Project 1 - Formal Report

Oluwabukola Emi-Johnson

2023-11-09

Abstract

Accurately predicting house prices is important for the real estate sector. In this report, we explore the process of estimating house prices using two predictive techniques; Linear Regression and K-Nearest Neighbors (KNN). By meticulously comparing these methods, we determined that KNN outperforms Linear Regression, with a lower Root Mean Square Error (RMSE) of 47.96. This translates to an average prediction deviation of approximately $47,962.86. The study details the steps taken and compares the results of both models and based on the analysis, the adoption of KNN for precise and reliable house price predictions is recommended. We also note that it is crucial to fine-tune the model for the best results as our data evolves.

Section 1: Linear Regression

To predict house prices, the analysis started with Linear Regression, utilizing all features and observations (rows) from the dataset to train the model and obtain estimates for the coefficient of each feature. The resulting estimates, as presented in the table below, provide a preliminary insight into the influence of each feature on the response variable: House prices. This analysis represents a critical step towards achieving the project's objectives.

Table 1: Table 1: Coefficients of Linear Regression Model

	Coefficients
(Intercept)	-1207.4570561
LotFrontage	-0.1734013
LotArea	0.0005522
OverallQual	27.6612095
OverallCond	8.0404681
YearBuilt	0.3138738
YearRemodAdd	0.2237385
MasVnrArea	0.0341827
TotalBsmtSF	-0.0017929
X1stFlrSF	0.0564686
X2ndFlrSF	0.0426149
GrLivArea	-0.0055312
BsmtFullBath	17.2534384
BsmtHalfBath	8.6509068
HalfBath	-2.3694722
BedroomAbvGr	-6.1049160
KitchenAbvGr	-36.8901725
TotRmsAbvGrd	8.2417254
Fireplaces	10.7534820

	Coefficients
GarageArea	0.0507471
WoodDeckSF	0.0142397
OpenPorchSF	-0.0275486
EnclosedPorch	0.0081792
X3SsnPorch	0.1315806
ScreenPorch	0.0603385
PoolArea	-0.0665365
MoSold	-0.0531326

Analysis of Table 1 above reveals that the most influential features on house prices include:

- KitchenAbvGR - How many kitchens are not in the basement?
- OverallQual - Quality score, 0 -10
- BsmtFullBath - How many full bathrooms are in the basement?
- BsmtHalfBath - How many half bathrooms are in the basement?
- TotRmsAbvGr - How many rooms in the home are not in the basement?

To assess the predictive accuracy of our regression model, two cross-validation (CV) techniques were employed, utilizing all features within our dataset. The chosen CV techniques for this analysis are 10-fold Cross-Validation (10-fold CV) and Leave-One-Out Cross-Validation (LOOCV).

In 10-fold CV, the dataset is divided into 10 equally sized folds, although the size of the dataset could also influence this equal division, with the model being trained and evaluated 10 times. Each iteration employs a different fold as the test set and the remaining 9 folds as the training set. This ensures every row in the dataset is used for both training and testing. On the other hand, LOOCV adopts a granular approach, using each row as the test set once, while the rest of the data forms the training set. However, LOOCV can be computationally intensive for large datasets, as it requires fitting the model as many times as there are rows in the dataset, making it better suited for smaller datasets.

Table 2: Table 2: Test RMSE and Time taken for 10-fold CV and LOOCV

	10-fold CV	LOOCV
test RMSE	56.92719	55.83159
Time (secs)	0.08000	3.22000

Referencing Table 2 above, the computational time for our 10-fold CV technique was approximately 0.08 seconds, while the LOOCV technique took about 3.22 seconds. Utilizing 10-fold CV saved us approximately 3.14 seconds in computation time. However, it is worth noting that despite the time difference, LOOCV yielded a better test Root Mean Square Error (RMSE) of 55.83 compared to 56.92 for 10-fold CV.

