NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Data-driven Evaluation of Real Estate Liquidity

Predicting days on market to optimize the sales strategy of a startup

Maria Lubomirova Dobreva

Project report presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DATA-DRIVEN EVALUATION OF REAL ESTATE LIQUIDITY

Predicting days on market to optimize the sales strategy of a startup

by

Maria Lubomirova Dobreva

Project report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information Systems and Technologies Management

Advisor / Co Advisor: Professor Roberto Henriques

Co Advisor*:* Mauro Castelli

August 2019

ABSTRACT

This is a research project for applying data mining techniques on Real Estate data in cooperation with Homeheed, a startup in the area of real estate, providing a platform solution as a single source of truth in Sofia, Bulgaria. This project suggests the development of a predictive model by using LASSO regression with the premise to determine days on market. As a consequence, the discoveries are expected to contribute to the Startup by providing insights about more attractive listings, and so will support faster return on investment. Additionally, the paper provides an experimental part where misleading and fake listings are targeted in order to support fraud and real availability of a listing detection. The project's main objectives and assumptions are that advanced statistics and information management can build such a synergy with data and business models that allows enhancement of both market entry strategy and quality of service.

KEYWORDS

Real Estate, Predictive model, Data mining, Days on market, LASSO regression

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
CRISP- DM	Cross-industry standard process for data mining
DOM	Day on market
DT	Decision tree
LR	Linear regression
MAE	Mean absolute error
MSE	Mean squared error
POI	Point of interest
R2	R squared
RMSE	Root mean squared error
ROI	Return on investment
SEMMA	Sample, Explore, Modify, Model, Assess

1. INTRODUCTION

The real estate market in Eastern Europe and former Soviet Union countries is emerging. In Bulgaria, the situation does not differ. On the ground of the country's political and economic situation, the development of Bulgarian property market can be presented in three main stage – during the socialism, the transition to a market economy, and the current internationally attractive market. The last stage is the period when the real estate market registered a double-digit annual growth due to the international investment interest. Later, between 2003-2008, the sector was blooming which led to the creation of a price balloon formed by 40% drop in the housing prices. After this crisis, property investments have register again a gradual increase. Statistics show that the housing sales increased with 11.5% for the first quarter of 2018 and the interest rates remained at their low levels. Together with that, the planned new building constructions reveal the growth of 6.3% (Stoykova, 2018).

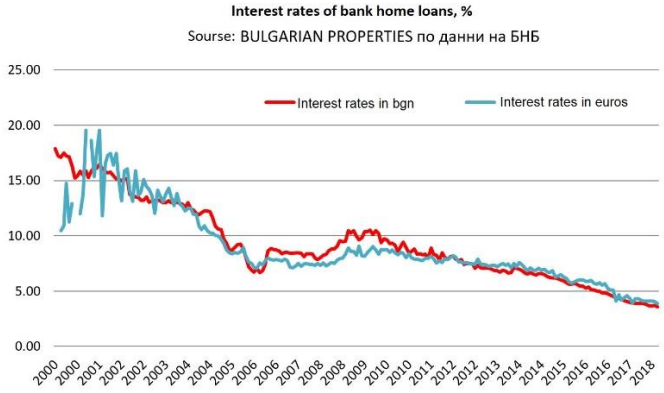

Figure 1 – Loans Interest Rates Bulgaria (Stoykova, 2018)

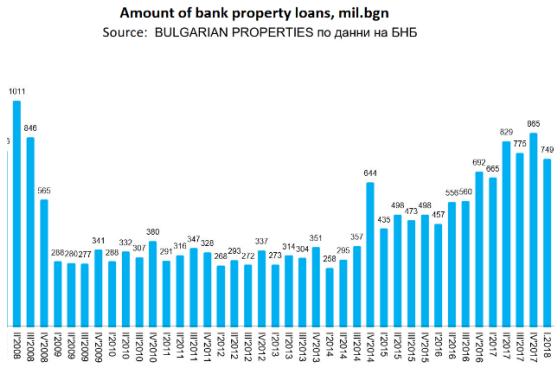

Figure 2 – Amount of home loans (Stoykova, 2018)

All these fluctuations in the market and factors imply liberalization and lack of regularities which lead currently to the easy entrance and exit of the market for brokers who compete for customers. The market is not exclusive and a single property can be offered on the market several times, in different sources and by a variety of brokers. Often brokers keep outdated or unreal, but attractive listings online in order to increase the chance to acquire new customers. This usually creates wrong expectations and bad customer experience.

Homeheed is a Bulgarian Start Up, which tries to solve this issue by centralizing the redundant listings in one single platform. In technical terms, the company uses Key points matching technique to identify duplicates of a listing by image recognition and then summarizes the listings in one central unit. Currently, one apartment can be found online listed by different brokers and/or with changes in the description. This results in difficulties to extract a unique identification key for duplicated listings. Homeheed found out that images remain the only part of a listing offer by which one and the same apartment can be tracked.

The value proposition of this process is to act as a single point of truth and to enable the customer to see all listings of a property, as well as to understand whether it is available or not. Homeheed entered the market recently with a first prototype to validate the idea and the demand. The team provides the potential customers with a demo version of the platform where the listings are filtered from fake offerings and only properties matching the individual preferences are received by email.

Homeheed collected information about the property market and listings from 2015 to 2018. The Startup aims to analyze this data in order to optimize its market entry program and to forecast return on investment (ROI). This work will apply data mining techniques, based on this historical information, in order to forecast how many days a listed property with specific characteristics will be online. This will help Homeheed to provide customers firstly with the most attractive offers and so to optimize the revenue stream.

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

The topic about the irregularities and frauds in the real estate market has been raising heated debates around the Bulgarian media channels. The market does not imply rigorous laws and rules, which leads to the easy entrance of real estate agencies. Some agencies frequently publish unreal or unavailable apartment listings, often on a price below the average for the market, as a method to acquire customers looking for a new living property. Customers either never see the desired place, or are even misled with fraud schemes for advanced payment before the deal. Not only creates this bad customer experience and non-satisfaction, but also makes the process of finding a living property challenging and time-consuming. This instabilities and misappropriations in the property sector necessitate the development of a more transparent platform, as Homeheed aims to be, and the establishment of better methods for assessing homes availability (Vasilev, n.d.) (BTV, 2012).

In the core of Homeheed's value proposition is the transparency of listings authenticity and availability. The Start-up goal is to provide a solution which will support the process of fixing the market irregularities, as well as will lead to better customer experience. Currently, Homeheed team works on the technology which can identify unreal listings, but there are still some limitations. Building a model which can predict with certain accuracy how many days a published listing will be online, can contribute to the technology of Homeheed and support their market entry strategy. Additionally, the first prototype of the system can be optimized by prioritizing a home listing by its attractiveness and respectively certain probability this listing to be rent. This will support the process of developing a brand identity, market growth and customer satisfaction of the platform and the service.

The technical aspect of the project work contains also several issues which should be identified and mentioned beforehand. The research of the market and interviews with real estate agents revealed that when a listing has been published by owner, the probability that a customer will face frauds and misappropriations is really low. The problem appears when a property is listed by agencies since the track of its authenticity and availability based on the provided historical data is challenging[1]. To solve this issue, several assumptions have to be made and some procedures prior to the modelling part have to be defined, both later explained in detail.

1.2. STUDY OBJECTIVES

The following paper aims to present a systematical approach on data analysis techniques, in particular, predictive modelling, applied for real estate market issue in favor of Homeheed market entry strategy, technology optimization, and value proposition. The core study objectives of this project are:

(1) Predicting days-on-market for housing

(2) Identifying features which make a property more attractive

(3) Contribute to the Homeheed's technology identifying homes real availability and supporting fraud detection

Given the complexity of the problem, regarding the first question should be mention that attaining a highly accurate model which can predict how long a given property will be on the market is a compound task. One of the main reasons is that data set which contains all required information is not available currently and in general is difficult to collect due to the high amount of not quantitatively measurable factors. Secondly, days-on-the-market is a variable highly influenced by a variety of dynamics, dependencies and features such as location, price, details regarding the condition of an apartment, to name but three. Despite these challenges, the project and the analysis will be executed under several assumptions, which are explained in more detail in the Methodology section of this paper. With the limited data, the goal of this project is to assess days-on-the-market to support the process of Homeheed in building a technology which can prevent property frauds by identifying the real availability of a home listing.

The answer of the second question is closely related to the findings of the first one. Different studies which focus is on predicting housing prices identify and measure the effect of common housing

[1] The collection of this information is currently in process together with different agencies and institutions.

attributes on the price. Here the point of interest is to measure the effect of such features on days-on-the-market and identify what makes an apartment more attractive to a customer. The answer to this question will support the product development of Homeheed plus will allow the team to provide customers with listings with a higher probability of being sold/rent. As a result, the ROI will be optimized and the revenue stream for Homeheed will be enhanced.

The third objective is part of the Startup value proposition. For the purpose of the project, it is interesting to investigate whether the data contains signals which can support the detection of unreal listings. If misleading and fake listing properties are filtered by Homeheed's technology, the value for the customer will increase and the platform will be able to propose better user experience in the process of searching a home.

2. STUDY RELEVANCE AND IMPORTANCE

In respect to the mentioned in the previous chapter real estate market challenges in Bulgaria, this project will allow Homeheed to (i) benefit from its own data by exploring historical market data and gaining valuable insights which will allow a more accurate estimation of the listings, (ii) streamline its market entry program, significant for the revenue stream and ROI planning, and (iii) further support the design of the technology which can assess a property availability. The outcome of the project will help to determine important housing attributes and so will serve as a proposal for restructuring of the database by introducing new features for future data mining projects in Homeheed.

Further, the work aims to contribute to a platform which serves as a tool to achieve more fair competition on the Bulgarian unregulated real estate market. Additionally, it is assumed that the findings can enhance the business model, the technology, and the market entry strategy. Data analysis techniques can influence positively the development of the system and enhance it by making it more sustainable, efficient and transparent, as well as by improving customer satisfaction and general citizens experience in the process of searching for a new home.

Regarding the subject of the study, should be mentioned that in the area of data science application in real estate sector problems, multiple studies about housing price prediction have been found. In different periods when the real estate market worldwide has recorded changes, bloom or descent, questions regarding the accuracy of property value assessment have been raised. The instabilities made housing predictive models subject of research among scholars. Literature review shows methods which can estimate the price of a property based on different features and in comparison to similar objects. However, the question of how long a listing will be on the market is not so extensively studied problem. This work aims not only to be a complementary project for the

business development of Homeheed but also to contribute to the study area with the concept that days-on-market as a feature has significance in terms of investment and ROI planning.

3. RESEARCH METHODOLOGY

To select the most relevant publications for this project, the focus was set on finding scientific papers and literature about the application of data science in real estate area, in particular the development of predictive models based on housing data. In-depth understanding of factors which cause fluctuations in the different variables of the similar dataset, as well as further investigation on the real estate sector in Bulgaria, will allow proper application of assumptions and data analysis approaches. Additionally, a literature review on data mining topics and theory is an essential part of the following project in order to provide a better explanation of the selected approach and chosen later predictive model.

The accuracy assurance of this project requires a set of qualitative methods to be executed. To fulfil the objectives of the project, the stakeholders should be interviewed. In this case, these are the founders of Homeheed, as the party providing the rights for the data, real estate agents, as the directly influenced by the marker irregularities; professor Roberto Henriques, as adviser; and professor Mauro Castelli as co-advisor for this Master work and representatives of NOVA IMS.

The technical requirements of the project will be fulfilled by packages of R language as a tool which provides a simple collection of functions created especially for statistical computing and data analysis. R will be used as the main tool to perform all the visualizations, transformations and modelling tasks required to execute the project. R language has good graphical capabilities which will support the exploration of the data and the decision-making regarding the data preprocessing phase. Additionally, R is specifically created for statistical tasks and data manipulation and so provides great list of libraries which support data analysis as well as the training and evaluation of models.

The project will follow the SEMMA methodology for the organization of the work. It consists of sequential execution of phases allowing the understanding, preparation, modifying, modelling and assessment of the data. The first and foremost step is the data preparation which includes understanding, visualization and cleansing. This process is the most time-consuming and essential processes in every data mining project. Having been determined, the anomalies and outliers in the data set can be further filtered and/or transformed in order to make the data suitable for model development. The main objective is not only to create an algorithm to fit the data observations but to develop such a model which will predict with as great as possible probability the future behavior of the target data points. However, the nature of the project requires approaches considered by the CRISP-DM framework. Before applying data mining procedures, a clear understanding of the

business, market and data is required. For that reason, the following figure illustrates better the procedure applied at this work.

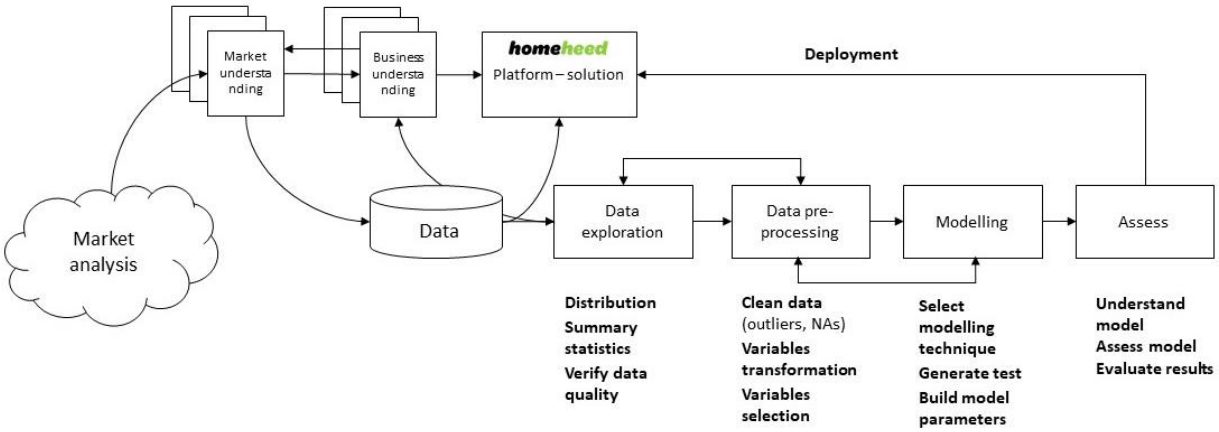

Figure 3 – Work methodology

In this project, several predictive models will be developed and compared by statistical measures and graphical illustrations in order to choose the best option which can predict the target – *days-on-the-market*. Referring to the problems identified previously in the chapters, it should be mentioned that in order to execute the project precisely, several assumptions will be made: (1) the dataset will be divided by owner and others, and (2) for the dataset subset by agency additional research should be executed to determine the authenticity of the agencies which published a particular listing. Additionally, the data set contains information not only about homes but also restaurants, places, lands, garages for sale or rent. In this project, the main focus will be on homes since this is the point of interest (POI) for the first prototype of Homeheed.

For the purpose of the project a data set provided by Homeheed will be used. The data set consists of 19 variables characterizing listings, contains more than 550 000 observation points and represents the time period from 01.07.2015 to 01.07.2018. The data set will be presented in detail in the chapter Methodology.

For the purpose of this project, a work structure is additionally created and provided. The figure below shows a sequential diagram of steps which will be executed and further presented in the report. The first two parts of the work have the purpose to build the theoretical context needed for the appropriate execution of the project. The main part contains the data mining procedures which will provide the answers to the research questions by developing a predictive model. Further, the model will be tested and evaluated.

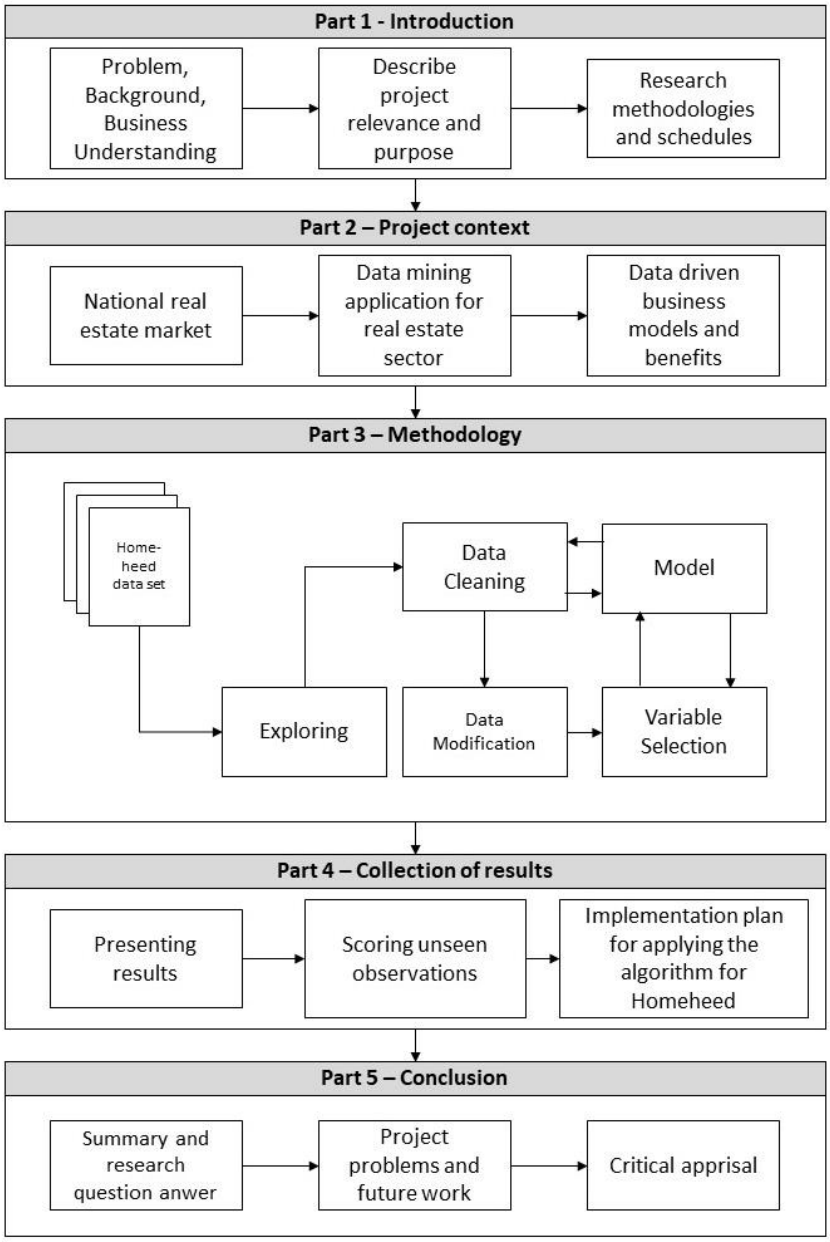

Figure 4 – Work Structure

4. LITERATURE REVIEW

For the proper execution of the project the literature review will not be limited related works and publications, but part of the scope will be also theoretical literature covering the topics included in the execution phase of this work. This chapter will first focus on the presentations of publications based on the application of data analysis techniques for real estate problems and predictive analytics approaches for similar questions. Afterwards, the importance of data-driven, rather than data-informed business solution, in particular for the scope of start-up development, will be mentioned. It is interesting to research whether, and if so at which point start-ups collect and analyse data to verify their ideas, to assess the prototypes and the market, and to forecast ROI. In the last part of this chapter, a review of the main data mining concepts included in the work will be presented.

