Task 1:

House Price Prediction Model Report

Prepared for: XYZ Limited

Prepared by: Ajakaiye, Oluwadamilola Oreofe

Date: 16th October, 2024

Executive Summary

This report details a machine learning model designed to predict house prices using features from a King County, USA, house sales dataset. Our goal was to develop a linear regression model to accurately forecast house prices, providing insights valuable to real estate stakeholders.

Introduction

In real estate, understanding price determinants is essential for buyers, sellers, and investors. Using a dataset with 18 attributes, including bedrooms, bathrooms, living area square footage, and waterfront availability, this project aimed to forecast house prices.

Methodology

1. Data Preparation

We began by importing the dataset, examining key statistics, and evaluating the distribution and correlations. Linear regression was chosen for its interpretability and efficacy in regression.

2. Feature Selection

Key attributes selected were bedrooms, sqft_living, bathrooms, grade, waterfront, yr_built, view, and latitude.

3. Data Splitting

The data was split into 70% training and 30% testing sets, with a random state of 75 for consistency.

4. Model Training

We trained a linear regression model on the training dataset.

Results and Evaluation

The model produced the following coefficients, indicating how each feature impacts price:

- Bedrooms: -£32,539.25
- Bathrooms: £41,567.47
- SqFt Living Area: £176.80
- Waterfront: £622,085.35
- View: £49,898.23
- House Grade: £110,207.63
- Year Built: -£2,937.78

- Latitude: £555,312.72

Key insights include:

1. **Waterfront premium**: Properties with waterfront views command a significant premium, increasing values by approximately £622,085.

2. **Bedrooms alone may reduce value:** Additional bedrooms correlate with a price decrease of £32,539, possibly indicating that without other high-value features, added bedrooms alone don't drive higher property values.

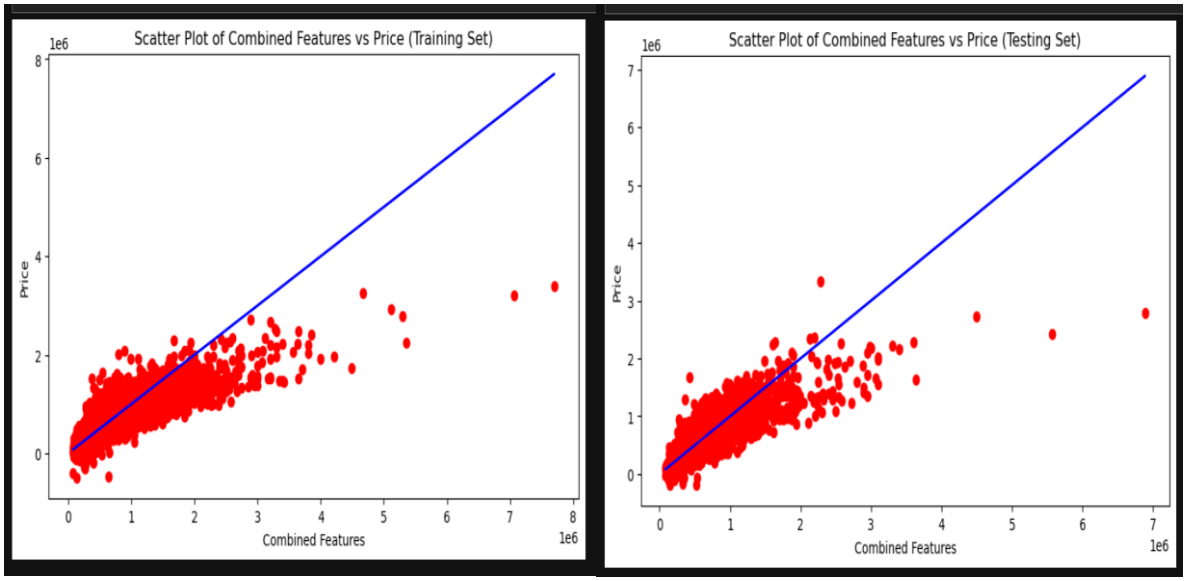
3. **Newer homes are more valuable**: Older homes tend to decrease in value, with prices dropping by about £2,937 for each additional year.

Model Evaluation

The model was assessed using Mean Squared Error (MSE) of 562,636,675.48 and an R^2 score of 70%. While these metrics indicate fair predictive power, there is potential for improvement.

