Final Report

Introduction

The goal of this machine learning project is to develop two machine learning models that predict outcomes in the housing market: estimating house prices and determining whether a house will sell within 30 days. To predict these outcomes, data was prepared, transformed, and analyzed using regression and classification techniques. The project entailed a model comparison, which included measuring the impact of LASSO regularization on finding the best-performing solution.

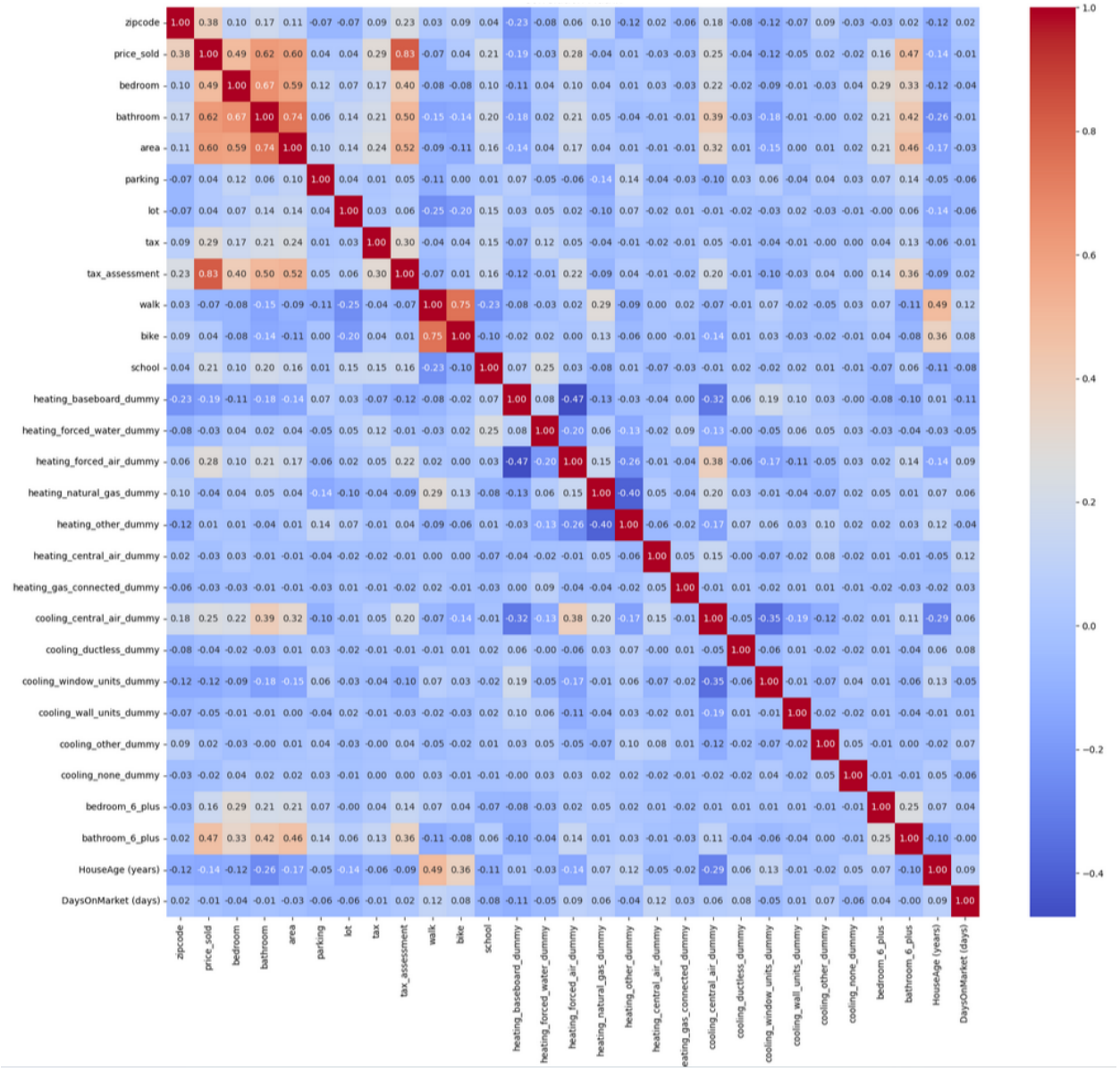
Data Preparation and Transformation

To begin the project, there was a process of data cleaning and transformation to ready the dataset for analysis. Many unneeded and irrelevant columns were removed to reduce clutter in the dataset. Further, two new variables were created based on the existing variables in our dataset: firstly, a calculation of the age of each house was made by subtracting its year built from the current year (2024) and second, the number of days a house was on the market by subtracting the listing data from the sale date.

