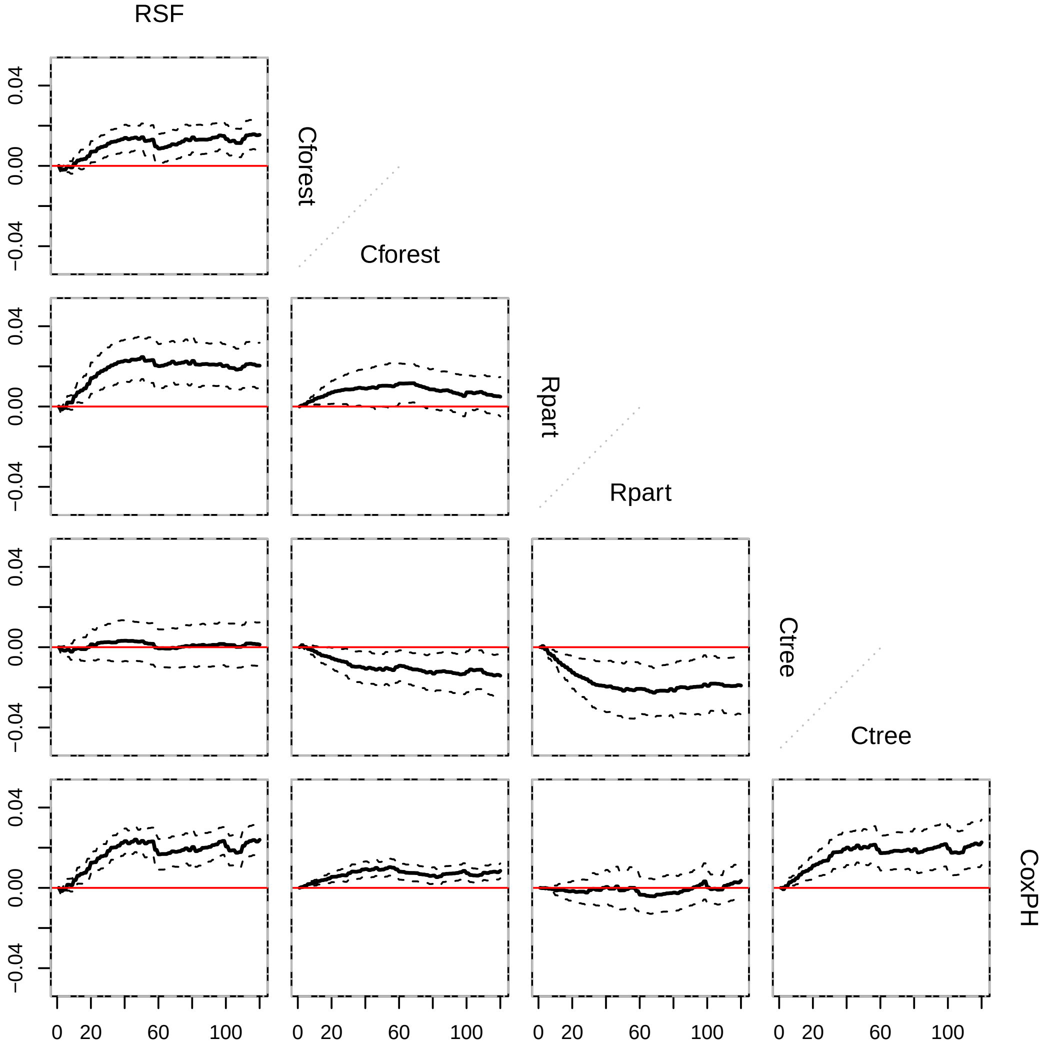


Figure 6: Pairwise model comparison - Brier Score for right censored data. Mean and 95% confidence interval for paired differences between selected survival models. Estimation from 100 disjoint training samples.

Random survival forest and the conditional inference tree show the best calibration performance. Their performance does not differ significantly, as shown in the panel ”RSF - Ctree”. However, random survival forest and conditional inference tree significantly outperform other models in terms of the Brier Score. The Cox proportional hazards model shows the worst performance, which we attribute to violations of model assumptions for the data employed here. Figure 7 depicts pairwise comparative results in terms of the

C-index. Note that unlike for Brier Score, higher C-index indicates better discriminative performance. Thus, a difference in means below the red line indicates better performance of the column model compared to the row model.

