# Revealing the Potential of Survival Analysis at Online Market Places

Seminar Paper submitted

to

**Prof. Dr. Stefan Lessmann, Johannes Haupt and Annika Baumann**

Humboldt-Universität zu Berlin

School of Business and Economics

Chair of Information Systems

by

**Arvid Reiche and Kamarhulrhizwan Benjamin Jaidi**

(568292, 584018)

Berlin, March 15, 2018

# Abstract

The application of survival analysis methods in order to predict defaults of devices and survival rates of patiens has been common practice for many years. Apart from survival analysis' popularity in predicting events one would like to avoid, it can also be applied to predict positive events such as a sale to a customer. We evaluate different survival models with respect to their ability to predict a deactivation of a listing at online market places. The goal is to initiate a discussion about the potential usage of survival techniques to find Pareto optimums between the user and the lister's satisfaction and the companies revenue in case of a freemium business model.

# Contents

# List of Abbreviations

| | | | |
|---|---|---|---|
| ATF | accelerated time failure | AUC | area under the curve |
| PH | proportional hazard | ROC | receiver operating characteristic |

# List of Figures

# 1 Introduction

Survival analysis encompasses a variety of techniques to predict at which probability and especially when a subject under observation will experience a certain event. The aim is to find out the hazard of an event going to happen or the probability that a subject will survive, meaning the event did not occur within the given time frame. In contrast to predictive regression models such as logit or least squares which are suitable to find out whether a certain event is going to happen, survival analysis will provide information about the time to event and the chance of that event, at any point in time.

With the ever-increasing amount of data being generated in almost every business sector, the sensible employment of survival analysis techniques has become a key asset in ensuring future success for many companies. Typical use cases range from application in medical studies to assess the impact of various factors on the patients life expectancy to marketing and finance, where crucial information such as customer churn rates and credit default risks are calculated. Moreover, survival analysis has also found application in predictive maintenance which is aimed at detecting and preventing impending failure in technical devices while avoiding unnecessary replacement of healthy parts.

Apart from application by practitioners, survival analysis has also been subject to extensive research, focusing on benchmarking and improving existing techniques in several different settings such as gradient boosting[1], deep learning and the application of neural networks.

The goal of this paper is to evaluate several full, semi - and non-parametric survival analysis techniques using data taken from an international online market place for car sales. We aim at providing helpful advice to practitioners when it comes to choosing a model for predicting future events to enable them to make information-based product decisions.

[1]Chen, Y., Jia, Z., Mercola, D. and Xie, X., 2013. A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. Computational and Mathematical Methods in Medicine, 9 September.

# 2 Related Work

A broad selection of different techniques has been featured and analyzed in scientific literature. The widely used Cox PH model, for example, was featured in 1972 and has originally been proposed for usage in a medical context[2].

Backiel, Baesens and Claeskens evaluate the performance of logistic regression, neural networks and Cox PH in the field of churn prediction in the prepaid mobile market by ROC, AUC and lift in combination with a cost function for the expected maximum profit. Instead of assuming that all customers are independent, the authors also consider social network effects in their predictions. Their findings suggest that incorporating social network effects in customer churn prediction can improve prediction results and expected maximum profit although models are likely to become more time sensitive. However, Cox PH can deliver good results for the expected maximum value even when no network effects are incorporated in the model. Thus, the cost function has a decisive effect on prediction quality[3].

Another benchmark has been conducted by Dirick, Claeskens and Baesens[4]. In this paper, the authors benchmark accelerated failure time models, Cox PH and mixture cure models by AUC, default time prediction, as well as the ratio of expected future value and true future value, calculated by a cost function. According to their findings, the best overall performance is achieved by Cox PH. Mixture cure also shows good results in future loan estimation while AFT models are outperformed regularly.

[2]D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society B, vol. 34, no. 2, pp. 187220, 1972

[3]Backiel, A., Baesens, B. and Claeskens, G., 2016. Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. Journal of the Operational Research Society, 16 March, p. 11351145.

[4]Dirick, L., Claeskens, G. and Baesens, B., 2015. Time to default in credit scoring using survival analysis: a benchmark study, Leuven: KU Leuven Faculty of economics and Business.

# 3 Methodology

## 3.1 Scope of Paper

Taking an approach similar to the one featured in the papers in section two, we analyze models from the parametric, semi-parametric and non-parametric branches of survival analysis. Namely, we compare Kaplan Meier curves, Cox PH, an adaption of the Cox model to work with a gradient boosting model named Cox GBM, Accelerated Failure Time and a decision tree based model. In order to measure our results, AUC values and Concordance Index are used as benchmark metrics.

All calculations where conducted using the free statistical programming language R.

## 3.2 Method Overview

Survival analysis methods can be categorized into three main branches:

- Non-parametric models that rely only on the observed outcomes in the data to calculate the hazard and estimate future development without considering any covariates. Therefore, no assumption is imposed on survival time to follow any distribution.

- Semi-parametric models that take into consideration both - the underlying hazard as well as the relative influence of the covariates, the parametric part. These models assume a distribution only for the parametric part.

- Fully parametric models that assume a distribution for both the survival time and the covariates.

The techniques compared in the benchmark can be seen in the table below:

**Table 1:** Scope of Paper - Applied Method Overview

| Model | Feature |
|---|---|
| Kaplan Meier | Non-parametric |
| Decision Trees | Non-parametric |
| Accelerated Failure Time | Fully parametric |
| Cox Proportional Hazard | Semi-parametric |
| Gradient Boosting with Cox | Semi-parametric |

## 3.3 Techniques

### 3.3.1 Kaplan Meier

The Kaplan Meier Estimate is based on the survival function S(t) and therefore purely non-parametric[5]. The curve displays the probability that an event did not occur in i, hence one can observe the chances of surviving in a given length of time while considering many small time steps. The survival function at i is defined as:

$$\hat{S}(t) = \prod_{i:t_i \leq t}(1 - \frac{d_i}{n_i})$$

- $d_i$: Number of events
- $n_i$: Number of subjects at risk at time i
- $t$: Total time

The Kaplan Meier Estimate at point t is the sum of all survival probabilities of the previous time steps:

$$KaplanMeierEstimate = \prod_{i:t_i \leq t}(S_i)$$

Kaplan Meier is based on three strong assumptions:

- Censored subjects have the same survival prospects.
- Survival probabilities are the same for all subjects regardless of the timing when the subjects enter the observations.

[5]Kaplan, E. L.; Meier, P. (1958). "Nonparametric estimation from incomplete observations". J. Amer. Statist. Assn. 53 (282): 457481. doi:10.2307/2281868. JSTOR 2281868.

- Events happen at the time specified.

Especially the first and second assumption are hard to find since the survival does not always only depend on time only and seasonality is completely ignored. Nevertheless, Kaplan Meier is a good choice for a first visualization of the survival object and its survival probabilities. The survfit function in R-package "survial" allows to model the survival objects by characteristics, such as treatment method for a patient or customer type. These first insights can be very useful to understand the observed data.

