MODELING THE FAIR VALUE OF BANKING COLLATERAL USING REAL ESTATE WEBSITE DATA AND MACHINE LEARNING TECHNIQUES - THE CASE OF CASABLANCA CITY, MOROCCO

[1]YOUSSEF TOUNSI, [2]HOUDA ANOUN, [3]LARBI HASSOUNI

RITM Laboratory, CED Engineering Sciences, Ecole Superieure de Technologie

Hassan II University of Casablanca, Morocco

E-mail : [1]tounsi@gmail.com, [2]houda.anoun@gmail.com, [3]lhassouni@hotmail.com

ABSTRACT

In recent years, big data has been widely used to understand emerging trends in several markets. With this paper, we consider the use of real estate websites data and machine learning in the problem of knowing the current fair value of banking collateral in Morocco. In our study, we are going to utilize the online data, containing housing sales advertisements, extracted from three websites: (Avito.ma, Mubawab.ma and Sarouty.ma), popular online portals for real estate services in Morocco, in order to develop forecasting models of apartment prices using artificial intelligence for the city of Casablanca. Each record of the database contains information on the listed housing unit (asking price, district, floor, area, number of rooms, balcony, terrace…), on the building (elevator, garage, garden, etc.), on the ad (publication date, apartment sector) and a short description. After collecting the database and the deduplication process, we make additional controls on the dataset to address potential incoherence errors in the data. After validating the clean dataset against official statistical sources, we construct forecasting models for real estate sale prices, based on artificial intelligence and statistical methodologies.

Keywords: *Housing Price Prediction, Mortgage lenders, Machine Learning Algorithms, Big Data.*

1. INTRODUCTION

Housing is particularly important for financial stability due to the banking sector's huge exposure to real estate. Indeed, mortgage underwriting at many financial institutions is an important part of general risk management. Therefore, most lenders would have to estimate the actual value of a real estate when deciding whether or not to extend credit, as well as at any point during the loan's repayment period, since that constitute the mortgage collateral.

In the current context of the Covid-19 epidemic and like many other economic sectors, the real estate sector has suffered greatly from this crisis even if this sector has benefited from certain fiscal measures in Morocco to revitalize the market during 2020 and 2021, knowing that it was weakened long before. This stresses the need that bankers should practice cautious risk management by obtaining a thorough understanding of all sources of repayment. In the event that the borrower doesn't repay the credit in accordance with the loan agreement, cash from the sale of collateral is a secondary source of repayment.

Furthermore, the real estate sector became a center of attention over the last few years in Morocco, given the extent of its effects on financial spheres and its implications for monetary policy decisions and financial stability. In fact, the sector's share in Morocco's GDP exceeds 6% and employs over a million people, according to the last report of the Ministry of National Territory Planning, Land Planning, Housing, and City Policy. In the absence of reliable indicators for the Moroccan properties prices, The Central Bank of Morocco has constructed in 2010 a quarterly real estate price index (REPI) based on the office's Databases which contain detailed information on all property transactions registered at the national level. However, in the presence of multiple obstacles, such as the absence of detailed physical characteristics for houses sold and the hopelessness of any possible timely analyses at a detailed geographical level, the making and the development of such indicators remains difficult. Hence, the use of a new type of data analysis, big data; these large amounts of data open new opportunities to improve processes and make more informed decisions.

In this perspective, this paper attempts to build models for real estate evaluation using several machine learning techniques such as linear regression family (Ordinary Least Squares, Ridge, Lasso, Elastic Net, Huber, and RANSAC), RandomForest (RF), and the new generation of gradient boosting algorithms (Xgboost, Lightboost and Catboost).

This study focuses on the housing market in Casablanca. We collected property price information from the most used websites (Avito.ma, Mubawab.ma, and Sarouty.ma) in Morocco and we made the following contributions:

- We propose an intelligent automatic prediction system for real estate by adopting different machine learning algorithms. Our system involves several Python modules, which allow the implementation of an effective method of recognition of repeated housing data source, detection of outliers and reliability of information for improving data quality, the feature engineering and improving the robustness and accuracy of the proposed model.
- Various regression techniques are proposed for apartment price prediction using the data of poplars from online real property portals in Morocco (Avito.ma, Mubawab.ma, and Sarouty.ma) for Casablanca city.
- We show that the information from the “Apartments for Sale in Casablanca” dataset, is consistent with existing official statistical sources such as the National Agency of Land Registry, Cadaster and Cartography (ANCFCC), and The Central Bank of Morocco.
- The granularity of online data also makes possible timely analyses of the trends of real estate market at a very detailed geographical level.
- Results are validated on real life data.

However, the studies that were conducted in the past, only look at a limited number of models. In fact, different regression techniques, including the new generation of gradient boosting algorithms, have been experimented with using real data from a North African country for the first time. The experimental results show that the performance of the RF model exceeds the other methods used.

This paper is organized as follows: The first section gives an overview of related work; The second section illustrates the details of the methodology. In the third section a case study is presented and analyzed. Section 4 discusses the results, draws conclusions, as well as proposes further potential directions to study the problem.

2. RELATED WORKS

Over the past years, a number of papers have focused on understanding emerging trends in the real estate market and predicting housing prices using machine learning algorithms. Eduard Hromada carried out a statistics comparison of the market prices of real estate in the Czech Republic using an innovative proposed software for professionals and researchers. the data processed in this study is retrieved from the real estate advertisements published on the internet: over 650,000 price quotations concerning the sale or rental of apartments, houses, business properties, and building lots. This study revealed that prices have fallen since 2008 and that this negative trend does not seem to be significantly changing [1]. Yeonghwa Park and Jae Kwon Bae showed that the performance of RIPPER is superior to that of the C4.5, Naïve Bayesian, and AdaBoost models for housing price prediction problems by analyzing the housing data of 5359 townhouses in Fairfax County, Virginia [2]. Michele Loberto, Andrea Luciani, and Marco Pangallo analyzed a dataset consisting of more than one million online sales advertisements for residential units posted on the website Immobiliare.it between the beginning of 2015 up to June 2017 in all Italian provincial capitals [3].