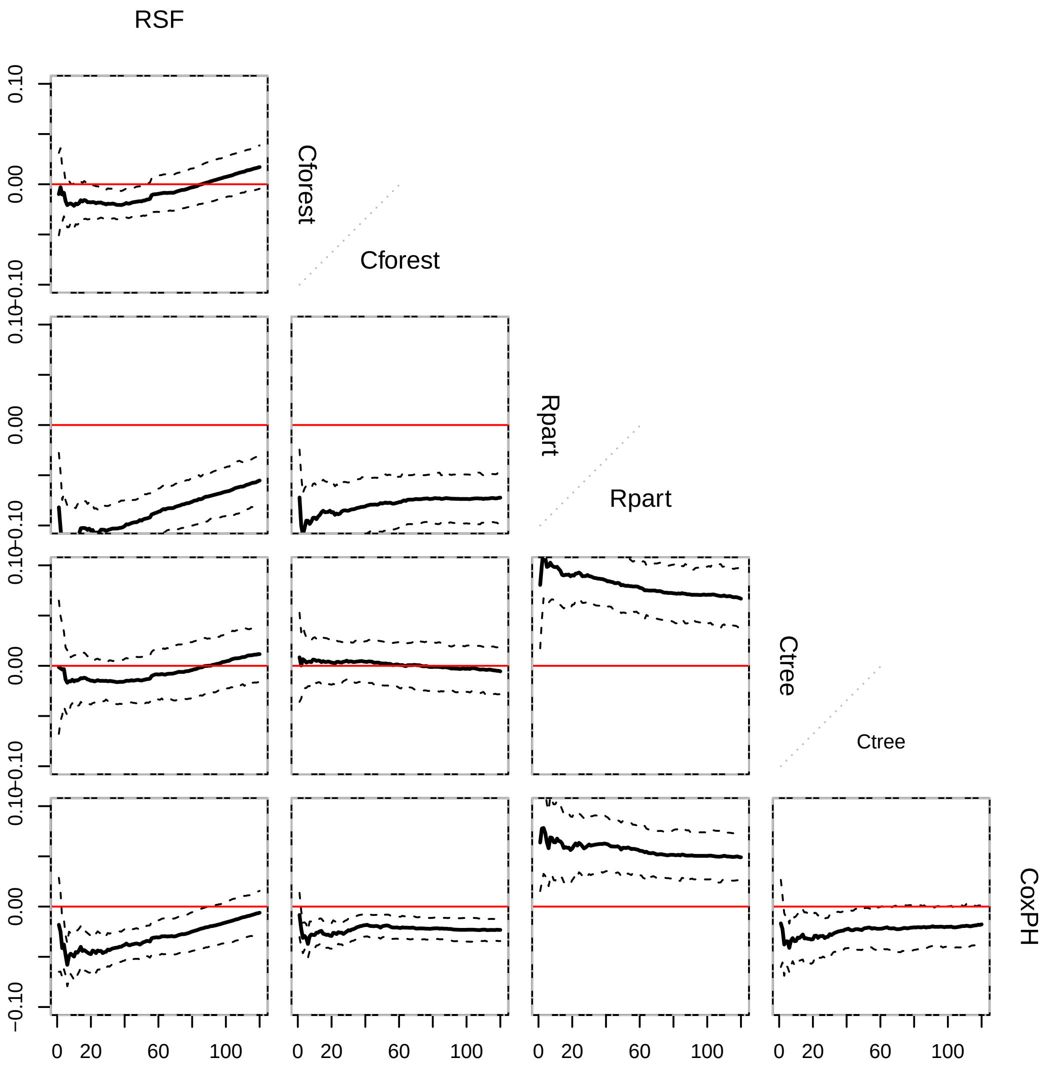


Figure 7: Pairwise model comparison - C-Index for right censored data. Mean and 95% confidence interval for paired differences between selected survival models. Estimation from 100 disjoint training samples.

Figure 7 suggests that the difference in model performances decreases with longer TOM. This might come from lemons (i.e., cars that are hard to sell) showing common patterns, which are easily detected by each model. RSF shows superior performance up to 60 days on market. The confidence intervals indicate that the difference in performance is not significant for conditional inference forest and conditional inference tree. For cars with TOM higher than 60, a change in performance in favor of both conditional inference models, tree and forest, is present. Nevertheless, the confidence intervals include zero. Practitioners consider day 60 of a car's TOM as an unspoken threshold before the price of car needs to be reduced in used car retail. Single survival tree shows the worst performance, followed by the Cox proportional hazards model. Contrary to the calibration analysis, a single conditional inference tree does not outperform its ensemble version.

Overall, results of Figures 6 and 7 clearly confirm the need for data-driven survival methods for PRF estimation. The Cox proportional hazards models cannot compete with the data-driven alternatives considered here; neither in calibration nor in discrimination performance. More specifically, we find the random survival forest to be the best model in the comparison. In terms of both calibration and discrimination performance, random survival forests perform at least competitive to and typically better than other survival methods; often outperforming them with statistically significant margin. Surprisingly, the single conditional inference tree shows superior performance to its ensemble version in terms of calibration and at least on par performance for discrimination. This contradicts the view ensemble methods outperform their single model counterparts, which is widely adopted in predictive modeling (Lessmann and Voss, 2017, e.g.,), and provides some first evidence that empirical findings obtained in the broader scope of supervised learning might not necessarily generalize to the more specialized area of survival analysis.