### 3.3.2 Accelerated Time Failure

In contrast to the Kaplan Meier Estimate, the AFT model is characterized by being a fully parametric survival model. The explanatory variables act as accelerating factors to speed up or slow down the survival process as compared to the baseline survival function $\hat{S}(t)$[6].

$$S(t|X) = S_0(t * exp(-\beta \times X))$$

It follows that the event rate is slowed down whenever $0 < exp(-\beta \times X) < 1$ and is speeded up when $exp(-\beta \times X) > 1$. The hazard function of AFT models is therefore obviously different from the hazard function of Cox models. The individual hazard at time t is:

$$h(t|X) = h_0(t * exp(-\beta \times X)) * exp(-\beta \times X))$$

The advantage of the AFT model is that the interpretation of the coefficients is straight forward. While positive coefficients imply that the hazard rate is increasing, negative coefficients force it to decrease, hence the survival time is lengthened. This is due to the dynamics of the regressors to act as accelerators. If a coefficient is for example 0.2, it means the variable is reducing the survival time, hence, the event of death is observed five times faster in this case.

Within our final model set up, we are using the function *survreg* from the Rpackage *survival* in order to regress the AFT models. Three different versions of the error term

[6]Kalbfleisch and Prentice (2002). The Statistical Analysis of Failure Time Data (2nd ed.). Hoboken, NJ: Wiley Series in Probability and Statistics.

distribution are tested: weibull, exponential and loglogistic. The model performance of the AFT versions on our data set perform poorly. Models including a similar regressor set like our Cox models described below are resulting in almost random prediction while models including a small set of performance and car features such as the sum of all detail and app result page views and price are working but overall produce medium results.

### 3.3.3 Cox Proportional Hazard

The Cox PH model, introduced in 1972[7], is one of the most popular models used in survival analysis. Its goal is to estimate the influence of independent co-variates on the hazard of an item under observation to experience an event or on the time-to-event, respectively. It is also suitable for repeated measurements for each subject under observation, i.e. time varying covariates. The hazard function is composed of the product of a baseline hazard $\lambda_0(t)$ that has the same value for all objects under observation at a certain time and the proportional hazard $\exp\left\{x^T\theta\right\}$ which represents the effect of the covariates on the life time of a subject.

The hazard rate $hr$ for each covariate $\exp(x_i)$ can be interpreted as follows:

- $hr = 1$: covariate has no effect
- $hr > 1$: covariate increases hazard
- $hr < 1$ : covariate decreases hazard

The hazard function is expressed by:

$$\lambda\left(t|x,\theta\right) = \lambda_0\left(t\right) exp\left\{x^T\theta\right\}$$

Cox proposed, that in order to find out the relative effect of the covariates, $\hat{}_0(t)$ does not need to be specified. Instead, partial likelihood is used:

$$L_p\left(\theta;\{x_i,t_i,\delta_i\}_i^n = 1\right) = \prod_{i\in E} \frac{exp\left\{\theta^T x_i\right\}}{\sum_j : t_j \geq t_i} exp\left\{\theta^T x_j\right\}$$

[7]D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society B, vol. 34, no. 2, pp. 187220, 1972

Within our model set up, we are using the coxph function from survival package to produce the regressions. We tested various regressor sets and decided on the following model:

$$Surv(dateinteger, status == 1) =$$
$$savedorcontacted + mileagegroups + bodycoloridfilled + bodytypeidfilled + fuelid +$$
$$log(price) + imagesnumbergroup + detailandappresultpageview$$

### 3.3.4 Gradient Boosting with Cox

Since the performance of the final model is not satisfying, we decided to improve the model further by gradient boosting method. Gradient Boosting comes from the family of ensemble learning techniques. A number of weak learners is sequentially fitted and added to the model, so that a cost function is minimized. The implementation that was used in the benchmark was the *gbm* Rpackage, the cost function corresponds to the Cox PH distribution. We are using the following code snippet to boost our above described cox model:

```
gbm_1_3 = gbm(data_final_clean.cox_1_3,
              data = train_1_3,
              distribution = "coxph",
              n.trees = 2500,
              shrinkage = 0.02,
              n.minobsinnode = 4)
```

**Figure 1:** Code Snippet for Gradient Boosting

After the boosting the cox model produces acceptable results in terms of our benchmarking indicators described below.

### 3.3.5 Decision Trees

Decision trees are a type of supervised learning algorithm often used for classification problems[8]. The general idea is that the root, the entire set or sample population, is split by the decision tree model into two or more homogenous sets by the variable that achieves the most homogenous split. The same process is then applied to the new decision nodes which are splitting the data again into subsets, creating new branches of the tree. Thus the generation of the new sub-nodes increases the homogeneity of resultant sub-nodes further with respect to the target variable. Once the results in terms of outcome are homogeneous and no further improvement can be achieved, the tree ends in the terminal nodes.

The advantages of decision tree models are numerous. Decision trees are easy to understand, useful in data exploration, can handle numeric and categorical parameters and is considered to be non-parametric as decision trees have no assumptions about the space distribution and the classifier structure. However, the risk of overfitting exists, and decision trees cannot handle continuous variables as information is lost when categorizing variables in different categories.

We are using the function *rpart* from the Rpackage "rpart" to estimate the tree models. Our plots of the decision tree are generated by the Rpackage "partykit" which is very helpful as it produces not only the tree but also visualizations of the corresponding Kaplan-Meier curves at the terminal nodes. We are testing the model on the similar set of regressors like in our Cox model. The resulting decision tree reveals the dominance of the performance features.

[8]Tree-Based Methods for Survival Data Atanu Biswas, Sujay Datta, Jason P. Fine and Mark R. Segal Mousumi Banerjee and Anne-Michelle Noone, Published Online: 23 MAR 2007DOI: 10.1002/9780470181218.ch16.

