INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 2 - V8I2-1176)
Available online at: https://www.ijariit.com

House price prediction using Machine Learning

V. V. Sai Pavan Pindiproli
pvvsp1234@gmail.com
The ICFAI Foundation for Higher Education, Hyderabad, Telangana

A. Abhishek Reddy
abhishekreddy0421@gmail.com
The ICFAI Foundation for Higher Education, Hyderabad, Telangana

ABSTRACT

The main purpose of the paper is to show the use of linear regression to estimate the house prices prediction in Hyderabad a city in Telangana state India. Nowadays, home is a basic amenity for most of human beings. The biggest dream of all middle-class people is to have their own home. But today, the prices of plots, flats and homes have become so high that the general public could not afford them. At the coin second side the sellers are unable to find genuine buyers and the prediction of house prices is nightmare for middle class people, these all issues are created because of land brokers and land brokers sell land or homes or land for more money and take more commission. The applications like magic bricks and no broker are providing data to public, that data is to estimate the current prices of any locality. In this article the prediction of house price taken as problem. A home price forecasting method that collects past home prices and it predicts the current price of the house. The machine learning techniques are used in the analysis to estimate the prices of future. The Python language is used to analyze the problem and various regression models to ensure accurate predictions. The Home Price Index usually represents the sum of price fluctuations in residential real estate. However, to predict the price ofa home, based on location, house type, size, year of construction, local amenities, and some other factors that may affect the supply and demand of the house. We need a more accurate method in this analysis we developed more accurate method by applying various algorithms.

Keywords—*House Price, Random Forest, Xgboost, Regression Methods, Gradient Boosting*

1. INTRODUCTION

Computer learning algorithms are used in modeling, where machines learn from data and apply what they learnto predict new data. Regression is the most common model used for predictive analytics. The approaches proposed to reliably predict future outcomes apply to business, economics, finance, healthcare, e-commerce, entertainment, sports and more. Several criteria are considered in the strategy for predicting real estate prices. In metropolitan areas like Hyderabad, potential homebuyers consider location, property size, accessibility to parks, schools, hospitals, and power plants. Most important is the cost of the house. Multiple regression is a statistical approach for determining the relationship between a large number of independent variables and (dependent) target variables. To predict pricing, it is common to use regression techniques to develop models based on large numbers of inputs. In this study, we sought to create a regression model for predicting home prices. We also looked at common least squares, random forests, gradient boosting models, and XGBoost regression models. A comparative study of evaluation indicators was also conducted. You can use this model to predict the monetary value of that particular Hyderabad residential property after finding a reasonable fit.

2. LITERATURE SURVEY

To extrapolate tweaks in mortgage lending, receivables, and housing affordability in specific geographic locations, housing economists are used [1]. HPI is ineffective in predicting the price of a given property becauseit is a reference number based on all transactions. Instead of only duplicate purchases from earlier decades, manyfactors such as district, age, and the number of floors must be addressed. Machine learning has become a critical prediction approach in recent seasons, thanks to the significant rise toward Big Data, because it can estimate property values more correctly based on their qualities, regardless of prior year's data. Numerous publications investigated this issue and shown the capabilities of the machine learning approach [2],[3],[4], but the vast number of them compare the effectiveness of the models without taking into account the combination of several machine learning models. S. Lu et al. carried out an experiment on home price forecasting using a hybrid regression approach, although finding the best answer involves extensive parameter adjustment [5]. The Stacked Function approximation strategy [6],[7], a machine learning ensemble technique, was used to maximize the projected ability to make appropriate to the relevance of model combination. Q. Qiu retrieved and uploaded the "Housing Price in Beijing" dataset to Kaggle [8]. We were able to examine the reliability of each