6 Discussion

The empirical results provide insights into the dynamics of car resales in the second-hand car market. Furthermore, we elaborate on the appropriateness of data-driven methods for survival analysis in the context of this market, and more specifically the support of pricing decisions. We consider deeper market understanding and more precise methods as essential enablers for a dynamic pricing approach in a broader revenue management framework. Such framework is needed to increase margins and overcome the ”zero-profit-problem” that affects car manufacturers and independent dealerships in the huge second-hand car business (Jerenz, 2008). Against this background, we analyze the interactions of variables and estimate survival curves for used cars, depending on the car's intrinsic features, price related variables, and structural effects related to dealership size. Supported by the empirical results, we can answer our research questions.

1. **Which factors influence the time a used car spends on the market before it gets sold?**

This study analyzes factors related to offer prices (DOP and Quantile), a factor representing to the car's competition within its market segment (MS), a factor capturing maturity of front-end business processes and market insight, which we originally introduce in this work, (SOD), and the age of a used car. Table 5 summarizes relevant results across factors and survival models in terms of variable importance ranks. We assess the magnitude of the exponentiated coefficients for the Cox proportional hazards model and assign the importance rank accordingly. For survival trees and conditional inference trees, we focus on the rank of the splits, which are shown in plots in the Appendix. For random survival forest and conditional forest, we rank the importance of the variables based on the integrated Brier score. We motivate our decision to use different statistics to assign importance ranks by the underlying differences in the models and lack of a common importance statistic. Nevertheless, as we are able to find meaningful interpretations for the models' behavior, we propose to use the overview as a starting point for the discussion. We assign rank 1 to the variable with the highest importance and 5 to the variable with the lowest importance for a model. We note variables with no impact according to the model as 'none'.

Table 5: Comparison of importance rank of variables across used models.

Variable	Model				
	Cox p. h.	**Survival tree**	**RSF**	**C. i. tree**	**C. i. forest**
DOP	1	none	5	1	3
MS	none	2	4	3	none
Quantile	2	none	3	4/5	2
Age	none	3	2	4/5	1
SOD	none	1	1	2	none

Random survival forest and conditional inference tree are the only models that regard all variable as important. These models are also the ones with the best predictive performance. This suggests that all variables carry explanatory and predictive information for the mechanisms of the used car market.

Random survival forest and conditional inference tree both assign high importance to SOD. This leads to the conclusion that structural effects related to a dealership's size are an important factor for a successful car sale. We argue that this finding is important because prior work on used car revenue management has limited its attention to variables related to the car (Jerenz, 2008). Our new variable SOD evidences the relevance of other sources of information to model sales probabilities and identifies a potentially fruitful direction along which to search future improvements in the form of additional variables. For example, SOD itself is a gross measure that encompasses multiple factors. Future research could examine the explanatory and predictive value of more specific characteristics of used car vendors.

DOP shows contrary importance for both models. To correctly interpret this result, we recall Table 3. The integrated Brier scores for all variables are very similar, indicating that a low importance rank is by far not equivalent to no importance. Further, we recall that price is often the only steering mechanism available to a decision-maker. Therefore, we conclude that DOP, as a price related variable, has a significant predictive and explanatory power.

2. **What is the most accurate statistical method to predict the time on market?** The empirical results show superior performance in calibration for random survival forest and conditional inference tree. In terms of discrimination, random survival forest shows best performance for used cars with TOM up to 60 days. Models based on the conditional inference may outperform random survival forest for used cars with higher TOM.
   A potential implementation as a decision support system for a used car dealership have to take into account the scarcity of internal statistical resources and high costs of external resources. Thus, even if our results indicate an on-par performance of two different models, we understand the importance of a final reference towards one of the potential candidates. Considering our target to estimate the market demand for further use in a revenue management framework, we suggest to regard additional model characteristics prior to a final decision. Random survival forest as an ensemble method has a lower tendency towards overfitting (Breiman, 2001) in comparison to a single tree. With the empirical results and additional considerations, we suggest to use random survival forest as the model of choice for large datasets with right-censored observations.

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. From additional considerations we suggest to select the random survival forest for an implementation in a possible decision support system or revenue management framework. Here, our contribution is not limited to the application of advanced data driven methods, but also includes survival analysis modeling with significantly larger dataset than used in previous work. Such combination of large, real-world data, innovative statistical modeling, and deeper understanding of the market is precisely what promises to be the new success driver in the ever more Big Data obsessed business world of tomorrow.

References

, 2017. Dat report 2017.

Aalen, O., 1978. Nonparametric inference for a family of counting processes. The Annals of Statistics, 701–726.

Adams, C. P., Hosken, L., Newberry, P. W., 2011. Vettes and lemons on ebay. Quantitative Marketing and Economics 9, 109–127.

Akerlof, G., 1970. The market for lemons: qualitative uncertainlyand market mechanism. Quarterly Journal of Economics 89.

Andersen, P. K., 1982. Testing goodness of fit of cox's regression and life model. Biometrics, 67–77.

