1 Introduction

This exam paper examines the Danish market for used cars by utilizing publicly available data on the supply of used cars at Bilbasen.dk, which is by far the largest online platform in Denmark. Across different markets and models our scraper allows us to retrieve data on all current listings on Bilbasen.dk. The number of listings is around 50,000 (average during one week of August 2017). In comparison, the entire Danish fleet consists of around 2,500,000 cars[1]. Hence, during an average week 1.5 pct. of the entire fleet is for sale at Bilbasen.dk. Further, according to Statistics Denmark a total of 45,000 used cars were sold during July 2017. This substantiates the importance of Bilbasen.dk in the market for used cars and the representativeness of our sample. For many families, cars probably constitute the largest investment among durable goods. Hence, this market is of great importance, and not only for the consumers: the taxation of cars is a heavily politicized field, which is shown in the number of different taxation types (i.e. weight, usage, efficiency, price).

Our research question relates traditional economic theory with methods within social data science: First, we examine the market for used cars and in particular spatial price differences. Second, we evaluate out of sample predictions on the prices of used cars.

Cars as a consumer good has been examined in depth in the literature. A classical reference is Akerlof's celebrated paper about asymmetric information and 'lemons'. It is not clear how the rise of online platforms as Bilbasen.dk has changed the informational asymmetry between buyer and seller. On one hand, buyers might learn to read the signs and identify the 'lemons'. Sellers, on the other hand, might learn how to hide the 'lemons', for example with information overload on car characteristics. Given the format of listings on Bilbasen.dk, this effect is probably not present and runs rather in the opposite direction: sellers are forced to post detailed information about the cars for sale, which increases the level of competition in the market.

Whereas the effect of online platforms on asymmetric information can be discussed, there are numerous examples of online platforms greatly reducing consumers' search costs. Hence, a consumer can instantly compare similar cars and their attributes. The influence of online platforms on equilibrium prices is interesting, but beyond the scope of this paper. What this paper sets out to do is more modest: it looks for differences in prices across similar cars. This entails looking at spatial differences as well as product level differences. Large spatial differences in prices would, for example, be an indication of consumers still having significant costs associated with car purchases (search costs, transportation to dealer, lost income, imperfect information etc.). This would also indicate smaller markets, whereas small differences would indicate an integrated market across Denmark. Hence, spatial differences can help us understand consumers' behavior in the market for used cars. Further, by examining the market structure, this paper contributes with a deeper understanding of a durable good, which requires a large investment for most households. This demand is extremely interesting for policymakers, for example in relation to the optimal taxation scheme and environmental policy. Lastly, it is of interest to all consumers how the informational barrier between buyer and seller can be alleviated. This paper provides a method to predict car prices based on observable characteristics. Although the informational barrier typically relates to unobservable characteristics (at least for the econometrician), this analysis still gives the consumer a tool to better understand the market for used cars.

In the Industrial Organization literature, most studies have focused on the purchase of new cars and not used cars. This is due to data scarcity, but also the complexity in modeling secondary markets. A good example is Berry, Levinsohn and Pakes (1995) who analyzes equilibrium prices in the U.S market for new cars based on aggregated product-level data. However, it is unclear if the market for used cars can build on similar aggregated methods. In standard economic theory, a good can be represented by a set of characteristics.

[1] DST: BIL6

In the market for used cars, every good is different. Similar used cars (i.e. Audi A4) cannot be described by the same set of characteristics. A noteworthy example of an analysis of the used car market is Gavazza, Lizzeri and Roketskiy (2014), who empirically investigates welfare effects of secondary markets for used cars by explicitly modelling the heterogeneity.

The rest of this paper is structured as follows: section 2 provides an overview of the web-scrapping procedure and elaborates on ethics governing our work. Section 3 provides a descriptive analysis of the market for used cars. Section 4 presents a baseline fixed effects regression for estimating spatial price difference as well as the estimates. Section 5 examines predictions based on the baseline model. Finally, section 6 concludes.

2 Web-scraping Procedure and Ethics

The data for the analysis was retrieved from Bilbasen.dk on 19th of August 2017. The scraper iterates over all URLs from the total supply of used cars on the 18th of August 2017. Given the structure of the target website, it was not possible to extract all the information needed from the overview page. Hence, the scraper requests information from each individual car for sale, which amounts to requesting approximately 50,000 URLs. This includes information on price, the seller's contact information, as well as car characteristics such as model, make, year of registration, weight, horse power etc. The scraper can be examined in the attached Jupyter Notebook file section 1 and accounts for differences in HTML scrips across listings on Bilbasen.dk. The listings are, generally, quite consistent, which eases the manipulation needed.

