PREDICTING HOUSE PRICES IN ANKARA USING MACHINE LEARNING

CİHAN ERSOY

DECEMBER 2022

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING
MASTER'S THESIS IN
INFORMATION TECHNOLOGIES

PREDICTING HOUSE PRICES IN ANKARA USING MACHINE LEARNING

CİHAN ERSOY

DECEMBER 2022

ABSTRACT

PREDICTING HOUSE PRICES IN ANKARA USING MACHINE LEARNING

ERSOY, Cihan
Master of Science in Information Technologies

Supervisor: Assistant Professor Dr. Serdar Arslan
December 2022, 47 pages

The focus of this thesis is to investigate whether machine learning predictions are accurate and viable enough to replace traditional real estate appraisal reports. To do this, we compare two datasets, one scraped from a real estate website and the other created from appraisal reports, and use various machine learning and neural network methods to find the best performing one and to determine the practicality of the study. Bagging and boosting ensemble methods are compared with the implementation of Extreme Gradient Boosting and Random Forest Models. Also, an Artificial Neural Network with five layers and Relu activation function is built as well as ensemble learning models. Hyperparameters of all models built throughout the study are chosen diligently for a comprehensive comparison. We evaluate the success of the models using root mean square error and accuracy score. Findings suggest that this approach has potential for improving the real estate valuation process, but further research is needed to determine its viability in the real world.

Keywords: House Appraisal, Machine Learning, Random Forest, Artificial Neural Network

ÖZ

MAKİNE ÖĞRENMESİ KULLANARAK ANKARA'DA EV FİYATLARI TAHMİNİ

ERSOY, Cihan
Bilgi Teknolojileri Yüksek Lisans

Danışman: Dr. Öğr. Üyesi Serdar Arslan
Aralık 2022, 47 Sayfa

Geleneksel gayrimenkul değerleme süreci, bir değerleme uzmanının gayrimenkulü görmesi ve evin sahip olduğu değerlere göre içerisinde ev fiyatının da bulunduğu bir rapor oluşturması üzerine kuruludur. Ancak, bu yöntem zaman alıcı ve yüksek maliyetli olarak nitelendirilebilir. Makine öğrenmesi, bu süreci hızlandırmaya ve maliyetleri azaltmaya yardımcı olabilecek bir araçtır. Bu nedenle, bu tezde amacımız, makine öğrenimi tahminlerinin ev fiyatı değerlemesi sürecinde gerçekçi ve yeterli olup olmadığını araştırmaktır. Bu amaçla, çalışmada biri gayrimenkul web sitesinden toplanmış, diğeri ise değerleme raporlarından oluşturulmuş iki veri seti çeşitli makine öğrenimi yöntemleri kullanılarak karşılaştırılmaktadır. İnşa edilen tüm modellerin hiper parametreleri dikkatli bir şekilde seçilmiş, modellerin başarısı ise kök ortalama kare hatası ve netlik skoru kullanılarak değerlendirilmiştir. Bulgular, yaklaşımın varolan değerleme sürecini iyileştirme potansiyeline sahip olduğunu, ancak uygulanabilirliğini göstermek için daha öteye araştırma gerektiğini öneriyor.

Anahtar Kelimeler: Gayrimenkul Değerleme, Makine Öğrenmesi, Karar Ağaçları, Yapay Sinir Ağları

ACKNOWLEDGEMENT

Special thanks to my supervisor Dr. Serdar Arslan for the excellent guidance and providing me with an excellent atmosphere to conduct this research.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION AND OUTLINE OF THE STUDY

Information Technology plays a key role to shorten and improve business processes. Various steps in business processes that were conducted manually before have been transformed into an automated way thanks to information systems. Checking a client's credit in banking or preparing an invoice in retail sector are some fundamental examples of these kind of implementations. [1]

In this study, home appraisal process is subject to automation. Using machine learning on a relevant data set to set a realistic value for a house is expected to be faster and less prone to vulnerabilities as it would involve less human intervention. Accordingly, this study asks the question how accurate would it be to use machine learning models to predict the real value of a house rather than appraisals report. To be precise, accuracy and viability of machine learning is discussed with its all dimensions to decide if it is all set to replace appraisals report in a standard mortgage loan process.

Machine learning, which is the technique used in this thesis study, has enabled researchers and practitioners to implement complex mathematical calculations easily in order to predict of a real estate, commodity or stock prices. Among these prediction works, many of them have appeared to be focusing on house prices. Because there are many parties who are interested in an accurate prediction of a house such as buyers, sellers, investors, banks, local governments etc., different approaches are much needed to satisfy all kind of market needs. Investors need the current value of a real estate as well as rental value to make a wise investment. Local governments require house prices to set taxes. Banks need the current value of a house to make a reasonable decision for mortgage applications. This study aims at meeting financial organizations' needs, specifically banks, rather than local governments or investors.

While there has been much research on technical background of a price prediction, fewer researchers have taken the organization factor into consideration and

adapt the conclusion for the real world. In this study, the prediction of a machine learning model and the appraisal of an expertise report are compared to understand if the former is accurate enough to replace latter. To elaborate, two data sets for those two cases have been created meticulously from different resources, analyzed and then compared eventually regarding the 13 features of houses.

Adding a further point, although academic studies of price prediction are mushrooming, there are still some roadblocks which prevent researchers from doing studies about it. Difficulty in acquiring relevant data appears to be a major impediment for enthusiasts at this stage of the game. Companies that own the data make it quite hard for others to scrape them for commercial concerns, and data businesses set the price bar notably high beyond an individual's affordability. Besides all these, data businesses are not always available in developing countries.

Before we get into much detail, describing a standard mortgage loan process would be practical as our ultimate goal is to reduce business steps involved in it by automating house evaluation by appraisers. Mortgage loan process starts when a customer that has intention of buying a house comes to the bank. It is expected from the customer to have an acceptable credit score by meeting some other financial obligations of the bank. Although some details may change from bank to bank, the main principles are as follows: The amount that can be taken out as mortgage credit from the bank is up to 80 percent of the value of the house which is intended to be bought. When it comes to installments of loan repayment, it cannot be higher than 70% of one's (or households if applicable) salary. If customer meets the requirements, then bank sends a house appraiser to see the house and set a price for it. It takes usually more than one day for an appraiser to see the house and prepare a report which includes the details. Although house appraisals report is quite exhaustive, loan officer has only time to check if price is high enough for the demanded credit and the condition information of house. House appraisal is the most demanding process of all in the course of a mortgage loan and if it can be automated the whole process could be shorten one or maybe even two days.

1.1.1. Personal Motivation

Emerging new technologies and tools draw lots of attention by most. Same practices are implemented by enthusiasts over and over again in order to comprehend it and gain first movers' advantage. However, most of these technologies gain meaning

and find its potential once they are used by those who already have some domain knowledge and those who try to solve a specific problem in a certain domain. Machine learning, too, is said to enjoy huge potential if people with domain knowledge use it. I have been positively encouraged by these reflections of experts as an ex-banker.

It is a common knowledge that to work in the desired area in your thesis, first you need to acquire data in that specific field. Without the data, your desires do not take you long. Data scraping skills of someone helps a lot on this, and hearten to collect your own data for your study. I consider myself lucky to gain this skill beforehand, which, I believe, added much value on this research. On top of that, trying to solve a problem that you already encountered gives you an additional motivation which one would need a lot throughout the whole research process.

1.1.2. Research Aims and Objectives

As stated in the introduction part, this research aims to shorten business processes in banks when a mortgage loan applicant needs a house appraisal with regard to his or her mortgage application by replacing conventional valuation methods with state-of-the-art Machine Learning models. Whole process is described below in detail:

An expertise report is an inseparable part of a mortgage process. Once a customer applies for mortgage to buy a house, bank wants to know the value of the house to prevent fraud and determine the limitations of the credit. Expertise reports are prepared by an independent organization to establish reliability, and mostly slow down the whole process. Bank checks mainly two things in this report: first whether the house is in a good condition to sell, and second the value of the house. Therefore, eliminating the second criterion in this process by automating the home appraisal stage can lead to shorten the required time and reduce the expenses. This standpoint establishes the main objective for this study.

To be able to eliminate the process, viability of the suggested automation also matters as well as the statistical success in the prediction. Therefore, viability has been studied thoroughly to cover all angles to this problem. Besides, other factors such as which characteristic of a house affects most on a house price, or popularity of towns have been investigated as well in order to give a clear comprehension of this domain to reader within the scope of research aims.