Aranda-Ordaz, F. J., 1983. An extension of the proportional-hazards model for grouped data. Biometrics, 109–117.

Avi, H., 2018. Optimal pricing and replenishment of an expiring inventoried product under heterogeneous consumer sensitivities. Decision Sciences (doi:10.1111/deci.12276).
URL `https://www.onlinelibrary.wiley.com/doi/abs/10.1111/deci.12276`

Ayres, I., Siegelman, P., 1995. Race and gender discrimination in bargaining for a new car. The American Economic Review, 304–321.

Bapna, R., Jank, W., Shmueli, G., 2008. Consumer surplus in online auctions. Information Systems Research 19 (4), 400–416.
URL `https://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0173`

Belleflamme, P., Peitz, M., 2014. Asymmetric information and overinvestment in quality. European Economic Review 66, 127–143.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al., 2011. A review of survival trees. Statistics Surveys 5, 44–71.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press.

Chen, J., Esteban, S., Shum, M., 2013. When do secondary markets harm firms? The American Economic Review 103 (7), 2911–2934.

Cox, D. R., 1992. Regression models and life-tables. In: Breakthroughs in statistics. Springer, pp. 527–541.

Crowley, J., Leblanc, M., Gentleman, R., Salmon, S., 1995. Exploratory methods in survival analysis. Lecture Notes-Monograph Series 27, 55–77.
URL `http://www.jstor.org/stable/4355862`

Desai, P., Purohit, D., 1998. Leasing and selling: Optimal marketing strategies for a durable goods firm. Management Science 44 (11-part-2), S19–S34.

Dirick, L., Claeskens, G., Baesens, B., 2017. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society 68 (6), 652–665.
URL `https://doi.org/10.1057/s41274-016-0128-9`

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 (4), 637–644.
URL `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 (22), 2867–2885.

Eren, D., 2014. A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. Decision Sciences 45 (5), 849–880.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/deci.12094`

Genesove, D., 1993. Adverse selection in the wholsesale used car market. Journal of Political Economy 101 (4), 644–665.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M. J., 2006a. Survival ensembles. Biostatistics 7 (3), 355–373.

Hothorn, T., Hornik, K., Zeileis, A., 2006b. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics 15 (3), 651–674.

Hothorn, T., Lausen, B., 2003. On the exact distribution of maximally selected rank statistics. Computational Statistics & Data Analysis 43 (2), 121–137.

Hothorn, T., Lausen, B., Benner, A., Radespiel-Tröger, M., 2004. Bagging survival trees. Statistics in medicine 23 (1), 77–91.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., 2008. Random survival forests. The annals of applied statistics, 841–860.

Ishwaran, H., Kogalur, U. B., Chen, X., Minn, A. J., 2011. Random survival forests for high-dimensional data. Statistical Analysis and Data Mining 4 (1), 115–132.
URL `http://dx.doi.org/10.1002/sam.10103`

Jerenz, A., 2008. Revenue management and survival analysis in the automobile industry. Springer.

Johnson, J. P., Waldman, M., 2003. Leasing, lemons, and buybacks. The RAND Journal of Economics 34 (2), 247–265.
URL `http://www.jstor.org/stable/1593716`

Kaplan, E. L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American statistical association 53 (282), 457–481.

Kleinbaum, D. G., Klein, M., 2006. Survival analysis: a self-learning text. Springer Science & Business Media.

LeBlanc, M., Crowley, J., 1993. Survival trees by goodness of split. Journal of the American Statistical Association 88 (422), 457–467.

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.

Levin, J., 2001. Information and the market for lemons. RAND Journal of Economics, 657–666.

Lin, M., Jr, H. C. L., Shmueli, G., 2013. Research commentary—too big to fail: Large samples and the p-value problem. Information Systems Research 24 (4), 906–917.

Naftel, D., Blackstone, E., Turner, M., 1985. Conservation of events. Unpublished notes.

Nelson, W., 1969. Hazard plotting for incomplete failure data. Journal of Quality Technology 1 (1), 27–52.

Nelson, W., 2000. Theory and applications of hazard plotting for censored failure data. Technometrics 42 (1), 12–25.

Olivares, M., Cachon, G. P., 2009. Competing retailers and inventory: An empirical investigation of general motors' dealerships in isolated us markets. Management Science 55 (9), 1586–1604.

Prado, S. M., 2010. Macroeconomics of the new and the used car markets. Economics Bulletin 30 (3), 1862–1884.

Ranganathan, P., Pramesh, C., et al., 2012. Censoring in survival analysis: potential for bias. Perspect Clin Res 3 (1), 40.

Ratchford, B. T., Srinivasan, N., 1993. An empirical investigation of returns to search. Marketing science 12 (1), 73–87.

Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model. Biometrika 69 (1), 239–241.

Segal, M. R., 1988. Regression trees for censored data. Biometrics, 35–47.

Tang, L., Thomas, L., Fletcher, M., Pan, J., Marshall, A., 2014. Assessing the impact of derived behavior information on customer attrition in the financial service industry. European Journal of Operational Research 236 (2), 624–633.

