HUMBOLDT-UNIVERSITÄT ZU BERLIN

MASTER'S THESIS

Is This a Good Time to Sell? Liquidity Estimation of Residential Properties Using Survival Analysis with Non-Proportional Hazards

Author:
Batuhan IPEKCI

Supervisors:
Prof. Dr. Stefan LESSMANN
Dr. Benjamin FABIAN

A thesis submitted in fulfillment of the requirements
for the degree of
Master's Program in Economics and Management Science (MEMS)

in the

School of Business and Economics

March 8, 2020

Declaration of Authorship

I, Batuhan IPEKCI, declare that this thesis titled, “Is This a Good Time to Sell? Liquidity Estimation of Residential Properties Using Survival Analysis with Non-Proportional Hazards” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:

Date:

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Abstract

Economics and Management Science

Master of Science

Is This a Good Time to Sell? Liquidity Estimation of Residential Properties Using Survival Analysis with Non-Proportional Hazards

by Batuhan IPEKCI

This thesis aims to pave the way towards a decision support system for setting the asking price of real estate properties based on the relationship between liquidity and price. It is shown that survival analysis with non-proportional hazards allows for tracking survival curves of individual properties across time with alternative pricing scenarios. Hence dynamic pricing strategies can be developed by determining whether to overprice, until when to overprice, and the liquidity cost of overpricing. A deep neural network is optimized on the architecture of DeepHit (Lee et al., 2018) to learn the distribution of survival times, allowing for the effects of variables on liquidity to change over time. The model is then interpreted by SHAP values following the calculations of Strumbelj and Kononenko (2014). Spatially aggregated contributions of individual input features enrich the decision making by identifying the places that are most sensitive to price reductions. The German residential real estate market is presented as a use case through hedonic, temporal, spatial, and population variables, together with the degree of overpricing (DOP). DOP is calculated by the estimation of a random forest model (Breiman, 2001a) as opposed to the ordinary least squares (OLS) regression used by Anglin, Rutherford, and Springer (2003). DeepHit is demonstrated to achieve a better calibration and discrimination performance than the state-of-the-art Cox proportional hazards (Cox PH) (Cox, 1972), and random forest is demonstrated to have a lower mean squared error than OLS regression.

The code containing the calculations can be found at: https://github.com/Batuhanipekci/ITGTS

Acknowledgements

I would first like to thank my thesis advisor Prof. Dr. Stefan Lessmann for all his help and guidance over the past two and half years. He encouraged me to find my own solutions at the times I felt stuck during the thesis process. His contribution to the thesis is invaluable.

The thesis is inherently speculative and it is difficult to find a real life implementation for that. Nonetheless, I had the opportunity to discuss and present my ideas multiple times to different audiences. The criticism that are addressed to my ideas have greatly helped me to mature them. I am indebted to the Valuation and Data Science / Business Intelligence teams at McMakler; especially Kevin Konings, Dr. Andreas Baudisch, Dr. Duo Zeng, Dr. Michael Kieweg, Dr. Yuval Winkler, Fitz Li, Valeriy Arsentyev, and Marcel Huth.

Thanks also Begüm Pekalp, for her love and support in times of difficulty.

Contents

List of Figures

List of Tables

List of Abbreviations

AFT	Accelerated Failure Time
Cox PH	Cox Proportional Hazards
DOP	Degree of Overpricing
MSE	Mean Squared Error
OLS	Ordinary Least Squares
SHAP	Shapley Additive Explanations
TOM	Time on Market

Chapter 1

Introduction

1.1 Problem Statement

There is a joint determination between liquidity and transaction price in residential real estate markets. Although real estate properties are generally characterized as illiquid assets, real estate liquidity varies over time dramatically. Hence liquidity should be considered as a determinant of the shocks to the fundamental value of housing (Krainer, 2001). This determination is partly due to the seller's trade-off between too much time for selling versus too low of a price (Anglin, Rutherford, and Springer, 2003).

This thesis focuses on an automated decision support system which may improve the seller's bargaining power by informing her about the changing state of the market so that she can capture a "liquidity-increasing" moment to overprice her house. The system is made possible by a price estimation through random forest (Breiman, 2001a) and a liquidity estimation through DeepHit (Lee et al., 2018). A simple post-processing on the estimated survival functions at different overpricing levels of a property defines the strategy of whether to evaluate a property over its estimated market price and how long to overprice before making a discount to its actual valuation. Besides, the cost of the overpricing in terms of liquidity loss can be computed. Consequently, decisions can be made by defining heuristic thresholds to the market state of the property, the liquidity cost, and the difference between the week of highest opportunity to overprice and the expected week of sale. The decisions are, then refined by examining regional sensitivity to reduction in the degree of overpricing (DOP, henceforth), using spatially aggregated contributions of individual input features.

The proof-of-concept is done, at first, by demonstrating the relevance of the liquidity estimation in the economic theory of real estate markets. Then the empirical setting of the German housing market, together with the assumptions that are made beforehand, is discussed. The performance of the machine learning models are, then compared against their classical statistical counterparts. It is observed that random forest performs better than the multivariate linear regression in terms of mean squared error. Furthermore, DeepHit performs better than the state-of-the-art Cox proportional hazards (Cox PH in short) (Cox, 1972) in terms of calibration by the Brier score and the integrated Brier score (Gerds and Schumacher, 2006) and discrimination performance by the time dependent concordance index (Antolini, Boracchi, and Biganzoli, 2005). Furthermore, a version of Shapley values are used to interpret the "black box" DeepHit model using the python package SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) which calculates approximate Shapley values (Shapley, 1953) with local regression approach (KernelSHAP) resulting in local explanations that are consistent with global interpretations (Molnar, 2019). Explanatory analysis of SHAP values, even though they are calculated approximately, allows for a sanity check to the DeepHit model. Also, they serve to build a temporal element in the decision support system.

Survival analysis is originated in health statistics, where the researcher is concerned with determining the effects of variables on the survival probability of patients encountering a fatal disease. Recently, the application field of survival analysis has been broadly extended to many disciplines like social sciences and finance (See Collett, 2014; Dirick, Claeskens, and Baesens, 2017; Box-Steffensmeier and Jones, 2004). Cox Proportional Hazards (Cox PH) or Accelerated Failure Time (AFT) (Wei, 1992) are traditionally the most popular models because of their interpretability. Nevertheless, AFT assumes a fully parametric statistical model on the error term and linearity on the covariates, whereas Cox PH assumes linearity and proportionality, although being semi-parametric. With the advance of enhanced data-driven modeling, various adaptations of deep learning models are introduced to the setting of survival analysis (See Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2019a; Fornili et al., 2014; Katzman et al., 2018; Fotso, 2018). DeepHit is a promising model among them, as it allows to specify a weight on loss function depending on whether the discrimination or the calibration performance is aimed to optimize. Although DeepHit is originally designed for the setting of "competing risks", it performs quite well in the single risk setting of the sale of a real estate property.

Survival analysis can be adopted to estimate a proxy to the demand, a time measure of liquidity in the housing market. Throughout the paper, the definition of liquidity by Wood and Wood (1985) as the inverse of time on market (TOM) is considered. The author disregards alternative measures derived from transaction cost, price, or volume.

The purpose of this thesis is to derive tools for the strategic action of the seller. Therefore, it is crucial to estimate the effect of the seller's individual decision making separately from the effect of the current market state on the liquidity. These two effects are examined separately by including both the expected price of the property and DOP to the model. The definition of DOP is adopted from Anglin, Rutherford, and Springer (2003) as the ratio of the listing price to the expected price of the property. By doing so, the seller's trade-off between selling fast versus selling cheap is analyzed.

1.2 Research Questions

The following research questions are addressed to contribute to the literature:

1. Is it a feasible strategy to estimate the liquidity in German residential real estate market by state-of-the-art Cox PH modeling? Do the assumptions of Cox PH even hold?

2. Is there a better way of estimating the liquidity than Cox PH, while not making any explicit assumptions? Does DeepHit perform better than the Cox PH in both calibration and discrimination performances?

3. Is there also a way to interpret the results of the DeepHit model with SHAP values, so that we can be more sure about the sanity of the model? How important are spatial, hedonic, and population characteristics in determining liquidity? What additional information can be gained by using DeepHit instead of Cox PH?

4. Could DOP be better estimated by random forest than OLS?

5. How could the output of DeepHit be utilised for strategic overpricing? Is it possible to speculate about the market state of a specific property and to derive heuristics about whether to overprice, until when to overprice, and the cost of strategic overpricing in terms of liquidity loss? Could the SHAP values of the model yield an additional support to the decision making?

1.3 Contributions

The paper contributes to the literature in at least five ways. At first, random forest is suggested as a better way of engineering the variable DOP in comparison to OLS estimation. The results of the random forest regression are particularly important as we are interested in the direct effect of strategic overpricing, separated from the bias arising from modeling, as much as possible. Consequently, the results of a well-tuned random forest model allowed more data to use in implementing the survival analysis estimation, since the data is restricted to contain overpricing degrees only between 0.8 and 1.2, in order to keep results as realistic as possible.

Previous studies on the determinants of the real estate liquidity were generally restricted to the use of Cox PH. The second contribution of the paper states that the assumptions of linearity and proportionality inherent in Cox PH do not reflect the reality of the German residential real estate market. The explanatory residual analysis in Chapter 5 demonstrates that there is a non-linear relationship between the covariates and the target, and the hazards associated with the covariates are non-proportional over time.

To the author's knowledge, DeepHit is applied for the first time in the context of liquidity estimation in residential real estate markets. The paper's third contribution is demonstrating that DeepHit is indeed a feasible model for this estimation. Although the assumptions of Cox PH on linearity and proportionality are usually violated for many real-life situations, surpassing its performance is still a challenging task due to its semi-parametric design. The enhanced data-driven modeling by DeepHit results in better discrimination and calibration scores than state-of-the-art Cox PH model. During the estimation, both the expected price of a property and DOP are included in the covariates. By doing that it is attempted to analyze the effect of strategic decisions in different scenarios for a property whose estimated price is also known.

The fourth contribution is the extensive use of SHAP values as a modern technique not only to interpret machine learning models, but also to be integrated in a decision support system. SHAP values are mainly used to reveal the spatial patterns in the effects on liquidity. Using SHAP values, it is possible to study geographic submarkets, and observe how the opportunities of selling houses develop over districts or neighborhoods.

An accurate estimation of the liquidity brings about important informational advantages to market participants during the bargaining process. The fifth contribution of the paper is a post-processing heuristic for capturing the opportunities for the seller side of the market. As a direct result of defining the liquidity estimation in the domain of survival analysis, the sellers have information about the expected TOM of the properties bounded by the chosen lower and upper bounds on survival probabilities. The additional benefit of using DeepHit is its ability to capture non-proportional hazards during the estimation of survival curves. This enriches the seller's knowledge to develop overpricing strategies for any individual real estate property.

The proposed decision support system to overprice real estate properties consists of temporal and spatial components. Estimations from DeepHit give rise to the temporal component, where a time series of hazards can be extrapolated for any individual property. In Chapter 6, the focus will be on how to calculate market states, cost of overpricing in terms of liquidity loss, and until when to overprice. Automated decision making is made possible by defining rules for these three instruments. Finally, the spatial component which is aggregated by SHAP values refines the strategies by ordering the locations where decreasing the level of DOP would yield a fast sale.

Chapter 2

Theoretical Background and Related Literature

2.1 The Bargaining-and-Search Game

Real estate markets are characterized by infrequent transactions, heterogeneous goods, and imperfect information. The matching process in the market is outlined by a bargaining-and-search game: The seller seeks potential buyers by setting a listing price and advertising the property through intermediaries. The buyer searches for the utility-maximizing good, having specific valuations for each particular features of a property. The buyer's valuation is only observed after the seller sets a listing price, hence it is a random variable. Price setting initiates a bargaining game where the buyer determines her valuation for the property in question, while also looking around for outside opportunities. The bargaining game is then complemented by a search game where the asking price affects the rate at which potential buyers are contracted. The equilibrium of the game depends on the buyer's and seller's discount rates, the buyer's outside opportunity, and the value of search to the seller (Arnold, 1999). The transaction is realized after the seller optimizes a trade-off between selling the property for as high a price as possible, or as quickly as possible. The interval for the transaction price is bounded above by the listing price, and below by the seller's reservation price (Yavas and Yang, 1995).

The seller's bargaining power is restricted by Akerlof's adverse selection problem, when the buyer suspects the quality of the property which remains on the market for a long time, even though there may be no other reason to suspect (Akerlof, 1970). The seller's pricing strategy does not only determine the rate at which potential buyers are attracted, but also the buyer's valuation process. A property that is not selected during the buyer's search process is subject to price revision. A property with ever declining prices and ever prolonged time-on-market (TOM) spirals into negative herding, as it becomes „stigmatized“ (Taylor, 1999). Price revisions following initial price setting are determined by a prolonged TOM and the degree by which the property is overpriced initially (Knight, 2002).

2.2 Temporal and Spatial Market Dynamics

The liquidity of real estate markets puzzled economists for many years at least since Cubbin (1974) who pointed out that "the housing market behaves in somewhat unexpected manner" as the higher the price the quicker the house was sold. There exists a complicated relationship between transaction prices and TOM. If we consider the duration of TOM as the value of houses perceived by the buyers in the market (as fast-selling houses are more valuable and slow-selling houses are less valuable), it is difficult to conform the housing market to the efficient market hypothesis in economics:

> " [...] most economic theory predicts that fluctuations in fundamentals should be immediately reflected in prices. That is, if the value of a house changes by a certain amount, the price of the house should change by the same amount. The real estate market does not appear to work this way. Rather, when house values decline, sellers are slow to drop their prices. Thus, marketing times increase and the volume of sales declines. These are all features of a cold real estate market. The hot market has just the opposite characteristics. Real estate prices are typically rising during hot markets. However, prices do not appear to rise fast enough, as suggested by the fact that houses are quickly snapped up after they are brought to market." (Krainer, 2001)

Hot and cold market states are usually attributed to the lagged responses of individual buyers and sellers to aggregate shocks in the market. The lag in responses has been attributed by Krainer (2001) to the asymmetric information where the seller knows the value of the home much better than the buyer. The status of credit markets are discussed to be another factor causing the market frictions (Berkovec and Goodman, 1996; Stein, 1993; Follain and Velz, 1995; Hort, 2000). Haurin (1988) demonstrates the importance of the housing characteristics in determining TOM, where atypical houses have a longer marketing time. The role of real estate brokers are studied by Yavas (1995), the commission of whom increases the transaction price but decreases the search intensity of the buyers.

The effect of market states on liquidity in the purchase market depends on the functionality of the rental market (Krainer, 1999). The German real estate market is characterized by low homeownership rate and widely dispersed homeownership. This peculiar stability is a result of a historical path-dependency bound by cultural and political processes such as a careful city planning (Kohl, 2016), an extensive social housing sector, an established legal protection on tenant rights, high transfer taxes for buyers, and lack of subsidies for homeowners (Voigtländer, 2009; Kaas et al., 2017). Therefore, estimating the effect of the legal circumstances between the tenant and the homeowner is an important task too, as it may severely affect the expected returns from the property.

2.3 Survival Analysis in Real Estate Market

Survival analysis is a popular method in analyzing the liquidity of real estate properties. Table 2.1. summarizes selected applications in the field with their relation to this study. Blank cells imply that the effect of the variable is not examined in the respective study. In all the reported studies, Cox PH is the main model in estimating liquidity because of its semi-parametric structure and its ability to incorporate censored observations. Estimating TOM directly by a usual regression approach would result in biased estimations. Not all of the properties which are listed are sold, and some properties are sold with price revisions through multiple advertisements.

