DOI: https://doi.org/10.54338/27382656-2022.2-012

Henrik Tigran Sergoyan[1*], Grigor Vahan Bezirganyan[1]

AUTOMATED REAL ESTATE VALUATION WITH MACHINE LEARNING: A CASE STUDY ON APARTMENT SALES IN YEREVAN

Technical University of Munich, Germany

Real estate is one of the major sectors of the Armenian economy and has been developing dynamically since Armenia transitions from planned to market economies in early 1990s. More recently, large online platforms have been developed in Armenia to advertise real estate offerings, thus reducing information asymmetry, and increasing liquidity in both sales and rental markets. Simultaneously, granular geospatial data became increasingly affordable via platforms such as OpenStreetMap, Google Maps and Yandex Maps. With granular data concerning a representative portion of the real estate offering available online, it is increasingly tenable to monitor the real estate market in real time and develop analytical tools that can automatically and accurately estimate the value of real estate assets based on their internal and external features. This paper sets out to analyze Armenia real estate market and assess the performance of a special class of machine learning models while predicting the price of a square meter of apartments in Yerevan. Furthermore, it is presented the way to determine the most decisive factors which have an influence on the price of apartments on sale.

Keywords: *Real-estate market, machine learning, automatic valuation, feature importance calculation, XGBoost.*

Introduction

Real estate is a major sector in the Armenian accompany. As of 2020, it has a market capitalization of nearly 1 billion USD, and outstanding mortgage loans amounting to approximately 226 billion AMD [1,2]. While the rate of growth of total economic activity fell to 4.4% during the first three quarters of 2021, the same indicator for the real estate market rose to exceed 16% . Economists largely attribute the increasing activity in this sector with the relative lack of alternative investment opportunities, an issue that is exacerbated by the COVID pandemic and the conflict in Ukraine [3].

Increasing transaction volumes and price volatility are the primary motivation for building a holistic, scalable automatic valuation framework. A highly accurate and scalable framework for automatic valuation can help stakeholders in the real estate sector detect potential market opportunities in real time and reduce overhead cost by automating significant portions of the valuation workflow. Furthermore, automated valuation reduces the likelihood of human error or malfeasance negatively impacting the valuation process [4]. The goal of this paper is to analyze the real estate market in Armenia and implement an apartment price prediction model. We will assess the performance of two different classes of tree – based machine learning ensembles, named XGBoost and Random Forests [5,6]. Related works show that these two classes exceed the performance of other classical machine learning models [7]. Furthermore, their primary advantage over neural networks, which often match them in accuracy, is that their hyperparameters are more easily tunable and model performance is more explainable. These factors are especially important for application in banking, since automatic valuation that is used in mortgage loan must be explainable [8]. As an end point of our work we have adopted SHAP (SHapley Additive exPlanations) approach to interpret model predictions and calculate features importance [9].

Data

Data concerning conventional apartment features is collected from the platforms presented in Table 1. These platforms host the largest numbers of real estate offerings relative to any other in Armenia where list.am is the overwhelming leader. It is important to note that these numbers were determined following detailed

deduplication – the process whereby multiple announcements for the same apartment are detected and discarded. Table 2 describes the data gathered from these platforms. This set of features is a distilled version of all the data scraped from the platform, since some of the data was lacking veracity and was too sparse to be used as features in a dataset.

In addition to the conventional data recovered from the abovementioned sources, data concerning the geospatial features of the neighborhood surrounding the apartment was recovered using the *OpenStreetMap* API [10]. A neighborhood is defined as the set of points that can be reached in seven minutes if travelling at the average walking speed via paths designated for walking (such as sidewalks, etc.) and neighborhood isochrones are calculated using the openrouteservice API [11]. Neighborhood features are grouped into the categories described in Table 3. The number of each of these features in the neighborhood immediately surrounding a particular apartment is likely to influence the price of said apartment. In his paper entitled Interpretable Machine Learning for Real Estate Market Analysis, Felix Lorenz applied a hedonic pricing approach to automatic valuation which indicated that distance from the city center and density of neighboring amenities had a positive impact on rent prices [12].

