VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

**UNIVERSITY OF ECONOMICS AND LAW**

****



**FINAL REPORT**

**Course: BIG DATA ANALYTICS**

**Topic: HOUSE PRICES PREDICTION WITH PYSPARK**

**AND MACHINE LEARNING**

**GROUP 06**

***List of members:***

| **ID** | **Names** |
| --- | --- |
| K194141740 | Tran Thanh Phuc (Team leader) |
| K194141724 | Pham Quynh Huong |
| K194141718 | Ho Thuy Dung |
| K194141736 | Vo Thuy Uyen Nhi |
| K194141730 | Le Nguyen Phu Loc |

***Lecturer*:** Nguyen Thon Da Ph.D

*Ho Chi Minh City, 2022*

# ABSTRACT

This study uses Machine Learning models to predict and evaluate price houses whether the price house is above or below the median, and also identify factors that generally influence price houses in three counties of USA including Los Angeles, Orange and Ventura, California in 2016. The observations are all the real estate transactions in the USA which are publicly available. We proceeded to build 2 classification models including Logistic Regression and Random Forest. The results with the default threshold show that the Logistic Regression model gives a better forecasting result for the goal of minimizing FALSE NEGATIVE observations. Then, we proceeded to change the threshold for Logistic Regression models. The results show that when the threshold is reduced to 40%, the model works better as the Logistic Regression model with threshold of 40% gives less FALSE NEGATIVE observations than with the threshold of 60%. Besides, TRUE POSITIVE and TRUE NEGATIVE are also improved. So the Logistic Regression model with a threshold of 40% is better for this purpose which helps the investors, or property buyers can anticipate their budget. Also, it helps real estate agents figure the suitable customer segment out.

**TABLE OF CONTENT**

[**ABSTRACT**](#_kilgg5q0oyzq) **2**

[**LIST OF TABLES**](#_iftxutyowm6a) **5**

[**CHAPTER 1. INTRODUCTION**](#_q9ofmfkrtnwi) **6**

[**1.1. Background and justification for the research project**](#_rst4a5ddd3yx) **6**

[**1.2 Objectives of research**](#_cfx4go9k0aq7) **6**

[**1.3 Research subjects**](#_5h9qob91xnd4) **6**

[**1.4 Research scopes**](#_l50qcoz6swgg) **6**

[**1.5 Related work**](#_pdepg3wtiedo) **6**

[**CHAPTER 2: PRELIMINARIES**](#_gca3ttisn8qv) **8**

[**2.1. Theoretical basis**](#_ksiae9xd8c9w) **8**

[2.1.1. Machine learning](#_k62r0lcf7y9f) 8

[2.1.2. Supervised learning](#_fcnr3eaf3jwa) 8

[2.1.3. Logistic regression](#_71s6hifrcigk) 9

[2.1.4. Random forest](#_o1527djjyyq) 10

[2.1.5. ROC - AUC score](#_n2sryobt4qn3) 10

[2.1.6. PySpark](#_ek39ay2n6mmh) 11

[**2.2. Real state of affairs**](#_phz563878qfg) **11**

[**CHAPTER 3. DATA EXPLORATION AND DATA VISUALIZATION**](#_xkjm6hma1rh5) **13**

[**3.1. Data resource**](#_dc0pyh6l8olq) **13**

[**3.2. Variables**](#_2n7p3q1cvr3g) **13**

[3.2.1. Target](#_p0gvay6o2cce) 13

[3.2.2. Independent Features](#_atow5i4h42ry) 13

[**3.3. Data preprocessing**](#_9g9avudtb49q) **13**

[3.3.1. Check missing value](#_408eo572bp6s) 13

[3.3.2. Count number of each target element.](#_8y1d6zv50l3r) 14

[3.3.3. Transform data type from string to integer.](#_m3o9uww4kepi) 14

[3.3.4. Descriptive statistics.](#_flx5rd94y9ps) 15

[3.3.5. Correlation matrix.](#_wnb44jb3xari) 16

[3.3.6. Features’ importance through correlation coefficients.](#_dx0g6j9u4ch8) 16

[3.3.7. Scaling data set.](#_p977iogr25od) 17

[3.3.8. Splitting train-test sets.](#_u4qq0zl5t9m1) 17

[**3.4. Data visualization.**](#_fwy61dlf6obc) **17**

[3.4.1. Visualization of numeric variables.](#_tuyboqxmx59m) 17

[3.4.2. Visualization of categorical variables.](#_pjafaxjetf86) 19

[**CHAPTER 4. PROPOSED MODELS AND EXPERIMENTAL RESULTS**](#_pfc2rpy5vdl9) **25**

[**4.1. Logistic regression**](#_xixl7dymwf50) **25**

[4.1.1. Model summary](#_42m7ejq9l1g) 25

[4.1.2. Model performance](#_22dxsdim2j8i) 25

[**4.2. Random forest classifier**](#_cf4krgn4an9s) **27**

[4.2.1. Model summary](#_cv17w0avlv55) 27

[4.2.2. Model performance](#_3pvle55fme98) 27

[**4.3. Logistic regression and Random Forest comparison**](#_bm8pgn7mtii3) **28**

[**4.4. Advance Logistic Regression**](#_fufibl83olxp) **28**

[**4.5. Conclusion**](#_vdt4155uatc5) **30**

[**CHAPTER 5. CONCLUSION AND DISCUSSION**](#_sw0782v1bbg3) **32**

[**5.1. Conclusion**](#_4wfguy4qwvnq) **32**

[**5.2. Discussion and future study**](#_e73qdf3g4z75) **32**

[**REFERENCES**](#_9dg83g3vqe9w) **33**

# LIST OF TABLES

| Figure 1. | Number of missing values |
| --- | --- |
| Figure 2. | Number of each target element |
| Figure 3. | Data type of each initial column |
| Figure 4. | Data type of each column after transformation |
| Figure 5. | Descriptive statistics |
| Figure 6. | Correlation matrix |
| Figure 7. | Features’ importance |
| Figure 8. | Scaling dataset |
| Figure 9. | Lot area |
| Figure 10. | Total basement square feet |
| Figure 11. | Garage area |
| Figure 12. | Overall Quality |
| Figure 13. | Overall Condition |
| Figure 14. | Overall Fullbath |
| Figure 15. | Overall Halfbath |
| Figure 16. | Bedroom above ground |
| Figure 17. | Total number of rooms above ground |
| Figure 18. | Fireplaces |
| Figure 19. | Density charts |
| Figure 20. | Features importance of Logistic regression model |
| Figure 21. | ROC Curve - AUC of Logistic regression model on the training set |
| Figure 22. | Precision - Recall Curve of Logistic regression model on the training set |
| Figure 23. | Confusion Matrix Logistic regression model on the test set |
| Figure 24. | Features importance of Random forest classifier model |
| Figure 25. | Confusion matrix of Random forest classifier model |
| Figure 26. | ROC Curve - AUC of Advance Logistic Regression model on the training set |
| Figure 27. | Confusion matrix of Advance Logistic Regression model on the training set |
| Figure 28. | ROC Curve - AUC of Advance Logistic Regression model on the test set |
| Figure 29. | Confusion matrix of Advance Logistic Regression model on the test set |

