ISSN 1829-0043 PROCEEDINGS OF ENGINEERING ACADEMY OF ARMENIA (PEAA). 2019. V.16, N 2

UDC 519.862 **ECONOMICS AND MANAGEMENT**

A.A. Asryan

PREDICTING THE APARTMENT PRICES IN YEREVAN CITY

In this paper we carry out a cross-sectional study to design an apartment price estimation model for Yerevan city. The data is scraped from a real estate website. Several machine learning algorithms are compared based on their goodness of fit and predictive power on training and testing sets respectively, the best one being the random forest algorithm. Recommendations for further research are provided.

Keywords: *apartment price estimation, goodness of fit, predictive accuracy, cross validation, random forest algorithm.*

Introduction. Apartment price estimation is a widely researched topic as it is of great importance to the real estate market participants (prospective homebuyers, real estate investors, appraisers, mortgage lenders, etc.). Having accurate and up-to-date apartment price estimation models helps the interested parties conduct timely evaluations of various apartments in hand, thereby increasing the efficiency of their economic decisions.

Buying an apartment is considered one of the most stable and profitable investment alternatives in Armenia, especially in Yerevan. There were 978 apartment buy/sell transactions in Yerevan only in the April of 2019, up to 18.4% from March, 2019 and up to 18.4% from April, 2018. Meanwhile, the price of a square meter of apartment space in Yerevan in April 2019 increased by 0.8% from March, 2019 and 9.8%, when compared to April, 2018 [1]. As the apartment market in Yerevan is thriving, accurate and timely estimation of the apartment prices becomes a crucial factor for stakeholders. Traditional price estimation is based on cost and sale price comparison and lacks timeliness. Therefore, the availability of an apartment price prediction model would help the interested parties make rapid estimations and would improve the efficiency of the overall real estate market.

Apartment price estimation is a widely researched topic. Early works in the field tended to take a single-model approach, focusing on the choice of the determinants. Can (1992) examined the traditional and spatial autoregressive hedonic urban housing price models corresponding to different conceptualizations of the housing price determination process [2]. Day (2003) designed distinct hedonic pricing models for different property submarkets in Glasgow that were identified using the hierarchical clustering approach [3]. The more recent approaches, however, are inclined toward the use of machine learning algorithms and their combinations for the purpose: support vector regression [4], random forest [5, 6], support vector machine [7,8], artificial neural networks [9] and others.

In this paper, we attempt to design an apartment price estimation model for Yerevan city. To achieve this, we employ several machine learning algorithms and compare them based on their predictive performance on the given data. The best performing algorithm is suggested as the basis for the apartment price estimation model.

Data. The data is obtained by collecting information about the apartments on sale in Yerevan from the website www.myrealty.am as of April 30, 2019. It consists of 6049 observations and 70 features. Some of the features (such as **ID**, **Date_Added**, **Date_Edited**, etc.) are dropped as they had no effect on the prices of apartments. We perform greedy feature elimination (first, we identify the independent variable pair with the highest absolute correlation coefficient, then the variable of the pair which has the lower correlation with the dependent variable is eliminated) using Cramer's V statistic and remove the redundant features. Using the **Latitude** and the **Longitude** features of the data, we engineer a new feature **dist** by calculating the distance of each apartment from the selected center (see Fig. 1) of the city.

Fig. 1. The selected center (40°11'13.9"N 44°30'54.7"E) of Yerevan

We can see from the plots depicted in Fig. 2 that there are clearly outliers in the data. In order to identify and remove these observations, we employ the Mahalanobis distance, which is a measure of the distance between an observation x and the distribution of observations. To conclude, after feature engineering and data cleaning, we end up with 5750 observations and 57 attributes (see Table 1).