Since the "unexpected" movements in the housing liquidity are generally attributed to the bargaining-and-search theory, Haurin (1988) elaborated on the idea of house atypicality as a factor determining the variance of the offers to a home that are likely to be received. From the coefficients of a hedonic price regression an atypicality measure is constructed, on the assumption that the more unusual is the house, the greater the house can be overpriced. A Cox PH is estimated on only the variables of the atypicality index, the broker effect and season dummies. As a result, atypical houses are found to be sold faster.

Anglin, Rutherford, and Springer (2003) have examined the bargaining-and-search game from the viewpoint of the seller. The seller has a trade-off between selling fast versus cheap. Her strategies under changing market conditions can be estimated by the degree of overpricing (DOP), defined as the ratio of the listing price over the estimated price of the property. It is found out that the houses with smaller DOP sells faster, although a lower list price is not necessarily related to a shorter TOM. A heteroskedasticity correction was included to the liquidity estimation and it is concluded that the houses in a niche segment having a low predicted variance of prices are suffer from overpricing more than others in terms of the liquidity loss. It is noted that TOM varies more with spatial location and market conditions than it does with hedonic characteristics.

The seller adjusts her strategy with changing market conditions, where she tries to capture „value-increasing“ or „liquidity-increasing“ moments (Anglin, Rutherford, and Springer, 2003). It can be so that the state of the market can theoretically allow for higher transaction prices than the listing prices during a housing boom period, as opposed to the general bargaining-and-search framework discussed above (Haurin et al., 2013).

The marketability of a specific property vary by location within the housing market (Smith, 2009). It is usually accepted that the inclusion of the spatial variables, either as the Cartesian coordinates or dummy variables, improves the explanatory power of the liquidity modelling of real estate properties. Cajias and Heller (2018) argue in favor of a persistently overrented state of market equilibrium determined by a multitude of geographic submarkets in Germany, notably the cities surrounding top 7 economically advanced urban regions.

Cajias and Freudenreich (2017) assert that the exceptional levels of rental demand in German houses cannot be explained only by the moderate appreciation in listing rents. Clustering of spatially aggregated regions inform investment strategies on the market states arising from the joint classification of price and liquidity. For instance, a high liquidity in a low price-and-rent cluster might indicate an opportunity to buy or rent, as it might signal for rising prices for the region, and in its neighbors by spillover effects.

Kluger and Miller (1990) are the among the first who suggested to study factors affecting liquidity with the use of Cox PH. Nevertheless, they warned future researchers that one must be careful in designing pricing strategies. During their study, Cox PH predicted reasonable hazard rates even for houses with arbitrarily high price levels out of the range of the training set. This effect can still be confirmed with using DeepHit to develop pricing strategies as it is the aim of this project. Therefore, both the experimental design and the optimization design was done carefully.

The output of the survival analysis modeling has been used to improve decision making processes. Haurin et al. (2010) tested the effects of atypicality of properties on overpricing and liquidity by means of simulating the search-and-bargaining game, where the simulation is validated by a Cox PH estimation.

Jerenz (2008), informed by the discussions around the liquidity and price of real estate properties, has framed the seller's problem in the second hand car market as a single-product dynamic pricing problem with a finite stock of items and stochastic demand. The optimization module takes the price-response functions as the main input. The price-response functions were derived from the functional form of the survival analysis modeling, where the variable DOP was a main

Author	**Spatial**	**Temporal**	**DOP**	**Population**	**Property Type**	**Market**
Haurin (1988)		✓				Sale
Kluger and Miller (1990)	✓	✓				Sale
Anglin, Rutherford, and Springer (2003)	✓	✓	✓			Sale
Smith (2009)	✓	✓	✓			Sale
Haurin et al. (2013)			✓			Sale
Cajias and Freudenreich (2017)	✓	✓	✓	✓	✓	Rental
Cajias and Heller (2018)	✓	✓	✓			Rental
Our Study	✓	✓	✓	✓	✓	**Sale**

TABLE 2.1: Survival Analysis in Real Estate Studies

predictor for demand. In other words, a price forecast informs demand estimation which in turn feeds an optimization engine.

There is a high potential of advanced data-driven methods to improve the effectiveness of applications which can be derived from survival analysis estimations. Lessmann and Voß (2017) examined an exhaustive list of alternative learning methods (linear, nonlinear, and ensemble models) on various datasets to overcome the inefficiencies in the traditional price estimation by multivariate linear regression. Random forest regression was found to be particularly effective. Further, a similar study has been done to increase the efficiency of the demand estimation step Lessmann et al. (2018) where random survival forests are shown to be performing significantly better than Cox PH in both discrimination and calibration scores.

This thesis examines pricing strategies for overpriced properties in changing market circumstances. Therefore, a careful price and liquidity estimation is carried out. Hedonic, spatial, temporal, and population variables together with the DOP and property type of residential properties for sale is analyzed. Unlike some studies, the actual sale prices were not available in our data set; hence we decided to use the expected prices as another variable in order to capture the effects of hypothetically changing DOP levels with the same expected price. The effects of covariates are interpreted as a model sanity check. A heuristic for decision makers is presented on whether to overprice, until when to overprice, and the liquidity cost of overpricing. The assumptions on proportionality and linearity of Cox PH severely restricts this application. Therefore, DeepHit, a deep learning model which do not assume proportionality and which could identify non-linear interactions, is deemed to be suitable. The price estimation is done by a random forest, which is shown to be superior to ordinary least squares (OLS) regression, in constructing DOP.

Chapter 3

Methodology

3.1 Survival Analysis

Survival analysis is used to study the time until the occurrence of some event. In analyzing the time on the market of real estate properties, the main interest lies in exploring the determinants of the probability of sale during the marketing period. A major advantage of survival analysis regression to usual multivariate regression models is that the response can be incompletely determined for some subjects. Whenever it is not possible to be sure about the subject's status of sale, it is counted as "right censored" after the time of its last appearance in the data set. Since survival analysis prioritizes the estimating the probability of surviving after a certain time than the expected survival time itself, it is still possible to operate on incomplete information. There are various types of censoring and truncating in the literature of survival analysis, the focus will be only on right censoring in this thesis, as the nature of the problem suggests so (See Harrell, 2006 for a complete understanding of regression techniques).

The survival function at time t $S(t)$, the probability that a subject survives at least until time t, is given by

$$S(t) = Prob\{T < t\} = 1 - F(t)$$

where $F(t)$ is the cumulative distribution function for the response variable T. $S(t)$ is always 1 at t=0, and must be non-increasing as t increases. The expected failure time is the area under the survival function for t ranging from 0 to infinity:

$$E(T) = \int_0^\infty S(v)dv$$

In practice, we are rather interested in instantaneous failure rate in a small interval around t, i.e. the hazard function at t, which is defined as

$$\lambda(t) = \lim_{u->0} \frac{Prob\{t < T \leq t + u|T > t\}}{u}$$

3.2 Kaplan Meier Estimator

Kaplan-Meier (the product-limit) estimator (Kaplan and Meier, 1958) is a nonparametric maximum likelihood estimator for survival function $S(t)$. Let k denote the number of failures in the sample and let $t_1, t_2, ..., t_k$ denote the unique event times. Let d_i denote the number of failures at t_i and n_i be the number of subjects at risk at time t_i. The estimator is then defined as follows:

$$S_{KM}(t) = \prod_{i:t_i \leq t} (1 - d_i/n_i)$$

If one is interested in computing Kaplan-Meier survival curves for multiple groups, one would estimate a different survival function for each group. The log-rank test (Goel, Khanna, and Kishore, 2010) can be used to test the equality of different survival functions for subgroups within the data. The failed log-rank test imply that survival curves cross each other (hence non-proportional hazards for the splitted variable exist).

Kaplan-Meier estimator can produce very flexible survival curves. However, it cannot incorporate the effect of multiple covariates. The estimations are restricted only for the groups of a single covariate.

3.3 Cox Proportional Hazards Model

Cox proportional hazards (Cox PH) (Cox, 1972) model is a semi-parametric model; it restricts the effects of the predictors on the hazard function by a parametric assumption, although it makes no assumption on the nature of the hazard function $\lambda(t)$ itself. Cox PH is more popular than fully parametric alternatives like accelerated failure time (AFT) models because it is useful in situations when the true hazard function is unknown.

Cox PH is usually defined by the multiplicative relationship between the linear combination of predictors $X\beta$ and the baseline hazard function $\lambda_0(t)$:

$$\lambda(t|X) = \lambda_0(t)exp(X\beta)$$

where $\lambda(t)$ is the baseline hazard which can be estimated by an arbitrary estimator (for instance, the Kaplan-Meier, although there is a variety of alternative baseline estimators). Note that $\lambda(t)$ is defined independent of the coefficient vector β, that means a valid estimate of β does not presuppose the estimation of $\lambda(t)$. As the coefficient vector β is also independent of time, it assures a constant individual hazard ratio over time.

The proportionality assumption follows allows the survival times between distinct individuals to be independent. The hazard ratio of $\frac{\lambda(t|X_1)}{\lambda(t|X_2)}$ does not depend on either on the baseline hazard, $\lambda_0(t)$, or time, t.

$$\frac{\lambda(t|X_1)}{\lambda(t|X_2)} = \frac{\lambda_0(t)exp(\beta^T X_1)}{\lambda_0(t)exp(\beta^T X_2)} = exp(\beta^T X_1 - \beta^T X_2)$$

A log-rank test based on Cox PH is developed to discover if the proportionality assumption holds, hence, the derivation above is valid (Harrell, 2006).

The assumptions of linearity and proportionality severely restrict the performance of the model mainly because they do not generally hold in reality. Different model selection procedures like stepwise regression are used to determine non-linear effects of variables. The nonlinearities are incorporated via basis functions. Choosing the functional form of nonlinearities and identifying the interaction variables rely on an extensive field knowledge. On the other hand, the adverse affects of the proportionality assumption on the predictive performance is usually addressed by stratifying continuous variables into different binary categories. Then an approximate effect corresponding to a discrete category is estimated.

Cox PH model is useful to isolate the effect of a single variable in the absence of the knowledge about the hazard function $\lambda(t)$. Nevertheless, it is problematic to assume a constant relationship between the variables and the baseline hazard across the time. This constant hazard rate inhibits the specific application discussed in this thesis, which relies on changing relationships between variables and hazard risks across time.

3.4 DeepHit

DeepHit (Lee et al., 2018) uses a deep neural network to learn the survival functions directly, without making any assumption on the underlying baseline hazard or its relationship with the variables unlike classical survival analysis models such as Cox PH and AFT. It is specifically developed to deal with the case of competing risks, i.e. in the context of subjects facing multiple events. In this thesis, the use of DeepHit is restricted to only a single event, as the setting of the problem (the sale of the property) suggests doing so.

The survival time is considered to be discrete and finite, denoted by $T = \{0, ..., T_{max}\}$ for a predefined maximum horizon T_{max}. There are $K \geq 1$ events. The situation of right-censoring is also regarded as a separate event, and denoted by $\varnothing$. So the set of events are defined as $\kappa = \{\varnothing, 1, ..., K\}$. In case of a single event, this set reduces to $\kappa = \{\varnothing, 1\}$. The output layer consists of 2 nodes in case of a single event. The input node for receives the data set in tuples (x, s, k) , where $x \in X$ is a D-dimensional vector of covariates, $s \in T$ is the time at which an event or censoring occurred, and $k \in \kappa$ is the event or censoring occurred at time s. The input variable s is named as the first hitting time, as it could suggest either the time of failure or the time of censoring.

The model attempts to estimate the conditional probability $P(s = s^*, k = k^* | x = x^*)$ for each tuple (x^*, s^*, k^*) with $k^* \neq \varnothing$. A deep neural network is trained to estimate $\hat{P}$. The architectural difference of the deep neural network from multi-task network variants is a single output layer with softmax activation, and a residual connection from the input covariates into the input of each cause-specific sub-network. The output layer ensures that the joint distribution of censoring and event is learned instantaneously. The residual connections capture the latent representation that is common to the event and censoring.

DeepHit optimizes the loss function L decomposed by $L = \alpha L_1 + (1 - \alpha) L_2$. L_1 is the negative log-likelihood of the joint distribution of the first hitting time and the corresponding event, and L_2 is a loss function based on the idea of concordance which incorporate the ranking of the subjects. As L_1 allows for learning of the general representation of the joint distribution of the first hitting time and the event, L_2 lead to an estimation of the event-specific cumulative incidence function at different times.

The implementation of DeepHit is done through the python library pycox (Kvamme, 2019). During the implementation, it was possible to tune the parameters alpha and sigma, the former of which controls the linear combination of L_1 and L_2, where alpha equals to 1 gives a loss only containing L_1. Sigma is a parameter in the ranking loss L_2.

3.5 Shapley Values and SHAP

The main concern with using machine learning models is that although they are usually powerful in predictions, it is difficult to quantify how they behave. Fortunately, any predictive model can be explained by calculating the Shapley values of its predictors, which are a proxy to their marginal contribution to the model output (Strumbelj and Kononenko, 2014). The setting is formulated in coalitional game theory, where each predictor obtains a payoff by getting included to the model. The Shapley value is the average marginal payoff of a feature value across all possible coalitions. Shapley values has a direct interpretation as individual conditional expectation curves (Molnar, 2019). Feature importance can be derived by the average of absolute value of Shapley values.

Local explanations which are derived from Shapley values are consistent with global interpretations like feature importance. This is due to the game theoretical foundations of Shapley values. They give a fair distribution of payoffs, satisfying the properties of efficiency, symmetry, dummy and linearity. Competing interpretation methods like LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, and Guestrin, 2016) do not automatically promise the consistency of local explanations and global interpretations. Lundberg and Lee (2017) combines LIME and Shapley values under the name of SHAP values by introducing a Shapley kernel to locally weighted regressions of LIME. Nevertheless, the exact computation of SHAP values is computationally expensive. Therefore, an approximate computation method is developed by Lundberg and Lee (2017) in the python package SHAP. Therefore, the original properties of Shapley values as discussed here are compromised. It is argued that the approximate Shapley values suffer from the bias occurring from highly correlated features, as other permutation-based feature importance measures do (Molnar, 2019). Therefore, one should be careful in interpreting the effects of correlated features.

3.6 Performance Measures

The performance of a survival analysis model is measured usually in two aspects; discrimination and calibration. The measures for the discriminatory performance quantify how the long-surviving and short-surviving subjects are ranked accurately. Harrell's Concordance Index (C-Index) is a popular measurement for the discrimination performance (Harrell et al., 1982). However, the C-Index does not deal with non-proportional hazards naturally. Therefore, a time-dependent C-Index is developed by Antolini, Boracchi, and Biganzoli (2005) as an alternative. C-Index is between 0.5 and 1.0. The higher C-Index is, the better is the discrimination performance. On the other hand, calibration performance is measured by how much an error occurs given a specific time period. The Brier score is computed by the squared residual between the observed status and expected status at a given time, weighted by the inverse probability of censoring weights. Computing the Brier score at each time period gives prediction error curves, and integrating the area under this curve is the integrated Brier score. The lower the Brier score is, the better the calibration performance is.

3.6.1 Time Dependent C-Index

C-Index is an extension of the area under operating characteristic curve (AUC) to the case of right-censoring data. During the calculation of Harrell's C-Index, any two time points are considered to be 'comparable' if the minimum of them is uncensored. The C-Index also postulates a notion of concordance which requires a subject who developed the event to have a less predicted probability of surviving beyond her survival time than any subject who survived longer. In the presence of censoring, the C-Index is the probability that two time points are concordant given they are comparable. Nevertheless, the usual C-Index specifies a 'one-to-one correspondence' between predicted times and predicted survival probabilities in any time point. As the models like Cox PH guarantees this correspondence, the classical C-Index can be computed with hazard rates instead of the whole survival curves.