# CHAPTER 1. INTRODUCTION

## 1.1. Background and justification for the research project

We are moving closer to experiencing housing scarcity as a result of the world's population growth. The loop is a finite resource even though the population is expanding daily. Housing costs keep rising over time as a result of the brief incident. Due to this, a large number of individuals are forced to deal with the issue of homelessness, and the cost of owning a home is steadily rising. Some investors and middlemen have made every effort to generate short-term profits by grasping the market's ups and downs. This has the additional effect of accelerating the rate of housing price bubbles and driving up housing costs. For investors, middlemen, and anybody else interested in the real estate industry, this paper will be a useful resource.

## 1.2. Objectives of research

Investigate the prices of real estates and the features that affect the prices, this report will provide specific recommendations for investors, brokers, and anybody who cares about the real estate market.

## 1.3. Research subjects

The real estate prices and the conditions that affect the house prices.

## 1.4. Research scopes

Space: The above topic is studied in the territory of USA.

Time: Research data was collected in 2016 from three counties (Los Angeles, Orange and Ventura, California).

## 1.5. Related work

Previous studies on the real estate market using machine learning approaches can be categorized into two groups: the trend forecasting of house price index, and house price valuation.

According to Changchun Wang and Hui Wu (2018), they found that the Random Forests can capture the nonlinear hidden relationship between house price and house location and give an overall better estimation than benchmark linear regression. This simple model can be scaled up for larger data with more features and captures the nonlinear information traditional models used to neglect based on attributes like: Area, bedrooms, bathrooms, …

According to CH.Raga Madhuri, Anuradha G, M.Vani Pujitha (2019), the applied machine learning models are Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and AdaBoost Regression, the results show that Gradient Boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions. Phan, T. D. (2018), Regression tree delivers a prediction result as good as linear regression, while Polynomial regression results in lower errors which is acceptable. Furthermore, the Neural Network doesn't seem to work effectively with this dataset. This may not represent the effectiveness of modern deep learning methods based on several important attributes such as: House Type, number of bedrooms, number of bathrooms, number of Car slots, and Land size. NEELAM SHINDE, 2KIRAN GAWANDE (2018) showed that Decision Tree overfits our dataset and gives the highest accuracy of 84.64%. Lasso gives the least accuracy of 60.32%. Logistic Regression and Support Vector Regression giving an accuracy of 72.81% and 67.81% respectively with their dataset. Pow, Nissan, Emil Janulewicz, and L. Liu (2014), scraped data set from duProprio.com and Centris.ca. It consists of 25000 rows and 130 columns. Approximately 70 columns of 130 were scraped. Rest 60 are based on the location of the premises. The author then implemented PCA (Principal Component Analysis) for reducing dimensions. The four techniques used by the author were Support Vector Machine, Linear Regression, K Nearest Neighbors (KNN) and Random Forest Regression and an approach to combine KNN and Random Forest Regression.

# CHAPTER 2: PRELIMINARIES

## 2.1. Theoretical basis

### 2.1.1. Machine learning

Machine Learning, according to Arthur Samuel, is "the branch of research that makes computers capable of learning without being explicitly programmed.". It is the activity of learning from data in an iterative fashion using various algorithms to develop models and predict outcomes

Machine Learning, as defined by Tom Mitchell, is "a computer program that learns from experience E to do task T, and its effectiveness is evaluated by P, if its efficacy in executing task T is measured by performance P, enhanced by experience E."

There are many popular machine learning algorithms such as: Artificial Neural Networks - ANN, Support Vector Machines - SVM, Genetic Programming - GPN, K-nearest neighbors (KNN), Logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest…And it is popular recommended uses include fraud detection, spam filtering, malware threat detection, business process automation (BPA) and Predictive maintenance.

Reason for choosing machine learning to predict models, especially in finance, is that machine learning can help financial companies make better pricing, risk, and customer behavior decisions. The technology can build models that improve understanding large data sets and uncover patterns that facilitate new business systems and processes.

### 2.1.2. Supervised learning

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In this assignment, we use supervised learning to predict the price of houses in the area.

Supervised learning: In this type of machine learning is supplied algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified with direct feedback to check if it is predicting the correct output or not. The goal of supervised learning is to train the model so that it can predict the output when it is given new data.

Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through datasets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.

One practical example of supervised learning problems is predicting house prices and this is the research subject of our report. To achieve the result, first, we find out data about the houses such as square footage, number of rooms, features, whether a house has a garden or not, and so on. Then we need to know the prices of these houses, i.e. the corresponding labels. And the results are shown in the next chapter.

### 2.1.3. Logistic regression

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not. Since we have two possible outcomes to this question - yes they are infected, or no they are not infected - this is called binary classification.

The three types of logistic regression

* Binary logistic regression - When we have two possible outcomes, like our original example of whether a person is likely to be infected with COVID-19 or not.
* Multinomial logistic regression - When we have multiple outcomes, say if we build out our original example to predict whether someone may have the flu, an allergy, a cold, or COVID-19.
* Ordinal logistic regression - When the outcome is ordered, like if we build out our original example to also help determine the severity of a COVID-19 infection, sorting it into mild, moderate, and severe cases.

Training data assumptions for logistic regression. Training data that satisfies the below assumptions is usually a good fit for logistic regression.

The predicted outcome is strictly binary or dichotomous. (This applies to binary logistic regression).

The factors, or the independent variables, that influence the outcome are independent of each other. In other words there is little or no multicollinearity among the independent variables.

### 2.1.4. Random forest

According to the definition of Breiman (2001): random forest is “a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k = 1, . . .} where the {k} are independently identically distributed random vectors and each tree casts a unit vote for the most popular class at input x”.

The random forest classification algorithm is an enhanced version of the decision tree classification technique. This algorithm, which is built from several decision trees, aids in overcoming the problem of overfitting. To pick trees in the forest, the algorithm utilizes a voting mechanism.

There are two ways to vote on a random forest. One is to choose the decision tree with the highest number of votes. The second is to select the results based on the proportion of votes, these votes are the weight of the results.

Random forest generates random trees by: bootstrapping and criteria selection.

* Bootstrapping technique: Each tree is created with a unique data set that is made up of a subset of the same size of the available data.
* Criteria: Decision trees will choose features for the tree to branch.