4.1. REFERENCE WORKS FOR REAL ESTATE PREDICTIONS

The application of data mining in real estate has become widely popular. Researchers and companies use a variety of prediction techniques to capture fluctuation periods and the factors influencing them, to analyse the market trend through regression and machine learning algorithms, to describe property types by clustering heterogeneous housing data, including house attributes and geo-social information, and to find customer habits to determine sales strategies (Xian Guang LI, 2006).

As mentioned earlier in the Research Methodology chapter, data mining is a process which follows mainly an interactive and iterative procedure constructed of several steps - problem definition combined with useful data sample; preprocessing of the data, requiring its coherence check and proper manipulation; choosing appropriate method for data modelling; results assessment and final knowledge implementation. Regarding prediction in the real estate sector, the procedure will be the same, but a proper collection of data and market knowledge, as well as proper estimation of the different factors influence on the problem, should be also carefully considered.

Many studies analyse the real estate prices while analysing the days on market (DOM) and the popularity of a property is still understudied area. DOM is an essential factor, although challenging to measure, for real estate listing since it is highly correlated with the popularity of a housing object. The literature review showed that some publications are focused on studying the relationship between DOM (or time on the market) and different factors, such as prices, brokers/ broker agencies, marketing strategy, and others (Jud, 1996) (Eddie C.M. Hui, 2010). The results show contradictory findings. For example, Belkin, Hempel, and McLeavey (1976) suggest that DOM and

sale price of a housing object have no relationship, while Miller (1978) uses DOM to explain sales prices and show a positive effect. Other studies illustrate that DOM and sale price has a associated connection due to various factors such as quality, listing strategy, and real estate agency, which adds complexity to the relationship (Catherine Tucker, 2013). In this thesis, the focus is not on describing the existence of such relationship, and so the referred papers will not be reviewed in detail, but on providing an accurate prediction of DOM in order to support Homeheed development. Some paper works regarding the studied problem have been found and considered as a reference for this work.

Zhu et al (2016) present a study in which the authors measure the liquidity of the real estate market by developing an approach for predicting DOM. The authors use Multi-task learning based regression to overcome to the problem of location dependency and further compare the results by using baseline models such as Linear Regression (LR), Lasso, Location-specific Linear Regression, Decision Tree (DT), and others. Their result illustrates also the importance of different features in the model. The performance of the method is presented with real-world real estate data and a designed prototype of a system showing the practical use of their analysis, which can be used as a reference for Homeheed software (Hengshu Zhu, 2016).

Sergey V. Ermolin (2016) uses a Decision Tree algorithm to predict DOM within 7 days. In this case, Linear regression was not applicable due to the lack of linear correlation between numerical features and the target. The author makes the assumption that any accuracy for more than a week should be considered arbitrary due to the seasonality of the housing market. In Ermolin's work it was concluded that geospatial features did not add value to the prediction (Ermolin, 2016).

Mou et al (2018) propose a system which can predict short DOM from which can benefit both buyers and sellers. This work provides a framework which can serve as a reference to estimate the market value of a housing property. The authors make the assumption that true market value can be approximated to the listing price when real estate agents have similar offers because few brokers would be willing to sell a property on a much lower price. Further, housing with short DOM is detected by comparing their listing prices and estimated market values (Chao Mou, 2018).

4.2. DATA-DRIVEN BUSINESS MODELLING

This paper could be seen as a work which illustrates how data is changing the traditional business paradigm. Data is increasingly becoming a dominant factor not only in high-tech and profitable companies but also in more low-tech sectors. Many industries shift in direction of data-driven strategies. Social, economic, and technological changes have made possible the implementation of technologies into business models, to which the availability of data gave a limitless flood of new

products and services that in sequence affect society and business. However, to be competitive does not mean to generate as much as possible data, but to collect and extract data in a way which can generate competitive advantage (Spijker, 2014).

Regarding the business development and the market entry strategy of a start-up, such as Homeheed is, should be mentioned the importance of data analytics for these both. If great companies such as Amazon, Google, Facebook, etc. are taken into consideration, the pattern of them all, using data in a new, intelligent way will be captured. Croll and Yoskovitz (2018) state that most of the start-up founders have a fundamental doubt of developing their businesses only on numbers and figures. An essential part of the business model design in their intuition, incorporated in the company's soul, and market vision, projected into the business model. Such quantitative measures are useful and crucial for hypotheses test, but limited and inferior for creating new ones (Yoskovitz, 2018). Clearly, in business, especially in the stage of its development, data analytics and future predictions not only can assist with planning and strategy, but also can change the whole direction and model of a business. Insights gained from data mining can guide decision-making, business experiments and prototype design, at which stage Homeheed is, and so to make them more effective (Marr, 2017). The analysis of the provided data itself potentially will help Homeheed to reveal relationships between different factors and so the generating of different scenarios for alternative outcomes.

Referring to the market entry strategy, data analytics, in this case predictions about the future, can enhance business process, customer offering, and marketing. Data is considered extremely useful for detection of misappropriation and fraudulent activities, such as Homeheed has already identified for the market. The analysis of the data helps to gather patterns of certain activity/ behaviour and to recognize it as an act of deception/ scam. In case of achieving good model accuracy, the information can be used to be integrated into the technology and so to improve the rate of highlighted cases (house availability). Data (in this case modelling) can support not only business operations optimization, but also the delivery of better service (Marr, 2017). In the particular case, a whole new customer value proposition might be discovered, e.g. providing customers with the most attractive offers which can increase the probability of taking an action.

4.3. DATA MINING THEORETICAL CONTEXT

Data mining is the process of extracting and transforming data into knowledge based on discovered patterns, valuable insights and structures, complex and significant relationships between features. The techniques offered by data mining serve as an in-depth analysis of unexpected relationships not

related to the original purpose of the data collection. In general, data mining tasks can be grouped by descriptive and predictive.

The main focus of this work is the predictive functionalities of data mining. Predictive modelling is a data mining technique which make possible to discover relationships between explanatory variables and a target one. The target variable in this work is DOM for real estate related data. This implies several assumptions about predictive modelling techniques which can be trained.

4.3.1. FEATURE SELECTION AND REDUCTION METRICS

A fundamental issue in machine learning is the mapping between input and output data based on the recollection of observation points. Usually, the output is not perfectly determined by all of the input collection which makes some features irrelevant. In practice, using irrelevant or redundant features in a model, imply several problems.

Including variables with little or no effect results in high computational cost and significant time for training the algorithm. The greater issue is overfitting because the model will not be able to work correctly on new data. And to underestimate that the right subset of data features ameliorates the process of interpreting the model (Deng, n.d.). Additionally, variables selection facilitates data visualization and understanding, cuts the dimension and storage necessities, and solves the curse of dimensionality issue, which all support the prediction accuracy (Guyon & Elisseeff, 2003). In this chapter, three main practices to subset features and optimize space will be presented.

At the pre-processing step of the data, the first method which can be applied to reduce dimensionality is filtering. This method requires the usage of statistical tests which can examine whether there is a specific relationship between variables. For numerical variables, one of the steps is to check for correlation by using Pearson and Spearman coefficients in order to explore the existence of a linear or monotonic relationship between variables. When two variables are highly or perfectly correlated, one of them can be eliminated since it does not add any additional information. For categorical variables, one filtering method is the analysis of variance in case of examining whether the means of two or more groups are equal or not. Such analysis of variance is performed by the Kruskal-Wallis Test, a non-parametric test which measures for significant dissimilarities on continuous target variable by a categorical independent one. Kruskal-Wallis test does not require the assumptions which ANOVA requires, namely normal distribution and equal variance between groups, which in this case is not observed (Ostertagova, Ostertag, & Kovac, 2014).

The second approach is covering wrapping techniques. These include training an algorithm and exploring the effect of the variables in the model. Such feature selection algorithm is forward selection – an iterative procedure which adds variables, starting from zero, in each iteration until the point when an additional feature does not add value. An alternative procedure is backward elimination which is the opposite of the forward selection – all variables are included in the algorithm and iteratively variables which do not add value to the performance of the algorithm are removed one by one.

The last feature selection method revealed in this paper is embedded technique, which includes algorithms which have pre-build feature selection procedures. Such an algorithm is LASSO regression, which has the goal to minimize the prediction error and so to enhance the accuracy of a model (D.R. Cox, 1972).

Apart from feature selection, there are other additional techniques which can support the dimensionality reduction task. For instance, a significant percentage of missing values in a variable could be a reason for its exclusion from the data set, since any imputation technique could cause biased results. In the chapter covering the pre-processing of the data, all these steps will be covered and additionally explained with examples based on the project case.

4.3.2. TEXT PRE-PROCESSING

This project requires to cover some basic techniques regarding text mining, because one of the variables, which exploration will be considered in this work, includes text and supposes some alterations. Text mining is a whole area which has an emerging role in different applications such as information retrieval, natural language processing, information extraction, categorization and others. In this chapter, only basic concepts and procedures will be covered to give the reader an overview of the steps executed in the project regarding the text variable. To make data more effective and text more suitable for analysis, string format variables should go through several transformations.

The first step of the process is cleaning and removal of any noise which can make it hard to extract insights from the data. Before transforming text, punctuation, stop-words, and expressions should be cleaned. Textual data may contain many symbols or characters which do not add any valuable information for the analysis and should be dealt with. In general, text requires lookups for apostrophes, slangs and attached words, as well as HTML characters or URLs. All these are unnecessary and meaningless text characteristics which will make difficult to understand the data and to later build a model with proper features.

Tokenization of the data is the next step which is essential when dealing with character variables. This is the process of breaking text into single words or other meaningful elements which enhance the exploration of a sentence (Kannan & Gurusamy, 2014). Stemming is also a crucial step which helps to identify the root of a word so all version of the same word can be stemmed in a single one. The goal of this method is to eliminate numerous suffixes, to decrease the number of words, to have precisely matching stems, and to reduce the time for processing and memory space needed for the data (Vijayarani, Ilamathi, & Nithya, n.d.).

After the cleaning step, the text should undergo a process supporting feature extraction. N-grams, which is a sequence of items, in this case words, captures sentences structure and provide some additional information about the sentiment in a text structure. For example, such analysis can show that two or more words occur always together and should be extracted as one whole or only one of them can be considered (Aase, 2011).

Creating tokens by n-gram is a useful way to search pairs of co-occurring words. However, co-occurring is not highly meaningful alone since the words of n-gram could be also the most common separately. Examine correlation among words, which designates how frequently they occur together compared to how often they appear unconnectedly, by using the phi-coefficient which is measured for the association of binary variables (Robinson & Silge, 2017).

Cleaning, tokenization, n-grams and stemming are only some of the topics regarding text mining. The aim of this chapter is only to cover the basic steps executed in this project, because the work does not require more sophisticated analysis on the textual variable since it provides only keywords description, but not a complete one which can be organized and analysed in depth.

4.4. Machine Learning Algorithms

In this section, a literature review of the theoretical concepts for the models used in this work will be presented. The chapter covers main theory about the algorithms which will be trained, both for model building and feature selection, in order to provide the reader with a bit more detailed overview on what has been done in this project work. Survival analysis, multiple linear regression, and some regularization models, and the neural network will be discussed.

4.4.1. Multiple Regression and Regularization Models

Linear models are mainly statistical technique for prediction or evaluating the relationship between dependent and independent variables. Simple linear regression is a process which allows making

predictions about target variables based on the information about another variable. Most problems are too sophisticated to be modelled only with two variables. Multiple regression is a statistical tool which includes several independent variables and their relation to a dependent one. The equation of the regression includes the intercept, which is the point where the regression crosses the Y-axis, and coefficients representing the slopes of the regression in direction to the X-axis. This equation is only efficient and useful in measuring linear relationships between variables. If the data form a non-linear shape regression analysis would not be able to discover an association. By plotting each independent variable with the target one, observing for curves, distant points, alterations in the quantity of variability, and different other irregularities that may occur.

There are a set of assumptions which should be considered when applying multiple regression analysis. Firstly, the relationship between the variables should be a straight line. Additionally, the analysis requires constant variance between variables and deletion of outliers. Normality of the data and scale optimization are also part of the underlying motivations for applying regression analysis. However, multiple regression could be applied also as variables selection method by analyzing the added value for the performance of the model of every additionally added variable (NCSS Statistical Software, n.d.).

Apart from multiple regression, there are other types of regression analysis which perform better in the case of multivariate data samples. These models are known as penalized for having too many features and a linear model is built by including a limitation in the equation. This is called also shrinkage or regularization. The regularization reduces the values of the coefficients towards zero, which allows variables which do not add or add less value to the model to have coefficients equal or close to zero (Gareth, Witten, Hastie, & Tibshiran, 2014). The most widely used models which impose penalties are lasso, ridge, and elastic net regression.

Ridge regression equation allows the decreasing of coefficients and as a consequence, the variables which have an insignificant impact on the outcome have coefficients close to zero. The penalty term which supports the shrinkage is known as L2-norm – the sum of the squared coefficients (Gareth, Witten, Hastie, & Tibshiran, 2014). The regularization term includes also a constant lambda for which the value selection is of high importance. If the value is too high, the impact of shrinkage penalty increases and this will lead to under-fitting (Gareth, Witten, Hastie, & Tibshiran, 2014).

Ridge regression performs better compared ordinary least square regression in case of a large number of variables, but it will include all predictor in the final model, which is considered a

drawback. Additionally, this method changes the coefficients towards zero, but it does not adjust them exactly to zero (Gareth, Witten, Hastie, & Tibshiran, 2014).

The lasso regression is a model which overcomes this disadvantage. Lasso stands for Least Absolute Shrinkage and Selection Operator. The difference with the ridge regression is the penalty term used, namely known as L1-norm, which adds the sum of the absolute value of the coefficients (Fonti, 2017). Variables which do not have a contribution to the model have coefficients equal to zero. For this reason, lasso regression is an efficient tool for variables selection and complexity reduction. The smaller number of predictor makes the interpretation of the model easier (Gareth, Witten, Hastie, & Tibshiran, 2014). Elastic net regression uses both L1 and L2 norm penalty term and so does both - decreases coefficients effectively and set some of them to zero.

4.4.2. SURVIVAL ANALYSIS

In respect to the problem studied in the paper, the concepts of survival analysis are interesting to be covered as well. The main elements of survival analysis and its relation to the problem studied in this work are presented. Survival analysis is mostly described as a set of methods for studying data where the product variable is the time until the occurrence of an event of interest. The survival time (time to the event) can be measured in days, weeks, years, etc. (Despa, n.d.).

Survival regression analysis resembles linear and logistic regression, because (1) typically there is one dependent variable and one or several predictors; (2) examination of confounding variables is essential; (3) and model selection and estimation is a key requirement. However, survival analysis has several advantages over these standard statistical techniques. A fundamental attribute of survival data is that the target variable is a positive discrete or continuous one and represents the time interval from a well-defined starting point to the exact event. Secondly, the survival analysis can accommodate and handle censoring in the observation. Censoring occurs when start or end events are not observed for some observation points during the study. Three main types of censoring could be described – right, left and interval. Right censoring occurs when the end event exceeds a particular value. Left censoring is observed when an observation point enters a study after the endpoint. Interval censoring characterizes survival time of an event of interest which occurs between two studies (Moore, 2016) (Vermeylen, 2005). The neglect of censoring can lead to a biased sample and incorrect conclusions (Brian D. Kluger, 1990).

Survival analysis methods use well information from both censored and uncensored data to estimate dependent variable which is composed by time to an event or event status. To describe the survival distribution two main functions are defined – the survival and the hazard function. The survival

function gives the probability of an event to occur up to a given time, and the hazard function gives the potential of event to occur when an individual has survived up to the defined time (Vermeylen, 2005).

A well-known and widely used survival regression model is the Cox regression or Proportional Hazard Model. The purpose of this model is to estimate at the same time the influence of several factors on the time to an event. This is known as the hazard rate. The Cox regression is usually expressed with a hazard function (the chance event to occur at a specific point in time) which includes the survival time; set of predictor variables which determine the function; the coefficients which measure the impact of the predictors; and the baseline hazard. The Cox model can be described as multiple linear regression of the logarithm of the hazard on the independent variables with baseline hazard (intercept) which changes in time (D.R.Cox, 1972).