Table 1. The number of announcements in each portal

Platform	Number of Apartments Scraped
Myrealty.am	3834
Real-estate.am	5217
List.am	39062

Table 2. Data types

Variable	Variable Type
Latitude and Longitude	Float
Number of rooms	Integer
Number of bathrooms	Integer
Building type	Categorical
Ceiling Height	Float
Condition	Categorical
Floor/Total Floors	Integer
New Construction	Categorical
District	Categorical
European Window	Bool
Furnished	Bool
Sunny	Bool
Home Area	Integer
Balcony	Bool
Price	Integer

Table 3. Neighborhood feature categories

Category	Examples
Sustenance	restaurants, cafes, bars, clubs, etc.
Education	schools, universities, kindergartens, etc.
Transportation	bus stops, parking lots, charging stations, etc.
Financial	banks, ATMs, currency exchanges, etc.
Healthcare	pharmacies, hospitals, clinics, etc.
Entertainment, Art & Culture	movie theater's, museums, conceit halls, etc.
Public Service	courthouse, post office, town hall. etc.

Data exploration

Fig. 1 illustrates the distribution of the price of a square meter in a Yerevan apartment based on data scraped from the abovementioned sources. This is the target variable that we aim to predict. The histogram peaks around the 800 USD to 1200 USD price category and declines in reverse proportion to price, as expected. It is important to note that the prices quoted in online announcements are often not the final transaction value of the apartments. In fact, our analysis of data reported in the annual report by the Cadastre Committee of the Republic of Armenia indicates that online prices are 1.3-1.9 times higher than those reported by Cadastre, where the coefficient fluctuates based on the administrative district. One issue that is raised because of this distribution shape, is that there is relatively less data to represent apartments on the higher end of the market, therefore leading to relatively inaccurate results on the higher price range. Finally, online prices are often

quoted irrationally, and based on the whim of the seller or agent representative, who may have sentimental value and inadequate understanding of market conditions.

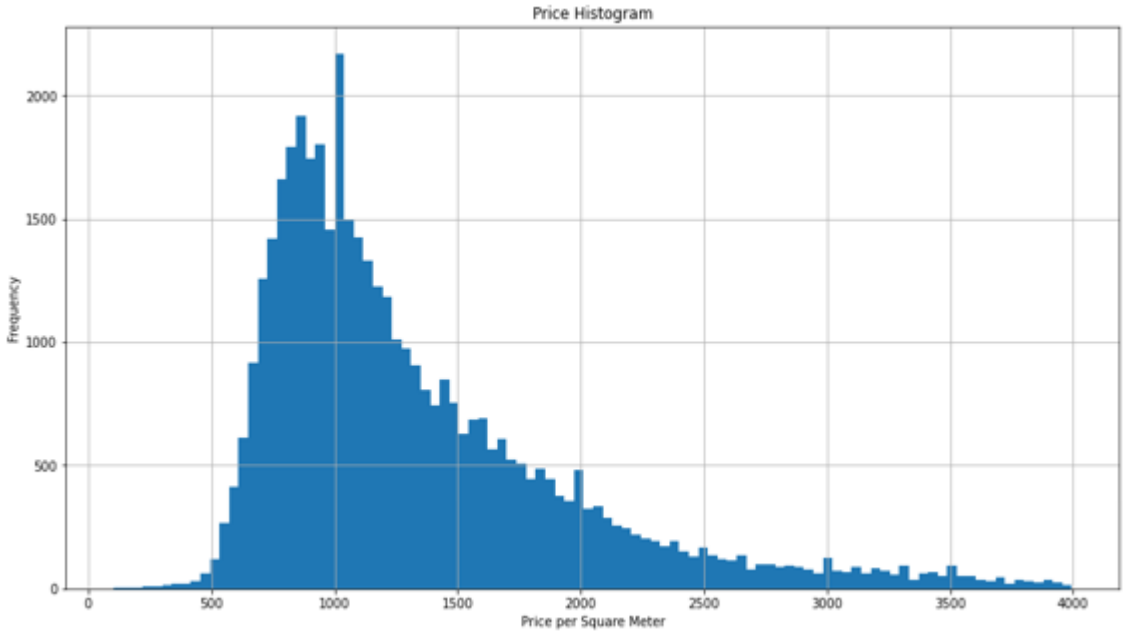

Fig. 1. *The Distribution of the Price of a Square Meter*

Fig. 2 illustrates the distribution of surface areas for apartments in Yerevan. A vast majority of apartments have less than 200 square meters, but some outliers have over 1000 square meters. Detailed analysis of these obvious outliers indicate that they are most often commercial properties that fell in the apartment category in the online platforms due to human error. These outliers are carefully filtered to increase representativeness of the dataset.

