

sa

**Diandaru Suchrady**

**ID-21-0105**

**Institut Teknologi Bandung**

# CHAPTER I

**Professionals Opinion**

Before making any assumptions about the features contained in the dataset, the proper way to find a clue of whether a feature is considered necessary or not is by asking the professionals. According to Susi Sundari (2021), a senior property marketer, there are three factors of a customer’s preference for buying a house:

1. Price

The price depends on the customer’s budget.

1. Location

Customers are very likely to buy a house near their workplace, relatives, and their old residence.

1. House Shape

House shape affects the customer’s interest.

Susi also stated that there is no correlation between customer’s preferences and date of purchase. On the other hand, Dady Suchrady (2021), a public notary inferred that dates play a role in the event of house or property transactions. From his past experience, documents for transactions are usually done by the month of April, hence it is inferred that transactions mostly occur in the second semester of the year. In addition to that, many of Dady’s clients have more budget during the second semester since they gain more income that time of the year (In Indonesia, when the government gives projects to companies, their employees may gain bonuses). Therefore, it can be assumed that there is a chance that a house with a higher price will have more demand in the second semester of a year.

**Research Study**

According to the results of a study conducted by Amrin Fauzi (2012) about what affects a consumer's decision in buying a house using a regression analysis, consumer’s decision in buying a house depends with the rate of 64.2% on location, economic status, and lifestyle.

In another study with the same topic conducted by Sutianingsih (2010), quality of a building, pricing, location, and promotion determine the decision making of a housing consumer by 61.1% and the 38.9% is affected by other variables.

**Assumptions**

Since the test dataset ‘Test.csv’ does not contain the ‘cnt’ and ‘dealing’ columns, backed up by the decision to make the machine learning model to only evaluate categorical variables, the ‘cnt’ and ‘dealing’ columns are not to be included in the model construction. In addition, the ‘buyer\_id’ will not be involved in the model to prevent any overfitting to happen since a buyer’s identity is deemed irrelevant.

**Decision**

It is concluded that it is best to not include the ‘cnt’, ‘dealing’, and ‘buyer\_id’ column. In the further process of preparing the dataset, data reduction and feature selection will be held to reduce the dimension of the dataset, so there is a chance of excluding other columns.

# CHAPTER II

**Preprocessing**

### Data Cleansing

### In this process, firstly the information of the data must be seen.

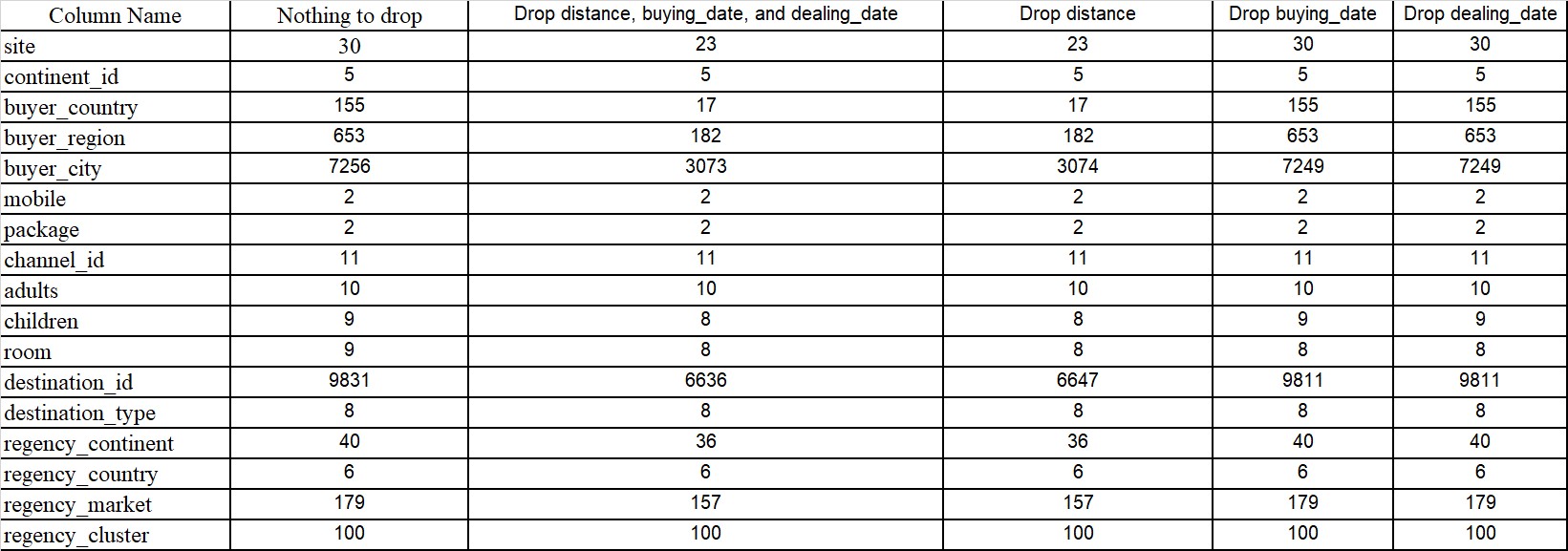
### 

As the ‘cnt’, ‘dealing’, and the ‘buyer\_id’ have been decided to be dropped, the dataset will remain the rest columns (21 columns).