Table 1

List of Attributes

Feature	Type	Feature	Type
Price (dependent)	integer	Fireplace	binary
Room quantity	integer	Garage	binary
Area (in sqm)	integer	Laminate flooring	binary
Floor	integer	Roadside	binary
Storeys	integer	Fence	binary
Bathroom quantity	integer	Tile	binary
Building type	Factor with 4 levels	High first floor	binary
Ceiling height	Float	Park	binary
Condition	Ordered Factor with 4 levels	Building existence	binary
District	Factor with 12 levels	Equipment	binary
Hot water	binary	Heater	binary
Heating	binary	Storage room	binary
Gas	binary	Security system	binary
Central heating	binary	Sunny	binary
Electricity	binary	Close to the bus station	binary
Irrigation	binary	Bilateral	binary
Sewerage Canalization	binary	Balcony	binary
Internet	binary	Iron door	binary
Hot Water	binary	Heated floor	binary
Air conditioner	binary	Playground	binary
Water 24/7	binary	Attic	binary
Parking	binary	Basement	binary
Gate	binary	Grating	binary
Loggia	binary	Elevator	binary
Gym	binary	View	binary
Sauna	binary	Parquet	binary
Open balcony	binary	Furniture	binary
Swimming pool	binary	dist	float
Euro windows	binary		

In order to prepare the data for some of the models, we employ the Ordered Quantile (ORQ) normalization transformation [10] for the continuous variables in hand (*Price_total*, *Area_sqm*, *dist*). To complete the data preparation phase, we also examine the correlations between the features. In particular, the associations between the continuous variables are of great importance for the modeling phase. As we can see in Fig. 3, there is a statistically significant pairwise correlation between the independent continuous variables *Area_sqm* and *dist* and the dependent variable *Price_total*.

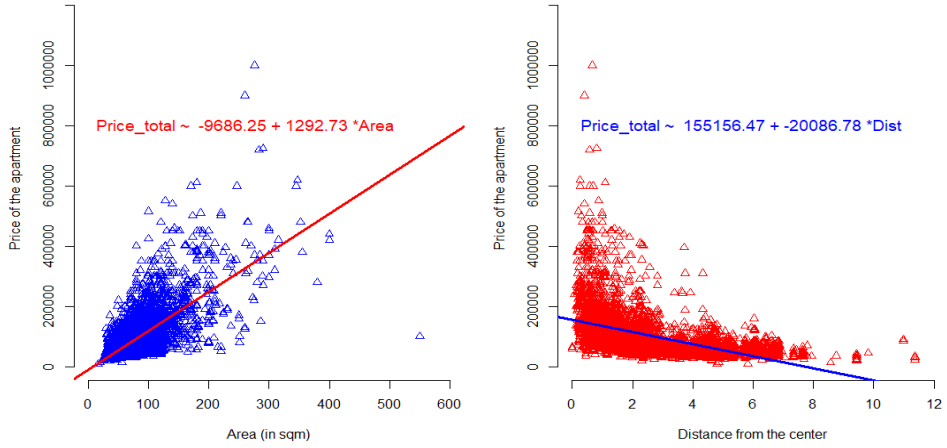

Fig. 2. Plotting Price against Area and Distance

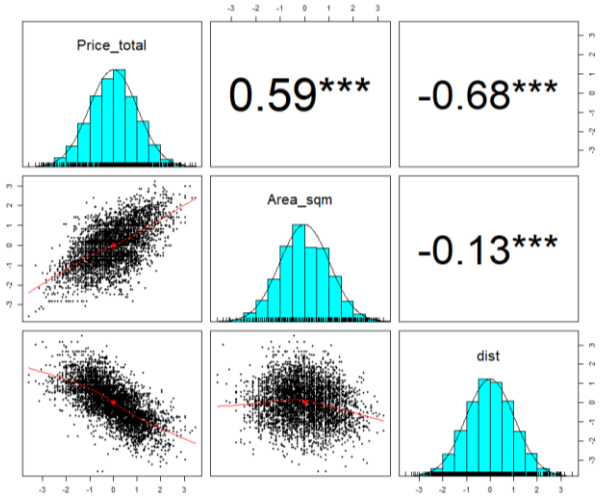

Fig. 3. Correlation between the Normalized features

Model Fitting (Summary and Visualization). In this section, we provide short descriptions of models and their fitting details (where applicable), and in the next section we compare the models based on their performance on both training and testing sets. The cleaned data is split into training and testing sets (training:testing ratio = 0.6:0.4). The training set is used for fitting the models, while the testing set is later utilized for comparing the fitting accuracy and the predictive power of the models in question.

1. Linear Regression (LinReg): The first model we employ for explaining and predicting the apartment prices is the multivariate linear regression. The results are shown in Table 2 below.