The scraper acted in accordance with the robots.txt file provided by Bilbasen.dk specifying the disallowances. The disallowances include, among others, user reviews and information on users. Further, it is not allowed to redirect listings to another website, thereby capitalizing on Bilbasen.dk's business model. Also, we do not extract any metadata from the URLs as this could, potentially, harm the business model. Finally, we extract information about the size of each dealer (measured as total listings). This information is only used in such a way that no individual dealer can be identified.

3 Descriptive Analysis and Data Collection

After cleaning the data and filtering erroneous observations our dataset consists of 42,000 observations. This includes, for example, handling discrepancies in the HTML structure but also deleting erroneous observations such as an extreme mileage and new cars listed as used cars. To ensure comparison (i.e. from different taxation schemes) hybrid and electric cars are disregarded. All in all, the data cleaning process removed 10,000 observations. It is out of scope for this paper to try and remedy the possible selection bias related to this but it is of cource kept in mind. Besides Bilbasen.dk, also data from Postnord is used to convert zip-codes to municipalities and regions. The variables are constructed using only listed characteristics.

3.1 Bag of Words (BoW) and Dictionaries

Alongside with the table on each listing, there is a description of the car written by the seller. This is mainly a listing of the car characteristics, but it also contains non-neutral wording not related to car characteristics but a promotional purpose. We want to investigate whether the variance in the wording and magnitude hereof can improve the precision of our regression and prediction models, as these descriptions might include observations that are otherwise unobserved for the econometrician. We tokenize the descriptions of cars and excludes any Danish stop words, that are present in great numbers in most texts. For each description of

a car we extract a bag of words, a multiset, with which the frequency of the words is obtained. To find differences in the bags of words we implement a dictionary method following Grimmer and Stewart (2013). We choose and group the adjectives that constitutes a specific wording, e.g. connotating positivity, warranty etc. To maximize the variance and precision we omit words occurring in more than 80 % and less than 2 % of the description. Only words with no ambiguity in their connotations are chosen. Further, we carefully ensure that e.g. the positive words are unlikely to appear in a negative context (e.g. big could be 'big trunk' or 'big scratch'), such that the connotation is altered. Similarly, we have ensured, by looking at the descriptions, the words are not preceded by adverbs or prefixes reverting the connotation, e.g. 'not pretty'. This does not seem to constitute a problem as the purpose is selling and the majority of descriptions only contain various degrees of positives wordings and not negative. Generally, the dictionary method is very hard to validate, but we believe we have solved the most vivid problems. Our later regression analysis shows that only the positive connotations are significant, therefore all other dictionaries are omitted from in the prediction. The positive dictionary consists of the following:

Dictionary: *attraktiv, flot, orig, pæn, fin, godt, fantastisk, yderst, fejlfri, ikke-ryger, ikke ryger, velhold, velkør, facelift, premium, veludstyr, lækker, skarp, komfort, kørekomfort, bedst, dynamisk, performance, tysk, unik, tilbud, utrolig, nysyn, rumme, yderst, synet, super, hysterisk, god stand, økonomisk*

From this we compute a variable counting the number of words from the dictionary present in each listing's description. 35 pct. of the observations include more than one of the words in the dictionary. In what follows, we present some descriptive statistics to examine the market for used cars further.

3.2 Descriptive Statistics

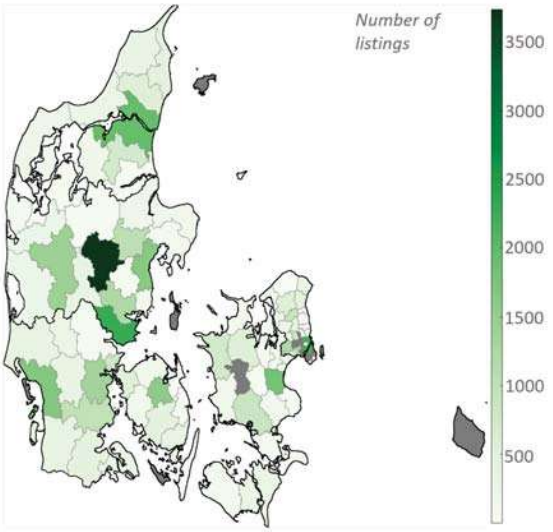
To get a first impression of the data and the cars for sale we list some summary statistics for car characteristics below. The typical car has driven around 100,000 kilometers, is 7 seven years old, has manual transmission and is among the top five manufacturers, who jointly has 40 pct. share in the market for used cars. Table 1 below shows the summary statistics for the cleaned dataset. This dataset contains 42,592 observations.

