Revving in Profit from the Used Luxury Car Market using Survival Analysis Techniques:

A Case of the German Used Luxury Car Market

N. Kovachka¹, S. Lessmann¹ and H.V. Seow²

¹Faculty of Economics and Business Administration
Humboldt University of Berlin.

²Nottingham University Business School
The University of Nottingham Malaysia Campus.
• Introduction
• Methodology
• Data and Results
• Conclusions
Introduction
• The German automobile industry
  • one of the most developed production sectors
  • also generates additional turnover on the used car market
  • above 53 billion Euro in 2008 raising to more than 63 billion Euro in 2012
• The German used car market is organized in three business segments:
  • new car dealers
  • used car dealers
  • private offers
The used car landscape significantly changed since the introduction of the new warranty law (Gewährleistungsgesetz) in 2002:

• Privately sold car sales go down from 53% in 2002 to 39% in 2015) and gives way to professional dealers
• Professional dealerships acquire used cars:
  • rental and leasing disposals
  • trade-ins for a new car
  • cars directly bought from private owners

• Challenges:
  • Inflexibility of prices
  • Oversaturated markets

An optimization query arises:
How do we maximize profits by carefully balancing the asking price for a car with the number of days it spends on the market before being resold?
Figure 1.1: Revenue management tool for the used car sector (Jerenz, 2008, p.28).
Methodology
Survival Analysis (SA):
• "survival analysis" came into being from initial studies, where the event of interest was death
• after World War II, there was new area of SA emerged originating from reliability of military equipment
• Nowadays, the framework is constantly conquering new fields of application, especially in the business domain
• is modeling price response functions for price optimization on markets with restricted goods, such as the used car market
Common terms in Survival Analysis (SA):

• Event: a car being sold at a specific time point

• Time-to-event/Time-on-market (days): the time between the first day a car entry was uploaded online and:
  a) an event (e.g. car being sold), or
  b) end of the study.
  c) loss of contact or withdrawal from the study
Common terms in Survival Analysis (SA):

Right censoring:
Subjects for which we do not observe the full time-to-event are described as right censored

Left truncation/interval censoring:
Not fully observed from beginning to end during the study where only the status of the event is known
**Survival Function:**
The probability a car is still not sold after a certain period online.

\[ S(t) = P(T > t) \]

For car entries which are not censored:

\[ P(T = T_i | \theta) = f(T_i | \theta) \]

where \( T_i \) refers to the measured time a car spent on the market. The real time for which the car was sold (T) coincides with \( T_i \) and \( \theta \) refers to a set of covariates that influences the time-to-event.
Survival Function
Right censored data (probability of survival time being more than observed survival time):

\[ P(T > T_i | \theta) = 1 - F(T_i | \theta) = S(T_i | \theta) \]

Hazard Function:
the instantaneous probability of an event to occur, given that the event has not taken place until now.

\[ \lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t+dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} \]
Hazard Function
the probability of a car being sold at next instant:

\[ H(t) = \int_0^t h(u)du = -\ln S(t) \]
Standard Statistical Approaches in Survival Analysis (SA):

• Kaplan-Maier Estimator
• Nelson-Aalen Estimator
Kaplan-Maier Estimator

- non-parametric model for estimating the survival function of time duration data

- provides a step-wise non-increasing function that is based on the difference between total number of survivors $n_i$ and total number of censored observations prior to time point $t_i$

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$
Cox Proportional Hazard Model

• ability to estimate the effect of multiple covariates $X_i$ on survival (via the hazard function $h(t|X_i)$), as well as its non-parametric estimation of the baseline hazard functions $h(0)$, which grants more flexibility and thus better fit to the data in comparison to fully parametric models
Data and Results
Data

- Data was originally assembled in the context of a cooperation between a large German automobile manufacturer and mobile.de

- A database of cars from a mentioned brand, offered online for the period between 18 September 2008 and 18 December 2012.

- This includes a thorough record of daily prices, as well as a number of attributes (e.g. car technical specifications or vendor characteristics).

- The data contains 5,915,774 unique car IDs and 747,102 unique vendor IDs. The prices tables contain 190,323,612 price observations in total and the cars tables - 8,323,386 distinct car attribute descriptions.
Data Preparation

- Analyze Prices, Cars and Vendors datasets
- Merge Data set
- Aggregate Unnecessary Observations
Assessing the effects of Covariate on TOM

- Kaplan Meier Curves
- Cox Proportional Hazards
Kaplan Meier Curves

- useful for obtaining an initial overview of the trends in survival data
- Challenge for complex datasets with a number of continuous variables, since it requires creating a factor level for each continuous value recorded
Kaplan Meier Curves

- useful for obtaining an initial overview of the trends in survival data
- Challenge for complex datasets with a number of continuous variables, since it requires creating a factor level for each continuous value recorded
- Degree of Over Pricing (DOP), Market Size (MS), Quantile of Car Price, Age, Dealership size grouped into quantiles
Kaplan Meier Curves

(a) Survival functions for DOP in quartiles.