1.1.3. Structure of the Thesis

The outline of the whole thesis is given in this section to get readers familiarized the course of chapters. The following parts contain the details below.

Chapter two is the literature review section, related researches that have been conducted in this area so far are examined. Different approaches and similar standpoints are discussed to give a comprehensive insight along with necessary fundamental knowledge for following chapters.

Third chapter is the methodology part; it presents the underlying reasons for methods chosen. Statistical knowledge that is required throughout the study is also explained here. Boundaries of data cleansing implementation and the study itself for that matter was analyzed within the scope of this dissertation. Tools and software beginning from the data collection up to machine learning evaluation took the well-deserved credit in this chapter, too.

The result has been explained in chapter four, it explains at what rate the objective of the study is covered. Whether the findings of the research are sufficient to accomplish goals and also some real estate market analyses that has been studied on the data set. Also, features that have the biggest effect on setting the price of a house are presented as well.

Finally, last chapter is reserved for the discussion part, where conclusions are confronted and idea for further studies are specified. The data set scraped from web portal and data set that involves appraisal reports are compared to comprehend if the aim of the study is practical yet to be implemented.

CHAPTER II

LITERATURE REVIEW

A remarkable number of studies have been done on house prices over the course of many years. While some of them implemented statistical methods and artificial intelligence on the purpose of utilization of new technical approaches, others examined financial bubbles and inflations from economics standpoint. Apart from all these, urbanization, purchasing power etc. have been other subjects in this regard. However, this research investigates if house price prediction which is generated by using machine learning can be applied to an organization in real life from information technology perspective.

Various different approaches exist to estimate the price of a residential property. One of those popular approaches is hedonic price model. Hedonic price model regards a house as aggregation of individual attributes and it has both advantages and drawbacks. [2] One of the main advantages of HPM is that it is relatively easier to have a better grasp of relationships between inputs and outputs. [3] Non-linearities problem, on the other hand, can be counted as the major disadvantage for it. Besides, due to the difficulties of linear approach such as constrictions of a generic prediction, other approaches have been investigated further to handle the non-linearity problem. [4]

In 2010, a new dimension was added to the game as distinct from the available ones, fuzzy logic. A questionnaire was set by Ozdemir and his friends in their study and taken it into account along with attributes of the house to calculate the estimated value, believing that even salesman and regional aspects play a vital role in this process. Hence, they divided all house characteristics into 4 categories: house, environmental, transportation, and regional/socio-economic aspects. Then, they investigated all factors which have an impact on the house value in detail. [5]

Another approach for a price estimation task is machine learning. Machine learning systems learn from data and consist of 3 main steps: representation, evaluation and optimization. Primary objective of machine learning is to generalize what have

been learned from training data set. [6] Machine Learning has been used many times by researchers with regard to prediction problems which is not all about residential commodities but also car, vinyl or stock prices.

Table 1: Price classification based on price ranges

From	To	Class
500	2000	500-2000
2000	3500	2000-3500
3500	5000	3500-5000
5000	6500	5000-6500
6500	8000	6500-8000
8000	9500	8000-9500
9500	11000	9500-11000
11000	14000	11000-14000
14000	17000	14000-17000
17000	20000	17000-20000
20000	25000	20000-25000
25000	30000	25000-30000
30000	60000	30000-60000

Car price prediction conducted 2019 is also worth mentioning at this stage. 1105 car samples were scraped from the web for car price prediction task. [7] Like this study and all other similar studies, researchers encountered a messy data set that needs cleaning. After all cleaning work, the size of their data set was reduced to 797 samples. They removed brands that had less than 10 samples, also they eliminated cars that were priced above a particular limit because of the skew class problem. Price prediction is a regression task by its nature. However, this car price prediction task was transformed into a classification problem by creating price ranges as stated in Table 1.

Three machine learning models were built on the car prices data set namely Support Vector Machine Model, Artificial Neural Network and Random Forest Model. 90% of the data was utilized for training and 10% was to test for all those 3 models. Features of the data set were as following: model, brand, fuel, age, car condition, transmission, kilowatts, miles, doors, drive, color, leather seats, navigation, alarm, aluminum rims, digital AC, manual AC, xenon, remote unlock, seat heat, parking sensors, cruise control, abs, panorama roof, asr, esp and price. Their model achieved 92.38% accuracy, and they attributed this success to rigorous data cleaning work and applying multiple models rather than a single one.

To be more elaborative on the conclusion of their prediction, Support Vector Machine performed well with highest accuracy score for relatively cheap cars and most expensive ones, whereas Artificial Neural Network worked best for moderate cars. It is understood from their study that they used a voting classifier on those 3 models by combining them and increased the performance beyond all 3 single accuracy scores. [7] Although the final performance of the work seems to be satisfactory, we need to keep the initial transformation operation in mind before we acknowledge the performance of the work to be right.

Another study that was subject to machine learning implementation in real estate context was conducted for the houses in the Salamanca region of Madrid. [8] They tried wide range of models, namely k-nearest neighbors' model, regression trees, and support vector machine in their study but they indicate in the conclusion part of their research that best performing results are ensembles of regression trees. They also emphasize the superiority of more complex models to classical linear regression models.

Mean absolute error and median absolute error was focused in the study to evaluate the model rather than r score. They found the smallest mean absolute error score as 338,715 euros and the best performing number is 94,850 euros for median absolute error evaluation. Researchers justified the relatively high error scores by stating the fact that only houses priced over one million were subject to study. Besides, they compared the mean and median of the distribution of prices to mean and median absolute errors and calculated 16.80% and 5.71% relative errors respectively. The significant difference between mean and median was explained by the researchers saying that presence of outliers.

Artificial neural network, which copies the learning process of a human brain, is one of the main state-of-the-art concepts in terms of prediction tasks. Limsombunchai ‘s study that compares hedonic price model and artificial neural networks has significant amount of insight to have a better grasp of it. ANN contains three fundamental layers which are input data layer (attributes of the house in this case), hidden layer/layers, and output layer. Price prediction is brought by the output layer of the model. The strength of artificial neural network comes from its trial-and-error strategy. [9]

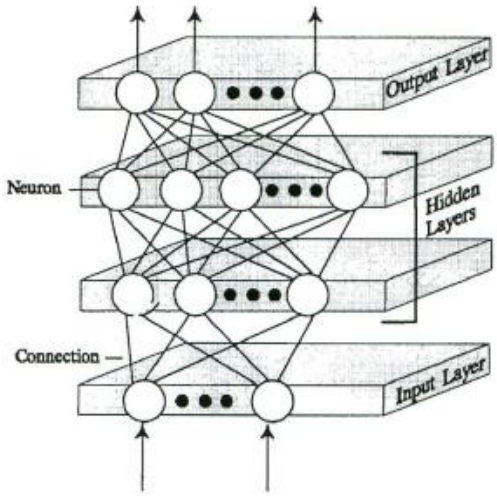

Figure 1: Feed Forward Neural Network Structure with Two Hidden Layers

The figure above visualizes the fundamentals of a neural network structure. Each computational unit (or as known as neuron) in the figure has some input connections which enables getting signals from other neurons. Besides, a set of weights are utilized for bias adjustment as well as a transfer function which rebuilds the total amount of weighted inputs and bias. [10]

Limsombunchai claims the process of neural network model is not much different than the process of hedonic price model, yet a data set for training is needed for neural networks. He compares artificial neural networks to hedonic price model by using 200 houses and their attributes. Artificial neural networks outperform hedonic price models in his study. He concludes that potential reason for poor performance of hedonic price could be the lack of some environmental attributes and non-linear relationship between house attributes and house price.

Analyzing the early days of this breakthrough would help comprehending the foundation from beginning till now regarding the conceptual differences between ANNs and conventional methods as well as ANN implementation then and now. Early researches, such as the one by James and Lam had the false impression that of neural networks appear to be a better model for smaller data sets, whereas regression performs superior for larger data sets in their study. However, no direct relationship has been found between the size of a data set and model choice in coming years. At the time of this study, a model that was built on a data set with 223 observations and 42 features took 23 hours to establish. Even then, because of an unfortunate power outage the model did not end as it was planned. [11]

Thanks to the recent increase of the popularity of artificial intelligence, many other researchers have got their hands dirty to contribute to this area with more accurate and well put results. Neural Network researches on house price estimation are worth examining to experiment their high performance on real estate market. A study conducted in 2016 by Wang and his friends expressed that neural network model successfully maps the non-linear relationship between price and attributes, in other words inputs and output. The delayed neural network model was built to estimate the trends of the resale price index for Singapore houses. They came up with a conclusion that delayed neural networks are capable of generating a convincingly high performance in their research.

