

**School of InfoComm Technology**

**Machine Learning**

Diploma in Data Science (DS)

Diploma in Information Technology (IT)

October 2022 Semester

**INDIVIDUAL ASSIGNMENT 1**

(30% of Machine Learning Module)

**Deadline for Submission:**

**17th Dec 2022 (Saturday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| Student Name | : | Kiara Avendano |
| Student Number | : | S10219186a |
| Video Presentation Link | : | <https://youtu.be/qzLF6wb8rDs>  <https://youtu.be/qzLF6wb8rDs> |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 24th Dec 2022, 23:59.

Contents

[Overview 3](#_Toc122212946)

[HR Analytics 3](#_Toc122212947)

[Loading and EDA 3](#_Toc122212948)

[Cleansing and Transformation 7](#_Toc122212949)

[Correlation Analysis 8](#_Toc122212950)

[Final Dataset 9](#_Toc122212951)

[Airbnb 10](#_Toc122212952)

[Loading and EDA 10](#_Toc122212953)

[Cleansing and Transformation 12](#_Toc122212954)

[Correlation Analysis 13](#_Toc122212955)

[Final Dataset 14](#_Toc122212956)

[Summary 15](#_Toc122212957)

[Further improvements 15](#_Toc122212958)

[References 16](#_Toc122212959)

# Overview

This report explores and understands the two given datasets, “hr\_data.csv” and “listings.csv”, through visualizations and applying statistical concepts. With consideration to the columns present in each dataset, the report will showcase how the data has been wrangled and processed to prepare each dataset fit for machine learning modelling. The final datasets produced at the end of the report would be used to train a classification model and a regression model respectively. Throughout the report, trial and error were conducted to determine the best methods to cleanse and transform the dataset. These methods were then assessed through suitable correlation analysis techniques, not by explicitly training a simple machine learning model and evaluating it. Thus, due to the lack of concrete mathematical evidence, the techniques used to create the final dataset will be assessed and chosen through consideration of the logic behind such actions and the overall total correlation the dataset will have to the target variable.

# HR Analytics

## Loading and EDA

As described in the brief overview of “hr\_data.csv”, this dataset consists of employee information, educational background, performance at work, and whether they had been promoted before or not. In fact, the end goal for cleaning this data set would be to train a classification model to predict whether an employee will be promoted or not. So, the target variable in this case would be the column, “is\_promoted” and the other columns would be treated as the dependent variables.

To facilitate smoother exploratory data analysis (EDA), both PowerBI and Jupyter notebook was employed to visualize the dataset. A quick overview of the dataset’s first 5 rows depicts how there are 5 categorical features and 9 features of the numeric datatype. However, not all these 9 features are truly continuously numerical, as some should be considered as categorical or even ordinal due to their values. For example, the columns “KPIs\_met >80%” and “awards\_won?” have only 2 unique values, “0” to signify “No” and “1” as “Yes”. Since the values are binary by nature, the columns are not continuous features. Additionally, the column “previous\_year\_rating” is intrinsically ordinal, where its values have a ranking associated with them and is not technically continuous. Unfortunately, since these features are already numeric, it would be unnecessary to encode these numbers into their respective labels, so they were kept as such and mostly untouched moving forward.

**Figure 1**

*A pair plot constructed using seaborn to visualize and compare the relationship among all the numerical columns.*

A picture containing text, crossword puzzle

Description automatically generated

Key insights derived from Figure 1 would be each numerical feature’s value distribution and how there is a clear correlation between a high average training score of an employee and such an employee being promoted, signified by the orange colored points clustering together on the higher end of the “avg\_training\_score” axis. Additionally, none of these continuous features are normally distributed, which makes logical sense considering the context of such data.