Table 1: Summary statistics

	Mean	Std. error	Min.	Max.
Price (DKK)	184,471	184,185	902	7,077,500
Kilometers driven	99,506	87,751	1000	847,000
Age	6.32	4.49	0	62
HP	137.93	72.37	26	963
Number of doors	4.67	0.77	2	6
Dictionary	1.42	1.50	0	16
Km/l	19.42	5.09	4.4	33.3
Manual transmission	0.71	0.45	-	-
Diesel	0.51	0.50	-	-
ESP	0.82	0.39	-	-
Popular makes	VW (10 %), Audi (8 %), Peugeot (7.5 %), Ford (7.5 %), Mercedes (7%)			
Popular models	Passat (2 %), Octavia (2 %), A6 (1.5 %), A4 (1.5 %), Focus (1.5 %)			

To get a deeper understanding of the supply of used cars, figure 1 shows the distribution of used cars across the Danish municipalities in a choropleth map. Denmark has 98 municipalties and our dataset includes

Figure 1: Distribution of listings across municipalities

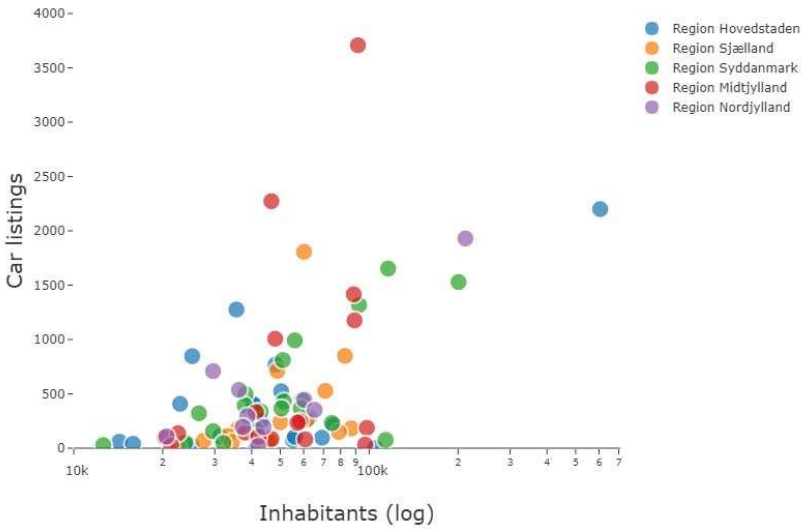


data for 86 of them, with listings ranging from 1 to 3,737. Note that 14 municipalities having more than 30% of the listings. Included in the map are region boundaries, which shows that listings, despite their location in a few municipalities, are somewhat evenly spread out across the 5 regions of Denmark. This might be explained by large dealers clustering in just a few municipalities, with the surrounding municipalities being served a few smaller dealers. The number of listings vary from 4,800 in Region Nordjylland to 13,000 in Region Midtjylland. Interestingly, Silkeborg has the maximum number of listings among the municipalities. This is despite being only the 11th most populous municipality.

In order to say something about the actual supply of cars, one need to look at the demand as well. Figure 2 plots the number of listings in the municipalities against the log of inhabitants in the municipalities. One would expect a proportional relationship between the number of listings and how populous a municipality is. This is only partly true and there is no obvious pattern across the regions either. For a lot of municipalities, the number of listings does not seem to vary with the population size. This seem to confirm the hypothesis that dealers cluster in just a few municipalities, regardless of that municipalities population, which indicates that buyers are willing to travel a certain distance when buying a car. When examining this, one should keep in mind that there are other things than population size that determines the aggregate supply of used cars. For instance, one would need to control for other means of transportation as well: if a municipality is more populous, they will probably have better public transport solutions.

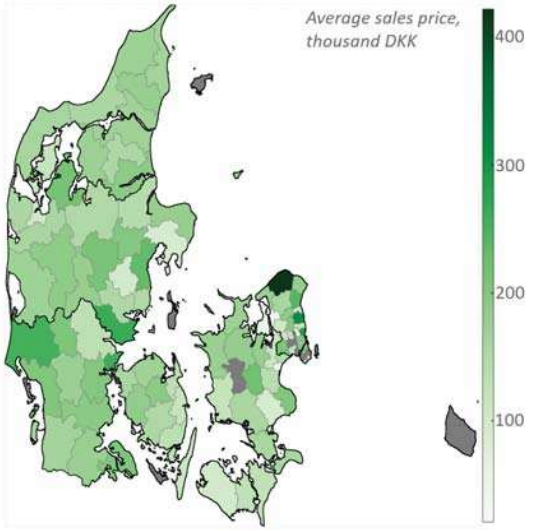
The prices of the listed cars vary less than listings, which is clear from figure 3. The odd one out is Gribskov municipality, with an average sales price of 421,000 DKK, is explained by Gribskov only having 3 listings. This means that despite a large gap in number of listings from municipality to municipality, prices are in a somewhat narrow range. The overall distribution of prices seem to vary little from region to region, but with distinct variation in regions. The distribution across regions can be seen from figure xx in the app