Table 2

Summary of the Linear Regression

Observations	3,450
R2	0.810
Adjusted R2	0.806
Residual Std. Error	0.440 (df = 3380)
F Statistic	208.551*** (df = 69; 3380)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2 indicates that the linear regression with the 3450 observations and 57 features provides a good fit to the training data (adjusted R^2 = 0.806).

2. Improved Linear Regression (LinRegCook): In order to make the LinReg model more robust, we used Cook's Distance [11] to identify and remove 146 influential outliers from the training set. The results are shown in Table 3 below.

Table 3

Summary of the Improved Linear Regression

Observations	**3,304**
R2	**0.843**
Adjusted R2	**0.840**
Residual Std. Error	**0.390 (df = 3237)**
F Statistic	**262.842*** (df = 66; 3237)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

It can be noted from Table 3 that the linear regression with 3304 observations and 57 features now provides a slightly better overall fit to the training data (adjusted R^2 = 0.840).

3. Ridge Regression (RidgeReg): The next model employed is the Ridge regression, which uses $L2$ regularization to weight/penalize residuals in the parameter-learning phase, returning a model that generalizes better because it is less sensitive to extreme variance in the data, such as outliers [12]. Technically, the Ridge regression is formulated as the penalized residual sum of squares (PRSS) shown in equations (1) and (2):

$$PRSS(\beta)_{L2} = \sum_{i=1}^{n}\left(y - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = (Y - X\beta)^T(Y - X\beta) + \lambda\|\beta\|_2^2, \quad (1)$$

where Y – independent variable, X – independent variable(s), β – coefficients, λ – penalty term.

The solution to the PRSS above is:

$$\hat{\beta}_{\lambda}^{ridge} = \left(X^TX + \lambda I_p\right)^{-1}X^TY. \quad (2)$$

4. Random Forest Regression (RandForest): The next algorithm employed is the Random Forest - a supervised learning algorithm that operates by constructing a multitude of decision trees at the training phase and outputting the mean prediction (regression) of the individual trees. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents overfitting [13].

5. Regression Tree (RegTree): The next algorithm – the regression tree is built through binary recursive partitioning: an iterative approach that divides the initial data into branches, and then keeps on splitting each partition into smaller groups until each node drops to the minimum node size (defined by the user) and turns into a terminal node [14]. Fig. 4 shows the regression tree fitted to the training set.

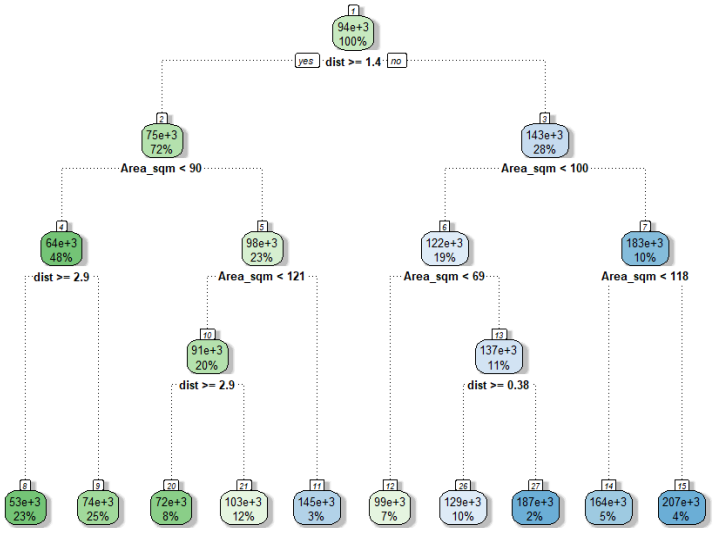

Fig. 4. Fitted Regression Tree

6. Neural Network (NeuralNet): For the neural network, we choose a three-layer structure {5, 3, 2} using the pruning algorithm. We select the resilient backpropagation algorithm with weight backtracking [15] to adjust the connection weights. The neural network fitted to the training set is shown in Fig. 5.