Antolini, Boracchi, and Biganzoli, 2005 extended the C-Index to make use of the whole survival curve by rendering it time dependent. The functional form can be described as follows:

$$C^{td} = \frac{\sum_{k=0}^{K} AUC(t_{(k)}).w(t_{(k)})}{\sum_{k=0}^{K} w(t_{(k)})}$$

where the weight $w(t_{(k)})$ denotes the probability that any pair of time points are comparable at t(k)

3.6.2 Brier Score

The evaluation of the calibration performance is done by administrative Brier score proposed by Kvamme and Borgan (2019b). This version of the Brier score does not assume the conditional independence between the censoring and the event time given features, and in particular does not require estimation of the censoring distribution. Therefore, it deals with the bias of the earlier versions of the Brier score.

The formulation of the administrative Brier score is as follows:

$$BS(t) = \frac{1}{\hat{n}(t)} \sum_{i=1}^{n} [\mathbb{1}\{T_i^* > t\} - \pi_i(t)]^2 \mathbb{1}\{C_i^* \geq t\}$$

where T_i^* is the event time, C_i^* is the censoring time,$\pi_i(t)$ is the estimation at time t, and the term is scaled by $\hat{n}(t)$ which is the sum of the probability of censoring weights defined as

$$\hat{n}(t) = \sum_{i=1}^{n} \mathbb{1}\{C_i^* \geq t\}$$

Unlike the mean squared errors in usual regression models, the prediction error of the survival curves is not a number, but a process in time. A number summary of the Brier error curve can be derived by integrating the Brier error curve.

Chapter 4

Experimental Design

4.1 Data Source

The raw data set contains 624480 records in the period 1st January 1st, 2018 to May, 27th 2019, collected in September 2019. The data is web scraped from the largest platforms for residential real estate offers in Germany. Only the real estate properties which are listed for purchase and non-commercial use are selected.

DOP is estimated in a subset of 25 percent of the data set and predicted on the unseen 75 percent of the data set. Then the predicted data set is restricted only to contain DOP in the range between 0.8 and 1.2. The resulting cleaned data set contains 297131 records.

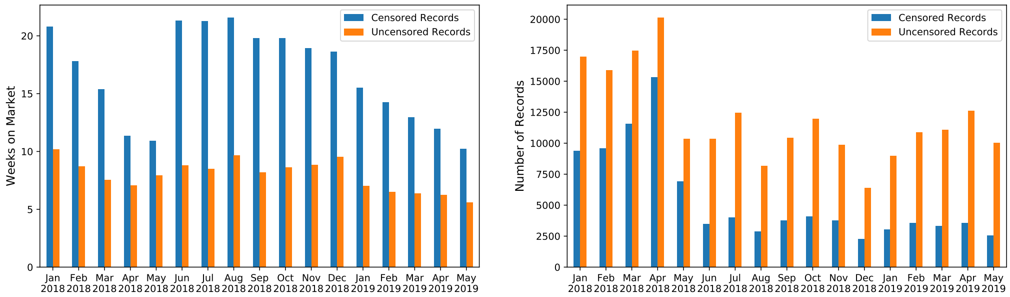


FIGURE 4.1: Distribution of Censored Records and their Weeks on Market

City	Count	%	Mean Weeks on Market	Average Price per m^2
Berlin	18014	0.061	11.73	4175
Munich	7766	0.026	8.83	7075
Hamburg	6442	0.021	9.76	4225
Cologne	3495	0.012	7.80	3487
Leipzig	3436	0.011	13.49	2245
Frankfurt	2863	0.010	8.93	4774
Dresden	2842	0.009	12.71	2256
Bremen	2765	0.009	9.11	2242
Stuttgart	2714	0.009	7.34	4307
Hannover	2507	0.008	8.48	2793

TABLE 4.1: Distribution of Records in Top 10 Cities

Variable	**Min**	**Max**	**Median**	**Mean**	**Std**
Location					
Latitude	47.473	54.922	50.825	51.028	1.696
Longitude	5.878	15.001	9.510	9.020	2.172
Distance to the City Centroid	0.000	4.537	0.080	0.032	0.251
Hedonic					
Construction Year	1850.000	2025.000	1983.539	1989.136	33.389
Living Area	20.020	996.000	125.478	115.000	71.1307
Number of Rooms	1.000	20.000	4.376	4.000	2.368
Balcony Available	0.000	1.000	0.600	1.000	0.489
Parking Available	0.000	1.000	0.600	1.000	0.489
Market					
Population Density (Zip Code)	14.282	26731.848	1894.522	750.427	2943.876
Market Size (Zip Code)	1.000	2641.000	427.555	363.000	259.044
Price per m^2	515.464	10000.000	2763.461	2454.550	1387.932
Expected Price per m^2	541.969	9787.515	2753.826	2463.637	1306.0942
Degree of Overpricing	0.800	1.200	0.997	0.996	0.105
Actively Rented	0.000	1.000	0.212	0.000	0.408
Property Type					
Bungalow	0.000	1.000	0.016	0.000	0.128
Farmhouse	0.000	1.000	0.001	0.000	0.041
House (Type Unknown)	0.000	1.000	0.084	0.000	0.277
Loft	0.000	1.000	0.001	0.000	0.034
Maisonette	0.000	1.000	0.023	0.000	0.152
Mansion	0.000	1.000	0.007	0.000	0.087
Multifamily House	0.000	1.000	0.048	0.000	0.215
One or Two Family House	0.000	1.000	0.243	0.000	0.430
Penthouse	0.000	1.000	0.011	0.000	0.103
Regular Apartment	0.000	1.000	0.326	0.000	0.469
Semi-Detached House	0.000	1.000	0.051	0.000	0.221
Special Building	0.000	1.000	0.023	0.000	0.152
Terraced House	0.000	1.000	0.045	0.000	0.206
Object Type (Ordinal)	0.000	13.000	2.340	2.000	2.594
Temporal					
Date of First Advertisement	0.000	1.000	0.421	0.356	0.301
SPRING	0.000	1.000	0.420	0.000	0.494
SUMMER	0.000	1.000	0.140	0.000	0.346
WINTER	0.000	1.000	0.292	0.000	0.4936
Target Variable					
Weeks on Market	1.000	90.000	10.477	6.000	12.165
Status	0.000	1.000	0.687	1.000	0.463

TABLE 4.2: Summary Statistics

The data set is spatially well-spread. The total number of observations in the 10 most represented cities corresponds to around 17.6 percent of the all observations. The three most populated cities in Germany; Berlin, Hamburg, and Munich account for the 10.8 percent of the whole data set and have a clear distance apart to the rest of the cities. An East/West divergence is apparent even in the data set, where the mean weeks on market of the cities Berlin, Dresden, and Leipzig is 12.64, far higher than that of the western cities Munich, Cologne, or Stuttgart.

4.2 Censored Observations

It is crucial for any survival analysis model to distinguish between the observations which are censored and those which are uncensored. Unfortunately, the knowledge of whether a delisted property is sold for sure was not available, although the properties which went offline could be tracked. Therefore, A bold assumption must be made: The listings having a high degree of pairwise text similarity in their descriptions are regarded as "censored" observations (Please refer to Appendix A for the detection procedure). A high degree of pairwise text similarity might either be caused by an updated price information, a correction of information, or a generic corporate advertisement which uses similar descriptions for multiple properties. Whenever repeated listings are detected in the property descriptions, the one which is published the first is selected and the remaining descriptions are disregarded. The uncensored observations are assumed to be sold at the time of delisting, while the date of sale for censored observations are assumed to be unknown. The censored observations are denoted by the status 0 and uncensored observations by the status 1.

The distribution of the censored records in months are demonstrated in Figure 4.1. The count of uncensored records is always higher than the count of censored records. The data set contains in general less records with incomplete information. Censored records have always higher mean weeks on market than uncensored records. This might suggest that censoring is caused due to relisting of the property which might be a response to an extended time on market. There seems also be a seasonality in liquidity, as the decreasing pattern of weeks on market in the period between January 2018 and May 2018, which has repeated in the period between Jan 2019 and May 2019, suggests.

4.3 Variables

Each record in the cleansed data set corresponds to an advertisement made for a real estate property with hedonic, temporal, and spatial variables together with their listing prices per square meter and their estimated degree of overpricing (DOP). Table 4.2 shows the summary statistics of the variables. The variable "Market Size (Zip Code)" represents the number of real estate properties in the zip code region, as a proxy to the supply of properties. The variable "Distance to the City Centroid" is the Euclidean distance between the centroid of the postal code region and the centroid of the city region. Since the locations of some properties are not exact and are approximated by the centroids of the zip code regions, sometimes the centroid of the postal code region of a property is the same as the centroid of its city region. therefore this variable is zero at the minimum. The variable "Population Density (Zip Code)" is calculated via dividing the population living in the postal code region by the area of the respective postal code region.

"Date of First Advertisement" is included as a numeric time trend. It is scaled between 0 and 1, where 0 represents the date January 1st, 2018, while 1 represents the date May 27th, 2019. Seasons are included to the models as dummy variables. Most properties in the data set are advertised in spring and winter.

A rich diversity of property types are analyzed. They are used as dummy variables in OLS regression, random forest and Cox PH. However, for the sake of reducing the computational training costs, this variable is encoded as a numeric for DeepHit, where integers represent the types of properties. Approximately 57 percent of the data set consists of either regular apartments or one-or-two family houses. The richness of property types enables estimating niche markets.

The dummy variable "Actively Rented" signals if the apartment is currently rented. Although, this variable is not fully informative about the legal status of the rental contract, it might yield an approximate estimate of the effect of a residing tenant. The extensive tenant rights in Germany might make the returns on investment lower than the market rates, as the new owner could sometimes not raise the rent of the property or cancel an existing contract.

Both "Number of Rooms" and "Living Area" are included as variables in the estimation, although they are highly correlated with a correlation coefficient of 0.84. The intention of using both of them is to capture the effect of "atypicality" as Haurin (1988) argues to be an important determinant of liquidity. The atypicality is not directly measured as suggested in (Haurin, 1988), but the nonlinear interaction of those two variables with the price might reveal an atypicality structure automatically during the application.

It can be said that approximately 68 percent of the data is uncensored, that is they are not repeatedly listed and their offline time could be tracked. The target variable, weeks on market, has a highly skewed distribution, where its mean is 6 while its median being 10, indicating the effect of censored observations.

It is a difficult task to infer on the price levels and liquidity by using linear models. One reason is the spatial segregation of the housing market as described above. It is also not feasible to use spatial dummies for the cities or postal code regions at the scale of this data set. Another reason can be thought as the zeroes in the "Distance to the City Centroid" which requires higher degrees of interaction between the latitude and longitude of the observations. The high correlation between the number of rooms and the living area is also problematic for a linear model.

4.4 Degree-of-Overpricing Estimation

The degree of overpricing (DOP) is the proportion of the listing price to the estimated price of the property. The price estimation of the property has been typically done by a log-linear regression as in Anglin, Rutherford, and Springer (2003).

$$\log(price) = \beta^T X + \epsilon$$

where $\epsilon \sim (0, \sigma)$

then

$$DOP = \frac{price}{E(price)}$$

However, the assumptions of linearity and Gaussian distributed errors made by the multivariate linear regression are not realistic. In addition, the data set, as discussed in the previous section, is not very suitable to be inferred by a linear model.

As an alternative to OLS estimation, DOP is estimated by random forest. Also, the expected price has been used together with DOP during the liquidity estimation. The reason behind is that the seller's strategies can be made visible when we control for changing levels of overpricing for a property whose valuation is already known.

The predictive performances of OLS regression and random forest are compared next. They are trained on the training set split as the 25 percent of the data to predict the other 75 percent. One third of the training data is used as a validation set for the hyper-parameter tuning of random forest. Their performances on the unseen test data are compared in Table 4.3. Modeling with random forest gives a considerable boost in the performances in both the training and test sets, although we can observe overfitting when we inspect the difference between the training and test errors of random forest.

Model	**MSE (Training Data)**	**MSE (Test Data)**
OLS Regression	0.216	0.216
Random Forest	0.008	0.063

TABLE 4.3: Out of Bag Mean Squared Error (MSE)

Chapter 5

Validation of the Liquidity Estimation

5.1 Kaplan-Meier Curves

The Kaplan-Meier estimate is the simplest way to visualize grouped differences in survival times given a single variable. An exploratory analysis is carried out by estimating separate survival curves for each group of a chosen variable and inspecting the visual distance between them across time. Kaplan-Meier estimations can be considered as the actual survival curves of property groups, on which it can be examined whether Cox PH is able to capture the patterns which are seen in the curves. The analysis of Kaplan-Meier curves also shows how estimating them by DeepHit might make a difference, as DeepHit is designed to capture crossing survival curves of a variable.

Figure 5.1 demonstrates the survival curves, which are estimated for the quartiles of all continuous variables and binary categories of all dummy variables. Continuous variables often have non-monotonic survival curves, suggesting nonlinear effects. The changing distance between curves across time suggests non-proportionality.

The highest survival time in the data set is 90 weeks. Hence the curves are drawn up to 90 weeks. Dummy variables such as Balcony Available, Actively Rented, and Parking Available show a clear separation in their survival curves, although the degree of the separation decreases with prolonged time on market. Properties having a balcony, a parking lot, and a residing tenant seem to sell faster in general. There is a changing order in season dummies as prolonged time on market, hence non-proportional hazards. For instance, properties which are listed in spring show a higher probability of selling up to the 50th week, but then they lose this clear advantage. This effect is somewhat expected because as TOM increases also the seasons change, and property enters a different market state, hence the liquidity effect of the changing season is observed. This effect is difficult to be captured by Cox PH, by construction. Survival curves of Latitude and Longitude suggest spatial patterns in a non-monotonic manner. Cox PH cannot explain the lack of monotonicity because it is fundamentally a linear model, hence yields estimations only in one direction.