Exploratory Data Analysis

Once data preparation and transformation were completed, a preemptive analysis of the data was conducted to better understand and identify relationships among the independent and dependent variables. Outliers were identified and removed in key columns like tax values, house area, and lot size using the interquartile range of each data column. Next, a correlation matrix

was used to evaluate the relationships between features for feature selection and multicollinearity identification.



Feature Engineering

Once complete with the exploratory analysis, various feature engineering techniques were used to prepare the dataset for modeling. A new binary feature was made to describe whether a house had been sold within 30 days which later serves as the target variable for the

classification model. Interaction terms were introduced to demonstrate the relationships between important independent variables such as bedrooms, bathrooms, house area, tax assessments, and walkability scores.

Polynomial features were also created to more accurately identify true relationships which included square root functions, logarithmic functions, and squared functions. These changed relationships helped capture necessary non-linear correlation within the data, especially for data like number of bedrooms, number of bathrooms, area, parking, tax and more.

Model Development

As stated previously, this project involves two modeling tasks: predicting house prices using regression and determining the likelihood of a house selling within 30 days using classification.

For the regression model, a linear regression was first developed using features that were determined to be relevant by logical inference. As can be seen by the image below, the model performed quite well, getting a high accuracy and strong predictive ability on the testing data.

```
Model 1: Without Lasso Regularization

Training Metrics:
MAE: 329116.84
RMSE: 586061.64
R²: 0.78
Adjusted R²: 0.77

Testing Metrics:
MAE: 327159.91
RMSE: 519737.68
R²: 0.84
Adjusted R²: 0.81
```

To improve on the metrics of our model, LASSO regularization was used. LASSO selected a subset of all of the features, trying to gather the most influential variables in the

dataset. Although the updated model using LASSO did not affect our analyzed metrics significantly, it improved the adjusted R^2 on the testing metrics, as can be seen from the figure below.

```
Model 2: With Lasso Regularization

Training Metrics:
MAE: 330649.53
RMSE: 586162.78
R²: 0.78
Adjusted R²: 0.77

Testing Metrics:
MAE: 333702.36
RMSE: 528073.76
R²: 0.83
Adjusted R²: 0.82
```

In the classification task, two models were used to predict whether a house would sell within 30 days: logistic regression model and decision tree. The logistic regression model had a strong performance with high precision. However, the model did have some difficulty in predicting houses that did not sell within 30 days.

```
Logistic Regression Model Performance:
Accuracy: 0.9070
Precision: 0.9112
Recall: 0.9949
F1-Score: 0.9512
AUC-ROC: 0.6917

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        19
           1       0.91      0.99      0.95       196

    accuracy                           0.91       215
   macro avg       0.46      0.50      0.48       215
weighted avg       0.83      0.91      0.87       215

Confusion Matrix:
[[  0  19]
 [  1 195]]
```

Next, the decision tree classifiers were applied to the same task, which resulted in a perfect performance. Even though the tree accurately predicted every outcome, the results are a suggestion of overfitting. In other words, the model could have captured specific noise or patterns within the data that were reflected in the outcome.

A decision tree classifier was then applied to the same task, producing perfect performance metrics. While the decision tree accurately predicted all outcomes, its flawless results suggested overfitting, indicating that the model may have captured noise or specific patterns in the training data.

Conclusion

This project was successful in developing two machine learning models to predict house prices and the likelihood of a house being sold within 30 days. The regression models had strong performances and the logistic regression model gave balanced results with high precision. Although the decision tree model was likely subject to overfitting, it demonstrates the importance of human intervention in developing models. If, for example, the decision tree model were to then be implemented, it could see major failures and inaccuracies on foreign data. Overall, this project is a good example of the capabilities of machine learning in making effective predictions that can drastically improve the decision-making process.