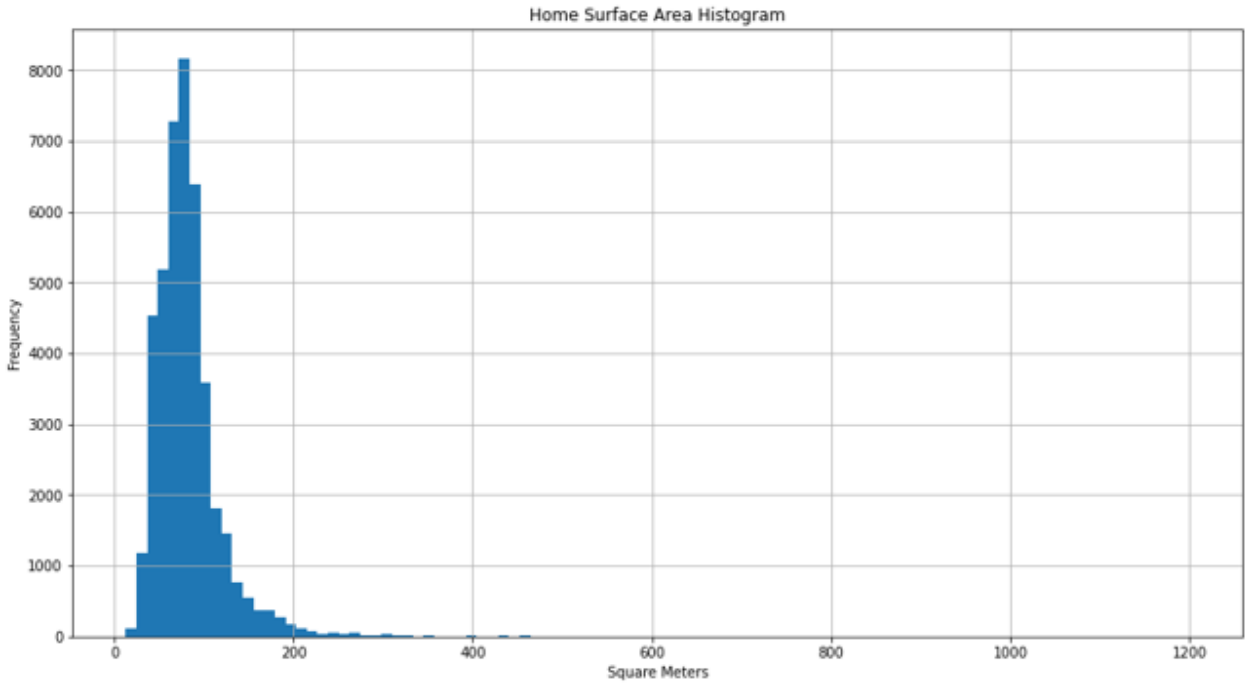

Fig. 2. *The Distribution of the Apartments' Surface Areas*

Fig. 3 illustrates that the most popular number of rooms in an apartment is three. Of note is the fact that there are more apartments that have four rooms than apartments that have one room. Bathroom counts are ordered intuitively in ascending order, with apartment with one bathroom being the most frequent and apartments with five bathrooms being the least frequent.

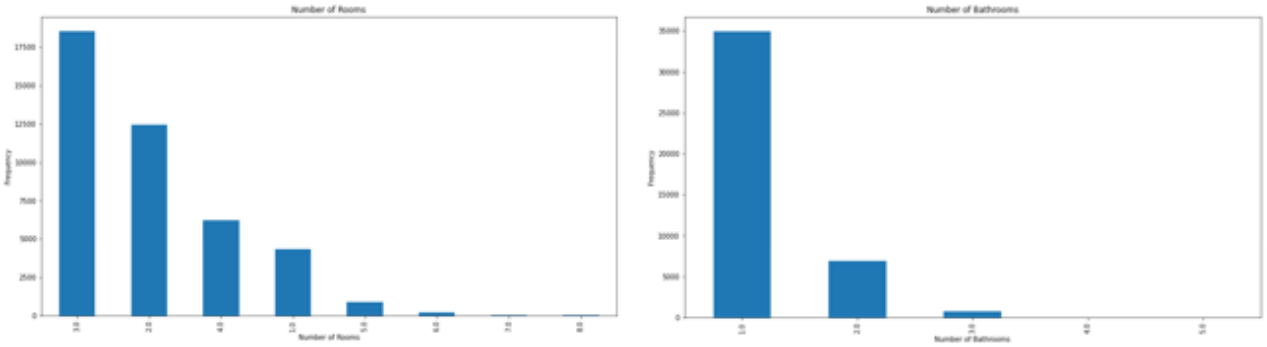

Fig. 3. *Room and Bathroom Count Frequency*

Fig. 4 illustrates that apartment floor frequencies are also ordered largely with respect to an ascending order, with higher floors being increasingly less frequent. Some interesting exceptions to the trend hold. For instance, while the four most common floors are two, three, four and five, apartments on the first floor are in fifth place by frequency. Similarly, apartments on the eighth and ninth floors are more common than those on the seventh floor. According to the Fig. 4, apartment in buildings with five floors are the most common. Second, third, fourth and fifth by frequency are buildings that have nine, four, 14, and 16 floors respectively. The trend is otherwise irregular except for buildings with exceptionally high floor number. In fact, the trend in building floor frequencies is in line with Soviet building standards.