Moreover, the research presented in reference [4] is an exploratory attempt to estimate house prices using three machine learning algorithms (SVM, Random Forest (RF), and Gradient Boosting Machine (GBM)), and then compare their results using 18 years of housing property data. The authors showed that RF and GBM can generate comparably accurate price estimations with lower prediction errors. To predict real estate prices using actual transaction data, the study in reference [5] utilized four machine learning models, namely, least squares support vector regression (LSSVR), classification and regression tree (CART), general regression neural networks (GRNN), and backpropagation neural networks (BPNN). Numerical results indicated that the least squares support vector regression outperforms the other three machine learning models in terms of forecasting. The study of Quang Truong et al. [6], investigates different models for housing price prediction of the “Housing Price in Beijing” dataset fetched from the Kaggle platform, which contains more than 300, 000 data with 26 variables representing housing prices traded between 2009 and 2018. Several different types of Machine Learning methods including Random Forest, XGBoost, and LightGBM, and two techniques in machine learning including Hybrid Regression and Stacked Generalization Regression are compared and analyzed for optimal solutions. One of the main findings is that the Stacked Generalization Regression method has a complicated architecture, but it is the best choice when comparing accuracy. Since it influences the loan and insurance markets, the tax system, accounting, and infrastructure construction programs, the real estate valuation process is critical for the economy [7].

3. METHODS

Regression is the process of learning relationships between inputs and continuous outputs from example data, which enables predictions for novel inputs. The objective is to build a model based on training data, which will allow us to predict the Y output associated with a new input X. The Y output in our case is quantitative (price per square meter of the apartment). Some machine learning techniques adapt to all types of explanatory variables while others are specialized. For this regression process, there are many techniques:

3.1 Linear Regression

A linear regression model is a model that seeks to establish a linear relationship between a variable X and a variable Y. The idea is to then be able to make predictions on Y when X is measured.

$$Y = f(X) = \beta_0 + \beta_1 x_1 + \cdots .. + \beta_n x_n + \varepsilon \quad (1)$$

$$Y = f(X) = X\beta + \varepsilon \quad (2)$$

Where $X = \{x_1, x_2, ... x_n\}$ is known as input space, $Y = \{y_1, y_2, ... y_n\}$ as output space and β(β_0, β_1, …, βn) are the regression coefficients, and ε is the random error.

x_1 is called a sample, and x_{ij} means the j-th feature of the i-th sample. For each observation i = 1,..,n the estimated $f(x_i)$, should be as near to the actual response y_i as possible. The residuals are the differences y_i - $f(x_i)$ for all observations i = 1,.., n.

The Method of Ordinary Least Squares (OLS) aims to find the best predicted weights that correspond to the smallest residuals. The sum of squared residuals (SSR) for all observations i = 1,.., n is commonly minimized to obtain the optimal weights:

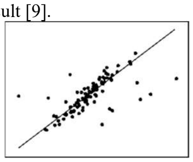
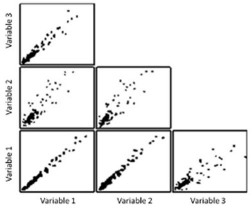
$$SSR = \sum_{1}^{n} (y_i - f(x_i))^2 \quad (3)$$

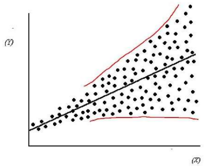
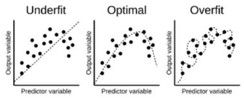
OLS Problem comes down to:

$$minimize\ (\beta)\ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \quad (4)$$

The ordinary least squares estimator may face certain challenges, such as the existence of [8]:

Table. 1. OLS challenges

Problems With OLS	Description	How to deal with?
Outliers	An outlier can be described as an observation that deviates so much from the rest of the observations that it can be suspected that it was produced by a different mechanism. Training techniques include sequential presentation of the initial set of data to the recognition system and model adapting to these data. That is why the presence of distorted objects can affect the final result [9].  Fig. 1. Outliers	Use of the methods: Huber, RANSAC, Theil-Sen
Multicollinearity	When two or more predictors are associated, a phenomenon known as multicollinearity occurs, and the standard error of the coefficients rises [10]. Because of multicollinearity, certain variables are statistically insignificant when they should be relevant. Multicollinearity should not be present in the dataset because it causes issues when ordering variables based on their significance for several types of regression approaches.  Fig. 2. Multicollinearity	Dropping variables, combining variables, use of Ridge Regression technique, application of PLS, PCA and Regularization methods

Heteroscedasticity	When the standard deviations of a predicted variable are non-constant over different values of an independent variable or as compared to previous time periods, this is known as heteroscedasticity [11]. The occurrence of residual heteroscedasticity causes the following issues: • The standard errors computed for the regression coefficients are overstated or understated. • The model's tendency for producing fitted values and forecasts of the dependent variable with varying degrees of average error at different points within the sample. • The forecasted regression has an incorrect functional form. • Missing explanatory variables that are significant. • Within the estimated relationships, structural changes occur. • The nature of the variables in the model is constantly changing.  Fig. 3. Heteroscedasticity	Log-transformation
Underfitting and Overfitting	Overfitting occurs when the algorithm over-learns, in other words, when it learns from data but also from patterns that are not related to the problem, such as noise and Underfitting is defined as a model that cannot both model and generalize to new data [12].  Fig. 4. Underfitting and Overfitting	**Underfitting**: Get more training data, Optimize Model Parameters, etc. **Overfitting**: Cross-validation, Regularization, Pruning (decision trees), etc.