The random forest can indicate the importance of the model's features. It helps us to know the most important features and the unimportant features. This helps to focus more on the essentials and can be considered to remove low-impact features.

### 2.1.5. ROC - AUC score

ROC Curve stands for Receiver Operating Characteristics Curve, which is a metric to evaluate the performance of a classification model. Each point along the ROC Curve corresponds to a classification model. The ROC curve shows the sensitivity of the classification model to the ratio between true positive and false negative. The ROC that compares two operating characteristics as the criteria change is called a relative operating characteristic curve. These two performance characteristics are True Positive Rate (TPR) and False Positive Rate (FPR). An ROC with a ratio between True Positive and False Positive corresponding to 1:0 is considered ideal.

Important points in ROC Curve:

* TPR = 0, FPR = 1: the model predicts all cases to be negative class
* TPR = 1, FPR = 1: The model predicts all cases to be positive class
* TPR = 1, FPR = 0: Ideal model with 0 false classifications.

AUC, also known as Area Under Curve, is a metric commonly used to evaluate machine learning models, which is the area below the ROC curve. The AUC helps the classifier to distinguish between classes. The higher the Area Under Curve, the better the positive and negative discrimination performance of the model.

The classifier is considered perfect when AUC = 1. And AUC=0 if the algorithm only makes random guesses.

### 2.1.6. PySpark

PySpark is an interface for Apache Spark in Python. And also Spark comes with an integrated framework for performing advanced analytics that helps users run repeated queries on sets of data. Among the components found in this framework is Spark’s scalable Machine Learning Library (MLlib). MLlib can work in areas such as clustering, classification, and dimensionality reduction.

One of the reasons that we choose to use a framework like PySpark is because of how quickly it can process big data. It is faster than libraries like Pandas and Dask, and can handle larger amounts of data than these frameworks. When having over petabytes of data to process, for instance, Pandas and Dask would fail but PySpark would be able to handle it easily.

While it is also possible to write Python code on top of a distributed system like Hadoop, Spark and the PySpark API are faster and can handle real-time data. And with PySpark, we can write code to collect data from a source that is continuously updated, while data can only be processed in batch mode with Hadoop.

Furthermore, PySpark provides fault tolerance, which means that it has the capability to recover loss after a failure occurs. The framework also has in-memory computation and is stored in random access memory (RAM). It can run on a machine that does not have a hard-drive or SSD installed.

## 2.2. Real state of affairs

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a houseHousing price trends are not only the concern of buyers and sellers, but it also indicates the current economic situation. Therefore, it is important to predict housing prices without bias to help both the buyers and sellers make their decisions. There are three factors that influence the price of a house which include physical conditions, concept and location.

This assignment examines the impact on pricing of the relative size of houses. Using models which are combinations between PySparks and Python to figure out the variable AboveMedianPrice. And with simple models and getting results takes hours, not months or even days to give accurate results. Therefore, with a huge amount of variables and data, it will take a lot of time and effort. On the other hand, when the study is completed, the results may be outdated and no longer useful and timely.

Relative size impact refers to the effect on home values of house size relative to the average size of property in the same area. Unlike absolute size, which is a property characteristic, relative size is a type of neighborhood effect. There may be a number of rationales for such a pricing effect. And we found some of them such as lot area (in sq ft), overall quality (scale from 1 to 10), overall condition (scale from 1 to 10), total basement area (in sq ft), number of full bathrooms, number of half bathrooms, number of bedrooms above ground, total number of rooms above ground, number of fireplaces and garage area (in sq ft), which are described more detail in the next section.

However, there has been very little analysis of the potential effects of relative size and the appraisal profession has not developed any speciﬁc deﬁnitional or procedural guidelines for the concept. So this study is an exploratory attempt to better quantify the pricing effects associated with relative size.The results show a statistically signiﬁcant negative effect from the rela-tively larger home sizes and a hypothesized positive effect for the sales of relatively smaller homes. The use of quartile groupings results in statistically sig-niﬁcant effects with relative size effects increasing as distance from the average-sized house increases.

# CHAPTER 3. DATA EXPLORATION AND DATA VISUALIZATION

## 3.1. Data resource

Zillow Prize competition, a one million dollar grand prize competition in 2017, challenging the data science community to further enhance Zestimate's accuracy. In it, you are provided with a complete list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data for 2016. However, the dataset we will use in this research is adapted from Zillow’s Home Value Prediction Kaggle competition data. We’ve reduced the number of input features and changed the task into predicting whether the house price is above or below median value.

## 3.2. Variables

### 3.2.1. Target

* Above median price or not.

The median home price is the price of a home in the middle of a list of properties ranked from highest to lowest for a specific area and time period. The median house price in our dataset is a binary variable with two values: 0 and 1. Where 0 represents a selling price below the median and 1 represents a higher price.

### 3.2.2. Independent Features

1 - Lot area (in square feet).

2 - Overall quality (scale from 1 to 10)

3 - Overall condition (scale from 1 to 10)

4 - Total basement area (in square feet)

5 - Number of full bathrooms

6 - Number of half bathrooms

7 - Number of bedrooms above ground

8 - Total number of rooms above ground

9 - Number of fireplaces

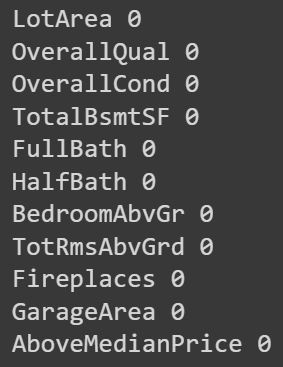
10 - Garage Area (in square feet)

## 3.3. Data preprocessing

### 3.3.1. Check missing value

We should first check if our dataset is missing many values to know the quality of the dataset.

**Figure 1. Number of missing values**

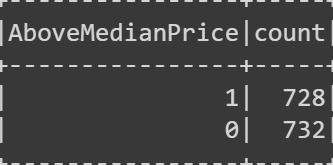
****

As we can see there is no missing value in all columns. So our dataset must be very well collected.

### 3.3.2. Count number of each target element

Our target variable in this dataset is the AboveMedianPrice column. So we will test the number of observations of each value in this variable to see if they are balanced or not?