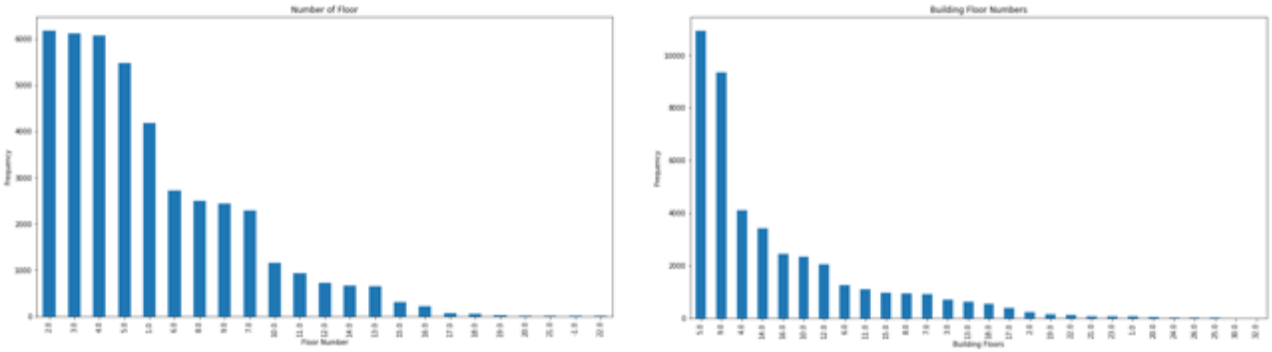

Fig. 4. *Apartment Floor and Building Floors Frequency*

According to Fig. 5, the three main types of buildings are: *panel*, *stone and monolithic*, in that respective order. The top renovation condition category frequencies are largely ordered with respect to a descending level of renovation, with the most common being *"Renovated"*. An exception to this trend is that *"Renovated in Modern style"* and *"Euro renovated"* have low frequencies, which most likely reflects their level of exclusivity. It is important to note that this indicator is somewhat subjective, since home owners often inflate the degree to which a property is renovated to justify higher prices.

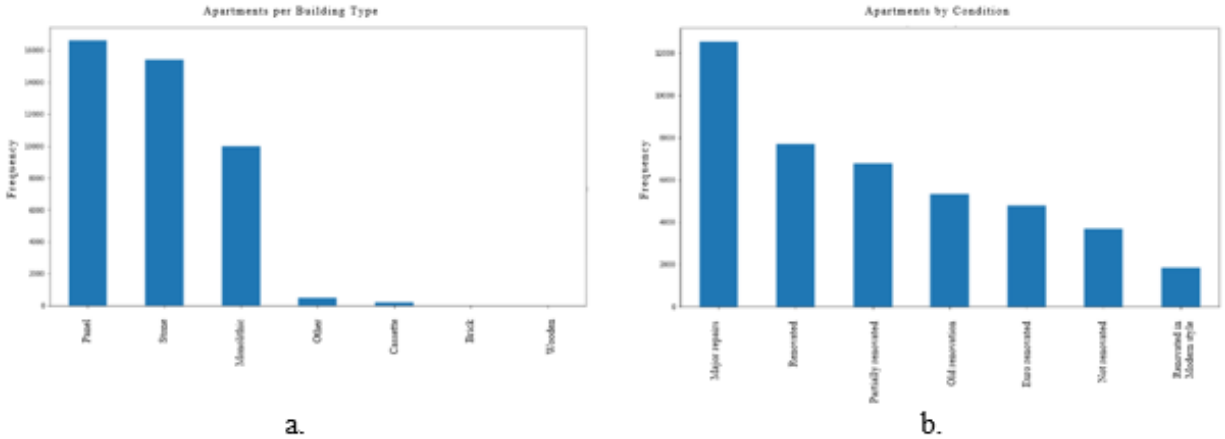

Fig. 5. *Building Type and Apartment Condition*
a. *Building Type*, b. *Apartment Condition*

Fig. 6 illustrates that the center of the city is the highest in both the density and the price of apartments. It is interesting to note that entire sections of the south – eastern portion of the city have almost no apartments for sale, and that those that are for sale have relative low prices.