From the data information, it is seen that the ‘distance’, ‘buying\_date’, and ‘dealing\_date’ contain null values. Since the majority of the data types are categorical, before deciding to drop the null values, it is best to check the class loss when the null values are dropped.

Following are the class numbers for each feature before and after dropping the null values:

**Table 2.1. Class Value Counts of Each Feature**



Preventing the loss information in categorical features, it is decided to fill the null values and not drop the outliers.

**Data Transformation**

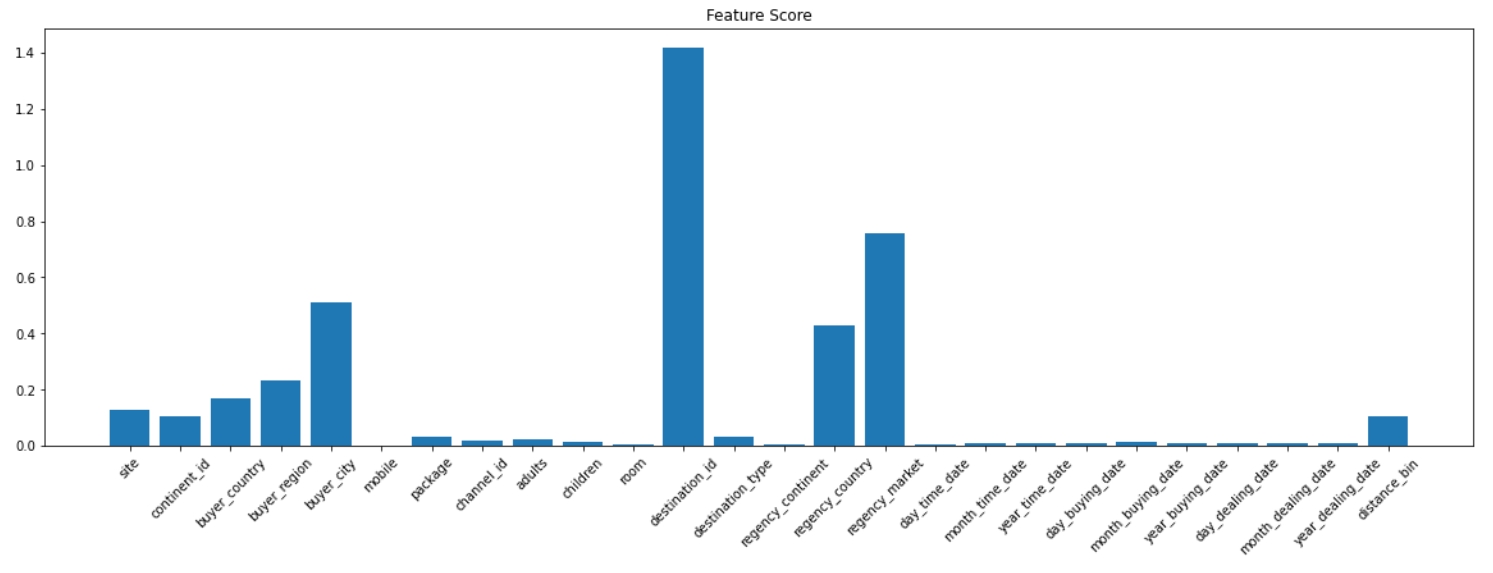
In the data transformation part, the distance column is grouped into 5 corresponding bins with the interval of every bins are as the following:

1. (0.0046, 326.593] labeled as 0
2. (326.593, 1100.383] labeled as 1
3. [(1100.383, 2808.036] labeled as 2
4. (2808.036, 11761.396]] labeled as 3
5. Null values labeled as -1

The date columns (time\_date, buying\_date, and dealing\_date) are extracted to day, month, and year. This process results in all columns bening of categorical type.

**Data Reduction**

To reduce the burden of the model while fitting/training the dataset, feature selection should be conducted. Mutual information feature selection is chosen because the dataset’s majority columns are categorical or numeric.

