**Basic Or Luxury Paris**

Matthew Carrico, [mcarrico2@bellarmine.edu](mailto:mcarrico2@bellarmine.edu)

**ABSTRACT**

The goal of this project was the use Python to analysis a dataset from Kaggle and Appling a machine learning Classification model of our choice. I chose a dataset about Paris Housing, which contains information like the location, size, number of rooms, etc. My main objective for this project was to predict whether a housing property is Luxury or Basic based on the features. To find this out I preformed exploratory analysis of the dataset, data preparation, and doing experiments with different parameters for the Logistic Regression model. The results from the experiments are presented in the form of confusion matrices and classification reports.

1. **INTRODUCTION**

This project is about an analysis of the Paris Housing dataset, which has information about different housing properties in Paris. The primary objective of this project is to predict whether a housing property is Luxury or Basic using the variables from the dataset.

1. **BACKGROUND**
   1. *Data Set Description*

The dataset used in the project can be found here <https://www.kaggle.com/datasets/mssmartypants/paris-housing-classification>. It is a dataset made up of imaginary data of housing prices in a fictional Paris. It has 10000 entries and 18 columns.

* 1. *Machine Learning Model*

The Logistic Regression model is used for binary classification problems. It estimates the probability of a binary outcome based on the inputs. The model uses a logistic function to map the input to a probability value between 0 and 1. The closer to 1 the more accurate the model is.

1. **EXPLORATORY ANALYSIS**

This dataset has 10000 samples with 8 columns of integer, float, and objective data types. No variables had missing values.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| squareMeters | Int |
| numberOfRooms | Int |
| hasYard | Int |
| hasPool | Int |
| floors | Int |
| cityCode | Int |
| cityPartRange | Int |
| numPrevOwners | Int |
| Made | Int |
| isNewBuilt | Int |
| hasStormProctector | Int |
| basement | Int |
| attic | Int |
| garage | Int |
| hasStorageRoom | Int |
| hasGuestRoom | Int |
| price | Float |
| catergory | Object |

1. **METHODS**
   1. *Data Preparation*

Luckily this dataset was complete and I didn’t have to mess with anything.

* 1. *Experimental Design*

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | 60/40 split for train and test |
| 2 | 80/20 split for train and test |
| 3 | 90/10 split for train and test |

* 1. *Tools Used*

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1.

I chose these tools because our time using them in class makes them feel like a comfortable environment.

1. **RESULTS**
   1. *Classification Measures*

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

**Figure 1: (a) Confusion Matrix for Experiment 1 (b) Classification Report for Experiment 1**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

**Figure 2: (a) Confusion Matrix for Experiment 2 (b) Classification Report for Experiment 2**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

**Figure 3: (a) Confusion Matrix for Experiment 3 (b) Classification Report for Experiment 3**

* 1. *Discussion of Results*

The model was good for predicting if a property was basic or luxury. With report one having an accuracy of 88% and reports two and three having an accuracy of 87%. There were also no false negatives or positives.

* 1. *Problems Encountered*

The main problem was finding a dataset that had the right amount of data in the dataset.

* 1. *Limitations of Implementation*

One of the limitation of the model is that it is only binary for this dataset having results that are non binary would be beneficial.

* 1. *Improvements/Future Work*

Using some different models with this dataset would be a cool experiment. Like using the Random Forest Classifier and Gradient Boosting Classifier algorithms.

1. **CONCLUSION**

Overall the Logistic Regression model is a good fit for this dataset and a good predictor of if the property is basic or luxury based off the variables from the dataset. Experiment one provided the most accurate result. Even though our model had a good fit our results could be improved by using other models.

**REFERENCES**

<https://www.kaggle.com/datasets/mssmartypants/paris-housing-classification>