Visualisations

Scatter plots were used to show correlations between selected features and prices, assisting in identifying trends and highlighting the model's predictions.



Recommendations for Improvement

1. Feature Engineering: Add relevant features or interaction terms to refine predictions.

2. Model Tuning: Test alternative regression models, such as Lasso or Ridge, to improve performance and manage multicollinearity.

3. Data Quality: Further cleaning, such as outlier management, may improve prediction accuracy.

4. Visual Analysis: Ongoing use of visualisations can deepen understanding of feature impacts on prices.

Task 2: Clustering

Conclusions about the clustering?

Based on the provided cluster centroids and means, three distinct groups of data points were identified by the K-Means clustering algorithm. These centroids represent the average values for each feature in the dataset across the different clusters:

Cluster 0 (C0) has higher values in features such as income, life expectancy, and lower values in child mortality. This suggests that countries in this cluster are likely more developed, with better healthcare, higher living standards, and stronger economies.

Cluster 1 (C1) has moderate values across most features but is characterised by a balance between life expectancy and moderate-income levels. This could represent developing countries that are progressing economically but still face challenges in areas like healthcare and child mortality.

Cluster 2 (C2) shows lower values in income and life expectancy but higher values in child mortality and total fertility. This cluster likely includes underdeveloped nations with weaker economic conditions and lower standards of living.

In conclusion, the clustering effectively groups countries into three categories: developed, developing, and underdeveloped nations. Each cluster is associated with unique socio-economic characteristics, and the centroids give a clear idea of how these groups differ across key factors like income, child mortality, and life expectancy.

Advice I would you give, in the context of the data, based on the clustering

Given the clustering results, here's some advice:

- **For Developed Nations (Cluster 0)**: These countries exhibit high income and life expectancy but lower fertility rates and child mortality. The main focus for these nations should be on maintaining these positive indicators, perhaps focusing on innovation in healthcare and economic sustainability, as they are already in a strong position. Policies around education, workforce skill development, and ageing populations might be relevant.
- **For Developing Nations (Cluster 1)**: The countries in this cluster show moderate progress in areas such as income and life expectancy. They should focus on strategies that promote economic growth, access to healthcare, and improvements in infrastructure. Investments in education, technology, and industrial sectors could be crucial to transitioning to the next level of development. Addressing moderate child mortality rates could also help improve overall living conditions.
- **For Underdeveloped Nations (Cluster 2)**: These nations face significant challenges, particularly with low income and high child mortality. Immediate actions should be taken to improve healthcare access, reduce fertility rates, and enhance education. International aid and investments in critical infrastructure, such as sanitation and education, may also play a vital role in supporting their growth and stabilisation.