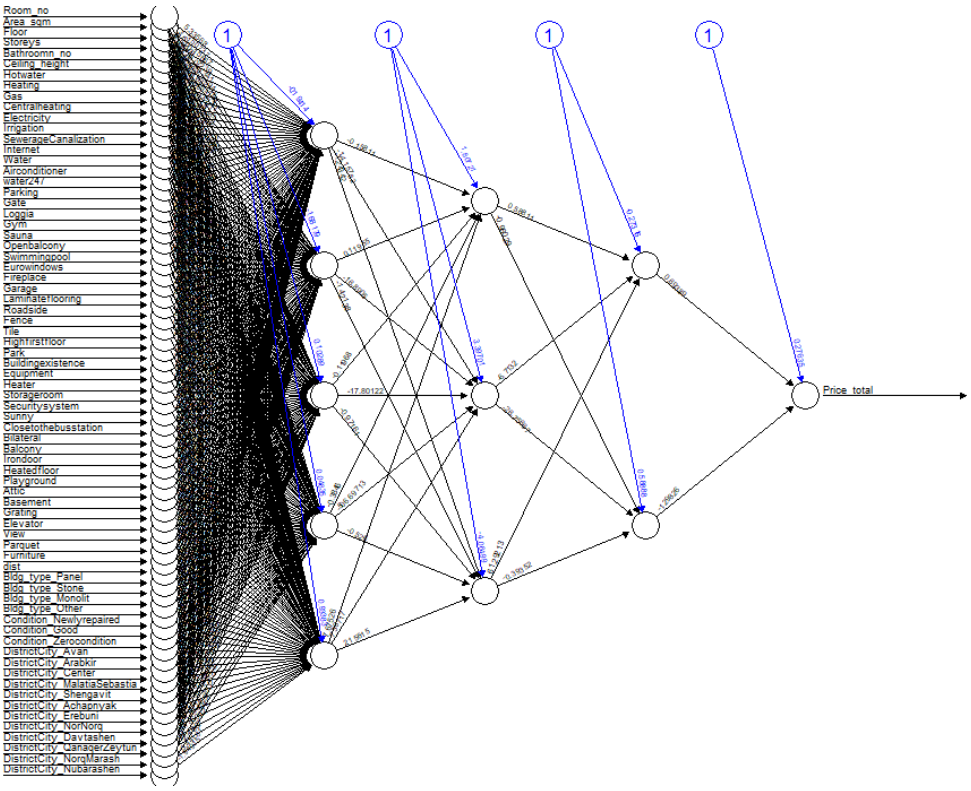

Fig. 5. Fitted Neural Network {5, 3, 2}

Model Performance Evaluation. In this section, we examine the performance of the aforementioned models. The accuracy measures used for comparing the employed models in terms of the goodness of fit and the predictive power are the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Squared Error (MSE). Table 4 shows that the best fit to the training data is achieved by the Improved Linear Regression (LinReg+CooksD) model, with the Random Forest algorithm being the second one. This is mainly due to the fact that the influential outliers (with respect to the multiple linear regression model) have been removed from the training data using the Cook's D.

Table 4

Fitting Accuracy of the Employed Models

	Fitting Accuracy Measure			
Model	**RMSE**	**MAE**	**MAPE**	**MSE**
LinReg	25126.81	16255.74	0.164476	631356741
LinReg+CooksD	**22078.95**	**14664.19**	**0.1492674**	**487480058**
RandomForest	23803.52	15167.73	0.1608418	566607450
RegressionTree	30203.20	21004.92	0.2409165	912233362
RidgeRegression	25148.19	16230.76	0.1642565	632431511
NN_ResProp+	25751.71	15788.91	0.1698422	663150636

However, when comparing the predictive power of the models (see Table 5), the Random Forest model achieves the best result, with the Improved Linear Regression being the second one. This is due to the fact that the Random Forest model can capture hidden non-linear relations between the price and features of apartments. In addition, it is worth mentioning that, despite increasing the fitting accuracy of the linear regression model, removing the 146 influential observations (using the Cook's Distance) from the training data causes the model to have less predictive power due to valuable information loss.