Table 2: Experimental Results of the Relevant Study

ANN Architecture	Distribution Ratio	Average MSE	Overall R-value
6-8-1	60:20:20	3.882	0.9554
6-8-1	60:20:20	39.90	0.9062
6-10-1	60:20:20	21.46	0.9209
6-15-1	70:15:15	13.03	0.9579

Time was one of the features in that study due to having a historical data. Other than that, both the details of house and also economic variables were used for the estimation task. Ten economic variables, namely Singapore Real Gross Domestic Product, Population, Unemployment Rate, Avarage Monthly Wages, Labour Cost, Straits Time Index, Prime Lending rate, Interbank Rate, Singapore Customer Spending, Singapore Customer Price Index (CPI) were put in the model with the intention of finding a correlation with output. Throughout the model building process, many different architectures of ANN and different distribution ratio between training, validation, and testing were tried and results are in Table 2.

According to the researchers, the limitations of the above-mentioned study is having inconsistent results. It was also concluded that initial conditions of network were completely different for each training attempt having no rules or formulas. N-fold cross validation and trying different initial weights may result in better performance and worth experimenting the writer suggests. [12]

In addition to all these; some studies focused on real life applications of forecasts in this regard. No matter how accurate the result of the forecast is if the idea behind the research is not viable to real life scenarios, then there is a considerable

amount of risk that all efforts to make the model better will have little to no effect on improving business processes. From this perspective, reviewing a couple of research related to direct comparison between a price prediction result and real estate appraisal's will be worth mentioning.

In a study that had 288 sales of single-family residential properties in Colorado, a Neural Network model implemented to calculate an educated guess for house prices. Researchers obtained the data set from Fort Collins (Colorado) Board of Realtors' Multiple Listing Services (MLS). They trained their model with 217 houses and tested it with the rest of the data, which contains 71 houses. First thing that attract one's attention regarding the difference of prediction implementations then and now is the size of data sets. Thanks to ever-increasing amount of data in the world, current researches have a better chance to get a higher accuracy score in comparison to ones from the previous decade. Another significant difference is the method of acquirement of data set. This also indicates the main difference in approach for the problem. The data that already exist on the internet can suffice to solve the problem of determining the right valuation of a real estate from my point of view. However, using previous appraisal's reports does not fit for purpose in this regard considering the fact that ultimate goal of this study is to be able to apply the study to real life.

Worzala found inconsistent results from many different aspects and warned appraisers who plans on using neural network in their routine workflows. Their approach seems more appraisers-centric in contrast to this study which aims at transforming house appraisal process into a more data–driven way. In conclusion part of their study, they explain that although neural network models outperform multiple regression models in some of the cases, difference is not either large or convincing to conclude something meaningful. [13]

Another innovative method for real estate valuation was introduced in 2016 by Eduard Hromada. Model, which is called historical market price, was based on mathematics, statistical and database-founded algorithms. Data come from a specialized software which was fed more than 650.000 real estate price offers every six months. Yet, the method suggested by him only used exact information gathered from the last purchase. Besides, the paper suggest that same method could be used for determining the rent of a real estate and for calculating the market value even for past dates.

On the other hand, limitations of HPMs were a concern for them to overcome, one of which was trustworthy data. [14] It is obvious to an intelligent layman that without a trustworthy data set, even the best model will fail regardless of the structure of it let alone cross validation or the chosen hyperparameter tuning method. With the researcher's concerns in mind, it can be said that data quality is always a limitation for any model. As any data set or research, for that matter, is contaminated with typos or intentional faulty inputs, the accuracy of the model will fall down. That's also the fundamental reason why data cleansing the first and foremost step to do when it comes to increasing the performance before adjusting the structures, parameters or hyperparameters of the model.

CHAPTER III

METHODOLOGY

Main goal of this study is to experiment reducing workload in a mortgage loan process to help both financial institutions and home-buyers by eliminating some cumbersome steps involved. Therefore, the question we try to find answer is if we can replace appraisal's report with machine learning predictions. To be able to do this, we needed to compare machine learning price prediction of a house and the price determined in appraisal's report to understand if former is accurate enough to replace latter.

As clarified in various academic studies, machine learning systems automatically learn programs from data. Therefore, first we needed some data to build a machine learning model on. Then, which learning algorithm to use should be decided. There is a wide variety of algorithms around, and choosing the right one is a daunting task. Although ensemble learning methods appear to have an advantage over conventional methods, there still is a stiff competition between neural network approaches and ensemble learning methods. In this study, random forest method was utilized together with some complementary techniques and also along with other models, namely cross validation, hyperparameter optimization, Linear Regression, Support Vector Machine, Extreme Gradient Boosting, and Artificial Neural Network.

Throughout the study, although the structure of machine learning models has been examined and explained, the major focus of the study was kept within the scope of research question, which is whether machine learning predictions are accurate and viable enough to replace appraisal's report. For this reason, evaluation metrics of the model is not only criterion for the success of the research, but viability also matters to reach out the goals of the study. Therefore, increasing the accuracy score is an important constituent for this study in that regard but not the only one, and left for further studies if not found sufficient.

3.1. DATA COLLECTION

Due to the fact that machine learning models are built on a data set, there needs to be one for this study as well which involves price and some other details of it. A proper way to collect this data was to get it from a real estate website. However, collecting a data set that has house details all across the country was not an effective method because second data set involving appraisal report details has only houses in Ankara. Hence, location-related restrictions and some other filters were applied during the data collection process.

Figure 2: Location Information of Observations in the Data Set

Although collecting the relevant data from the web is a painful stage with its all difficulties, a workflow has placed below to show the main steps involved. Houses without price value was left behind due to the fact that it is chosen as the dependant value and it cannot be null. Apart from that, some other variables such as room and maintenance fee etc. were cleansed in the course of scraping to reduce the workload.

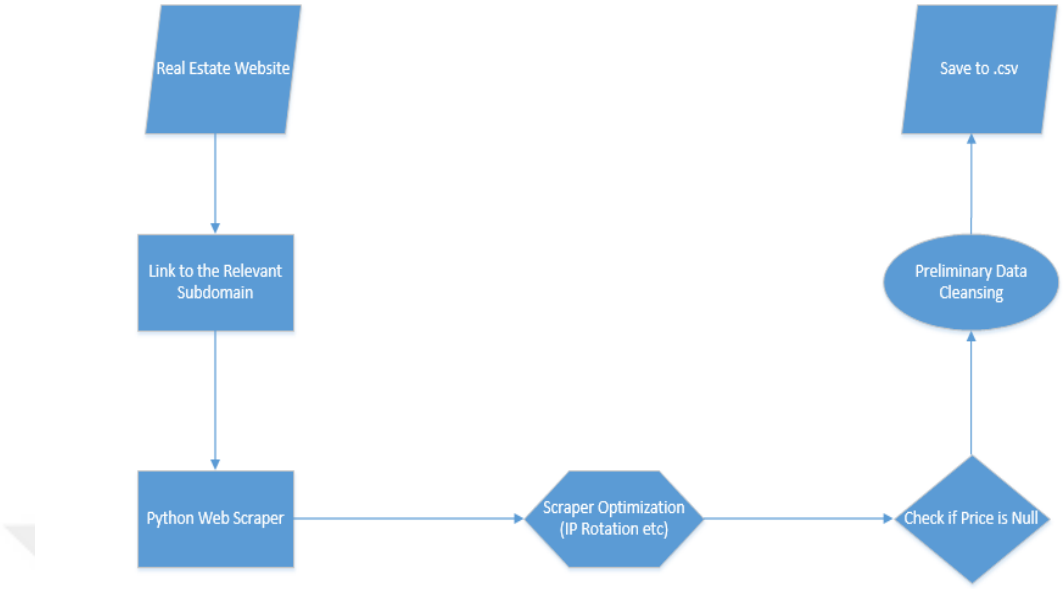

Figure 3: Data gathering and transformation workflow diagram

After the data scraping process, data set collection process was completed in October 2021 from the real estate website. It had 20.381 rows and 15 features in the first place. Column headers and the explanation of it is given in Table 3. This is the first version of data set, another column which shows if house has terrace was added to it afterwards along with some column drops which explained in the coming paragraph.