The properties which have a recent construction year or which are an ongoing construction project are separated clearly with a higher survival curve from others. It is difficult to say anything for Property Type, as they are bundled into 4 categories arbitrarily, although the Q1 represents Apartments, One-or-Two-Family Houses, and Multi-Family Houses. Likewise, there is no discernible

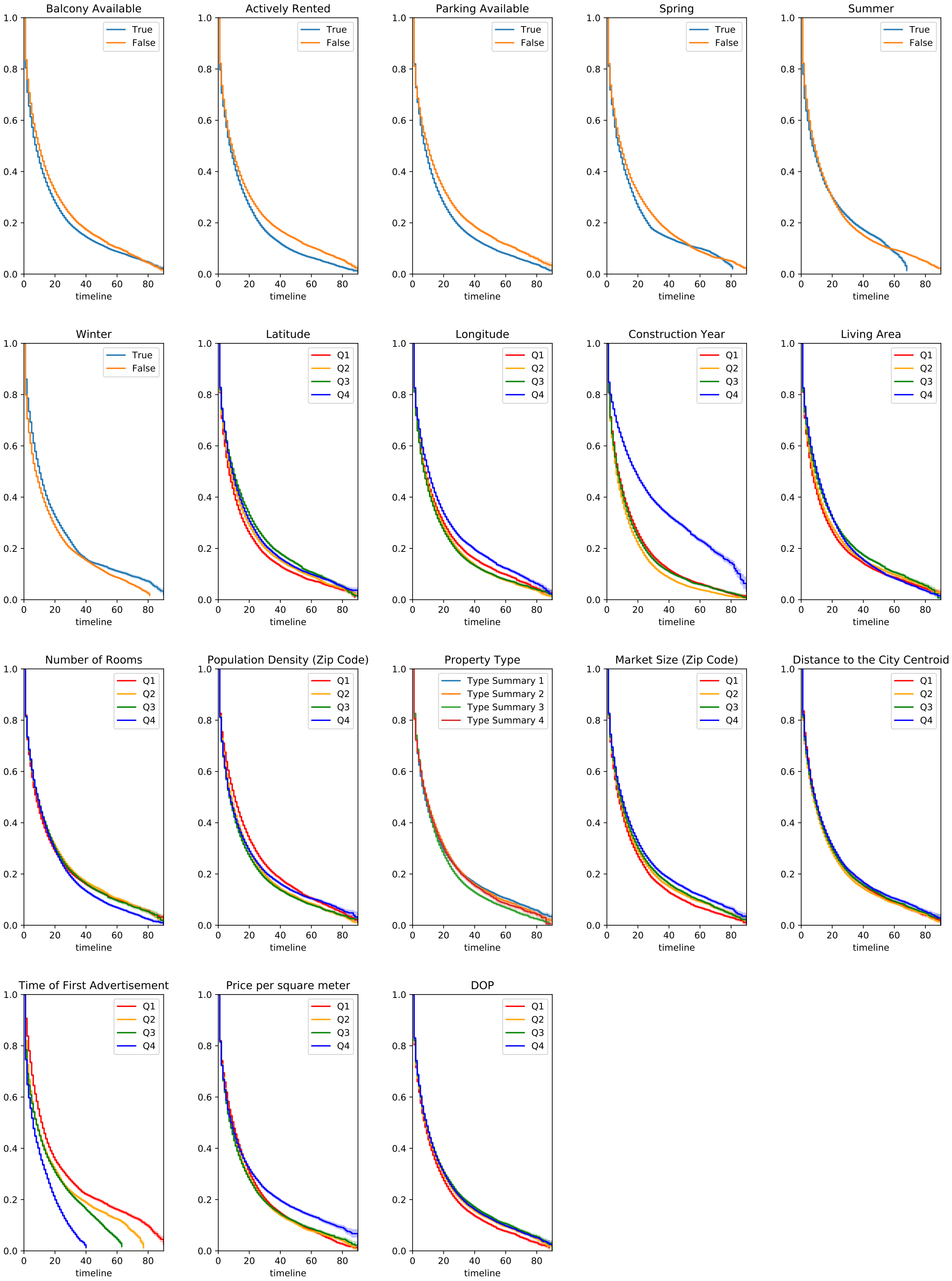


FIGURE 5.1: Kaplan-Meier Estimates for Survival Curves, all Variables

pattern for Distance to the City Centroid. Non-proportional hazards are visible in Number of Rooms and Living Area. If the property has a big Living Area (Q4), it is harder to sell among the properties up to week 40, thereafter it is easier to sell. A similar pattern applies to Number of Rooms, where properties with a high number of rooms (Q4) have higher liquidity than others after the 30th week. Population Density and Market Size shows a surprisingly contrasting pattern. Lower levels of liquidity is observed for properties in sparsely populated regions (Q1). However, regions with more properties per zip code (Market Size, Q4) seem to sell slower than those with less Market Size. This can be attributed to the bargaining-and-search framework, where buyers are more hesitant to buy a property if they have more options, as they tend to evaluate alternative options before buying a property (Arnold, 1999). The interaction of those two variables might give insights for theoretical discussions.

The trend variable Date of First Advertisement shows a monotonic separation of survival curves; the later a property is advertised, the faster it is sold. This can either suggest an increasing liquidity in the German real estate market, or a bias in data collection if the censored observations are not well separated from non-censored observations. Because Figure 4.1 in the previous chapter indicates a clear identification of censored and uncensored observations, it can be argued that the pattern in survival curves of Date of First Advertisement suggests an increasing liquidity level.

At last, non-proportional hazards are visible in Price per Square Meter and DOP, although very high Price per Square Meter (Q4) usually leads to low liquidity, and very low DOP (Q1) usually leads to high liquidity. These separations are in harmony with the theoretical background, and likely to be captured by Cox PH, at least partially.

5.2 Cox Proportional Hazards Assumption Checks

The coefficients of Cox PH estimation are demonstrated in Table 5.1. All of the estimated coefficients except Distance to the City Centroid, Actively Rented and some property types are significant at 0.01 level. Surprisingly, the effect of some variables like Expected Price per Square Meter, Distance to the City Centroid, Population Density, and Market Size are estimated to be near zero. Perhaps, in the absence of explicitly stated interactions, the effects of Population Density and those of Market Size have eliminated each other, as they had reverse directions in their Kaplan-Meier curves.

The direction of the effects of dummy variables corresponds to Kaplan-Meier curves of the respective variables. The ordinal time variable Date of First Advertisement has an impact with the highest magnitude. This is followed by DOP, season dummies, property types, hedonic, and location variables in decreasing order. Anglin, Rutherford, and Springer (2003) observed that hedonic variables are usually less important in explaining liquidity than temporal and spatial variables which signify market states. The estimation of this study shows that hedonic variables are by far less important than DOP which is supposed to demonstrate the seller's strategy. They are less important than property types which is supposed to estimate market segmentation. They are also less important than seasons, which signify market states. The only relationship that is not explained is that hedonic variables are more important than spatial variables. This can be accounted for the fact that the presence of the variable Distance to the City Center might have counteracted the effect of Longitude and Latitude.

The rest of the coefficients might only capture the reality partially, due to the non-linear effects that are visible in Kaplan-Meier curves. Non-linear effects of continuous variables are examined

Variable	Coefficient	exp(Coefficient)	P-value
Balcony Available	0.0519	1.0533	<0.005
Actively Rented	0.0127	1.0128	0.010
Parking Available	0.1424	1.1531	<0.005
Farmhouse	0.0159	1.0160	0.600
House (Unknown)	0.1282	1.1368	<0.005
Loft	-0.0325	0.9680	0.590
Maisonette	0.0039	1.0039	0.820
Mansion	0.1720	1.1877	<0.005
Multi-Family House	0.0790	1.0822	<0.005
One-or-Two Family House	0.0209	1.0211	0.090
Penthouse	-0.2277	0.7963	<0.005
Apartment	-0.0672	0.9349	<0.005
Semi-Detached House	0.1266	1.1349	<0.005
Special Building	-0.1797	0.8354	<0.005
Terraced House	0.2177	1.2432	<0.005
SPRING	0.1452	1.1563	<0.005
SUMMER	0.2330	1.2624	<0.005
WINTER	0.0580	1.0597	<0.005
Latitude	-0.0290	0.9713	<0.005
Longitude	-0.0186	0.9815	<0.005
Construction Year	-0.0019	0.9980	<0.005
Living Area	-0.0019	0.9980	<0.005
Number of Rooms	0.0198	1.0200	<0.005
Population Density (Zip Code)	0.0000	1.0000	<0.005
Market Size (Zip Code)	-0.0001	0.9999	<0.005
Distance to the City Centroid	0.0008	1.0008	0.900
Date of First Advertisement	0.8235	2.2784	<0.005
Expected Price per Square Meter	-0.0000	0.9999	<0.005
DOP	-0.3686	0.6916	<0.005

TABLE 5.1: Cox PH Coefficients

in Figure 5.2, where the smoothed martingale residuals from Cox PH model are plotted against single features to test for a non-zero slope. The martingale residual for each variable is computed by fitting a Cox PH model omitting that variable. The lines with a slope near zero indicate a linear relationship (Therneau, Grambsch, and R., 1990). Strong non-linear effects are seen in DOP, Latitude, Longitude, Construction Year, Market State, Date of First Advertisement. Other variables exhibit a pattern closer to linearity in their higher or lower values.

Kaplan-Meier curves in Figure 5.1 also shows signs of non-proportional hazards. The proportionality assumption is further examined by a χ_2 goodness of fit test, which reports the Pearson product-moment correlation (rho) between the scaled Schoenfeld residuals and the logarithm of time for each variable (Grambsch and Therneau, 1994). The last row indicates the global test for all the variables tested at once. A p-value less than 0.05 is usually regarded as a violation of the proportionality assumption, although even smaller p-values are suggested for large data sets (Lin, Lucas, and Shmueli, 2013). Table 5.2. shows that all variables with the exception of Distance to the City Centroid, Date of First Advertisement, and some property types have non-proportional survival curves, as the null hypothesis is rejected for them at 0.05 or even below p-values. The null hypothesis is also rejected globally.

The Cox PH assumptions do not seem to hold for the given data set. The smoothed martingale residuals and scaled Schoenfeld residuals suggest that non-linear and non-proportional modeling of the survival curves would be more appropriate for the real estate liquidity estimation problem.

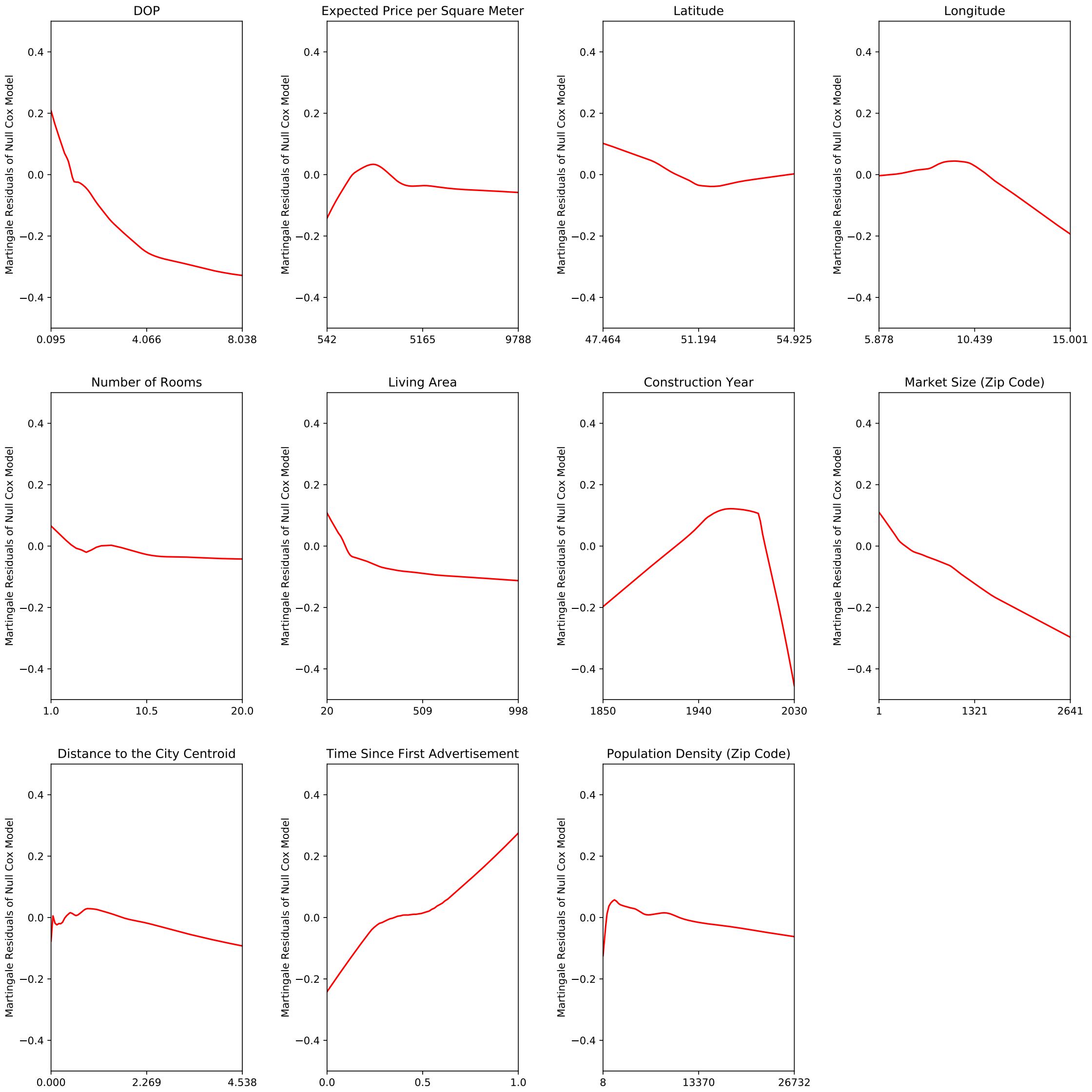


FIGURE 5.2: Martingale Residuals for Single Features

DeepHit is not restricted by these assumptions, although it requires a careful hyper-parameter tuning in order to surpass the performance of Cox PH in calibration and discrimination.

Variable	Rho	Chi-Squared	P-value
Balcony Available	-0.0066	14.41	<0.005
Actively Rented	0.0159	85.22	<0.005
Parking Available	0.0420	586.96	<0.005
Bungalow	-0.0063	12.98	<0.005
Farmhouse	-0.0056	10.35	0.001
House (Unknown)	0.0134	59.20	<0.005
Loft	-0.0059	11.51	<0.005
Maisonette	-0.0012	0.54	0.461
Mansion	-0.0044	6.47	0.011
Multi-Family House	-0.0115	43.19	<0.005
One-or-Two-Family House	-0.0160	83.36	<0.005
Penthouse	0.0005	0.11	0.735
Apartment	-0.0138	62.04	<0.005
Semi-Detached House	-0.0107	37.33	<0.005
Terraced House	-0.0101	33.54	<0.005
SPRING	0.0310	315.84	<0.005
SUMMER	-0.0122	49.14	<0.005
WINTER	0.0542	956.44	<0.005
Latitude	0.0124	51.07	<0.005
Longitude	0.0070	15.77	<0.005
Construction Year	-0.0375	406.32	<0.005
Living Area	0.0248	241.68	<0.005
Number of Rooms	-0.0129	60.23	<0.005
Population Density	-0.0195	121.73	<0.005
Market Size	-0.0085	24.20	<0.005
Distance to the City Centroid	-0.0002	0.02	0.8701
Date of First Advertisement	-0.0022	1.60	0.2048
Expected Price per Square Meter	-0.0179	102.22	<0.005
DOP	0.0301	321.73	<0.005
GLOBAL		**5493.11**	**<0.005**

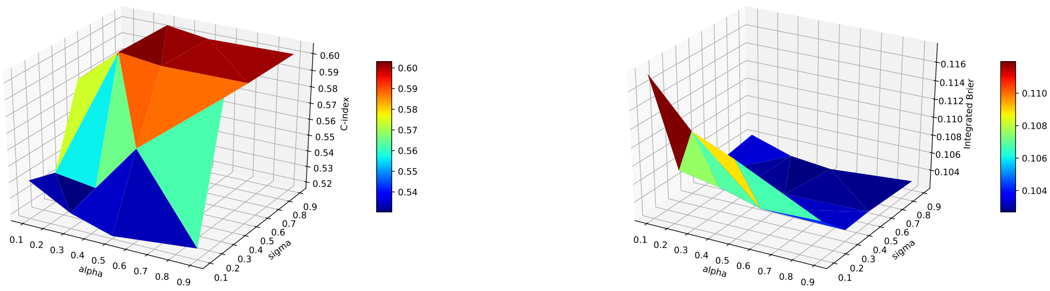
TABLE 5.2: Proportional Hazards Test

5.3 Hyper-Parameter Tuning for DeepHit

Deephit has many parameters to optimize; in addition to usual neural network hyperparameters, it is also needed to optimize the type of the loss function. There are two additional hyperparameters, alpha and sigma to this end. The parameter alpha, being the weight between L_1 and L_2 losses discussed in Chapter 3, controls whether the model is optimized towards the ranking ability (C-index) or the minimum likelihood (I-brier), while sigma is a parameter only in L_2.

