Kingdom of Saudi Arabia
Ministry of Education
College of Computer
Department of Computer Science

المملكة العربية السعودية
وزارة التعليم
كلية الحاسب
قسم تقنية الحاسب

Course: IT 351 "Internet of Things & Data Science"

First Semester 1443 / 1444

Date of Submission: (21-10-2021: 11.00 AM)

Real Estate Price Prediction by Linear Regression

Student Name:	**ID**
Asmaa Moallim	381226390
Yara Al-Tuwaijri	381210460

The Problem and Goal of This Project

Housing is one of the essential human necessities, and property pricing is a crucial agent that encourages or discourages a purchaser's decision of acquiring a home. Real estate prices are inextricably tied to the status of our economy, residential area, facilities offered, and other multiple variables. Therefore, it is almost impossible to anticipate real estate worth with an idealistic degree of accuracy. Nonetheless, great efforts are established into developing viable real estate trends predicting ML models, so it is the goal of this project. More precisely, Riga city housing prices are anticipated by linear regression approach, seeded with roughly a big dataset of 13 features, 8846 entries, consisting of two merged data-sets, one consists of 4689 and the other of 4157; both of the same city, but had been uploaded to Kaggle after a one-year gap. More details of our trials are discussed later on.

Dataset description

Two data sets have been combined and used in our experiments, also the one made up of 4689 samples has been tested separately as well. Having said that, motivations will be stated later on in the experiment setup section of this report. Both of these data sets are titled Riga real estate data sets and consist of 13 features as presented in the forthcoming table. A total of 8846 samples make up this dataset that is collected of, as the name implies, Riga city, which is the capital city of Latvia and it's near the Baltic Sea.

Note: in house_seria column: (602, 119, 103, 467, and 104) are merely odd names for construction projects in Riga, not calculable values as it might be wrongly assumed.

Column Title	Description
op_type	offer type ('For rent', 'For sale', 'Buying', 'Renting', 'Change', 'Other').
district	The district where real estate objects are located .
street	address of real estate object.
rooms	number of rooms.
area	living area of real estate objects.
floor	floor of real estate object.
total_floors	total number of floors in building.
house_seria	house design ('LT proj.', '602.', 'P. kara', 'Jaun.', 'Specpr.', 'Hrušč.', '119.', 'M. ģim.', 'Renov.', '103.', nan, 'Priv. m.', '467.', 'Staļina', '104.', 'Čehu pr.').
house_type	type of building ('Brick–Panel', 'Panel', 'Wood', 'Masonry', 'Brick', 'Panel–Brick').
condition	stuffing premises ('All amenities', 'Partial amenities', 'Without amenities').
price	price in EUR.
lat	latitude of real estate objects.
lon	longitude of real estate objects.

Machine learning Model

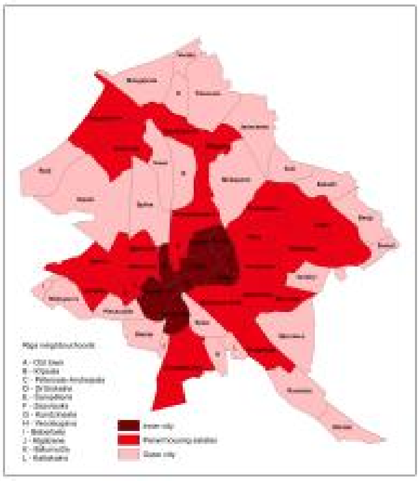
The ML model utilized in this project is built on linear regression, which is a method for predicting or visualizing the connection or the cause and effect relationships between two characteristics or features. There are two types of variables evaluated in linear regression tasks: the independent and the dependent variable. The former is which exists independently of the other factors, whereas the latter is the investigated variable predicted by the model, which varies according to the independents. This method goes by plotting a line that best represents the variations of data points based on the training data samples. The linear equation presented below is used to describe the regression line : y = a_0 + a_1 * x.