Table 3: Data Set Columns' Explanations

Column Header	Explanation
Age	How old the building is
Area	The area of the flat
CreditAvlb	Is it eligible to take out mortgage loan?
Firsthand	Has the flat ever been settled before?
FloorCount	How many floors are there in the building totally?
FloorNumber	The floor number of the flat
Orientation	The direction that flats face
Fuel	What kind of fuel is used
Furniture	If the house is for sale with its furniture or not
Heating	What appliance is used to heat the flat
MaintanenceFee	The amount of fee that building applies to each flat
Room	Total number of rooms in the flat
Town	Location information of the building
Typeofit	Is it a flat, or a detached house
Price	Amount in the advertisement

Price feature was set to be dependent variable of the study and other features are utilized as independent variables to calculate the prediction of price feature. As this research was planned to be conducted only on flats rather than detached houses or so, Typeofit column which has only one different value was dropped. CreditAvlb column which shows the eligibility of the house for mortgage loan was dropped after eliminating ineligible houses from the data set.

3.2 DATA CLEANSING

Data cleansing is an inseparable part of a Machine Learning implementation. Because having a high accuracy score plays a vital role in this kind of researches, beginning with a well-prepared data set is considered to get a head start for this objective. In the light of these; houses with unrealistic details were detected in the first place for cleansing. Then, houses that cannot be subject to mortgage loan because of their age or ongoing construction.

On the one hand, having more data results in higher accuracy. Therefore, dropping as few rows as possible should be the aim. On the other hand, there are some houses that are not eligible for mortgage credit such as houses that are older than 40 years old.

Floor count feature is kept between the range of 1-50 due to the fact that the highest residential building in Ankara has 50 storeys. Higher values in floor count distribution are also examined diligently in case of inconsistent rows.

Houses whose price are lower than 40.000 TL or higher than 10.000.000 TL are also examined further for data cleansing because they are more like to be subject to typo for example a missing or an extra trailing 0.

Area column is also examined with the extreme values of 10 and 750. For instance, 10 square meter house having more than 5 rooms was subject to cleaning. Inconsistent samples are removed diligently.

3.3 TOOLS AND SOFTWARE

Mentioning the tools and software that have been utilized throughout the research is necessary and useful at this point. First of all, scrapy library of Python is applied to collect the first data set. An interval was set into the automation script in order not to do any harm to website during the process. Second one is pandas library. Pandas is go-to tool in terms of any exploratory data analysis task. It was used in the scope of this research for this purpose as well. Apart from these, sklearn and matplotlib were used for machine learning implementation and visualization respectively. All the scripts from these libraries implemented on Jupyter Notebook. As stated earlier, a second data set was created for appraisal's report details and google forms came to our rescue for that task. While training the machine learning models, Lenovo v55T AMD Ryzen 5 3400G 16GB 512GB SSD with the Cinnamon edition of Mint Distribution of Linux Operating System has been utilized.

3.4 MACHINE LEARNING MODELS USED IN THE STUDY

Recent studies show that ensemble learning methods outperform other old-school methods such as decision tree or linear regression vast majority of the time. Therefore, decision to be made at this stage appears to be which ensemble model should be used. XGBoost, Random Forest, Voting Classifier are all available options to be utilized. Although random forest is not considered as state-of-the-art anymore, it is still in the game when it comes to an acceptable rate of accuracy. Besides, the ultimate goal of this study is not to find the best performing model by comparing them but instead examining if machine learning ready to supersede conventional business processes. Accordingly, effort was made to discuss its practicality and present solutions as well as accuracy score. Be that as it may, a necessary outline of random forest model that has been utilized in this research is given below to complete a knowledge base to assess the research thoroughly.

As stated in Analysis of a Random Forests, Random Forest models are a scheme proposed by Leo Breiman in the 2000's for building a predictor ensemble along with a predefined number of decision trees which exist in a random generated subset of data.

To build a random forest model, we used the required knowledge below and well-adjusted hyper parameters for these concepts:

The number of trees in the random forest: More trees may result in a higher accuracy, however the probability of facing an overfitting problem increases together with tree numbers.

The max depth of each tree: Fine tuning max depth of trees is also vitally important. To balance bias-variance trade-off well, you have to choose this value carefully.

The minimum number of leaves in each tree: The minimum number of samples required to be a leaf node is determined as well. min_sample_leaf syntax is used to assign a value. Potential range of this parameter is mostly 1 to 10. Smaller values could lead to overfitting with high variance while greater values are most likely to underfit suffering from high bias low variance.

The minimum number of splits in each tree is determined with the help of min_samples_split syntax. It stands for the least number of samples needed to split an internal node. This can differ greatly whether one sample at each node to use or all of the samples at each node to use is better choice.

Hyper Parameter Tuning: All above-stated hyperparameters have an optimal value in a specific random forest model. Tuning them cautiously is key to success of the model with a high accuracy from an outcome-oriented point of view.

It is a well-known fact that all machine learning models have its own advantages and drawbacks. For instance, some of them are easy to tune hyper parameters, others have more complex structure and hard to digest all details. For this reason, all models are executed on the same data without tuning too many hyperparameters. Then best performing model is found and most of the effort is made to tune that model, which is Random Forest.

Although Random Forest stands out the best performing model among others and compared to appraisal report, there still are some other solid work in the course of machine learning model building. Result section covers all details in this regard to answer the question of why random forest model is chosen to compare to traditional

house appraisal methods. Also, it is important to mention that model tuning experiences that was gained when building one model was used to improve others, too. For example, if a specific number for depth of a tree-based model shows strength, then it is tried on another tree-based model to measure whether it can be generalized for both models or not.

3.5 LIMITATIONS OF CHOSEN APPROACH

A price prediction implementation without time series may seem to be lacking. One can easily detect the gap here which needs to be filled from a time-wise point of view. Relevant data is scraped, then a machine learning model is implemented on those data, and it is ready to productize for today. However, there still is a significant amount of work to make this more sustainable and age-proof. Machine Learning is fed from data and when data gets outdated due to a notorious high inflation rate, this would be a serious problem. Here appear two solutions to this issue: One is applying inflation rate to predicted price for the relevant timeline. The other is renewing data set frequent enough so that inflation is not a concern for available data. The latter resonates with the outputs of this work more.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 RESULTS

First, the result of a random forest model was put below to show the accuracy and viability of the most effective model, which helps us conceive the very core of hypothesis gauging whether the score is enough to replace appraisal's report. Model was built on real estate web site data and tested on appraisal's report data which was collected through an online form. Accuracy score was found as 0.828 along with root mean square error of 160.353. Root mean square error finds its real meaning when other details of target variable is given as well. Therefore; mean, median, standart deviation, interquartiles of target variables are also important to state at this point. Mean value of target variable (house price) is 590.644, median is 445.000, standard deviation is 508.863, and 25% quartile is 298.000 and 75% quartile is 700.000. Also, it is important to mention that unit for all these values is Turkish Lira, and data is collected for the interval of August 2021 – December 2021.

Table 4: Scores for Validation and Test Data Sets

	Validation Data Set	**Test Data Set**
R Score	0,811	0,828
RMSE	211.276	160.353

Above table indicates the evaluation metrics of the model which includes accuracy score and root mean square error for both validation and test data set. Validation data set is a 20% random sample of training data set, and test data set is appraisal's report data. Interpretation of these results are discussed in coming chapters, and numbers are presented in results section as is without any inference for the sake of simplistic structure of this dissertation.

Random Forest Model appeared as the best performing model among a variety of others, namely Linear Regression, Artificial Neural Network, Support Vector Machine

and XGBoost, and its result was evaluated throughout the study. However, mentioning other underperformed models will give readers a clue regarding the journey of this study. Although each model has been revised many times with different hyper parameters and layers when applicable, only best accuracy scores of each was thought to be useful to exhibit.

Defining root mean square error (RMSE) here helps us through conceiving how well our models performed. Root mean square error is the standard deviation of the errors of our predictions (residuals), and these prediction errors are found by calculating the distance between regression line and data points. Below we see the comparison of RMSE evaluation metrics of each model implemented on validation data set.