particular technique by using various ways on this dataset. the test set in which uses the Stack Generalization approach, the lowest Root Mean Squared Logarithmic Error (RMSLE) is 0. 16350.Asset prices are affected by a number of things. Categories these elements into 3 number of categories for this process: investigation, thinking, and territory [9]. States of being are conductivity characterized by a house that can be seen by humans, such as the value of the house, the number of bedrooms, the easy accessibility of kitchen and parking space, the approachability of the yard nursery, the zone of land and configurations, and the age of the house [10], whereas an assumed is an idea rendered by design professionals to incentivize prospective customers, such as the likelihood of a moderate home, strong and green environment, and high - end restrictions [11].The expenditures of a home is heavily influenced by the zone in which it is located. This is because the zone determines the typicalland cost. Additionally, the remaining chooses the most direct route to open places of employment, such as institutions of higher learning, as well as grounds, emergency facilities, and prosperity centers. Workplaces that are focused on families, such as commercial areas, cuisine trips, and other entertainment venues, presentmagnificent beauty. [12, 13].

3. METHODOLOGY

The term "methodology" refers to the framework that is being used. It comprises different stages that must be met to attain the goal. In this analysis several data mining and machine learning techniques are used.

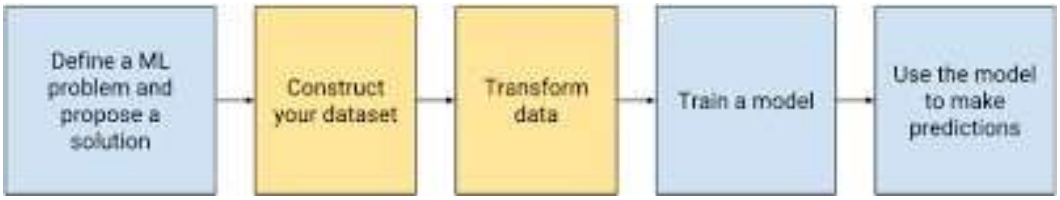

Figure 1: Preprocessing and Testing Model

The path to a social event is paved with data collection. the to estimate the data analysis on specific factors in a built in framework, which then enables one to address pertinent inquiries and assess outcomes. Information gathering is a part of research in various fields of study which includes the physical science , sociological sciences, humanities, and business. While strategies differ by discipline, the emphasis on ensuring precise and

legal selection remains the same. This dataset was discovered after reviewing a variety number of datasets. The selected data set is describes the house prices in the city of Hyderabad.

3.1 Data set:

The dataset exploited in this simulation is a realistic dataset. It has 1259 records with 11 characteristics that have the potential to influence property prices. However, the 11 characteristics were picked as being likely to effect house prices. Area in square meters, the data is collected from Hyderabad a city in Telangana State India, this data covers all parts of the city as the city is expanding in all sides, this city has small cities across all directions like Ibrahimpatnam, Shankarpalli, Patancheruvu, Ramoji film city, there is tremendous growth in the regions of the shakerpalli and patancheruvu regions the prices and other factors like room size number of wash rooms, and other 11 important parameters are taken in the dataset and the analysis is performed, these all parameters are assessing the house's overall condition and finish. Location, the number of bedrooms and baths, as well as the total number of people who will be staying in the house, The garage and the number of cars that can fit in it, status of the apartments, and finally prices are all factors to consider.