3.2 Ridge Regression

Hoerl and Kennard (1970) presented the ridge regression estimator as an alternative to the ordinary least squares (OLS) estimate in the presence of multicollinearity [13]. This approach aims to optimize the following problem:

$$E(\beta) = minimize_{\beta}\,(\alpha,\beta)\ \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \alpha\,\|\beta\|^2 \quad (5)$$

It is a linear regression with a quadratic constraint on the coefficients. This is useful when the variables are highly correlated, which often skews the numerical resolution. The solution can be expressed exactly:

$$\beta^* = (XX' + \alpha I)^{-1}X'Y \quad (6)$$

where X' is the transpose of X and I identity matrix, we see that it is possible to choose an α for which the matrix $XX' + \alpha I$ is invertible, it is also useful when there are a lot of variables because the probability of having correlated variables is high. It is possible to choose α large enough so that the matrix $XX' + \alpha I$, is invertible and the solution is unique.

3.3 LASSO Regression

Tibshirani (1996) introduced LASSO (Least Absolute Shrinkage and Selection Operator), an innovative variable selection method for regression that minimizes the residual sum of squares when the total absolute value of the coefficients is less than a constant and it's a well-known sparse regression approach for regularizing the parameter under the sparse assumption [14]. This regression method performs Both regularization and variable selection.

Lasso regression technique aims to optimize the following problem:

$$E(\beta) = minimize_{\beta}\,(\alpha,\beta)\ \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \alpha\,\|\beta\| \ (7)$$

It is also a linear regression with a linear constraint on the coefficients. This is useful when the variables are highly correlated, which often skews the numerical resolution. The solution is not expressed exactly and the resolution uses a gradient-based method.

Finding β^* which minimizes the expression requires finding the parameter β such that $E'(\beta) = 0$. There is no explicit matrix formula of β^* for LASSO method, unlike ridge regression. Although the quadratic programming techniques from convex optimization (efficient approximation algorithms) are used to find the solution.

3.4 Elastic Net Regression

In 2005, Zou and Hastie suggested an Elastic Net Regression compromise between LASSO and Ridge regression, in order to overcome some drawbacks of the LASSO regression technique [15], in which the penalization takes the form of:

$$\alpha\|\beta\|^2 + \alpha\,\|\beta\| \quad (8)$$

The elastic net method's estimates are defined by:

$$minimize_{\beta}\,(\alpha,\beta)\ \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \alpha\,\|\beta\|^2 + \alpha\,\|\beta\|\ (9)$$

The prediction can be formulated as an elastic net regression issue, which has two advantages. On the one hand, the Elastic Net retains the regression model's advantage of being able to employ more background information. The Elastic Net regression, on the other hand, can adaptively pick stable features for training, resulting in a more accurate appearance model [16].

3.5 Huber Regression

Peter Jost Huber (1964) introduced Huber Regression; it is a supervised learning approach that is robust when it comes to outliers [17]. The concept is to apply a special loss feature instead of the conventional Least Squares:

$$minimize\ (\beta)\ \sum_{i=1}^{n}\emptyset(y_i - x_i^T\beta) \quad (10)$$

For the variable $\beta \in R^n$, where the loss $\emptyset$ is Huber function with threshold M > 0,

$$\emptyset(\mu) = \begin{cases} \mu^2 & if\ |\mu| \le M \\ 2\,M - M^2 & if\ |\mu| > M. \end{cases} \quad (11)$$

This ensures that outliers have a little impact on the loss function while not fully ignoring their impact.

3.6 RANSAC Regression

This algorithm was first published by Fischler and Bolles in 1981 [18]. It is an iterative method for estimating mathematical model parameters from a set of observation data. RANSAC's basic idea is to eliminate outliers from a batch of data before fitting the model for the remaining data. The RANSAC steps are as follows:

- Select the smallest number of points needed to calculate the model parameters at random
- Compute the model's parameters
- Estimate how many points in a set of all points meet a particular tolerance
- Re-estimate the model parameters using all the identified inliers if the fraction of inliers over the total number of points in the set exceeds a predetermined threshold
- If not, repeat steps 1 through 4 again (maximum of N times).

The number of repetitions, N, is chosen to ensure that the probability p (typically set to 0.99) that at least one set of random samples does not contain an outlier, is high enough. Let μ denote the probability that every given data point is an inlier. The probability of seeing an outlier is given by $\vartheta = 1 - \mu$. There must be N iterations of the minimal number of points designated m, where:

$$1 - p = (1 - \mu^m)^N \quad (12)$$

As a result:

$$N = \frac{\log(1-p)}{\log(1-(1-\vartheta)^m)} \quad (13).$$

3.7 RandomForest Regression

The random forest technique was first proposed by Ho in 1952 and was formally proposed in 2001 by Leo Breiman [19]. This model is founded on a decision tree algorithm. The difference between the decision tree and the random forest algorithm is that the decision tree method uses only one tree to make decisions, but the random forest algorithm uses numerous trees to create a forest.

Indeed, many decision trees make up a random forest algorithm. The random forest algorithm's generated 'forest' is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them together. In fact, the random forest algorithm determines the outcome based on decision tree predictions. It forecasts by averaging the output of various trees. The precision of the result improves as the number of trees grows. The following stages will help us understand how the Random Forest algorithm works:

- Stage 1: Begin by selecting random samples from a dataset.
- Stage 2: For each sample, this algorithm will create a decision tree. The forecast result from each decision tree will then be obtained.
- Stage 3: Voting will be done for each expected outcome in this step.

- Stage 4: Finally, choose the prediction result with the most votes as the final forecast result.

3.8 XGBoost

XGboost is an advanced implementation of Gradient Boosting Algorithms (GBM). They differ on numerous points:

- The GBM do not have regularization which avoids the over-learning while XGboost has an automatically integrated into its procedure.
- GBMs stop separating nodes when they find a negative loss in the limb. XGBoost, as for this approach, separates up to a certain depth (given in parameter) and begins to carve in the opposite way and removes the branches which do not have positive gain.

3.9 LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.

3.10 CatBoost

The CatBoost algorithm, recently launched by the company Yandex, is an implementation of gradient boosting that handles categorical data. As the ensemble of trees can generally only handle numeric features, converting categorical features to numbers requires major preprocessing efforts such as the one-hot encoding technique that transforms each category into binary variables. Instead of these time-consuming preprocessing steps, CatBoost handles categorical data efficiently as after performing randomly permutation, an average label value is computed for each example when the same value was set before the permutation. In addition, overfitting is prevented by using multiple permutations for training different models.