**Figure 2**

*A collection of graphs created using PowerBI to visualize the categorial features, specifically “gender”, “recruitment\_channel”, and “department”*

Chart

Description automatically generated

Graphical user interface

Description automatically generated

Chart, bar chart

Description automatically generated

Interestingly, there seems to be minimal human bias present in the dataset, where the percentages for those who received promotions and for those who didn’t are relatively within the same range, regardless of the employee’s gender, recruitment channel, and assigned department. Additionally, the percentages make logical sense, where among the departments the Legal department, one of the smallest departments in the organization, had the smallest percentage of promoted employees.

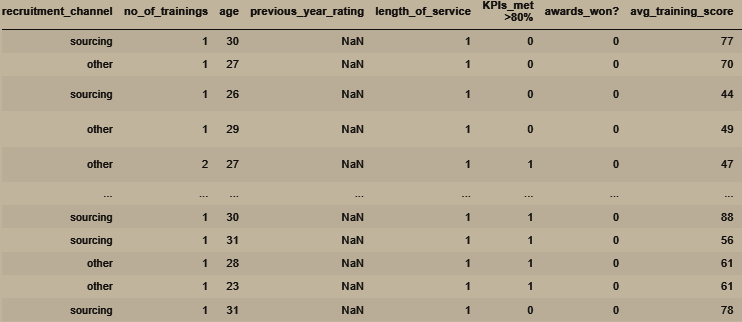
With this, these features can be retained as they do not contain much bias, so the eventual model trained by the final dataset would not regurgitate such bias and remain ethical.

**Figure 3**

*Filtered pandas dataframes to understand the NaN values present in the given dataset.*

**Table

Description automatically generated**

****

Seen in Figure 3, there are 2 features with NaN, aka null or missing, values, which are “education” and “previous\_year\_rating”. So far, there is no clear evidence that the missing values under the “education” column is related to the other columns in the dataset. Thus, it is possible that there is a meaning behind the null values, for example that employee may have not received any educational certification at all, or the employee may have received education elsewhere that is not recognized by the company to be under one of their pre-defined categories. In respect to the ambiguity of the reason behind the missing values in the column “education”, it is best to represent the fact it is missing in the dataset. In the “previous\_year\_rating” column, there would always be a NaN value if the employee has had 1 year of service under the company. So, this meant that the null value is used instead when there is no record of the employee’s rating the previous year. After all, these affected employees had not been hired into the company as of then. Thus, to best represent how such employees did not a rating the previous year, the value of “0” should be used. Especially since how the column had an ordinal range starting from 0 to 5.

## Cleansing and Transformation

During exploration, it was also noted that the dataset contained some NaN values, specifically in the columns of “education” and “previous\_year\_rating”. Based on insights gathered regarding that, the features were imputed with a missing value indictor and a constant value of 0 respectively. Additionally, the column “employee\_id” was dropped at this stage since in the end, it would not offer any value to the dataset or model, since there is no intrinsic meaning in its values by nature.

**Figure 4**

*Code snippets from Jupyter Notebook showcasing how the missing value imputation and column dropping were conducted.*

Graphical user interface, application

Description automatically generated

For the numerical features, specifically “avg\_training\_score” and “age”, no mathematical transformations or clipping were performed. Despite how the features may not follow a normal distribution, their distributions and values make logical sense regarding the dataset’s context and background. In respect to that, those features should be untouched as of now, since applying any transformations would alter the value’s underlying meaning too. The features’ ranges do not pose any anomalies so there is no need to clip or trim them.

For the categorical features, specifically those with a datatype of “object”, they need to be encoded into a numeric value. In this case, the columns affected are “department”, “region”, “education”, “gender”, and “recruitment\_channel”. For “education”, there is a sense in ranking among the categories, where “Master’s & above” are ordinally considered higher than “Bachelor’s”. Thus, a mapping dictionary is created to help encode the column, as seen in Figure 5.

**Figure 5**

*A dictionary created for reference when the values are being mapped during encoding.*

**A picture containing shape

Description automatically generated**

The other features were experimented iteratively on, using different encoding techniques. Such methods include, dropping the features, one hot encoding them, ordinal encoding, and encoding using the mean of the target variable. These will be delved deeper into the following section.