1	Area	BHK	Bathroom	Furnishing	Locality	Parking	Price	Status	Transaction	Type	Per_Sqft
2	800	3	2	Semi-Furnished	kondapur	1	6500000	Ready_to_move	New_Property	Builder_Floor	
3	750	2	2	Semi-Furnished	kondapur	1	5000000	Ready_to_move	New_Property	Apartment	6667
4	950	2	2	Furnished	kondapur	1	15500000	Ready_to_move	Resale	Apartment	6667
5	600	2	2	Semi-Furnished	kondapur	1	4200000	Ready_to_move	Resale	Builder_Floor	6667
6	650	2	2	Semi-Furnished	tellapur	1	6200000	Ready_to_move	New_Property	Builder_Floor	6667
7	1300	4	3	Semi-Furnished	tellapur	1	15500000	Ready_to_move	New_Property	Builder_Floor	6667
8	1350	4	3	Semi-Furnished	tellapur	1	10000000	Ready_to_move	Resale	Builder_Floor	6667
9	650	2	2	Semi-Furnished	tellapur	1	4000000	Ready_to_move	New_Property	Apartment	6154
10	985	3	3	Unfurnished	shadnagar	1	6800000	Almost_ready	New_Property	Builder_Floor	6154
11	1300	4	4	Semi-Furnished	shadnagar	1	15000000	Ready_to_move	New_Property	Builder_Floor	6154
12	1100	3	2	Semi-Furnished	gachibowli	1	6200000	Ready_to_move	New_Property	Builder_Floor	6154
13	870	3	2	Semi-Furnished	gachibowli	1	7700000	Ready_to_move	New_Property	Builder_Floor	6154
14	630	2	2	Semi-Furnished	gachibowli	1	5500000	Ready_to_move	New_Property	Builder_Floor	6154
15	660	2	2	Semi-Furnished	kukatpally	1	5000000	Ready_to_move	Resale	Builder_Floor	6154
16	344.4448	2	2	Semi-Furnished	kukatpally	1	3310000	Ready_to_move	Resale	Builder_Floor	6154
17	660	2	2	Semi-Furnished	kukatpally	1	4700000	Ready_to_move	New_Property	Builder_Floor	6154
18	550	2	2	Semi-Furnished	kukatpally	1	4500000	Ready_to_move	New_Property	Builder_Floor	6154
19	1100	4	3	Semi-Furnished	patancheru	1	17000000	Ready_to_move	New_Property	Builder_Floor	6154
20	1150	3	3	Semi-Furnished	patancheru	1	25000000	Ready_to_move	Resale	Apartment	6154

Figure 2: Data Set for House rate prediction in Telangana

3.2 Visualization of Data

The illustrative or image type representation of classified information is generally known as data visualization. It enables the analysis skills of difficult concepts or the identification of recent patterns. Many organizations regard Data Visualization as a modren likeness visual correspondence.

3.3 Cross Validation:

Cross validation is a strategic approach in which our network is developed using a subset of the dataset and thensurveyed using the basic subset of the dataset a short time later. In validation, training is performed on 50% of the dataset, while the remaining 50% is used for testing. The significant disadvantage of the approval strategy is that after it has been prepared for half of the dataset, it is possible that the remaining half may contain some useful data that was overlooked at the time of preparing the model.

4. ALGORITHM IMPLEMENTATION

Once the data is clean and we have gained insights into the dataset, we can apply a machine learning technique that is acceptable for our dataset. In our set of data, we chose four algorithms to estimate the model. The algorithms we chose are essentially classifiers, but we are training them to predict categorical data. Linear regression, random forest, gradient boosting Regression Technique, and XGBoosting Regressor are the four algorithms. These algorithms were created using Python's SciKit-learn Library.

4.1 Multiple linear regression

Multiple regression analysis is used to determine whether there is a statistically significant association between two sets of variables. It is used to find patterns in individuals' sets of data. Several relapses the exploration will be very similar. Likewise, the basic straight relapse. The main distinction between straight relapse and straight relapse is in the middle. There are also many relapses in the number of predictors ("x" variables) used within those relapses. Simple relapse examination employments Each subordinate "y" variable has an absolute x variable. Consider the following example: (x1, y1). Many relapses use multiple "x" variables for each free variable: (x1), (x2), (x3), (y1).

In one-variable straight regression, you can compare a subordinate variable (for example, "sales") to an autonomous variable (for example, "profit"). The x1 as a specific case type for claiming sales, and your x2 in the same way as well as sorting out deals, the multiple regression estimate is shown in the below equation

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_n x_n$$

The following figure3 represents the train and test data of Multiple linear regression