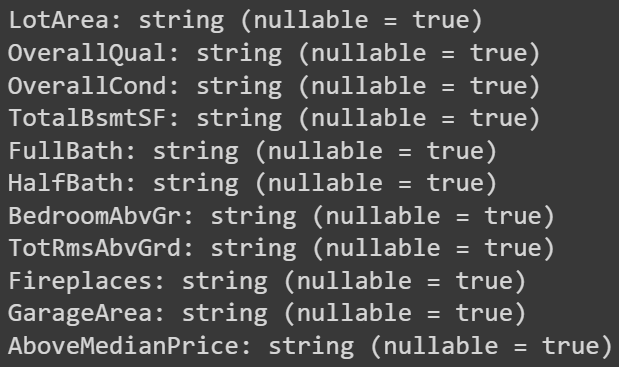
**Figure 2. Number of each target element**

****

We can easily see that the target variable has a total of 728 observations with a value of 1 and 732 observations with a value of 0. They account for almost equal proportions. This is great and we don't need to resampling like over-sampling, under-sampling, etc.

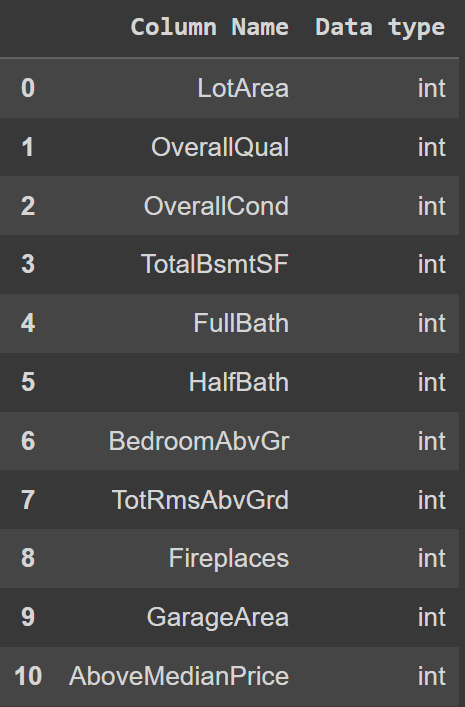
### 3.3.3. Transform data type from string to integer

**Figure 3. Data type of each initial column**

****

In the original dataset, all columns have a data type string. We will convert all data types to integer to accommodate the following processes.

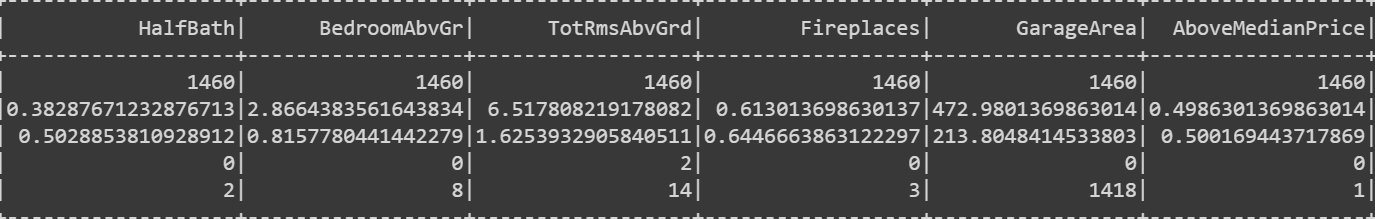
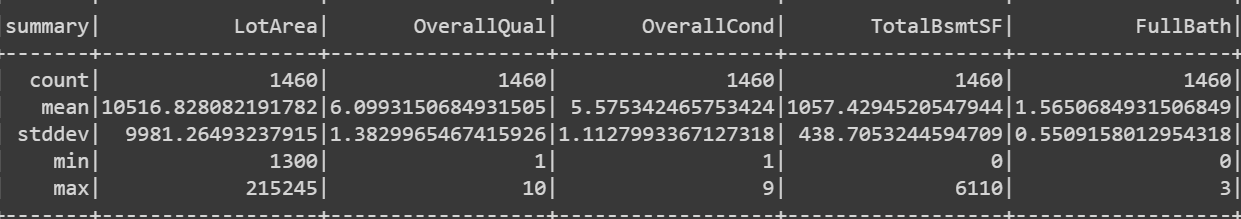
**Figure 4. Data type of each column after transformation**



It's easy to see that after transformation, they've all become Interger (int) types.

### 3.3.4. Descriptive statistics

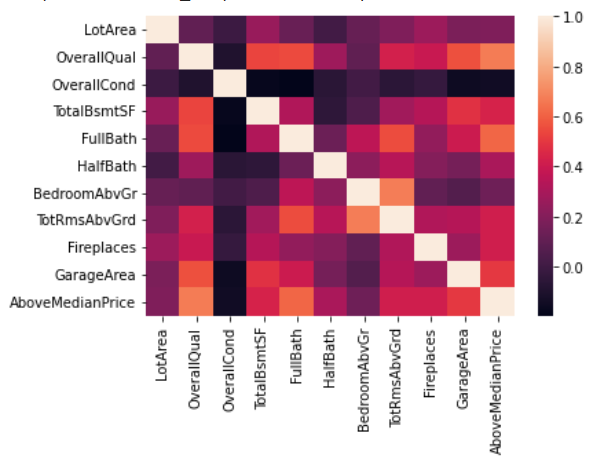
**Figure 5. Descriptive statistics**



We ran descriptive statistics for all the variables in the dataset as you can see in the image above.

### 3.3.5. Correlation matrix

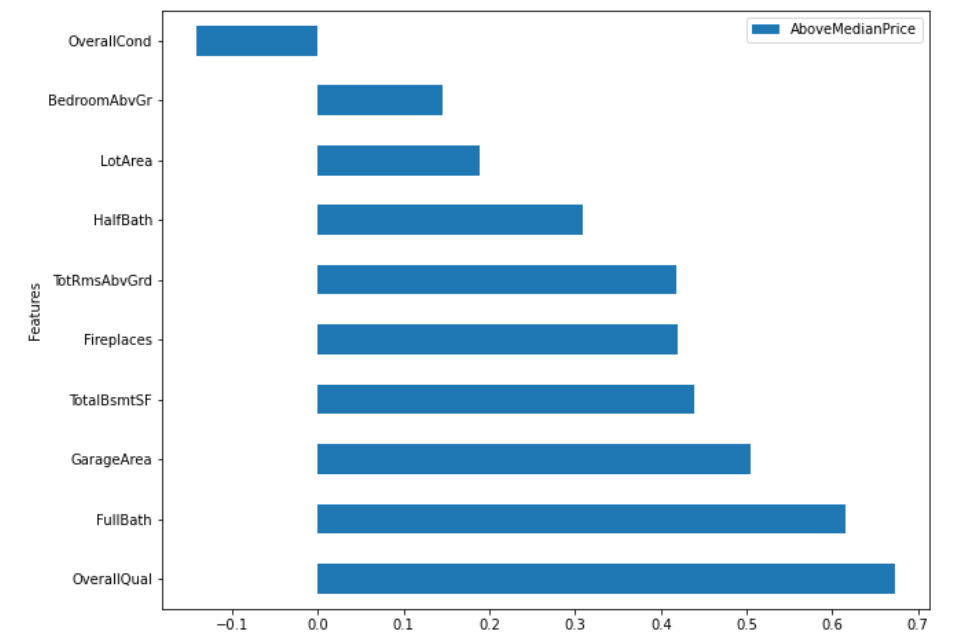
**Figure 6. Correlation matrix**



We also ran the correlation matrix between the features. As can be seen, the OverallCond feature has no or very little correlation with the other features. In addition, the LotArea feature is also relatively uncorrelated with the rest of the features. On the other hand, the remaining features are correlated quite similarly and at a decent level.

