**MIS780 Advanced AI For Business - Assignment 2 - T2 2024**

**Task 1: Predicting Residential Property Prices for Strategic Real Estate Decision-Making**

**Student Name:** [Your Full Name] **Student ID:** [Your Student ID]

Table of Contents

[Executive Summary 2](#_Toc176648154)

[Business Problem: 2](#_Toc176648155)

[Dataset: 2](#_Toc176648156)

[Methods: 2](#_Toc176648157)

[Data Preprocessing 3](#_Toc176648158)

[Data Cleaning and Preparation: 3](#_Toc176648159)

[Feature Engineering: 3](#_Toc176648160)

[Predictive Modeling 3](#_Toc176648161)

[Linear Regression: 3](#_Toc176648162)

[MLP Model 1 (50, 50): 3](#_Toc176648163)

[MLP Model 2 (100, 100, 100): 3](#_Toc176648164)

[MLP Model 3 (20, 20): 4](#_Toc176648165)

[MLP Model 4 (200, 200): 4](#_Toc176648166)

[MLP Model 5 (100, 100, 100, 100): 4](#_Toc176648167)

[MLP Model 6 (50, 50, 2000 iterations): 4](#_Toc176648168)

[MLP Model 7 (30, 30, 1500 iterations): 4](#_Toc176648169)

[MLP Model 8 (50, 50, 50, 50): 4](#_Toc176648170)

[MLP Model 9 (300, 300, 500 iterations): 4](#_Toc176648171)

[MLP Model 10 (100, 100, 100, 100, 100, 2000 iterations): 4](#_Toc176648172)

[Experiments and Results 4](#_Toc176648173)

[Analysis: 5](#_Toc176648174)

[Deployment Considerations: 5](#_Toc176648175)

# Executive Summary

Business Problem: The real estate market is dynamic and highly competitive, with property prices subject to fluctuations driven by a wide range of factors. For real estate companies, investors, and homeowners, accurately predicting property prices is essential for making informed decisions regarding buying, selling, and investment. Incorrect price predictions can result in significant financial losses due to overpricing, missed sales opportunities, or undervaluing properties. The business problem addressed in this project involves developing predictive models capable of accurately estimating residential property prices based on various house features, such as the number of bedrooms, bathrooms, square footage, lot size, and other important characteristics. These accurate predictions provide stakeholders with valuable insights, enabling them to optimize their strategies for market entry, property evaluation, and investment portfolio management. Furthermore, having the ability to forecast price trends offers a competitive advantage in the constantly evolving housing market.

Dataset: The dataset used for this project includes a variety of features associated with residential properties. These features include the number of bedrooms, bathrooms, square footage of living area, lot size, number of floors, and the presence of a waterfront, among others. These attributes provide essential information about the characteristics of each property. The target variable for this analysis is the house price, which the models aim to predict. By analyzing the relationships between these features and the target house price, the models offer stakeholders accurate predictions to inform their decisions within the real estate market.

Methods: Six different models were developed and compared for their effectiveness in predicting house prices:

1. **Linear Regression:** This baseline model assumes a linear relationship between the features and the target variable, providing an initial benchmark for comparison with more complex models.
2. **MLP Model 1 (50, 50):** A simple multi-layer perceptron (MLP) neural network with two hidden layers, each containing 50 neurons.
3. **MLP Model 2 (100, 100, 100):** A more complex neural network architecture with three hidden layers, each containing 100 neurons, designed to capture more intricate patterns in the data.
4. **MLP Model 3 (20, 20):** A smaller neural network with two hidden layers of 20 neurons each.
5. **MLP Model 4 (200, 200):** A larger neural network architecture with two hidden layers, each containing 200 neurons.
6. **MLP Model 5 (100, 100, 100, 100):** A deeper multi-layer perceptron neural network with four hidden layers, each containing 100 neurons.
7. **MLP Model 6 (50, 50, 2000 iterations):** A neural network with two hidden layers containing 50 neurons each, but trained for a longer duration (2000 iterations) for enhanced accuracy.
8. **MLP Model 7 (30, 30, 1500 iterations):** A neural network with two hidden layers containing 30 neurons, trained for 1500 iterations.
9. **MLP Model 8 (50, 50, 50, 50):** A neural network with four hidden layers, each containing 50 neurons.
10. **MLP Model 9 (300, 300, 500 iterations):** A neural network with two hidden layers, each containing 300 neurons, trained for 500 iterations.
11. **MLP Model 10 (100, 100, 100, 100, 100, 2000 iterations):** A neural network with five hidden layers, each containing 100 neurons, trained for 2000 iterations.

# Data Preprocessing

Data Cleaning and Preparation: The dataset underwent extensive cleaning to ensure high-quality data input for the models. Missing values were handled appropriately, numerical features were normalized, and categorical variables were encoded. For instance, features such as the number of bedrooms, bathrooms, square footage, and lot size were scaled to ensure they contributed equally to the model training process. This step was crucial for improving model convergence and prediction accuracy.

Feature Engineering: Feature engineering was conducted to boost the predictive power of the models by creating new features based on domain knowledge. Interaction terms and polynomial features were generated to capture non-linear relationships in the data, enhancing the models' ability to understand complex patterns within the dataset.

