Predicting London House Prices

Kenny Hunt

1. Introduction

1.1 Background

The London property market is leading the way for regeneration, many property developers choose to invest in London because the stable growth of house prices will pretty much guarantee a solid return on investment. Investors and developers adding capital to areas of London has enabled the regeneration of many London boroughs, this has in turn made many areas in the suburbs of London more desirable to live, which attracts more businesses to trade. This cycle has lead to the constant growth of the London property market. For developers, residential buyers and property investors, London continues to be the flagship property investment hotspot in the UK.

There are many variables that can impact the price of a property in London; size, number of bedrooms, location, proximity to travel links, proximity to other amenities e.g. supermarkets, restaurants, bars.

Price varies wildly between different boroughs and areas of London, for this reason a single borough will be selected for the analysis, although the experiment will be applicable to all areas. The selected area for this experiment is the London Borough of Ealing. Ealing is generally recognised as a sought after place to own property, this is largely down to the significant amount of regeneration that has been taking place over the last decade throughout the entire borough, coupled with its superb transport links into the city.

1.2 Problem

The project will identify and analyse the impact of different variables - both structural and environmental - and the effects they have on the price of a house in London Borough of Ealing. Once the key variables are identified, the project aims to utilise this analysis to predict the sales price of future property sales based on both current and historical data.

1.3 Interest

Understanding the key drivers behind property value in a given area presents many potential benefits to property developers, buyers and investors and alike.

2. Data Acquisition and Cleaning

2.1 Data Sources

The project utilises the following data sources:

Name	Description	Extract Method
Zoopla	• Current properties on sale in the selected area. • Historical sales prices of properties sold in the selected area.	Web Scraper
Foursquare	• Data surrounding the amenities in the local area	API
Transport for London	• Locations of Transport Links	Web Scraper
Zoopla	• Longitude and Latitude of London postcodes	API

2.2 Data Preparation

Data scraped from multiple sources was combined into one table for ease of analysis. There was a limitation on the number of records that could be exported through the zoopla API at any one time, therefore the decision was taken to break the entire borough down into 7 separate postcodes and call the api separately for exact data set, then merge them together into one table.

The data set required steps to cleanse. Firstly there were a number of data quality errors;

2.2.1 Number of Bedrooms

Where the property for sale was a piece of land, this was regarded as a row with zero bedrooms. As it was believed that the number of bedrooms was influential on the total price of the property, it is reasonable to assume that any calculations would be skewed by these figures - therefore any row with zero bedrooms was dropped.

There were a number of properties containing more than 4 bedrooms which were available at very low prices - analysis evidenced that these properties were “shared equity” therefore were filtered from the dataset at source.

2.2.2 Nulls

The data set contained nulls in the post_town column only, these records were left alone as post town was consistent throughout the data set, therefore this feature will be dropped later on.

	%	Total # Records
post_town	3.0	21
street_name	0.0	0
property_type	0.0	0
price	0.0	0
outcode	0.0	0

Fig1: Shows the number and percentage of nulls in the dataset

2.2.3 Price

There were a number of properties with the price set as zero, these are effectively “price upon enquiry” properties, and were therefore removed from the analysis.

2.3 Feature Selection

After cleansing there were 653 records and 28 columns in the data. Many of the columns were deemed irrelevant for our analysis and will require dropping.

Feature	Action	Justification
agent_address	Drop	Irrelevant
outcode	Keep	
num_recepts	Keep	
first_published_date	Drop	Irrelevant
displayable_address	Keep	
details_url	Drop	URL Irrelevant
num_bedrooms	Keep	
price	Keep	Target Variable
post_town	Drop	Identical throughout analysis
price_modifier	Drop	Irrelevant
property_type	Keep	
street_name	Keep	
floor_plan	Drop	Irrelevant
image_url	Drop	URL Irrelevant
last_published_date	Drop	Date Irrelevant

latitude	Keep	
listing_id	Drop	Primary key not needed
listing_status	Drop	Filtered to ‘Sale’ at source
longitude	Keep	
num_bathrooms	Keep	
country_code	Drop	Identical throughout analysis
agent_name	Drop	Irrelevant
agent_logo	Drop	Irrelevant
agent_phone	Drop	Irrelevant
category	Keep	
county	Drop	Identical throughout analysis
country	Drop	Identical throughout analysis