Table 5

Prediction Accuracy of the Employed Models

Model	Prediction Accuracy Measure			
	RMSE	**MAE**	**MAPE**	**MSE**
LinReg	24979.71	16296.38	0.1616112	623985973
LinReg+CooksD	24859.3	16209.86	0.161051	617985040
RandomForest	**23134.53**	**14700.86**	**0.1518078**	**535206358**
RegressionTree	31642.13	21746.19	0.2399154	1001224501
RidgeRegression	25017.16	16291.29	0.1613606	625858312
NN_ResProp+	25531.65	15503.68	0.1638713	651865172

In order to solidify our claim on the Random Forest being the best predictive algorithm for the given data, we conducted cross validation which is a resampling method used to evaluate the predictive capabilities of given models given a limited data set. We chose the following parameters for the cross validation: training:testing ratio = 0.9:0.1, number of resamplings $k = 10$. The choice of k is based on the fact that $k = 10$ has been found through experimentation to generally result in a model capability estimate with modest variance and low bias. As can be seen from Figs. 6, 7, 8, 9 below, the prediction model based on the Random Forest algorithm shows the best predictive power with the least mean predictive errors calculated upon 10 resamplings.

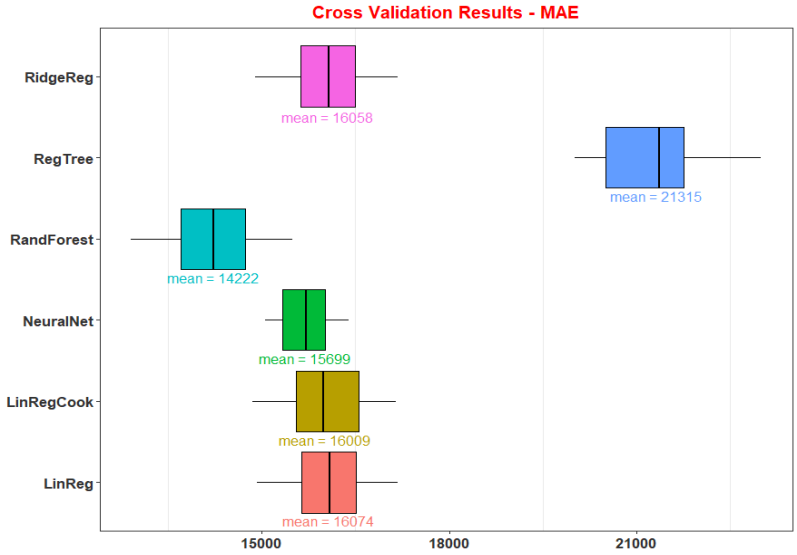

Fig. 6. Cross Validation Results – MAE Comparison

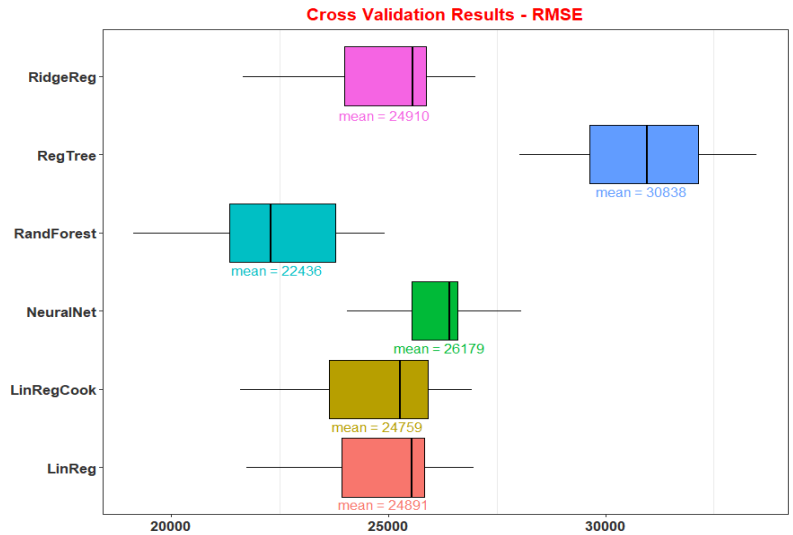

Fig. 7. Cross Validation Results – RMSE Comparison

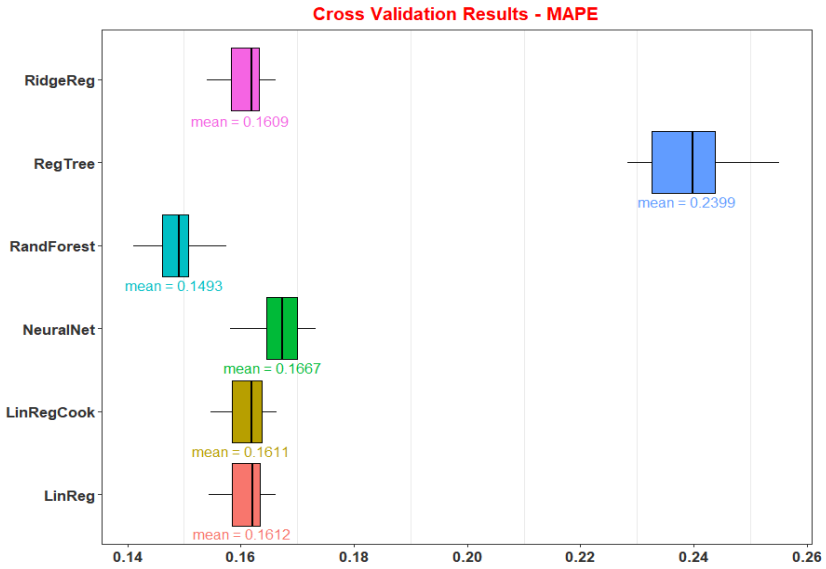

Fig. 8. Cross Validation Results - MAPE Comparison