Figure 2: Listings and population size in the municipalities



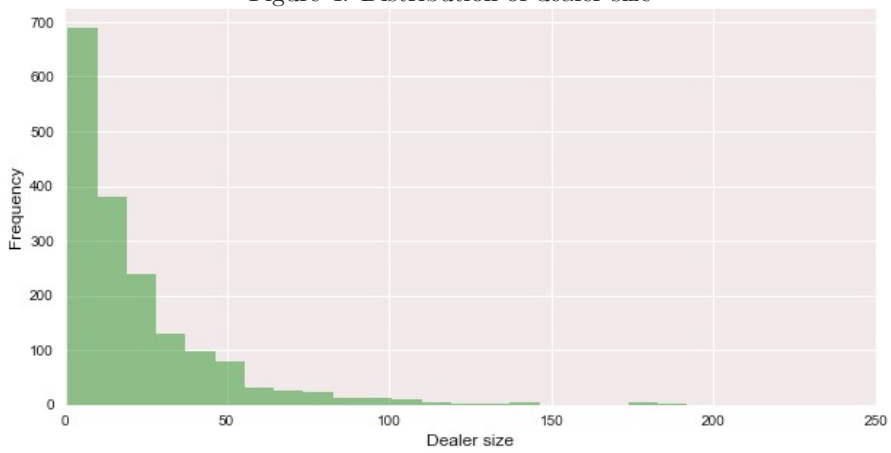
Note: Copenhagen has been omitted.
Link to interactive plot here:

Figure 3: Distribution of sales prices across municipalities



The above analysis showed a somewhat large variance in the market for used cars, at least when one focus on a municipal level. To better understand the market structure, it is informative to look at the distribution of car dealers (figure 4) and not only the listings themselves. Around 1,800 car dealers have active listings on Bilbasen.dk. Of these, only 10 pct. has a portfolio of more than 50 cars. This picture is constant across regions, where the average number of listing per dealer is around 25. This is in contrast to an average of 100 listing for the 10th quantile in the distribution. This gives the impression of a fragmented market with few big dealers. These few large dealers seem to cluster in just a few municipalities as mentioned previously.

Figure 4: Distribution of dealer size



The number of dealers also vary across regions, although this difference disappears when one controls for inhabitants.

The preceding descriptive analysis confirms a market for used cars that is concentrated in just a few municipalities, yet somewhat evenly distributed across regions. If the existance of a few large dealer increase the asymmetry, this will only increase the possible benefit from predictive analysis. The listing prices are more evenly distributed, but with clear variation which might be explained by a lack of observations for a few municipalities. Looking at regions the price variation is smaller, but still with distinct differences between regions. Based on the descriptive analysis of the market for used cars in the section, we next turn towards spatial price differences as discussed in the introduction.

4 Analysis of Spatial Price Differences

In the following we will use a linear regression model with fixed effects to examine spatial price differences. That is, we estimate

$$price_{jk} = \alpha_j + \alpha_k + \beta' X + u_{jk} \tag{1}$$

Where α_k contains fixed effects across regions, α_j contains fixed effects across car makes, which allow us to control for brand specific effects (marketing, prestige etc.), and β is vector of car characteristics including weight, horse power, fuel type, mileage, year of registration and a dummy for positive adjectives. u_{jk} is the idiosyncratic error terms and the usual exogeneity assumptions are required for identification. $price_{jk}$ is the actual price in DKK. Although it would be interesting with fixed effect across municipalities, as section 3 discussed, the variation in the data doesn't allow this. Also, it seems reasonable that local markets for used cars span across municipal borders, whereas a regional border is a more appropriate market definition. This is also indicated by the descriptive analysis in section 3.