At first, the usual neural network parameters are tuned through random search on a parameter set. The chosen network architecture is a DeepHit with 2 dense layers of size 128, a learning rate of 0.001, batch size of 128, a dropout rate of 0.1, and batch normalization. Parameter tuning has been validated by a separate validation set, along with early stopping. Early stopping mechanism was activated after 125 epochs.

After choosing appropriate neural network settings, the parameters alpha and sigma are tuned through grid search for values 0.1, 0.3, 0.5, and 0.9. Alpha and sigma are tuned with the abovementioned settings with a validation set and early stopping. As a result, alpha is chosen as 0.3 and sigma as 0.5.



(A) Time dependent C-Index

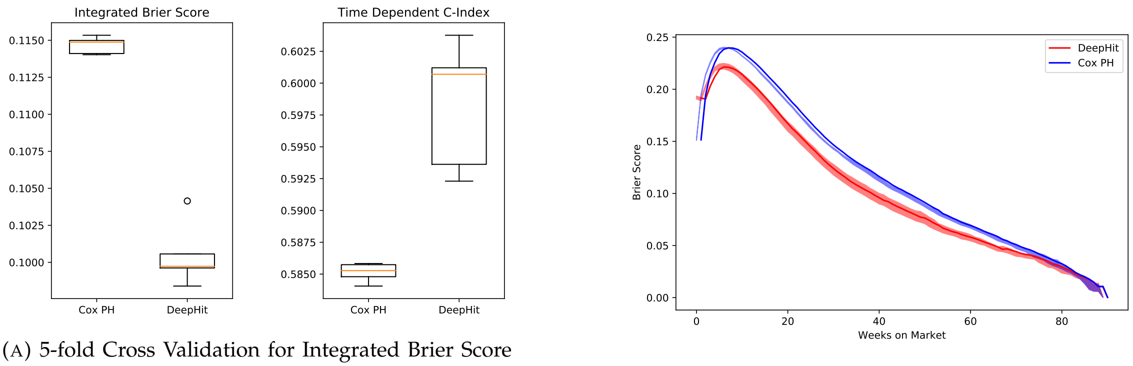
(B) Integrated Brier Score

FIGURE 5.3: Tuning DeepHit for Alpha and Sigma

Finally, the tuned DeepHit is compared to Cox PH in 5-fold cross validation. This time early stopping is not used and the number of epochs for Deephit was fixed to 125 epochs. It is done in order to ensure that both models use the exactly same data set for training.

A combination of discrete values of alpha and sigma are compared in Figure 5.3A (C-Index) and Figure 5.3B (I-Brier). The less the I-Brier score and the more C-index, the better is the performance. Small values (near 0) of alpha leads to an estimation that theoretically favors improving the C-index, and large values (near 1) favors improving the I-Brier. In Figure 5.3B a higher sigma generally results in good estimations without a trade-off between C-index and I-Brier. A too-small alpha causes high C-index while poor performance in I-Brier, although there is no visible trade-off in the region where alpha is greater than 0.3.

The 5-fold cross validated results are shown in Figure 5.4A. DeepHit has lower integrated Brier Score and higher time-dependent C-Index than the state-of-the-art Cox PH model. Nevertheless, the Brier error curve in Figure 5.4B reveals that DeepHit has a worse performance than Cox PH in predicting the properties which are sold in less than 6 weeks, which is the mean weeks on market of all properties in the data set. When a property is expected to sell near the average value, Cox PH might yield better results; however DeepHit is more capable of capturing properties which have an outlier duration on the market. A property which stays longer than 6 weeks is more likely to be captured by DeepHit, and the performance markup decreases in the limit cases where the property is on the market for more than 65 weeks.



(A) 5-fold Cross Validation for Integrated Brier Score and C-Index

(B) Brier Scores for DeepHit and CoxPH (Error curves)

FIGURE 5.4: Model Selection

5.4 Feature Importance and Partial Dependency Plots

The survival data set is split into training and test data sets in order to calculate SHAP values. DeepHit is trained on the training data set with the above-mentioned setting. The approximate SHAP values are calculated on the test data set. The default values which are needed for permutations are obtained by 30 clusters estimated in the training data set by k-means clustering algorithm. The payoffs for the SHAP are 1 minus mean survival probabilities within the 90-week period. Therefore, the effects of variables are indeed the effect on the inverse mean survival probability. This choice was made to study the effects on the average survival. Note that it is also possible to obtain the effects on a different payoff function, for instance the effect on the hazard at the first week.

A summary of partial dependence plots are drawn in Figure 5.5. The covariates are ranked according to their importance in explaining liquidity in decreasing order from above to below. There are fundamental differences in both variable importance and the direction of the effects between DeepHit and Cox PH. The Cox PH coefficients which are reported in Table 5.1 are used for the comparison.

The most important determinant of liquidity is the ordinal time variable Date of First Advertisement, which suggests a general increase in liquidity in the data set as time goes on. This was also visible in the Cox PH model. Distance to the City Centroid has the least importance, followed by Population Density. The weakness of these two spatial variables are in parallel with the estimated coefficients in the Cox PH model. The hedonic variables Living Area, Construction Year, Parking Available, and Number of Rooms affect the liquidity dramatically and their importance are in stark contrast with the Cox PH coefficients. Coefficients of hedonic variables were estimated to be the one of the smallest in magnitude among all variables in Cox PH. The specific information on market segmentation is also captured by DeepHit, as the ordinal Property Type variable is among the most important variables. The variables for season dummies, Cartesian coordinates, DOP, and Market Size are located in the middle in explaining liquidity. The coefficient of Market Size in the Cox PH was close to zero. DOP used to be the second most important variable in Cox PH. DeepHit estimates the Expected Price per Square Meter as a much more important determinant of liquidity than DOP.

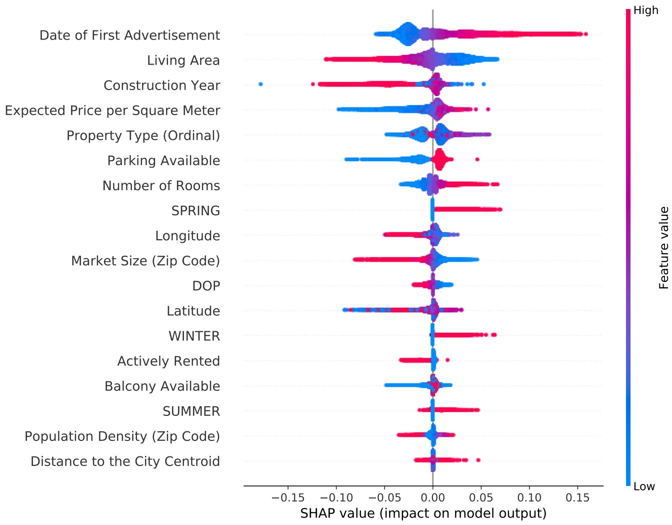


FIGURE 5.5: SHAP Values

In closer inspection, the graph in Figure 5.5 also reveals the direction of the partial dependence of the target to each variable. Red dots represent above-average values of the variable, while blue dots represent below-average values. The dots which are on the right side of the graph show an increasing effect on liquidity, whereas those on the left side a decreasing effect.

An increase in Date of First Advertisement induces an increase in liquidity when its value is above average, but the same pattern is not apparent when its value is below average. The importance of this variable should be attributed to its highest values, suggesting the market might have entered into a faster-selling market state in the period after roughly October 2018 in comparison to before. There is an opposite sign of the effects between Living Area and Number of Rooms; and this pattern was also visible in Cox PH model. A property with a high living area is sold slower, although a property with more rooms is sold faster. Location effects might have played a role in this contrast.

More interestingly, an increase in Market Size seems to affect liquidity adversely, meaning, if there are more properties on the market, it becomes slower to sell, other factors excluded. This effect was also visible in Kaplan-Meier curve of Market Size in Figure 5.1. It can be attributed to the bargaining-and-search framework.

The reverse direction of the effects between Expected Price per Square Meter and DOP was also in harmony with Kaplan-Meier curves. Figure 5.6 shows a clear partial dependence plot of DOP in interaction with Expected Price per Square Meter. Red dots represent data points where the expected price is above average, and blue dots represent below-average prices. DOP is located on the x-axis and its SHAP value (Liquidity) is located on the y-axis. For both expensive and cheap properties, it can be observed as they are overpriced, their liquidity decreases (tracking the red and blue lines respectively). However, underpricing a property which is expected to have an above-average price would increase its liquidity, while overpricing it would decrease it. Similarly, underpricing an already ’cheap’ property would not increase the liquidity, but overpricing some ’cheap’ properties might increase liquidity on average.

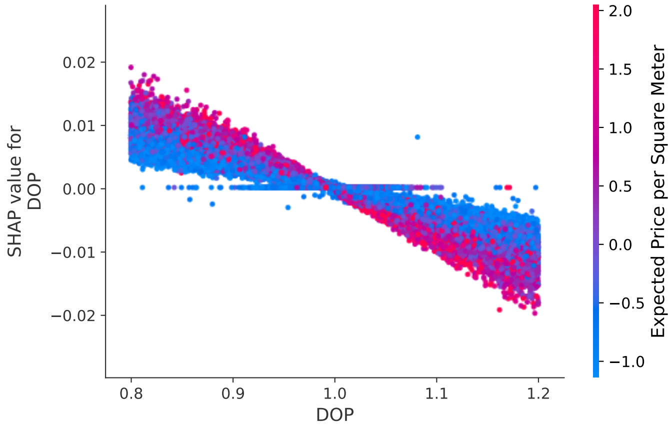


FIGURE 5.6: Partial Dependence Plot of DOP in interaction with Expected Price per Square Meter

DeepHit is able to estimate non-proportional and non-linear effects of variables in complex interactions and results in better integrated Brier Score and time-dependent C-Index than Cox PH. The calculated SHAP values reveal that DeepHit is able to capture the complexity of the real estate liquidity in a more detailed way than Cox PH. Most importantly, the role of DOP in this bargaining-and-search game was better stated in DeepHit model. DOP is not a fundamental factor determining the liquidity in DeepHit, as it has a mediocre importance rank as seen in Figure 5.7. DOP is rather a strategy that is chosen upon the fundamental variables that affect liquidity. As Figure 5.8 shows, overpricing properties might result in different liquidity results for properties with different valuations.

Chapter 6

Applications

The proposed decision-making design consists of temporal and spatial components. The temporal component defines opportunities to overprice across time given market states. Strategies for individual properties are developed by calculating whether to overprice, until when to overprice, and the liquidity cost of overpricing. The spatial component defines geographic submarkets. The strategies are, then, diversified by analyzing the locations where a price reduction would lead to a fast sale and where overpricing should be applied more carefully.

The temporal component is developed with the help of the non-proportional hazards which are estimated by DeepHit. A simple Cox PH model does not promise this application by construction, since the assumption of proportional hazards forces the effect of overpricing on liquidity to be constant over time. As a side note, this thesis is not concerned with making suggestions on "by how much to overprice". This question should be addressed by an appropriate price estimation process with reliable intervals. Therefore, it is avoided from giving a full price optimization model where a life-time value is maximized while liquidity counteracts the price, as opposed to the approach taken by Jerenz (2008) in suggesting a revenue management system for second-hand automobiles. The excuse is the observation that non-parametric estimation of liquidity does not give reliable estimates for extreme values of overpricing levels, as it was also discussed in Kluger and Miller (1990), who advised against designing pricing strategies by Cox PH.

The suggested strategies here are, instead, derived from a speculation depending on the distance between the hazard curves of a given property and its overpriced version at 10% level. The level of overpricing is chosen arbitrarily, but it can also be thought as an estimated upper bound of a price interval for a property. The distance between hazards that is changing across time will be referred as the "opportunity function" later on. The temporal component speculates on the upcoming time by suggesting in which time periods it is preferable to overprice a single property and until when to keep the price at the overpriced level. It is also possible to define intuitively if a property will be in a hot state or cold state in the upcoming weeks, in the sense argued by Krainer (2001). Knowing if the property is in a hot state and how long will it remain there would give the seller an informational advantage.

It is possible to make automated decisions by only using the temporal component. Nevertheless, applying the automated strategies everywhere would be risky because it can reinforce negative herding as discussed by Taylor (1999). A property whose price decreases after a long waiting time on market might be deemed as an inferior good. The spatial component is designed to restrict the application area of the automated overpricing system via ordering the locations by the sensitivity to DOP. Geographic submarkets are revealed by spatial aggregation of SHAP values.

The application in this thesis is restricted to utilizing only the SHAP values of DOP, in order to estimate regional sensitivity to DOP.

6.1 The Temporal Component: Pricing Strategies with Non-Proportional Hazards

Temporal strategies for selling individual properties can be developed by determining whether to overprice, until when to overprice, and the cost of overpricing in terms of the prolonged time on market. Hazard functions for a given property are estimated at chosen lower and upper bounds of DOP levels. The hazard function $\hat{h}_t(X)$ is simply computed by the contribution of the week t to the decrease in the estimated survival function $\hat{S}_t(X)$ at week $t-1$. A selling horizon for the property is defined, denoted by T. It is the first time when the survival function goes lower than 0.25. The median survival time is the considered as the expected week of sale, denoted by M. It is the first time when the survival function goes lower than 0.50.

At each time point, the difference between the hazard curve of the property which is overpriced at some level and that of the same property which is valued at its expected price is collected. This function is called the opportunity function for the given property, denoted by $\omega(t)$. The market state of a given property is hot if $\omega(t)$ is greater than 0 for at least one time before the selling horizon T. Otherwise, it is said that the property is in a cold state. The week of highest opportunity, denoted by H, is defined as the week t when $\omega(t)$ reaches its maximum. The time until overprice is defined as the maximum of the expected week of sale M, and the week of highest opportunity H. At any following time point, the seller should consider making a discount from the overpriced level to the estimated price level.

The hit times $\rho(t, \psi|t)$ arise for properties in the hot market state at the weeks when $\omega(t)$ is greater than ψ. ψ equals 0 for the properties in the hot state and equals $\frac{1}{T}\sum_t \omega(t)$ for properties in the cold market state. It is also suggested here a way to overprice a property in a cold market state, as it is the case when sellers try to "fish" buyers even in the periods marked by slow market transactions (Krainer, 2001).

The cost of overpricing c in terms of the risk of prolonged time on market is estimated by the ratio of the hit times $\rho(t, \psi|t > M)$ which fall in the time after M to all hit times $\rho(t, \psi|t \leq M) + \rho(t, \psi|t > M)$.

Depending on the risk appetite of the seller, thresholds for the cost of overpricing c for the hot and cold market states can be determined. Any property whose cost of overpricing is above the threshold is not overpriced. Similarly, additional restrictions can be placed on the distance between M and H; for instance a risk-averse seller might want to choose to overprice only the properties whose week of highest opportunity H is less than whose expected week of sale M.

The procedure is outlined as follows:

Step 1. A real estate property is chosen from the unseen test data set. A new data set with the chosen lower and upper bound of DOP values for the property is prepared, while keeping other covariates constant. The estimated survival functions of this data set is denoted as $\hat{S}_t(X_i)$ where X_i for $i \in \{L, U\}$ and L stands for the lower bound, U for the upper bound.

The hazard rate is the instantaneous risk of selling a property at time t, that is the proportion of the contribution of the period t to the previous period $t-1$ relative to the survival at $t-1$.

$$\hat{h}_t(X) = \frac{\hat{S}_{t-1}(X_i) - \hat{S}_t(X_i)}{\hat{S}_{t-1}(X_i)}$$

Step 2. A selling horizon T and the expected time on market M are defined for the chosen property with its DOP equal to L. In this study, T is chosen by the third quartile of the individual survival function, while M is the median of the survival function.

