**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](#_Toc176372570)

[Business Problem: 2](#_Toc176372571)

[Dataset: 2](#_Toc176372572)

[Methods: 2](#_Toc176372573)

[Experiments and Results: 2](#_Toc176372574)

[Data Preprocessing 2](#_Toc176372575)

[Data Cleaning and Preparation: 2](#_Toc176372576)

[Feature Engineering: 3](#_Toc176372577)

[Predictive Modeling 3](#_Toc176372578)

[Linear Regression: 3](#_Toc176372579)

[Decision Tree Regression: 3](#_Toc176372580)

[MLP Model 1: 3](#_Toc176372581)

[Random Forest Regression: 3](#_Toc176372582)

[MLP Model 2: 3](#_Toc176372583)

[Experiments Report 3](#_Toc176372584)

[Model Comparison: 3](#_Toc176372585)

# Executive Summary

Business Problem:  
The real estate market is highly competitive and dynamic, with property prices fluctuating based on a variety of factors. For real estate companies, investors, and homeowners, accurately predicting property prices is essential to make informed decisions regarding buying, selling, and investing. Failure to predict prices correctly can lead to overpricing, lost sales opportunities, or undervaluing properties, which may result in significant financial losses. In this project, the business problem addressed is developing a predictive model that can accurately estimate residential property prices based on various house features, such as the number of bedrooms, bathrooms, square footage, lot size, and other relevant characteristics. These predictions enable stakeholders to optimize their strategies for market entry, property evaluation, and investment portfolio management. Furthermore, the ability to forecast price trends provides a competitive edge in a rapidly evolving housing market.

Dataset:  
The dataset consists of various features related to residential properties, such as the number of bedrooms, bathrooms, square footage of the living area, lot size, number of floors, and whether the property has a waterfront. These features provide crucial information about the characteristics of each house. The target variable in this analysis is the house price, which the models are designed to predict. By analyzing the relationships between these features and the house price, the models aim to offer accurate predictions, helping stakeholders make informed decisions in the real estate market.

Methods:  
Five different models were developed and compared for their performance in predicting house prices:

1. **Linear Regression:** A baseline model that assumes a linear relationship between the features and the target variable.
2. **Decision Tree Regression:** A non-linear model that splits the dataset into different segments and fits a simple model in each segment.
3. **MLP Model 1:** A neural network with a simple architecture consisting of two hidden layers, each with 50 neurons.
4. **Random Forest Regression:** An ensemble model that uses multiple decision trees to improve the accuracy and robustness of the prediction.
5. **MLP Model 2:** A more complex neural network architecture with three hidden layers, each containing 100 neurons.

Experiments and Results:  
The performance of the models was evaluated using Mean Squared Error (MSE) and R-squared (R²) metrics. The results indicate that MLP Model 2, with its more complex architecture, significantly outperformed both the Linear Regression model and MLP Model 1.

# Data Preprocessing

Data Cleaning and Preparation:  
The dataset was cleaned by handling missing values, normalizing numerical features, and encoding categorical variables. Features such as bedrooms, bathrooms, square footage, and others were scaled to ensure they contributed equally to the model.

Feature Engineering:  
New features were created based on domain knowledge to enhance the model's predictive power. This included interaction terms and polynomial features that could capture non-linear relationships.

# Predictive Modeling

## **Linear Regression:**

The linear regression model was trained as a baseline to assess the performance of more complex models. While simple and interpretable, it was limited in its ability to capture non-linear patterns in the data.

## **Decision Tree Regression:**

The decision tree model provided better performance than linear regression by capturing non-linear relationships. However, it was prone to overfitting and lacked the ability to generalize as effectively as more complex models.

## **MLP Model 1:**

A simple multi-layer perceptron (MLP) was developed with two hidden layers. This model showed improvement over linear regression but struggled with capturing complex relationships due to its limited architecture.

## **Random Forest Regression:**

The random forest model, an ensemble of decision trees, significantly improved prediction accuracy and reduced overfitting by averaging multiple trees' predictions. It demonstrated strong performance but was slightly less accurate than the most complex neural network model.