The above estimation is in many ways naïve, although it does give an indication towards spatial price differences, which is part of the purpose of this paper. Before we proceed, a short discussion of the estimation procedure and possible biases seems appropriate. The data shows the supply of used cars and not equilibrium interactions: the prices gathered resembles a price ceiling on the actual equilibrium price. Likewise, the car

may not be sold. This is by far the largest bias in our dataset. Note, however, that it is still possible to conclude something about the direction of price differences as long as the possible bias in the supply of used cars is consistent across regions. That is, the price reduction should be equal across regions (in relative terms).

Table 2 shows the results from eq. (1). The reference group is an Alfa Romeo from Region Hovedstaden. The coefficients of the make dummies can be seen in table A1 in the appendix. The specification of (1) is based on a general to specific approach, where insignificant variables are gradually removed, as well as an intuitive specification of significant price drivers for used cars. This does not reflect any underlying structural relationships. Instead, the model with the best fit is chosen. In subsequent sections this model will be used as well.

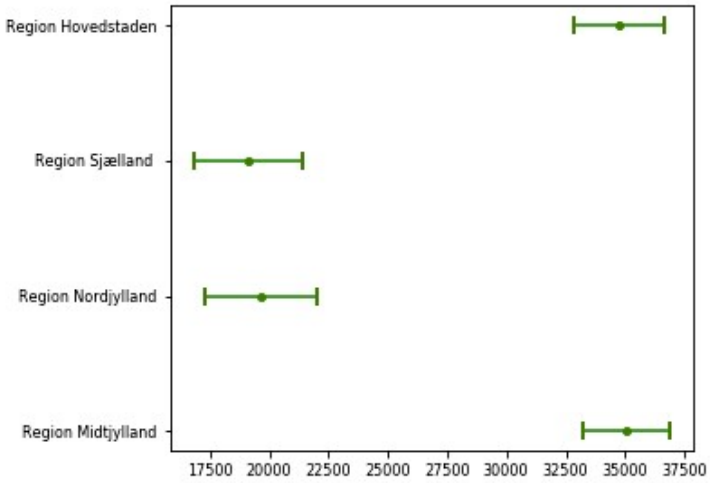
Table 2: OLS estimation in eq. (1)

	Parameters	Std. Errors
Intercept	99,763.57	11,061.09
Region Midtjylland	35,056.11	1,825.12
Region Nordjylland	19,636.98	2,374.54
Region Sjælland	19,095.73	2,294.36
Region Syddanmark	34,748.48	1,918.19
Diesel	9,895.06	1,720.07
Manual transmission	-42,068.49	1,948.02
No ESP	31,228.56	2,359.19
Kilometers driven	-0.50	0.01
Weight	151.93	4.30
# Doors	-9,572.17	986.76
HP	281.49	18.08
Age	-10,298.52	289.33
Dictionary	7,741.42	916.34
Dictionary, squared	-615.13	149.49
R squared	0.45	
Observations	42,592	

Note: reference is an Alfa Romeo from Region Hovedstaden

It is comforting that variables on characteristics of the cars have the expected signs and reasonable coefficients. Diesel cars with automatic transmission and a low millage are expensive, whereas ligther cars with less horsepower are cheaper. In relation to our research question, we see significant price differences across regions in Denmark. This is also plotted in figure 5 below. In particular, used cars turn out to be least expensive in Region Hovedstaden, somewhat more expensive in Region Sjælland and Region Nordjylland and most expensive in Region Syddanmark and Region Midtjylland. This indicates a substantial price difference across regions, and does not resemble well with the hypothesis of one large integrated market. In a world with perfect information and a correct specification on the supply of used cars, this implies arbitrage opportunities across the regions in Denmark. Although it's out of scope for this paper to examine what causes the price differentials, it's puzzling that the difference is that large, which can be evidence towards misspecification and biased parameters. Future work should focus on specifying a model that is better in capturing the features of the used car market, for example by implementing a model, that only predicts positive sales prices (for instance a Tobit model).

Figure 5: Regional price differences from (1)



Finally, listings including words from our list of positive adjectives turn out significant as well. That is, if a listing includes one more word from the list the price rises with 7,500 DKK, which might seem a bit too large. Note though, that this is effect is decreasing in the number of words (i.e. $Dictionary^2 < 0$). This is interesting as a more positive description imply a higher price. This could be a way for the dealers to send a signal about the quality of the car, thereby separating the 'good' cars from the 'bad' cars. However, as we only observe sales prices, a positive wording can also be way for dealers to justify higher sales prices, but with an equilibrium price similar to cars without a positive wording.

With the flaws of our model in mind, we next turn towards prediction. This can be of great value to the consumer, as it enables her to instantly get a sales price for her car, which increases her bargaining power and reduces the informational barrier between buyer and seller in the market for used cars as discussed in section 1.