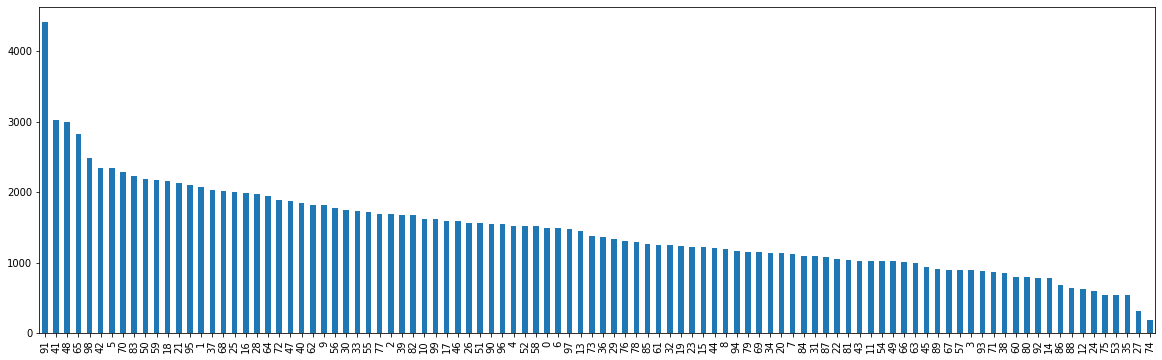


**Graph 2.1. Feature score bar plot**

After running the feature selection, it is concluded that the chosen columns for the modelling process are ‘site’, ‘continent\_id’, ‘buyer\_country’, ‘buyer\_region’, ‘buyer\_city’, ‘destination\_id’, ‘regency\_country’, ‘regency\_market’, and ‘distance\_bin’.

**Data Balancing**

In order to minimize overfitting in our model, we need to take a look at the frequency of each class in the target feature. A dominant class in the target feature may result in a model that assumes classes that have small frequencies as an insignificant class, thus making it less considered by the model. The following is the bar plot for the frequency distribution of the target class ‘regency\_cluster’ of the dataset Train.csv



**Graph 2.2. Target class frequency distribution**

As seen on the bar chart, there is an imbalance between the distribution of class frequencies. Over-sampling is done to resolve this issue, that is extrapolating classes of the target attribute that have less entries using Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) and Synthetic Minority Over-sampling Technique for Nominal (SMOTEN) through scikit library. Thus, every class from the target attribute now has 8077 entries.

**Modelling and Validation**

An experiment is done to determine the best algorithm to predict the target attributes. The experiment compares the performance of Random Forest Classifier and the Decision Tree Classifier using the accuracy as a metric. Each of the algorithms will use both the entropy and gini function. The models evaluate the result using ten fold cross validation.

There are twelve combinations of model:

1. Random forest classification with entropy function splitting using the SMOTEN transformed dataset
2. Random forest classification with gini function splitting using the SMOTEN transformed dataset
3. Random forest classification with entropy function splitting using the SMOTENC transformed dataset
4. Random forest classification with gini function splitting using the SMOTENC transformed dataset
5. Random forest classification with entropy function splitting using the imbalanced dataset
6. Random forest classification with gini function splitting using the imbalance dataset
7. Decision tree classification with entropy function splitting using the SMOTEN transformed dataset
8. Decision tree classification with gini function splitting using the SMOTEN transformed dataset
9. Decision tree classification with entropy function splitting using the SMOTENC transformed dataset
10. Decision tree classification with gini function splitting using the SMOTENC transformed dataset
11. Decision tree classification with entropy function splitting using the imbalanced dataset
12. Decision tree classification with gini function splitting using the imbalanced dataset

# CHAPTER III

The result of the experiment is as follows

|  |  |
| --- | --- |
| Model | Accuracy |
| 1 | 49.63 % |
| 2 | 49.63 % |
| 3 | 47.23 % |
| 4 | 47.23 % |
| 5 | 28.22 % |
| 6 | 28.22 % |
| 7 | 49.81% |
| 8 | 49.77% |
| 9 | 47.29% |
| 10 | 47.19% |
| 11 | 28.40% |
| 12 | 28.38% |

It can be inferred from the table that models trained with unbalanced dataset (5, 6, 11, 12) performed worse than models that were trained using balanced datasets either using SMOTEN or SMOTENC. This may be due to the overfitting of the model due to the dominance of some target attribute class that makes up a large percentage of the training dataset. It can also be concluded from the results that SMOTEN (1, 2, 7, 8) data balancing works better than SMOTENC (3, 4, 9, 10), meaning that it is better to assume all attributes in this dataset as categorical. Although only by a small percentage, the entropy function works better than the gini function in terms of feature selection.

# CHAPTER IV

**Conclusion**

Based on the conducted experiment, it is found that the balance of the data greatly affects the performance of the model, therefore data balancing is a crucial part of data preprocessing. In addition, mutual information can be used to determine which categorical attributes are best to be put into the model which concludes that in this case, consumer preference in choosing the residence cluster is highly dependent on location represented by continent\_id, buyer\_country, buyer\_region, buyer\_city, destination\_id, regency\_country, and distance\_bin.

**Suggestion**

More data preprocessing methods may be done to improve the performance of the model and the experiment may involve more varying algorithms.