### 3.3.6. Features’ importance through correlation coefficients

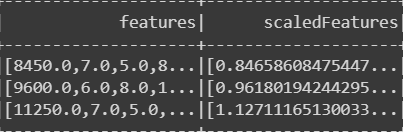
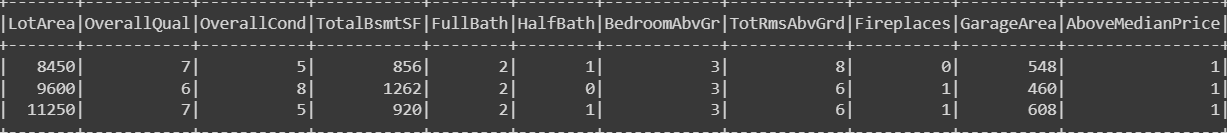
**Figure 7. Features’ importance**



Only one OverallCond feature has a negative effect on the results with a magnitude of about -0.13. The remaining features have a positive effect. In which, OverallQual and FullBath feartures have the greatest influence, 0.67 and 0.61. The two features with little positive influence are BedroomAbvGr and LotArea of 0.15 and 0.19.

### 3.3.7. Scaling data set

**Figure 8. Scaling dataset**

****

You are aware that if the data in any condition has data points that are far apart, scaling is a technique to bring them closer together. In other words, scaling is used to generalize data points so that the distance between them is reduced. Therefore, we have scaled the dataset and the results are as you can see in the image.

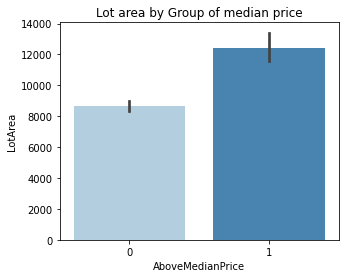
### 3.3.8. Splitting train-test sets

After scaling the dataset, we proceed to split the dataset into two datasets, the train dataset and the test dataset. We divide by 0.8 for the train dataset and 0.2 for the test dataset. As a result, there are 1157 observations in the train dataset and 303 observations in the test dataset.

## 3.4. Data visualization

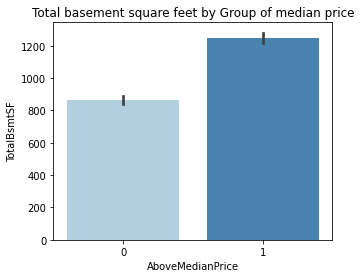
### 3.4.1. Visualization of numeric variables

**Figure 9. Lot area by group of median price**



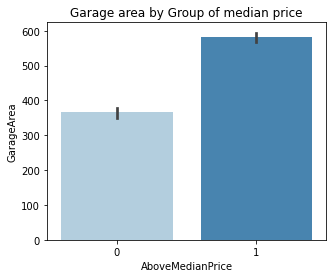
From Figure 9, it is easy to see that the number of observations with a higher than median price will have a larger value in the Lot area variable, the maximum value is up to nearly 14000 square feet. On the other hand, observations that cost less than the median have a maximum value of only about 9000 square feet. Thereby, we see that houses with prices greater than the average value will usually have a larger lot area.

**Figure 10. Total basement square feet by group of median price**



Similarly, the variable Total basement square feet also has a maximum value of more than 1200 square feet with observations costing more than the median price. Observations that cost less than the median price have a maximum value of only about 900 square feet.

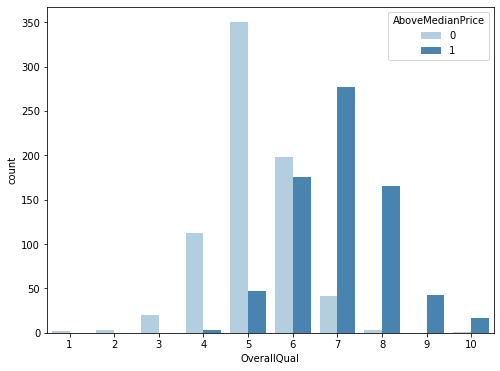
**Figure 11. Garage area by group of median price**

****

Just like the previous two figures, the Garage area variable also has its maximum value up to nearly 600 square feet with observations worth more than the median. Observations that are priced below the median have a maximum value of only about 400 square feet.

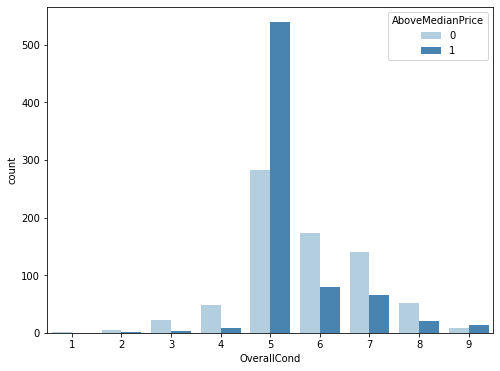
### 3.4.2. Visualization of categorical variables

**Figure 12. Overall quality by group of median price**

****

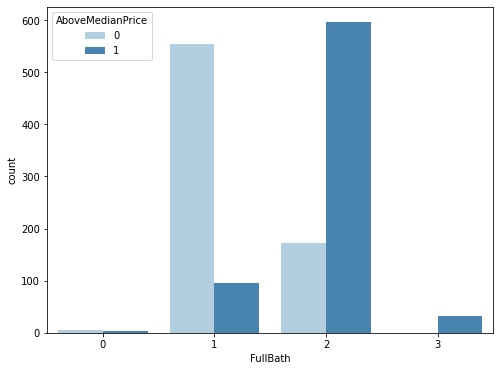
In observations with a target variable of 0, we see that the Overall Quality variable mainly has a value of 4,5,6. In which, the number of observations with a value of 6 is up to about 350 observations. On the other hand, in observations with a target variable of 1, mainly the Overall Quality variable has the value of 6,7,8 and the value of 7 has the highest number of more than 250 observations.