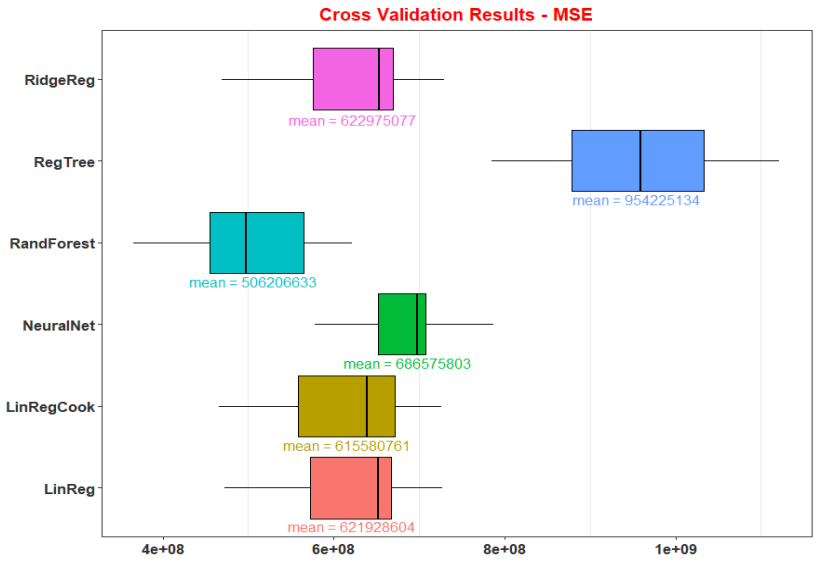

Fig. 9. Cross Validation Results - MSE Comparison

Conclusion and Future Research. We conclude from the findings above that the random forest algorithm performs best and should be chosen as a base for a large-scale apartment price estimation model for Yerevan. However, the findings of this study have to be seen in light of some limitations. First, we used data from only one real estate website (chosen source was the only one providing full and standardized information on features of apartments), which may bias the predictions of our model towards higher/lower values than the real prices in the market. Second, the features influencing the price formulation of the apartments were limited to those presented by the website, meaning that there might be other factors influencing the price of the apartments that were not included in the apartment listings online.

The research framework adopted for this study did not take into account a very important category of information held by the website – apartment photos. In particular, with the use of deep learning techniques, these images can be utilized to engineer and employ new features (for instance, the convenience level of the apartment) which might significantly increase the predictive power of the selected models. Therefore, it is suggested that, for future research, apartment images should be examined for possible incorporation into the price prediction models.