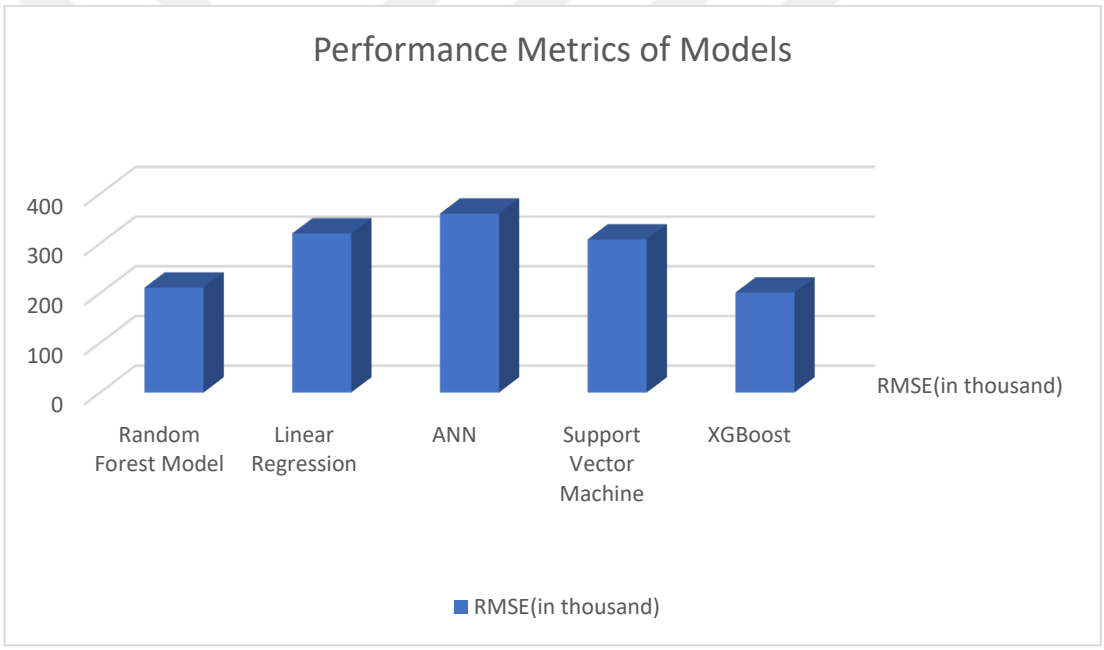

Figure 4: RMSE Scores of Each Model

Mean Absolute Error (MAE) is also another evaluation metric, especially when target variable contains relatively large values and outliers play a big role in it. Further performance metrics including Mean Absolute Error and R Scores of all models are provided below in a table to get a better grasp of whole study. Only well-performed models of all, such as Random Forest, XGBoost and ANN were subject to hyper parameter tuning. Linear Regression and Support Vector Machine Models performed poorly in the first place by using some default hyper parameter values. Although RMSE scores of XGBoost and Random Forest quite close to each other when taking

other evaluation metrics into consideration, Random Forest appears to have best performance.

Table 5: Model Comparison

	R Score	RMSE	MAE
Random Forest Model	0,811	211.276	114.234
Linear Regression	0,560	320.511	184.880
ANN	0,582	360.172	238.032
Support Vector Machine	0,598	308.570	165.012
XGBoost	0,775	201.816	119.098

Duration of model trainings were not subject to comparison, due to the fact that different tools were utilized to train different models, such as Google Colab for Artificial Neural Network and computational power of a personal device for Random Forest Model. ANN model requires relatively high computational power, having much more hyper parameters. Hence, taking advantage of cloud GPU usage seemed a reasonable approach along with abstaining from potential version conflict issues of python libraries in mind. Taking a look at hyper parameters which were utilized during the model training in ANN can be useful with the help of visual below.

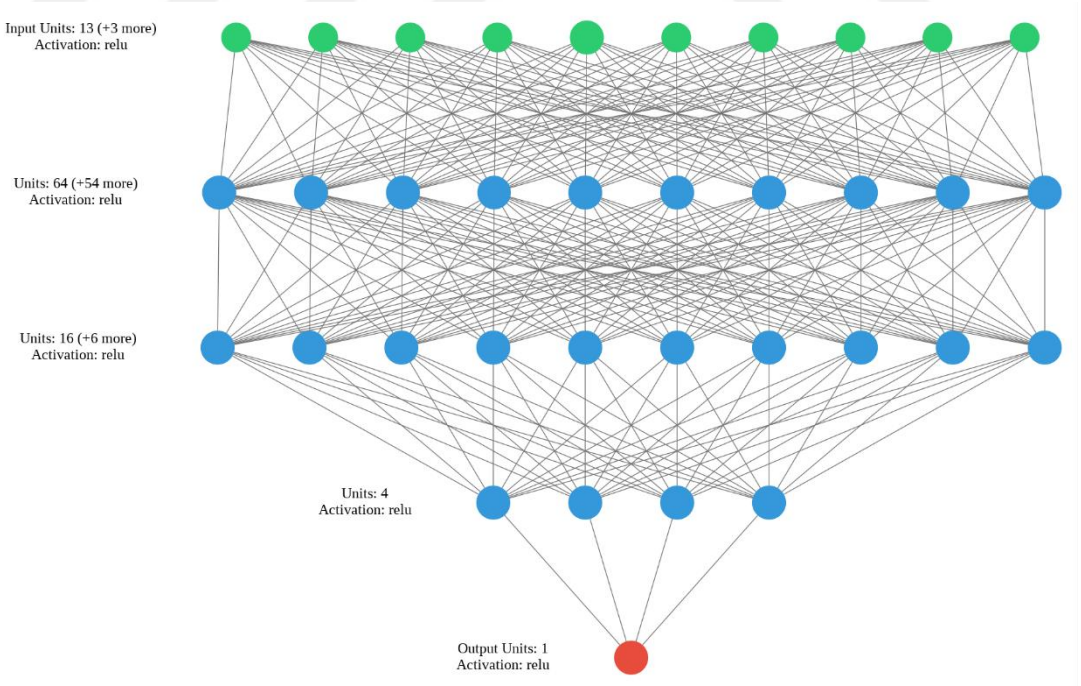

Figure 5: Structure and Hyper Parameters of ANN Model

Size of the data is always an issue when it comes to effectiveness of the model. Although various different approaches have been studied from this standpoint, recent studies, one of which belongs to Domingos, states that more data beats cleverer algorithm. That was also experimented in this study and findings related to the size of the data are as follows:

Table 6: Effectiveness of the Size of Data Set

Sample Model (RF)	R Score	RMSE
250 rows	0,601	266.196
1000 rows	0,766	193.751
10000 rows	0,797	161.659

Car price prediction study whose details were given in literature review part (by Gegic and his friends) attribute the high accuracy score of the model to rigorous data cleaning work. To comprehend effectiveness of data cleaning part, an experiment conducted and displayed below. We see how much effect data cleaning has on accuracy scores of Random Forest Model. However, it is important to keep in mind that some data cleaning job was implemented in the very first part of study while scraping. Therefore, there still is some preliminary cleaning to a modest degree in the column called ‘without data cleaning’ even though distinction is observable.

Table 7: Effectiveness of Data Cleaning

Sample Model (RF)	Without Data Cleaning	With Data Cleaning
R Score	0,798	0,811
RMSE	212.436	211.276

Some early researches, as in the one by James and Lam (1996) concluded that neural networks seem to perform better for smaller data sets. This idea was also experimented in this study to find an answer to following question: Is artificial neural network model performance better for smaller data sets? To find an answer to this question, 250 sample data were chosen and 3 models were compared to one another to understand if a model is superior in this context. However, ANN did not stand out from the rest in repetitive experiments, one of which has details in the table below.

Table 8: Model Building on Small Data Set -RMSE Score Comparison

	250 Rows	10000 Rows	Change Percentage
ANN	448.863	381.638	0,176
XGBoost	241.198	202.090	0,193
Random Forest	185.193	161.659	0,145

To comprehend the results at a glance chart below will cooperate with us. No direct relationship is seen between the model type and data set size. Choosing different samples of 250 rows brings different results each time.