Set

$$T = min_t(\hat{S}_t(X_{DOP=L}) \leq 0.25)$$

$$M = min_t(\hat{S}_t(X_{DOP=L}) \leq 0.50)$$

Step 3. Compute the difference between the hazard functions $\hat{h}_t(X_{DOP=U})$ and $\hat{h}_t(X_{DOP=L})$ at each time point t from 1 to T. Call this function the opportunity function $\omega(t)$.

$$\omega(t) = \hat{h}_t(X_{DOP=U}) - \hat{h}_t(X_{DOP=L})$$

Step 4. Define the overall market state for the given property s as "hot" ($s = 1$) if the opportunity function is greater than zero for at least one time, and "cold" ($s = 0$) if it is always less or equal than zero.

$$s = \begin{cases} 1 & \sum_t \mathbb{1}(\omega(t) > 0) \geq 0 \\ 0 & else \end{cases}$$

Step 5. Given $w(t), M, T, s$, calculate the cost of overpricing in terms of the risk of prolonged time on market.

$$c = \begin{cases} \dfrac{\rho(t,0|t > M)}{\rho(t,0|t > M) + \rho(t,0|t \leq M)} & s = 1 \\ \dfrac{\rho(t, \frac{1}{T}\sum_t \omega(t)|t > M)}{\rho(t, \frac{1}{T}\sum_t \omega(t)|t > M) + \rho(t, \frac{1}{T}\sum_t \omega(t)|t \leq M)} & s = 0 \end{cases} \tag{6.1}$$

where

$$\rho(t, \psi|t > M) = \sum_t (\omega(t|t > M)\mathbb{1}(\omega(t|t > M) > \psi) - \psi)$$

Step 6. Define the week of highest opportunity H as the time point t when the opportunity function is at its maximum. The time until when to overprice is the maximum of H and M.

$$H = \underset{t \in \{1,\ldots,T\}}{\text{argmax}}\, \omega(t)$$

$$t^* = \max\{H, M\}$$

The time t^* is the last time point recommended for overpricing at the level U with the associated market state s, and the cost of prolonged time on market c.

There is no clear methodology to define the strategy given the calculated s ,c, and t^*. It certainly depends on the expert knowledge of the market and the opinion on the property. In this thesis, the strategy is defined ad-hoc as follows:

$$Overprice \quad the \quad property \quad until \quad T \quad if \quad s = 1, \quad c \leq 0.5 \quad and \quad H \leq M \tag{6.2}$$

Figure 6.1 demonstrates the opportunity curves for 8 chosen properties from the unseen test data set. The opportunity curves are drawn on a background of mixed colors. The value of the opportunity function is drawn on the y-axis and the weeks on market on the x-axis. The region colored by green shows the weeks before the expected week of sale, M. The time horizon is up to T. The red region shows the weeks after the expected week of sale. The hit times of hot-market properties are seen when the opportunity curve goes above 0. If the background is all gray, then the property is in the cold market state.

Following the strategy in (6.2) leads to overpricing A, F, G; and not overpricing B, C, D, E, and H in Figure 6.1.

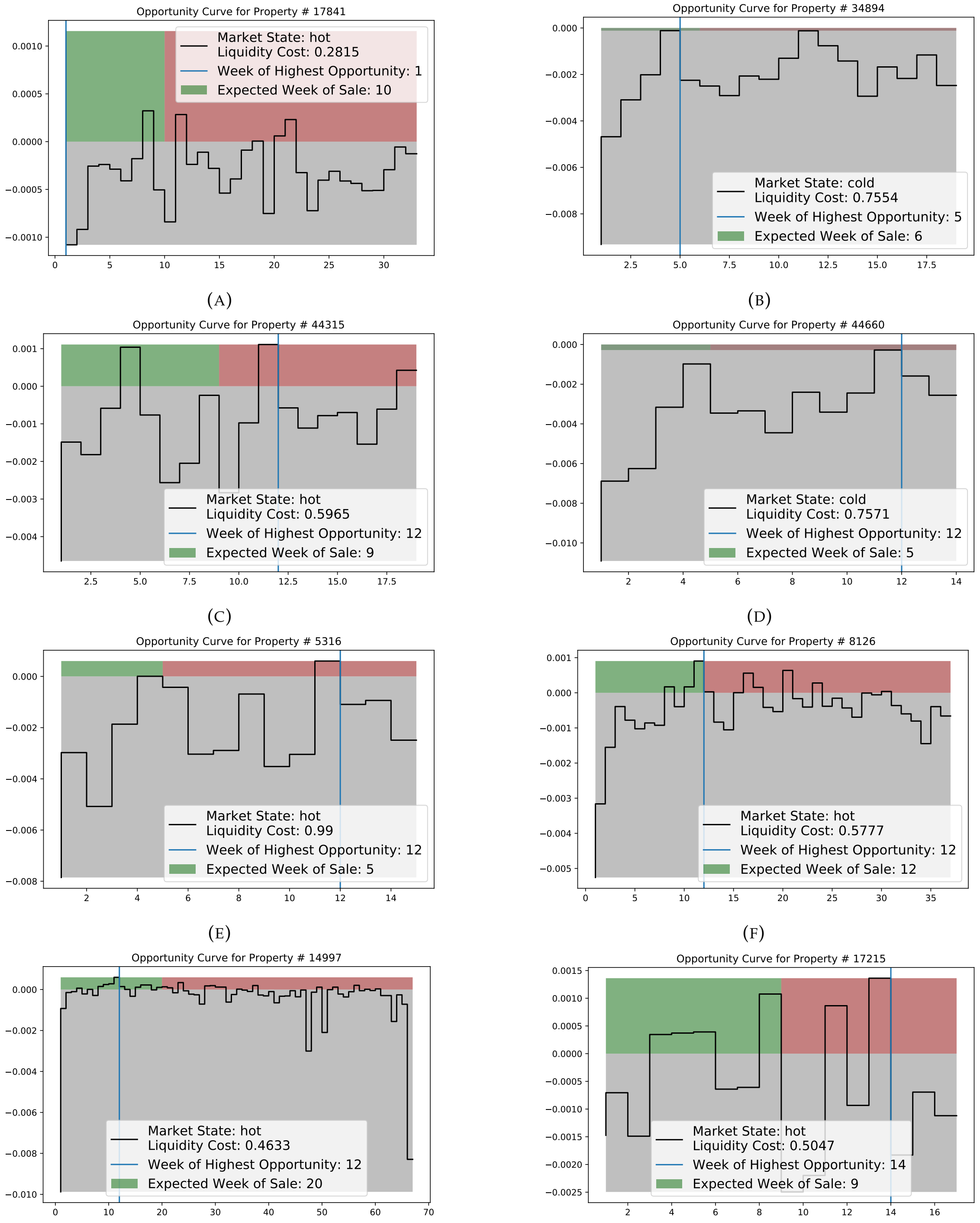
6.2 The Spatial Component: Refining Strategies with Aggregated SHAP values

SHAP values can be clustered and aggregated to derive global relationships from individual interpretations. In this application, they are aggregated on the district level to estimate the location's liquidity sensitivity to DOP.

Figure 6.2. shows the sensitivity to DOP in Berlin's districts, as an illustrative example. Dots represents the properties which are suggested to overprice with respect to the strategy (6.2). District colors which are close to red signifies that if DOP decreases, the liquidity increases faster; whereas district colors which are close to blue signifies that this behavior occurs slower.

Empirically, once the real estate property is on the market, price changes usually occurs in the downward direction (Anglin, Rutherford, and Springer, 2003). Therefore, SHAP values are rather interpreted in this context as the sensitivity to decrease the overpricing level. As seen in Figure 6.2, making a discount on the price of an overpriced property in Mitte (the red region) would lead to a faster sell than making a discount in Marzahn-Hellersdorf (the blue region). Overpricing in the regions whose colors are close to blue entails a higher risk of getting trapped in negative herding as in Taylor (1999). These region might cause the seller to struggle to sell with ever decreasing price levels.

In summary, the temporal component estimates speculative opportunities for a given property, while the spatial component restricts overpricing to the locations where decreasing the DOP level results in the fastest additional speed of sale. The spatial component allows the seller to differentiate the overpricing strategies; different thresholds of c can be applied to properties in different locations.



(A)

(B)

(C)

(D)

(E)

(F)

(G)

(H)

FIGURE 6.1: Opportunity Functions of Selected Properties

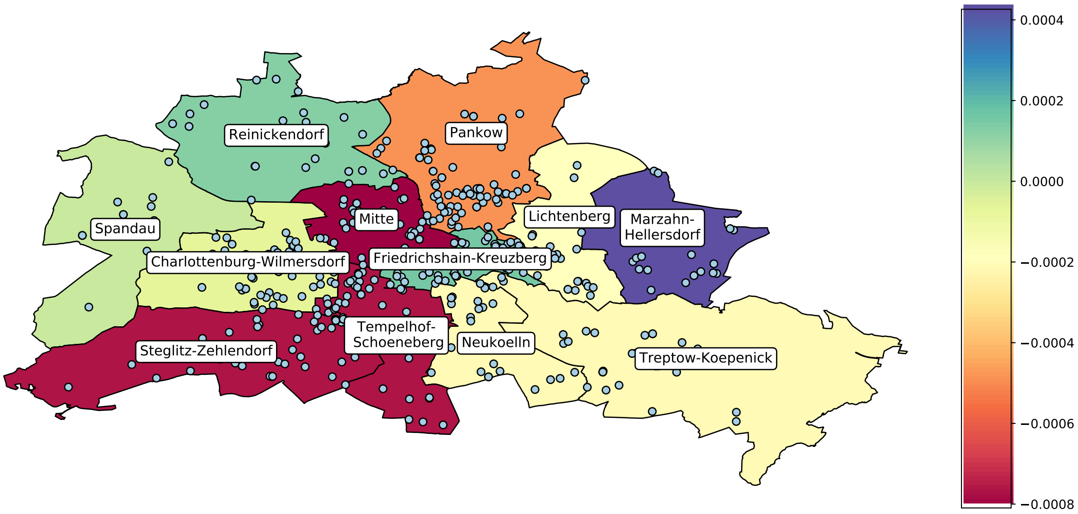


FIGURE 6.2: Refining Overpricing Strategies in Berlin

Chapter 7

Conclusion

7.1 Summary

The paper demonstrates the first attempts to build an automated decision support system in real estate markets by exploiting the survival curves of individual properties that are estimated without making the assumption of proportional hazards. The state of the art Cox PH is inadequate to estimate the survival curves which are necessary for this task. This is the temporal part of the system, that is individual and speculative. The temporal part supports the seller in her auction-like behavior (Haurin et al., 2013).

Spatially aggregated SHAP values of DOP supports the temporal component by differentiating overpricing decisions for different locations. The SHAP values are estimated through the interaction of a data set and the model, hence they are empirical explanations. The spatial component is aggregate and empirical. It can prevent the seller from falling in the trap of negative herding (Taylor, 1999). This is also an alternative use case of SHAP values, which is not only used to explain the model predictions, but also to be integrated into the decision support system.

DeepHit is tested among the recent generation of deep learning models in survival analysis which can estimate non-proportional hazards. It is shown to be superior to Cox PH in calibration and discrimination performances (except some cases close to the mean weeks on market). The inability of Cox PH in its basic form to evaluate complex non-linear interactions and non-proportional hazards in the data (demonstrated in Kaplan-Meier estimations) is a reason for its weaker performance.

7.2 Implications and Limitations

The decision support system is likely to improve the bargaining power of the sellers in the bargaining-and-search game framework. Nevertheless, further applications can be developed by extensively modeling and simulating the buyer side of the market; consequently developing strategies for them, as well. The non-existence of the buyer side of the market is a major drawback of the decision support system. This makes the system inherently speculative.

The decision support system can be integrated to available price estimation tools with ease. As a side effect, it still requires the expert knowledge in fine-tuning the parameters such as the liquidity cost of overpricing in hot and cold market states.

Due to above-mentioned issues, this thesis should be considered rather a speculative exercise of understanding the seller's side on the market than a complete tool for pricing decisions.

Appendix A

Appendix A: Data Cleaning

A.1 Duplication Removal

As described in Chapter 4 - Explanatory Design, a somewhat bold assumption is made that any repeated record is regarded as censored, and the rest as uncensored observation. The repeated records are found by comparing the description texts of properties located in the same postal code area. The cosine distances between tokenized descriptions are calculated, and any description which has a pairwise cosine distance higher than 0.9 is regarded as ’repeated’.

The calculation of the cosine distances between the descriptions of properties are done through calculating the term frequency–inverse document frequency (TF-IDF) matrix for all properties and multiplying the respective rows representing the listings (Salton and McGill, 1986). If the cosine similarity is greater than 0.9, the listings are thought to be repeated. In the next step, the repeated listings are clustered back in order to separate if they identify more than one listings. In order to separate groups of listings from the properties which are identified by the same description text, one dimensional K-means clustering is applied on the multiplication of living area and number of rooms. The number of clusters are determined by the number of clusters which maximizes the silhouette score (Peter, 1987).

A.2 Missing Values

Data cleaning steps consisted of variable selection, missing value imputation and outlier removal.

Only the variables whose proportion of missing values are less than 60 percent are included to the analysis. Therefore, there is a loss of predictive power that might have been gained by the omitted variables like the quality rating of the property or its restoration year.

Following the initial variable selection, the variables which contains the most missing variables are Balcony Available (60 %), Actively Rented (57 %) and Construction Year (17 %). Missing Balcony Available is imputed by a dummy variable whether the description text contains the words "balkon" or "terrasse" in German or not. Similarly, the Actively Rented is imputed by a dummy variable whether the description text contains the word "vermietet" in German.

Imputation of missing Construction Year required more effort. Whenever possible, it is recovered from the descriptions via text mining, and the rest is imputed by the postal code average.

Appendix B

Appendix B: Price Estimation

B.1 OLS Regression Estimation

Table B.1 demonstrates the coefficients estimated by OLS regression. All the variables except Loft, WINTER, and Date of First Advertisement are significant on 0.01 level. The regression results demonstrate that properties which have a balcony, a parking lot, a big living area; which are recently built and of house types except apartment, farmhouse, special building, multi-family house and one-or-two-family house; which are advertised on summer and winter tend to have higher prices. Conversely, properties which are actively rented, that is which have a residing tenant tend to have a lower price. Unexpectedly, properties which have a higher distance to the city centroid also tend to sell at a higher price. This can be explained by the feature engineering; as many small cities with a single postal code have zero distance to the city center, having some distance to the center is an indication that property is in a well-populated urban region. Therefore, this unexpected coefficient might have been so. Another unexpected situation is the reverse signs of Number of Rooms and the Living Area. This should be due to the effect of multicollinearity, as there is a high correlation between these variables, at a level of 0.85.

According to OLS results, Population Density and Market Size have a coefficient near zero, i.e. have no effect to price of a property. This can be explained by the inability of the OLS framework to capture interactions which are not explicitly identified. It should be the case that Distance to the City Centroid have already captured the linear effects of Population Density and Market Size, due to the reason explained above.

In addition to the lack of automatic identification of interaction effects, OLS is restricted by Gauss-Markov assumptions, which can be summarized as linearity, homoscedasticity, normality of errors, strict exogeneity, and full-rank variable matrix (perfect multicollinearity) (Hayashi, 2000). There is a high multicollinearity in the variable matrix, so that even implementing test for heteroscedasticity was not possible. The invalidity of the normality assumption in errors can be visually inspected in Figure B.1. The Q-Q plot is obtained by plotting the quantiles of the residuals on the quantiles of a theoretical normal distribution. If residuals are normally distributed, the blue points should lie on the red line. The high deviations from the red line shows that the residuals from the OLS is clearly not normal distributed. Note that the deviations are much higher at low and high values and the graph is cropped for the residuals between 0 and 1, in order to ensure that the graph is visually appealing.