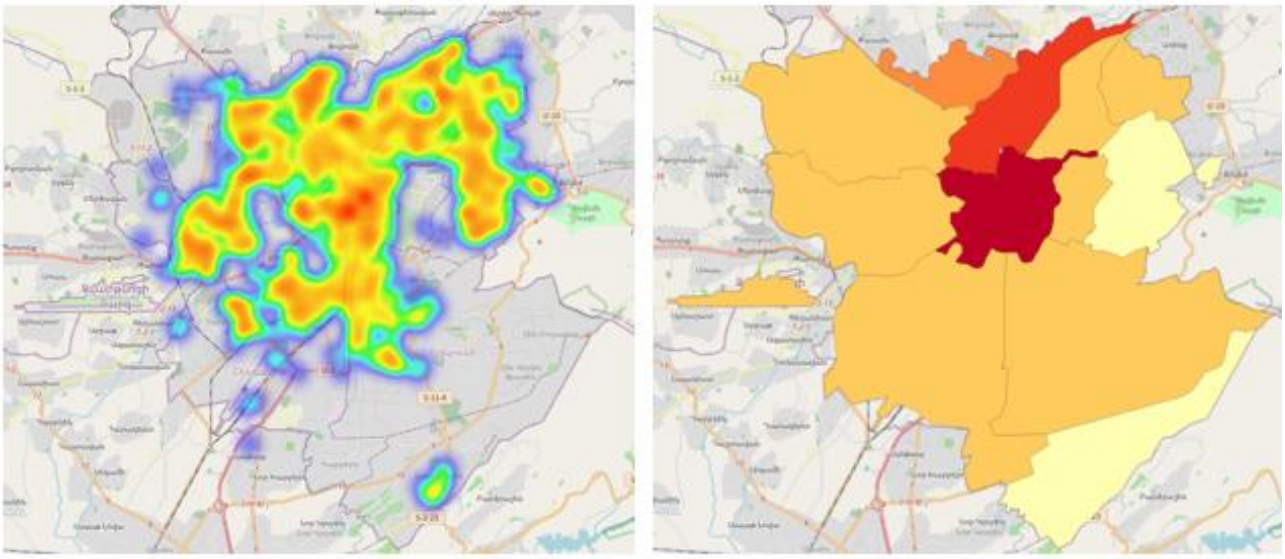

Fig. 6. *Density and Price Heatmaps*

Research methodology

Automated real estate valuation is a popular topic in applied machine learning, and recent work illustrates the effectiveness of various tree-based ensembling techniques for accurate value prediction.

Guliker et al. in their recent work compared the performance of three pricing techniques - linear regression, geographically weighted regression, and extreme gradient boosting - XGBoost) [13]. The performance of XGBoost exceeded that of the other models, with an explained variance of approximately 83% and Mean Absolute Percentage Error (MAPE) of 6.35%.

Metrics to measure the performance of real estate valuation models are also a major topic of discussion since each metric highlights a particular aspect of a model's performance. Steurer et al. suggested ways to assess the effectiveness of several metrics in measuring the performance of machine learning models for automated valuation [15]. A total of 48 metrics were divided into seven classes, and seven final metrics (one per each group) were determined to be the most effective to evaluate automated valuation models. Table 4 describes the final metrics proposed for model assessment where p_n and $\hat{p}_n$ respectively denote actual and predicted price.

Since the proliferation of affordable and granular geospatial data on platform such as OpenStreetMap, Google Maps, Yandex Maps, etc., the use of geographical location features has become increasingly prevalent in automated real estate valuation. Tchuente and Nyawa in their paper evaluate the impact of including granular geographical location features on model performance in the context of the French real estate market [14]. Table 5 illustrates the impact on modelling performance for tree – based ensembling models when including geographic location features data. The paper also illustrates that although neural networks and random forest techniques outperform other algorithms without geographic features, performance gain was larger for random forests and other tree – based methods benefit more than other architecture from the additional data.