## Correlation Analysis

As the target variable is a categorical feature, the dataset was assessed on the features’ values from its chi-squared Test, its Pearson’s Coefficient, and the mutual information between its dependent variables and its target variable. Unfortunately, both the chi-squared test and mutual information works best when the compared variables are all categorical, so as of now only the categorical features are to be assessed. Even though this issue could be solved by applying the ANOVA test for the numerical features, the target variable is binary, which does not work well with ANOVA tests.

The statistical chi-squared test is meant to test and calculate the independence between categorical variables, like the calculation of mutual information (Brownlee, 2019) (Roepke, 2022). For both measures, the higher the returned value, the higher the feature’s importance is calculated to be and the more correlated it is with the target variable (Brownlee, 2019) (Roepke, 2022). The Pearson’s Correlation Coefficient works best with purely numeric data on the contrary (Roepke, 2022) (Shetye, 2019). When visualized in a heatmap, the blocks with a lighter shade have a higher value, indicating a stronger correlation.

**Table 1**

*Encoding methods compared by the categorical training data’s total feature importance from chi-Square test.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dropping features | One hot encoding | Ordinal encoding | Using mean of target |
| 3118.03710372287 | 3577.406680410906 | 3168.4426577499185 | 3121.2221527109896 |

**Table 2**

*Encoding methods compared by the categorical training data’s total feature importance from calculating mutual information.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dropping features | One hot encoding | Ordinal encoding | Using mean of target |
| 0.05451304600889872 | 0.08393606007321974 | 0.063209194629563 | 0.06634170777936621 |

From both Table 1 and Table 2, it can be surmised that using one hot encoding yields the dataset with the highest total feature importance. Thus, this method would be used

The chi square test recommended using the following columns : 'education', 'previous\_year\_rating', 'KPIs\_met >80%', 'awards\_won?', 'department\_HR', 'department\_Sales & Marketing', 'department\_Technology', 'department\_Procurement', 'department\_Legal', 'region\_region\_7', 'region\_region\_32', 'region\_region\_22', 'region\_region\_17', 'region\_region\_24', 'region\_region\_28', 'region\_region\_25', 'region\_region\_23', 'region\_region\_26', 'region\_region\_31', 'region\_region\_29', 'region\_region\_4', 'region\_region\_9', 'region\_region\_6', 'region\_region\_5'.

The mutual information calculation recommended using the following columns too: 'education', 'previous\_year\_rating', 'KPIs\_met >80%', 'awards\_won?', 'department\_Operations', 'department\_Procurement', 'region\_region\_2', 'region\_region\_22', 'region\_region\_17', 'region\_region\_14', 'region\_region\_8', 'region\_region\_1', 'region\_region\_26', 'region\_region\_31', 'region\_region\_4', 'region\_region\_20', 'region\_region\_27', 'gender\_f', 'recruitment\_channel\_other'.

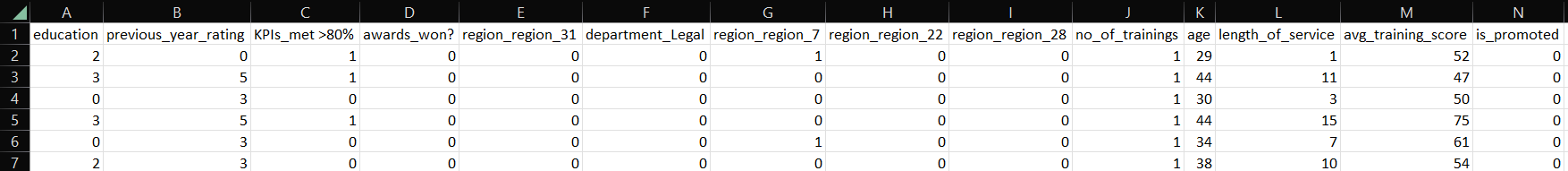
Among all 4 methods used, the Pearson Correlation Coefficient returns the same results, where the same features possess the strongest correlation to the target variable. All above the coefficient of 0.1, the features are “previous\_year\_rating”, “KPIs\_met >80%”, “awards\_won?”, and “avg\_training\_score”. Thus, it is important to retain these features in the final dataset.