## **MLP Model 2:**

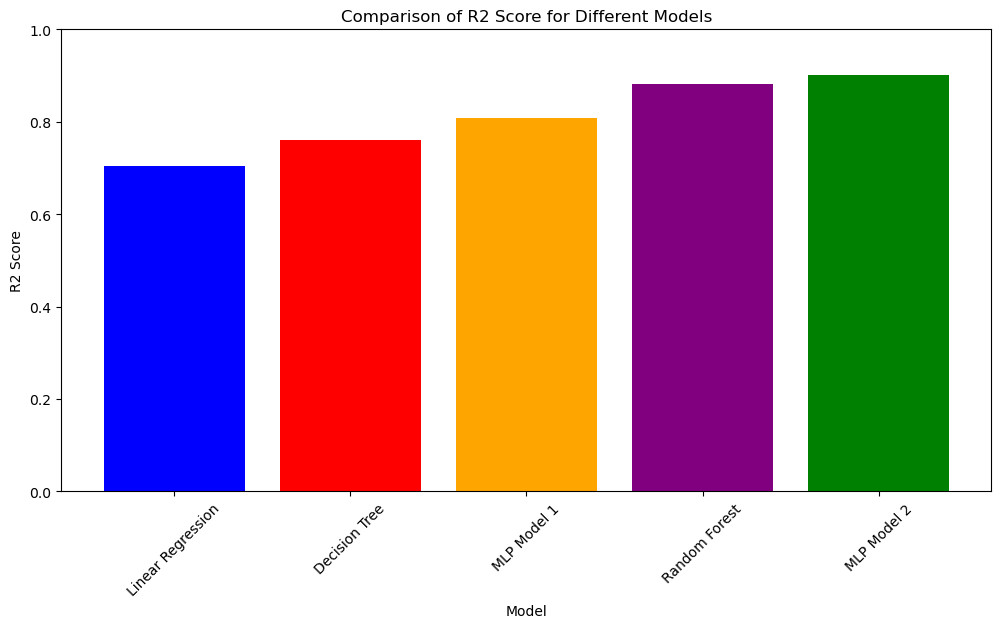
A more sophisticated MLP model was built with three hidden layers, each containing 100 neurons. This model demonstrated the best performance, capturing the nuances in the data more effectively than the simpler models.

# Experiments Report

Model Comparison:  
The following table summarizes the performance of the five models:

| **Model** | **MSE** | **R²** |
| --- | --- | --- |
| Linear Regression | 41,333,765,808.39 | 0.7035 |
| Decision Tree Regression | 33,341,603,650.97 | 0.7609 |
| MLP Model 1 | 26,806,736,596.51 | 0.8077 |
| Random Forest Regression | 16,405,203,002.27 | 0.8824 |
| MLP Model 2 | 13,783,950,875.72 | 0.9011 |

Here is attached Visual to check difference between models

****

**Analysis:**  
MLP Model 2 outperformed the other models due to its deeper architecture, which allowed it to learn more complex patterns in the data. The Random Forest Regression model also performed well, demonstrating the effectiveness of ensemble methods in capturing non-linear relationships while reducing overfitting. The improved performance of these models justifies their use in real-world applications where predictive accuracy is critical.

**Deployment Considerations:**  
When deploying MLP Model 2 or Random Forest Regression in a real-world scenario, several factors must be carefully considered to ensure optimal performance and practicality. First, computational cost is a crucial aspect, as both models, particularly MLP Model 2 with its deeper architecture, require significant processing power and memory. This may limit their application in resource-constrained environments. Second, model interpretability is vital, especially for stakeholders like investors or real estate agents who need to understand how certain features influence property prices. While Random Forest offers better interpretability compared to the more complex neural network, this trade-off between accuracy and transparency should be evaluated. Third, the need for continuous retraining must be addressed as market conditions and property features change over time. Regularly updating the model with new data will help maintain its accuracy. Finally, ongoing performance monitoring is essential to ensure that the models adapt effectively to any shifts in the real estate market.