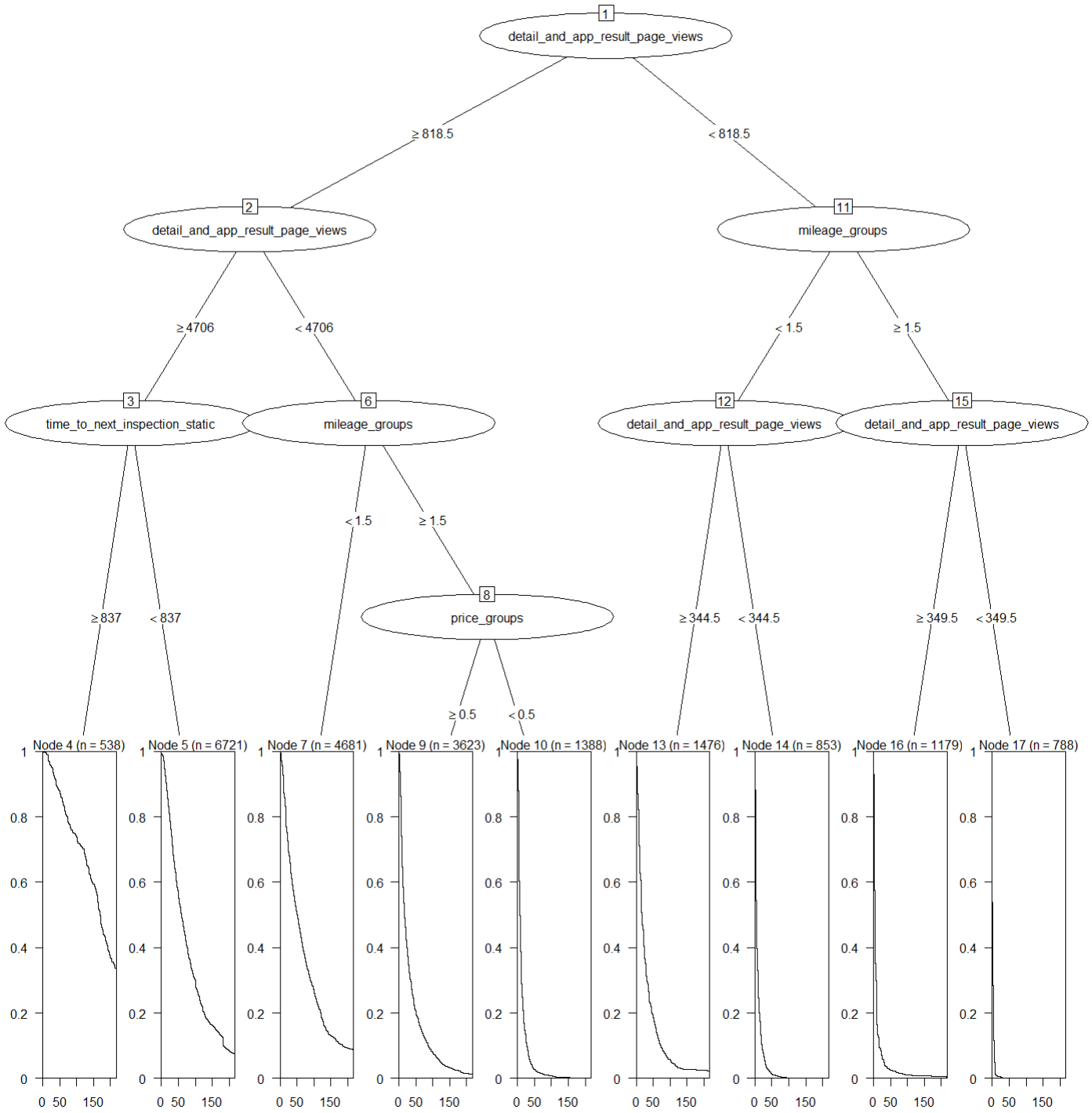


**Figure 2:** Decision Tree Plot with package partykit

## 3.4 Dataset Description

We are using a data set from an international online market place for car listings for private and professional sellers. The original data set is from 2017 and 2018 and contains 216 days with 1.380.814 observations of 23.393 cars that all have been listed at the same day on the platform by 7.951 sellers. Originally the set contained 57 features, which we extended to 74 features by grouping and clustering some data and by calculation. Moreover the status of the listing, the date of de-listing and other important features required to perform survival analysis techniques had to be calculated and added to the original data set. The result is a cleaned final data set which includes only 21.247 observations of the same number of cars. The set is simplified by just including the event of de-listing the car from the page, excluding the observations of a listing being active except for the end date after 216 days. Since all listings are activated on the same day, this modification is feasible.

There are four types of features included in the data sets. The first type are the characteristics of the car such as price, mileage, power, brand, co2emission. The second type are characteristics of the listing such as whether the color of the car is declared within the listing and for example whether or not the information on the displacement is filled. In most cases this is a boolean value indicating 0 to be unfilled and 1 to be filled. The third type are performance features such as how often the listing was shown in the result list on different devices, how often it was bookmarked or saved and how often searchers did click on the listing to see details. The forth type are calculated features that are based on the existing 57 features such as age of the car and the time that is left till the next inspection is due and groups such as *mileagegroup* that cluster different mileage numbers into groups to control for outliers. Below is a selection of features

| Dataset | Nr. Observations | Nr. Features | Time Varying Covariates | Censored | Type |
|---|---|---|---|---|---|
| Raw data set | 1.380.814 | 57 | yes | right | Online market place for cars |
| Final clean data set | 21.247 | 74 | no | right | Online market place for cars |

**Figure 3:** Data Set Overview

| Feature name | Description | Feature Type | Data Type | Min | Max | NAs |
|---|---|---|---|---|---|---|
| body_color_id_filled | Whether or not the seller declared the color of the car listed | Second type | Boolean | 0 | 1 | no |
| customer_type | Private or professional (0,1 respectively) | First type | Boolean | 0 | 1 | no |
| detail_and_app result_page_views | Sum of all app and detail resultlist page views over all days by listing | Third type | Int | 0 | 573325 | no |
| fuel_id | Diesel or petrol, (0,1 respectively) | First type | Boolean | 0 | 1 | yes |
| images_group_number | Number of images attached to a listing, grouped | Forth type | Int | 0 | 3 | no |
| last_belt_service_filled | Whether or not the seller declared the last belt service of the car listed | Second type | Boolean | 0 | 1 | no |
| make_group | Brand of the listed car, grouped | Forth type | Int | 0 | 3 | no |
| mileage_groups | Mileage of the listed car, grouped | Forth type | Int | 0 | 9.999.999 | yes |
| transmission_id | Tranmission of the car, manual, automatic or other (1, 2, 0 respectively) | First type | Int | 0 | 2 | no |
| price | Sale price of the listed car | First type | Int | 0 | 2.796.500 | yes |
| saved_or_contacted | Sum of all bookmarks or contact emails over all days by listing | Third type | Int | 0 | 3592 | no |
| status | Status of the listing, active after end period or de-listed (0, 1 respectively) | Forth type | Boolean | 0 | 1 | no |



**Figure 4:** Selected Feature Overview

The final clean data set is no longer including NAs for price, customer type, saved or contacted, mileage groups, fuel id, last belt service filled, body color id filled, images number group, make group and detail and app result page views. Moreover, we did exclude de-listings of cars that have never been contacted by email or bookmarked and received less than five clicks in total. We decided on five as a threshold as the seller of the car will click on his own listing at least once to check whether it is live or not. Without the filtering of listings that were deactivated without experiencing interaction with the searcher by clicks or bookmarks, the impact of some features will be under or overestimated.