## Final Dataset

In the end, the following dataset containing 54808 rows and 14 columns had been exported to a csv file, called “cleaned\_hr.csv”.

**Figure 6**

*A snippet of the new cleaned\_hr file in Excel.*



# Airbnb

## Loading and EDA

The dataset retrieved from “listings.csv” has a defined target variable of “price” that is continuous, and it has a mix of both categorical and numerical features.

**Figure 7**

*The visualizations of some of the categorical features in PowerBI.*

*Graphical user interface

Description automatically generated with medium confidence*

*Chart, box and whisker chart

Description automatically generated*

Interestingly, majority of the listings are located around the central region. Drilling down into which neighborhood shows how Kallang and Geylang is the top 2 towns with the most amount of Airbnb listings available there. In response to the top 2 neighborhoods being in the same central region, it might be more practical to only consider one of the 2 features since they would be highly correlated to each other after all. Majority of the listings also are the entire home or apartment being rented out. Even though there is a clear class imbalance among the classes for “room\_type”

**Figure 8**

*Pair plot visualizing the numerical features’ distribution and relationship with other features.*

*Table

Description automatically generated with medium confidence*

Honestly, there is nothing notable from this visualization. There is no clear new correlation or insight that can be derived from Figure 8. Even though Figure 8 depicts each numerical variable’s distribution to not be normal, it might be unwise to transform and try to rectify such a distribution.

**Figure 9**

*Snippets of dataframes used to investigate the presence of missing values.*

A picture containing graphical user interface

Description automatically generated

It seems that there are 2 features with missing values, that may have caused each other to have a null value. Rows that do not have any average reviews per month also do not have a date stated in their “last\_review”. Since this is highly caused by having no reviews, the listing must have been brand new in the marker, and thus would not be able to calculate the average number of reviews.

## Cleansing and Transformation

At the very start, the following columns are slated for being dropped since they act as an identifier :

"id","name","host\_id","host\_name". Moving to cleansing the dataset, there are 2 features with NaN values, “last\_review” and “review\_per\_month”. As investigated earlier, these two features seem to be best suitable to have their null values imputed with “0”.

For the numerical features, no mathematical transformations or clipping also were performed. Despite how the features may not follow a normal distribution, their distributions and values make logical sense regarding the dataset’s context and background. However, the price range for these listings was rather absurd, ranging from $0 to $10 000. The inter-quartile range was quite puzzling too, ranging from $65 to $199. Considering how majority of the listings are cheap, the model should be trained from that set of listings. Thus, the entire dataset was “trimmed” to only include listings with a price less than or equal to $400.

**Figure 10**

*Snippet of code from Jupyter Notebook to sample a subset from the entire dataset.*

Graphical user interface, text, application

Description automatically generated

Moving onto the categorical features, the concerned columns were "last\_review", "neighbourhood\_group", "neighbourhood", and "room\_type". For the latter three, 4 encoding methods were explored, which would later be assessed in the correlation analysis section.

For the remainder of the categorical features, it is challenging as the feature “last\_review” was of the date data type, which require extra processing. To replace this column, the number of days since the last review was calculated into a new column using the difference between “last\_review” and “10 Nov 2022”, which had been extracted from the file’s metadata to be its creation date.

Lastly, the entire dataset was normalized, or rather scaled using a standard scaler. As the target variable is continuous, it is highly likely that the model trained by the final dataset would be using a Euclidean distance based algorithm. In anticipation of this, the dataset needed to be scaled, regardless. Especially since scaling the training dataset would usually provide better results.