The test RMSE is an important metric for assessing our model's predictive accuracy. In the context of our analysis, LOOCV indicates that, on average, our predicted house prices deviate from the actual prices by approximately $55,831.59. This insight provides valuable feedback on the precision of our predictions.

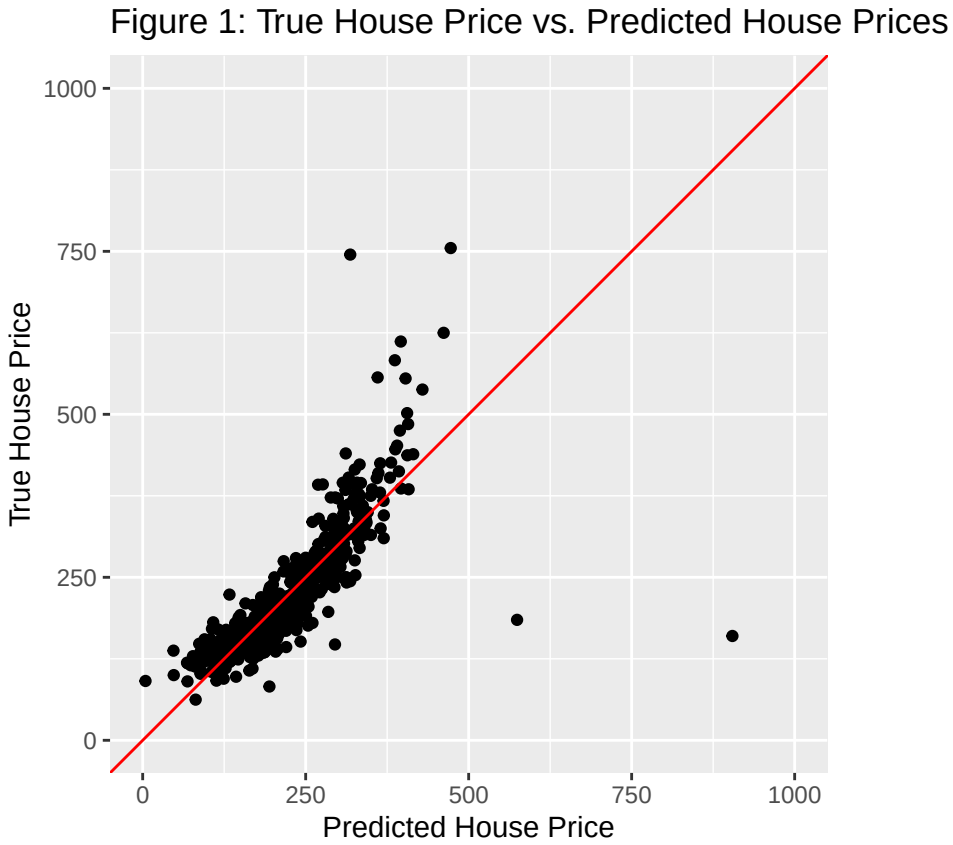


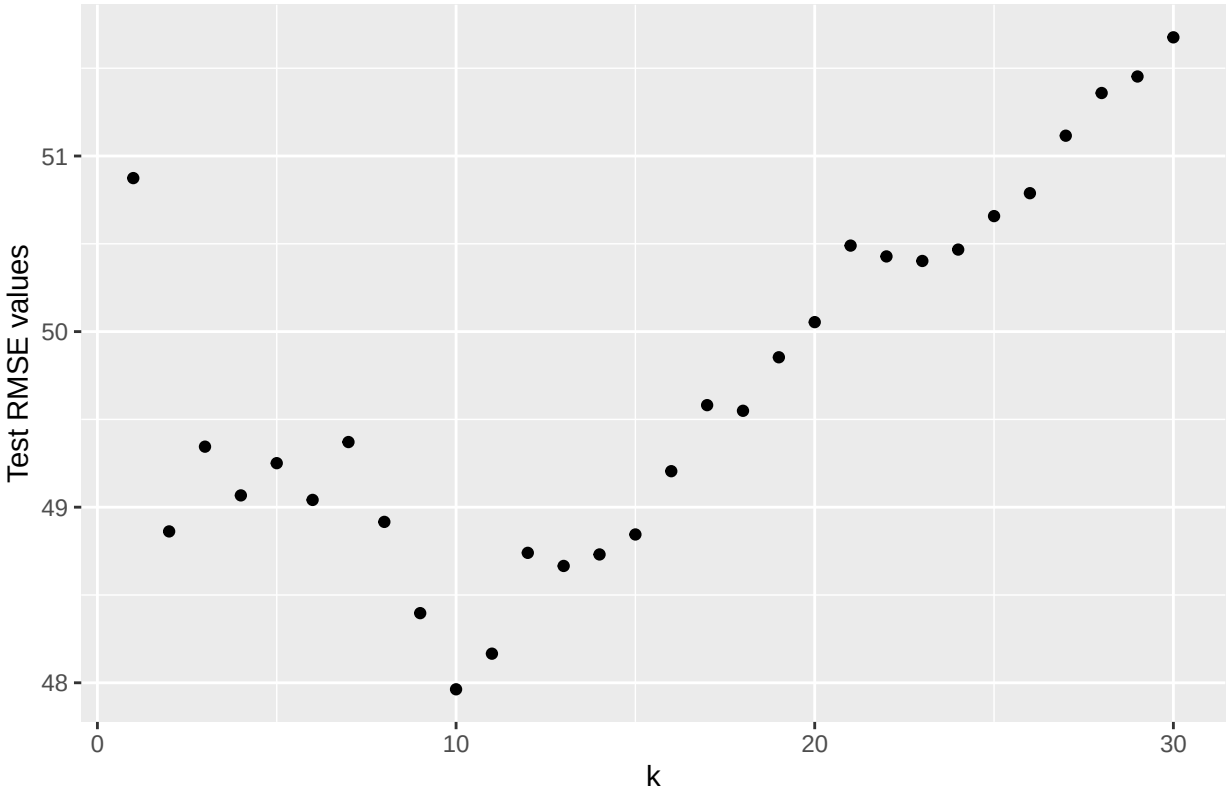
Figure 1 shows the relationship between predicted house prices and their corresponding true values. A noticeable concentration of points along the diagonal line signifies relatively accurate predictions, demonstrating the model's reliability for a significant portion of the dataset. However, the presence of outliers (extreme values) highlights areas where the model struggles to predict effectively. These outlines signify specific instances where the model's accuracy falls short.

Section 2: KNN

In our pursuit of predicting house prices, we explored an alternative approach: the K-Nearest Neighbor (KNN) algorithm. KNN predicts outcomes based on similarity, identifying the 'k' nearest data points in the training set and averaging their response variable values to make predictions.

The choice of 'k' in KNN significantly impacts prediction accuracy and requires careful consideration. To determine the optimal 'k,' we experimented with values ranging from 1 to 30. Employing 10-fold Cross-Validation (10-fold CV) for evaluation, we calculated the test RMSE for each 'k'. Figure 2 illustrates the test RMSE values corresponding to different 'k' values. Analyzing the plot, we identified a trend and found that the lowest test RMSE occurs at 'k = 10.' This insight guides our selection of the optimal 'k' for our predictive model.

Figure 2: test RMSE for respective K



After obtaining the optimal 'k' value for our K-Nearest Neighbor (KNN) algorithm, we conducted another round of testing using both cross-validation (CV) techniques in this analysis. The test RMSE values obtained were 10-fold CV (47.96) and LOOCV (48.52). Comparing the two, 10-fold CV produced a lower test RMSE, making it the preferred choice. Hence, utilizing 10-Nearest Neighbors (10NN), our predicted house prices deviate from the true prices by an average of $47,962.86. This insight showcases the accuracy level of our predictions.

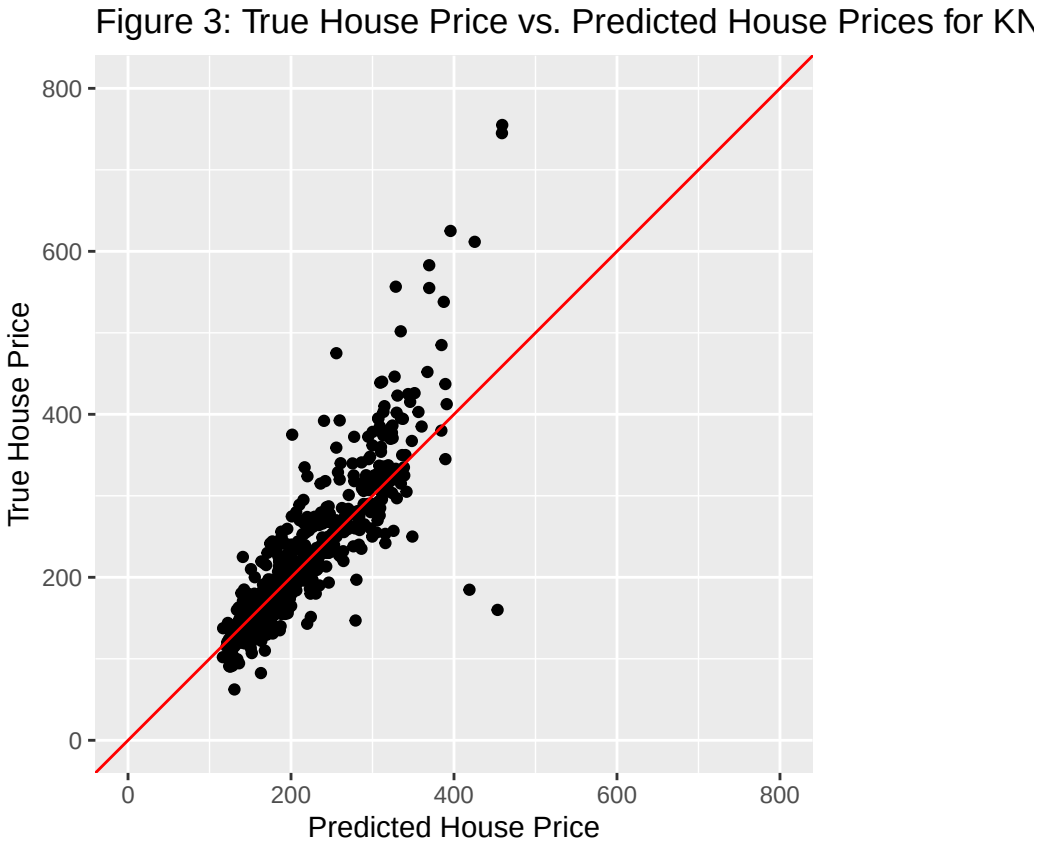


Figure 3 displays the relationship between predicted and actual house prices. Points cluster closely around the diagonal line, especially for prices around $300,000, indicating accurate predictions within this range. However, the model's accuracy diminishes for prices above $300,000, revealing its limitations in higher price brackets.

Section 3: Conclusion and Recommendation.

After thorough analysis, it is evident that K-Nearest Neighbors (KNN) outperforms Linear Regression, achieving the lowest test RMSE of 47.96. This implies that, on average, our predicted house prices deviate from the true values by approximately $47,962.86. The significantly lower RMSE of KNN signifies higher predictive accuracy compared to Linear Regression, making it the superior choice for our analysis.

When selecting a model, it is crucial to consider the data's structure, features, and analysis goals. Our primary objective is to make precise predictions. Given KNN's superior predictive accuracy over Linear Regression, it aligns with our objective. Therefore, I strongly recommend utilizing the KNN approach for predicting house prices in this context.

Furthermore, it is important to note that KNN's effectiveness depends on the choice of 'k'. Fine-tuning this parameter might further enhance the model's accuracy. Also, regular refinement and adjustment of the model, as the dataset evolves or expands, will ensure the continued reliability of our predictions.