**Figure 13. Overall condition by group of median price**

****

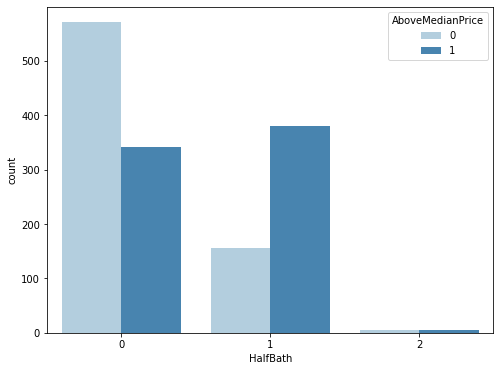
Regarding the Overall Condition variable, the observations with a target variable of 0 spread quite widely, ranging mostly from the value 4 to 8 and the value 5 had the largest number of observations. For observations with a target variable of 1, most of the Overall Condition variable focuses on the value 5, the rest is quite small but mostly values greater than 5.

**Figure 14. Overall fullbath by group of median price**

****

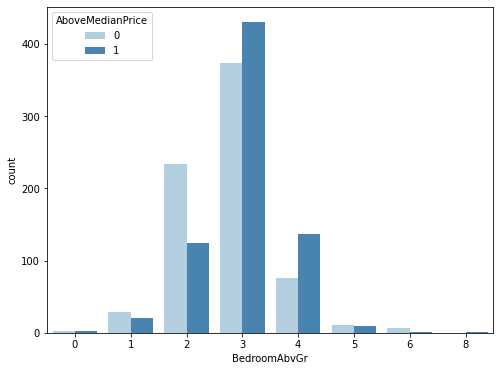
For the variable Overall Fullbath, mainly with the value 1 are observations with a target variable of 0 and with a value of 2 for observations with a target variable of 1. Observations with values of 0 and 3 are very few. In which, all the observations with the value 3 are the observations with the target variable of 1. This means that only the houses with the value higher than the median value can have the Fullbath variable of value 3.

**Figure 15. Overall halfbath by group of median price**

****

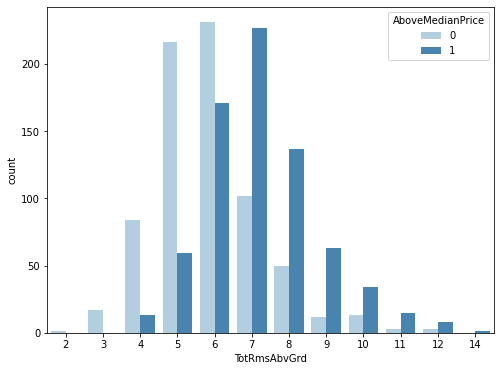
With the Overall Halfbath variable, observations with a target variable of 0, the Overall Haflbath variable has a weak value of 1 and 0. This all shows that homes priced below the median price often do not have Half Bath. For observations with a target variable of 1, the number of 0 and 1 values is quite equal. Observations with a value of 2 are very few.

**Figure 16. Bedroom above ground by group of median price**

****

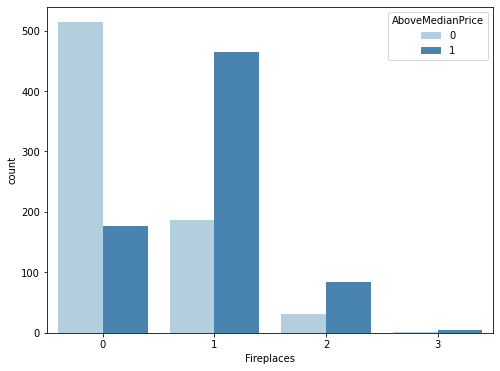
Most of the houses in the variable Bedroom above ground have a value of 3. In which the observations with the target variable of 0, there are mainly values in 2 and 3. For the houses with the target variable of 1, have slightly broader values from 2 to 4. The remaining values have a very small number of observations.

**Figure 17. Total number of rooms above ground by group of median price**

****

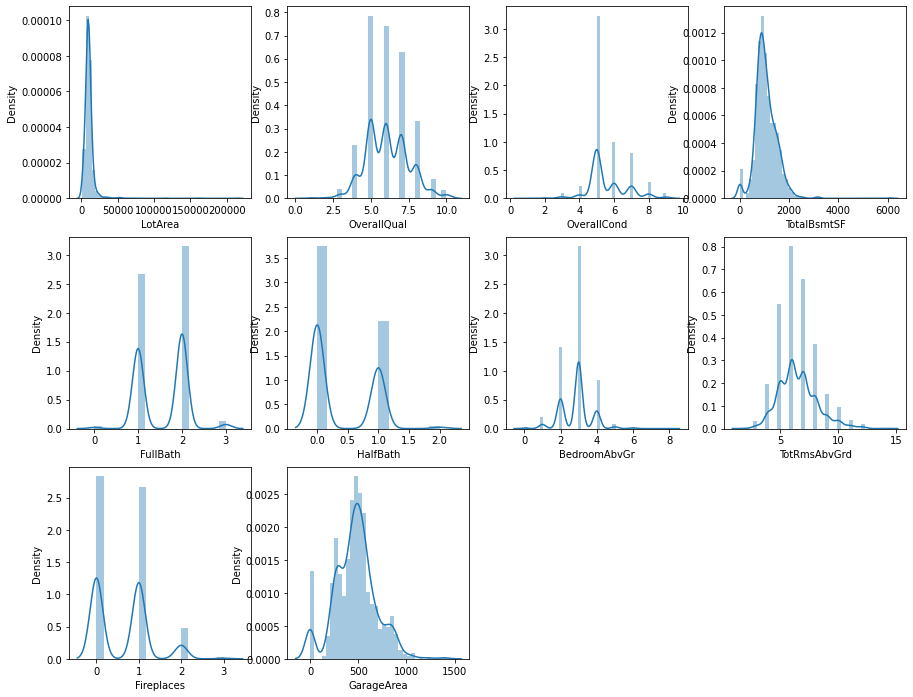
Most homes priced below the median have a Total number of rooms above ground from 3 to 8. For homes with a higher value, they have a value between 5 and 9. It is easy to see, the Higher value homes will often have more room.

**Figure 18. Fireplaces by group of median price**

****

Most homes priced below the median have no or only one fireplace. For homes with higher value, there will be 3 cases of no, 1 and 2 fireplaces. Very few observations have 3 fireplaces.