Experimental Setup & Results

The experiment started with selecting a project that already has been built by another data scientist, as instructed by the requirements of this project. Then we have studied the method used which is linear regression and figured out another prediction problem, stated in the problem section of this report. Thus, the data set used in the linear regression method is different from the original project. However, during our work, we faced some issues with the size of the data, which made us find another data set to the same studied city, Riga City, and merged it with the first one. Although, this step was one of the things which had improved the results achieved from our experiment, analyzing results had got more interesting once we discovered duplication in the data used, as well as understanding more statistical properties.

As our code notebook is structured, after importing necessary libraries as well as reading the dataset, we have explored the data in various steps, both mathematically, and visually, noting down any obtained observations. the coordinates of some records turned out to be out of the studded region of Riga city, thus excluded; some operation types of (opt_type) column, as other and change values had no relevance to the objectives of our prediction task, so excluded as well along with two more values, remaining with only two types: for rent and for sale. Then we have managed to fix all missing values found in our data sample. In terms of districts, Google Maps has been utilized to fill out missing cells. Whilst, in regards to handling missing, or invalid room count, a certain statistical estimation method has been used. However, the missing values in the two columns (Lon and Lat), which determine the coordinates of the real states, are ignored, as they have not been useful along with the development of our analysis and eventually are not used further ahead.

At this stage, missing data has been resolved, so it is time to move to feature engineering and data preprocessing, where we have converted categorical variables, at least those intended to be used, to numerical values. Initially, we have worked on categorizing the district values, as in the following figure, into three classes: inner_city, panel_housing_eastate, and outer_city. Next, dummy encoding has been used on the District column, along with house_type, and house_seria columns. Despite this conversion of the house_type, and house_seria independents has been beneficial to the regression model, it did not perform well with the district column. Therefore, we reconstructed the experiment with the ordinal encoding of the original district values, without using any classes. Lastly, we have also executed ordinal conversion on the condition columns as well. Besides getting rid of all non-numerical values in 4 columns by now, we have additionally dropped the rest of the useless categorical columns which are the following: op_type, street, lat, lon, district, and total_floors.

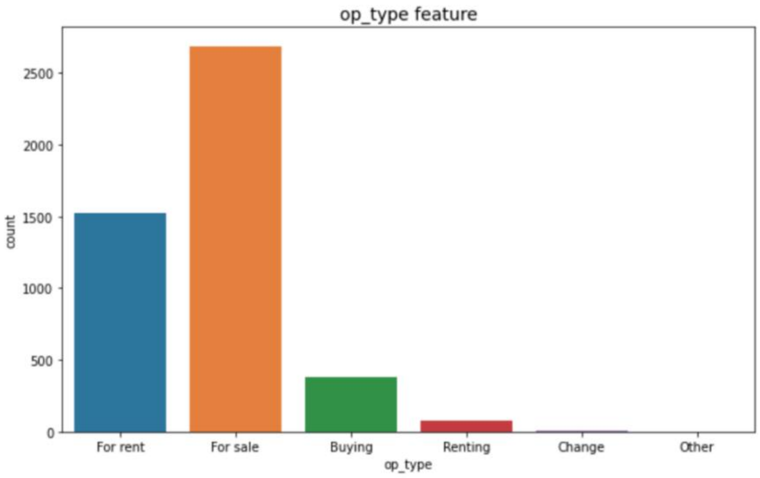


The following table presents the feature engineering and data preprocessing steps performed.