Finally, certain feature groups should be explained in order to understand the Kaplan Meier plots and the parametric part of the final models.

| **Feature name** | **Mapping** |
|---|---|
| age_groups | 0 = car is less than one year old<br>1 = car is between one to three years old<br>2 = car is between three to ten years old<br>3 = car is older than ten years |
| images_group_id | 0 = no images uploaded<br>1 = 1 – 5 images uploaded<br>2 = 6 – 10 images uploaded<br>3 = more than 10 images uploaded |
| make_group | 0 = all other brands<br>1 = selection of biggest low budget brands like Suzuki, Kia, Hyundai<br>2 = selection of biggest mid budget brands like Toyota, VW, Ford<br>3 = selection of biggest high budget brands like Mercedes, Porsche, Audi |
| mileage_groups | 0 = mileage is under 5.001 kilometres<br>1 = mileage is between 5.001 and 45.000 kilometres<br>2 = mileage is between 45.001 and 105.000 kilometres<br>3 = mileage is more than 105.001 kilometres |
| previous_owners | 0 = no previous owner<br>1 = one previous owner<br>2 = two previous owners<br>3 = more than two previous owners |
| price_groups | 0 = sales price is below 5.001€<br>1 = sales price is between 5.001€ and 15.000€<br>2 = sales price is between 15.001€ and 35.000€<br>3 = sales price is more than 35.000€ |

**Figure 5:** Selected Group Overview

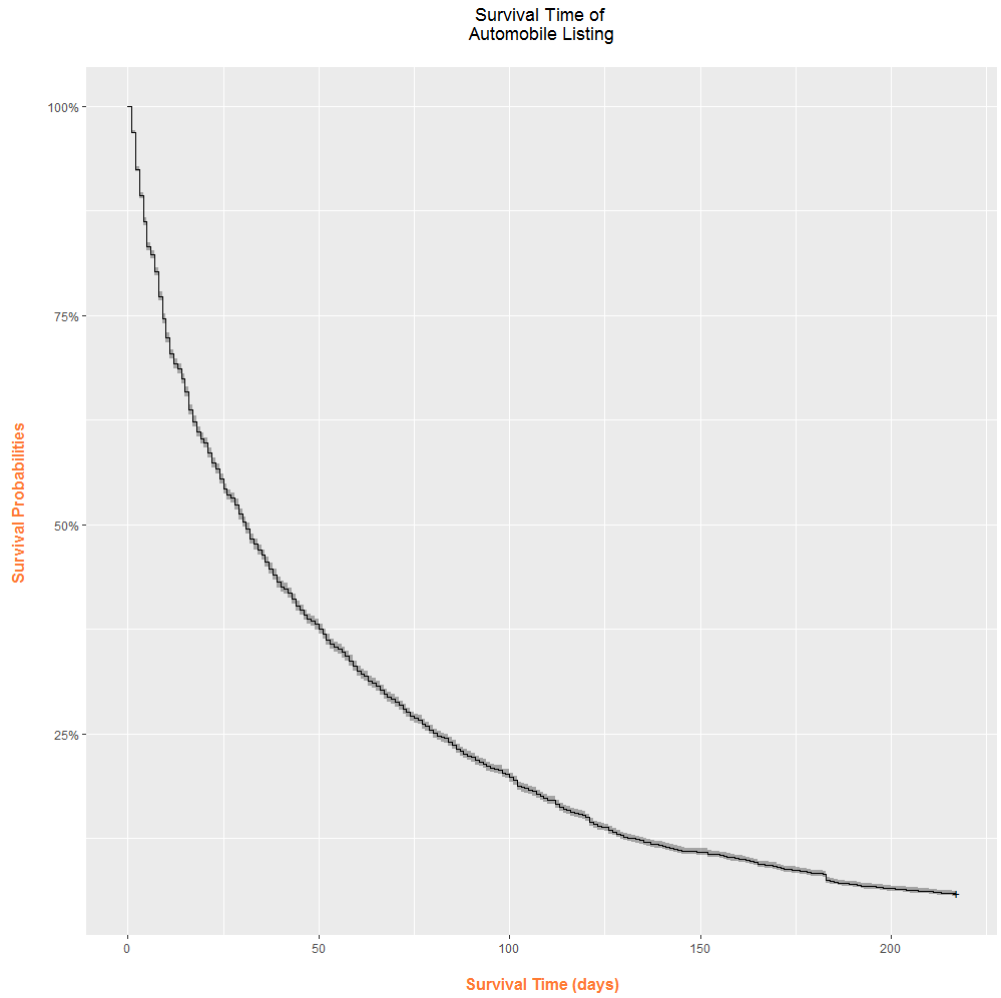


**Figure 6:** Kaplan Meier Curve

The general survival curve is smooth and has no great variations at any point in time. A small dip appears at around 185 days representing a bulk deletion of listings of private customers at this given point in time by the host. The reason for this deletion is unknown. Overall almost all listings are deactivated within the time horizon. Out of 21.247 observed subjects only 1181 are still active after 216 days of observation. The data looks very homogenous but when plotting Kaplan Meier curves based on single features, differences exist.