## Correlation Analysis

As the target variable is a continuous feature, the dataset was assessed on the results from calculating Pearson’s Correlation Coefficient, conducting backward elimination, Recursive Feature Elimination (RFE) and Lasso Regularization (LassoCV) (Shetye, 2019). The latter 3 techniques are all meant for feature selection of numerical datasets (Shetye, 2019).

**Table 3**

*Encoding methods compared by whether the categorical features are to be kept after conducting backward elimination.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dropping features | One hot encoding | Ordinal encoding | Using mean of target |
| - | Majority to be dropped | Drop both relating to neighbourhood | Not neighbourhood\_group |

**Table 4**

*Encoding methods compared by whether the categorical features are to be kept after conducting RFE.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dropping features | One hot encoding | Ordinal encoding | Using mean of target |
| - | Yes | Yes | Yes |

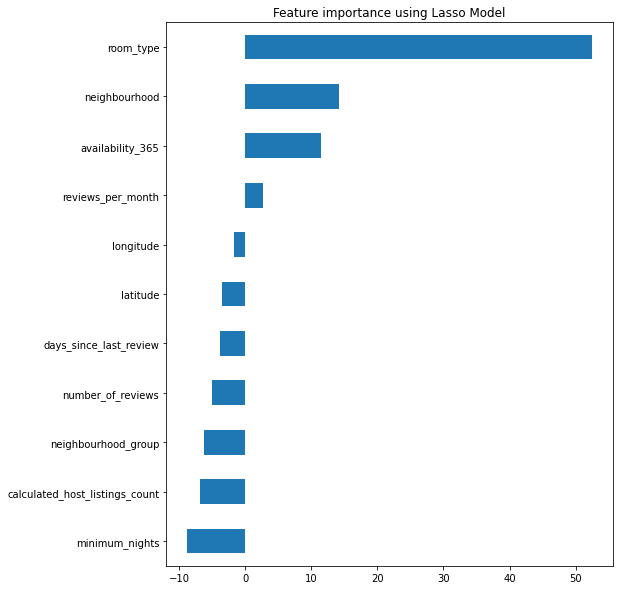
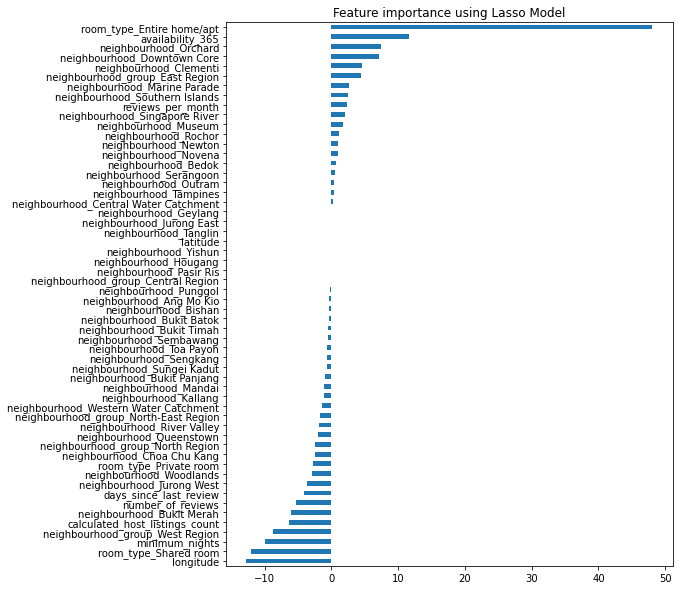
**Table 5**

*Encoding methods compared by the features’ score after conducting LassoCV.*

|  |  |  |  |
| --- | --- | --- | --- |
| Dropping features | One hot encoding | Ordinal encoding | Using mean of target |
| 0.093001 | 0.462652 | 0.176705 | 0.447245 |

**Figure 11**

*Side by side comparison of the feature importance of 2 datasets encoded using one hot encoding and the mean of the target variable respectively.*



Overall, it seemed using the mean of the target variable to encode the categorical features, so as to avoid having too many features for a regression problem.

