Predicting House Rental Prices using Machine Learning Model

Abstract

Machine Learning is a power tool that uses large amounts of data to train a mathematical model to predict a certain outcome. This technique was used to predict the price of housing rental using data such as floor size, number of bathrooms/bedrooms, type of housing, leasing agency and occupation availability month. It was shown that using a XGBoost ensemble method produced the best results with mean squared error, MSE, of R7129.

Keywords: Machine learning, ensemble, rental price, MSE

1. Introduction

A dataset containing the following columns, was provided:

- Id: unique identifier for each property
- Property_description: a text description for the property
- Area_jhb: area of johannesburg where property was located
- Listing_agency: agency through which the property is listed and rented
- Rental_amount: the price in Rands to rent the property
- Street_address: the street address of the property
- Occupation_date: the date on which the property will be available for occupation
- Bedrooms_1: ?
- Bedrooms_2: Number of bedrooms on/in property
- Bathrooms: number of bathrooms on/in property
- Area: floor size in m^2
- Feature_1: First additional feature of property
- Feature_2: Second additional feature of property

The goal is to predict the Rental_amount using the above columns of data using machine learning.

A combination of novel python code and an AutoML solution (Botha, 2020) will be used to perform all the steps of the machine learning process. All of the code used for this analysis can be found on github.

2. Data Structure/Preparation/Cleaning/Manipulation

2.1 Data Structure/Type

The data was first opened using excel/sheets to interrogate the format of the data. Data types can be inferred simply by looking at the data and it was determined that it had the following structure:

Table 1: Column data types and Examples

Column	Data Type	Example Data
id	Unique Identifier	20193121
property_description	Text	R 5 500 1 Bedroom Apartment / Flat to Rent in Emmarentia 1 bedroom apartment on the second floor. Separate lounge and dinning room, near busy shopping centre. Close ... Available: 16 October 2018 1 | 1 Floor Size: 104 mB2
area_jhb	Text/Categorical	1 Bedroom Apartment / flat to rent in Emmarentia - Johannesburg
listing_agency	Text/Categorical	Property.CoZa - Randburg
rental_amount	Numeric/Currency	R 5 500
street_address	Text	Villa Nova, Ferreira Street, Turffontein
occupation_date	Date/Categorical	Available: 16 October 2018
Bedrooms_1	Text	Beds
Bedrooms_2	Numeric	1
bathrooms	Numeric	1
area	Numeric	104 mB2

Feature_1	Text	1 Bedroom Apartment / Flat to Rent in Emmarentia
Feature_2	Text	118 mB2

2.2 Data Cleaning/Manipulation/Missing, erroneous Data

The data provided was very inconsistent and needed to be cleaned/transformed. Many values were missing and needed to be fixed or imputed.

In cases where the column value was found in the property description, see Table 1, or the columns were shifted, the values were extracted using various pythonic techniques and pandas.The following columns were left after extraction: id, Extracted Bedrooms, Extracted area, Extracted Bathrooms, Extracted Area_JHB, Extracted listing agency, Extracted Type, Extracted Available Month.

Where data was not present in the correct column and was not in the description, numeric data types were replaced with the mean value of the column and text/categorical columns were replaced with None.

2.3 Summary Statistics

The summary statistics of the numeric values can be seen below:

Table 2: Summary Statistics for Numeric Columns

Statistic	Extracted Price	Extracted Bedrooms	Extracted area	Extracted Bathrooms
count	1916	1352	1.90E+03	1519
mean	11449.51305	1.846524	4.99E+03	0.94865
std	29196.65824	0.763191	5.09E+04	0.862975

min	1300	0.5	0.00E+00	0
25%	5500	1	0.00E+00	0
50%	7500	2	1.00E+00	1
75%	13000	2	8.43E+01	1.5
max	980000	5	1.64E+06	4.5