(b) Survival functions for MS in quartiles.

(c) Survival functions for Q in quartiles.

(d) Survival functions for age in quartiles.
Kaplan Meier Curves

DOP and Quantile
• cars priced relatively low in comparison to the market have the lowest chance of survival
• the probability for survival increases with each ascending quartile.

Age
• the youngest cars (first quartile = 14 months or less) do demonstrate a tendency for faster sale.
• cars in the higher quartile ranges (28 months or more) are more likely to be sold than cars with age in the second quartile
Kaplan Meier Curves

Market Size

• cars with small competition on the market (first and second quartile) have lowest survival probabilities

• The survival curves for market size in the third and fourth quartile have interchanged

• the difference between a large market size due to general high production for a car type, which is related to its popularity on the market (for values in the fourth quartile) and a mere large supply with lower customer base (third quartile).
Kaplan Meier Curves

Vendor Size
• reveals that the largest dealerships (with more than 2500 cars on the market for the examined period) manage to sell their cars fastest on average.

• this might be explained by a tacit know-how and better marketing for large dealerships

• the lowest survival curve crosses both the curves of the second and third quartile.

• the presence of a number of dealerships with less than 10 cars on the market (private offers).
• The observed intersection of survival curves and non-monotonic arrangement of survival probability along quartiles, hint at possible non-linear relations between variables and a lack of proportionality in hazards for different data groups.

• might be due to the univariate examination of the effects of covariates on survival.

• Cox Proportional Hazards framework
### Cox Proportional Hazards Model

**Comparing buyers’ preferences of a major Mercedes Benz classes**
Cox Proportional Hazards Model

Market Size
• a negative influence on the car’s probability of being sold at next instant.

DOP
• which decreases the probability of a car being sold at next instant by a value between 50% to above 70% for one unit increase in DOP, depending on the car class
Cox Proportional Hazards Model

Quantile
• increase in the variable decreases the car’s probability of being sold at next instant by somewhere between 10% and 40% depending on the car class

Age
• an additional month in age, decreases the chance of a car purchase by 0.5%

Dealership
• An increase of 10 cars in the vendor portfolio, increases the chance of a car being sold by up to 1%. 
Cox Proportional Hazards Model

- However, linearity and proportional of hazards were a concern with this dataset.
- Data-driven techniques were looked into:
  - Decision trees and Random Forest
Data and Results

Decision Trees and Random Forest

Advantages

• potentially reveals interactions between covariates (their non-parametric nature)

• binary splitting rules facilitates easy identification of vehicle profiles with similar survival times
Decision Trees and Random Forest

Advantages

• potentially reveals interactions between covariates (their non-parametric nature)

• binary splitting rules facilitates easy identification of vehicle profiles with similar survival times
Decision Trees and Random Forest
Martingale (linearity) and Schoenfeld residuals (proportionality) for single covariates
Decision Tree

- 13500 observations
Random Forest

- Tune parameters
- Apply model to training data
Random Forest

Advantages

• potentially reveals interactions between covariates (their non-parametric nature)

• binary splitting rules facilitates easy identification of vehicle profiles with similar survival times
Random Forest

(a) Tuning tree parameters - check predicted error stability across changing number of trees and variables at each split. Number of observations - 20,000.

(b) Variable importance for the five covariate mean values from a sample of 100 models. Red points indicate maximal and minimal values.
## Random Forest Covariates Analysis

### Brier Scores

<table>
<thead>
<tr>
<th></th>
<th>IBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>0.090</td>
</tr>
<tr>
<td>w/o vendor size</td>
<td>0.096</td>
</tr>
<tr>
<td>w/o age</td>
<td>0.096</td>
</tr>
<tr>
<td>w/o Quantile</td>
<td>0.096</td>
</tr>
<tr>
<td>w/o DOP</td>
<td>0.094</td>
</tr>
<tr>
<td>w/o MS</td>
<td>0.094</td>
</tr>
</tbody>
</table>
Random Forest

• Predictions are made on a random sample of 1,500 observations from a validation set.
• Results: vendor size, age and Quantile as the variables which lead to highest improvement of explanatory power.
• DOP and MS
Conclusion

Statistics Approach
• DOP and Quantile

Data Driven
• Dealership and Age
Thank you
Terima Kasih
Xie Xie