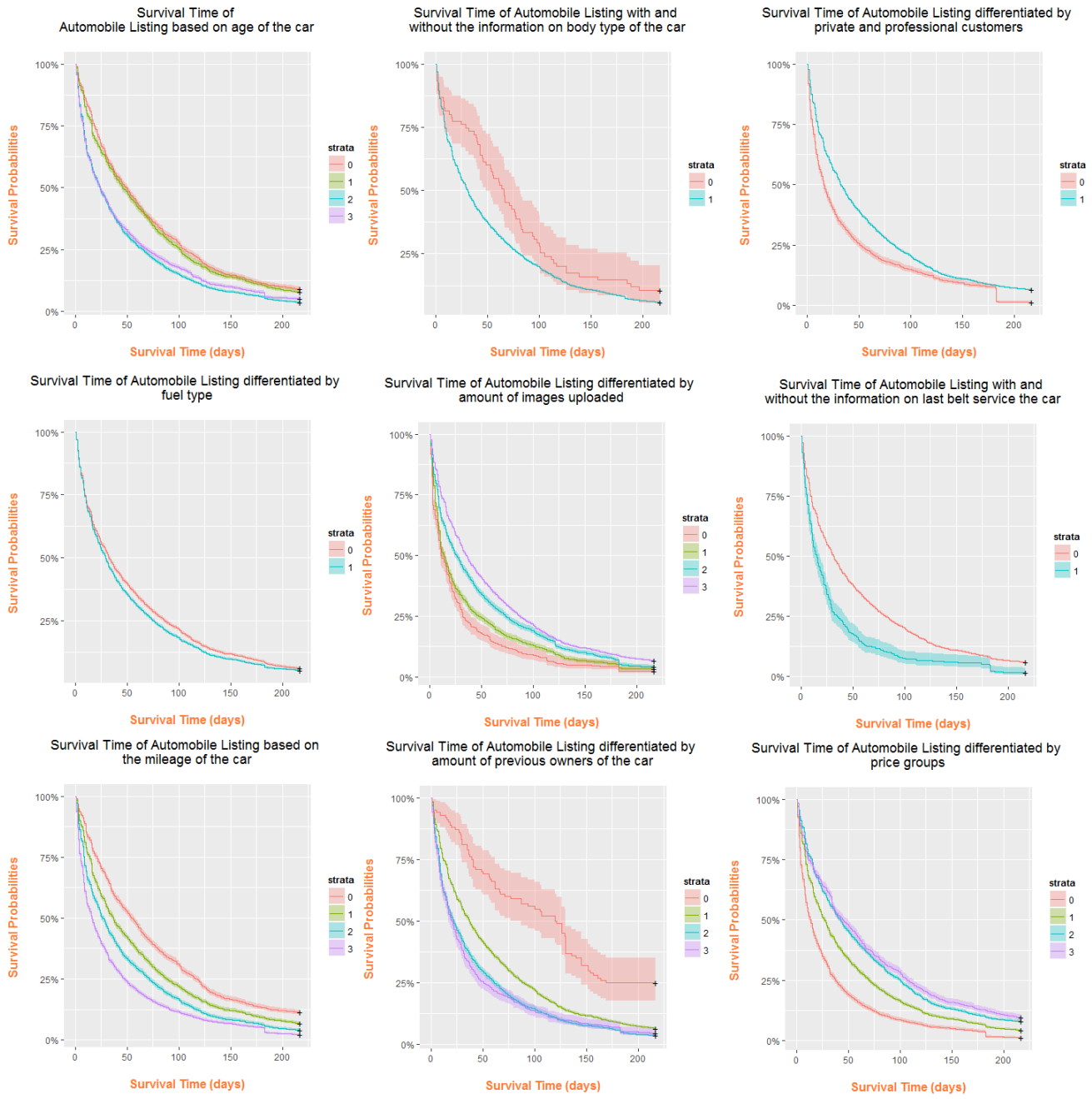


**Figure 7:** Kaplan Meier Curves on Selected Features

First of all, the age of the car seems to drive the survival probabilities. The older the car, the less time it takes to see a deactivation of the listing, hence the lower are the survival chances. The same holds for number of previous owners, number of images and mileage of the car. It seems that the older the car, the more mileage it has, the less

images are uploaded and the more previous owners it had, the higher the chances that the listing is deactivated. This is most likely due to the fact that all these features are correlated with the price of the car. Taking a look at the Kaplan Meier curve on price groups, it is obvious that the lower the sales price of the car, the higher the probability of death of the observed subject. Most cars being sold online are used cars. There is a large group of users that do look for cheap cars online to re-sell them offline. Moreover most of the car sales in Germany are resales of existing cars. The number of new cars registered is increasing within the last year to around 3.44 Mio cars[9] resulting in about 47 percentage of all car sales in Germany[10]. This could be the reason why listings of professionals are less likely to be deactivated faster than listings of privates. Dealers usually offer more expensive cars, partially even new cars. These kind of listings require detailed pictures of the features than cheaper cars that are offered for a fraction of the sales price. Hence, it is not unfeasible to observe these relationships.

It is surprising that the difference between diesel and petrol is very little. Listing features such as if the last belt service field is filled or if the shell type is filled impact the survival rate. This is reasonable as the information is useful to the user and if it is missing, the listing may appear incomplete, less attractive or even likely to be fraud.

## 3.5 Evaluation Metrics

### 3.5.1 Area Under The Curve

It is very popular in survival analysis research to benchmark models based on the AUC value of the model as the value is easy to understand and can be computed for many different kind of models. The following steps are always the same:

- Train the model based on a train set.
- Save the predictions of the performed model.

[9] $https://www.kba.de/DE/Statistik/Fahrzeuge/Neuzulassungen/n_jahresbilanz.html$, Accessed 13.03.2018.

[10] $https://www.kba.de/DE/Statistik/Fahrzeuge/Besitzumschreibungen/besitzumschreibungen_node.html$, Accessed 13.03.2018.

- Compare the saved predictions of the model with the actual values of the train set.
- Plot the true positive rate versus the false positive rate.
- Calculate the integral.

The AUC value is the integral of the plot of true positive rate versus false positive rate, hence the area under the curve. The advantage of AUC is that evaluation is possible at any point in time in the survival curve. Moreover, it is possible to impose restrictions on the false positive rate. If for example one false positive prediction is costlier than a true positive prediction, it is possible to restrict the AUC calculation only on a certain max false positive rate. This feature is very often used in clinical studies as a false positive prediction can cause costs and even death of the patient. The plot below the receiver operating characteristic curve is shown for the Cox model without gradient boosting on a 50 percentages train set out of the full data set.

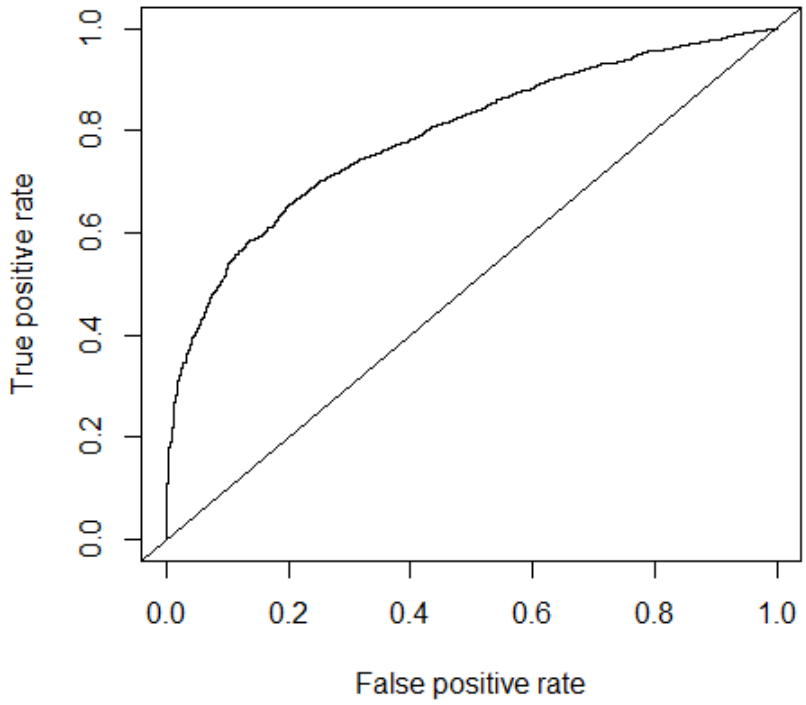


**Figure 8:** ROC of Cox Proportional Hazard Model

### 3.5.2 Concordance Interval

The CI is a global index to measure the performance of survival models. When predicting survival times, the output of a model can be interpreted as a ranking based on survival times. The subset of correctly ordered pairs of observations divided by the number of observations that can be ordered results in the CI.

$$CI = \frac{1}{|\rho|} \sum_{(i,j)\in\rho} I\left(F(x_i) < F(x_j)\right) = \frac{1}{|\rho|} \sum_{i\in E} \sum_{j:t_j>t_i} I\left(F(x_i) < F(x_j)\right)$$

- $\rho$ : The set of orderable pairs where $t_i < t_j$
- $|\rho|$ : number of pairs in $\rho$
- $F(x)$:function to predict survival time
- $I$: Indicator whether condition in (...) has been met