Variable	**Coefficient**	**P-value**
Constant	6.800	0.000
Balcony Available	0.1078	0.000
Actively Rented	-0.0989	0.000
Parking Available	0.1286	0.000
Farmhouse	-0.1758	0.000
House (Unknown)	-0.0262	0.000
Loft	0.0466	0.124
Maisonette	0.429	0.000
Mansion	0.1652	0.000
Multi-Family House	-0.1993	0.000
One-or-Two-Family House	-0.0803	0.000
Penthouse	0.2622	0.000
Apartment	-0.0488	0.000
Semi-Detached House	0.0791	0.000
Special Building	-0.1612	0.000
Terraced House	0.0369	0.000
SPRING	0.0102	0.001
SUMMER	0.0098	0.005
WINTER	-0.0016	0.631
Latitude	-0.0758	0.000
Longitude	0.0199	0.000
Construction Year	0.0022	0.000
Living Area	0.0002	0.000
Number of Rooms	-0.0237	0.000
Population Density	0.0001	0.000
Market Size	0.0003	0.000
Distance to the City Centroid	0.0077	0.000
Date of First Advertisement	0.0909	0.084
R-squared: 0.312, non-robust covariance		

TABLE B.1: OLS Coefficients

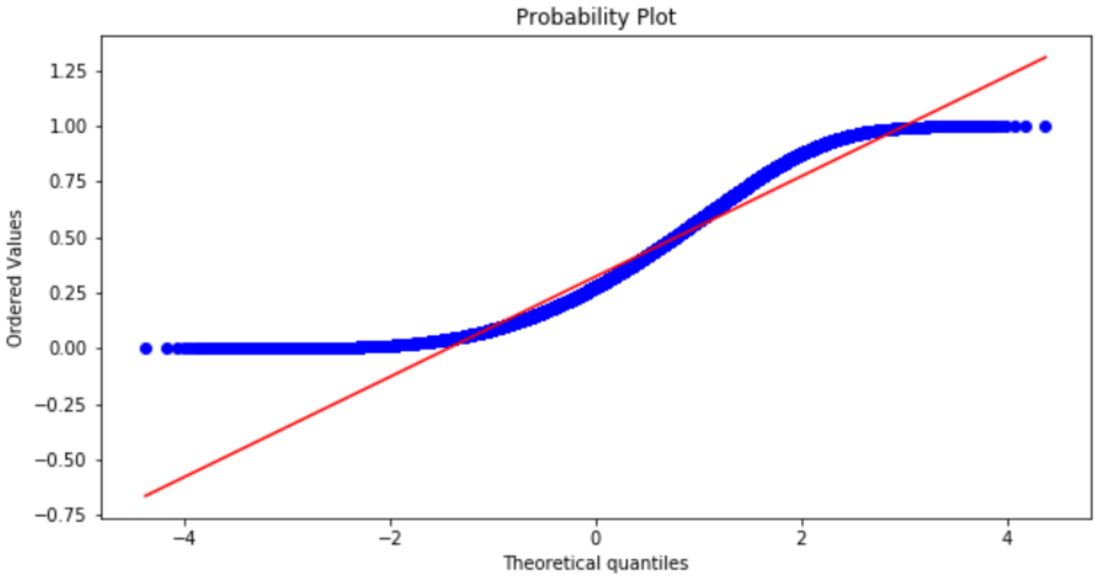


FIGURE B.1: Q-Q Plot for the OLS Residuals, cropped to ensure visibility

B.2 Hyper-Parameter Tuning for Random Forest

Random forest regression is a tree-based homogeneous ensemble model. Decision trees are known to be a good detector of interactions, but they tend to overfit the data. Random forests are developed to decrease the generalization error of decision trees by two main innovations: First, a randomly drawn sample with replacement (bootstrap sample) of the data is used to grow a tree. Then a randomly selected subset of variables is chosen as candidate variables for splitting. Averaging over trees yields the final result (Breiman, 2001a).

The most important hyper-parameters in random forest regression are the number of variables which are randomly sampled as candidate parameters at each split (Variables at Split) and the number of trees to grow. More trees and more variables at splits might cause overfitting, hence they should be controlled by checking the out of bag errors.

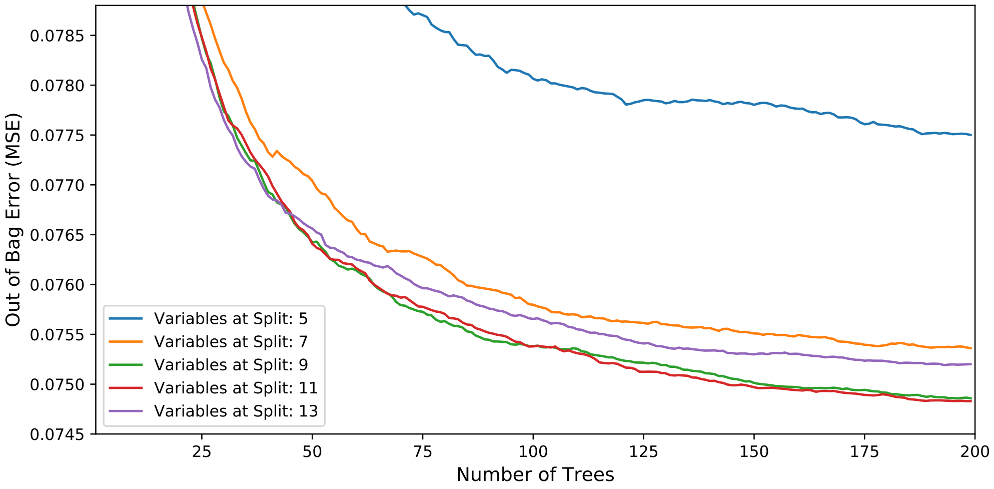


FIGURE B.2: Hyperparameter Tuning for Random Forest

In this setting, different values for 'Variables at Split' (5,7,9,11,13) are tried with increasing number of decision trees by 10 in the random forest. It is demonstrated in Figure 4.2 that the out-of-bag error in the validation set decreases with increasing Variables at Split up to 11, then begins to decrease from 11 to 13. After 200 trees, the increase in trees does not yield additional out-of-bag performance. Hence, the hyper-parameters are chosen as 11 for Variables as Split and 200 for the number of trees.

B.3 Feature Importance in Random Forest

It is possible to derive importance scores for the variables during the implementation of the random forest model. However, this importance measure inflates continuous variables. Usually the dummy variables are assigned to low feature importance scores. Instead, using SHAP values gives a more accurate ordering of features in terms of their importance in explaining prices of properties (Figure B.3). It is also possible to see the direction of the effects (Figure B.4). Note that the SHAP values are calculated on a small data set of 5000 records sampled randomly from the test data set.

The least important variables are the season dummies and Date of First Advertisement, while the most important variables are the Cartesian coordinates, Population Density, Construction Year, and Living Area, as seen in Figure B.3. In addition, Market Size is also an important predictor, as ranked at 9th place. Those variables had coefficients near zero in OLS results. Therefore, the effects of them is said to be captured better by random forest.

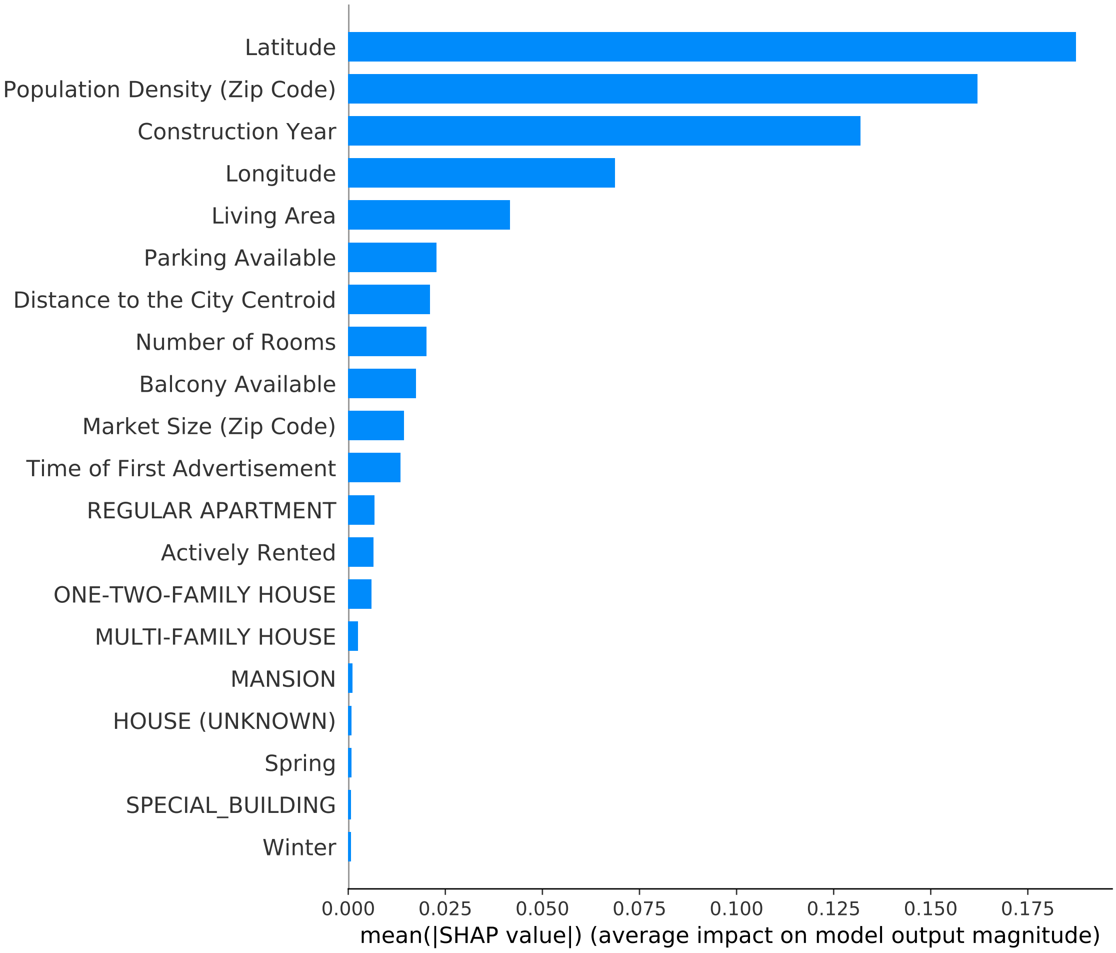


FIGURE B.3: Shapley Values for Random Forest (Absolute)

In Figure B.4, the direction of nonlinear effects are plotted. It can be said that the direction of the effects are usually not uniform in the distribution of the variables. Many variables have mixed effects, such as Latitude, Longitude, Number of Rooms, Population Density, and Construction Year. Date of First Advertisement shows that the first middle of the year shows a negative effect on price, while the rest of the year has a positive effect on price. This explains the insignificant coefficient estimation of this variables in OLS results, the direction is not surely determined.

In general, there is a positive effect on price is seen on Balcony Avaliable, Date of First Advertisement, and Distance to the City Centroid. The variable Actively Rented shows a discernible negative relationship with price.

B.4 Discussion on the Comparison of Performances

Random forest is fit with chosen parameters on the training set, and likewise the OLS regression. Their performances on the unseen test data are compared in Table B.2. Modeling with random

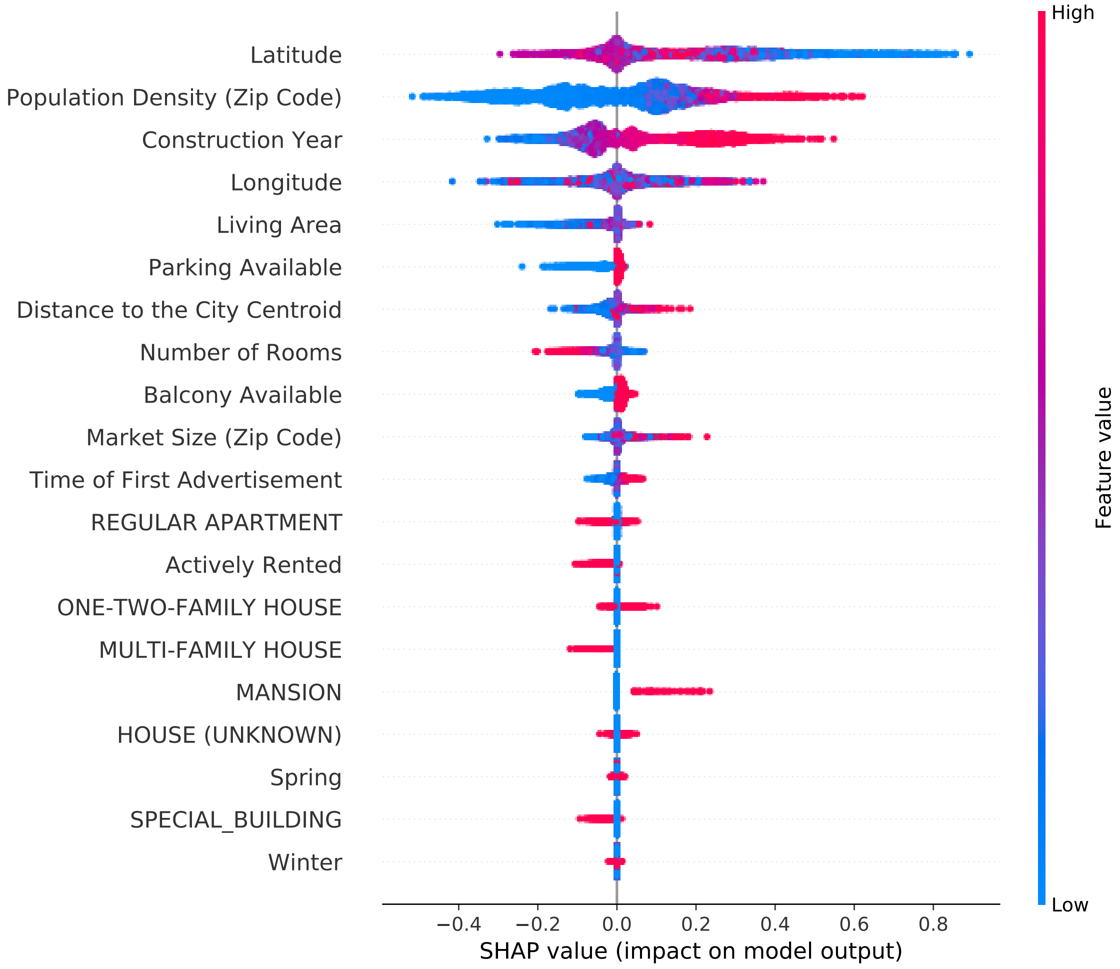


FIGURE B.4: Shapley Values for Random Forest (Relative, Direction)

forest gives a considerable boost in the performances in both the training and test sets, although we can observe an overfitting when we inspect the difference between the training and test errors of random forest.