**Figure 19. Density charts**

****

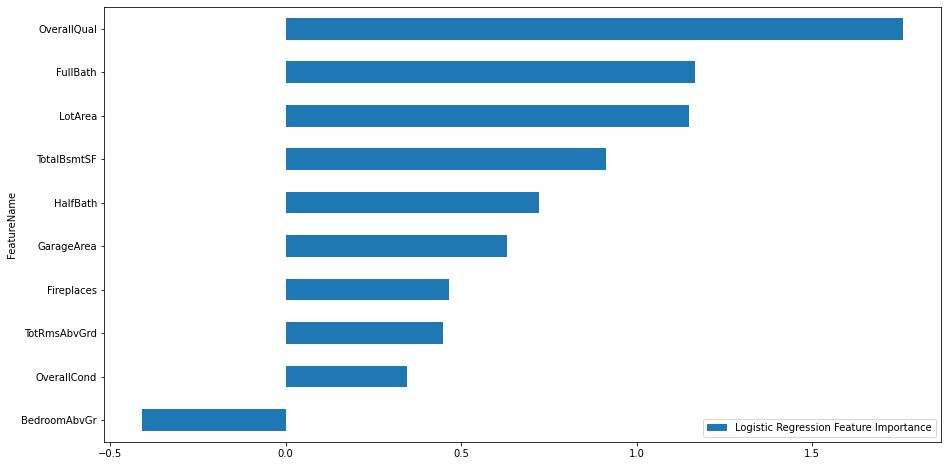
This is a composite image showing the weighted values for all independent variables. You can zoom in to see more clearly the distribution proportions of each variable.

# CHAPTER 4. PROPOSED MODELS AND EXPERIMENTAL RESULTS

## 4.1. Logistic regression

### 4.1.1. Model summary

**Figure 20. Features importance of Logistic regression model**

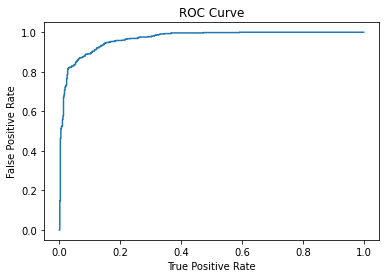


The magnitude of features are shown above. The result implies that only the BedroomAbvGr has a negative effect while the OverallQual has the largest effect when predicting the real estate price.

### 4.1.2. Model performance

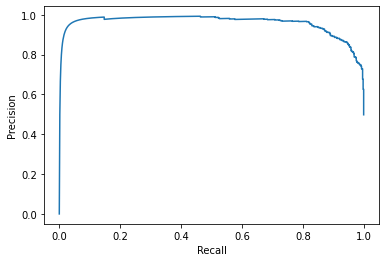
* On the training set:

**Figure 21. ROC Curve - AUC of Logistic regression model on the training set**



The curve above shows that model performance is very good since the AUC score is almost 1 (0.967).

**Figure 22. Precision - Recall Curve of Logistic regression model on the training set**



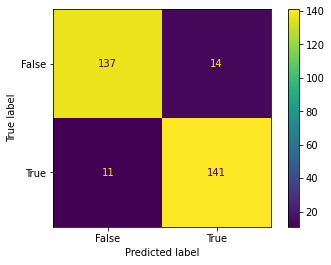
The curve above shows that the optimized performance of the model can be achieved with both precision and recall scores that are higher than 0.8.

* On the test set:

ROC Curve - AUC

The AUC when predicting on the test set is even higher when predicting on the training set with the score 0.973. It implies that the model performs very well on the test set.

**Figure 23. Confusion Matrix Logistic regression model on the test set**



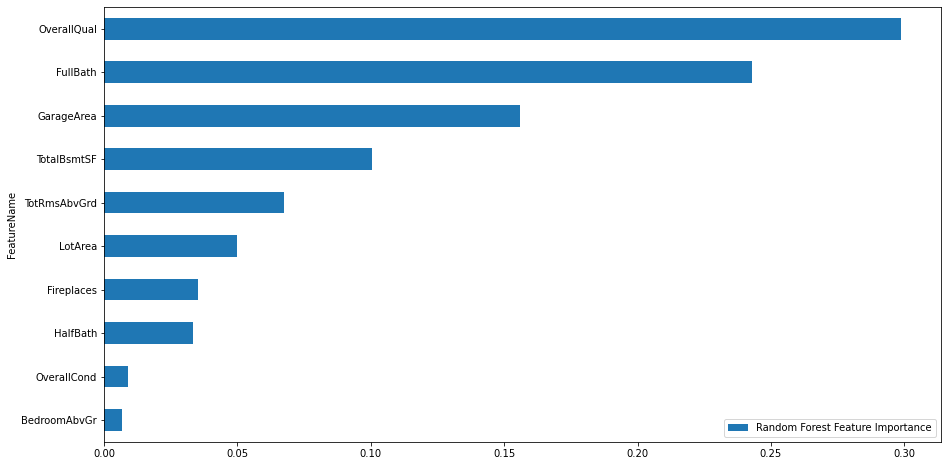
=> The matrix above shown that there are only small mistakes when the model tries to predict this dataset and two classes are classified very well.

Conclusion: Based on the performance on the training set and test set, the model performs very well with the accuracy approximately 92% and doesn't have problems about overfitting or underfitting.

## 4.2. Random forest classifier

### 4.2.1. Model summary

**Figure 24. Features importance of Random forest classifier model**



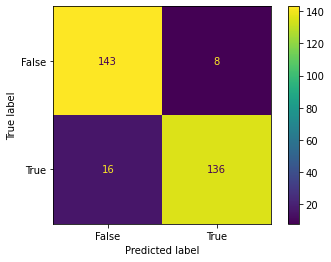
=> The figure above shows that OverallQual, FullBath and GarageArea are the most important features since they contribute more than 70% to the prediction decision.

### 4.2.2. Model performance

The model predicted very well overall with the accuracy score approximately 92% and AUC score 0.978.

The performance of model with each class are shown below by the confusion matrix:

**Figure 25.**  **Confusion matrix of Random forest classifier model**



=> The matrix above implies that the model also predicts well on both classes and makes very few mistakes.

From the confusion matrix above we can infer that the precision and recall score of this model is also very high (0.94 and 0.89). These scores imply that the model is better at predicting the true values than trying to predict all the cases of the true values.

## 4.3. Logistic regression and Random Forest comparison

Both Logistic Regression and Random Forest models give similarly high scores on the 'house price data' dataset.

* Logistic Regression: Accuracy (91.75%), AUC (97.23%), Precision (91%), Recall (93%)
* Random Forest: Accuracy (90.76%), AUC (97.24%), Precision (94%), Recall (87%)

Since the purpose of this model is to predict house prices whether they are higher (TRUE) or lower (FALSE) than the Median Price.

Therefore, we should minimize the number of FALSE NEGATIVE (FN) forecasts. It helps the investors, or property buyers can anticipate their budget. Also, it helps the real estate agents can figure the suitable customer segment out. Based on the confusion matrixes, we see that the Logistic Regression model gives less FALSE NEGATIVE observers than the Random Forest model (11 wrong predictions of Logistic Regression are less than 20 wrong predictions of Random Forest). --> Logistic Regression model is better for this purpose.