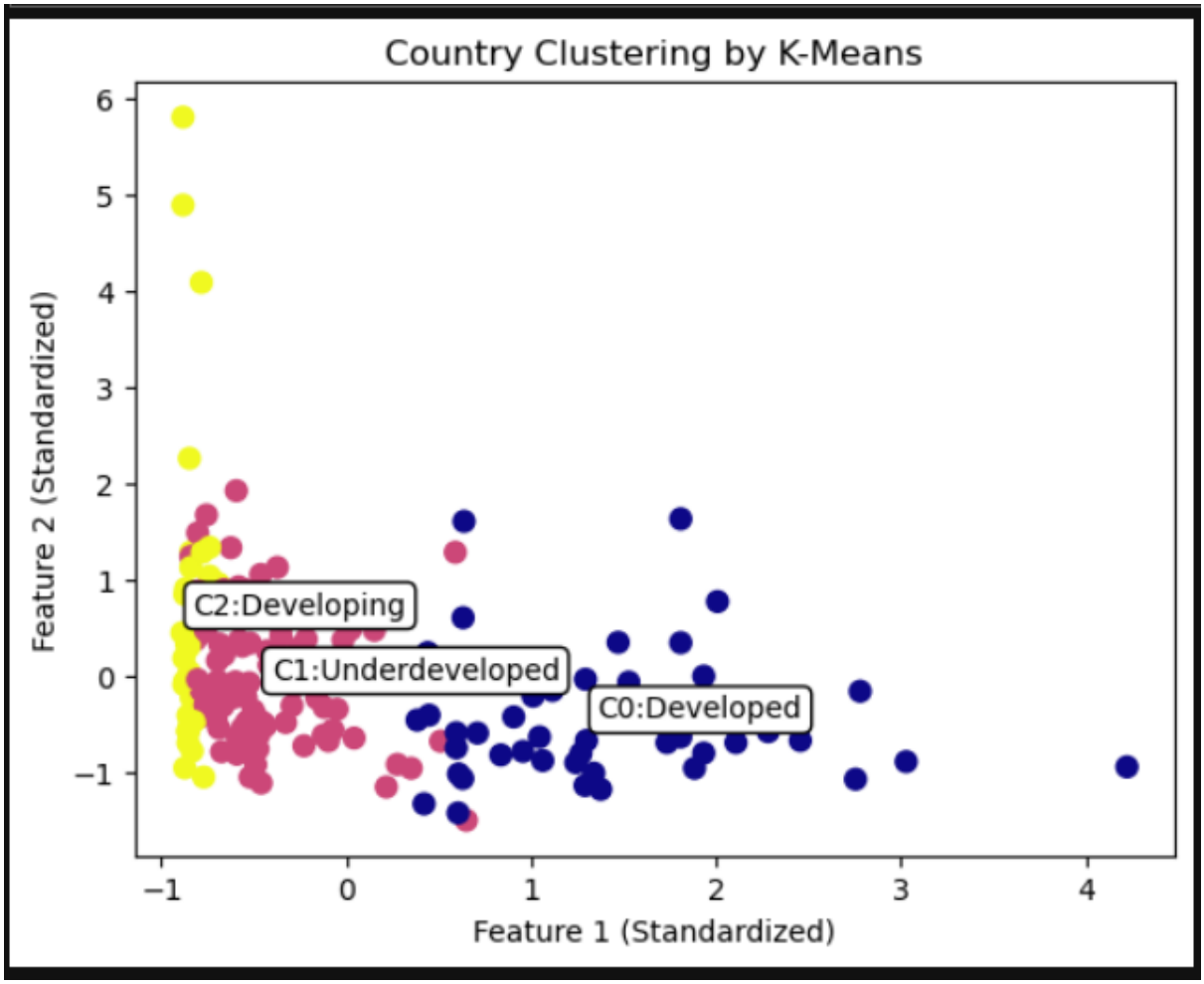


Image showing the K-Means Clustering Model

In summary, the clustering reveals key areas of focus for each group. Developed countries should work on sustaining their progress, developing nations should aim for economic and healthcare improvements, and underdeveloped countries need urgent support in addressing fundamental socio-economic challenges.

Task 3: NBA Rookie Career Prediction Analysis

The purpose of this analysis is to assess the likelihood of an NBA rookie achieving a career of five years or more, using various machine learning models to predict the target variable `Target_5Yrs`. We implemented and evaluated five models: Logistic Regression, Gaussian Naive Bayes, Decision Tree, Random Forest, and Neural Networks. Below is an overview of each model's performance and its respective strengths and weaknesses.

Model Evaluations and Interpretations

1. Logistic Regression

- **Metrics**: Accuracy: 74%; Precision (Class 0): 0.62, Precision (Class 1): 0.79; Recall (Class 0): 0.57, Recall (Class 1): 0.82.

- **Performance**: Logistic Regression produced satisfactory results, achieving a balanced performance with higher accuracy in predicting successful outcomes (players with careers ≥5 years). The precision and recall metrics reflect that this model better identifies players likely to succeed (Class 1) while maintaining acceptable accuracy for predicting shorter careers.

2. Gaussian Naive Bayes

- **Metrics**: Accuracy: 61%; Precision (Class 0): 0.46, Precision (Class 1): 0.88; Recall (Class 0): 0.88, Recall (Class 1): 0.48.

- **Performance:** With an accuracy of 61%, Gaussian Naive Bayes showed limitations in capturing complex data patterns, particularly underperforming in identifying players with longer careers. While it was accurate in identifying players with shorter careers, its high false-negative rate for long careers highlights its shortcomings in comprehensive prediction accuracy.

3. Decision Tree

- **Metrics**: Accuracy: 62%; Precision (Class 0): 0.44, Precision (Class 1): 0.73; Recall (Class 0): 0.51, Recall (Class 1): 0.67.

- **Performance**: The Decision Tree model identified some trends but demonstrated overfitting, evident in its struggle to consistently predict long-lasting careers. Despite reasonable accuracy in detecting successful careers, its misclassification rate and lower precision for predicting shorter careers limited its reliability.