References

1. **Cadastre** Committee of the Republic of Armenia, The Real Estate Market of the Republic of Armenia in April of 2019, https://www.cadastre.am/storage/files/news/news_5859301067_1_Hodvac_04.2019.pdf, Retrieved 19.07.2019.
2. **Can, A.** Specification and estimation of hedonic housing price models // Regional science and urban economics.- 1992. - 22(3). - P. 453-474.
3. **Day, B.** Submarket identification in property markets: a hedonic housing price model for Glasgow (No. 03-09). - CSERGE Working Paper EDM, 2003.
4. **Li, D. Y., Xu, W., Zhao, H., & Chen, R. Q.** A SVR based forecasting approach for real estate price prediction // IEEE International Conference on Machine Learning and Cybernetics. – 2009. -Vol. 2. - P. 970-974.
5. **Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B.** Estimating the performance of random forest versus multiple regression for predicting prices of the apartments // ISPRS International Journal of Geo-Information. - 2018. - 7(5). - P. 168.
6. **Wang, C., & Wu, H.** A new machine learning approach to house price estimation // New Trends in Mathematical Sciences. - 2018. - 6(4). - P. 165-171.
7. **Jirong, G., Zhu, M., & Jiang, L.** Housing price forecasting based on genetic algorithm and support vector machine // Expert Systems with Applications. - 2011. - 38(4). - P. 3383-3386.
8. **Wang, X., Wen, J., Zhang, Y., & Wang, Y.** Real estate price forecasting based on SVM optimized by PSO // Optik-International Journal for Light and Electron Optics. – 2014. - 125(3). - P. 1439-1443.
9. **Selim, H.** Determinants of house prices in Turkey: Hedonic regression versus artificial neural network // Expert systems with Applications. – 2009. - 36(2). - P. 2843-2852.
10. **Bartlett, M. S.** The use of transformations // Biometrics. – 1947. - 3(1). - P. 39-52.
11. **Cook, R. D.** Detection of influential observation in linear regression // Technometrics. - 1977. - 19(1). - P. 15-18.
12. Regularization: Ridge Regression and the LASSO (29.11.2006), http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf, Retrieved 21.04.2019.
13. **Chakure** A. (29.06.2019), Random Forest Regression. https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f, Retrieved 23.07.2019.
14. Regression Trees, https://www.solver.com/regression-trees, Retrieved 20.04.2019.
15. **Riedmiller, M., & Braun, H.** A direct adaptive method for faster backpropagation learning: The RPROP algorithm // Proceedings of the IEEE international conference on neural networks. – 1993. - P. 586-591.

14.01.2020.

Ա.Ա. Ասրյան

ԵՐԵՎԱՆ ՔԱՂԱՔԻ ԲՆԱԿԱՐԱՆՆԵՐԻ ԳՆԵՐԻ ԿԱՆԽԱՏԵՍՈՒՄ

Սույն հետազոտության նպատակն է նախագծել բնակարանների գների կանխատեսման մոդել Երևան քաղաքի համար: Տվյալները հավաքագրվել են անշարժ գույքի առքուվաճառքի մասնագիտացված կայքից: Համեմատվում են մեքենայական ուսուցման մի քանի ալգորիթմներ՝ հիմք ընդունելով դրանց համապատասխանության որակը և կանխատեսման ճշգրտությունը տվյալների համապատասխանաբար ուսուցողական և փորձարկային բազմությունների վրա, որոնցից լավագույն արդյունքն ապահովում է պատահական անտառի ալգորիթմը: Ներկայացվում են այս ուղղությամբ հետագա ուսումնասիրությունների առաջարկություններ:

Առանցքային բառեր. *բնակարանի գնի կանխատեսում, համապատասխանության որակ, կանխատեսման ճշգրտություն, խաչաձև վավերացում, պատահական անտառի ալգորիթմ:*

A.A. Асрян

ПРОГНОЗИРОВАНИЕ ЦЕН НА КВАРТИРЫ В ГОРОДЕ ЕРЕВАНЕ

Целью данного исследования является разработка модели прогнозирования цен на квартиры для города Еревана. Данные были извлечены из сайта о недвижимости. Проведено сравнение нескольких алгоритмов машинного обучения на основе их качеств подгонки и точности прогнозирования на обучающих и тестовых наборах соответственно, лучший результат из которых показывает алгоритм случайного леса. Представлены рекомендации для дальнейших исследований.

Ключевые слова: *прогнозирование цены на квартиру, качество подгонки, точность прогнозирования, перекрестная проверка, алгоритм случайного леса.*

Asryan Arman Ashot – Ph.D. student, Chair of Management and Business (YSU)