4. EXPERIMENTAL DESIGN

4.1 Analysis Procedure

The machine learning procedures we used on our dataset are described in this section. The primary goal of this research is to forecast the fair price of banking collateral as house value based on alternative data (online Real estate data) for the given features in order to maximize prediction accuracy by using the proposed approach Fig.5. This housing issue can be classified as either a regression issue. This study covers multiple regression models. Data collection, data preparation, data preprocessing, feature engineering, model training, and model evaluation are essential stages for data mining to build housing price prediction model.

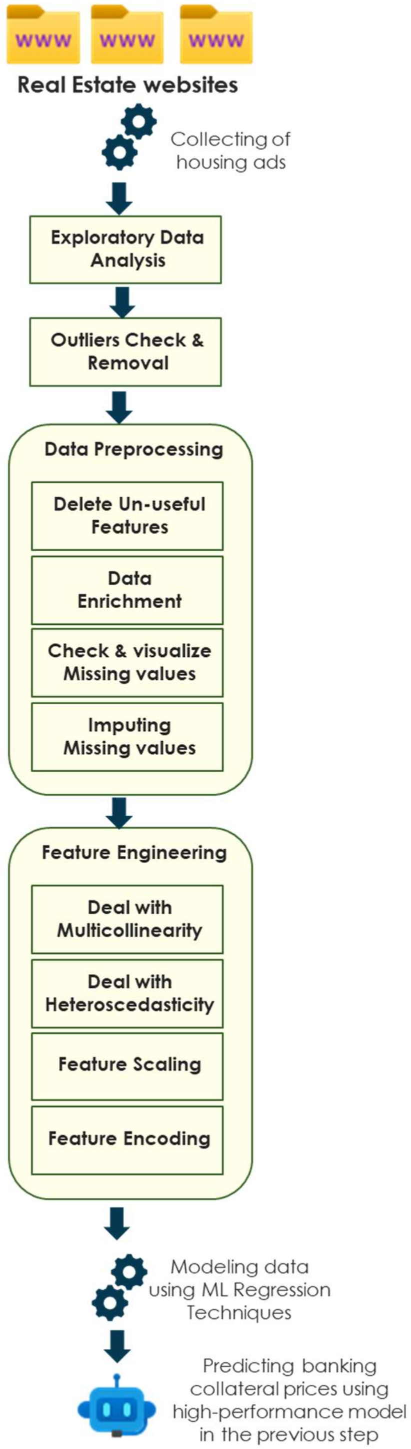

www
www
www
Real Estate websites
Collecting of housing ads
Exploratory Data Analysis
Outliers Check & Removal
Data Preprocessing
Delete Un-useful Features
Data Enrichment
Check & visualize Missing values
Imputing Missing values
Feature Engineering
Deal with Multicollinearity
Deal with Heteroscedasticity
Feature Scaling
Feature Encoding
Modeling data using ML Regression Techniques
Predicting banking collateral prices using high-performance model in the previous step

Fig. 5. Research Methodology for the housing price problem

First, our dataset is built from public data from popular online portals for real estate services in Morocco (Avito.ma, Mubawab.ma and Sarouty.ma) and we encountered several challenges. Indeed, private individuals or real estate agencies are the sellers. They post an ad for the property they're selling, and if the ad is set to visible, anyone can see it without having to register with the website. Sellers, on the other hand, must first create an account. Potential buyers can look for houses based on their location as well as their physical characteristics. They can also look at the photographs or read the textual description of the ad and specify a price range. Potential buyers can contact the seller who posted the ad once they have identified properties that they are interested in. The three Real Estate Websites provide contact information, including phone numbers and the possibility of online conversation.

For the reason that several sellers publish their ad more than once, so the regression algorithm considers them to be more significant, we constructed the main dataset by keeping unique ads, as selected by using python programs to detect duplicate ads. The data consists of all sales of "apartment" type real estate in the city of Casablanca published on popular online portals for real estate services in Morocco (Avito.ma, Mubawab.ma and Sarouty.ma) between April 2019 and March 2021. This dataset “Apartments for Sale in Casablanca” contains more than 18, 000 data observations with 20 variables (Before data processing).

The variables of our dataset listed in the table below, will be used to predict the average price per square meter of each apartment.

Table. 2. List of the variable names and definitions

Attribute	**Description**
Area	Apartment size in square meters
Ad title	Ad title
Neighbourhood	Physical locations within the prefecture of Casablanca limits
Zone	Code Zone according to the repository of National Agency of Land Registry, Cadastre and Cartography (ANCFCC)
living rooms	Number of living rooms
Bedrooms	Number of bedrooms
Bathroom	Number of bathrooms
Floor	Which floor?
Parking	Whether the apartment has any parking
Elevator	Whether the apartment has any elevator
Air-conditioning	Whether the apartment has any Air-conditioning
Balcony	Whether the apartment has any balcony
Terrace	Whether the apartment has any

Equipped kitchen	Whether the kitchen is equipped or not
Security services	Whether the Apartment has any security services
Age	Age of the apartment (range)
Description	Description (this data was used to make some other reliable data, but it was not used to build the model)
Phone	Phone number of the person/ real estate agency who published the ad (this data was not used to build the model)
Phone_rep	Number of repetitions of the telephone number on the total list to find out if it is published by individual or an intermediary

Having a large amount of data does not inherently imply that it is of high quality. In order to obtain some relevant insights from the existing data, it becomes more vital to focus on quality as the amount grows. It is in this sense that we performed data cleansing based on the concepts of commonly used dimensions to measure data quality [20]. The International Data Management Association (DAMA) provides a comprehensive list of the data quality dimensions as represented in Figure 6. DAMA (2013) defines the six data quality dimensions in the following way [21]:

Table. 3. Data Quality dimensions & Data cleansing activities

Dimension	**Definition DAMA (2013)**	**Data cleansing activities**
Completeness	The proportion of stored data against the potential of "100% complete";	Investigate missing data and removing variables with missing values more than 50%
Uniqueness	Nothing will be recorded more than once based upon how that thing is identified	Remove duplicate elements: The major problem with this type of dataset is that it contains a significant number of duplicates, i.e., more than one advertisement for the same appartement. We correct this distortion using some rules based in the criteria that identify the duplicates.
Timeliness	The degree to which data represent reality from the required point in time	This criterion is respected since it is new data retrieved
Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition	Ensure that the data is correct according to the checking of data type, format, code and Range.
Accuracy	The degree to which data correctly describes the "real world" object or event being described	Check the reliability of variables (District, floor, area, number of rooms, balcony, terrace, elevator, garage, garden) based on the description of the ad.
Consistency	The absence of differences, when comparing two or more representations of a thing against a definition	Check the consistency of our dataset based on statistical analysis of outlier detection and verifying of forgotten zeros or extra zeros at the asked price information level.

COMPLETENESS
CONSISTENCY
UNIQUENESS
Data Quality Dimensions
ACCURACY
TIMELINESS
VALIDITY

Fig. 6. Data Quality Dimensions adapted from (DAMA, 2013)

4.2 Data Analysis

After we applied a number of cleaning techniques, 15,000 data observations with 18 variables were kept and then used for the rest of this work. Before creating a regression model, an exploratory data analysis is required. Researchers can find the data's implicit patterns in this way, which aids in the selection of relevant regression techniques.

Figure 7 shows a representation of the number of ads by neighborhood in our dataset; Oulfa, Maârif and Racine, at the head of the districts with the largest number of advertisements for sale in Casablanca.

Thus, neighborhoods like Oulfa, Sidi Bernoussi, Sidi Moumen, and Aîn Sebaâ are rather low- and middle-class neighbourhoods. On the contrary, Maârif, Racine, Bourgoune and Gauthier, are central districts with a large number of modern apartments.

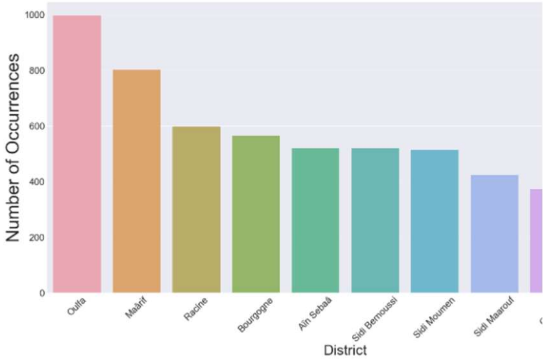


Fig. 7. The 10 Top Neighborhoods With The Highest Number Of Ads In The City Of Casablanca

The map of Casablanca (Google maps image figure 8) below allows us to locate the residential districts under study.

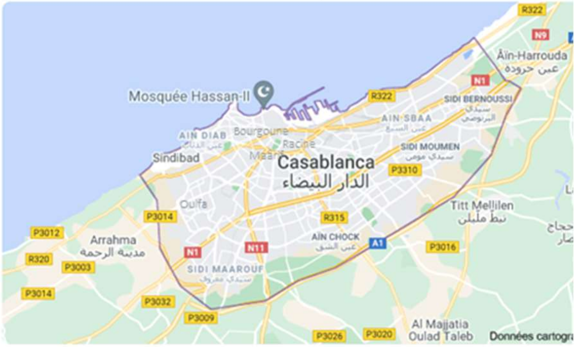


Fig. 8. Location Of Residential Districts Under Study On The Map Of Casablanca

The graphs 9 and 10 below show the distribution of different surfaces of apartments for sale in Casablanca and distribution of number of rooms.

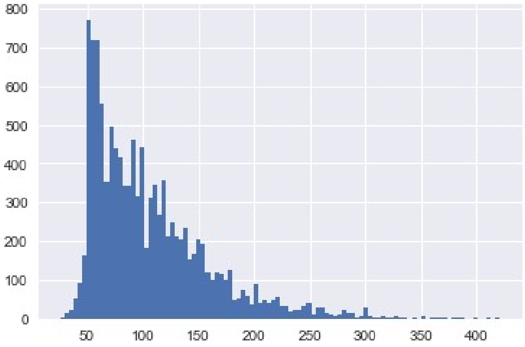


Fig. 9. Distribution Of Apartment Size

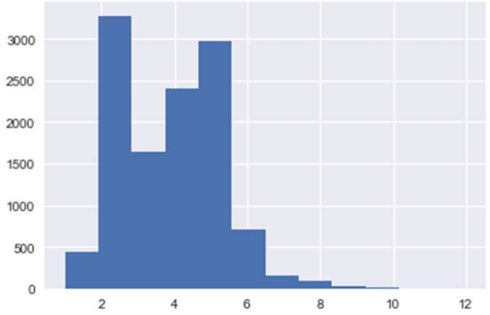


Fig. 10. Distribution Of Number Of Rooms

Figures 11 and 12 illustrate the distribution of asking prices per m2 and a representation of area against asking prices per m2, indicating the availability of exceptionally expensive apartments.