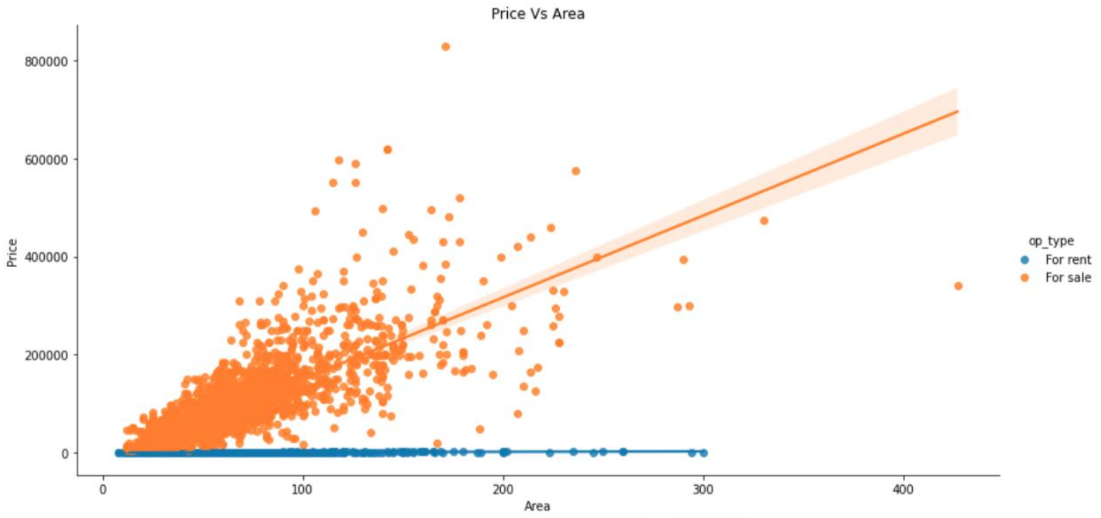
Step	Affected Columns	Resolving Procedure
Handling missing values	205 missing values of (lon and lat) columns	Ignored as ultimately not used
	2 missing values of (district) column	handled by looking for entries that have the same street address as befitting from Google Maps
	1 and eventually 15 missing values of (room) column	Statistical estimation according to the most frequent number in a set constructed based on the variation of room count found in the same area.
Feature engineering & handling categorical data	(District) column	Dropped after creating a new column named (district_code), which is the ordinal encoding of the original (district) column.
	(condition) column	Each type of the three condition values is replaced by an ordinal range (1,2,3) and 0 if no value is available.
	(house_type, house_seria) columns	Both handled by dummy encoding
Dropping unwanted features	(lon , lat, total floors, district, street, op_type) columns	all dropped since they had no impact on our model

Furthermore, the data visualization section lies between the two feature engineering and data preprocessing part, as some was also utilized at the data exploration stage. Here are some of the plotted figures, which expose so much information, mainly the fact that data at hand are not normally distributed, which had a crucial impact on the model outcome.

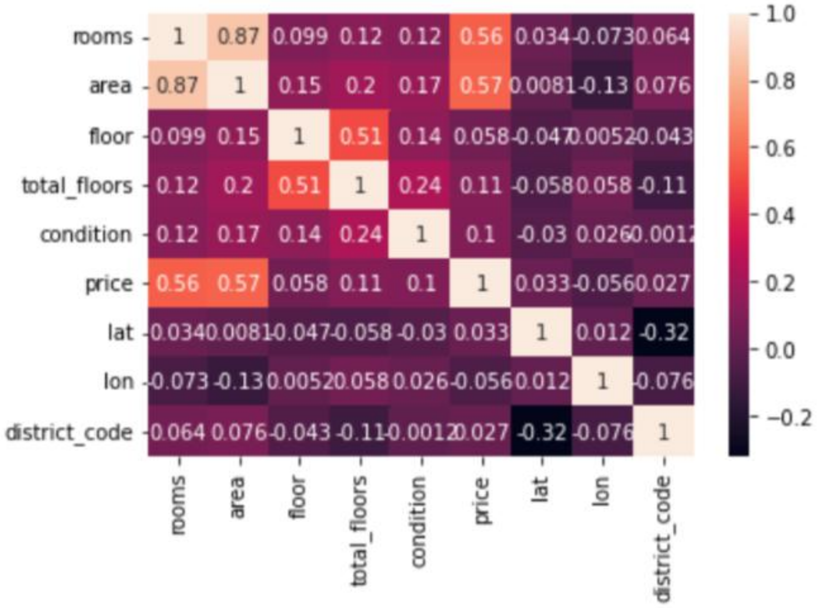
In the following chart, it can be seen that the cont values of columns: "Buying", "Renting", "Change" and "Other" are drastically smaller than the data available for for-rent and for-sale data