4. Random Forest

- **Metrics:** Accuracy: 76%; Precision (Class 0): 0.64, Precision (Class 1): 0.83; Recall (Class 0): 0.67, Recall (Class 1): 0.81.

- **Performance:** As the top-performing model with 76% accuracy, Random Forest was effective in balancing precision and recall for both career classes, successfully capturing linear and non-linear relationships. The high accuracy and balanced performance make it a strong choice for predicting players' career longevity.

5. Neural Network

- **Metrics:** Accuracy: 70%; Precision (Class 0): 0.55, Precision (Class 1): 0.79; Recall (Class 0): 0.62, Recall (Class 1): 0.74.

- **Performance**: Neural Networks showed competitive results, with a 70% accuracy, adept at capturing complex data structures. While effective in identifying players likely to succeed, it struggled

slightly in predicting shorter careers, indicating potential for improvement in overall prediction across both classes.

Summary and Conclusion

Among the five models tested, Random Forest emerged as the most reliable model for predicting whether NBA rookies would have careers lasting five years or longer. Its high accuracy and balanced precision-recall performance underscore its suitability for this task, especially given its ability to manage both linear and non-linear patterns.

While Logistic Regression provided a balanced yet simpler approach, Gaussian Naive Bayes and Decision Tree models struggled with consistent accuracy. Neural Networks, though promising, would benefit from further tuning to achieve balanced performance.

The insights from this analysis emphasize the importance of algorithm selection in predictive modelling, particularly in sports analytics. For future work, enhancements could involve hyperparameter tuning, advanced feature engineering, or ensemble approaches to improve predictive performance across diverse career lengths.

Task 4: The Trolley Problem in the Self-Driving Vehicle Context

The trolley problem is an ethical issue which challenges individuals to make difficult ethical choices. It describes a situation where a person decides letting a trolley continue on its course, leading to the deaths of five individuals, or redirecting it into another track, where one person will be killed. The problem illustrates the tension between aiming for the greatest good of the people, or choosing the moral duty to avoid direct harm. In the context of self-driving vehicles (SDVs), the trolley problem poses substantial ethical difficulties regarding how artificial intelligence (AI) systems should make life-and-death scenarios during critical events

Self-driving cars depends on it, to operate and make decisions instantaneously. These decisions often involve intricate ethical dilemmas, especially when they face unavoidable accident situations. For example, if an autonomous vehicle faces a situation similar to the trolley problem – choosing between colliding with a group of pedestrians or veering off to strike a single pedestrian – how should its AI be programmed to respond. This raises important questions about the moral frameworks that should guide such decisions (Nyholm & Smids, 2016).

One of the key issues in applying the trolley problem to Autonomous Vehicles (AVs) is the difficulty of programming ethical decision-making into AI systems. Unlike humans, AI lacks the capacity for moral reasoning and empathy, which are crucial in navigating ethical dilemmas. As a result, the ethical decisions made by AVs would be the product of predefined rules and algorithms programmed by engineers. This raises the questions whether lawmakers, engineers or society as a whole should decide the ethical principles that self-driving vehicles can follow. Various ethical systems such as utilitarianism (choosing greatest good) and deontology (avoiding direct harm) could result in significantly different decisions in similar scenarios (Lin, 2016).

Traditionally, the trolley problem in the context of self-driving vehicles prompts discussions about responsibility and accountability. If a self-driving vehicle causes harm in an unavoidable accident, who should bear the responsibility? Is it the software developer, the manufacturer or the owner of the vehicle? These questions complicate the ethical landscape of Avs development and deployment as societies may have different views on how responsibilities should be assigned in such cases (Goodall, 2014).

In conclusion, the trolley problem highlights the ethical complexities involved in the development of autonomous vehicles. Which AI can make quick and rational decisions, it lacks moral judgement that human apply to life-and-death situations. As self-driving vehicle technology advances, it is essential for ethicists, policy makers and engineers to engage in discussions on how best to incorporate ethical principles into these systems. Establishing clear ethical frameworks will be essential to ensuring that AVs operate in ways that align with societal values and protect human lives.

References

Goodall, N. J. (2014). Machine ethics and automated vehicles. *In Road Vehicle Automation* (pp. 93-102). Springer.

Lin, P. (2016). Why ethics matters for autonomous cars. In *Autonomes Fahren* (pp. 69-85). Springer Vieweg, Berlin, Heidelberg.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice*, 19(5), 1275-1289.