ЖАРАТЫЛЫСТАНУ ЖӘНЕ ТЕХНИКАЛЫҚ ҒЫЛЫМДАР

NATURAL AND TECHNICAL SCIENCES

IRSTI 27.47.23

D. Zhumabek[1], A. Zhailaubek[2], A. Temirali[3], Zh. Rayev[4]
[1,2,3,4]Suleyman Demirel University, Kaskelen, Kazakhstan

USING MACHINE LEARNING CLASSIFICATION ALGORITHMS TO STUDY HOUSE PRICE FOR ALMATY

Abstract. In real estate valuation and house market research, house prices and rental value are generally analyzed by decision tree regression and random forest regression model based on machine learning. Regression model examines the effect of characteristics of goods on their prices. Factors that determine the house prices in Almaty are analyzed in this paper using real dataset from legal site. The most important variables that affect house rents are type of house, type of building, number of rooms, size, and other structural characteristics such as water system, pool, natural gas. Also used jupyter notebook, numpy, pandas, matplotlib, scipy and scikit-learn.

Keywords: house price, decision tree, random forest regression model, Almaty.

***

Аннотация. В оценке недвижимости и исследовании рынка жилья цены на жилье и стоимость аренды обычно анализируются с помощью регрессии дерева решений и модели случайной регрессии леса, основанной на машинном обучении. Модель регрессии исследует влияние характеристик товаров на их цены. Факторы, которые определяют цены на жилье в Алматы, анализируются в настоящем документе с использованием реального набора данных с юридического сайта. Наиболее важными переменными, которые влияют на арендную плату дома, являются тип дома, тип здания, количество комнат, размер и

другие структурные характеристики, такие как система водоснабжения, бассейн, природный газ. Также используется блокнот jupyter, numpy, pandas, matplotlib, scipy и scikit-learn.

Ключевые слова: цена дома, дерево решений, модель случайной лесной регрессии, Алматы.

***

Аңдатпа. Жылжымайтын мүлікті бағалауда және тұрғын үй нарығын зерттеуде тұрғын үй бағасы мен жалдау құны әдетте регрессия ағашы және машиналық оқытуға негізделген кездейсоқ регрессиясы орман моделі арқылы талданады. Регрессия моделі тауарлар сипаттамаларының олардың бағасына әсерін зерттейді. Алматыдағы тұрғын үй бағасын анықтайтын факторлар осы мақалаға заңды сайттан нақты деректер жинағын пайдалану арқылы талданады. Үйдің жалға алу ақысына әсер ететін ең маңызды айнымалы болып табылады: үй түрі, ғимарат түрі, бөлме саны, мөлшері және басқа да құрылымдық сипаттамалары сияқты су мен жабдықтау жүйесі, бассейн, табиғи газ. Сондай-ақ, jupyter, numpy, pandas, matplotlib, scipy және scikit-learn блокноты пайдаланылады.

Түйін сөздер: үй бағасы, регрессия ағашы, кездейсоқ регрессиясы орман моделі, Алматы.

I. Introduction

Facing the upcoming era of big data, more and more people begin to engage in data analysis and mining. Machine learning [1], as a common means of data analysis, has gotten more and more attention. People of different industries are using machine learning algorithms to solve the problems based on their own industry data [2, 3]. Experts in the field of industry used machine learning in pattern recognition [4] and fault diagnosis [5, 6]. People in the field of economy began to use machine learning algorithms in economic modeling [7, 8]. The advantages of these algorithms were taken by the specialist in the field of aerospace in the aspect of classification and prediction [9]. Researchers in the field of construction combined machine learning methods with the professional domain knowledge of construction industry. Many intelligent systems are used in the construction industry, and lots of them have achieved good economic and social benefits. In general, the budget of construction project and benefit analysis of construction are usually gotten by the experience of the professionals and the construction of traditional models. If

the housing values can be accurately predicted, the government can make a reasonable urban planning [10].

In this paper, the background and current situation of the application of machine learning methods are firstly introduced. The development of its application in construction and real estate value is also expounded. Then, several machine learning algorithms which are involved in this paper are introduced. The mathematics process of them is described in detail. In this study, we apply the classification algorithms of machine learning known to us as Linear Regression. Also used dataset taken from legal site of Almaty, namely krisha.kz. According to the prediction results of these several methods, a discussion is made.