Table 4. Metrics to assess model performance

Class	Metric	Formula
Average Bias	Log Median Prediction Error	$LMPE = med\left[ln\left(\frac{p_n}{\hat{p}_n}\right)\right]$
Absolute Difference	Mean Absolute Error	$MAE = \frac{1}{N}\sum_{n=1}^{N}\lvert p_n - \hat{p}_n\rvert$
Absolute Ratio	Max-Min Mean Absolute Prediction Error	$mmMAPE = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\max(p_n,\hat{p}_n)}{\min(p_n,\hat{p}_n)} - 1\right)$
Squared Ratio	Logarithmic Root Mean Square Error	$LRMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[ln\left(\frac{p_n}{\hat{p}_n}\right)\right]^2}$
Squared Difference	Root Mean Square Error	$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(p_n - \hat{p}_n)^2}{N}}$
Percentage Ratio	Max - Min Percentage Error Range	$mmPER(x) = 100\left\lvert\frac{\max(p_n,\hat{p}_n)}{\min(p_n,\hat{p}_n)} - 1\right\rvert > x$
Quaniilc	Inter - Quanile Range in Ratios	$IQRat = ln\left(\frac{p_n}{\hat{p}_n}\right)_{75} - ln\left(\frac{p_n}{\hat{p}_n}\right)_{25}$

Table 5. Performance improvement with Geospatial features for each model

Algorithm	Average Performance Improvement with Geospatial Features
Random Forest	31.7%
Adaboost	40.85%
Gradient Boosting	39.77%

Our approach

Related works in automatic real estate valuation mostly contend that tree – based ensembling models are preferable to other algorithm families, while also documenting the improvement in performance when complementing conventional apartment features with geospatial neighborhood data [7]. We have therefore decided to concentrate our efforts in assessing the difference in performance between tree – based boosting and bagging ensembling models, using geospatial neighborhood data. The sections below provide performance summaries for each case.

XGBoost performance

We have used the *xgboost* Python library to train an XGBoost regressor with the following custom parameters (other parameters are default):

- Number of estimators: 3500,
- Maximum depth: 100.

The learning curves for the model are summarized in Fig. 7. The performance metric that the learning curves track is the RMSE, and the orange and blue curves represent the RMSE on the validation and training sets respectively. XGBoost

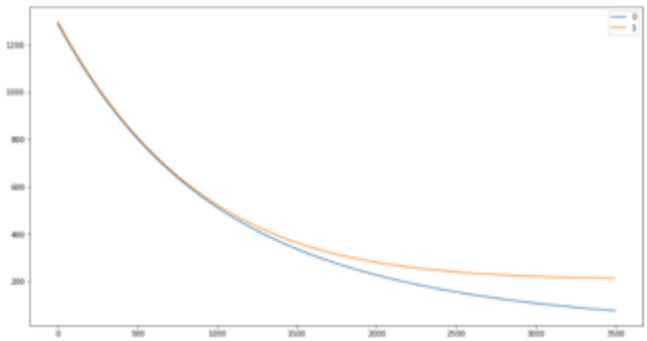

Fig. 7. *RMSE Loss for Training and Validation Sets*

is a sequential ensembling model and the curve illustrates the diminishing learning gains of each additional weak learner. It is evident that RMSE is no longer declining after 3500 estimators, meaning that additional estimators are no longer learning.

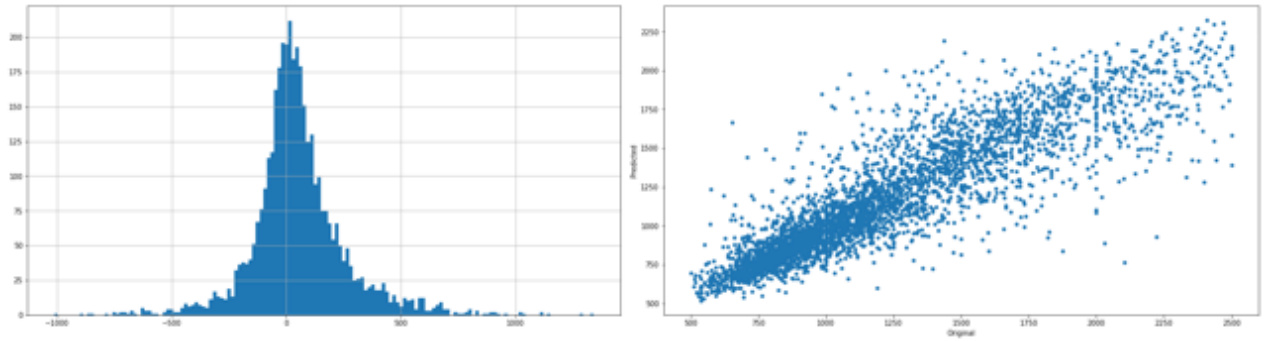

Fig. 8. *Error Histogram and Real vs. Predicted Scatterplot*

The error histogram, presented in Fig. 8, illustrates that the error is highly symmetric and centered around 0, which is an important characteristic of a well-balanced real estate valuation model. The scatter plot of real vs. predicted prices illustrates that a diagonal line has clearly formed, with some dispersion along this diagonal. The scatter plot not only illustrates the fact that test samples are more heavily concentrated around the 800 USD – 1300 USD price range, but also that error tends to increase as prices increase. This is likely a result of there being less training data for more expensive apartments and the fact that apartments which share traits that render them more expensive have a naturally higher variance in price.