5 Predicting Used Car Prices

Predicting a price is no easy task. The next section will describe the use of different models for prediction and recognize the challenges of predicting.

The aim of prediction models is to increase the precision of out-of-sample predictions by altering the bias-variance trade-off. In opposition to the standard linear regression model, prediction models are no longer unbiased. Hence, the estimates can't be interpreted as the usual causal relationships. Models are good at predicting outcomes, if they minimize the out of sample errors. This is hard to measure since out of sample observations are not observed per definition. One way to overcome this problem is to split the sample into a training set and a test set. The training set is used to estimate the model, and the test set is used to evaluate the training set. A problem with this method is that it is very sensitive of the split, which is where K-fold Cross Validation comes in handy: it splits the sample into k training and test sets using each training set to calculate the errors from each iteration on the test sets (BBDS page 174).

For this paper, the models are evaluated by the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) as a measure of accuracy. Both measures are indifferent to the direction of the errors and

Table 3: Prediction errors

	OLS	Ridge	Lasso	K Nearest Neighbors	Random Forest
MAE	66286.91	66253.46	66001.45	50302.10	45756.17
RMSE	132878.94	132874.93	132939.63	130207.97	131196.00

a small number can be interpreted as a more precise prediction model. The difference between the two measures is that RMSE imputes higher weight to large errors. Formally,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}, \qquad MAE = \frac{1}{k}\sum_{i=1}^{k}|y_i - \hat{y}_i|$$

In the following OLS, Ridge, Lasso and K-nearest-neighbors are used from the Python package 'SciKit-Learn'(sklearn) to find a prediction of prices on used cars. The models are evaluated on their MAE and RMSE. The results are shown in table 3. In the following each method will be presented.

5.1 OLS

When making a model to predict an outcome or variable there is a tradeoff between in sample fit and out of sample fit. The problem is that the model might overfit or underfit the data sample. If the model overfits the sample, the variance will be too high when new data is introduced. On the other hand, if the model underfits the sample, it may miss some important patterns in the data. Predicting thus becomes a tradeoff between bias and variance. The goal for a good prediction model is to find the point where the tradeoff is balanced.

OLS is very good at in sample predictions, since OLS is built to minimize the squared errors in the sample. However, OLS performs poorly when minimizing the out of sample squared errors, which is key for predictions.

5.2 Ridge

One way to solve the problem of overfitting is to introduce a penalty in the minimization of the OLS.

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \alpha\sum_{j=1}^{p}\beta_j^2$$

The ridge model uses a squared term of the coefficients as a penalty. The Ridge estimation method shrinks the coefficients introducing a bias, but lowering the variance.

5.3 Lasso (Absolute Shrinkage and Selection Operator)

Another model that penalizes the linear regression model is Lasso. The penalty term for the Lasso is the absolute value of the coefficients, which has the shape of a diamond that often gives corner solutions. Lasso estimates a sparse model by iterating over the variables and setting some coefficients to 0. The method is slower than Ridge since the method must search for the solution, whereas Ridge has an analytic solution.

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \alpha\sum_{j=1}^{p}|\beta_j|$$

For both the Ridge and the Lasso it is important to decide to what extent the OLS should be penalized. The respective α values for Ridge and Lasso, that gave the smallest MAE and RMSE were 900 for Lasso and 400 for Ridge. To get a grasp of what the size of α implies, the regression converges OLS as α converges to zero, and the coefficients will become zero if α tends to infinity.

5.4 K Nearest Neighbors

K Nearest Neighbors (KNN) uses the distance to the k nearest neighbors to guide the outcome of an arbitrary new observation. Sklearn uses the Euclidean distance, which is a straight line between two observations to find the nearest neighbors. Through an iterative process 4 was chosen as the neighbors the new observation should depend on, since this resulted in the lowest MAE.

5.5 Random Forest

The Random Forest model uses an ensemble method on different decision trees. In terms of used cars decision trees split the dataset into smaller and smaller pieces as the input is divided by e.g. fuel type. At the end of the decision tree there is a corresponding price. To predict a car price the Random Forest uses all the different decision trees and uses the variety in outcomes to calculate an average or weighted average of the observation. Sklearn uses the average from the outcomes of the different decision trees. A valuable attribute of the Random Forest is that it can find out which variables contributes most to the prediction, thus determining which variables are most important for the prediction (Varian (2014)).

There is a tradeoff between more decision trees and computational power. Here a 1,000 trees is used to estimate Random Forest.