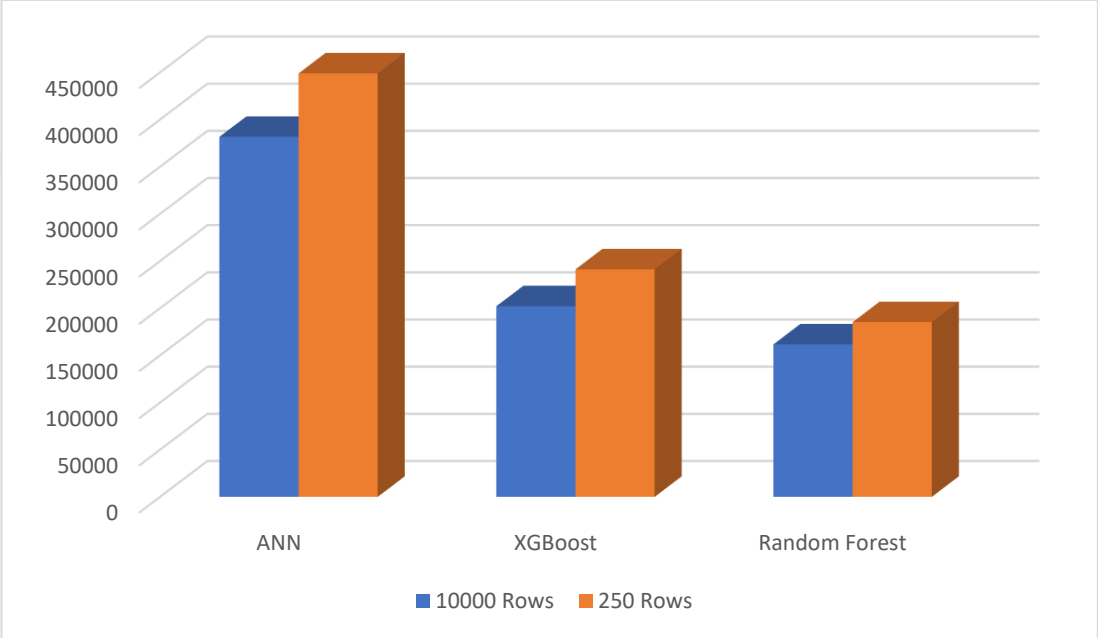

Figure 6: Model Comparison of Data Set Size Experiment by RMSE

Complexity of model preference by data set size remained as is after this experiment, as no direct relationship was found between those two.

4.2. DISCUSSION

4.2.1. ANALYSIS OF ANKARA HOUSING DATA SET

Before giving details about what has been found further, restating research question will be helpful to comply with the context: Is machine learning prediction on a data collected from web accurate and viable enough to replace house appraisal report in a mortgage loan process? To be able to answer this question adequately, getting a firm grip on data by analyzing it do so much help. These analyses also help those who

want to research this topic on the same data to take it to the next level. First visual in this part demonstrate us the heatmap of correlation among features.

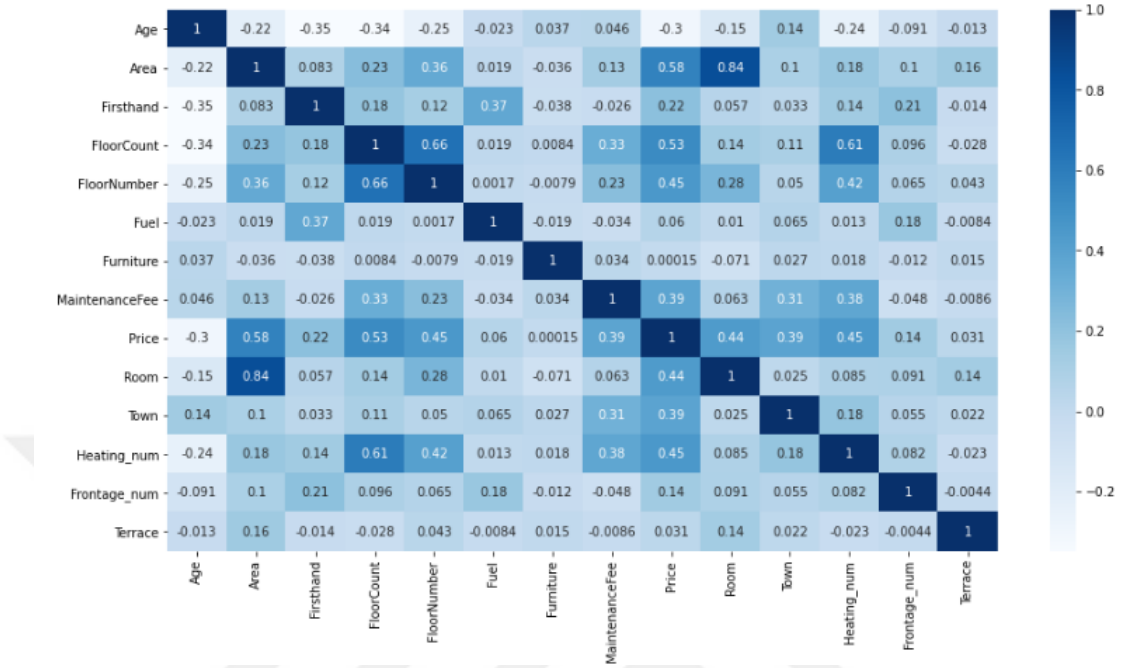

Figure 7: Heat Map of Correlations Among Features of the Data (After Cleansing)

From the heat map, it can be seen that which two variables have high correlation between them, and which feature has the highest correlation with house price. Area of the house appears to have the highest correlation score with price. This finding complies with the folk wisdom and is something expected. However, floor number feature, which shows the which story the flat belongs to, seems to have a higher correlation with price than floor count which is the total number of floors in the building. This is a bit of counter intuitive. When the reason behind this was investigated, it turned out that web page that data scraped from does not accept values higher than 21. Therefore, the effect of correlation that includes floor count stays limited for this reason.

Jointplots provides a decent way to visualize the relation of two variables for multiple reasons, one of which is outlier detection. Outliers can be seen with the help of jointplots, and then action to be taken needs to be decided. In the visual below, we see the relation between room and area variables.

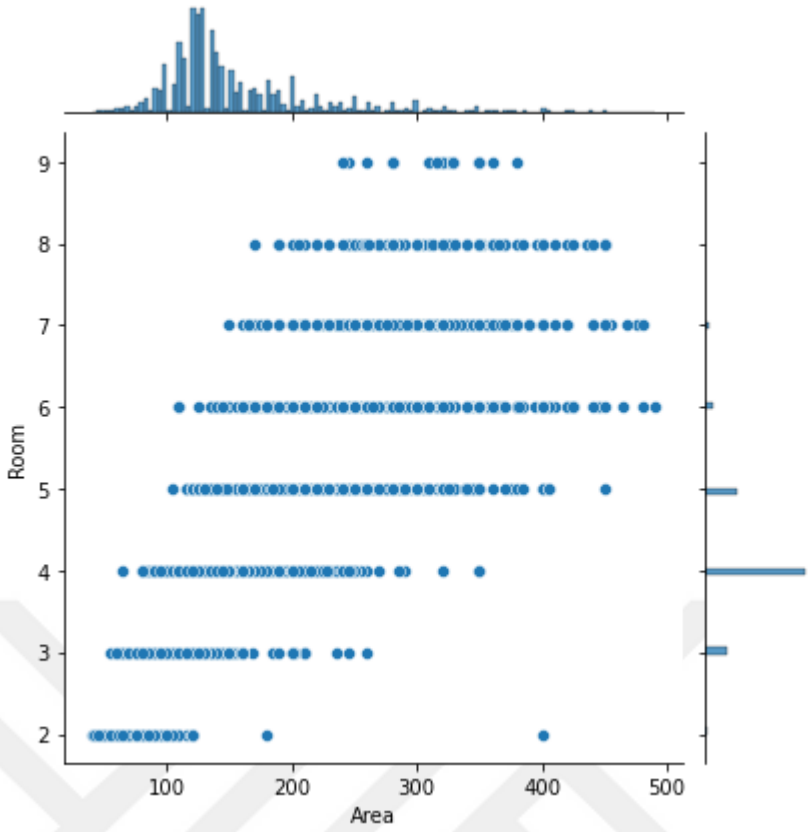

Figure 8: Jointplot of Area and Room

In the joint plot above, we see that there is a two-room house with the area value of 400. It can be easily concluded that there must be some type of faulty input during data entry process due to the unrealistic characteristics of the house. It was excluded from the data set for the sake of robustness expectancy of the model. It is also important to state that there are two types of area information in the web page which can be misleading for house owners during the entry. Additionally, knowing the exact value of area is quite challenging for most. Therefore, area is such an important factor that it deserves to be treated diligently even further by both model builders and data collectors.