Table 6 provides a regression report and accuracy comparison between XGBoost and Random Forest models that includes the metrics documented by Steurer, Hill and Pfeifer. From the Table it becomes clear that XGBoost outperforms its competitor. The model has a mean absolute error of 139.604 and root mean square error of 204.280, thus illustrating the relative impact of outlier errors. The median absolute percentage error is a popular metric to assess model quality. A score of 8.16% indicates that half of the estimates have an error below 8.16%.

Table 6. Model comparison on different evaluation metrics

Metric	XGBoost Scores	Random Forest Scores
Log median prediction error	-0.017	0.03
Mean absolute error	139.609	143.595
Max – Min Mean Absolute error	0.126	0.129
Logarithmic Root Mean Square Error	0.155	0.159
Root Mean Square Error	204.280	212.836
Interquartile Range in Ratios	0.162	0.16
Median Absolute Percentage Error	8.16	8.27

Of particular importance in automated valuation is the relative importance of each feature in the dataset in determining the price of the real estate asset. We have used the Shapley Additive exPlanations (SHAP) approach to obtain feature importance values. The method is based on the Shapley value as discussed in game theory and is often considered the most robust method for feature importance analysis in tree – based ensembling models or machine learning models in general. The SHAP scores can also be used for hedonic breakdown of apartment values as illustrated by L. Chen et al. [16]. According to Fig. 9, District is overwhelmingly the most important factor in predicting the price of a square meter, followed by the surface area of the apartment, whether it is a new construction or not, its renovation condition, building type, etc. It is interesting to note that the neighborhood indicator is not one of the top ten predictor variables.

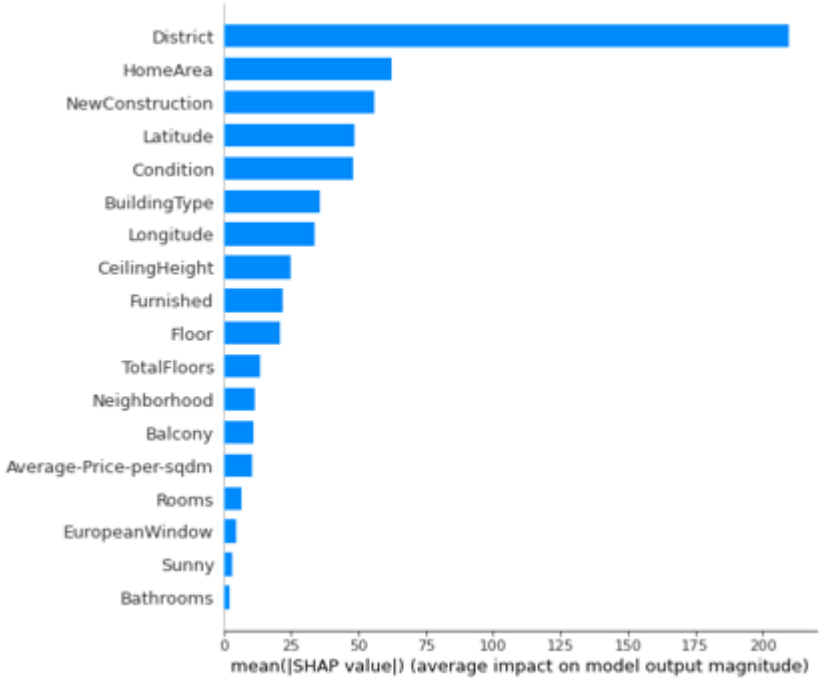

Fig. 9. *Mean SHAP Value*

Conclusion

To sum up, this work aims to provide data driven insights about current state of Armenian real estate market and present a novel approach for predicting house prices in Yerevan, Armenia. We identified the most important features which have an effect on the price of an apartment and elaborated on the explainability of models used for prediction.

We have assessed the performance of two families of tree – based ensembles and determined that XGBoost performs better than Random Forest on data from the Armenian real estate market. In collecting the data, we also developed a stable and scalable data collection pipeline that can function indefinitely to increase the size of our dataset and therefore improve model performance. We envision several future research directions to build on this work.