II. Main part

The house price estimation have a direct relationship with its quality and other sizes of estimation, for example house location, level, material play enormous role for house price. In this regard, this paper analyzes the main indicators that affect the estimation of prediction, which are the square feet, number of rooms, levels, location and house material made by.

General information about dataset.We can download all codes and dataset here:

https://github.com/ZhumabekDildar/diss/blob/master/data_scraping.ipynb.

The training data has 1426 observations and 6 explanatory variables and the test file has 117 observations and 5 explanatory variables. (Table1).

Here is the data description:

- Price: the property's sale price in tenge. This is the target variable that we are trying to predict.
- Square feet: lot size of home.
- Region: 6 regions of Almaty city.
- Levels: floor of house.
- Sanuzel: bathrooms of each house.
- States: plight of home.
- Security: video surveillance or fence of house.

Out[43]:

	price	square	region	levels	san	states	secur
0	30000000	278 м²	Наурызбайский	3	3 м	черновая отделка	4 сот.
1	110000000	270 м²	Медеуский	2	2 с/у и более	евроремонт	сигнализация
2	77000000	440 (219) м², кухня — 40 м²	Алатауский	4	2 с/у и более	евроремонт	сигнализация, видеонаблюдение, видеодомофон
3	73000000	320 м²	Наурызбайский	3	3 м	2 с/у и более	евроремонт
4	19500000	100 м², кухня — 18 м²	Жетысуский	1	3 м	раздельный	хорошее
5	47000000	225 (174) м², кухня — 18 м²	Медеуский	2	2.7 м	раздельный	хорошее
6	22000000	150 м²	Ауэзовский	1	2.5 м	раздельный	хорошее
7	10000000	52 (45) м², кухня — 6 м²	Турксибский	1	совмещенный	требует ремонта	5.28 сот.
8	269000000	470 м², кухня — 20 м²	Бостандыкский	3	3.2 м	2 с/у и более	евроремонт
9	18000000	100 (155) м², кухня — 75 м²	0	1	3 м	черновая отделка	мягкая кровля

Table1. General information about dataset

	price	square	region	levels	san	states
143	15000000	70.0	Медеуский	1	во дворе	среднее
654	21500000	78.0	Алмалинский	1	во дворе	среднее
963	8000000	90.0	0	1	во дворе	среднее
1039	15500000	202.6	Алатауский	2	во дворе	среднее
1040	30000000	200.0	Турксибский	2	во дворе	среднее

Table2. After preproccessing

In [88]:
```
dset = dset.drop(columns=['region_0'], axis=1)
dset.head()
```

Out[88]:

	price	square	region_Алатауский	region_Алмалинский	region_Ауэзовский	region_Бостандыкский	region_Жетысуский	region_Медеуский	region_
143	15000000	70.0	0	0	0	0	0	1	
654	21500000	78.0	0	1	0	0	0	0	
963	8000000	90.0	0	0	0	0	0	0	
1039	15500000	202.6	1	0	0	0	0	0	
1040	30000000	200.0	0	0	0	0	0	0	

5 rows × 23 columns

Table3. This table show us information about dates choose one explanatory variable

In [92]:
```
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.ensemble import RandomForestRegressor
```

In [101]:
```
acc_list_abs = []
acc_list_sq = []

counter = 1

for i in range(3,11):
    tree_ = DecisionTreeRegressor()
    grid_tree = GridSearchCV(tree_, param_grid=params, cv=i, n_jobs=-1)
    grid_tree.fit(X_train, y_train)
    acc_list_abs.append(mean_absolute_error(y_test,grid_tree.best_estimator_.predict(X_test)))
    acc_list_sq.append(mean_squared_error(y_test,grid_tree.best_estimator_.predict(X_test)))
    print(counter)
    counter+=1
```

```
1
2
3
4
5
6
7
8
```

```
In [102]: import matplotlib.pyplot as plt
          %matplotlib inline

In [107]: plt.plot(range(3,11), acc_list_abs, color='r')
          #plt.plot(range(3,11), acc_list_sq, color='b')
          plt.xlabel("Cross-val score")
          plt.ylabel("Errors")

Out[107]: Text(0, 0.5, 'Errors')
```