Location information is also quite crucial in this context. Minimum, maximum and mean values of each town was investigated to analyze outliers, and also to gain an insight.

Table 9: House Price Details by Town in Ankara

Town	count	min	max	mean
Elmadağ	12	125.000	440.000	295.250
Çubuk	67	130.000	750.000	301.858
Polatlı	451	125.000	900.000	324.643
Sincan	904	125.000	850.000	342.500
Mamak	3.349	125.000	1.750.000	359.055
Keçiören	3.065	125.000	2.500.000	513.254
Altındağ	816	125.000	2.300.000	521.263
Pursaklar	1.171	130.000	3.950.000	543.335
Yenimahalle	1.310	125.000	4.500.000	636.981
Etimesgut	1.723	125.000	3.650.000	688.008
Çankaya	5.120	125.000	5.000.000	828.406
Gölbaşı	254	185.000	3.650.000	882.342

Table displays many angles to data set by town. The upscale towns, total number of houses for sale and so on. Because of price range restriction that was implemented for outlier detection purposes, minimum and maximum numbers appear to be close to each other. Bar chart below also clarifies the price range among towns by providing some insight from the table.

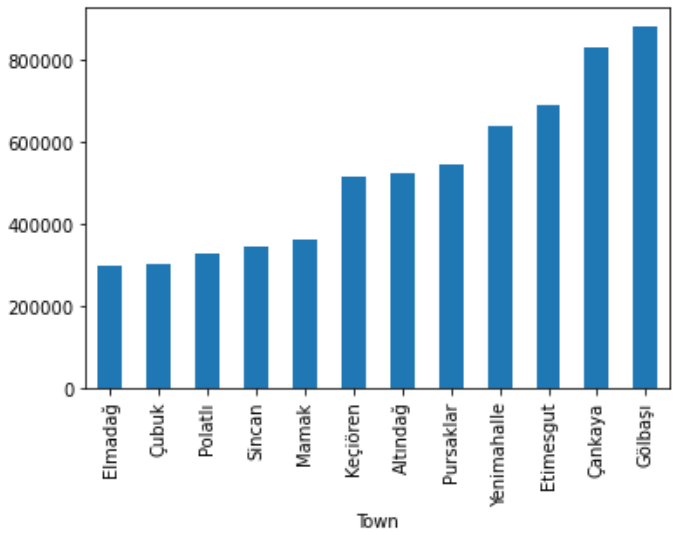

Figure 9: Bar Chart of Town and Price

Golbasi seems to have the most expensive houses for sale in Ankara. The mean value for that town is 882.342 TL. However, the number of total houses for Golbasi is only 254. Total number of houses by town can be said to comply with the population density of those towns as a ballpark estimate. Hence, no abnormality is detected in that

regard. Elmadag, Cubuk, and Polatli seems to have the least expensive houses in the city with close price ranges to one another.

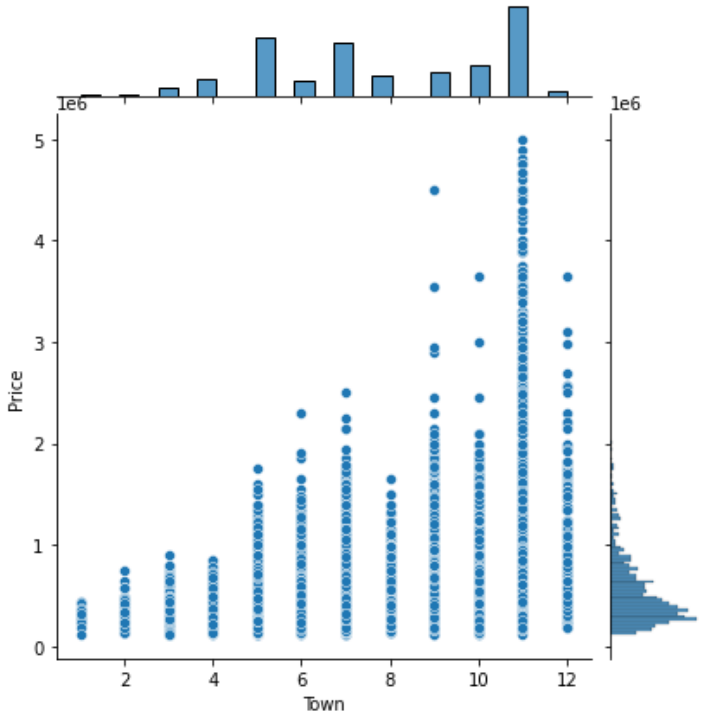

Figure 10: Jointplot of Town and Price

In the heat map of correlation matrix that was given at the beginning of the discussion chapter, town variable appeared to have a quite significant correlation with price. Therefore, examining these two variables in the above visual lead us to find out any outlier values. Town names were replaced with numbers for two reasons. First, machine learning models can be implemented on only numbers, and second reason seaborn library is not optimized to display long texts on x and y axes properly. However, numbers match the order of mean values of the house prices in that town. Accordingly, the town 1 shows the values of Elmadag, whereas 12 stands for Golbasi.

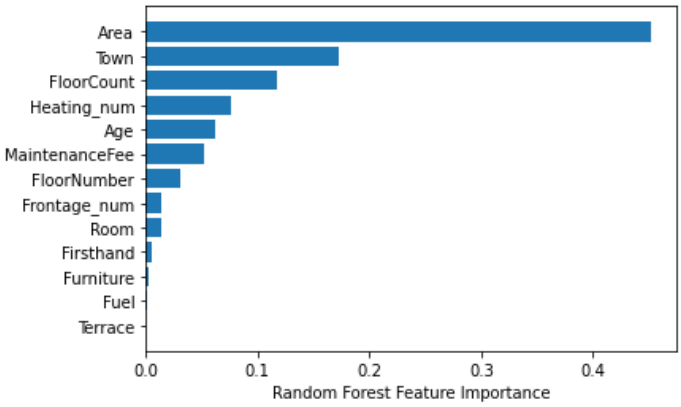

Figure 11: Random Forest Feature Importance Bar Chart

This horizontal bar chart above indicates the importance of features in Random Forest Model. The area is seen to be the most effective one as expected. Apart from that, Town, FloorCount, Heating are the ones that have biggest effect on the output, respectively. This feature importance bar chart has been utilized throughout the research a few times. For example, Orientation feature had much weaker affect in the first place, however some cleaning stuff and a proper label encoding did the magic and increased the effect of it along with a rise in overall performance. Hence, it can be said that further researches on this data set need to focus on developing fewer effective features by changing null value imputation methods and so on. The 3 features that have the least effect on the output are possession of a terrace, fuel type, and furniture.

4.2.1.1. Limitations

Limitations of present work include:

Lack of features such as thermal insulation, real estate agent commission, number of bathrooms, leaks and damages, locational advantage or disadvantages and garage possession have an effect on the performance of the model. Having these data too, would definitely help improve the result by having a better root mean square error and a higher accuracy score. Although some of the mentioned features are easy to add to this model, a couple of them such as locational advantage or leakage needs text parsing to be added. It is wise to keep in mind that not all features are given in filters but some of them are hidden in the text under the advertisement on relevant web page.

In addition, some features have limitations on their own. For instance, floor count feature varies between 2-22 although some buildings have more than 22 floors.

This issue stems from the limitation on the web site which data set has scraped from. If a building has more than 22 storeys then data entry is implemented as 22, which is maximum value for that value on the web site.

Another limitation of model is orientation feature of houses. It is well known fact that in the northern hemisphere the southern exposure of your house gets sunlight most. Research has been conducted for houses in the city of Ankara, and Turkey along with its all cities is located in northern hemisphere. Therefore, orientation that faces south is advantageous in the city. Although it is expected that it also affects the price of a house to a certain extent, feature importance of the model tells us that orientation is one of the weakest characteristics that has the least effect on the output. Potential reason behind this is sloppiness or intentionally erroneous entry of house details on real estate web site. If orientation is entered carefully or a subset of data that has a clean orientation information in it is chosen, a better output would be expected from the model.

Throughout the study, a machine learning model that is built on a web-based data set has been offered to shorten a business process. However, for the sustainability of the suggested model, a data pipeline and continuity of the data are needed. Because house appraisals are continuing processes and will never stop as the mortgage processes last, data always face the risk of being outdated. For this reason, a data pipeline and an automated machine learning model would be necessary in the case of implementation. Monthly renewal of data set would do the work with current inflation rate. Provided that inflation rate goes up further, frequency of data renewal process can change to once in fortnight or so.