Therneau, T. M., Grambsch, P. M., Fleming, T. R., 1990. Martingale-based residuals for survival models. Biometrika 77 (1), 147–160.

Xishu, L., Rommert, D., Christiaan, H., Mustafa, H., 2016. Assessing end-of-supply risk of spare parts using the proportional hazard model. Decision Sciences 47 (2), 373–394. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/deci.12192

Zhu, L., Sellers, K. F., Morris, D. S., Shmueli, G., 2017. Bridging the gap: A generalized stochastic process for count data. The American Statistician 71 (1), 71–80. URL https://doi.org/10.1080/00031305.2016.1234976

Appendix A Evaluation of tree-based methods

Appendix A.1 Survival trees

As described in the methodology section, survival tree relates to classification and regression trees (CART) with adjustments to censored data. Figure A.8 shows a survival tree plot from which we deduct our results. The first split accounts for the feature SOD, separating the underlying data in larger and smaller dealerships. Next split accounts for MS and the last split accounts for Age. The survival tree excludes DOP and Quantile and thus does not incorporate any information regarding the price of the used car. From this we deduct that survival tree may have low predictive performance on our data set as it cannot incorporate all information available.

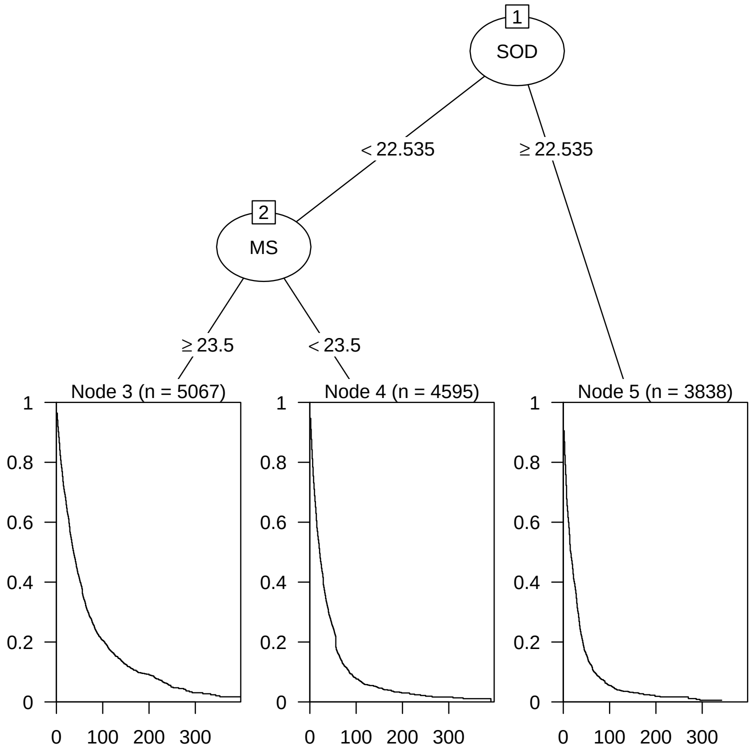
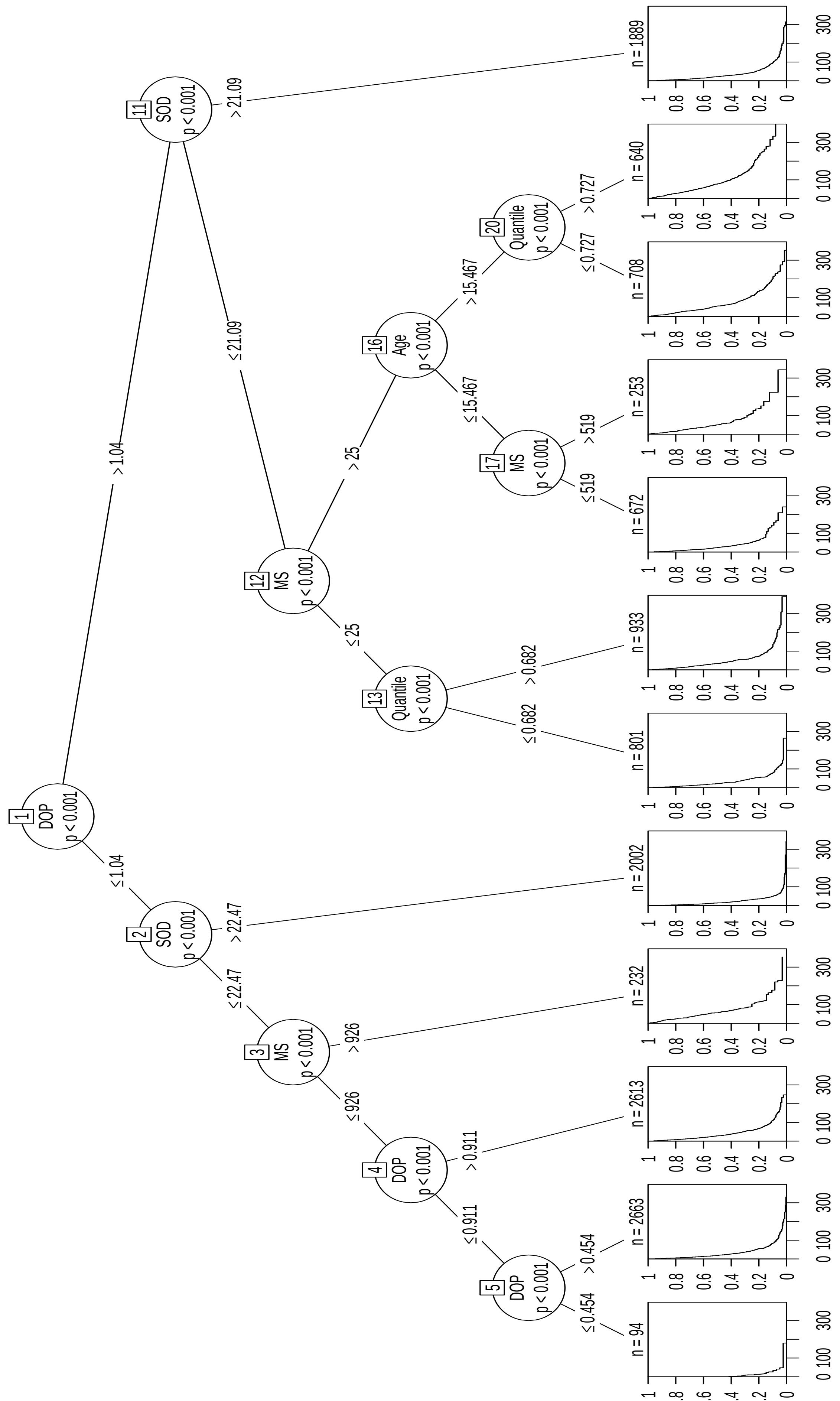