Table 3: Missing value statistics for data columns

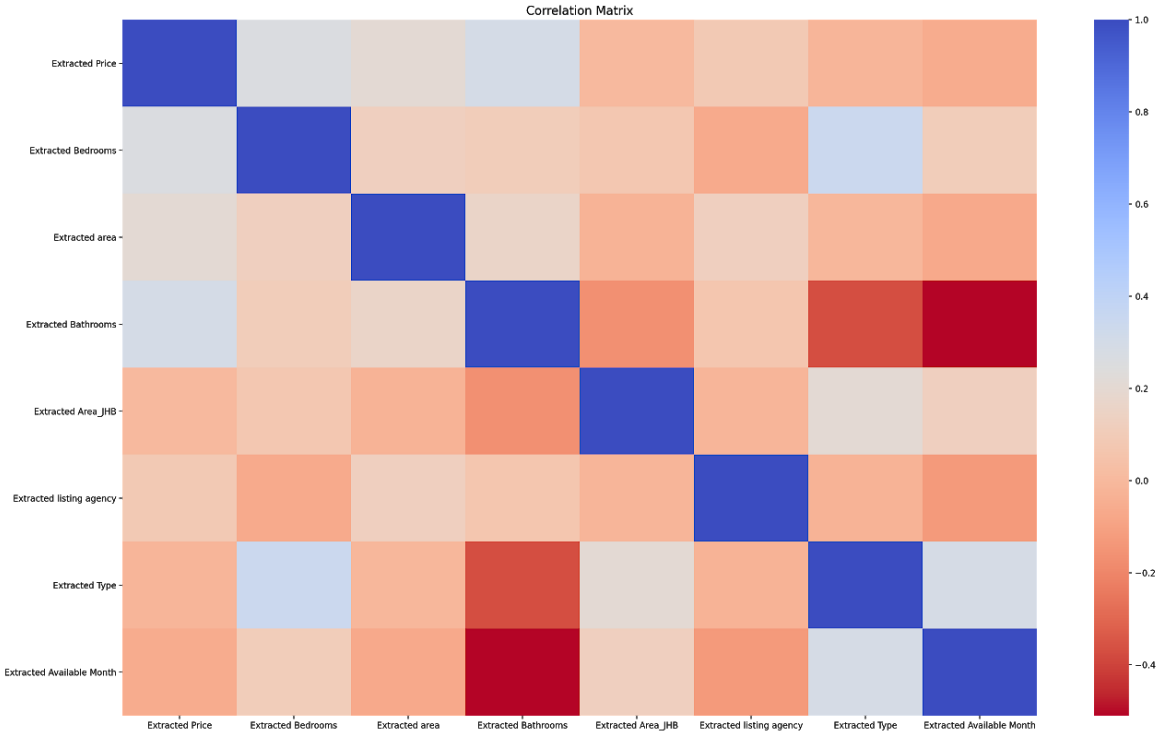
Column	Missing Values
Extracted Price	0
Extracted Bedrooms	564
Extracted area	20
Extracted Bathrooms	397
Extracted Area_JHB	0
Extracted listing agency	0
Extracted Type	0
xtracted Available Month	0

3. Data Visualisation

3.1 Correlation

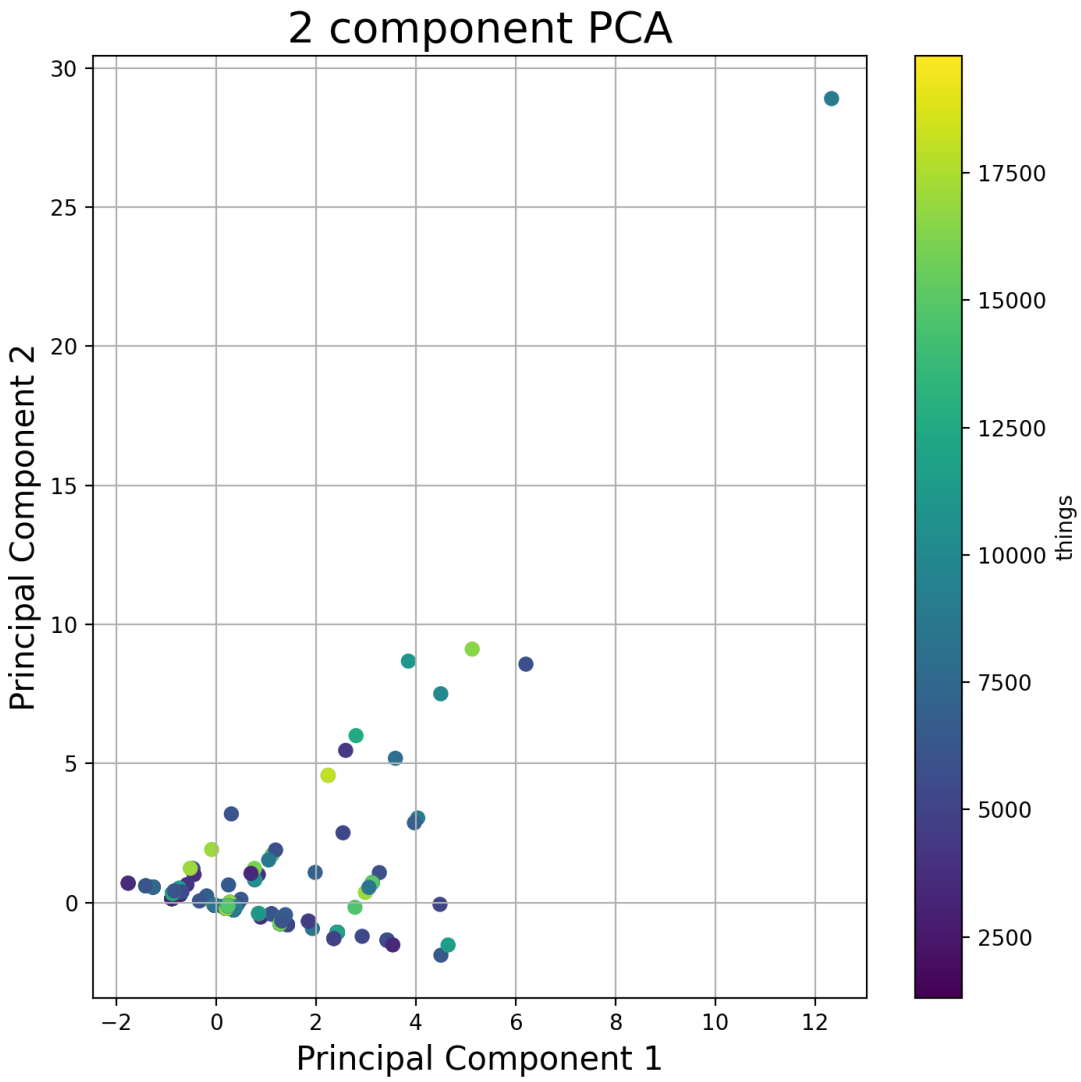
The data was correlated using the kendall method which can take into account categorical variables. The correlation plot is shown below:

Image 1: Correlation plot



3.2 Principal Component Analysis

A principal component decomposition of all the numeric values down to 2 dimensions was done.



If the variables showed high amounts of correlation it would show up as a straight line in the PCA. As we can see there are very few linear parts of the above graph, with the exception of the lower diagonal. Using the line along with the color mapping which shows the rental_amounts < 20000, we can see that rental_amounts at the lower end of the price spectrum do share some sort

of correlation. Due to the high dimensionality of the data it is difficult to ascertain which original features display this high correlation.

3.3 Histogram/Heatmap analysis of column correlation

Due to the number and size of the plots, please see the attached jupyter notebook for the heatmap analysis.

4. Feature Engineering and Feature Selection

4.1 Variable Importance

Automatic feature selection was done using SK-Learn's GenericUnivariateSelect using the f_regression regression test. It determined that only the Extracted Bedrooms variable was worth keeping. Using this, the mean absolute error for each model was significantly worse than including all of the variables.

4.2 High correlated variable combination

As seen in the correlation plot, Image 1, there are no highly correlated variables except perhaps the high negative correlation between the available month and the number of bathrooms, This seems dubious at best. It was decided not to combine any variables.

4.3 Feature Selection

All of the features shown in Image 1 were chosen as it performed the best.

5. Model Creation, Cross Validation and Metrics

5.1 Model Selection

The AutoML solution enables the user to specify multiple types of models to train and test. As this was a regression problem, DecisionTreeRegressor, RandomForestRegressor and XGBoostRegressor were chosen as they cover both linear/semi-linear and non-linear models.

5.2 Over-Underfitting

To avoid over or underfitting, cross validation was used during the training.

Looking at the metrics, discussed in 5.4 it can be assumed that some under fitting occurred due to the limited amount of training data.

5.3 Model ensemble

Two of the models mentioned in 5.1 are ensemble methods and make use of a linear or non-linear combination of smaller non-ensemble models.

5.4 Metrics

For the cross-validation training step, negative_mean_squared_error was chosen as it is suitable for regression models and is the preferred for cross validation and gridsearch.

For the testing/validation step, neg_mean_absolute_error was chosen as it is easier to interpret as the units are the same as the prediction value, in this case Rands.

5.5 Cross Validation/Hyperparameter

For hyperparameter optimisation GridSearchCV was used to test various combinations of hyperparameters for each model. The use of this technique and the benefit of using Cross Validation to determine the best combination of hyperparameters to prevent over fitting.

6. Prescription and documentation

6.1 Approach to data problem

With most data problems it is advised to spend the most time on data cleaning/preparation, this was no different.

Due to the very bad data quality most of the time was spent on extracting/cleaning the data to make it suitable for use with Machine Learning.

Understanding the data is also quite important and the above mentioned extraction part showed great insight when combined with Exploratory Data Analysis. It was determined that there are no solid correlations between the features and that selecting one above the other's could lead to a worse result.

Due to the AutoML solution that was used much time was saved not worrying about the actual model, over/under fitting and hyperparameter optimisation.

6.2 What insight can be derived from the data, what story does it tell

The data shows a high amount of variability, or low correlation between the features and the target and between features themselves. This leads on to think that method of simple estimation of the rental_amount of a property is not present within the dataset.

The data set also seems quite incomplete with many missing values and low data quality. Maybe an increase in the data quality of completeness would lead to first, and easier processing of the data, reducing the need for extraction as discussed in section 2 and second a more complete picture of the situation.

6.3 What other features should have been collected or engineered for a better prediction

A breakdown of income per capita of the jhb regions could show the level of affluence which would correlate with the price of housing and by extension, the rental_amount.

Geospatial data such as postal codes and relative distance to places of work or places of entertainment/shopping could also show some influence in the rental amount, the idea being individuals would “pay for convenience”