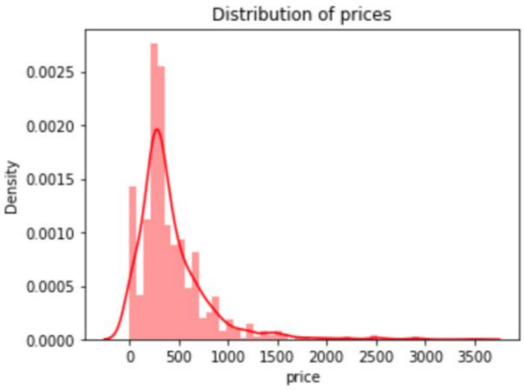
In the following regression graph, it can be seen that there is a clear regression line which is a good thing, yet suggests spitting the data into a for-rent dataset and for-sale dataset to be able to see the line clearly.

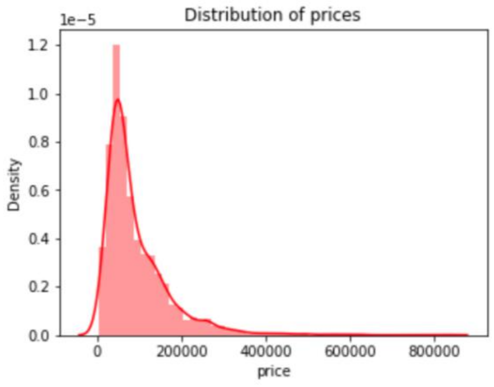


In the correlation chart, certain features are of more correlation indicates a higher impact on the price protection, which are the area and rooms independents.



These graphs display the distribution of price, one for sale and the other for rent data. Both data sets are positively skewed, which infers the necessity for the normalization of the data.





According to our observations and through a lot of experiments, correlation analysis, training, and testing the model. We have noticed that the data sets for the two types: for sale, and for rent, are not equally or similarly variant. This caused a lot of issues, despite conducting many normalizing methods such as excluding the outliers for each group, still, data samples of (for rent) types were extremely insufficient to predict the prices for rent, in all of our trials.

In contrast, the result of prediction for the for sale group of data, when the merged data was used, was 79.67% in the r-square score, and 80.55% when the smaller data, of 4689 entries, was used. This slight drop anyhow, of 0.88%, matches the rise of data items, which does not underrate the quality of the model. Nonetheless, the rent dataset was 1507, which is barely half of the sale data, 2651; also, rent data points were notably variant, thus suggesting higher error values when data increases. However, when data count doubles, variance decreases, equating to a raise from 26.60% to 33.63% in R-square score, when each dataset of data-sets 4689, 8846 was used, respectfully. These results show the good performance of the model, especially, when the data at hand are non-variant or properly normalized.

Lastly, model evaluation and validation follow the building and application of the model. Regardless, the Scikit Learn module has been hired as a comparison tool for verification purposes of the results achieved by the linear regression model built in this project. More of the statistical observation and clarification of the result trends are detailed in the notebook accompanied with this report.

Literature Review

The following tables illustrate the performance of five prices prediction studies achieved by different methods based on distinct data samples.