Figure A.8: Survival tree plot - inspecting splits for a random sample of 13.500 observations from the data.

Appendix A.2 Conditional inference tree

We fit a conditional inference tree model to our data and deduct our results from the visual observation of the plot presented in Figure A.9. We can observe effects of interactions between the features. The first two splits occur for DOP and SOD supporting results of random survival forest and Cox proportional hazards model. Next splits do not show a predominant feature. Rather groups of separate branches of the tree segregate the data.
DOP lower than 1 has the highest impact for a fast car sale. The fastest decline of the survival curves reflects this finding. The effect is less present for cars with very large MS values. This fact indicates that in presence of strong competition within a segment even a relative appealing price is not enough to offset the competition. The shortest TOM is present for cars with DOP of 0.8 or less and for cars sold by large dealerships. Contrary, the longest TOM is present for cars older than 18 months and the price set in the top 10% for the comparable cars. Highly overpriced cars with DOP over 1.4 also have the highest survival rates and thus are hard to sell. Further, large dealerships manage to sell cars with high DOP faster and can even offset the effect of competition corresponding to high MS.

Figure A.9: Conditional inference tree plot for a random sample of 13.500 observations from the data.



Appendix A.3 Conditional inference forest

We organize our conditional inference forest modeling approach similar to random survival forest model. First, we find the parameters for the model, which ensures the best predictive performance. Conditional inference forest depends on two hyperparameters, the number of trees in the ensemble and the threshold p to perform a split. Our initial examination on the suitable number of trees shows similarities to random survival forest. While higher number of trees improve the model and lowers the out-of-bag error, the effect of diminishing returns occurs at around 100 trees. Thus, we set the number of trees to 100. Note, that we do not provide the graphical results from analysis on the number of trees for conditional inference forest.

We decide to find a suitable p-value from a set of possible candidates from 0.95 to 0.99 in .01 steps. We use the concordance index as the measure of choice on bootstrapped data and select the p-value with the highest index value. A threshold value of 0.95 results in the best performance as shown in Figure A.10. We perform the analysis on feature im-