If a variable has much higher values than other variables it might end up dominating the regression. Therefore, normalizing the variables can be practical. For Ridge, Lasso, KNN and Random Forest the X-variables have been normalized to prevent large values from one variable to dominate smaller values from other variables.

5.6 Discussion of Prediction Results

As expected the models, performed better than OLS. The prediction of the models had a mean absolute error between 45,000 and 67,000, where Random Forest was the best predicter. This is a relatively high MAE since the average price of a used car is 184,000.

The RMSE will always give at least as negative a projection of the prediction as the MAE, but for our models it is around twice as high as the MAE for all models. The relatively higher magnitude of the RMSE suggests that the predictions have a problem with large errors. If some of the cars are very hard to predict given significantly isolated specifications, they will pull the models in a certain way that makes predicting harder. The high RMSE values may be an indicator, that the models cannot predict the variety of different cars. This suggests that future work should focus on specifying a model that better captures the heterogeneity in the market.

Another possible reason why the predictions are not that successful may be because too few regressors are included. From the features of importance from the Random Forest model (appendix, table A2), it is clear, that kilometers driven are important in predicting a cars price. Weight, age and horsepowers are also fairly important. The importance of the characteristics indicates, that variables such as engine size and type of engine should have been included, and that there could be an omitted variable bias. Supporting this

is the fact that car prices increases as the number of doors decreases in the Lasso and Ridge predictions (appendix, table A2). A variable stating whether the car is a sportscar (with few doors) or a family car (with many doors) could be a solution to this puzzle. Note however, that the prediction results merely constitutes correlations and causal relationships. Further research could look at a smaller sample of cars or try to include variables that were not available on Bilbasen.dk. This does, however, require a lot of manual work, escpecially in categorizing car models.

Even if our prediction was correct it is important to stress that viewing this as the correct predictor for all cars would be committing big data hybris (Lazer et al (2014) and Lazer and Radford (2017)). There may be a different set of characteristics for the cars that are for sale at another online platform. This selection bias is extremely plausible and should be examined in depth.

6 Conclusion

In this paper we set out to examine the market for used cars. This is a market of great importance for the consumers, but also for policy makers. During the recent week (21th-28th August 2017), policymakers have both discussed taxation on the usage of cars[2] and taxation on new cars[3]. This stresses the need for research about the market for used cars, both on the demand and supply side.

In this paper we use scraped data from Bilbasen.dk, the leading online platform for used cars sales. Our dataset consists of 42,592 observations, all of which were active on the 18th of August 2017. Our data on listings is solely based on information from Bilbasen.dk. Apart from basic car characteristics, we also create a dictionary of non-neutral words not related to car characteristics from the description on each listings.

Our descriptive analysis show a difference in the number of listings between the municipalities in Denmark that is not explained by the number of inhabitants. From a graphical inspection, there seem to be clusters in the market for used cars where some municipalities have a lot of listings and the surrounding have few. However, the distribution of sales prices across municipalities does not show the same pattern.

When examining the market for used cars this paper also looks at spatial price differences. Our analysis show somewhat large differences between the regions of Denmark when one control for observable characteristics with a fixed efffects regression. Further research should explore this, as it is an indication of arbitrage opportunities across regions.

Lastly, we perform a prediction analysis based on the regression analysis. Predicting the value of ones car can be of great value to the consumer, as it lowers the informational barrier in the market. Also, the value of used cars (and the distribution) is of great importance for policymakers when evaluating the effects of various policy proposals. The predictions shows, as expected, that OLS is the worst predictor among the set of predictors. Although, the predictive power of our model is not very strong, our work have laid a foundation for future research.

[2]Roadpricing on Storebæltsbroen
[3]Registreringsafgift (value tax)

References

[1] Akerlof, George (1970), 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism', *The Quarterly Journal of Economics,* Vol. 84, No. 3 (Aug., 1970), pp. 488-500

[2] Berrry, Steven; Levinsohn, James and Pakes, Ariel (1995), 'Automobile Prices in Market Equilibrium', *Econometrica,* Vol. 63, pp. 841-890

[3] Foster, Ian; Rayid Ghani Ron S. Jarmin, Frauke Kreuter, Julia Lane, *Big Data and Social Science: A Practical Guide to Methods and Tools*, CRS Press 2016

[4] Gavazza, Alessandro; Lisseri; Alessandro and Roketskiy, Nikita (2014), 'A Quantitative Analysis of the Used-Car Market, *American Economic Review,* 104(11): pp. 3668-3700.