Model	MSE (Training Data)	MSE (Test Data)
Linear Regression	0.216	0.216
Random Forest	0.008	0.063

TABLE B.2: Out of Bag Mean Squared Error

This paper follows the philosophy of Breiman (2001b) in defining the algorithmic modeling culture. Breiman distinguishes between two data cultures, one that assumes that data is generated by a given stochastic data model (data modeling culture) and focuses on estimation of parameters, and the other (including himself) uses the best-performing algoritmic models and treats the data mechanism as unknown (algorithmic modeling culture). Before constructing the variable DOP, the concept of what the real value of a property actually is and if any modeling approach captures this real value at all should be discussed. The assumption that the predictions resulting from OLS regression is the real value of a property, leads to properties which are overpriced more than 15 times which is not realistic. On the other hand, random forest as a better performing model attains a higher performance which reduces the bound of overpricing down. Finding a good performing algorithmic model for the price estimation is crucial for studying strategic pricing based on liquidity estimations. It should be accepted the effect of strategic overpricing is mixed with the estimation errors in the variable DOP. For the sake for obtaining a good data quality for DOP which is the most important variable in pricing decisions; we restrict the data in a way that DOP levels are between 0.8 and 1.2. Our assumption is that a property may at most 20 percent down or overpriced. The histogram of estimated DOP is demonstrated in Figure B.5, where red lines suggests the cutting points for the data set. Data points which correspond to outside of the red lines are removed.

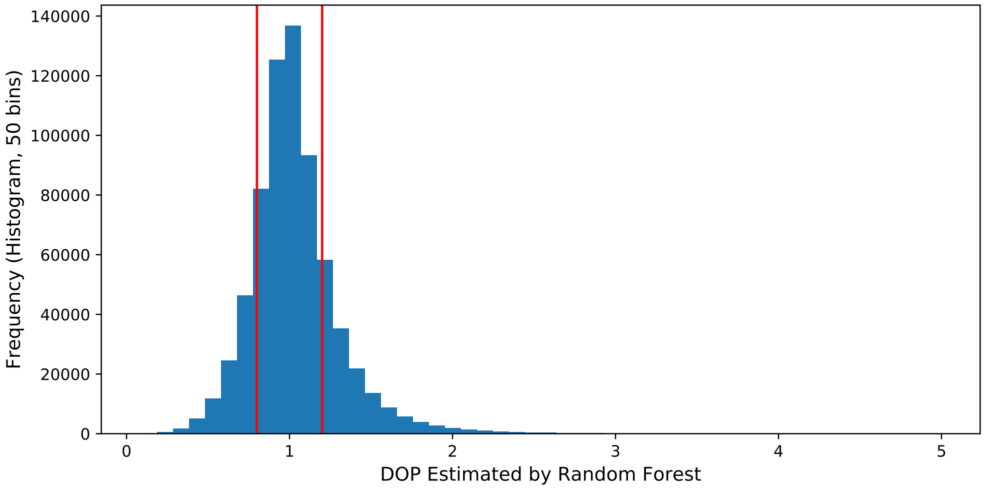


FIGURE B.5: Histogram of DOP Estimated by Random Forest

As the distribution of DOP depends heavily on the statistical model which is used to estimate it, the automated pricing decisions should be made carefully. As the warning of Kluger and Miller (1990) suggests, with data dependent semi- and non-parametric models like Cox PH, the unrealistic price levels would give realistic liquidity estimations just because the unrealistic price levels did not exist in the training set beforehand. It is tempting to construct an optimization mechanism as in Jerenz (2008) where he defined an automated revenue management system in second hand car market; with a price estimation by OLS and a liquidity estimation by AFT. Both of the models are fully parametric and their use can be classified in Breiman's data modeling culture. Designing pricing strategies for housing with non-parametric estimations of price and liquidity would be made carefully. Therefore, this paper designs them around the decision on "whether to overprice", but not around "how much to overprice".

Bibliography

Akerlof, G. A. (1970). “The Market for ’Lemons’: Quality Uncertainty and the Market Mechanism”. In: *The Quarterly Journal of Economics* 84.3, pp. 488–500.

Anglin, P., R. Rutherford, and T. Springer (Feb. 2003). “The Trade-Off Between the Selling Price of Residential Properties and Time-on-the-Market: The Impact of Price Setting”. In: *The Journal of Real Estate Finance and Economics* 26, pp. 95–111. DOI: 10.1023/A:1021526332732.

Antolini, L., P. Boracchi, and E. Biganzoli (2005). “A time-dependent discrimination index for survival data”. In: *Statistics in Medicine* 24.24, pp. 3927–3944. DOI: 10.1002/sim.2427. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2427.

Arnold, M. A. (1999). “Search, Bargaining and Optimal Asking Prices”. In: *Real Estate Economics* 27.3, pp. 453–481. DOI: 10.1111/1540-6229.00780. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.00780. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00780.

Berkovec, J. A. and J. L. Goodman (Dec. 1996). “Turnover as a Measure of Demand for Existing Homes”. In: *Real Estate Economics* 24.4, pp. 421–440. ISSN: 1540-6229. DOI: 10.1111/1540-6229.00698. URL: https://doi.org/10.1111/1540-6229.00698.

Box-Steffensmeier, J. M. and B. S. Jones (2004). *Event History Modeling: A Guide for Social Scientists*. Analytical Methods for Social Research. Cambridge University Press. DOI: 10.1017/CBO9780511790874.

Breiman, L. (Oct. 2001a). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32.

– (Aug. 2001b). “Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)”. In: *Statistical Science* 16. DOI: 10.1214/ss/1009213726.

Cajias, M. and P. Freudenreich (Dec. 2017). “Exploring the determinants of liquidity with big data - market heterogeneity in German markets”. In: *Journal of Property Investment Finance* 36, pp. 00–00. DOI: 10.1108/JPIF-01-2017-0006.

Cajias, M. and A. Heller (Jan. 2018). “Understanding the rent-liquidity co-movements in the real estate markets: Large sample evidence from German micro data”. In: eres2018$_5$4. URL: https://ideas.repec.org/p/arz/wpaper/eres2018_54.html.

Collett, D. (2014). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC. DOI: 10.1201/b18041.

Cox, D. (1972). “Regression models and life tables”. In: *Journal of the Royal Statistical Society B* 34 (2), 187–220.

Cubbin, J. (1974). “Price, quality, and selling time in the housing market”. In: *Applied Economics* 6.3, pp. 171–187. DOI: 10.1080/00036847400000017. eprint: https://doi.org/10.1080/00036847400000017. URL: https://doi.org/10.1080/00036847400000017.

Dirick, L., G. Claeskens, and B. Baesens (2017). “Time to default in credit scoring using survival analysis: a benchmark study”. In: *Journal of the Operational Research Society* 68.6, pp. 652–665. ISSN: 1476-9360. DOI: 10.1057/s41274-016-0128-9. URL: https://doi.org/10.1057/s41274-016-0128-9.

Follain, J. R. and O. T. Velz (1995). "Incorporating the Number of Existing Home Sales into a Structural Model of the Market for Owner-Occupied Housing". In: *Journal of Housing Economics* 4.2, pp. 93–117. URL: https://EconPapers.repec.org/RePEc:eee:jhouse:v:4:y:1995:i:2:p:93-117.

Fornili, M. et al. (2014). "Piecewise Exponential Artificial Neural Networks (PEANN) for Modeling Hazard Function with Right Censored Data". In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Ed. by Enrico Formenti, Roberto Tagliaferri, and Ernst Wit. Cham: Springer International Publishing, pp. 125–136. ISBN: 978-3-319-09042-9.

Fotso, S. (2018). "Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework". In: *arXiv e-prints*, arXiv:1801.05512, arXiv:1801.05512. arXiv: 1801.05512 [stat.ML].

Gensheimer, M. F. and B. Narasimhan (2019). "A scalable discrete-time survival model for neural networks". In: *PeerJ* 7:e6257. DOI: 10.7717/peerj.6257.

Gerds, T. A. and M. Schumacher (2006). "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times". In: *Biometrical Journal* 48.6, pp. 1029–1040. DOI: 10.1002/bimj.200610301. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200610301.

Goel, M., P. Khanna, and J. Kishore (Oct. 2010). "Understanding survival analysis: Kaplan-Meier estimate". In: *International journal of Ayurveda research* 1, pp. 274–8. DOI: 10.4103/0974-7788.76794.

Grambsch, P. M. and T. M. Therneau (Sept. 1994). "Proportional hazards tests and diagnostics based on weighted residuals". In: *Biometrika* 81.3, pp. 515–526. ISSN: 0006-3444. DOI: 10.1093/biomet/81.3.515.

Harrell, F. E. (2006). *Regression Modeling Strategies*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387952322. DOI: 10.5555/1196963.

Harrell F. E., Jr et al. (May 1982). "Evaluating the Yield of Medical Tests". In: *JAMA* 247.18, pp. 2543–2546. ISSN: 0098-7484. DOI: 10.1001/jama.1982.03320430047030. eprint: https://jamanetwork.com/journals/jama/articlepdf/372568/jama_247_18_030.pdf. URL: https://doi.org/10.1001/jama.1982.03320430047030.

Haurin, D. (1988). "The Duration of Marketing Time of Residential Housing". In: *Real Estate Economics* 16.4, pp. 396–410. DOI: 10.1111/1540-6229.00463. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.00463. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00463.

Haurin, D. et al. (2010). "List Prices, Sale Prices and Marketing Time: An Application to U.S. Housing Markets". In: *Real Estate Economics* 38.4, pp. 659–685. DOI: 10.1111/j.1540-6229.2010.00279.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6229.2010.00279.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6229.2010.00279.x.

Haurin, D. et al. (2013). "List price and sales prices of residential properties during booms and busts". In: *Journal of Housing Economics* 22.1, pp. 1–10. URL: https://EconPapers.repec.org/RePEc:eee:jhouse:v:22:y:2013:i:1:p:1-10.

Hayashi, F. (2000). *Econometrics*. Princeton, NJ [u.a.]: Princeton Univ. Press. XXIII, 683. ISBN: 0691010188.

Hort, K. (2000). "Prices and turnover in the market for owner-occupied homes". In: *Regional Science and Urban Economics* 30.1, pp. 99 –119. ISSN: 0166-0462. DOI: https://doi.org/10.1016/S0166-0462(99)00028-9. URL: http://www.sciencedirect.com/science/article/pii/S0166046299000289.

Jerenz, A. (Jan. 2008). "Revenue Management and Survival Analysis in the Automobile Industry". In: *Revenue Management and Survival Analysis in the Automobile Industry*, pp. 1–168. DOI: 10.1007/978-3-8349-9840-8.

Kaas, L. et al. (2017). "Low Homeownership in Germany - A Quantitative Exploration". In: 6775. URL: https://ideas.repec.org/p/ces/ceswps/_6775.html.

Kaplan, E. L. and P. Meier (1958). "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282, pp. 457–481.

Katzman, J. L. et al. (Feb. 2018). "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC Medical Research Methodology* 18 (1), pp. 1471–2288. DOI: 10.1186/s12874-018-0482-1.

Kluger, B. D. and N. G. Miller (1990). "Measuring Residential Real Estate Liquidity". In: *Real Estate Economics* 18.2, pp. 145–159. DOI: 10.1111/1540-6229.00514. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.00514. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00514.

Knight, J. R. (2002). "Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes". In: *Real Estate Economics* 30.2, pp. 213–237. DOI: 10.1111/1540-6229.00038. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.00038. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00038.

Kohl, S. (2016). "Urban History Matters: Explaining the German–American Homeownership Gap". In: *Housing Studies* 31.6, pp. 694–713. DOI: 10.1080/02673037.2015.1121213. eprint: https://doi.org/10.1080/02673037.2015.1121213. URL: https://doi.org/10.1080/02673037.2015.1121213.

Krainer, J. (1999). "Real estate liquidity". In: *Economic Review*, pp. 14–26. URL: https://EconPapers.repec.org/RePEc:fip:fedfer:y:1999:p:14-26:n:3.

– (Jan. 2001). "A Theory of Liquidity in Residential Real Estate Markets". In: *Journal of Urban Economics* 49 (1), pp. 32–53.

Kvamme, H. (2019). "PyCox - Survival Analysis with PyTorch". In: *Github*. URL: https://github.com/havakv/pycox.

Kvamme, H. and Ø. Borgan (2019a). "Continuous and Discrete-Time Survival Prediction with Neural Networks". In: *arXiv e-prints*, arXiv:1910.06724, arXiv:1910.06724. arXiv: 1910.06724 [stat.ML].

Kvamme, H. and O. Borgan (2019b). "The Brier Score under Administrative Censoring: Problems and Solutions". In: *arXiv e-prints*, arXiv:1912.08581, arXiv:1912.08581. arXiv: 1912.08581 [stat.ML].

Lee, C. et al. (Apr. 2018). "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks". In: *AAAI Publications, Thirty-Second AAAI Conference on Artificial Intelligence*.

Lessmann, S. and S. Voß (2017). "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy". In: *International Journal of Forecasting* 33.4, pp. 864 –877. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2017.04.003. URL: http://www.sciencedirect.com/science/article/pii/S016920701730050X.

Lessmann, S. et al. (2018). "Price Management in the Used-Car Market: An Evaluation of Survival Analysis". submitted.

Lin, M., H. Lucas, and G. Shmueli (Dec. 2013). "Too Big to Fail: Large Samples and the p-Value Problem". In: *Information Systems Research* 24, pp. 906–917. DOI: 10.1287/isre.2013.0480.

Lundberg, S. M. and S. Lee (2017). “A unified approach to interpreting model predictions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. 41. Curran Associates, Inc., pp. 4765–4774.

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Peter, J. R. (1987). “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20, pp. 53 –65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *CoRR* abs/1602.04938. arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.

Salton, G. and M. J. McGill (1986). “Introduction to modern information retrieval”. In:

Shapley, L. S. (1953). “A value for n-person games”. In: *Contributions to the Theory of Games* 28 (2), pp. 307–317.

Smith, B. C. (2009). “Spatial Heterogeneity in Listing Duration: The Influence of Relative Location to Marketability”. In: *Journal of Housing Research* 18.2, pp. 151–172. ISSN: 10527001. URL: http://www.jstor.org/stable/24861477.

Stein, J. C. (1993). “Prices and Trading Volume in the Housing Market: A Model with Downpayment Effects”. In: Working Paper Series 4373. DOI: 10.3386/w4373. URL: http://www.nber.org/papers/w4373.

Strumbelj, E. and I. Kononenko (2014). “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and information systems* 41 (3), pp. 647–665.

Taylor, C. R. (1999). “Time-on-the-Market as a Sign of Quality”. In: *The Review of Economic Studies* 66.3, pp. 555–578. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/2567014.

Therneau, T. M., P. M. Grambsch, and Fleming T. R. (1990). “Martingale-based residuals for survival models”. In:

Voigtländer, M. (2009). “Why is the German Homeownership Rate so Low?” In: *Housing Studies* 24.3, pp. 355–372. DOI: 10.1080/02673030902875011. eprint: https://doi.org/10.1080/02673030902875011. URL: https://doi.org/10.1080/02673030902875011.

Wei, L. J. (1992). “The accelerated failure time model: A useful alternative to the cox regression model in survival analysis”. In: *Statistics in Medicine* 11.14-15, pp. 1871–1879. DOI: 10.1002/sim.4780111409. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780111409.

Wood, J. H. and N. L. Wood (1985). *Financial Markets*. Harcourt Brace Jovanovich.

Yavas, A. (1995). “Can Brokerage Have an Equilibrium Selection Role?” In: *Journal of Urban Economics* 37.1, pp. 17 –37. ISSN: 0094-1190. DOI: https://doi.org/10.1006/juec.1995.1002. URL: http://www.sciencedirect.com/science/article/pii/S0094119085710029.

Yavas, A and S. Yang (1995). “The Strategic Role of Listing Price in Marketing Real Estate: Theory and Evidence”. In: *Real Estate Economics* 23.3, pp. 347–368. DOI: 10.1111/1540-6229.00668. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.00668. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00668.