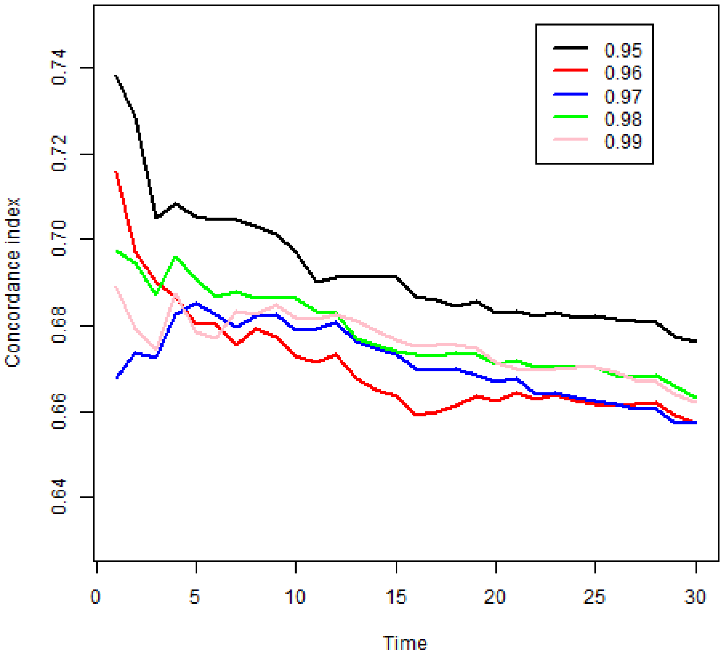


Figure A.10: Tuning tree parameters: minimal p-value criterion - check predicted error stability across changing lower p-values. Results from bootstrapped data.

portance using the integrated Brier score and show the results in Table A.5. Surprisingly, the exclusion of the features SOD and MS does not worsen the model performance. The features Age, Quantile and DOP show slight performance increase. The results from conditional inference forest show congruence with findings from Cox proportional hazards model. Again we assume, that a model which does not consider all variables as important may have lower predictive power.

Table A.6: Variables analysis. Integrated Brier score for conditional inference forest models under exclusion of variables

	Integrated Brier Score
Full model	0.087
w/o SOD	0.087
w/o Age	0.089
w/o Quantile	0.089
w/o DOP	0.088
w/o MS	0.087

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 ”Data Driven Value -at -Risk Forecasting using a SVR -GARCH -KDE Hybrid” by Marius Lux, Wolfgang Karl Härdle, Stefan Lessmann, January 2018

002 ”Nonparametric Variab le Selection and Its Application to Additive Models” by Zheng-Hui Feng, Lu Lin, Ruo -Qing Zhu as nd Li -Xing Zhu, January 2018

003 ”Systemic Risk in Global Volatility Spillover Networks: Evidence from Option -implied Volatility Indices” by Zihui Yang, Yinggang Zhou, January 2018

004 ”Pricing Cryptocurrency options: the case of CRIX and Bitcoin” by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou, Weining Wang, January 2018

005 ”Testing for b ubbles in cryptocurrencies with time -varying volatility” by Christian M. Hafner, January 2018

006 ”A Note on Cryptocurrencies and Currency Competition” by Anna Almosova, January 2018

007 ”Knowing me, knowing you: inventor mobility and the formation of technology -oriented alliances” by Stefan Wagner, Martin C. Go ossen, February 2018

008 ”A Monetary Model of Blockchain” by Anna Almosova, February 2018

009 ”Deregulated day -ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis” by Antanina Hryshchuk, Stefan Lessmann, February 2018

010 ”How Sensitive are Tail -related Risk Measures in a Contamination Neighbourhood?” by Wolfgang Karl Härdle, Chengxiu Ling, February 2018

011 ”How to Measure a Performance of a Collaborative Research Centre” by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018

012 ”Targeting customers for profit: An ensemble learning framework to support marketing decision making” by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018

013 ”Improving C rime Count Forecasts Using Twitter and Taxi Data” by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018

014 ”Price Discovery on Bitcoin Markets” by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018

015 ”Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance” by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018

016 ”Time -varying Limit Order Book Networks” by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018

017 ”Regularization Approach for NetworkModeling of German EnergyMarket” by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018

018 ”Adaptive Nonparametric Clustering” by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018

019 ”Lasso, knock off and Ga ussian covariates: a comparison” by Laurie Davies, May 2018

020 ”A Regime Shift Model with Nonparametric Switching Mechanism” by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, May 2018

021 ”LASSO -Driven Inference in Time and Space” by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

022 ”Learning from Errors: The case of monetary and fiscal policy regimes” by Andreas Tryphonides, June 2018

023 ”Textual Sentiment , Option Characteristics, and Stock Return Predictability” by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018

024 ”Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance” by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018

025 ”Construction of Non -asymptotic Confidence Sets in 2 -Wasserstein Space” by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018

026 ”Large ball pro babilities, Gaussian comparison and anti -concentration” by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018

027 ”Bayesian inference for spectral projectors of covariance matrix” by Igor Silin, Vladimir Spokoiny, June 2018

028 ”Toolbox: Gaussian comparison on Eucledian balls” by Andzhey Koziuk, Vladimir Spokoiny, June 2018

029 ”Pointwise adaptation via stagewise ag gregation of local estimates for multiclass classification” by Nikita Puchkin, Vladimir Spokoiny, June 2018

030 ”Gaussian Process Forecast with multidimensional distributional entri es” by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018

031 ”Instrumental variables regression” by Andzhey Koziuk, Vladimir Spokoiny, June 2018

032 ”Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective” by Li Guo, Yubo Tao, Wolfgang Karl Härdle, July 2018

033 ”Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present” by Natalie Packham, August 2018

034 ”A factor -model appro ach for correlation scenarios and correlation stress -testing” by Natalie Packham, Fabian Woebbeking, August 2018