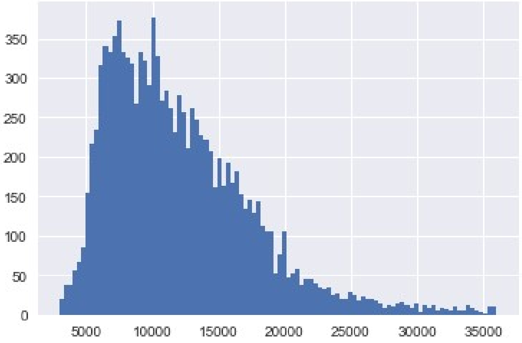


Fig. 11. Distribution Of The Asking Price Per Square Meter (Sqm)

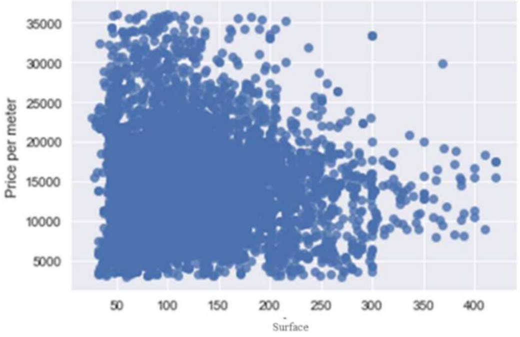


Fig. 12. Surface Vs. Apartment Price Per Sqm

4.3 Model selection

This study is carried out using several Python machine learning libraries including Pandas, Numpy, scikit-learn, Matplotlib, seaborn, etc. Besides, all experiments were executed on a desktop PC with 2.6 GHz (4 CPUs), Intel i7 CPU, 16GB RAM, and Microsoft Windows 10 operating system (64 bits). All used regressors, were used with their default configurations and utilizing the ten-fold cross validation procedure.

To evaluate the performance for every used regression technique of Casablanca real estate price forecasting, we utilized the following metrics:

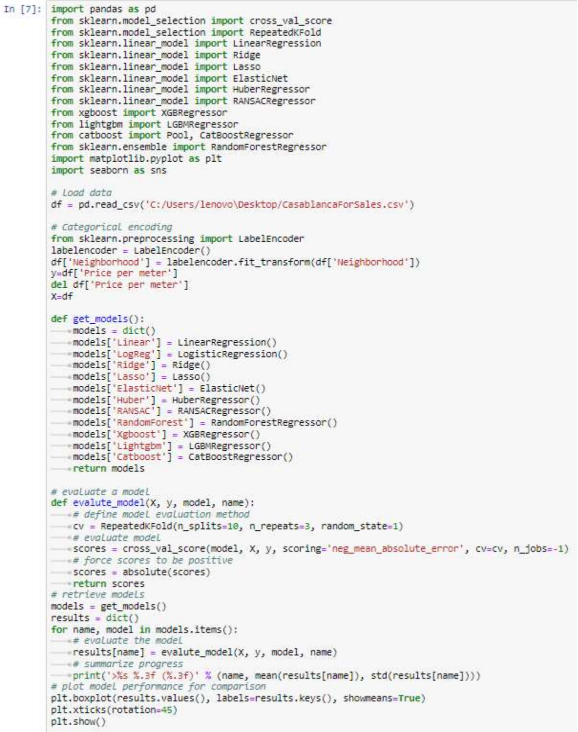
Table. 4. List of used metrics regression

Metric	The mathematical forms
R-Squared	$\mathbf{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$ (14)
Mean absolute Error	$\mathbf{MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}$ (15)
Root Mean Squared Error	$\mathbf{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}$ (16)
Mean Absolute Percentage Error	$\mathbf{MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i}}$ (17)
Mean Squared Log Error	$\mathbf{MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - (\log(y_i + 1))^2}$ (18)

where n is the number of observations, y_i actual value (asking price), $\hat{y}_i$ the corresponding predicted value and $\bar{y}$ the mean of the y_i values.

Our choice of MAE instead of MSE is justified by the fact that MSE metric penalizes large errors much more than MAE. Indeed, our dataset contains some observations with exaggerated asking prices.

Below a part python code used for the training, cross-validation and comparison of performances of models (Fig.13):



```
In [7]: import pandas as pd
        from sklearn.model_selection import cross_val_score
        from sklearn.model_selection import RepeatedKFold
        from sklearn.linear_model import LinearRegression
        from sklearn.linear_model import Ridge
        from sklearn.linear_model import Lasso
        from sklearn.linear_model import ElasticNet
        from sklearn.linear_model import HuberRegressor
        from sklearn.linear_model import RANSACRegressor
        from xgboost import XGBRegressor
        from lightgbm import LGBMRegressor
        from catboost import Pool, CatBoostRegressor
        from sklearn.ensemble import RandomForestRegressor
        import matplotlib.pyplot as plt
        import seaborn as sns

        # Load data
        df = pd.read_csv('C:/Users/lenovo\Desktop/CasablancaForSales.csv')

        # Categorical encoding
        from sklearn.preprocessing import LabelEncoder
        labelencoder = LabelEncoder()
        df['Neighborhood'] = labelencoder.fit_transform(df['Neighborhood'])
        y=df['Price per meter']
        del df['Price per meter']
        X=df

        def get_models():
            models = dict()
            models['Linear'] = LinearRegression()
            models['LogReg'] = LogisticRegression()
            models['Ridge'] = Ridge()
            models['Lasso'] = Lasso()
            models['ElasticNet'] = ElasticNet()
            models['Huber'] = HuberRegressor()
            models['RANSAC'] = RANSACRegressor()
            models['RandomForest'] = RandomForestRegressor()
            models['Xgboost'] = XGBRegressor()
            models['Lightgbm'] = LGBMRegressor()
            models['Catboost'] = CatBoostRegressor()
            return models

        # evaluate a model
        def evalute_model(X, y, model, name):
            # define model evaluation method
            cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1)
            # evaluate model
            scores = cross_val_score(model, X, y, scoring='neg_mean_absolute_error', cv=cv, n_jobs=-1)
            # force scores to be positive
            scores = absolute(scores)
            return scores
        # retrieve models
        models = get_models()
        results = dict()
        for name, model in models.items():
            # evaluate the model
            results[name] = evalute_model(X, y, model, name)
            # summarize progress
            print('>%s %.3f (%.3f)' % (name, mean(results[name]), std(results[name])))
        # plot model performance for comparison
        plt.boxplot(results.values(), labels=results.keys(), showmeans=True)
        plt.xticks(rotation=45)
        plt.show()
```

Fig. 13. Python Code Used For The Model Selection

5. EXPERIMENTAL RESULTS

The R^2, MAE, RMSE, MAPE and MSLE are used to estimate the performance of the ten techniques. The RandomForest has the ability to learn to predict the apartment price better than other algorithms, followed by the new generation of gradient boosting algorithms (Xgboost, Lightboost and Catboost) table.5.