CHAPTER V

CONCLUSION

This research aimed to investigate the viability and accuracy of house appraisal automation in a mortgage loan process by asking the question if automation is ready to replace traditional approach. Based on the quantitative and qualitative analysis of house appraisal, it can be concluded that house appraisals can be automated with some limitations in it for now. Those limitations were listed in the practical applications of findings section thoroughly. With this is mind, further improvements are needed for more comprehensive implementations.

Many other researches, as mentioned in literature review part, in this domain were introduced for both statistical-focused ones and the ones with information technology viewpoints. Early researches suffered from inconsistent results due to lack of data, while recent ones have had better performance lacking in the suggestions of real-life use cases. This research helps to develop a fresh point of view with using a new data set for the problem, and fills a gap by providing thoughts about the viability of theories. In this regard, research confirms the conclusion of previous studies that refer to the complexity of the problem while attempting to broaden the spectrum in terms of data in a result-oriented way.

As explained throughout the research, data used in this research are from a public resource. Real life implementations do not have to stick with only this data set. By utilizing the data that officials such as banks or government hold along with public resources, it can be expected to reach a higher accuracy score from a quantitative standpoint.

Based on these conclusions, practitioners and all parties involved should consider the dynamics and all angles to this issue in order to develop the robustness of the automation. With the contribution of all parties such as real estate agents, house appraisers, and bank authorities, theories can be implemented in real life for the common productivity of public. To better understand the implications of these results,

future studies could address not only statistical results with another data set but also the amount that could be saved as a result of this implementation.

All in all, in today's world, which becomes increasingly more data-oriented year by year, it is expected to see various data-driven transformations in coming years including house appraisals. This research opens this door to a wide range of possibilities to change traditional methods with upcoming breakthrough.

5.1. RECOMMENDATIONS FOR FUTURE WORK AND THE PRACTICAL APPLICATIONS OF STUDY

Some of the possible future works and real-life applications from the theory are mentioned below in order to get most out of this research and improve as well if needed.

5.1.1. Practical Applications of the Findings

Findings suggest that there still is room for improvement to apply the idea to the business steps of a financial institution as is for most cases. However, further studies can make the idea more viable in this regard. Suggestions for further studies are elaborated in the following parts, and here are some considerations on the viability of the study:

1- Although determination of the price of a house is vitally important and needs to be as accurate as possible, less accurate predictions could still have a place as a cost-effective alternative. Here is how: Financial institutions needs to know the price of a house with regard to the certain ratio between the value of a house and mortgage credit limit that can be taken out for that house. For example, in Turkey mortgage loan customers can take out credit from a bank up to 80% of the house value according to the latest regulations. Be that as it may, not all customers need full amount of their limitation. At this point, an accurate prediction loses its significance for those customers who do not need full credit limitation. Indeed, a customer can be offered two alternatives according to the above-mentioned percentage value. If 80% of the house value needed to be taken out then conventional methods can be applied, yet for those customers who only need half of the house value or maybe less this study can be a time-saving and economical application solution for them after all.

2- Another approach to viability of the study would be to find out a reasonable portion of the data set in terms of the sufficiency of accuracy. A data set which scraped

from web is quite likely to have some faulty values in it, and analyses throughout any research like this makes it clear to implementer that faulty values inclined to be in extreme values. Therefore, eliminating those faulty values with subsampling rather than outlier detection can actually increase the performance in a reasonable way, and make the study to one step closer to real life applications. For instance, provided that accuracy score is more convincing for houses that are priced within the range of 250.000 TL and 1.000.000 TL, then that subset of the data set can be applied in the first place. This kind of a scenario could be considered as a full-scale outlier detection in terms of price with little to no effort and may indicate a realistic performance of houses of which have fewer faulty data.

3- There are many players in the real estate game apart from house owners, banks and house appraisal companies. Insurance companies, real estate agencies etc. can be considered as some other parties for that matter. Although solving the problems of those other players in the game is not among the primary goals of this study, they can still benefit from the study by learning the estimated price of a house. Real estate agencies need to know a value estimation to set a realistic price and demand a satisfactory commission without increasing the market too much, and insurance companies need the current value and maybe trends within the value for their policy prices. These parties can use the output of this study maybe not as the sole foundation for their work but as one of the components that helps with their current workflow.

5.1.2. Suggestions for Further Research

Below-stated questions for further researches could be fruitful in real-estate price prediction studies which aim to replace the traditional approach.

. What is the effect of a real estate agency on the price of a house? We know that many different aspects can have an effect on house prices. What role a real estate agency plays in this regard?

. Can this research have a better accuracy with a cleaner and more relevant data? To be more specific, if there were more features in the data set or more houses for that matter, would it help with a more accurate output? Can data set be cleaner and better with only recent-published houses?

. Does state-of-the-art machine learning models such as XGboost or deep learning works better than random forest model if their hyperparameters are further tuned? Comparison between models would be a necessary study in this area. It is

beyond the scope of this study to address the question of which model performs better, yet those with technical background can investigate this in the context of real estate.

These ideas could be diversified, yet the above list would help encourage those who think of working in the field of real estate price estimation. What matters most is to find a real-life case from information technology standpoint.

REFERENCES

[1] LAUDON K. C. and LAUDON J. P. (2016), *Management Information Systems: Managing the Digital Firm*, Fourteenth Edition, Pearson Education Ltd, United Kingdom, pp. 33-61.

[2] GRAVES P., MURDOCH J. C., THAYER M.A. and WALDMAN D. (1988), "The Robustness of Hedonic Price Estimation – Urban Air Quality", *Land Economics*, No. 65, pp. 220 – 233.

[3] TAJANI F., MORANO P. and NTALIANIS K. (2018), "Automated Valuation Models for Real Estate Portfolios: A Method for the Value Updates of the Property Assets", *Journal of Property Investment Finance*, Vol. 36, No.2, pp. 324–347.

[4] KALLIOLA J., KAPOČIŪTĖ-DZIKIENĖ J. and DAMAŠEVIČIUS R. (2021), "Neural network hyperparameter optimization for prediction of real estate prices in Helsinki", *PeerJ Comput. Sci*., No.7, pp. 444-461.

[5] KUSAN H., AYTEKIN O. and OZDEMIR I. (2010), "The use of fuzzy logic in predicting house selling price", *Expert Systems with Applications*, No. 37, pp. 1808–1813.

[6] DOMINGOS P. (2012), "A Few Useful Things to Know about Machine Learning", *Communications of the ACM*, No. 55, pp. 78–87.

[7] GEGIC E., ISAKOVIC B., KECO D., MASETIC Z. and KEVRIC J. (2019), "Car Price Prediction using Machine Learning Techniques", *TEM Journal*, Vol. 8, No.1, pp. 113-118.

[8] BALDOMINOS H., BLANCO I., MORENO A., ITURRARTE R., BERNÁRDEZ O. and AFONSO C. (2018), "Identifying Real Estate Opportunities Using Machine Learning", *Appl. Sci.*, No. 8, pp. 2321, doi:10.3390/app8112321.

[9] LIMSOMBUNCHAI V. (2004), "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network", *NZARES Conference*, Blenheim, New Zealand.

[10] JAMES R. C. and CAROL E. B. (2000), "Artificial Neural Networks in Accounting and Finance: Modeling Issues", *International Journal of Intelligent Systems in Accounting, Finance & Management*, No. 9, pp. 119 – 144.

[11] JAMES H. and LAM E. (1996), "The Reliability of Artificial Neural Networks for Property Data Analysis", *Third European Real Estate Society Conference*, Belfast, Ireland.

[12] WANG L., CHAN F. F., WANG Y. and Chang Q. (2016), "Housing Price Prediction Using Neural Networks", *12th International Natural Computation and 13th Fuzzy Systems and Knowledge Discovery Conference,* Changsha, China.

[13] WORZALA E., LENK M., and SILVA A. (1995), "An Exploration of Neural Networks and Its Application to Real Estate Valuation", *The Journal of Real Estate Research*, No. 10, pp. 185-201.

[14] HROMADA E. (2016), "Real Estate Valuation Using Data Mining Software", *Procedia Engineering-Creative Construction Conference,* Prague, Czech Republic.