035 ”Correlation Under Stress In Normal Variance Mixture Models” by Michael Kalkbrener, Natalie Packham, August 2018

036 ”Model risk of contingent claims” by Nils Detering, Natalie Packham, August 2018

037 ”Default probabilities and default correlations under stress” by Natalie Packham, Michael Kalkbrener, Ludger Overbeck, August 2018

038 ”Tail -Risk Protection Trading Strategies” by Natalie Packham, Jochen Papenbrock, Peter Schwendner, Fabian Woebbeking, August 2018

039 ”Penalized Adaptive Forecasting with Large Information Sets and Structural Changes” by Lenka Zbonakova, Xinjue Li, Wolfgang Karl Härdle, August 2018

040 ”Complete Convergenc e and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables” by Ji Gao YAN, August 2018

041 ”On complete convergence in Marcinkiewicz -Zygmund type SLLN for random variables” by Anna Kuczmaszewska, Ji Gao YAN, August 2018

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche
Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

042 ”On Complete Convergence in Marcinkiewicz -Zygmund Type SLLN for END Random Variables and its Applications” by Ji Gao YAN, August 2018

043 ”Textual Sentiment and Sector specific reaction” by Elisabeth Bommes, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, September 2018

044 ”Understanding Cryptocurrencies” by Wolfgang Karl Härdle, Campbell R. Harvey, Raphael C. G. Reule, September 2018

045 ”Predicative Ability of Similarity -based Futures Trading Strategies” by Hsin-Yu Chiu, Mi-Hsiu Chiang, Wei-Yu Kuo, September 2018

046 ”Forecasting the Term Structure of Option Implied Volatility: The Power of an Adaptive Method” by Ying Chen, Qian Han, Linlin Niu, September 2018

047 ”Inferences for a Partially Varying Coefficient Model With Endoge nous Regressors” by Zongwu Cai, Ying Fang, Ming Lin, Jia Su, October 2018

048 ”A Regime Shift Model with Nonparametric Switching Mechanism” by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, October 2018

049 ”Strict Stationarity Testing and GLAD Estimat ion of Double Autoregressive Models” by Shaojun Guo, Dong Li, Muyi Li, October 2018

050 ”Variable selection and direction estimation for single -index models via DC -TGDR method” by Wei Zhong, Xi Liu, Shuangge Ma, October 2018

051 ”Property Investment and Rental Rate under Housing Price Uncertainty: A Real Options Approach” by Honglin Wang, Fan Yu, Yinggang Zhou, October 2018

052 ”Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective” by Qingliang Fan, Wei Zhong, October 2018

053 ”The impact of temperature on gaming productivity: evidence from online games” by Xiaojia Bao, Qingliang Fan, October 2018

054 ”Topic Modeling for Analyzing Open -Ended Survey Responses” by Andra-Selina Pietsch, Stefan Lessmann, October 2018

055 ”Estimation of the discontinuous leverage effect: Evidence from the NASDAQ order book” by Markus Bibinger, Christopher Neely, Lars Winkelmann, October 2018

056 ”Cryptocurrencies, Metcalfe's law and LPPL models” by Daniel Traian Pele, Miruna Mazurencu-Marinescu-Pele, October 2018

057 ”Trending Mixture Copula Models with Copula Selection” by Bingduo Yang, Zongwu Cai, Christian M. Hafner, Guannan Liu, October 2018

058 ”Investing with cryptocurrencies Œ evaluating the potential of portfolio allocation strategies” by Alla Petukhina, Simon Trimborn, Wolfgang Karl Härdle, Hermann Elendner, October 2018

059 ”Towards the interpretation of time -varying regularization parameters in streaming penalized regression models” by Lenka Zbonakova, Ricardo Pio Monti, Wolfgang Karl Härdle, October 2018

060 ”Residual's Influence Index (Rinfin), Bad Leverage And Unmaskin g In High Dimensional L2 -Regression” by Yannis G. Yatracos, October 2018

061 ”Plug -In L2 -Upper Error Bounds In Deconvolution, For A Mixing Density Estimate In R d And For Its Derivatives” by Yannis G. Yatracos, October 2018

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

062 ”Conversion uplift in e -com merce: A systematic benchmark of modeling strategies” by Robin Gubela, Artem Bequé, Fabian Gebert, Stefan Lessmann, November 2018

063 ”Causal I nference using Machine Learning. An Evaluation of recent Methods through Simulations” by Daniel Jacob, Stefan Lessmann, Wolfgang Karl Härdle, November 2018

064 ”Semiparametric Estimation and Variable Selection for Single-index Copula Models” by Bingduo Yang, Christian M. Hafner, Guannan Liu, Wei Long, December 2018

065 ”Price Management in the Used-Car Market: An Evaluation of Survival Analysis” by Alexander Born, Nikoleta Kovachka, Stefan Lessmann, Hsin-Vonn Seow, December 2018

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.