Study Title	Real Estate Investment Advising Using Machine Learning
Dataset	The data was collected by authors using Web Scraping from websites like 99acres.com, Magicbricks.com, Google.com.
ML method	Random Forest
Performance measures	Root Mean Squared Error (RMSE) = 0.007 Mean Absolute Percentage Error (MAPE)= 6.328 Mean Absolute Error (MAE)=0.062

Study Title	A Hybrid Regression Technique for House Prices Prediction
Dataset	Kaggle House Price Advanced Regression Techniques Competition Dataset
ML method	Hybrid regression with 65% Lasso and 35% Gradient boosting
Performance measures	Evaluation standard used is Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price = 0.11260

Study Title	Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring
Dataset	Dataset is collected by the authors from bProperty.com which include 3505 house entries and 19 features.
ML method	Ensemble Gradient Boosting
Performance measures	Accuracy = 88.75%

Study Title	Artificial neural networks for predicting real estate prices
Dataset	National Statistics Institute (INE) housing dataset
ML method	ANN
Performance measures	R-squared = 77.38%

Study Title	Forecasting spatial dynamics of the housing market using Support Vector Machine
Dataset	Gigahouse Taiwan's Real Estate Portal database (from 2007 to 2010)
ML method	SVM
Performance measures	Hit rate = 81.8%

Comparison with previous studies

Although our results are not, in total, as efficient as those found by the original project source, they are still considerable, specifically in terms of anticipating real state prices for the sale group of our data set, which is very different from the samples used in the original project. These outcomes, both in our experiment and the other mimicked project, support the efficiency of using linear regression in such prediction problems.

Despite everything, many other ML models are used in the quest of real estate prediction, at several locations and distinct dataset samples, as listed in the previous section titled literature review. It is hard to compare performance as different ML methods are used as well as the score scaling variation of each model. However, our model used the R-square measure similar to the fourth study above, which used the ANN-based model, and achieved 77.38%. Whilst, the linear regression model in this project had the higher result of R-square, that is 80.55%, which raises the positive potentiality found in the Linear regression models.

References

Forecasting spatial dynamics of the housing market using Support Vector Machine

- https://www.tandfonline.com/doi/abs/10.3846/1648715x.2016.1259190

A Hybrid Regression Technique for House Prices Prediction

- https://ieeexplore.ieee.org/abstract/document/8289904

Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring

- https://dl.acm.org/doi/abs/10.1145/3318299.3318377

Artificial neural networks for predicting real estate prices

- https://www.redalyc.org/pdf/2331/233127547002.pdf

Link to Dataset used (1)

- https://www.kaggle.com/trolukovich/riga-real-estate-dataset

Link to Dataset used (2)

- https://www.kaggle.com/dmitryyemelyanov/riga-real-estate-dataset-cleaned

The base code used to create this project

- https://www.kaggle.com/sudhirnl7/linear-regression-tutorial

Residential satisfaction and mobility behavior among the young: insights from the post-Soviet city of Riga

- https://journals.openedition.org/belgeo/28347?lang=nl

The Game of Increasing R-squared in a Regression Model

- https://www.analyticsvidhya.com/blog/2021/05/the-game-of-increasing-r-squared-in-a-regression-model/

Real Estate Investment Advising Using Machine Learning

- https://d1wqtxts1xzle7.cloudfront.net/53503829/IRJET-V4I3499-with-cover-page-v2.pdf?Expires=1632551437&Signature=Rq8paII-pKo1Vt-Q-zeJzgV~IWnjvZ5z2Dq2ajZ5sOzCfio5s-yEqtbQ0neFze91Agg7yAXSouNmkm3BINlehKdVmmsr~MQDvAjCqOyIG3ped7tewCgxA5dfAl6HV-hWRK6ny2vlpiyq0VY6MckqjoUklipWGa9YW~JtGms4IEZ-ri88qloTq1XE5a1pJrjyu2hhhrheT24iVEnt4d9uVgmFi6eBEPVsTv0h88Kdct-uhge3xkXDPXJzxgWNHPWdkm7nEBf6Dwd~muiOmoFVuhqpKugVhFuKxT0vmkP~F0Ga-QCF4fLJdANKuRqqirMBGJMQ7xAhAt244R0pQTytbw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA