**Summary:**

This project aimed to explore different modeling approaches to predict Miami home prices and their strengths and limitations. The study applied three modeling approaches: multivariate linear regression, KNN regression, and random forest regression. A summary of high-level results are show in the table below:

|  |  |  |
| --- | --- | --- |
| **Model** | **Explained Variance** | **Notes** |
| Linear Regression | 0.68 | Pro: easy interpretation  Cons: lowest model quality |
| KNN Regression | 0.71 | Pro: easy interpretation & improved model quality  Cons: room for model quality improvement |
| Random Forest Regression | 0.92 | Pro: highest model quality  Con: Low interpretability to drive decision making |

**Data:**

The main dataset contains information on 13,932 single-family homes sold in Miami in 2016 and contains 17 columns ([Kaggle: Miami Sold Home Prices for 2016](https://www.kaggle.com/datasets/deepcontractor/miami-housing-dataset)).

* PARCELNO: unique identifier for each property. About 1% appear multiple times.
* SALE\_PRC: sale price ($)
* LND\_SQFOOT: land area (square feet)
* TOT\_LVG\_AREA: floor area (square feet)
* SPEC\_FEAT\_VAL: value of special features (e.g., swimming pools) ($)
* RAIL\_DIST: distance to the nearest rail line (an indicator of noise) (feet)
* OCEAN\_DIST: distance to the ocean (feet)
* WATER\_DIST: distance to the nearest body of water (feet)
* CNTR\_DIST: distance to the Miami central business district (feet)
* SUBCNTR\_DI: distance to the nearest subcenter (feet)
* HWY\_DIST: distance to the nearest highway (an indicator of noise) (feet)
* age: age of the structure
* avno60plus: dummy variable for airplane noise exceeding an acceptable level
* structure\_quality: quality of the structure
* month\_sold: sale month in 2016 (1 = jan)
* LATITUDE
* LONGITUDE

(All columns except for avno60plus (dummy variable for airplane noise exceeding an acceptable level) are numerical variables.)

Additional data mergers were conducted to identify the quality of schools associated to each home based on latitude and longitude information and include that as a predictor. This task used Tableau and R to preprocess the data and prepare it for modeling.

**Results:**

The coefficients for each predictor variable (given by the linear regression) and their importance (given by the random forest regression) are found in the table below.

|  |  |  |
| --- | --- | --- |
| **Predictor Variable** | **Linear Beta Coefficient** | **Importance** |
| tot\_lvg\_area | 189.14 | 0.42 |
| ocean\_dist | -4.77 | 0.15 |
| cntr\_dist | -2.94 | 0.15 |
| water\_dist | -0.41 | 0.06 |
| subcntr\_di | 0.03 | 0.06 |
| structure\_quality | 58599.26 | 0.06 |
| lnd\_sqfoot | 2.53 | 0.02 |
| spec\_feat\_val | 2.92 | 0.02 |
| rail\_dist coef | 4.41 | 0.02 |
| hwy\_dist | 4.49 | 0.02 |
| age | -1874.45 | 0.02 |
| weighted\_avg\_school\_grade | 50075.53 | 0.01 |
| avno60plus | -72655.36 | 0 |
| month\_sold | 270.82 | 0 |

The results suggest that the top 5 most important factors to determining home prices in Miami are: total space of home (tot\_lvg\_area), distance to ocean (ocean\_dist), distance to city center (cntr\_dist), structure quality (structure\_quality), and distance to sub city center (subcntr\_di). A KNN regression model was also built, which predicted housing prices by taking the average of the price of the 5 closest homes with respect to geographical distance. This model performed better than the linear model but worse than the random forest model.