[5] Grimmer, Justin, and Brandon M. Stewart. 2013. 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political Analysis*, 21.3: 267-297

[6] Lazer, David, et al. (2014), 'The parable of Google Flu: traps in big data analysis', *Science,* 343.14.

[7] Lazer, David and Jason Radford (2017), 'Data ex Machina. Introduction to Big Data', *Annual Review of Sociology* vol 43, August

[8] Varian, Hal R. (2014), 'Big Data: New Tricks for Econometrics', *Journal of Economic Perspectives*, 28.2: 3-27.

7 Appendix

Table A1: OLS estimation in (1)

	Parameters	Std. Errors
Audi	35,792.54	9,258.02
BMW	27,285.73	9,336.47
Chevrolet	-50,467.80	11,122.03
Citroën	-51,543.86	9,416.51
DS	-69,788.65	14,487.53
Dacia	-63,923.79	14,593.71
Ferrari	-8,056.36	21,591.56
Fiat	-38,880.56	9,828.31
Ford	-20,386.91	9,256.55
Honda	-13,303.02	10,574.15
Hyundai	-47,166.99	9,519.71
Jaguar	32,784.56	14,771.80
Kia	-36,637.62	9,645.89
Land Rover	-31,124.64	13,984.56
Maserati	-157,348.52	20,236.18
Mazda	-3,143.70	9,814.61
Mercedes	52,610.51	9,353.44
Mini	14,730.40	12,700.82
Mitsubishi	-34,449.05	11,604.16
Nissan	4,013.46	9,639.97
Opel	-33,659.76	9,390.85
Peugeot	-39,129.34	9,271.96
Porsche	-10,569.17	11,176.30
Renault	-38,425.89	9,431.46
Saab	-18,230.93	14,502.35
Seat	-24,773.83	10,077.73
Skoda	-14,400.70	9,471.72
Smart	-32,089.12	17,478.54
Subaru	52,496.89	18,390.33
Suzuki	-31,173.51	9,807.00
Toyota	-10,743.24	9,445.85
VW	-2,698.36	9,179.08
Volvo	37,452.19	9,763.05

Note: reference is an Alfa Romeo from Region Hovedstaden

Table 4: Appendix: Prediction

Table A2: OLS, Lasso and Ridge coefficients. Random Forest prediction Random Forest feature importance

	Ridge	Lasso	Random Forest
Kilometers driven	-648.34	-590.60	0.27
Weight	359.47	0.00	0.16
Number of doors	-162.33	-43.29	0.02
Horsepowers	543.42	141.48	0.09
Age	-661.78	-200.06	0.10
Dictionary	-542.06	-35.49	0.05
Dictionary squared	1332.98	611.43	0.05
Region Midtjylland	-1078.84	-437.62	0.03
Region Nordjylland	698.49	357.70	0.02
Region Sjælland	-531.98	-123.67	0.02
Region Syddanmark	-376.68	0.00	0.03
Diesel	-1071.39	-184.76	0.01
Manual transmission	119.64	0.00	0.01
No ESP	-771.86	-466.16	0.01
Audi	226.53	27.00	0.01
BMW	490.55	5.73	0.01
Chevrolet	80.41	2.46	0.00
Citroën	445.11	87.17	0.01
DS	268.30	0.00	0.00
Dacia	-365.09	-30.30	0.00
Ferrari	914.55	620.67	0.00
Fiat	-505.52	-21.98	0.00
Ford	-542.26	-291.89	0.01
Honda	-951.32	-513.39	0.00
Hyundai	-567.47	-133.67	0.01
Jaguar	913.73	455.63	0.00
Kia	701.10	311.84	0.01
Land Rover	-1038.44	-523.13	0.00
Maserati	-694.87	-94.71	0.00
Mazda	-254.81	0.00	0.00
Mercedes	-488.30	-124.14	0.01
Mini	104.46	0.00	0.00
Mitsubishi	-919.29	-348.25	0.00
Nissan	218.99	20.52	0.01
Opel	256.35	0.00	0.01
Peugeot	1546.27	1075.40	0.01
Porsche	-70.32	0.00	0.00
Renault	280.98	0.00	0.01
Saab	951.14	455.90	0.00
Seat	1256.24	810.14	0.00
Skoda	-588.24	-357.18	0.01
Smart	-505.51	-229.19	0.00
Subaru	-724.75	-120.41	0.00
Suzuki	-914.48	-367.84	0.00
Toyota	-358.42	-56.82	0.01
VW	-32.92	0.00	0.01
Volvo	-96.26	0.00	0.01

Note: reference is an Alfa Romeo from Region Hovedstaden

Figure A1: Distribution of sales prices across regions