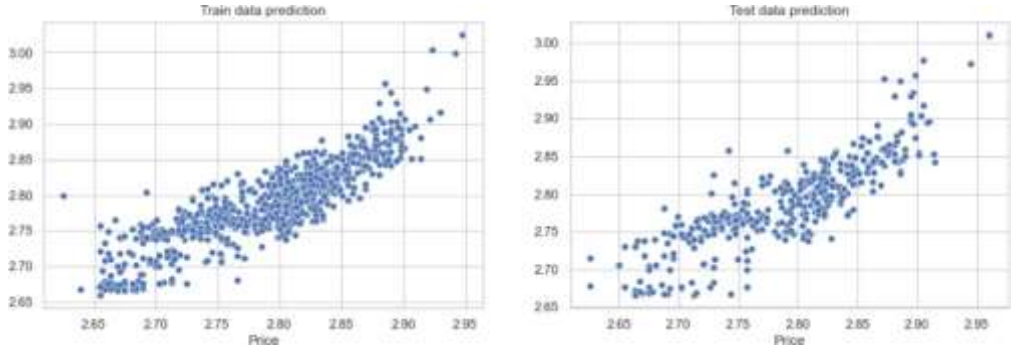

Figure 3: scatter plot of multiple linear regression

4.2 Gradient Boosting algorithm.

This is a learning algorithm that builds group-structured forecasting models from weak prediction models.A theoretical model's accuracy can be enhanced in two ways: Perhaps through comprehending core characteristic development on the other hand. Boosting calculations should be applied in a concise manner. Every boosting method requires its own theoretical underpinning. Furthermore, while using them simultaneously, a minor fluctuation may be noticed. Increasing measure will be a highlight among these. The overwhelming majority of people can recollect events from the prior twenty years. It was designed to solve issues, but anything that can be produced should also relapse. Gradient boosting might have been inspired by a method. This aggregates the results for large portions. classifiers that are "weak" to process A capable "committee." a disempowered classifier (e. G. Choice tree) will be the person whose slip rate is significantly better than random assuming. The following figure 4 represents the train and test data of Gradient boosting regression.

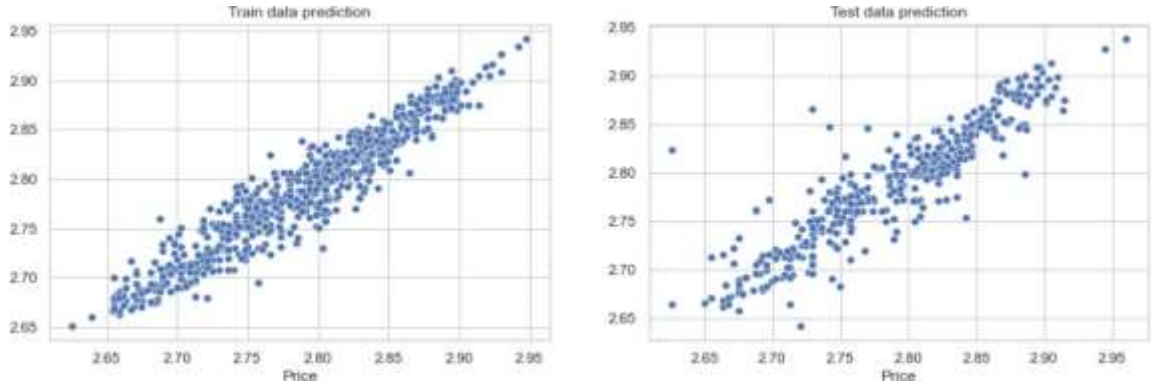

Figure 4: scatter plot of gradient boosting regression

4.3 XG boosting regression

Extreme gradient boosting (XGBoost) is an acronym for the most integrated strategy for either regression or classification tasks. The gradient boosting framework is used in this decision tree method. It includes featuresthat have a significant impact on parameter estimation. This method assists in the creation of a model that is less variable and more stable. Furthermore, the computational efficiency is rapid when compared to other algorithms. The trees are generated in a sequential order. In XGBoost, weights are quite significant. Weights are assigned to all of the independent factors, which are then input into the decision tree, which predicts outcomes. The weightof factors that the tree mistakenly predicted is raised, and these variables are put into the second decision tree. After then, the separate classifiers/predictors are integrated to create a more powerful and precise model. The following figure 5 represents train and test data of XG boosting.