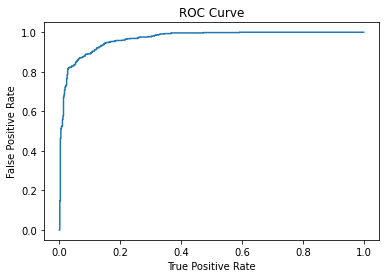
## 4.4. Advance Logistic Regression

* On the training set:

After selecting the logistic regression, we want to achieve the better prediction with this algorithm, so we try to adjust the threshold of the model from 0.5 (default) to 0.6 and 0.4 to test the performance.

***The results of model (t=0.6) are shown below:***

**Figure 26. ROC Curve - AUC of Advance Logistic Regression model on the training set**

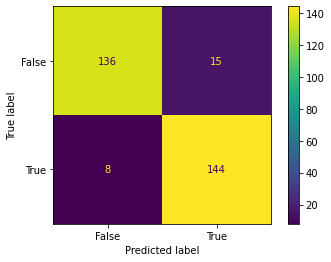


=> The AUC score of the training set of this model is quite similar to the default logistics regression model (approximately 0.967).

* On the test set:

ROC Curve - AUC (Test set): The AUC score of the test set is 0.972 that is also similar to the default logistic regression.

**Figure 27. Confusion matrix of Advance Logistic Regression model on the training set**

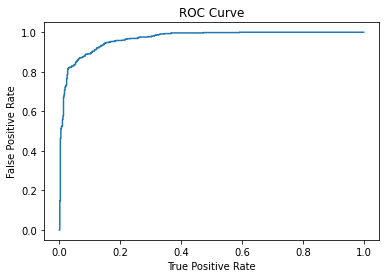


=> The matrix above shown that the precision of two models are the same as 0.91 but the recall of this algorithm is lower than the default algorithm (0.88 < 0.93).

Accuracy score: The accuracy of this algorithm is slightly higher than the default algorithm (0.924 > 0.9207), it implies that the prediction of this model is a little better than the default algorithm.

***The results of model (t=0.4) are shown below:***

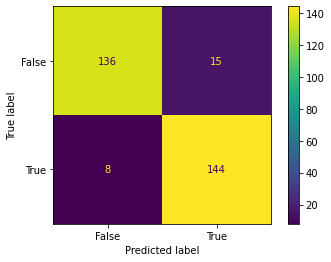
**Figure 28. ROC Curve - AUC of Advance Logistic Regression model on the test set**



=> The AUC score of the training set of this model is quite similar to the default logistics regression model (approximately 0.967).

ROC Curve - AUC (Test set): The AUC score of the test set is 0.972 that is also similar to the default logistic regression.

**Figure 29. Confusion matrix of Advance Logistic Regression model on the test set**



=> The matrix above shows that the precision of two models is the same as 0.91 but the recall of this algorithm is higher than the default algorithm (0.95 < 0.93).

Accuracy score: The accuracy of this algorithm is slightly higher than the default algorithm (0.924 > 0.9207), it implies that the prediction of this model is a little better than the default algorithm.

## 4.5. Conclusion

Both thresholds (40% and 60%) for Logistic Regression models give similarly high scores on the house price data' dataset.

* Threshold 40%: Accuracy (92.41%), AUC (97.23%), Precision (91%), Recall (95%)
* Threshold 60%: Accuracy (89.77%), AUC (97.23%), Precision (91%), Recall (88%)

Since the purpose of this model is to predict house prices whether they are higher (TRUE) or lower (FALSE) than the Median Price.

Therefore, we should minimize the number of FALSE NEGATIVE (FN) forecasts. It helps the investors, or property buyers can anticipate their budget. Also, it helps real estate agents figure the suitable customer segment out.

Based on the confusion matrixes, we see that the Logistic Regression model with threshold of 40% gives less FALSE NEGATIVE observes than the Random Forest model (8 wrong predictions of Logistic Regression with threshold of 40% are less than both of 18 wrong predictions of Logistic Regression with threshold of 60% and 11 wrong predictions of Logistic Regression with default threshold).

Besides, TRUE POSITIVE and TRUE NEGATIVE also are improved.

=> Logistic Regression model with threshold of 40% is better for this purpose.

# CHAPTER 5. CONCLUSION AND DISCUSSION

## 

## 5.1. Conclusion

The result of the implementation implies that the model also the features we select can be used for predicting the price of the real estate whether above or below the median value. After examining our model coefficients, we observed that features with the largest effect on housing price predictive power are Overall quality, Number of full bathrooms, Garage area, and Lot area. They are quite good factors that investors, brokers, and anybody who cares about the real estate market can consult before making decisions.

## 5.2. Discussion and future study

After realizing the good results of the model, we think a good next step can be to try to transform the problem into the regression matter that can predict the exact price of the real estate instead of predicting the trend of the price.

Although the model perform so well, there are some problems that we may consider such as the price of the real estate is momentary i.e. their price can be steadily increase or decrease after just one month, or there are still some unsystematic factors that we can not control (pandemic and inflationary).

To achieve things we discuss above, we can collect more and more data with a variety of features and try to build some complex algorithms (boosting, neural networks…)

# REFERENCES

1. The Effect of Zoning on Housing Prices Ross Kendall and Peter Tulip, <https://www.rba.gov.au/publications/rdp/2018/pdf/rdp2018-03.pdf>.
2. Predicting house prices with machine learning methods, ALAN IHRE, ISAK ENGSTRÖM, <https://www.diva-portal.org/smash/get/diva2:1354741/FULLTEXT01.pdf>.
3. House Price Prediction Using Machine Learning, 2021, The Journal of Philosophy Psychology and Scientific Methods 9 (6):2455-6211, <https://www.freecodecamp.org/news/how-to-build-your-first-neural-network-to-predict-house-prices-with-keras-f8db83049159/>.
4. Albouy D and G Ehrlich (2017), ‘Housing Productivity and the Social Cost of Land-Use Restrictions’, NBER Working Paper No 18110, rev December.
5. CEDA (Committee for Economic Development of Australia) (2017), Housing Australia, Research report, Committee for Economic Development of Australia, Melbourne.
6. Cheshire PC and CAL Hilber (2008), ‘Office Space Supply Restrictions in Britain: The Political Economy of Market Revenge’, The Economic Journal, 118(529), pp F185–F221.
7. Cheshire PC, CAL Hilber and HRA Koster (2018), ‘Empty Homes, Longer Commutes: The Unintended Consequences of More Restrictive Local Planning’, Journal of Public Economics, 158, pp 126–151.