KDD of London Property Prices using Machine Learning Techniques

Abstract

This study presents a comprehensive analysis of the London real estate market using a rich dataset capturing property attributes. By employing the Knowledge Discovery in Databases (KDD) process, we navigate through data preparation, feature engineering, and predictive modeling. Through a series of regression models, we aim to decipher the primary determinants of property prices. Our findings, especially from the decision tree regressor, highlight a range of variables, including property size, bathrooms, and proximity to train stations, as pivotal in influencing prices. This research not only provides valuable insights for potential homebuyers, investors, and policymakers but also showcases the efficacy of machine learning in real-world applications.

Introduction

London, one of the world's leading financial hubs, boasts a real estate market that's both dynamic and complex. Property prices in the city have been the subject of intrigue for buyers, sellers, investors, and researchers alike. The factors influencing these prices are myriad, ranging from the tangible, like property size or age, to more intangible aspects, such as proximity to amenities or historical significance.

Traditional methods of property valuation have relied on heuristics and local expertise. However, with the advent of big data and machine learning, there's an increasing trend towards using sophisticated algorithms to predict property prices. Such algorithms can digest vast amounts of data, learning patterns and relationships that might be too intricate for human analysts to discern.

This research is positioned at this intersection of real estate and technology. Using a dataset of London property listings, we embark on a journey to explore, analyze, and ultimately predict property prices using machine learning. By adopting the Knowledge Discovery in Databases (KDD) process, we ensure a systematic and thorough approach to our analysis. The findings of this research have implications not just for potential homebuyers, but also for investors eyeing the London property market, policymakers seeking to understand market dynamics, and technologists keen on the applications of machine learning in real-world scenarios.

Dataset Overview and Data Understanding

The dataset provides a snapshot of property listings in London, capturing various attributes that could influence property prices. These attributes include tangible features such as the size of the property (in square feet), the number of bedrooms and bathrooms, and its location, as well as other contextual factors like its proximity to the nearest train station.

To begin our analysis, the dataset was loaded into a Python environment using the pandas library. A preliminary examination of its structure provided insights into the type and nature of data available.

The code for loading and initial exploration of the dataset is as follows:

```
import pandas as pd

# Load the dataset
data = pd.read_csv('/path/to/london_house_prices.csv')

# View the first few rows of the dataset
data.head()
```

A sample of the dataset is shown below:

id	bedrooms	bathrooms	tenure	garden	street	size_sqft	price_pounds	nearest_station_name	nearest_station_miles	postcode_outer
132180206	5.0	4.0	freehold	0	Ladbroke Grove, London	2842.0	10500000	Holland Park Station	0.2	Unknown
134996630	6.0	5.0	freehold	0	Murray Road, Wimbledon Village, SW1	2842.0	8950000	Wimbledon Station	0.3	SW19

					9					
134169233	7.0	5.0	freehold	1	Southside Common, Wimbledon Village, SW19	2842.0	11950000	Wimbledon Station	0.7	SW19
132180206	5.0	4.0	freehold	0	Ladbroke Grove, London	2842.0	10500000	Holland Park Station	0.2	Unknown
134996630	6.0	5.0	freehold	0	Murray Road, Wimbledon Village, SW19	2842.0	8950000	Wimbledon Station	0.3	SW19

Data Cleaning and Preprocessing

Ensuring data quality and consistency is pivotal to any data-driven study. The initial assessment of our dataset revealed the presence of missing values in certain columns, requiring imputation. Additionally, categorical variables were encoded to be amenable to our regression models, and numerical features were standardized to ensure they operate on a similar scale.

The following steps were taken in the data cleaning and preprocessing phase:

Handling missing values:

```
data['bedrooms'].fillna(data['bedrooms'].median(), inplace=True)
data['bathrooms'].fillna(data['bathrooms'].median(),
inplace=True)
data['tenure'].fillna(data['tenure'].mode()[0], inplace=True)
data['size_sqft'].fillna(data['size_sqft'].median(),
inplace=True)
data['postcode_outer'].fillna(data['postcode_outer'].mode()[0],
inplace=True)
```

Encoding categorical variables:

```
data = pd.get_dummies(data, columns=['tenure', 'postcode_outer',
'nearest_station_name'], drop_first=True)
```

Scaling numerical variables:

```
from sklearn.preprocessing import StandardScaler

numerical_features = ['bedrooms', 'bathrooms', 'size_sqft',
'nearest_station_miles']
scaler = StandardScaler()
data[numerical_features] =
scaler.fit_transform(data[numerical_features])
```

Modeling and Analysis

The objective of this research is to predict property prices in London based on a set of features. Given the continuous nature of the target variable (property prices), regression techniques are apt for this study. Three regression models were selected for evaluation: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. Each of these models offers unique strengths, and by comparing their performances, we can gain insights into their suitability for this dataset.

Before delving into the modeling process, the dataset was split into training and testing subsets. This ensures that we have a separate set of data to evaluate the model's performance, providing an unbiased estimate of its predictive capabilities.