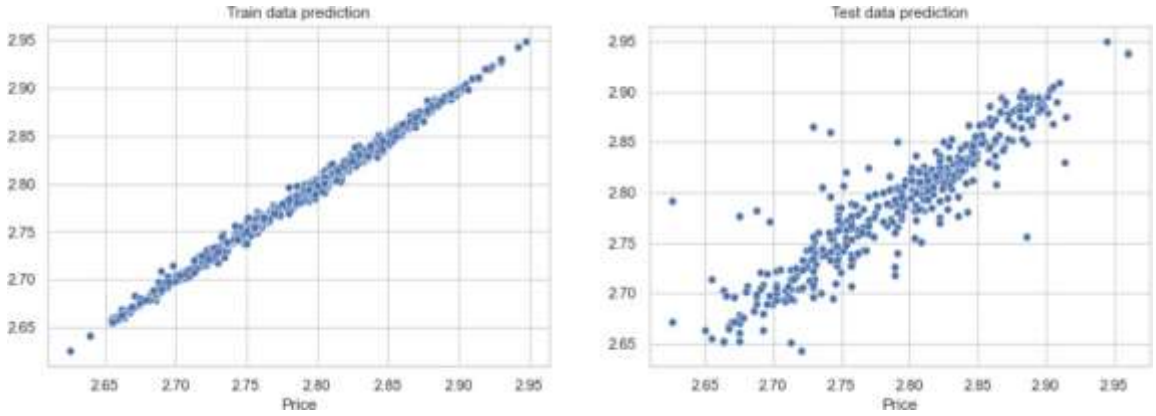

Figure 5 :Sscatter plot of XG boosting regression.

4.4 Random Forest Algorithm

By integrating numerous decision trees and a method known as Bootstrap and Categorization, often known as bagging, a Random Forest can do both regression and classification tasks. Instead than depending on individual decision trees, the primary idea is to combine numerous decision trees to decide the final result. Multiple decision trees are the foundational learning models of Random Forest. To construct sample datasets for each model, we randomly choose rows and attributes from the dataset.

In this study, we employed the Random Forest Classifier class. To specify how many trees to create, we set the Random Forest Classifier's n estimators' parameter to 1100. While raising the quantity of trees in the random forest improves dependability, it also lengthens the period of training for the model. Random forest subsets, on the other hand, will only employ a small number of attributes to add diversity into the trees. When we initialized the Random Forest Classifier, we looped the model numerous times and added a few criteria to enhance the productivity even more. The following figure represents the test and train data of random forest regression

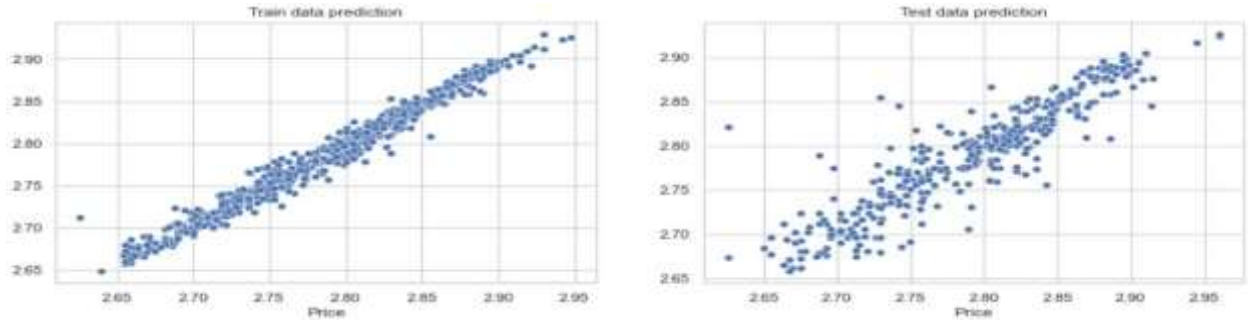

Figure 6 : Scatter plot of random forest regression

4.5. Results

The results of various algorithms are shown in the below table. In this paper various performance metrics such asaccuracy, R-squared value, Root Mean Squared Value (RMSE) were considered. The four models using these parameters of test and train data are compared and tabulated.