# Predictive Modeling

Linear Regression: Linear regression was trained as a baseline model to assess the performance of more complex models. While it is simple and interpretable, the linear regression model was limited in its ability to capture non-linear relationships in the data, as evidenced by its relatively high Mean Squared Error (MSE) and lower R² score compared to the other models.

MLP Model 1 (50, 50): The first multi-layer perceptron (MLP) model was developed with two hidden layers, each containing 50 neurons. This model demonstrated an improvement over linear regression in terms of predictive accuracy, as it was better suited to capture non-linear patterns. However, its relatively simple architecture limited its ability to fully understand the more complex relationships within the data.

MLP Model 2 (100, 100, 100): A more sophisticated MLP model was built with three hidden layers, each containing 100 neurons. This model demonstrated the best performance among the models tested. The deeper architecture allowed it to capture more intricate patterns in the data, leading to a significantly lower MSE and higher R² score compared to both MLP Model 1 and linear regression.

MLP Model 3 (20, 20): A smaller MLP model with two hidden layers, each containing 20 neurons, was also tested. However, it underperformed in comparison to the other neural network models, suggesting that its architecture was too simple to capture the complexities present in the dataset.

MLP Model 4 (200, 200): A larger MLP model was developed with two hidden layers, each containing 200 neurons. This model showed improved performance compared to MLP Model 1 but fell short of the performance achieved by MLP Model 2. The results suggest that while increasing the number of neurons enhances performance, the model architecture's depth also plays a crucial role.

MLP Model 5 (100, 100, 100, 100): The deepest MLP model tested, with four hidden layers, performed slightly better than MLP Model 4 but did not surpass the performance of MLP Model 2. This indicates that deeper architectures can lead to marginal performance improvements, but there are diminishing returns when the model becomes excessively deep.

MLP Model 6 (50, 50, 2000 iterations): This model was trained with the same architecture as MLP Model 1 but for 2000 iterations, allowing it more time to converge. While it performed better than MLP Model 1, it still lagged behind the more complex architectures such as MLP Model 2 and MLP Model 5, suggesting that model complexity plays a greater role in predictive power than extended training alone.

MLP Model 7 (30, 30, 1500 iterations): This model had smaller hidden layers with fewer neurons and was trained for 1500 iterations. It outperformed simpler models like MLP Model 1 but lagged behind more complex architectures like MLP Model 2 and 5.

MLP Model 8 (50, 50, 50, 50): This model had four hidden layers, each containing 50 neurons. It performed well, nearly matching the performance of MLP Model 2, highlighting the benefits of adding more layers.

MLP Model 9 (300, 300, 500 iterations): This larger model, with two hidden layers containing 300 neurons each, performed worse than smaller models like MLP Model 8 and MLP Model 2. Increasing the size of hidden layers alone did not improve performance.

MLP Model 10 (100, 100, 100, 100, 100, 2000 iterations): Despite its deep architecture, this model did not surpass the performance of MLP Model 2, suggesting diminishing returns for increased complexity beyond a certain point.

# Experiments and Results

The following table summarizes the performance of the eleven models:

| **Model** | **MSE** | **R²** |
| --- | --- | --- |
| Linear Regression | 41,333,770,000 | 0.7036 |
| MLP Model 1 (50, 50) | 26,806,740,000 | 0.8078 |
| MLP Model 2 (100, 100, 100) | 13,783,950,000 | 0.9012 |
| MLP Model 3 (20, 20) | 29,577,770,000 | 0.7879 |
| MLP Model 4 (200, 200) | 23,352,610,000 | 0.8325 |
| MLP Model 5 (100, 100, 100, 100) | 16,242,710,000 | 0.8835 |
| MLP Model 6 (50, 50, 2000 iterations) | 16,539,640,000 | 0.8814 |
| MLP Model 7 (30, 30, 1500 iterations) | 26,383,730,000 | 0.8108 |
| MLP Model 8 (50, 50, 50, 50) | 14,814,590,000 | 0.8938 |
| MLP Model 9 (300, 300, 500 iterations) | 25,364,580,000 | 0.8181 |
| MLP Model 10 (100, 100, 100, 100, 100, 2000 iterations) | 17,322,540,000 | 0.8758 |

Analysis: The results of the experiments reveal that **MLP Model 2** outperformed the other models, thanks to its deeper architecture that enabled it to capture more complex patterns in the data. The **MLP Model 8** also demonstrated strong performance, showing the benefits of deeper neural networks, but without surpassing the accuracy of Model 2. **MLP Model 1** offered a good balance between simplicity and accuracy but fell short of the deeper models. Meanwhile, the **linear regression model** had the highest MSE and lowest R², confirming its limitations in capturing the non-linear relationships that exist in real estate data.

Deployment Considerations: Deploying **MLP Model 2** or **MLP Model 8** in a real-world setting requires careful attention to several factors. First, the computational cost of these models, particularly **MLP Model 2**, must be considered, as deeper architectures require more processing power and memory. This may limit their use in environments with restricted resources. Second, model interpretability is an important factor, especially for stakeholders like investors and real estate agents who need to understand how certain features affect property prices. **MLP models** offer higher accuracy but are less interpretable than simpler models like linear regression. Therefore, a trade-off between accuracy and transparency must be evaluated. Additionally, these models will need regular retraining to stay updated with changing market conditions and evolving property features, ensuring that they remain accurate over time. Finally, ongoing monitoring is essential to track model performance and adapt to any shifts in the real estate market that may influence property prices.