# 4 Results

We performed various survival analysis techniques on the data. We are using a cox proportional hazard model, three versions of accelerated time failure models, a tree model and finally use gradient boosting to improve our cox model. We are using three versions of data to calculate the models. We split therefore our data set into a training set which is either one third or two third of the original data. Moreover, we are calculating the same models on the full set, too. Hence, we evaluate six versions of the cox model, including gradient boosting, nine versions of the AFT models and three versions of the tree model, resulting in 18 models in total. In this study the performance of the models is not evaluated on AUC solely, but the AUC is weighted equally to the Concordance Index.

| Model | Area under the curve | Concordance Indices |
|---|---|---|
| cox_1_3 | 0.779 | 0.706 |
| cox_2_3 | 0.790 | 0.702 |
| cox_3_3 | 0.788 | 0.702 |
| gbm_1_3 | **0.884** | **0.760** |
| gbm_2_3 | **0.875** | **0.756** |
| gbm_3_3 | **0.869** | **0.757** |
| tree_1_3 | 0.758 | 0.621 |
| tree_2_3 | 0.768 | 0.634 |
| tree_3_3 | 0.722 | 0.393 |
| atf_model_weibull_1_3 | 0.777 | 0.693 |
| atf_model_weibull_2_3 | 0.787 | 0.700 |
| atf_model_weibull_3_3 | 0.783 | 0.701 |
| atf_model_exponential_1_3 | 0.777 | 0.696 |
| atf_model_exponential_2_3 | 0.787 | 0.701 |
| atf_model_exponential_3_3 | 0.783 | 0.698 |
| atf_model_log_1_3* | 0.783 | 0.693 |
| atf_model_log_2_3 | 0.592 | 0.685 |
| atf_model_log_3_3 | 0.788 | 0.703 |

*The ATF log_1_3 includes all variables that the cox, tree and gbm models have
The three models with best values for both, AUC and CI are highlighted

**Figure 9:** Model Comparison by AUC and CI

The best performing models on the data set are the Cox models that are enhanced with gradient boosting algorithm. Although some of the AFT models have fair CI and AUC values, they generally underperform the Cox proportional hazard models. The tree models show below the average AUC values and have the lowest CI of all models.

The variables within the almost all AFT models are only a subset of variables used in the Cox and tree models. Our initial assumption was that the survival rate can be explained by characteristics of the car such as price, brand, mileage, fuel type and transmission id. It turns out that almost all of these variables are not driving the survival rate. Instead we had to include performance variables as proxy variables for the market demand, represented by search behavior and interest of users, indicated by interaction with the listing and presence of it to the users. Since many characteristics of a car are potentially correlated with these performance variables, we had to abandon several variables of interest such as model of the car and transmission.

After we encountered that our initial assumptions on relevance of variables were invalid, we found it very hard to come up with a valid set of parameters. One of the reasons is that model specification on semi-parametric models is different from fully parametric cases in which one can perform a RESET test for example to test for non-linear relationships. Another reason is the uncertainty on timing of the actual de-listing event. A listing can only be manually de-activated. We have to impose the strong assumption that all sellers de-activate their listing within a short period of time after they have actually sold it. This is a very strong assumption and we cannot control for outliers that leave the listing active for a longer period of time than necessary. Another reason is that the search behavior and intent of the searchers at online market places for cars is very heterogenous. There are at least three different groups of searchers. The first group is browsing very specifically for one or a small selection of cars with certain specs. The second group of users is less fixed on a specific car or brand. Their intent is usually to find cars that match their budget and fulfill the basic needs. This group is usually less informed and is actively using the market place to learn about the market supply and find a car. The third group of searchers is none of the above as their intent is not to buy a car, at least not in the active session. Instead the users only browse the market place to inform themselves, to check out

new cars or simply to inspire themselves. Thus, not every bookmark event in Google Analytics is of the same value. It is different for these user groups. A bookmark from the first group of users is more likely to convert to an actual sale than a bookmark of a user from group three.

Since we cannot distinguish between the searcher groups, we introduced the feature *makegroup* that would at least differentiate brands of cars. The assumption is that users from group one are more likely to browse for top end brands while users from group two are more interested in smaller, low budget cars. However, we cannot effectively control for the impact of the user groups. This leads to a bias of our results when we run the models on the entire data set. The difference becomes obvious when one plots Kaplan Meier curves on data subsets for make group $1, 2$ and $3$.