Table 1: Accuracy, R-square, RMSE, table

		Test Data		Train Data	
	Accuracy	R-Square	RMSE	R-Square	RMSE
LR	72.76	0.6689	0.0312	0.66	0.34
RIF	83.99	0.9717	0.0099	0.8146	0.026
GBR	83.25	0.8971	0.0187	0.8051	0.027
XBR	83.89	0.99	0.0014	0.823	0.0234

On comparing the various models, we find the random forest, gradient forest, xg boosting works the best with highest accuracy of 83% and linear regression perform least with an accuracy of 72%. These regression models hardly produce any errors in R-square, and RMSE. Thus, we can conclude these 3 regression models over fits our dataset and gives a very high accuracy. Various data mining techniques in Python are used to achieve the results. Various factors that influence house pricing are taken into account and

worked on further. To complete the desired task, machine learning has been considered. The first step is to collect data. The data is then cleaned

in order to remove all errors and make it clean. The data is then pre- processed. Then, using data visualization, different plots are created to resemble the distribution of data in various forms. Finally, the business costs of the houses were ascertained with precision and accuracy. This is possible because a simple stacking algorithm is used to improve the accuracies of the various regression algorithms used on our house pricing dataset, allowing them to produce better results.

5.0 CONCLUSION

Therefore, from this analysis can conclude that linear regression is one of the best ways to predict data by predicting it. You can analyze any number of models, but our main motto is to reveal independent variables or factors. How a variable represents a factor or a variable affected by a factor graphically. The linear regression model has all the features you need to analyze the combined data, making it easier and zero workload. Regression analysis is used in the broadest sense. However, it is primarily based on using the dependent variable data to quantify changes in the dependent variable (regression variable) due to changes in the independent variable. This is because all regression models, linear or non-linear, simple or multiples, associate the dependent variable with the independent variable. In the future, you may get many updates that predict the exact date, even before the original date was generated. It has many features and soon every company will try to introduce them into their system. Linear regression is one of the best ways to predict data or forecast data. Many models are free to analyze, but our main motto is to show in a graphical way how an independent variable or factor affects a dependent variable or factor.

The Linear Regression Model has all the features you need to analyze embedded data, making it easier and more ruined. Regression analysis is used in a broader sense. However, it is primarily based on using the dependent variable data to quantify changes in the dependent variable (regression variable) due to changes in the independent variable. This is because all linear or non-linear, single or multiple regression models associate the dependent variable with the independent variable. In the future, you may receive many updates that predict the exact date, even before the original data was generated. It has many features and soon every company will try to introduce them into their system.

6. REFERENCES

[1] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/(accessed September 1, 2019).
[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018. doi:10.1145/3195106.3195133.
[3] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Confer-ence on Machine Learning and Data Engineering (ICMLDE) 2018. doi:10.1109/icmlde.2018.00017.
[4] Mu J, Wu F, Zhang A. Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis 2014; 2014:1–7.doi:10.1155/2014/648047.
[5] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 2017. doi:10.1109/ieem.2017.8289904.
[6] Ivanov I. vecstack. GitHub 2016. https://github.com/vecxoz/vecstack (accessed June 1, 2019). [Accessed: 01-June- 2019].
[7] Wolpert DH. Stacked generalization. Neural Networks 1992;5:241–59. doi:10.1016/s0893-6080(05)80023-1.
[8] Qiu Q. Housing price in Beijing. Kaggle 2018. https://www.kaggle.com/ruiqurm/lianjia/(accessed June 1, 2019).
[9] H.L. Harter, Method of Least Squares and some alternatives-Part II. International Static Review.1972,43(2)
[10],pp. 125-190..
[11] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. York: John Wiley, 1980.
[12] J. R. Quinlan, "Combining instance-based and model- based learning," Morgan Kaufmann, 1993, pp. 236–243.
[13] S. C. Bourassa, E. Cantoni, and M. E. Hoesli, "Spatial dependence, housing submarkets and house price prediction,"eng, 330; 332/658, 2007, ID: unige:5737.[Online]. Available: http:// archive - ouverte. unige. ch/unige:5737.
[14] Bhalerao V., Panda S.K., Jena A.K. (2021) Optimization of Loss Function on Human Faces Using Generative Adversarial Networks. In: Bandyopadhyay M., Rout M., Chandra Satapathy S. (eds) Machine Learning Approaches for Urban Computing. Studies in Computational Intelligence, vol 968. Springer, Singapore. https://doi.org/10.1007/978-981-16-0935-0_9