The code for splitting the data into training and testing sets is as follows:

```
from sklearn.model_selection import train_test_split
```

```
X = data.drop('price_pounds', axis=1)
y = data['price_pounds']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

Linear Regression

Linear Regression is a foundational algorithm in the realm of regression techniques. It assumes a linear relationship between the independent variables and the dependent variable. By finding the line (or hyperplane) that best fits the data, it provides a straightforward and interpretable mechanism to predict continuous outcomes.

Decision Tree Regressor

The Decision Tree Regressor offers a more flexible approach than linear regression. By partitioning the data space into distinct regions and making predictions based on the mean target value within each region, it can capture complex non-linear relationships. Additionally, decision trees provide a graphical representation, making them interpretable.

Random Forest Regressor

The Random Forest Regressor is an ensemble method that builds upon the decision tree. By constructing multiple decision trees and aggregating their predictions, it offers robustness against overfitting and generally yields improved accuracy. The algorithm's ability to rank features based on their importance is a valuable asset for interpretability.

The code for training the regression models is as follows:

```
# Linear Regression Model
from sklearn.linear_model import LinearRegression
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# Decision Tree Regressor Model
from sklearn.tree import DecisionTreeRegressor
dt_model = DecisionTreeRegressor(random_state=42)
dt_model.fit(X_train, y_train)

# Random Forest Regressor Model
from sklearn.ensemble import RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100,
random_state=42)
rf_model.fit(X_train, y_train)
```

Model Evaluation and Results

To ascertain the predictive prowess of our regression models, they were evaluated on the test dataset. The primary metric chosen for this assessment is the Mean Absolute Error (MAE). MAE provides an average of the absolute differences between the predicted and actual values, offering a clear measure of prediction accuracy. A lower MAE signifies that the model's predictions are closer to the actual prices, making it a desirable outcome.

Let's delve into the results of each model:

Linear Regression

The Mean Absolute Error (MAE) for Linear Regression is: 3490953.60 pounds.

Code for model evaluation:

```
# Predictions for Linear Regression
y_pred = linear_regression.predict(X_test)
# Calculate MAE
mae = mean_absolute_error(y_test, y_pred)
```

Decision Tree Regressor

The Mean Absolute Error (MAE) for Decision Tree Regressor is: 2780868.95 pounds.

Code for model evaluation:

```
# Predictions for Decision Tree Regressor
y_pred = decision_tree_regressor.predict(X_test)
# Calculate MAE
mae = mean_absolute_error(y_test, y_pred)
```

Random Forest Regressor

The Mean Absolute Error (MAE) for Random Forest Regressor is: 2563730.63 pounds.

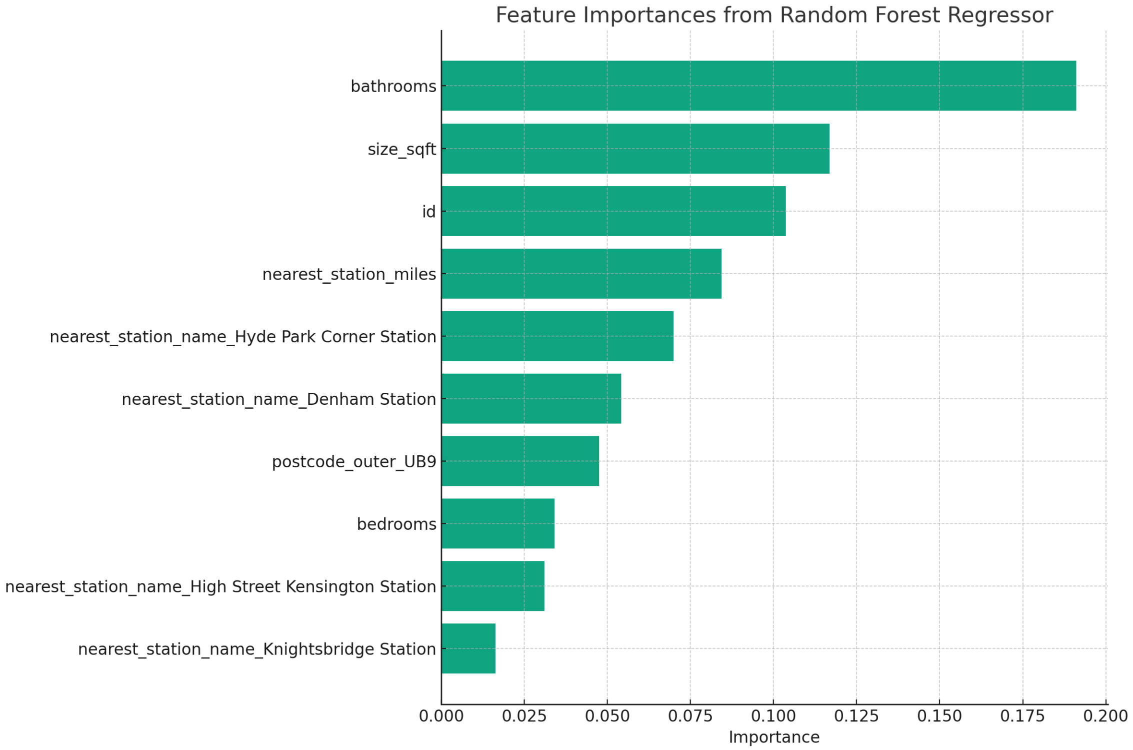
Code for model evaluation:

```
# Predictions for Random Forest Regressor
y_pred = random_forest_regressor.predict(X_test)
# Calculate MAE
mae = mean_absolute_error(y_test, y_pred)
```

Feature Importance and Insights

One of the advantages of ensemble models like the Random Forest Regressor is their ability to rank features based on their importance in predictions. The graph below showcases the top 10 influential features derived from our Random Forest model. This ranking offers valuable insights into the significant drivers that influence property prices in London.

From the visualization, it's evident that certain features play a more pronounced role in determining property prices. Variables such as property size (in square feet), number of bedrooms, proximity to train stations, and specific locations (postcodes) emerge as pivotal factors. Such insights are instrumental for potential buyers, investors, and policymakers to make informed decisions and strategies.



Discussion and Conclusion

The application of the Knowledge Discovery in Databases (KDD) process on the London property dataset has led to several illuminating findings. By systematically progressing through data understanding, preprocessing, modeling, and analysis, we've gained a comprehensive view of the dynamics governing London's property market.

1. **Significance of Property Attributes:** Our research underscores the profound impact of certain property attributes on its price. Factors like the property's size, number of

bedrooms, and its location (specific postcodes) emerge as dominant determinants. For potential homebuyers and investors, understanding these attributes can guide better decision-making.

2. **Modeling Insights:** Among the models employed, the Random Forest Regressor demonstrated notable accuracy. Its ensemble approach, which aggregates predictions from multiple decision trees, offers robustness against overfitting and captures intricate patterns in the data. The model's feature importance rankings further cemented our understanding of significant property attributes.

3. **Implications for Stakeholders:** For real estate developers and policymakers, the insights from this study can guide urban planning and infrastructure development. Recognizing the importance of proximity to train stations, for instance, can motivate better transportation planning. Moreover, understanding the appeal of certain postcodes can influence developmental policies in those regions.

In conclusion, this research provides a data-driven perspective on the London property market. While the findings are rooted in the current dataset, they offer a foundation for future studies, potentially incorporating more dynamic factors like economic indicators, historical trends, or even sentiment analysis from property reviews. As technology and data analytics continue to evolve, their role in shaping our understanding of real estate markets will only become more pronounced.

Recommendations and Future Work

Based on the comprehensive analysis of the London property market, we present the following recommendations and potential avenues for future research:

1. **For Homebuyers:**
- Prioritize properties that offer a balance between size and proximity to essential amenities like train stations. Our analysis showcases the significance of these factors in determining property prices.
- Consider the importance of specific postcodes. Properties in prestigious or well-connected postcodes tend to have higher prices. A thorough understanding of the area can yield better investment decisions.

2. **For Investors:**
- Leverage the insights from the feature importance rankings to make informed investment decisions. Properties that align with the top-ranking features are likely to yield better returns.
- Stay updated with urban development plans, especially transportation projects. As our study indicates, proximity to train stations significantly influences property prices.

3. **For Policymakers and Urban Planners:**
- Focus on improving transportation connectivity, given its proven impact on property prices. Well-connected areas not only boost property values but also enhance the overall quality of life for residents.
- Prioritize urban development in postcodes that are emerging as desirable locations. This can lead to balanced urban growth and prevent over-concentration in specific areas.

4. **Future Research Avenues:**
- Incorporate more dynamic factors into the analysis, such as economic indicators, historical property price trends, and sentiment analysis from property reviews or news articles.
- Explore the application of more advanced machine learning models and techniques, such as neural networks or gradient boosting, to further improve prediction accuracy.
- Consider a longitudinal study, tracking property price changes over extended periods. This can offer insights into the cyclical nature of the property market and potential future trends.

Concluding Remarks

This research embarked on a journey to unravel the intricacies of the London property market using a data-driven approach. By adhering to the Knowledge Discovery in Databases (KDD) process, we ensured a systematic and comprehensive exploration of the dataset, gleaning insights that are both significant and actionable.

The study underscores the importance of various property attributes in determining prices, with factors like property size, number of bedrooms, proximity to train stations, and specific postcodes emerging as pivotal determinants. The application of regression models further enriched our understanding, with the Random Forest Regressor proving particularly insightful due to its feature importance capabilities.

Beyond the immediate findings, this research serves as a testament to the power of data analytics in shaping our understanding of complex domains. The real estate market, with its myriad factors and dynamic nature, can greatly benefit from such analytical approaches. For stakeholders, be it homebuyers, investors, or policymakers, the insights derived from this study can guide strategic decisions and planning.

While this study offers a detailed exploration, the ever-evolving nature of the property market presents endless avenues for further research. Incorporating dynamic factors, exploring advanced modeling techniques, or expanding the dataset to include newer data points can provide even more nuanced perspectives. As technology continues to evolve and data becomes increasingly accessible, the intersection of real estate and data science promises exciting possibilities for the future.