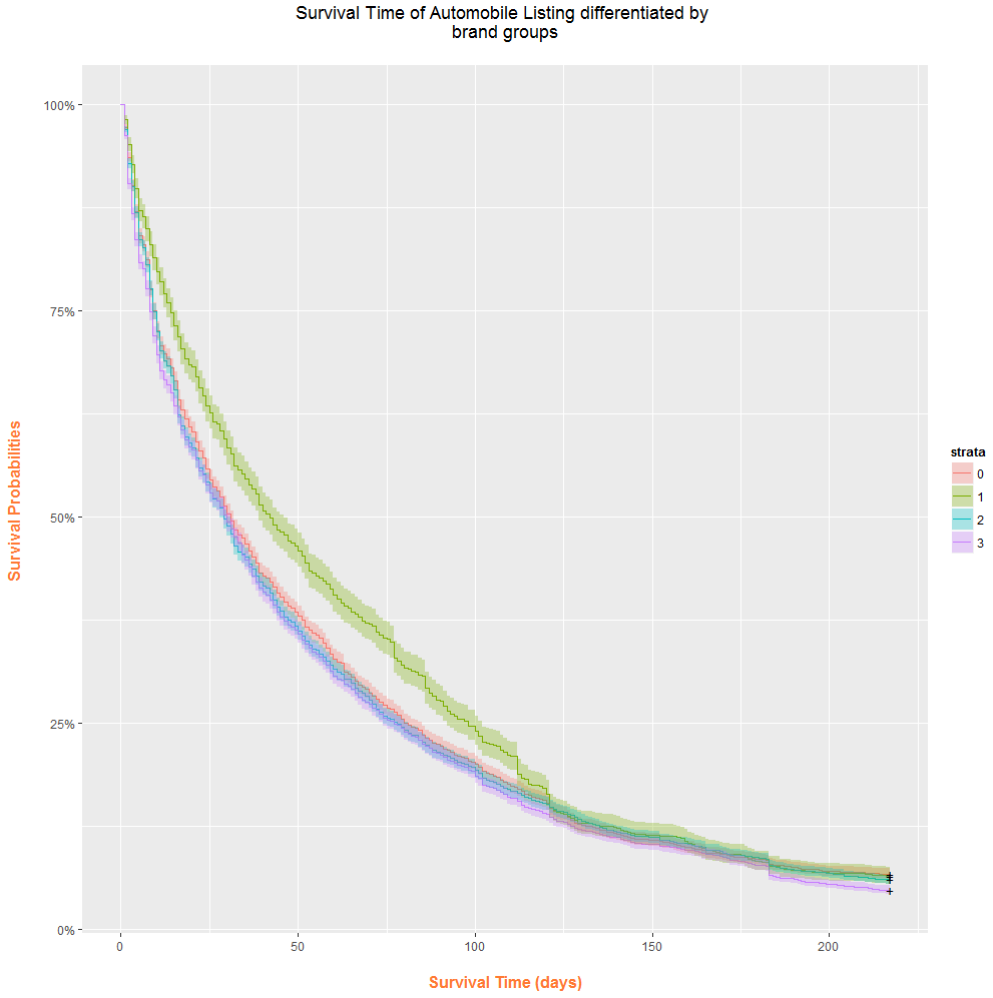


**Figure 10:** Kaplan Meier Curve based on Price Groups

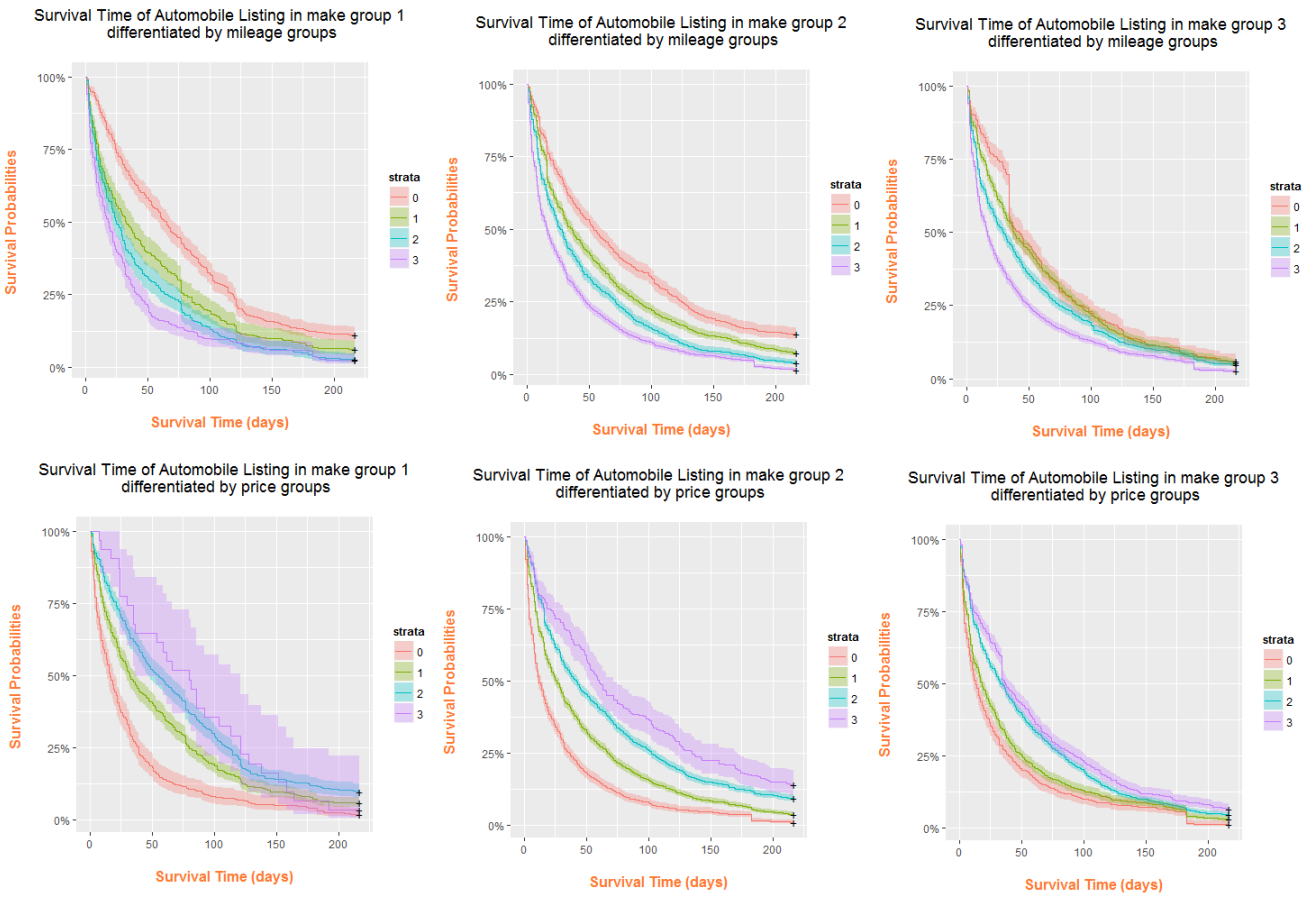


**Figure 11:** Kaplan Meier Curves of Price Groups

Especially the impact of the mileage groups and price group parameters is different for the three groups. Our models however overall perform equally in all three sub data sets. In comparison with the tree model on the full data set, the importance of price groups and mileage groups varies among the sets. For example, mileage groups are of more importance in make group 1 for survival times than in group 3. In group 2 the importance is split equally between price and mileage groups. However, the parameter with the greatest influence among all models and all subsets of data is how often a listing is displayed within the result list, regardless of the device type.