Firstly, it is highly advised to experiment with integration of computer vision models for object detection and image classification. By analyzing images in apartment advertisements, it is possible to ascertain features that describe furniture, interior design and renovation relying less on subjective descriptions by homeowners or real estate agents. For example, Poursaeed et al. in their work illustrates the effectiveness of convolutional neural networks for estimating the degree of luxury based on interior design, and it is possible to build on this work for more robust results.

Secondly, we can build on the results we obtained for feature importance using SHAP values, by building a hedonic pricing model that will estimate the dollar value of a feature of a property. These results will inform investors, home owners and other stakeholders in the real estate market with regards to the respective value of each feature, such as location, district, surface area or floor number, and will make financial decisions in the real estate sector more data - driven and well informed.

References

[1]. R. Sawsnt, O. Sumant, Armenia Real Estate Market by Property (Residential, Commercial, Industrial, and Land), and Business (Sales and Rental): Opportunity Analysis and Industry Forecast, 2019–2026. https://www.alliedmarketresearch.com/armenia-real-estate-market-A06057

[2]. S. Parsyan, Current trends in Armenia's real estate market. EVN Report. https://evnreport.com/economy/current-trends-in-armenia-s-real-estate-market/

[3]. A. Mejlumyan, Armenian real estate market booms. Eurasianet. https://eurasianet.org/armenian-real-estate-market-booms

[4]. J. Niu, P.Niu, An intelligent automatic valuation system for real estate based on machine learning. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing. ACM, 12, 2019, 1-6. https://doi.org/10.1145/3371425.3371454
[5]. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System. ArXiv, 2016. https://doi.org/10.48550/ARXIV.1603.02754
[6]. Tin Kam Ho, The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8),1998, 832–844. https://doi.org/10.1109/34.709601
[7]. Sheng Li, Yi Jiang, Shuisong Ke, Ke Nie, Chao Wu, Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model. Land, 10 (5), 2021. https://doi.org/10.3390/land10050533
[8]. A. Ghatnekar, A.D. Shanbhag, Explainable, Multi-Region Price Prediction. International Conference on Electrical, Computer and Energy Technologies. IEEE, 2021, 1-7. https://ieeexplore.ieee.org/document/9698641
[9]. S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions (version 2), 2017. https://doi.org/10.48550/ARXIV.1705.07874
[10]. M. Haklay, P. Weber, Openstreetmap: User-generated street maps. IEEE Pervasive computing, 7(4), 2008, 12-18.
[11]. S. Marciuska, J. Gamper, Determining Objects within Isochrones in Spatial Network Databases. Advances in Databases and Information Systems, 2010, 392–405. https://doi.org/10.1007/978-3-642-15576-5_30
[12]. F. Lorenz, J. Willwersch, M. Cajias, F. Fuerst, Interpretable Machine Learning for Real Estate Market Analysis, 2021. DOI:10.13140/RG.2.2.10990.13120
[13]. E. Guliker, E.Folmer, M. van Sinderen, Spatial Determinants of Real Estate Appraisals in the Netherlands: A Machine Learning Approach. ISPRS International Journal of Geo-Information, 11(2), 2022. https://doi.org/10.3390/ijgi11020125
[14]. D. Tchuente, S. Nyawa, Real estate price estimation in French cities using geocoding and machine learning. Annals of Operations Research, 308 (1), 2021, 1-38. DOI:10.1007/s10479-021-03932-5
[15]. M. Steurer, R.J. Hill, N. Pfeifer, Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38(2), 2021, 99-129. https://doi.org/10.1080/09599916.2020.1858937
[16]. L. Chen, X. Yao, Y. Liu., Y. Zhu, W. Chen, X. Zhao, T. Chi, Measuring impacts of urban environmental elements on housing prices based on multisource data — a case study of Shanghai, China. ISPRS International Journal of Geo-Information, 9 (2), 2020. https://doi.org/10.3390/ijgi9020106
[17]. O. Poursaeed, T. Matera, S. Belongie, Vision-based real estate price estimation. Machine Vision and Applications, 29 (4), 2018, 667-676. https://link.springer.com/article/10.1007/s00138-018-0922-2

Henrik Tigran Sergoyan *(Germany, Munich) – Technical University of Munich, Mathematics department, Master student, henrik.sergoyan@tum.de*

Grigor Vahan Bezirganyan *(Germany, Munich) – Technical University of Munich, Mathematics department, Master student, grigor.bezirganyan@tum.de*

Received: 06.04.2022
Revised: 22.04.2022
Accepted: 03.06.2022