Linear regression, Ridge regression and Lasso regression, have close results. Additionally, the same observation is found for the Elastic Net regression, Huber regression and RANSAC regression.

Table. 5. R^2, MAE, RMSE, MAPE And MSLE Values Of The Ten Used Regression Models

MODELS	R^2	MAE	RMSE	MAPE	MSLE
Linear	0.787	1420.00	1736.34	12.95%	0.024
Ridge	0.787	1420.27	1736.39	12.59%	0.024
Lasso	0.787	1420.69	1736.45	12.60%	0.024
Elastic Net	0.740	1628.31	1919.08	13.92%	0.026
Huber	0.747	1577.69	1893.29	13.65%	0.030
RANSAC	0.753	1477.30	1864.29	13.26%	0.028
RandomForest	***0.875***	***966.76***	***1323.94***	***8.35%***	***0.013***
Xgboost	0.853	1156.25	1442.89	10.01%	0.016
Lightboost	0.873	1057.73	1341.90	9.18%	0.014
Catboost	0.861	1122.76	1401.32	10.00%	0.015

a

Figure 14… 18 show the comparison of the ten regressions techniques in the form of a box plot, indicating the values of these measurement and the standard deviation.

Therefore

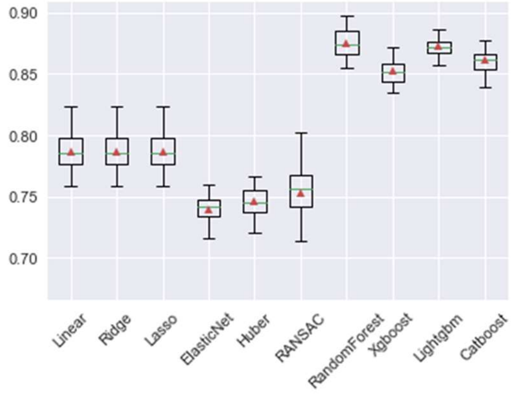


Fig. 14. Model $\mathbf{R^2}$ Performance Comparison

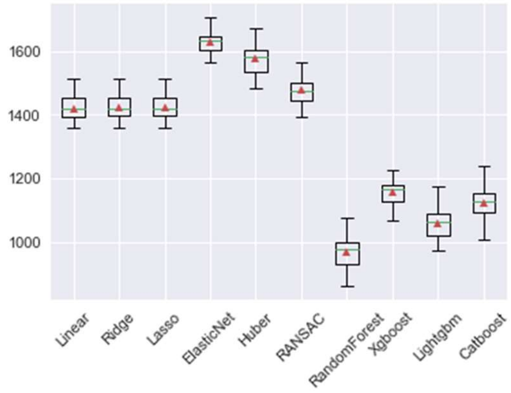


Fig. 15. Model MAE Performance Comparison

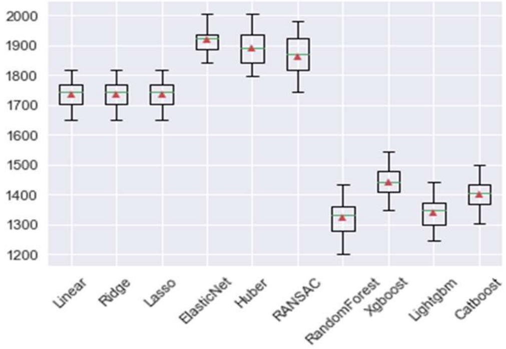


Fig. 16. Model RMSE Performance Comparison

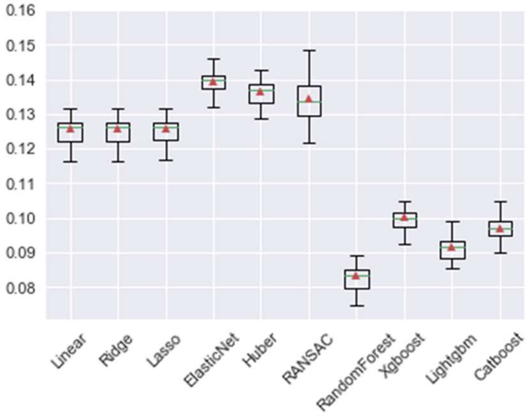


Fig. 17. Model MAPE Performance Comparison

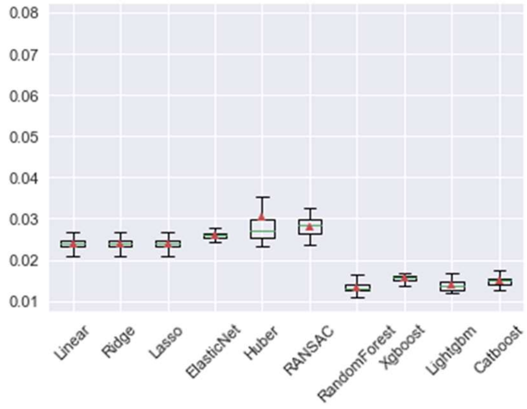


Fig. 18. Model MSLE Performance Comparison

Table. 6. Summary Of The Findings: Rank For Each Regression Technique

MODELS	R^2	MAE	RMSE	MAPE	MSLE	Rank Average
Linear	5	5	5	7	5	5.4
Ridge	5	6	6	5	5	5.4
Lasso	5	7	7	6	5	6.0
Elastic Net	10	10	10	10	8	9.6
Huber	9	9	9	9	10	9.2
RANSAC	8	8	8	8	9	8.2
RandomForest	***1***	***1***	***1***	***1***	***1***	***1.0***
Xgboost	4	4	4	4	4	4.0
Lightboost	2	2	2	2	2	2.0
Catboost	3	3	3	3	3	3.0

6. CONCLUSION AND FUTURE ENHANCEMENT

Getting an accurate price valuation of an apartment is vital for the bank to finance a mortgage. In this paper we present an approach to use machine learning techniques and alternative data from sources like real estate websites. Therefore, using efficient algorithms and fresh data can offer the bank the potential to get the fair value of banking collateral, avoid excessive risk-taking in the real estate sector and better manage the risk of current mortgage loans.