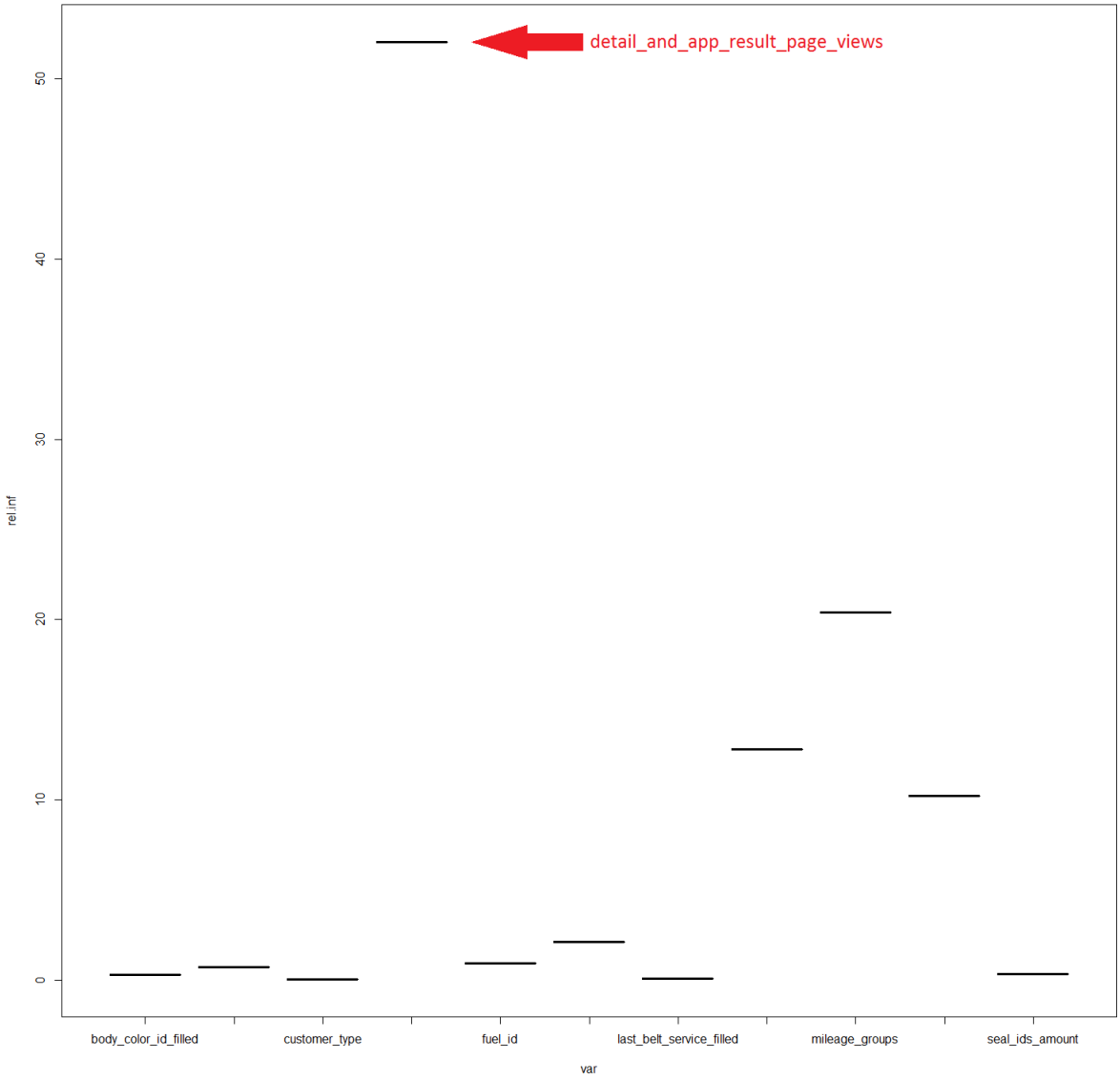


**Figure 12:** Gradient Boosting - Relevance of Features

Its dominance is revealed by the decision tree and the summary of the gradient boosting model above. Thus, the gradient boosting of the cox proportional hazard model is still the best model, although the impact of the different parameters may vary from make group to make group.

This an important finding because a company running an online market place for cars has limited options to generate incremental value for the user and the company.

The ultimate aim of the host of an online market place for commodities is to encourage sellers to publish frequently new listings and at the same time to reduce the time a listing is online to a certain threshold. The assumption is that once a listing is deactivated, the car is sold. Thus, reducing the survival time in this case is beneficial for the seller. If the sale speed is high, sellers are likely to return to list new cars. Searchers on the other hand always want to explore new cars when they return to the market place.

Thus it is clearly an desired objective to reduce survival time. Online market places are often forced to run freemium business models. Hence, the listing itself is for free. The company makes money by display media sale and on top products. The design of these products is complicated since the market demand determines how fast a car will be sold and if the listing itself is for free, it is hard to argue why a user should by an on top product. As our models show, various parameters impact the survival time. Thus, it is in the interest of the host to construct the freemium listing product in a way that the survival time is not minimized entirely. Instead it should be a pareto optimum between acceptance to wait for the seller, available listings to match market demand and revenue for the company by media and on top products. The searcher is monetized by media on the page and the seller can be monetized by offering on top products that feature even shorter survival time.

Although the parameter influence varies between the sets, it is certain that the performance feature on how often the listing was displayed to a user impacts the survival time. Certainly this number depends heavily on the search behavior of users. The demand for certain listings will naturally be significantly lower than for others. A product feature that could overcome this lag of interest is to display in a special location within the result list. The functionality could be that the user gets to see "*recommended*" cars that also fit, for example, in the selected price range. The impact on the user group one that is browsing for very specific cars will most likely be neglectable but the assumption of impact on the user group two is feasible.

We suggest a product feature that enables the listing to be shown 20 percent more often than the average result list page views of cars within the same price group. The

graph below shows the survival function based on Kaplan Meier for a mean listing in price group 1.

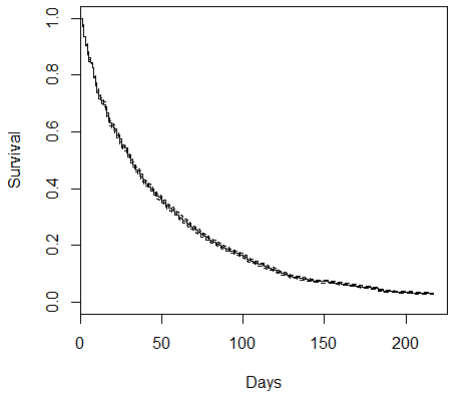


**Figure 13:** Survival Curve of a mean listing in group 1

The impact of such a feature is measurable. Assume the product is priced at 10.00 one-time payment. Further we naively assume user group one, two and three are equally distributed among the searchers. Thus about 33 percent of the 16.3646.022 page views are from group two users. The median on page views for listings in price group 1 is 2356.5. Thus, the weighted median of page views for price group 1 listings based on assumption that user group two makes 33 percent of the page views is 777.6. The average survival time of price group 1 is 43.6 days. Taking the cox proportional hazard model, the survival time of a mean listing is reduced by seven percent if the page views are increased by 20 percent more page views.

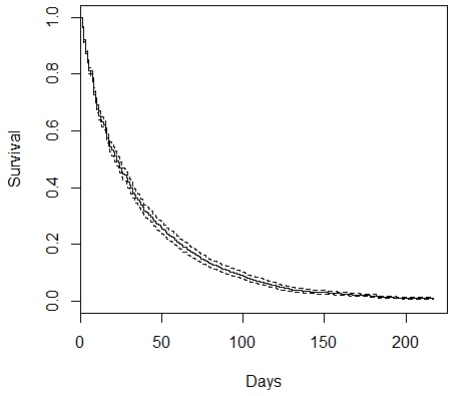


**Figure 14:** Survival Curve of a mean listing in group 1 with On-Top-Product

If we assume that only 10 percent of the sellers are interested in this product for the price group 1, we have 385 sellers with 770 listings per day, resulting in 7.700 income per day or about 2.8 million per annum.

# 5 Conclusions

We have tested five different survival techniques and encountered that all models in their normal form do not predict the survival rates accurate enough to be implemented. However, the enhancement of Cox proportional hazard model by gradient boosting algorithm leads to improved AUC and CI values, indicating a superior performance towards the other models. Moreover, the data shows that the demand patterns of searchers should be taken more into considerations. Different user groups convert by heterogenous speed. A classification of the searchers can eventually be performed once the performance data of the listings is enriched with search data. For example, information on filters being applied while a listing was displayed may help to cluster and differentiate users.

The analysis of the market place data shows that further research is required to reveal the potential of on-top-products designed by survival analysis. Potentially the market places should first join data from listing and performance features with revenue data from on-top-products and media display and customer satisfaction data. By this

a Pareto optimum between revenue, user satisfaction and sale time can be eventually found by survival analysis techniques, leading to higher sales, more income and greater user satisfaction.

# Declaration of Authorship

I hereby confirm that I have authored this seminar paper independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, March 15, 2018

Arvid Reiche, Kamarhulrhizwan Benjamin Jaidi