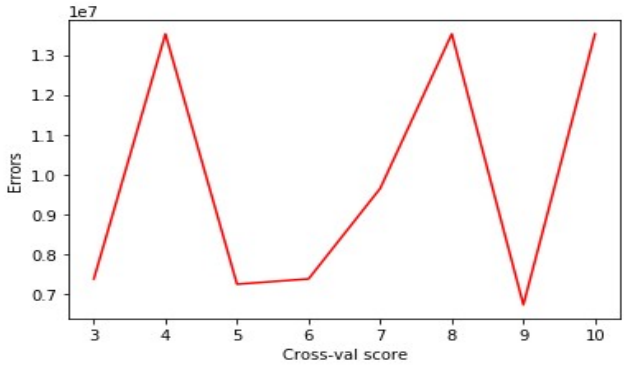

Table4. Error of Decision tree

```
In [110]: acc_list_abs = []
          acc_list_sq = []

          counter = 1

          for i in range(3,11):
              rand_ = RandomForestRegressor()
              grid_rand = GridSearchCV(rand_, param_grid=params, cv=i, n_jobs=-1)
              grid_rand.fit(X_train, y_train)
              acc_list_abs.append(mean_absolute_error(y_test,grid_rand.best_estimator_.predict(X_test)))
              #acc_list_sq.append(mean_squared_error(y_test,grid_tree.best_estimator_.predict(X_test)))
              print(counter)
              counter+=1
```

```
In [111]: plt.plot(range(3,11), acc_list_abs, color='r')
          #plt.plot(range(3,11), acc_list_sq, color='b')
          plt.xlabel("Cross-val score")
          plt.ylabel("Errors")

Out[111]: Text(0, 0.5, 'Errors')
```

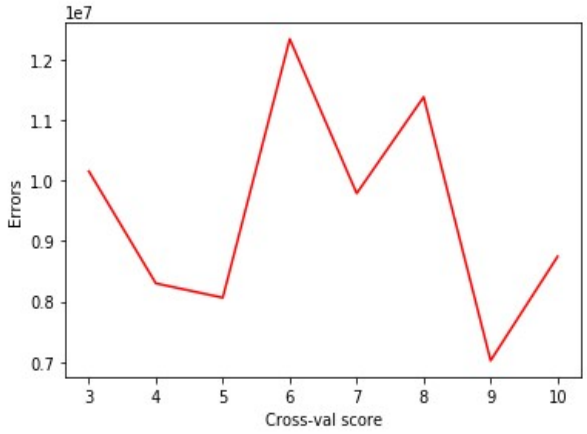

Table5. Error of Regression fores

III. Conclusion

In this paper, linear regression algorithm is used in the field of construction to predict the housing value. According to multiple characteristics, the housing value of Almaty suburb is forecasted. The models of several machine learning methods should be constructed and analyzed at first and then combined with the corresponding characteristics of testing data to predict the housing value. The prediction results of various machine learning approaches are not the same. Aiming at the regression decision tree and regression forest has better prediction effect, learning ability, and generalization ability. Table 4 and table 5 shows that when we divide the cross value score into 9 parts it gives us lowest error figure which is equal to 0.7.

Refenences

1 Bishop, C. *Pattern Recognition and Machine Learning*. Springer, 2006. pp. 738.

2 Yin, S., Wang, G., Karimi, H. Data-driven design of robust fault detection system for wind turbines. *Mechatronics*, 24 (4), (2014): pp. 298–306.

3 Yin, S., Yang, X., Karimi, H.R. Data-driven adaptive observer for fault diagnosis. *Mathematical Problems in Engineering*, (2012): pp. 21.
4 Bhagat, P. *Pattern Recognition in Industry*, Elsevier, 2005. pp. 220.
5 Widodo, A., Yang, B. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21 (6), (2007): pp. 2560–2574.
6 Yin, S., Li, X., Gao, H., Kaynak, O. Data-based techniques focused on modern industry: an overview. *IEEE Transactions on Industrial Electronics*, (2014): pp. 1-17.
7 Huang, C.L., Chen, M.C., Wang, C.J. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications,* 33 (4), (2007): pp. 847–856.
8 Ding, S., Yin, S., Peng, K., Hao, H., Shen, B. A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill. *IEEE Transaction on Industrial Informatics*, 9 (4), (2013): pp. 2239–2247.
9 Staszewski, W.J., Mahzan, S., Traynor, R. Health monitoring of aerospace composite structures—active and passive approach. *Composites Science and Technology*, 69 (11-12), (2009): pp. 1678–1685.
10 Mu, J., Wu, F., Zhang, A. Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis*, 4 (2014): p. 7.