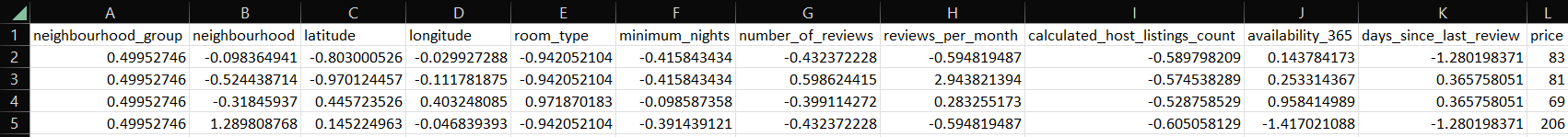
## Final Dataset

After aggregating the results from all 4 analysis techniques conducted, it seems that no column needs to be dropped.

In the end, the following dataset containing 7583 rows and 12 columns had been exported to a csv file, called “cleaned\_listings.csv”.

**Figure 12**

*A snippet of the new cleaned\_listings file in Excel.*



# Summary

With the goal to train a classification and regression model, the report explored, understood and transformed the given datasets fit for training. Both datasets were visualised in both Jupyter Notebook and PowerBI. For the dataset on HR information, after understanding the nature of each feature, missing values were imputed with a “missing” indicator and a “0”. After trial and error of 4 encoding methods, one hot encoding was assessed to be the best option. Lastly, after conducting chi-square test and mutual information, a subset of the dummy variables were selected, and then exported as “cleaned\_hr.csv” at the end of that section in the accompanying Jupyter notebook file. For the dataset on Airbnb listings, after seeing and understanding the direct causation of the “number\_of\_reviews” column on the columns with missing values, the null values were imputed with a simple “0” . The numerical variables were left untouched, except for only selecting a sample of the dataset to be used, specifically the listings where the price range was between $0 and $400. Similar to how they were dealt with for the earlier dataset, the categorical features were experimented on until deciding to encode them using the mean of the target variable. The sample also had a date column, which was replaced by a column calculating the number of days since the last review. After being normalised, the sample was evaluated using 4 feature selection techniques, Pearson’s correlation coefficient, backward elimination, RFE, and LassoCV. In the end, all the features were retained and the final sample was exported as “cleaned\_listings.csv”.

## Further improvements

For the dataset on HR information, the following avenues could be ventured in the future:

* Imputing missing values based on the most frequent educational level the department the employee is in
* Considering other encoding methods for categorical features
* Applying mathematical transformations onto the numerical features
* Creating new features based on the given features
* Explore fixing the class imbalances present in the dataset
* Explore using Spearman’s rank coefficient to assess the numerical variable’s correlation to the target variable

For the dataset on Airbnb listings, the following avenues could be ventured in the future:

* Considering other encoding methods for categorical features
* Applying mathematical transformations onto the numerical features
* Creating new features based on the given features
* Explore fixing the class imbalances present in the dataset
* Explore only selecting subset of data for certain regions

# References

Brownlee, J. (2019a, October 16). *Https://Machinelearningmastery.Com/Information-Gain-and-Mutual-Information*. <https://machinelearningmastery.com/information-gain-and-mutual-information>

Brownlee, J. (2019b, November 25). *How to Perform Feature Selection with Categorical Data*. <https://machinelearningmastery.com/feature-selection-with-categorical-data/>

Roepke, B. (2022, April 18). *4 methods that power feature selection in a machine learning model*. Data Knows All. <https://www.dataknowsall.com/featureselection.html>

Shetye, A. (2019, February 12). *Feature Selection with sklearn and Pandas*. Medium. <https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b>