In summary, the ability of machine learning in forecasting apartment prices in Casablanca taking advantage of the potential of housing online big data, is explored in this work. Therefore, ten different regression techniques including Linear regression, Ridge, Lasso, Elastic Net, Huber, RANSAC, RandomForest, Xgboost, Lightboost and Catboost are employed in this study to construct a Casablanca apartment price prediction model. We have discovered that the performance of RandomForest is greater than that of the Linear regression, Ridge, Lasso, Elastic Net, Huber, RANSAC and new generation of gradient boosting methods (XGBoost, CatBoost and LightGBM) in terms of R^2, MAE, RMSE, MAPE and MSLE. RandomForest outperforms the other house price prediction models in every test.

Recent research has attempted to compare traditional methods with machine learning techniques like decision trees, neural networks, Random Forest (RF), Gradient Boosting and SVM. However, this study evaluates the performance of multiple classifiers in machine learning algorithms and determines which classifier is best for predicting apartment prices.

The data constraints of this work should be improved to characterize specific information more exactly. For example, the property for sale advertisement data do not contain exact GPS latitude and longitude, and we cannot obtain the information whether the apartment is well oriented and sunny or not, especially in a city where the humidity is rather high, especially during certain parts of the year. Furthermore, we were unable to acquire photographs of house interiors in order to take house interior design into consideration. Finally, we could identify other influential elements for price prediction, such as satellite maps that can be used in conjunction with saliency maps. Prediction findings will be more exact if these aspects can be addressed.

Future research efforts will be devoted to:

- Predicting Real Estate Price using text mining techniques by exploiting the description field of the sale ad.
- using other types of models (such as deep learning model) using images to improve price estimation.

REFRENCES:

[1] H. Eduard, «Mapping of real estate prices using data mining techniques » Procedia Engineering, No 1123, (2015), pp. 233 – 240.

[2] P. Byeonghwa et K. B. Jae, «Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data» Expert Systems with Applications, No 142, (2015), pp. 2928–2934.

[3] L. Michele, L. Andrea et P. Marco, «The potential of big housing data: an application to the Italian real-estate market» Bank of Italy Temi di Discussione (Working Paper), No 11171, (2018), pp. 1171.

[4] K. H. Winky, T. Bo-Sin et W. W. Siu, «Predicting property prices with machine learning algorithms» Journal of Property Research, Vol. 38, No 1, (2020), pp. 48-70.

[5] P. Ping-Feng et W. Wen-Chang, «Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices» Appl. Sci, vol. 5832, No 17: 5832, (2020), pp. 17-43.

[6] T. Quang, N. Minh, D. Hy et M. Bo, «Housing Price Prediction via Improved Machine Learning Techniques» chez Procedia Computer Science 174, 2019 International Conference on Identification, Information and Knowledge in the IOT (IIKI2019), (2020), pp. 433–442.

[7] V. Gružauskas, A. Č. D. Kriščiūnas et N. Valentinas, «Analytical Method for Correction Coefficient Determination for Applying Comparative Method for Real Estate Valuation» Real Estate Management and Valuation, Sciendo, vol. 28, No 12, (2020), pp. 52-62

[8] R. Mahdi, B.-K. Saman et A. N. Sadigh, «A heuristic approach to combat multicollinearity in least trimmed squares regression analysis» Applied Mathematical Modelling, vol. 57, May (2018), pp. 105-120.

[9] L. Larisa, «Logical Analysis of Data for outlier detection» Procedia Computer Science, vol. 169, No 1ISSN 1877-0509, (2020), pp. 330-336.

[10] J. I. Daoud, «Multicollinearity and Regression Analysis» IOP Conf. Series: Journal of Physics, No 1012009, (2017), pp. 29-49.

[11] W. Jacek et J. R. E. Pedro, «Applied regression analysis for business: tools, Traps and Applications», Springer, (2018), pp. 1-6.

[12] Y. Xue, «An overview of overfitting and its solutions» Journal of Physics: Conference Series, . IOP Publishing, vol. 1168, (2019) pp. 22-42.

[13] G. Wenxing, Q. Shanshan et Z. Zhiwen, «Generalized ridge and principal correlation estimator of the regression coefficient in growth curve mode» Linear Algebra and its Applications, No1591, (2020), pp. 115–133.

[14] L. Ji Hyung, S. Zhentao et G. Zhan, «On LASSO for predictive regression» Journal of Econometrics, No 1ISSN 0304-4076, (2021), pp. 6-19.

[15] U. Yusuke, K. Takanori et N. Kei, «Verification of Lead-Lag Effect in Financial Markets by the Adaptive Elastic Net Regression», 8th International Congress on Advanced Applied Informatics (IIAI-AAI), doi: 10.1109/IIAI-AAI.2019.00143, (2019), pp. 693-696.

[16] Z. Shunli et X. Weiwei, «Object tracking with adaptive elastic net regression» 2017 IEEE International Conference on Image Processing (ICIP), doi: 10.1109/ICIP.2017.8296752, (2017), pp. 2597-2601.

[17] H. Xianfeng, Z. Yuyang et W. Yudong, «Forecasting the real prices of crude oil using robust regression models with regularization constraints» Energy Economics, vol.86, (2020), pp. 83-104.

[18] G. Morteza, W. Kevin, C. Fiona, T. Bernard, L. Yonghuai, W. Xiaofeng et D. John, «Direct and accurate feature extraction from 3D point clouds of plants using RANSAC» Computers and Electronics in Agriculture, Elsevier B.V, No1106240, (2021), pp. 187-200.

[19] DAMA (2017), «DAMA-DMBOK. Data Management Body of Knowledge», 2nd Edition. Technics Publications LLC, August (2017), pp. 20-45.

[20] DAMA-UK (2013), «The six primary dimensions for data quality assessment», October (2013), pp. 20-45.