4.4.3. ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANN) are another interesting area which gains prominence in a wide range of areas and applications. ANNs attempt to simplify and simulate the neural networks observed naturally in brains (Gurney, 2004). Artificial networks consist of connected neurons (nodes) and coefficients (weights) bound to these connections. Each neuron receives a set of weighted inputs and added all together these inputs obtain activation value, determining the output value by an activation function. The most common neural network structure consists of an input layer, hidden layer and output layer (Thomas, 2017).

Any type of ANN is described by three focal and significant elements – the neurons (nodes), the topology of the network and the learning algorithm defining the weights. The topology has a great impact of the model performance, even though there are no specified rules how to be defined – the ANN could be single or multilayer, with several hidden layers or only 1 (Rojas, 1996).

In terms of hidden layers, when a model has to be tuned, not only the number of layers but also of neurons have to be considered. The purpose of this layer is to obtain information about the inputs, join and associate them in a way which can establish it to the output neurons. Theoretically, an optimal and enough number of hidden layers for almost every problem is two. Regarding the number of hidden neurons, the rule of thumb is considered a suitable way to take the decision, but in the literature, the number varies from 6 to 12 (Baesens, 2015). To transform the inputs into outputs, activation function for the hidden units are needed. Based on the literature, the most used one is the sigmoid function as it gives values in the 0 – 1 interval (Guo, 2008) (Paasch, 2008).

The definition of the weights is also crucial for the success of the model. There are some methods which support the decision, for example, ranking method. However, ANN achieves good weights estimation by iterative optimization of the algorithm by initial random assignment of weights. It is common to use backpropagation – gradient descent method which adjusts the weights after comparing the derived output with the desired and calculating the error for each input neuron. However, one of the downsides of this approach is that it tends to define local, but not global minima. To minimize this, many studies show the application of modifications and mode of advanced techniques (Kriesel, 2005).

4.5. EVALUATION METRICS

In regression problems, which is the case and focus in this paper, the outcome of a model is continuous numbers. In respect of building and improving models, estimation of the performance of the prediction capabilities of a model is a focal point in the complete procedure. This chapter gives a theoretical overview of the two separate goals that will be observed later in this paper – model selection and model estimation. Here one of the most widely used metrics will be presented. This will allow the reader to build an understanding of how the trained models later will be evaluated and which are the metrics which will help to improve and select the one with the greatest performance.

In a situation with a significant amount of data, the most appropriate approach for both objectives is to randomly divide the dataset into three parts, namely training, validation, and test set. The training set is the one based on which the models are built, the validation serves as estimation tool of the prediction error for the purpose of model selection, and the test set is used to evaluate the generalization error of the chosen model (Hastie, Tibshirani, & Jerome, 2016).

In this work, the metrics which are point of interest are metrics which both can capture different error properties, e.g. noise, bias, as well as scaling, and provide the most robust results (Spuler, Sarasola-Sanz, Birbaumer, Rosenstiel, & Ramos-Murguialday, 2015). The most common metrics used in regression problems are a mean squared error (MSE), the average squared distance between the actual score and the predicted score, and root-mean-squared-error (RMSE), which is the square root of MSE. The higher the value for MSE, the worse the model performance is. These metrics have some issues because the equation results in average, which makes the metric not really robust to outliers. Even in the case of a single very bad prediction, the error will look significant (Zheng, 2015). Another metric a bit more robust to outliers is mean absolute error (MAE). It gives a value which is the average of the absolute difference between the target values and the estimated values.

However, the values of these metrics alone are hard to be analyzed whether they mean good model performance or not. To compare the model with a constant baseline may give a bit more clear understanding of the model performance. Regression analysis uses an equation which minimizes the distance between the fitted line and the data points. Considering the term goodness of fit, it can be said that a model fits well the data when the difference between the known values and the predicted ones is small. The coefficient of determination, R squared (R^2) is a metric which measures how close the data points are to the regression line. It is related to MSE but has the advantage of being scale-free. The higher the value of R squared, the better the model fits the data. However, this is not always an indicator for the model's accuracy and adequacy, as well as do not provide information on whether the predictions are biased.

After evaluating the model, a measure of its reliance and interpretation is needed. This paper will use the concept of permutation feature importance. This method suggests to rearrange randomly the values in a variable (iteratively one by one) and check whether the error will increase. If the error stays unchanged, then a variable could be considered as less important for the model. This step is considered valuable for this paper since it will provide an insight into the model's behaviour, for example in the case of ANN. Additionally, permutation importance does not require retraining of the model. This technique implies also some disadvantages. Permutation importance is linked to the error of a model. In cases when the point of interest is to examine how much the output varies for a feature without concentrating on the performance of this feature on the model. In a well-fitted model, the variance (how well a model is explained by the features) and feature importance correlate (Molnar, 2019). In the particular case, additional methods for model variance check might be applied, but this will be reviewed later in the chapters.

4.6. R PACKAGES

This chapter will give a brief outline of what R packages will be used to execute the project. The following table will present the main packages to be included, but not limited to, in this work based on their main topics.

Table 1 – R Packages[2]

PACKAGE	CATEGORY	DESCRIPTION	AUTHOR
dplyr	Data wrangling, data analysis	Powerful library for working with data frames, simple syntax.	Hadley Wickham
purrr	Functional programming	A functional programming kit, especially used for vector programming as mapping, modifying and transposing vectors.	Lionel Henry and Hadley Wickham

[2] The data for all references and authors were conducted from the online Community Cran (Cran, n.d.).

readr	Data sample	An efficient kit for loading efficiently rectangular data. It was mainly used to read csv files.	Hadley Wickham, Jim Hester, Romain Francois
tidyr	Data Cleaning	A collection of data tidying functions to efficiently reshape and clean data. Functions as replacing or dropping NAs were applied.	Hadley Wickham, Lionel Henry
stringr	Data Cleaning	Another library which provides different functions to target NAs and zero length vectors.	Hadley Wickham
gmodels	Data Modelling	A library which provides functions for model fitting. Especially used to estimate confidence metrics as confidence interval of the model coefficients.	Gregory R. Warnes, Ben Bolker, Thomas Lumley, and Randall C Johnson.
ggplot2	Data Visualization	One of the most used packages in the R environment. It provides functions to visualize data with different plots. Additionally, it provides the opportunity to customize any gives plot.	Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo
tm	Text Mining	The most famous package in the R environment for text mining approaches. It was used for analysing, cleaning and tokenizing text variables as well as for visualization of textual content.	Ingo Feinerer, Kurt Hornik
wordcloud	Text Mining	A library which provides functions to visualize textual data in form of frequency clouds, also known as word clouds.	Ian Fellows
pastecs	Time Series Analysis	A library which provides a set of functions to analyse as well as filter time series, for example the regularization and the decomposition of time-spaces.	Philippe Grosjean, Frederic Ibanez
moments	Data Exploring	A set of functions to calculate the skewness, Pearsons kurtosis and Geary's kurtosis.	Lukasz Komsta, Frederick Novomestky
reshape2	Data Transformation	A library designed for targeting complex data to restructure and aggregate by using two main functions, namely melt or decast.	Hadley Wickham
corrplot	Data Exploring	A library which provides functions to graphically analyse the correlations by using correlation matrices.	Taiyun Wei, Viliam Simko
car	Data Modelling	A complementary library to optimize the parameter settings of regression models, in particular box plots, etc.	John Fox, Sanford Weisberg, Brad Price
MASS	Data Exploration		Brian Ripley
caret	Data Modelling	The caret library is a common used set of functions to apply classification and	Max Kuhn. Contributions from Jed Wing, Steve

		regression training. Tools contain: data splitting; pre-processing; feature selection; model tuning using resampling; variable importance estimation	Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt
h2o	Data modelling	Scalable open source machine learning platform that offers parallelized implementations of many supervised and unsupervised machine learning algorithms such as Generalized Linear Models, Gradient Boosting Machines (including XGBoost), Random Forests, Deep Neural Network, Stacked Ensembles, Naive Bayes, Cox Proportional Hazards, K-Means, PCA, Word2Vec, as well as a fully automatic machine learning algorithm (AutoML).	Erin LeDell, Navdeep Gill, Spencer Aiello, Anqi Fu, Arno Candel, Cliff Click, Tom Kraljevic, Tomas Nykodym, Patrick Aboyoun, Michal Kurka, Michal Malohlava, Ludi Rehak, Eric Eckstrand, Brandon Hill, Sebastian Vidrio, Surekha Jadhawani, Amy Wang, Raymond Peck, Wendy Wong, Jan Gorecki, Matt Dowle, Yuan Tang, Lauren DiPerna, H2O.ai
neuralnet	Training of neural networks	Training of neural networks. The package allows flexible settings through custom-choice of error and activation function.	Stefan Fritsch, Frauke Guenther, Marvin N. Wright, Marc Suling, Sebastian M. Mueller
metrics	Models evaluation	An implementation of evaluation metrics in R that are commonly used in supervised machine learning. It implements metrics for regression, time series, binary classification, classification, and information retrieval problems.	Ben Hamner, Michael Frasco, Erin LeDell
glmnet	Data modeling	Procedures for fitting the regularization models	Jerome Friedman, Trevor Hastie, Rob Tibshirani, Noah Simon, etc.

5. METHODOLOGY

This chapter is organized in three main sections. In the first one, the original data set will be presented together with the problems and assumptions which can be inferred. A brief analysis of the variables will support both the identification of issues regarding data quality and data relevancy for the problem solution, as well as the decisions concerning the feature selection. Additionally, business and market insights will serve as a method to navigate further steps. The second section will present descriptive statistics of the features provided in the data set. Inference for both numerical and categorical data using statistical techniques and metrics will be the focal point in this part. Issues in the input data have to be considered before building a predictive model. The identification of data anomalies, noise and/or inflations, as well as their appropriate treatment, is a crucial part of the work which later guarantees more accurate model and better predictions. Discussion about outliers, missing values, dimensionality, scaling and distribution will be imparted. The third section will propose data-preprocessing procedures important for the modelling stage. Steps and techniques regarding different variable transformations will be presented in this part. The proper cleansing and preparation of the data set are essential to prevent inaccuracy of the final result. The chapter will be closed with a section describing the usage of different techniques for feature selection. In this part, variables will be selected on the basis of their results in diverse statistical tests for their correlation with the dependent variable.

5.1. EXPLORING THE DATA SET

The provided data set consists of more than 550.000 observation points and 19 variables, describing apartments, houses, stores, restaurants, garages, lands, etc., for rent or sale in Sofia, Bulgaria. The data is collected from the main online property listing website and contains historical information for the listings published in the period from 01.07.2015 until 01.07.2018. The table below lists the features which characterize a listing from the data set together with the accompanying description.

Table 2 – Variable list and description

VARIABLE NAME	DESCRIPTION
lid	Listing ID
date_first_seen	The date on which the listing of a housing object first appeared online
date_last_seen	The date on which the listing of a housing object was last seen online
rent_or_sell	Variable which indicates whether a housing object is for renting or selling
property_type	Identifies the type of property being for sale or rent
city	The city in which a property is located
neighborhood	The neighborhood in which a property is located
street	The street on which a property is located
space_m2	The area of a property in m2

price_in_bgn	The price of a property in national currency
price_in_currency	The price of a property in different currency
currency	Specifies the currency
build_type	Specifies the building material type
floor	Names the floor on which is a property
specials	Gives details about the condition of a property
description	Text description of a property
n_photos	Number of photos which a property has included in the listing
lister_type	Specifies whether the listing was made by owner, agent, investor, etc.
lister_username	The name of the account from which the listing was made[3]
broker_name	The name of the broker (company) which stays behind the listing

The data contains both qualitative and quantitative variables. The variables *date_first*/ *last_seen* describe the dates when a listing has been online and not available anymore. These two are used for the creation of the dependent variable of this research work – d*ays_on_market*. The variable *city* is constant for all observation points, namely Sofia city, and so it will be rejected as does not add additional information for the model. The variable *broker_name* will not be taken into consideration due to both poor quality (most of the names are in Cyrillic) and data privacy issues. For the variable *lister_username* also some data privacy issues could be inferred, but encoding of the characters into numbers could be executed. The relevance of this unique ids will be examined for the model development since this might provide further insights for fraud detection. The rest of the variables describe a property in terms of location, value, and specific attributes.

The variables *specials* and *description* contain details about the listed property. The variable *description* provides full text about the property amenities, while *specials* contains only keywords characterizing the exterior or interior of a property. A decision for the rejection of the variable *description* has been taken since the content is in Cyrillic and the alphabet cannot be handled by R and RStudio. However, the features provided by the variable *specials* summarize some main attributes of a property and will be further analyzed after translation in the next chapter with some text mining techniques.

The variable *floor* mainly informs about the floor on which a property is, as well as the total number of floors in the building, e.g. “5 of 12”. However, it contains also misplaced values regarding the area of the garden in m^2 for houses and villas, or some other words which purpose for the data set cannot be identified and are considered as mistakes. For explorative purposes, new variable called *space_m2_garden* was created.

[3] Almost all of the listings which appear online in Bulgaria, first appear in the main website imot.bg

The variable *build_type* contains several levels, namely brick, beams, MICCS, which stays for monolithic in-situ cast concrete structures, SF – sliding formwork, panel and under construction, together with the year when the building was constructed.

5.1.1. EXPLORING MISSING VALUES

The figure below provides an overview of the missing values in the original data set. The newly created variable *space_m2_garden*, since it corresponds only for houses and villas, has the greatest amount on NAs missing not at random (MNAR), followed by *street*, *broker_name* and *build_type* most probably missing completely at random (MCAR). At this stage, a decision to reject *street* and *broker_name* could be taken due to the reason that both variables have a high percentage of missing values, which cannot be even handled by imputation or other methods. Regarding *build_type*, additional analysis and decision for further steps will be completed in the next chapter.

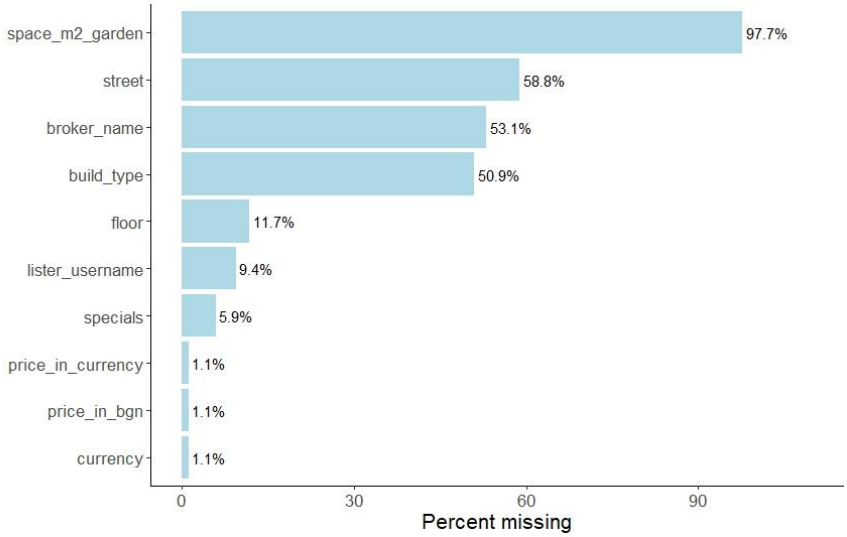

Figure 5 – Missing values in the input data set

5.1.2. DATA PARTITION

Homeheed currently provides a pilot version of their application for home seekers and their interest, as well as service, is limited to properties which can be clustered as 'home'. Therefore, the focus of this research work are only properties listed for living purposes and showing apartments (they could be found in the variable *property_type* named as 1,2,3,4[4], multiple rooms, studio, maisonette, room), which include a bit more than 400.000 observation points. At this point in the research, other listings will not be analysed and will be excluded from the data set. To provide the reader with a

[4] 1,2,3,4 represents the number of rooms which a flat has

visual understanding of the frequency distribution of the selected property types, a table and a graph illustrate not only the total amount listings of every property but also their distribution by real estate owner types. As can be seen, most of the listings are provided by real estate agencies.

Table 3 – Property types provided by real estate owner types

LABELS	1	2	3	4	MAISONETTE	MULTIPLE ROOMS	ROOM	STUDIO	GRAND TOTAL
agency	43632	147984	133338	24050	7145	9551	17987	3332	387019
agency (looks like)	100	671	749	174	65	94	1	51	1905
bank	2	24	25	3	2	8		6	70
builder	13	89	94	21	6	3			226
LABELS	**1**	**2**	**3**	**4**	**MAISONETTE**	**MULTIPLE ROOMS**	**ROOM**	**STUDIO**	**GRAND TOTAL**
investor	12	131	163	29	11	15	1		362
owner	4822	12601	9192	1425	509	584	1438	606	31177
GRAND TOTAL	**48581**	**161500**	**143561**	**25702**	**7738**	**10255**	**19427**	**3995**	**420759**

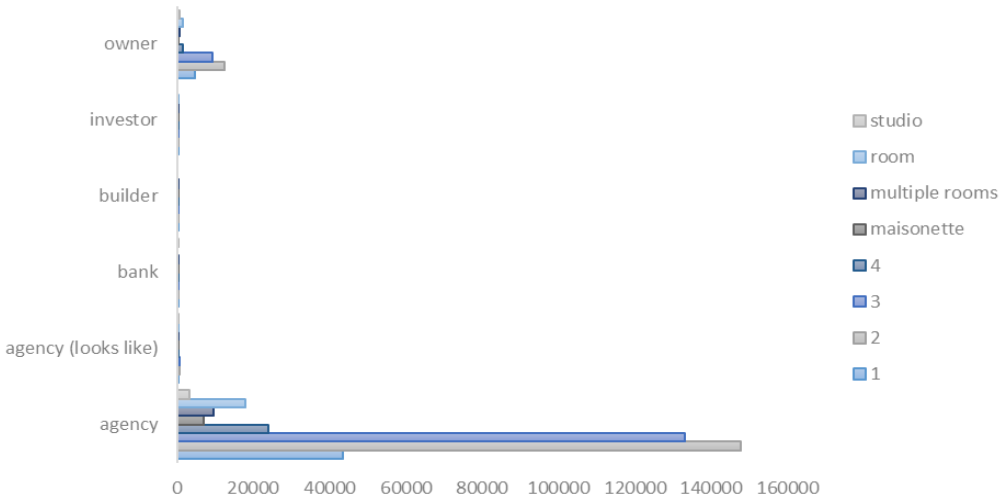

Figure 6 – Property types provided by real estate owner types

However, the market research made by Homeheed provides the information that the collected data about *DOM* of listings made by real estate agencies is not reliable enough and in some cases not real. The reason for that and the problem were stated in the introductory chapters. The missing piece is a variable which states whether a listing was really available or not at the moment when it was published. Since this information is not available and hard to be collected for all historical observations, building a model which predicts *DOM* for listings made by real estate agents will be highly biased. To overcome this issue, a decision to build a model which excludes listings made by agencies was taken.

5.1.3. INFERENCE OF NUMERICAL AND CATEGORICAL VARIABLES

This section will present the explorative analysis made on the data set which excludes the listings made by real estate agencies (includes the levels "agency" from *lister_type*)[5]. The scrutiny and evaluation will consider both numerical and categorical variables, as well as issues requiring further transformations.

Beforehand, to better understand the data and the market trend, the dependent variable *days_on_market* will be separately analysed through tables and dashboards summarizing graphically various pieces of the information extracted.

Different profiles of real estate owners/ agents who publish listings are assumed to have different behaviour. It is a point of interest to observe the distribution of *DOM* based on these profiles. The figure below shows a significant difference between the listings published by the owner and those published by other profiles. These first observations require further analysis later in the chapter where the effect of all variables on the dependent one, their capability to explain the target, will be considered in detail.

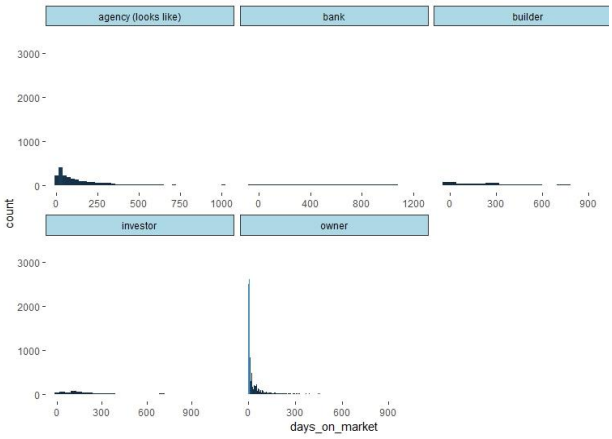

Figure 7 – DOM by lister_type

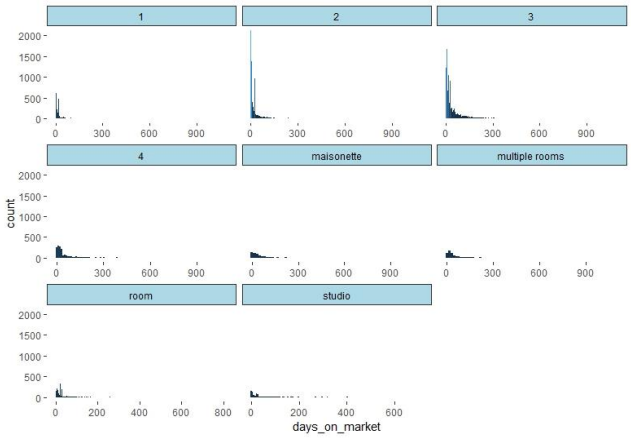

Figure 8 – DOM by property_type

An assumption for seasonality of the listings was made, so a major part is to check whether *DOM* of a listing can be determined as shorter or longer based on the month in which the listing has been published. For this reason, a graph which implements *property_type, rent_or_sell* and *month_start* was performed. The bar chart below shows that listing published in July has the maximum *DOM* for most of the property types.

[5] The level "agency (looks like)" is wrongly assigned during the data collection. Originally, all listings represented by "agency (looks like)" are from investors or builders.

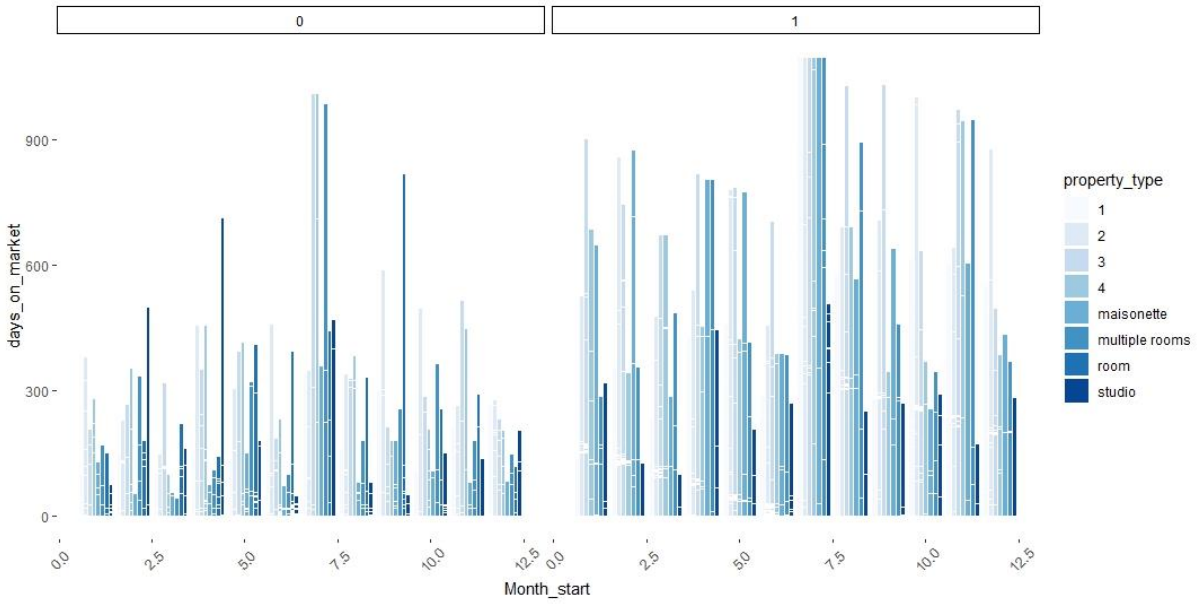

Figure 9 – DOM based on month when a listing was published

5.2. UNIVARIATE ANALYSIS

At this step, the chapter will propose a univariate analysis of the variables, as well as a bi-variate or multi-variate overview for some of them. Different statistics and methods will be used to understand individually both continuous and categorical variables. To examine the continuous variables, central tendency and spread of the data points have to be considered. Additionally, the skewness and kurtosis are appropriate statistics measures to further describe the symmetry and distribution of the data points. For the categorical variables, frequency tables facilitate to distinguish the distribution of each category. Afterwards, bi-variate analysis to identify the relationship between the different combinations of features has to be executed.

5.2.1. EXPLORING CONTINIOUS VARIABLES

These first type of features are the numerical ones. The focus here is the univariate analysis of the variables and their distribution. The figure below represents some basic statistics, describing the variables, including measures for central tendency, variability, standard deviation, to name but three. The measures provided include the numerical features available in the original data set (marked with red), as well as some additional features created for the purpose of the project. They will be discussed in detail in the section covering the data pre-processing part.

	space_m2	price_in_bgn	price_in_currenc	n_photos	Year_start	Month_start	Day_start	Year_end	Month_end	Day_end	floor_new	total_floors	year_built
nbr.val	33740.0	33398.0	33408.0	33740.0	33740.0	33740.0	33740.0	33740.0	33740.0	33740.0	32517.0	32517.0	10586.0
nbr.null	0.0	0.0	0.0	6104.0	0.0	0.0	0.0	0.0	0.0	0.0	1356.0	0.0	0.0
nbr.na	0.0	342.0	332.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1223.0	1223.0	23154.0
min	1.0	39.0	20.0	0.0	2015.0	1.0	1.0	2015.0	1.0	1.0	0.0	1.0	1900.0
max	8065.0	5280741.0	3617400.0	17.0	2018.0	12.0	31.0	2018.0	12.0	31.0	24.0	26.0	2021.0
range	8064.0	5280702.0	3617380.0	17.0	3.0	11.0	30.0	3.0	11.0	30.0	24.0	25.0	121.0
median	70.0	1271.0	850.0	8.0	2017.0	7.0	14.0	2017.0	7.0	15.0	3.0	6.0	2006.0
mean	79.6	81022.6	42467.9	7.9	2016.5	6.4	14.6	2016.6	6.5	14.8	3.9	6.9	1997.6
var	11536.7	14446727282.3	4742428195.7	32.2	1.0	10.5	82.3	1.0	10.5	83.4	7.7	9.7	385.6
std.dev	107.4	120194.5	68865.3	5.7	1.0	3.2	9.1	1.0	3.2	9.1	2.8	3.1	19.6

Figure 10 – Numerical variables basic statistics

The usual, useful and informative measures of central tendency are the (arithmetic) mean, median, and sometimes mode, which in this case is not included. For symmetrical distribution, the mean is a good measure, but when the skewed distribution is observed or when there is concern about outliers, the median is better to measure due to its robustness.

The variance and the standard deviation allow inference about how far from the center the values in the different variables. Considering these measures, first assumptions about the distribution of the observation points can be made. For normally distributed data, approximately 95% of the values lie within 2 standard deviations of the mean. In this case, only from *year_end* and *year_start* can be expected normal distribution. The standard deviation is not the most suitable measure when the values in a variable are not normally distributed.

Histograms are one of the most common visual tools to quickly investigate a lot about the data and make conclusions about central tendency, spread, modality, shape and outliers. Histograms support the illustration of the data distribution and serve as a method to envision skewness and kurtosis. Skewness measures the asymmetry, while kurtosis determines “peakedness” compared to the normal distribution. These measurements are useful for the differentiation of extreme values. In positively skewed (right-skewed) data values far from the mode are more regular and usually, the mean is greater than the mode. If the skewness is negative, then the mean is less than the mode. Regarding kurtosis, a positive one allows the interpretation that values which are far from the central tendencies are more probable, as well as that the shape is more centrally peaked, but the tail is greater. When the kurtosis is negative, then the peak is wider “shoulders” compared to the normal distribution (Seltman, 2015). The figure below illustrates the histograms and the distribution of some of the variables.

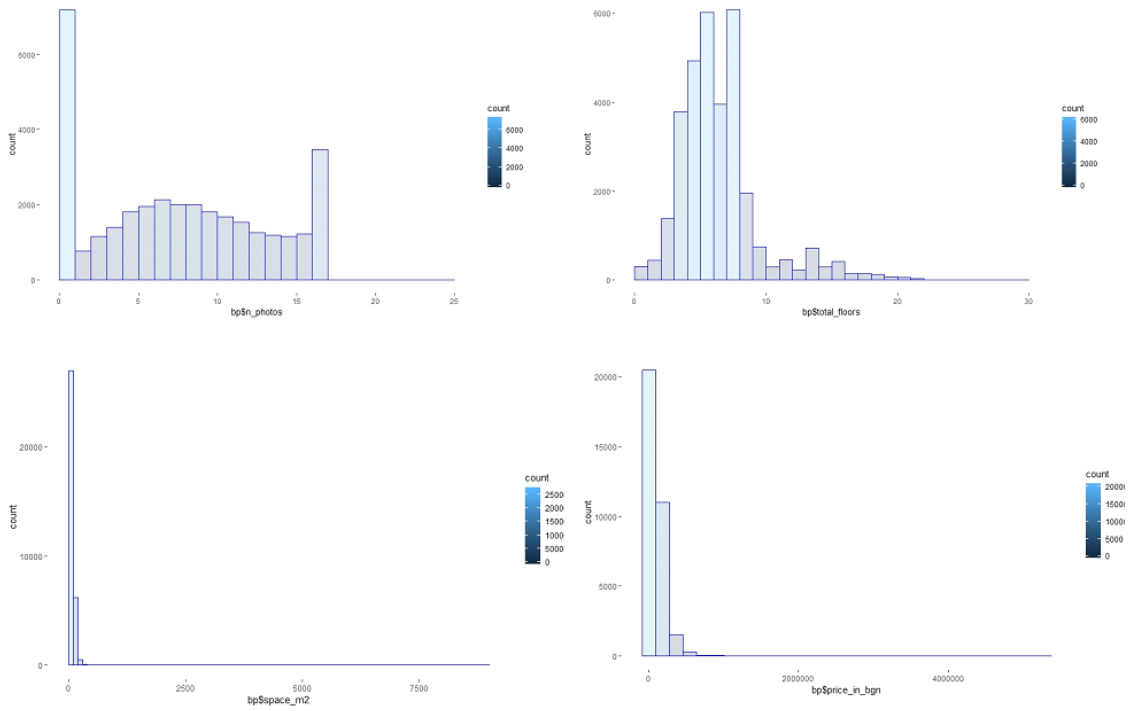

Figure 11 – Histograms of some numerical variables

Additionally, the table below represents the values calculated for the skewness and the kurtosis of the variables. The negative values imply that the distribution of the data is skewed to the left or negatively. The closer to zero, the slighter the skewness is, and likewise if the number is more distant from zero. Oppositely, when the value is greater than zero, the distribution of the variable data is positive/ skewed to the left. Concerning the kurtosis, a value less than 3 implies negative or flat and wide distribution, while a value greater than 3 should be interpreted as high and "slim" distribution (Asaad, 2013).

Table 4 – Skewness and kurtosis values of the variables

VARIABLE	SKEWNESS_OF_VARIABLE	KURTOSIS_OF_VARIABLE
space_m2	58.55332354	4076.35939
price_in_bgn	6.276439942	166.7471085
price_in_currency	13.76938472	556.4922087
n_photos	0.09699499	1.811480527
year_start	-0.019183275	1.928599009
month_start	-0.005586014	1.954408133
day_start	0.099231688	1.790969492
year_end	-0.085003855	1.933427314
month_end	-0.047972758	1.945445542
day_end	0.066149649	1.778833735
floor_new	1.464791477	6.643659231
total_floors	1.585019943	6.768881337

year_built	-1.06531323	3.845297287

5.2.2. EXPLORING TEXT VARIABLES

Another kind of features is extracted from the main description of the house, collected in the variable *specials*. They include keywords which describe a property regarding its construction and/or amenities. To give the reader an overview perspective on the features which generally are used in a listing, a word cloud is created. The figure demonstrates that "elevator" and "furniture" are from the main words which appear in the *description*, followed by "internet" and "brick". These features have to be extracted through text modelling as an essential part of the data preparation.

Figure 12 – Word cloud of the variable specials

Organized text format is considered to represent a table with one-token-per-row. A token is an important component of the text, for instance, word, which is noteworthy for analysis, and tokenization is the practice of separating the text into tokens. A token can be also n-gram or sentence. For example, in this particular case several combinations of words such as "entrance control" are observed and they are called bigrams. However, the variable *specials* itself contains multiple words which define property features, so the relationship and co-occurrence of words are interesting to be examined as well. The figure below illustrates the consecutive sequences of words which can be found in the *description* of a property.

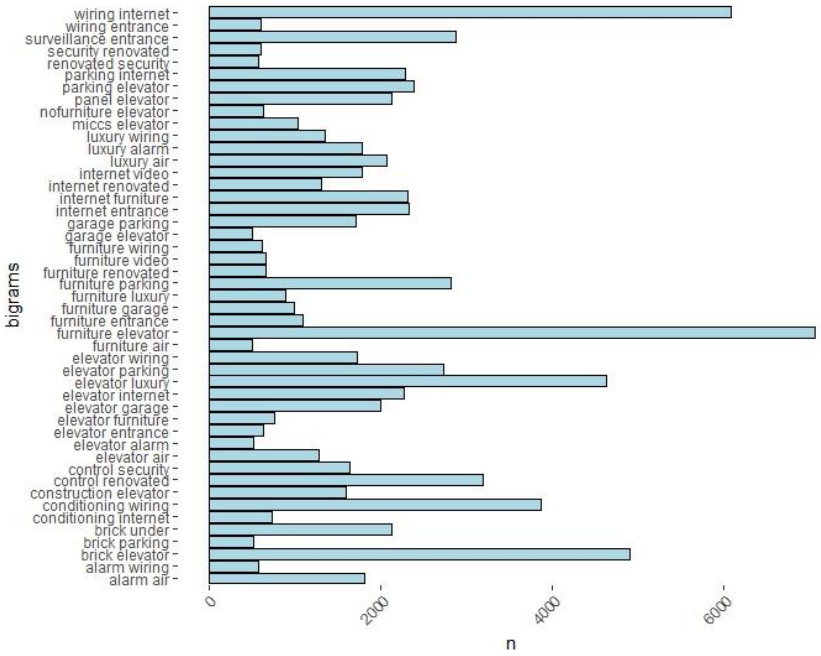

Figure 13 – Relationship between the words in the description

Not only “furniture” and “elevator” are the words which appear most frequent, but also the combination between these two words occurs repeatedly. Since this does not provide a lot of information, the correlation between words, which implies how often they are used in *description* together compared to how regularly they appear unconnectedly. To examine this, the phi coefficient, which is a measure for the binary association of features, is applied. This coefficient concentrates on how much more probable it is that either both words emerge, than that they can be seen independently. The figure below illustrates the four words which appear most often and the words which are most associated with them. Here should be mentioned, that e.g. “under” and “construction” have the same phi coefficient related to “brick” since “under construction” is a predefined special bigram. The same is valid for several more word combinations. Interesting point was to study the correlation between the words “furniture” and “elevator” due to their common occurrence, but the analysis showed a phi coefficient of only 0.096.

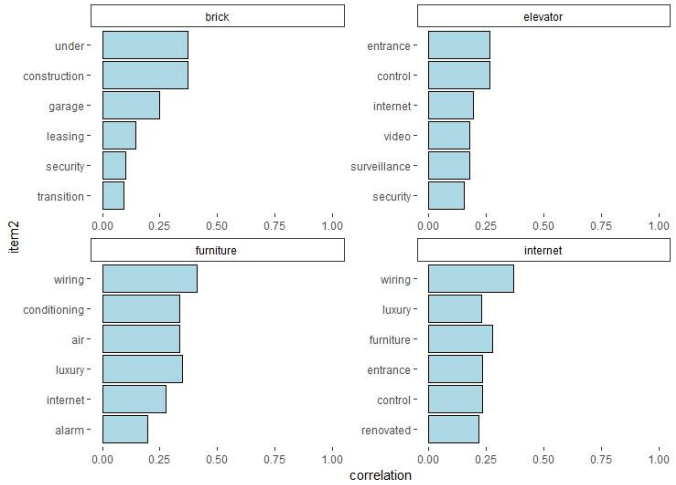

Figure 14 – Words correlation

The steps through which the variable *specials* was converted will be reviewed in more detail in the next chapter covering variables transformation. For the further development of this part should be mentioned that single words or bigrams were split as single binary variables which effect is interesting to be examined in the next pages.

5.2.3. EXPLORING CATEGORICAL VARIABLES

The last type of variables which can be observed in the data set are the categorical ones. To understand their distribution and the different categories, frequency tables and bar/pie charts can be used for visualization.

The first examined variable was *rent_or_sell* and the frequency table showed a small difference between both categories. In terms of *property_type*, as mentioned earlier, the levels which are studied in the project are 1,2,3,4, multiple rooms, studio, maisonette, room. The variable *lister_type* propose 5 levels – agency (looks like), investor, bank, builder, owner, where the agency (looks like) is wrongly assigned and should represent investor or builder, which will be aligned in the chapter considering variables transformation. The table below illustrates that owners are the main listing publishers among the studied ones and that flats with 2 or 3 rooms have the highest supply level for both rents or sell.

Table 5 – Cross table for property type by listing provider and by rent or sell

		1	2	3	4	MAISONETTE	MULTIPLE ROOMS	ROOM	STUDIO	TOTAL
0	AGENCY (LOOKS LIKE)	7	56	54	15	2	5	1	7	**147**
	BANK	0	0	0	0	0	0	0	0	**0**
	BUILDER	2	4	2	0	0	1	0	0	**9**
	INVESTOR	1	10	7	0	1	1	1	0	**21**
	OWNER	3251	7586	4027	455	174	191	1438	449	**17571**

		1	2	3	4	MAISONETTE	MULTIPLE ROOMS	ROOM	STUDIO	TOTAL
	AGENCY (LOOKS LIKE)	93	615	695	159	63	89	0	44	**1758**
	BANK	2	24	25	3	2	8	0	6	**70**
1	BUILDER	11	85	92	21	6	2	0	0	**217**
	INVESTOR	11	121	156	29	10	14	0	0	**341**
	OWNER	1571	5015	5165	970	335	393	0	157	**13606**
	TOTAL	**4949**	**13516**	**10223**	**1652**	**593**	**704**	**1440**	**663**	**33740**

The variable *neighborhood* contains a great number of levels. The center region offers slightly more listings, but still, none of the neighborhoods preponderates significantly. Nevertheless, steps to recode the variable will be executed in case the variable has a meaningful impact on the target one.

Appealing fact, showed on the figure below, observed during the analysis is that the variable *type_built* contains levels only when a property is listed for selling. When a listing is marked for renting, then the construction type is unknown. This should be considered during the missing value treatment.

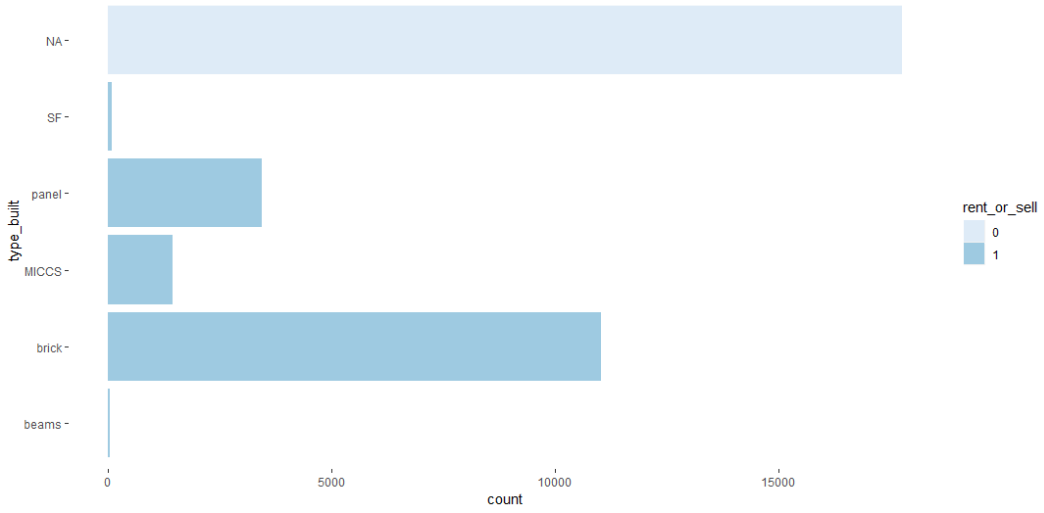

Figure 15 – Type_built by rent or sell

5.3. TRANSFORMING THE DATA SET

After the statistical exploration of the variables, an analysis regarding the main issues, which have to consider during the data pre-processing part, has to be executed. The train data suggests several questions and concerns which has to be studied before proceeding to the model in order to guarantee better performance.

5.3.1. EXPLORING MISSING VALUES AND OUTLIERS

Missing values is the first issue which is considered in this chapter. The significant level of NA-s can reduce the capabilities of the model because the behavior and the association with the features are not correctly considered and studied. The graph below provides an overview of the missing values in the data set (including some transformations which will be discussed in the next chapter). The *year_built* analysis shows that the data provide some information of the year for properties for sale purpose, but these for rent contain no year of the building construction. The variable *type_built* is also available only for properties for sale intention. These observations lead to an assumption that both represent missing values not at random. Maybe listing publishers have greater relevance to provide this information only when they want to sell a property, while when a customer searches to rent s/he by her-/himself is not so interested in this particular facts. The rest of the variables do not have a significant amount of missing values.

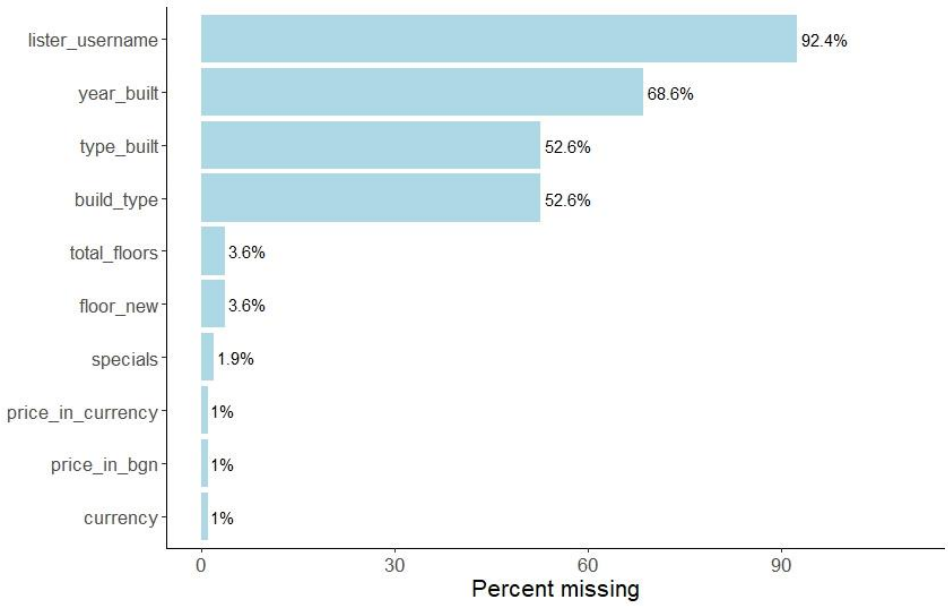

Figure 16 – Missing values in the data set

The next step before the transformation of the data is the detection of outliers. Outliers have an excessive impact on the data if no actions are taken. They increase the error discrepancy and decrease the supremacy of numerical tests. Outliers can affect normality, as well as the fundamental hypothesis of some statistical models. In some variables were detected long tails, whereas in others there were observation points which were significantly distant from the mean. In order to capture as much as possible any possible issues which will affect the model accuracy, both univariate and multivariate analysis for outliers have been executed.

The most conventional explanation of an outlier describes it is a value that is located distant from the major part of observations for a feature and could disfigure and misrepresent summaries of the distribution of data points. In reality, this can be interpreted to a value which is 1.5 times the IQR (interquartile range) more extreme than the quartiles of the distribution. The most applicable and useful way to detect an extreme value is boxplot. The four figures below illustrate the boxplots of four features. Extreme values which require attention can be seen, as well as long tails.

Figure 17 – Boxplot for price_in_bgn

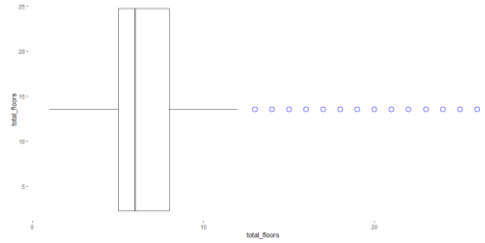

Figure 18 – Boxplot for total_floors

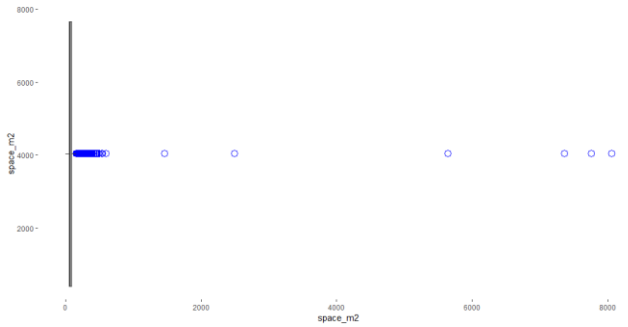

Figure 19 – Boxplot for space_m2

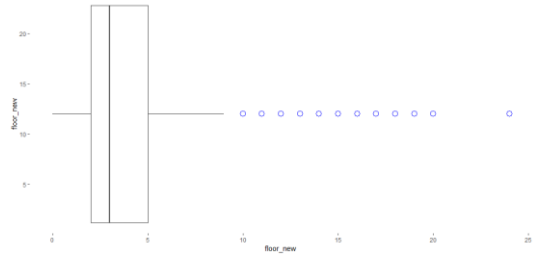

Figure 20 – Boxplot for floor_new

One method which can support the better understanding of these outliers is the breakdown of the variable which is observer based on the values in another feature. This is called multivariate analysis. In the figure below is demonstrated example with separating *space_m2* based on property type and by price in currency Bulgarian Lev. The plot shows that there are mainly outliers in 1,2 or 3-rooms apartments and that places with extreme space have extreme price.

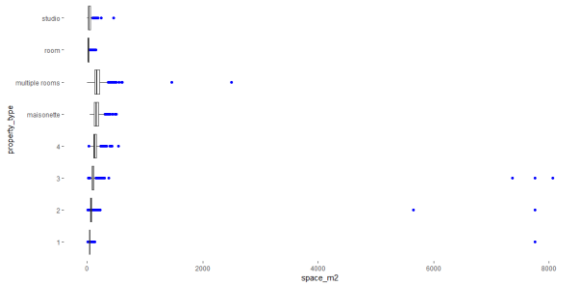

Figure 21 – Outliers of space_m2 based on property_type

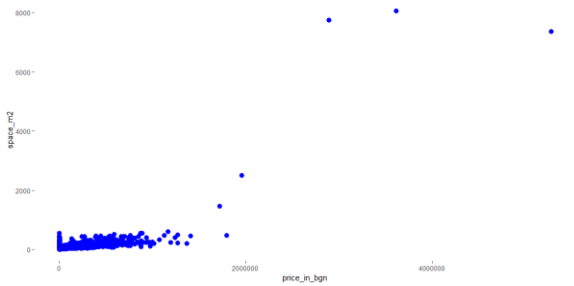

Figure 22 – Outliers of space_m2 based on price_in_bgn

Another data issue which has to be mentioned is the high dimensionality. The original data set includes 19 variables. Some transformations which will be presented in the next chapter will increase the number even to 54, so variable selection techniques have to be applied to choose the most valuable predictors for the model.

An important step which has to be stated is the scale of the variables. The range of values in the features varies and since a lot of algorithms use the distance between observation points to calculate prediction, a common scale is needed to assure that none of the features will be dominant. Further, as revealed previously, the distribution of the data in some variables shows skewness which might be a problem for some of the models. By scaling the data, normality might be achieved.

5.3.2. MODIFYING DATA ANOMALIES

In this section, the procedures regarding handling anomalies, noise, and missing data, as well as creating and transforming new variables. The features which have been created and transformed will be introduced first. Afterwards, the methods to control the missing values and outliers, as well as scale and distribution will be all reported. The purpose of this section is to present the work executed to prepare the data in a structured format which can allow the feature selection and the building of the model.

In the previous section, a graph illustrating the percentage of missing values was shown. The theory introduces various methods to treat this issue. In the case of the variable *lister_username*, the column has to be dropped because more than 92% of the observation points are missing. The same is valid for *year_built* and *type_built*. Even though both variables contain information for properties which are for sale, the imputation or prediction for 50% of the observation points will be either extremely time-consuming or the predicted values will behave better than the original values. For the case where the analysis shows missing values complete at random, deletion can be executed. In this particular data set, no specific explanation for the missing values was defined. This led to a decision for the deletion of all NA-s. Additionally, some variables are related, for example, floor and

total floors in a building – a flat cannot be on the 8th floor if the building has in total 6, or the *price_in_bgn* cannot be less than the *price_in_currency*, which means that imputation may cause incoherencies which have to be further examined. As the total number of missing values was also not significant, with no concern they have been removed.

5.3.3. MODIFYING OUTLIERS

The next step of the work is the treatment of outliers. The method used to reveal and analyse outliers is visualization. Box-plot, histograms, and scatter plots were built to understand the distribution of the observation points. To treat them several rules of thumb were applied to take a decision for deletion, namely that any point which is outside the range of -1.5 x IQR to 1.5 x IQR, or three or more standard deviations away from mean, is considered as an outlier. Extreme points were detected only in two variables, *price_in_bgn* and *space_m2*, and their total amount was small and inconsequential for the overall analysis, so these points were removed from the data set.

Once all this have been completed, specifically variables exploration, as well as missing values and outliers handling, a step towards variables transformation and engineering can be made. The first stage is to use the current data and make it a bit more useful for further tasks. This includes, for example, the separation of *date_first/last_seen* into the year, month and day start/end, because one of these might be a more important feature for the model. The table below serves as a structured guide which aims to clearly present the recoding of the original variables which was executed.

Table 6 – Recoding of original variables

VARIABLE	ORIGINAL VERSION	RECODED VERSION
date_firsl/last_seen	Both variables are in format "yyyy-mm-dd"	6 new variables were created, namely *year_start/end; month_start/end;* and *day_start/end*
rent_or_sell	"rent" and "sell"	Recoded to binary one – 0 rent, 1 for sale
property_type	1 2 3 4 maisonette multiple rooms room studio	Recoded completely to numbers - 1 2 3 4 5 6 7 8
lister_type		
neighbourhood	Contains many character values with the name of the neighbourhood	An id number was assigned for the different neighbourhoods
lister_type	Contains the levels "owner", "investor", "builder", "bank", "agency (looks like). The value "agency (looks like)" is a mistake made during data collection. It represents in reality either investors or builders.	Since no strict condition for the recognition between investor and builder was found, the value "agency (looks like)" was randomly replaced to be either builder or investor. New variables with codes from 1 to 4 were created.
build_type	Originally the variable contains year and building material	The variable was split in two new variables – *year_built* and *type_built*

specials	Text variable in the format [\word1\",\"word2\", \"word3\",….]	Binary variables for each word indicating the existence or lack of this feature.
floor	Originally in the format for example "5 of 10"	Split in two new variables *floor_new* and *total_floors*

Before feature extraction from the variable *specials*, the first part of the pre-processing is cleaning with some fundamental steps. The first task is the removal of punctuation because it has no added value to the information in the column. Further, all letters are converted to lower case. This prevents multiple extracted copies of the same word. Since the variable itself contains only keywords, some basic pre-processing procedures, such as stop words removal or stemming (removal of suffices), were not executed. However, the last step was the conversion of a single word into binary variables.

Two new variables were additionally introduced – *price_per_m2* and *n_features*. The first one is calculated based on *space_m2* and *price_in_bgn*. The second one represents the total number of features, including as keywords in the *description*, available for a listing.

As mentioned earlier, the range of the values differs. Most of the algorithms use the distance between the observation points to compute predictions. Feature scaling or data normalization is needed to assure that all features will have the same scale and will be comparable to each other, as well as none of them will dominate in a function. Optimization of the scale will affect the mean and the standard deviation, and so might help to achieve normality. Common normalization methods are Min-Max, which scales the range between 0 and 1, and Z-score, which scales values between -1 and 1. In this case, the values have been normalized using Min-Max function.

5.4. SELECTING AND EXTRACTING FEATURES

In this section, several methods for feature selection will be presented. Choosing the proper variables is a fundamental step in the data pre-processing. In this work, feature selection is described separately, because executing the right steps to select variables to train the algorithm, diminishes training and evaluation time. The later applied machine learning algorithm will be also easier for interpretation due to its reduced complexity. Furthermore, accuracy can be improved and overfitting prevented.

Feature selection is an important step when the data set contains a large number of variables, because on the one hand it allows proper and the best knowledge wrapping, and on the other hand – dimensionality reduction. The curse of dimensionality is a phenomenon which refers to the exponential growth of data which causes its high sparsity and degrade the classifier performance. By

reducing the dimensionality, the algorithm works faster and the model is easier for interpretation. Filter methods are usually used as a data preparation step. The selection of features is developed based on the scores in various statistical tests for the relationship/ influence of the independent on the outcome variable.

To begin with, the correlation check will be presented to illustrate the relationship between the continuous variables. The figures below illustrate both heat matrix of Pearson and Spearman correlation check.

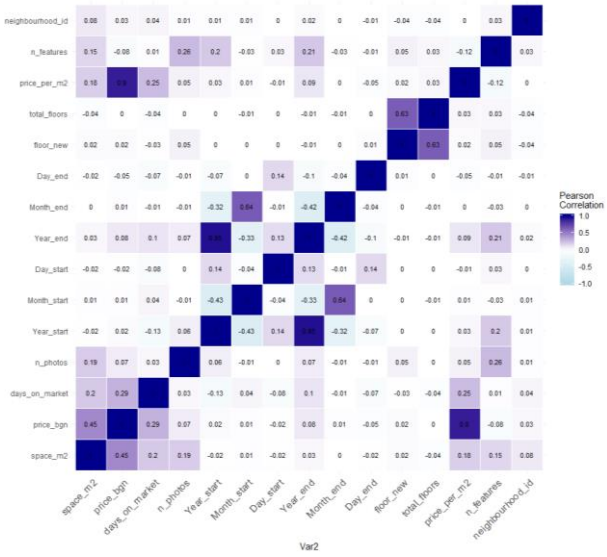

Figure 23 – Pearson Correlation Heat Map

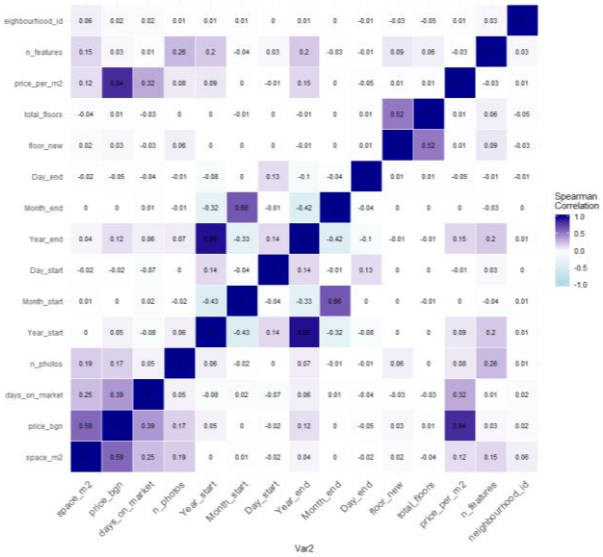

Figure 24 – Spearman Correlation Heat Map

The correlation plots helped to determine and measure the extent to which two variables are associated and the direction of their relationship. The Pearson correlation measures the linear relationship between two variables, while the Spearman correlation coefficient measures the strength and direction of the monotonic relationship. However, both show significant correlation coefficient only for the expected *year_start* and *year_end*, and *price_bgn* and *price_per_m2*.

The correlation check alone can limit the detection of multicollinearity since it is only pair-wise. One of the techniques which support the detection of more complicated correlation relationships is the usage of Eigenvalues. A small magnitude shows that there is no multicollinearity, while a high range between the values is a signal for significant multicollinearity, which is the case here. The Variance Inflation Factor (VIF), which indicates how much the variance of a regression coefficient is overestimated due to multicollinearity, can be calculated. The minimum result values for VIF is one and as a rule of thumb, results between 5 and 10 are considered as indicators for the problem. In

this case, *year_start* and *year_end* showed extreme results above 20 and *price_bgn* has a result around 9. To solve this issue, the variables should be excluded.

To examine the relationship, in particular, the significance level, between the categorical variables and the target, Kruskal-Wallis test as an alternative to one-way ANOVA test will be executed. One-way ANOVA test is used to calculate whether there are any statistically considerable and important dissimilarities between the means of two or more independent groups. This test requires several assumptions to be fulfilled, namely (1) the dependent variables is continuous; (2) the independent variable has two or more categories; (3) no relationship between observations in the groups or between groups; (4) lack of extreme observation points; (5) the target variables have normality for each group; (6) homogeneity of variance – the distribution of scores around the mean of the samples are equal (Laerd Statistics, n.d.). Since not all of these assumptions are satisfied, Kruskal-Wallis test, a non-parametric method, is considered as more suitable and is executed. In this case still four main assumptions have to be respected: (1) the dependent variable is measured in continuous level; (2) the independent variables contain of two or more categorical non-related groups; (3) there is no relation between the points of each group; (4) as Kruskal-Wallis test does not require normality, but takes care of the distribution in each group which needs the same shape (Laerd Statistics, 2018). Before performing this test, is has been checked whether the four assumptions are violated. A p-value which is less than 0.05 indicates a significance level between the groups. In this particular case, only the variables, extracted from the *description*, *telephone_exchange* and *elevator* had p-value higher than 0.05. Since almost all categorical variables showed the significance for the model, certain uncertainty about the correct choice requires further steps in the feature selection process.

Based on these filter methods, correlation and Kruskal-Wallis test, not a significant amount of variables can be excluded. To select the proper variables for the model, wrapped method (forward selection/backward elimination) and embedded method (e.g. LASSO regression) could be applied and inferences about variables importance can be made.

LASSO, which is an abbreviation for Least Shrinkage and Selection Operator, is a powerful method for two focal tasks, namely regularization and feature selection. The LASSO method applies a shrinking (regularization) process where it corrects the coefficients of the regression variables shrinking some of them to zero. During the variables selection process, features which have a non-zero coefficient after the shrinking process are selected as the best predictors. The objective of this procedure is to minimize the prediction error (Fonti, 2017). The figure below illustrates the variable importance based on the LASSO method.

Additionally, multiple regression using the stepwise algorithm of adding and dropping possible variables was computed. The model uses AIC (Akaike Information Criterion) to compare the improvement of the model by the elimination or supplement with different variables. The results of this method suggested a greater amount of significant variables than this proposed by the LASSO algorithm.

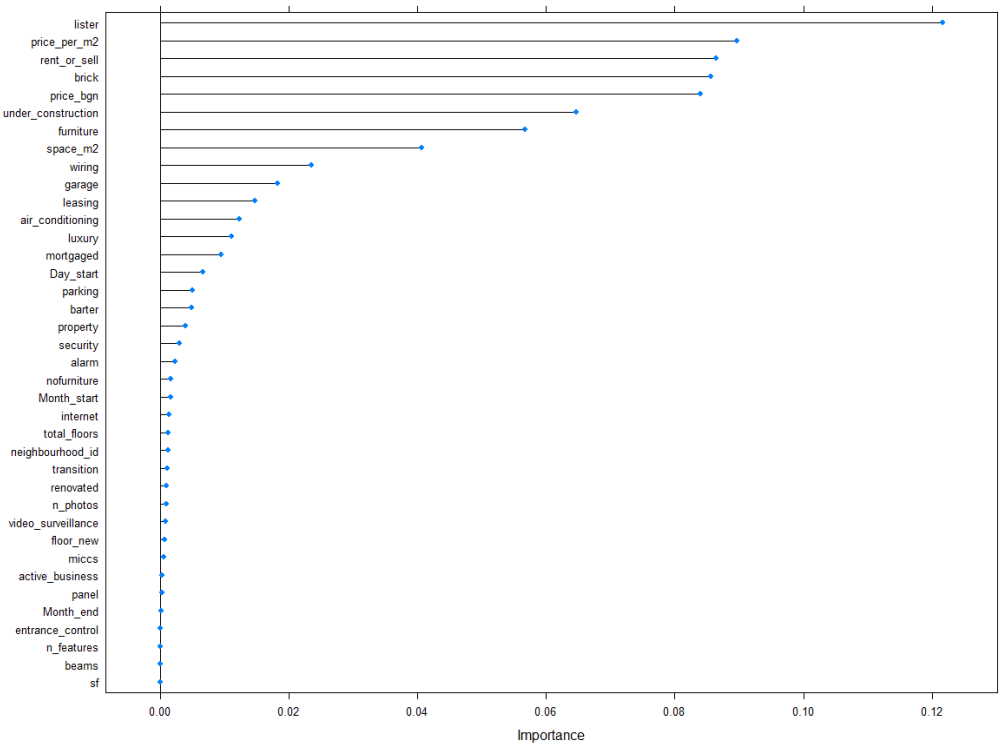

Figure 25 – LASSO variables importance

5.5. MODELLING

In this chapter, the modelling techniques executed for the project are presented. In the following pages the reader will have the opportunity to understand what models were trained and how they were analyzed by going through the main statistics used for evaluation and comparison. In addition, the interpretation of the selected model will be shown and explained.

The model selection has to be seen as an estimation process for evaluating the performance of several algorithms applied to the data set with the purpose to select the possibly most accurate one. The estimation of the model is performed on out-of-sample data set used to calculate the error. This approach is known as cross-validation and suggests as best practice to divide the data set into

subsets – training, testing and validation[6]. This step is crucial for the statistical analysis (building the model) because it aims to support the assessment of the model results in independent data. Additionally, issues like overfitting or underfitting are observable, as well as a comparison between models is possible since the target is to select the one which performs the best on the unseen data set. The dividing rule here is chosen to be 70% for the training set and 30% for the testing one.

5.5.1. PENALIZED MODELS

The first modelling techniques described in this section are the penalized models – lasso, ridge, and elastic net. These approaches are known as penalized (also regularization or shrinkage) regression models because they minimize the residual sum of squares by applying a penalty on the regression coefficients, which causes their continuous reduction towards zero. This shrinkage affects the variance and so the mean squared error (MSE), which results often in lower MSE for penalized regression than in simple linear regression model. Thus, regularization models sometimes have a better predictive performance on unseen data (Funda Gunes, 2015). The main difference between lasso and ridge regression is the penalty assignment. Lasso uses L1 norm, meaning that it adds a penalty equal to the sum of the absolute value of the magnitude, while ridge which applies L2 norm, uses the sum of square coefficients. Elastic net is a cross technique, which uses both L1 and L2 norms and by doing so effectively reduces coefficients and set some of them zero. The penalty which is lambda is what defines the amount of shrinkage and its setting is crucial for the model. The best lambda is the one which minimizes the error rate. In this case, a grid range of lambda values was defined. The other value which the model formula requires is alpha and it takes values between 0 and 1: 0 is for ridge regression; 1 for lasso; and everything in between is corresponding to the elastic net.

Each of the three models was trained through the grid of the predefined lambdas. By comparing RMSE results, the optimal value of lambda which minimizes the RMSE was selected. The following summary of the results comparing the models can be considered:

Table 7 – RMSE of regularization models

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
Ridge	0.059623733	0.0656467516	0.06861103715	0.067766678	0.07011292	0.07343618	0
Lasso	0.056119428	0.064287629	0.06688481	0.067545677	0.070250247	0.081202623	0
Elastic	0.059611997	0.065685412	0.068609881	0.067759895	0.070082334	0.073416531	0

[6] In this project the data will be split only in training and test set. The data selected to execute the project is not big enough to assure the accuracy of the result, as well as an assumption for low signal, based on the exploration part, was made.

The optimal value of lambda for the ridge regression model is equal to 0.0023101297 with RMSE of 0.06458345108.[7] This can be seen also from the graph – the higher the lambda (the regularization parameter), the worse the model (RMSE value is increasing).

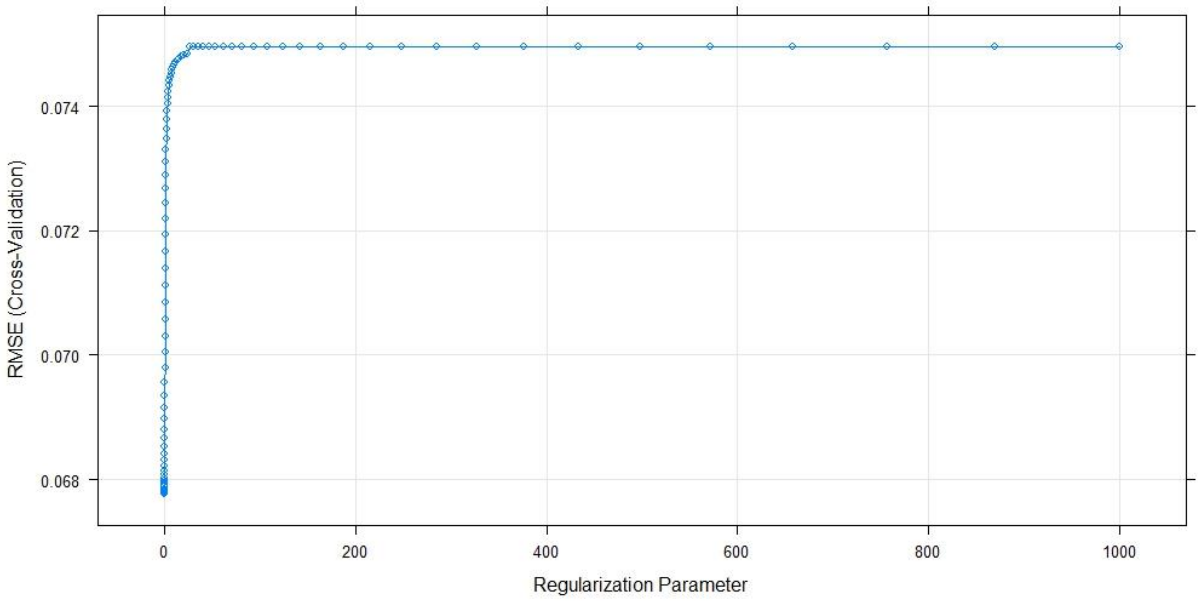

Figure 26 – Ridge Regression RMSE and regularization parameters

The table below illustrates the ridge regression coefficients and as can be seen, none of them is 0, which means that the model contains all the pre-selected variables. Similar to the simple linear regression, the coefficients give some visibility on how *days_on_market* will change when (1) a continuous variable increase by specific number (e.g. when *space_m2* increases it results in increase on average day_on_market by 0.087954307932) and (2) a categorical variable – e.g. on average *days_on_market* is higher with 0.017990042310 for properties for sale (dummy coding value = 1), than for the ones for rent.

Table 8 – Ridge regression coefficients

Variable	**Coef.**
(Intercept)	0.007812742613
rent_or_sell	0.017990042310
space_m2	0.087954307932
brick	0.006890894870
furniture	-0.005648094327
under_construction	0.023199993261
price_per_m2	0.012193552241
lister	0.126633821677

[7] The median value for RMSE is considered.

The analysis was done for both lasso and elastic net models. The following table illustrates the coefficients for both models. Afterwards, the graphs illustrate the relationship between the regularization parameters and the RMSE results. The best performing LASSO regression is the one with lambda 0.001 and RMSE equals to 0.06455596484. Regarding elastic net – lambda = 0.0001474162544, alpha = 0.1 and RMSE 0.06456125674. It can be concluded that the LASSO regression has the best performance.

Table 9 – Coefficients Lasso and Elastic Net

VARIABLE	COEF. LASSO	COEF. ELASTIC
(Intercept)	0.008262840044	0.006914658626
rent_or_sell	0.020773920773	0.021302049363
space_m2	0.081148777731	0.088121228126
brick	0.006277781636	0.006389871295
furniture	-0.004166719412	-0.004836976890
under_construction	0.020923002635	0.022619172277
price_per_m2		
lister	0.127546013431	0.130465654935

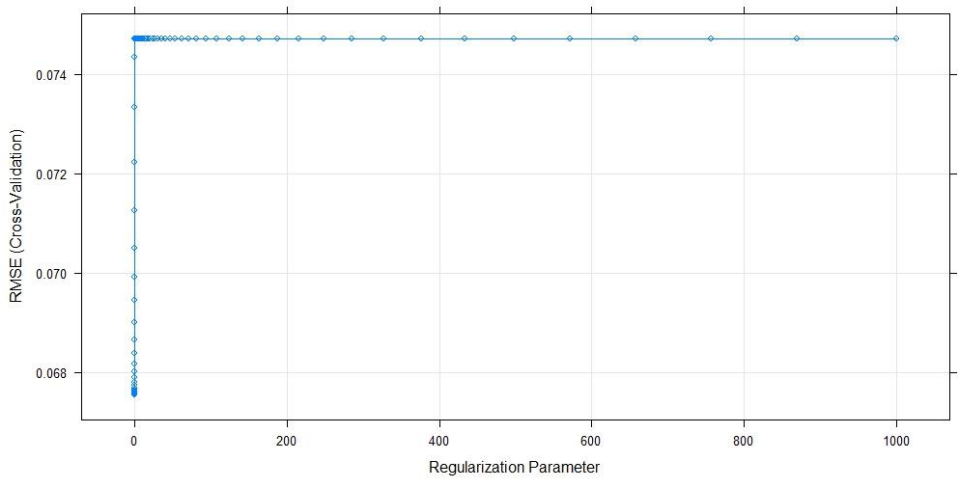

Figure 27 – LASSO RMSE and regularization parameters

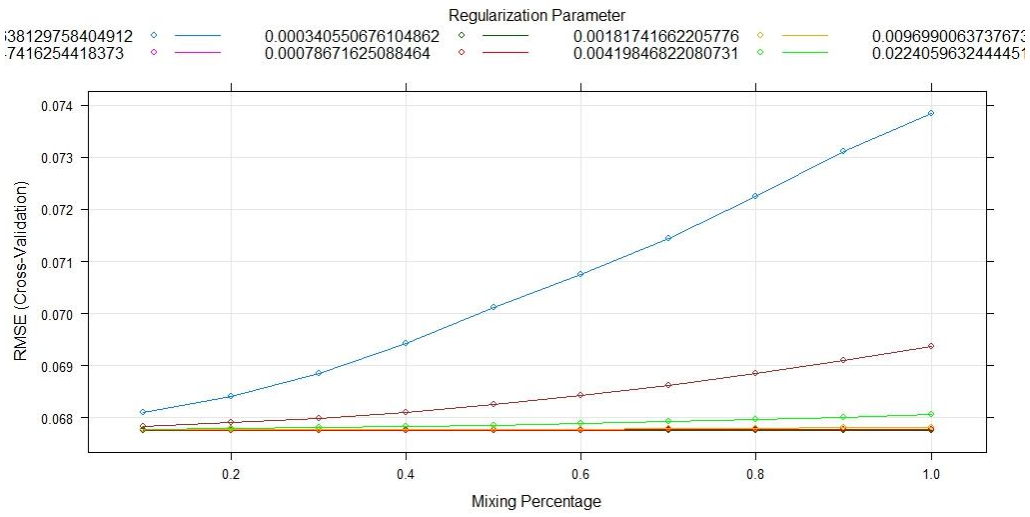

Figure 28 – Elastic Net RMSE and regularization parameters

To measure the strength of the relationship between the models and the target, goodness-of-fit can be measured with R-squared. This statistic reveals how well the predictors explain the target variable. R2 ranges from 0 to 1, where 0 suggests that the developed model does not have any impact on the prediction, while 1 indicates perfect prediction. As can be seen from the table below, the three models have R2 around 18-19%.

Table 10 – R-squared regularization models

Rsquared

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
Ridge	0.13978064	0.16879293	0.1814451	0.1882617	0.2024361	0.225172	0
Lasso	0.13948731	0.17052054	0.1875318	0.1879821	0.2027301	0.2244501	0
Elastic	0.14068227	0.16838816	0.1816644	0.1880144	0.2027459	0.2256157	0

19% is low R2 value, but here it should be highlighted that this does not mandatorily indicate an inadequate model. One of the reasons in this particular case is that, as mentioned in the beginning, there is an initial problem with the data and the *days_on_market* values due to incorrect property advertisements. However, the predictors in the model are statistically significant and the conclusions made regarding how the target will change when change with one unit occurs in the predictors are still valid and valuable. To further analyze the model, a residual plot will be helpful.

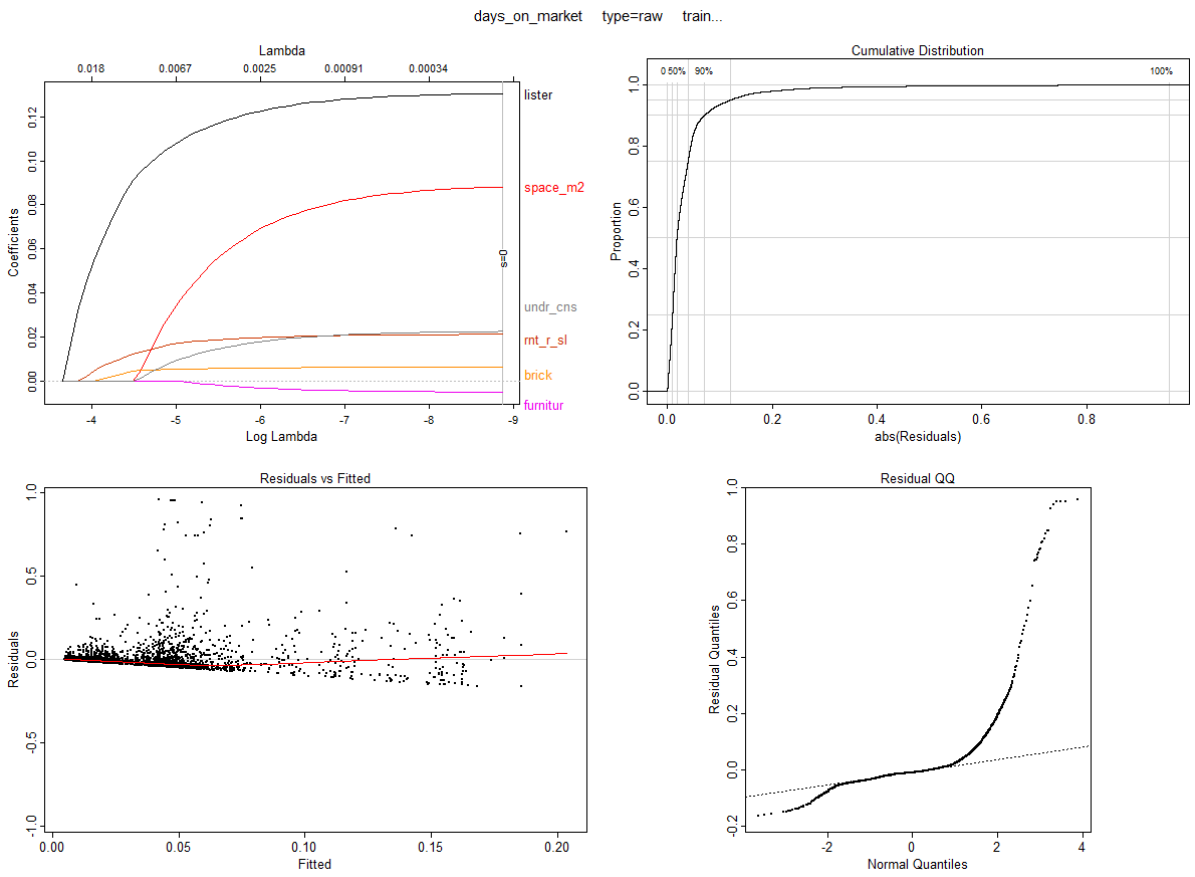

Figure 29 – Residual vs. Fitted Plot, QQ plot and Lasso coefficients trace

The residual vs fitted plot provides an overview of two main things – linearity (the relationship between the features) and homoscedasticity (the variance along the regression line). The "perfect" plot would not have any specific shape – the lack of data points pattern and their random positioning around 0 implies linearity. A sign for the relatively equal variance of the error terms is the "flat line" around the 0 line (The Pennsylvania State University, 2018). On this particular graph, some possible outliers can be seen, as well as that not all points are symmetrically distributed around the 0. The Residual QQ plot showing curving indicates some skewness with upward cavity denoting positive skewness. The shape of this QQ plot can serve as a visual tool to indicate the existence of some extreme values. The first graph shows the trace of the predictors. Price_per_m2, as seen also in summary of the coefficients, is rejected from the model. The first variables which enters the model, as well as the most significant one, is *lister* and it slowly and steadily reaches its final values. On a later stage, *space_m2* also enters the model and its effect is also visible, while the rest of the variables do not show so considerable impact.

5.5.2. SURVIVAL MODELS

Another modelling technique which was used and found suitable for this case is the survival model, in particular Cox regression. For one property listed in Homeheed, there is a set of feature specifying aspects of this property, such as rent or sell, price, neighbourhood, number of rooms, etc. Let property x appear on the website at time t1 and gets down at time t2. The time between t1 and t2 is the interval, days on market, or in survival analysis the time period in which a property "survives" before being sold or rent. The target in this project work is to predict the probability of a property, represented by a set of features, to be sold/rent in a given time period. The values below show the summary statistics of the Cox regression.

```
Call:
coxph(formula = Surv(days_on_market, rent_or_sell) ~ ., data = train_)

n= 21886, number of events= 10130

                      coef             exp(coef)          se(coef)      z          Pr(>|z|)
space_m2         -0.35217438          0.70315750         0.12035178   -2.92621      0.0034312 **
brick             0.19391552          1.21399372         0.02627833    7.37929      0.00000000000015913 ***
furniture        -0.46902011          0.62561500         0.02388139  -19.63956      < 0.00000000000000222 ***
under_
construction     -0.27012452          0.76328444         0.03283422   -8.22692      < 0.00000000000000222 ***
price_per_m2      4.33868827         76.60698586         0.08574912   50.59747      < 0.00000000000000222 ***
lister           -1.04442663          0.35189353         0.06085715  -17.16194      < 0.00000000000000222 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef)          exp(-coef)          lower .95          upper .95
space_m2         0.7031575          1.42215649          0.5554040          0.8902177
brick            1.2139937          0.82372749          1.1530504          1.2781582
furniture        0.6256150          1.59842714          0.5970068          0.6555941
under_
construction     0.7632844          1.31012758          0.7157113          0.8140198
price_per_m2    76.6069859          0.01305364         64.7557880         90.6271156
lister           0.3518935          2.84176868          0.3123270          0.3964724

Concordance                0.743 (se = 0.003)
Rsquare                    0.177 (max possible= 1)
Likelihood ratio test      4264.43 on 6 df,                    p=<0.0000000000000002
Wald test                  5408.09 on 6 df,                    p=<0.0000000000000002
Score (logrank) test       4754.14 on 6 df,                    p=<0.0000000000000002
```

Figure 30 – Cox Regression Summary Statistics

The "z" column provides the statistical significance or Wald statistic value. It resembles the ratio of the regression coefficient and the standard error, which helps to conclude how important is a feature for the model. Considering the statistics above, it can be estimated that all variables, except for *brick*, are statistically significant for the model.

The values to be observed next are in the first column – regression coefficients and in particular the sign in front of the numbers. Negative values show that the hazard is lower (in this case "risk of sell") for objects with lower values of this particular feature. For example, considering *space_m2* it can be

seen as an indicator that smaller apartments are more likely to be rent. The opposite can be seen for positive values – then is a higher probability a property to be for sale for higher price_per_m2.

The final rows of the summary present the global statistical significance of the model. The three listed p-values provide an overview of the three alternative tests for model significance. Since the data is of sufficient size, the values are similar.

However, the prediction evaluation gives a value of 0.6191472644 for RMSE and R2 of 0.03630297421. Both metrics indicate that the Cox regression model fits the data worse than the previous models.

The last model which was built is a neural network with configuration 7:3:2:1. As for the parameters, the number of hidden layers is 2 (since there are enough for a considerable amount of applications). The input layer has 7 inputs and so the number of neurons for the hidden layer was defined to 3 and 2. The output is a single neuron.

The first graph below illustrates the trained neural network. The black lines give visibility on the connections and their weights on which of the connections, while the blue lines and values represent the bias term added on each step.

The second graph represents a comparison between the neural network and the Lasso regression real vs predicted values. The closer the data points to the line, the better the model (in the best-case scenario the data points should align perfectly with the line, RMSE = 0). The visuals show that the NN has slightly more distant data point than the Lasso. Not only the graphical comparison but also the statistical results should be considered. The RMSE of the neural network is 0.06464663508, which means that the Lasso regression performs slightly better.

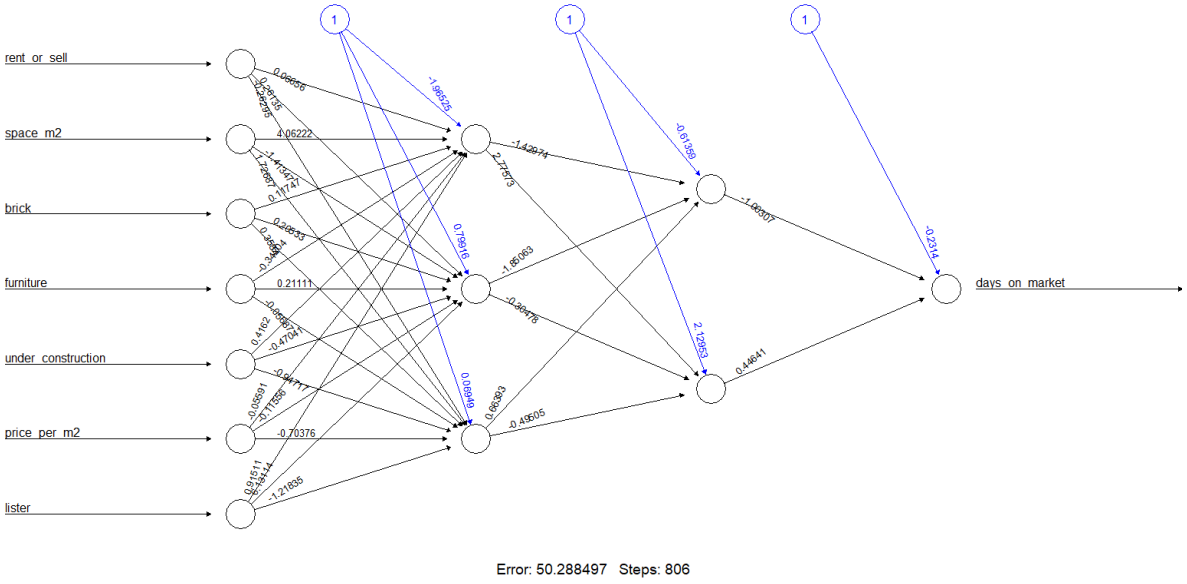

Figure 31 – Neural Network

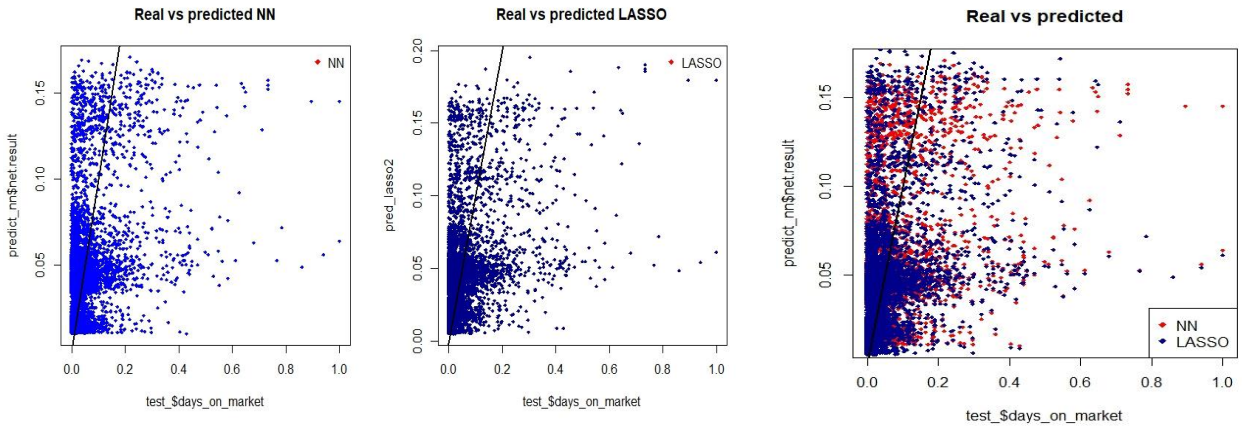

Figure 32 – Real vs. Predicted NN and Lasso

Neural networks are in general more complicated for interpretation and explanation. Considering this fact and the current application and purpose of the modelling part, the neural network's parameters are not tuned additionally, even though an assumption that slight changes can lead to better results is made. For the purpose of the project and based on the result Lasso is considered the best performing model in this case and so in the next part of this chapter an attempt to explain the model by extracting understandable insights will be done.

Usually complex, non-parametric machine learning algorithms perform really well on predictions, but they operate as black boxes for interpretability. Even though the winning model in this project is lasso regression, which is more intuitive for analysis and understanding, some tools can help to gain insights and to better interpret the model.

5.5.3. FEATURE IMPORTANCE

One of the methods suggests feature scoring by comparing their effect on prediction performance. The importance of the feature is measured by the increase in the error after permuted the variable values – this breaks into the relationship between the variable and the target variable. A given feature is considered less or not important if rearranging its values do not lead to any change in the model error. The algorithm works on several steps, starting with training a model. Afterwards, the values in a single column are rearranged and new predictions are calculated with the new data set. The error difference defines the variable importance. The same is done for every single column separately and then the features are scored.

The graph below illustrates the importance of the features, starting from the one with the highest weight in the model – lister, and gives the error measurement – RMSE, in this case, 1.068. The table afterwards gives detailed information on the results of the features importance test.

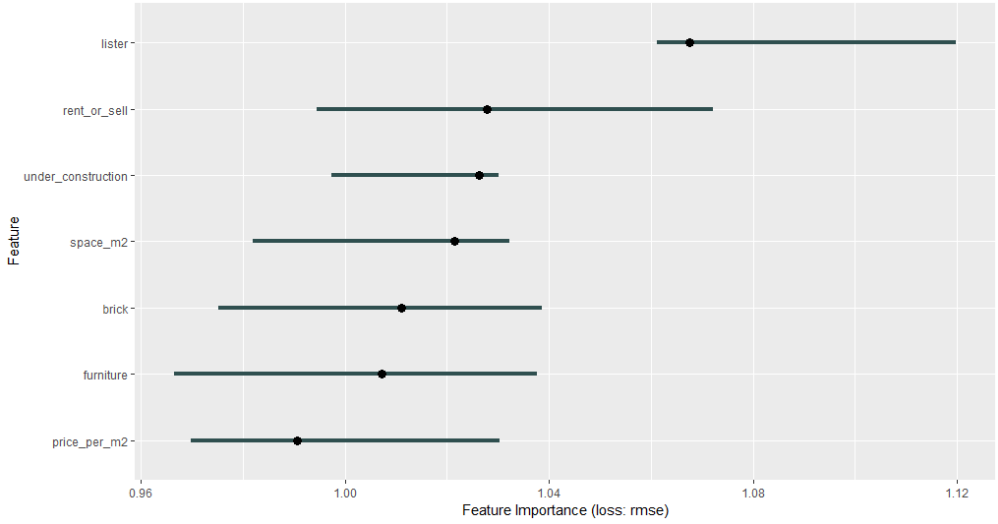

Figure 33 – Feature importance in the model

Table 11 – Feature Importance in the model

VARIABLE	IMPORTANCE.05	IMPORTANCE	IMPORTANCE.95	PERMUTATION.ERROR
lister	1.0611371	1.0676099	1.119868	0.07243278
rent_or_sell	0.9945100	1.0278727	1.072171	0.06973678
under_construction	0.9973978	1.0262011	1.030145	0.06962337
space_m2	0.9818909	1.0215203	1.032234	0.06930579
brick	0.9752772	1.0109712	1.038632	0.06859008
furniture	0.9664737	1.0071756	1.037656	0.06833257
price_per_m2	0.9698690	0.9905373	1.030250	0.06720373

This method has some positive aspects, including that it takes into account the relationship which a feature has not only with the output variable but also with all the rest of the features. Additionally, the permutation importance does not require retraining of the model to be applied but just shuffling a feature's values (Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable., 2019).

6. CONCLUSION

The present thesis project conducted research about various real estate data mining projects and procedures, in order to build fundamental domain-knowledge to optimize the AI component and customer satisfaction for a start-up company called Homeheed. The main goal was to develop a model to predict the *days_on_market* variable by applying several algorithms, in particular, regressions and neural network. The purpose of this prediction is to improve the customer experience by sorting and estimating the relevance of new income data entries in Homeheeds database. This initiative will allow Homeheed to benefit from its own data by exploring historical market data and gaining domain-knowledge to estimate the relevance of their listings, to enhance their customer experience by streamlining their market entries, as well as to support the design of the technology which can assess a property availability. To target this business problem and to structure the applied research methodology, this project declared the following research questions, which will be answered in the upcoming paragraphs:

(1) Can a machine learning algorithm predict the days-on-market variable for the housing units?

(2) Which features effectively influences the property attractiveness for the customer target?

(3) Is it possible to contribute to Homeheeds technology by identifying homes real availability in particular to support fraud detection?

The first research question (1) was researched by predicting *days_on_market* for property listings and thus enhancing the transparency on the market. The data containing properties listed by agencies were excluded due to its unreliability – great percentage of the values for DOM are unreal. The pre-processing of the rest of the data took 80% of the time dedicated to this project. The various features were investigated and transformed to identify the key factors that affect the attractiveness of a property, which resulted in the reduction of features used in the model to 7. Then several algorithms were trained and LASSO regression was selected as the winning one due to its best performance presented on a test set, namely RMSE 0.06688481. Even though good accuracy was achieved, it should be highlighted that if the model was developed on all the data, including listings from agencies, the accuracy would be significantly lower. By that, the research conducts the result that the following project was able to develop an effective machine learning algorithm, namely LASSO regression, to predict the independent variable *days-on-market* with a selection of discriminators, which will be discussed in the second research question.

The answer to the second question (2) was closely related to the findings of the first one – recognizing the features which make a property more interesting to the market. As many studies

focus on measuring the effect of factors on house price, here the point of interest measuring the effect of features on attractiveness. Based on this particular dataset, the features which have the most influence on the *days_on_market* is *lister* – who provided the property; whether it is for *rent_or_sell*; space in square meters; with furniture or not; whether the building is *under_construction* and *brick*.

The last question (3) was the ability of the data to provide a signal whether a property is fake or not and available in the time period when it is listed or not. To develop a technology which is identifying real property listings, image data is required and the features used in this project are not sufficient. At the beginning of the project, an assumption that a developed model on properties listed by owners, investors, banks, etc. can serve as fraud detection for agencies' listings was made. However, during the execution phase, an experiment was conducted and this assumption was rejected. The first reason is that *lister* is the main feature which has an effect on the prediction, but the agency's data this feature is constant so has no influence on the prediction and the results were extremely unreliable. Additionally, knowledge from the market and interviews with real estate agents stated that long *days_on_market* is not the only and enough factor to argue that a listing is unavailable or fake. In conclusion, this research neglects the assumption that the present database can be used to support fraud detection by identifying the real availability of housing units and consults to enhance the variable set of the database.

6.1. CRITICAL APPRAISAL

In this chapter, the limitations of the project will be revised and afterwards, comments on possible future improvements will be made.

One of the main limitations of this work is the data. A significant amount of the variable characters was in Cyrillic and while it was possible to translate some of them in English, others contained a major number of characters which made the automatic and correct translation impossible. For example, analysing further the full description of the listings or considering the names of the agencies/ owner who published the listings could provide not only deeper insights, but also better estimation for the *days_on_market* for properties listed by agencies and possible fraud detection based on account names.

With this work, it was manifested that predicting days on market for a property is a challenging and not only historical listings details data-dependent task to be solved. On the one hand, to provide a reliable model, the researcher needs a broad understanding of the market and its specifications on the local level. In conditions of unregulated market and lack of visibility, the proper estimation of

property attractiveness based on historical data seems to be perplexing, long-term research question. Additionally, real estate liquidity is influenced by various factors, including weather, macroeconomic and detailed geo-location (neighbourhoods, amenities, residential community profile, etc.) data. One of the studies in the area, from Hengshu Zhu, Hui Xiong, Fangshuang Tang and others, consider multiple types of heterogeneous real estate data, such as neighbourhood profiles and geo-social information of residential communities. On the other hand, it was proved that fraud detection requires properly collected data with various features which can support the estimation of the target. Furthermore, in this project, it was highlighted several times that agencies listings were excluded from the research, and while good accuracy was achieved for owners, most of the listings are published by agencies and enhancement in this area is required. Even though Homeheed has currently working technology for fake properties filtering based on image data, it still needs improvement and the consideration of additional factors due to market specifications.

In the modelling phase of this project, several models performance were tested, one of which artificial neural network. The literature recommends iterative optimization of the performance of this model, but a good result was achieved with the 2nd iteration and tuning of parameters, further modifications were not applied. For exploration and demonstration purposes, additional models could have been trained as well, showcased in similar projects found in the literature. However, in this project regression models satisfied the fulfilment of objectives and no further models were trained.

6.2. FUTURE WORK

To enhance this work, several supplementary steps can be included. In the data collection phase, which was not part of this project scope, additional data sources can be taken into account. For example, data for the neighbourhood and residential profile (schools, supermarkets, transport, etc,) can be collected and included in the research. The same is valid for other factors which influence the market. In addition, as mentioned previously, the real days on market for a property were not available and known in this data set. To assure the reliability of such project outcome, information about properties liquidity needs to be collected. This is not only time-consuming but also only long-term possible task though, as such information would be available only if it provided directly by agencies and owners. The data used here was only for one city, but a more complicated data set covering various cities and regions with their own specifications would be more interesting, not only for the growth of Homeheed, but also for the state-of-art to explore modelling techniques such as multi-task learning ANN or regression. In the light of Homeheed case, broader data collection will enrich the result and maybe provide some additional insights.

In long-term, Homeheed can consider the collection of demographics, users profile and in-app behaviour data. Such information together with macroeconomic statistics for purchasing power, banking interest rates, employment level, wage rates, etc. can provide a broader picture not only about the market, but about the factors which influences home preferences and attractiveness. Not to forget news and media data, which both can reveal interesting patterns for the customer behaviour and market fluctuations, as well as can provide some insights for the reputation of different agencies.

Another field of potential research involves the use of other machine learning algorithms noticed in similar papers, such as a k-nearest neighbour, support vector machines, random forest. As potentially most influencing features were identified, would be of interest to explore how different models with well-establish parameterization will perform. Furthermore, the training of the neural network in this work was stopped after the 3rd iteration as a good result was achieved. However, for the purpose of a deeper understanding of the model itself and the features, several more iterations could be executed in the future.

An extra line of research can be the application of classification techniques as Baldominos et al. (2018) suggest. The differentiation of different market segments in a bigger and more complex data set would be required as complexity and multi-factors dependency can lead to an unreliable result. In this case former to the modelling part, segmentation step should be added (Baldominos, et al., Nov 2018).

In conclusion, this project focused on the goal of the startup Homeheed by experimenting with modelling techniques to achieve prediction of real estate liquidity and collect insights on how market frauds can be detected. Contrary to the experiments, but not surprising in hindsight, fraud detection requires multi-level analysis, various techniques application and quantitative data, representing local real estate market complexity. Last but not least, this project added value to the study and research area of machine learning application in the real estate industry with the usage of LASSO regression.

7. BIBLIOGRAPHY

Aase, K. (2011). Text Mining of News Articles for Stock. Trondheim: Norwegian University of Science and Technology

Asaad, A. A. (2013). Measures of Skewness and Kurtosis. R bloggers. Retrieved 22nd January 2019, from https://www.r-bloggers.com/measures-of-skewness-and-kurtosis/

Baesens, B. V. (2015). Fraud Analytics Using Descriptive, Predictive,and Social Network Techniques: A Guide to Data Science for Fraud Detection. New Jersey: Wiley Ed.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, O., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. Applied Science. Retrieved 19th August 2019 from https://arxiv.org/pdf/1809.04933.pdf

Brian D. Kluger, N. G. (1990). Measuring Residential Real Estate Liquidity. New Jersey: AREUEA Journal.

BTV (2012). Bulgaria. BTV Novinite. Retrieved 1st August 2018 from https://btvnovinite.bg/bulgaria/falshivi-brokeri-zalivat-pazara-na-imoti.html

Catherine-Tucker, J. Z. (2013). Days on market and home sales. Boston: Rand Journal of Economics.

Chao Mou, Q. Z. (2018). Recommending property with short days-on-market for estate agency. Journal of Ambient Intelligence and Humanized Computing. Retrieved 1st January 2019 from https://doi.org/10.1007/s12652-017-0508-2

Cox, D. R. (1972). Regression models and life tables. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. (1972), p. 187-220. London: Royal Statistical Society.

Cran (n.d.). Cran. Cran.com. Retrieved 14th July 2019 from https://cran.r-project.org/

Deng, (n.d.). Chapter 7: Feature Selection. Carnigie Mellon School of Computer Science.

Despa, S. (n.d.). What is survival Analysis? Cornell Statistical Consulting Unit. Retrieved 4th September 2018 from https://www.cscu.cornell.edu/news/statnews/stnews78.pdf

Eddie C.M. Hui, J. T. (2010). Marketing Time and Pricing Strategies. Journal of Real Estate Research. Retrieved 4th Febraury 2019 from http://pages.jh.edu/jrer/papers/pdf/forth/accepted/marketing%20time%20and%20pricing%20strategies.pdf

Ermolin, S. V. (2016). Predicting Days-on-Market for Residential Real Estate Sales. Department of Computer Science Stanford University. Retrieved 5th January 2019 from http://cs229.stanford.edu/proj2016/report/ermolin_predicting_Days_on_market_for_Residential_Real_Estate_Sales_report.pdf

Fonti, V. (2017). Research Paper in Business Analytics: Feature Selection with LASSO. Amsterdam: VU Amsterdam.

Funda Gunes, S. I. (2015). Penalized Regression Methods for Linear Models in SAS/STAT. SAS Institute. Retrieved 12th December 2018 from https://support.sas.com

Gareth, J., Witten, D., Hastie, T., & Tibshiran, R. (2014). An Introduction to Statistical Learning: With Applications in R. Heidelberg: Springer Publishing Company.

Guo, T. A. (2008). Neural Data Mining for Credit Card Fraud Detection. In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. Kunming, p. 12-15. Frankfurt: Goethe University of Frankfurt.

Gurney, K. (2004). An introduction to neural networks. London: University College London (UCL) Press.

Guyon, I. & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, p. 1157-1183. n.d.: Journal of Machine Learning Research.

Hastie, T., Tibshirani, R. & Jerome, F. (2016). Model Assessment and Selection. The Elements of Statistical Learning, p. 219-230. Heidelberg: Springer Science and Business Media.

Hengshu Zhu, H. X. (2016). Days on Market: Measuring Liquidity in Real Estate Markets. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 393-402). Beijin: KDD '16 Press. doi:10.1145/2939672.2939686

Belkin, D. D. (1976). An Empirical Study of Time on Market Using Multidimensional Segmentation of Housing Markets. n.D.: Real Estate Economics.

Jud, G. D. (1996). Time on the Market: The Impact of Residential Brokerage. Journal of Real Estate Research. Retrieved 11th November 2018 from https://libres.uncg.edu/ir/uncg/f/D_Winkler_Time_1996(MULTI%20UNCG%20AUTHORS).pdf

Kannan, D. S., & Gurusamy, V. (2014, October). Preprocessing Techniques for Text Mining. *Proceedings of RTRICS* (pp. 1-6). India: Research Gate. Retrieved from https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining

Kriesel, D. (2005). A Brief Introduction to Neural Network. Dkriesel.com. Retrieved 11th February 2019 from http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-2col-dkrieselcom.pdf

Laerd Statistics. (n.d.). Take the Tour: Statistical Tests: One-Way Anova. Leard Statistics. Retrieved 4th January 2019 from https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Marr, B. (2017). Data Strategy: How to profit from a world of big data, analytics and the internet of things. London: Kogan Page Limited.

Miller, N. (1978). Time on the Market and Selling Price. San Diego: Real Estate Economics.

Molnar, C. (2019a). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Christopher Molnar. Retrived 11th December 2018 from https://christophm.github.io/interpretable-ml-book/

Molnar, C. (2019b). Model-Agnostic Methods. Feature Importance. In C. Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Christopher Molnar. Retrived 11th December 2018 from https://christophm.github.io/interpretable-ml-book/feature-importance.html

Moore, D. F. (2016). Applied Survival Analysis Using R. In D. F. Moore, Applied Survival Analysis Using R. Heidelberg: Springer.

NCSS Statistical Software. (n.d.). Data Analysis: Procedures. NCSS Statistical Software. Retrived 11th December 2018 from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Multiple_Regression.pdf

Ostertagova, E., Ostertag, O., & Kovac, J. (2014). Methodology and Application of the Kruskal-Wallis Test. Applied Mechanics and Materials, p. 115-120. n.d.: Scientific.net.

Paasch, C. A. (2008). Credit card fraud detection using artificial neural networks tuned by genetic algorithms. Hong Kong: Hong Kong University of Science and Technology.

Robinson, D., & Silge, J. (2017). Chapter 4. Relationships Between Words: N-grams and Correlations. In D. Robinson, & J. Silge, Text Mining with R, p. 45-67. London: O'Reilly Media.

Rojas, R. (1996). Neural Networks. Neural Networks. Berlin: Springer Verlag.

Seltman, H. (2015). Exploratory Data Analysis. Experimental Design and Analysis. Carnegie Mellon University. Retrieved 28th January 2019 from https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

Spijker, A. V. (2014). The New Oil: Using Innovative Business Models to turn Data Into Profit. New Jersey: Technics Publications.

Spuler, M., Sarasola-Sanz, A., Birbaumer, N., Rosenstiel, W. & Ramos-Murguialday, A. (2015). Comparing Metrics to Evaluate Performance of Regression Models for Decoding of Neural Signals. Tuebingen: University Tuebingen.

Stoykova, P. (2018). Основни показатели за жилищния пазар в България през 2018: Bulgarian Properties. Retrieved 2nd November 2018 from https://www.bulgarianproperties.bg/novini-za-imoti/pokazateli-imoten-pazar-2018-7555.html

The Pennsylvania State University (2018). STAT 501: Lesson 4: SLR Model Assumptions. The Pennsylvania State University. Retrieved 4th October 2018 from https://newonlinecourses.science.psu.edu/stat501/node/277/

Thomas, A. (2017). An introduction to neural networks for beginners. Adventures in Machine Learning. Retrieved 2nd November 2018 from https://adventuresinmachinelearning.com/wp-content/uploads/2017/07/An-introduction-to-neural-networks-for-beginners.pdf

Vasilev, V. (n.d.). Reportages. Gospodari. Retrieved 2nd October 2018 from https://gospodari.com/

Vermeylen, F. (2005). Censored Data. Cornell Statistical Consulting Unit. Retrieved 2nd October 2018 from http://www.cscu.cornell.edu/news/statnews/stnews67.pdf

Vijayarani, D. S., Ilamathi, M. J. & Nithya, M. (n.d.). Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks, p. 7-16. Tamilnadu: Bharathiar University.

Xian Guang LI, Q. M. (2006). The Applicaition of Data Mining Technology in Real Estate Market Prediction. Fraunhofer Information Center for Space and Construction IRB. Retrieved 5th October 2018 from https://www.irbnet.de/daten/iconda/CIB5807.pdf

Yoskovitz, A. C. (2018). Lean Analytics: Use data to build a better StartUp faster. London: O'Reilly Media.

Zheng, A. (2015). Evaluating Machine Learning Models: A beginner's guide to Key concepts and Pitfalls. Boston: O'Reilly Media.