Assessing Predictive Performance of Classification Models Used to Classify House Sale Prices

Michelle Cleary, Fionnuala Marshall, Ellen Crombie

Executive Summary
This report evaluates two models which predict whether a house will sell for more or less than the average market price. The models were formulated using data from the sale of houses in an American city over a 5 year period, which includes information about the features of each house, such as house style and lot area.

1. Model 1 uses all 29 house features from the dataset, and we estimate that when asked to make 100 predictions, 89 will be correct.
2. Model 2 uses 5 house features: lot area, year built, number of full bathrooms, number of bedrooms, and house style. We estimate that when asked to make 100 predictions using this model, 73 will be correct.

Considering this, both models performed well. To highlight a few key results, both models determined that houses which were newer, had larger lot areas, and a higher number of bedrooms were more likely to be above average price. Model 1 suggested that the timing of the sale can be a significant factor in determining a house's selling price. It indicated that houses sold during peak buying seasons (such as spring and summer) had a higher probability of selling for above average price than below average price, compared to winter and autumn.
We recommend Model 1 in most cases since it has a better prediction accuracy. However, Model 2 still has a good prediction accuracy and only requires information regarding 5 standard house features, which are likely to be readily available for most houses entering the market, even if there is limited information about more specific features such as kitchen quality.

Introduction

In this report, we implemented two models for predicting whether a house will sell for more or less than the market average. The dataset used to create our models included 1460 observations based upon 31 features. The data focused on the sale price of numerous houses in an American city over a 5 year period and included information about features of each house, such as lot area in square feet and neighbourhood.

We first fitted a logistic regression model to the data using all 29 available predictor features. We also fitted a Naive Bayes model, using five features: lot area, year built, number of full bathrooms, number of bedrooms, and house style. We assessed the performance of each model under 10-fold cross validation by measuring prediction accuracy and computing the Brier score. We defined prediction accuracy as the proportion of the test data which was correctly classified.

Preprocessing of the Dataset

When carrying out initial exploration of the data, we looked for outliers and any features which had a significant level of missing values. Lot area appeared to contain some values outside of the usual range. However, these

correlated with other features of the properties, and thus, we did not feel it was appropriate to remove them entirely from our dataset. The features garage type and electrical system had missing values. We were able to remove any entries corresponding to these values since Figure 1 shows that they only represented a small proportion of entries in the dataset and so did not majorly reduce the number of observations.

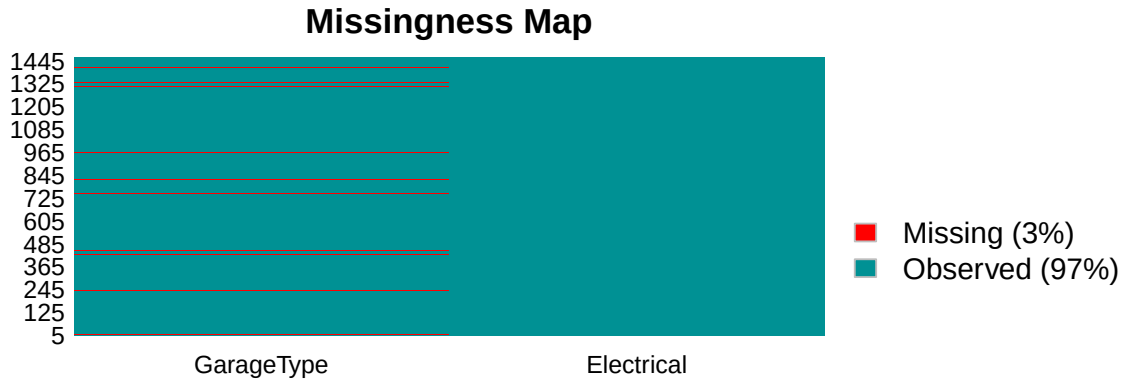

Figure 1: Missingness map showing missing values for features garage type and electrical system.

We created a binary feature called average price which took the value 0 if the house sold for less than the average price, and 1 if the house sold for greater than the average price, based on all sale prices. This allowed us to measure whether a house sold for more or less than the market average for classification within our models.

Next, we coded factor features and made simplifications based on similarity of level descriptions. For example, the house style levels "1.5 story finished" and "1.5 story unfinished" were combined into "1.5 story", and similarly for "2.5 story". We decided that features such as neighbourhood, which had a high granularity, could not be simplified any further because the dataset did not contain any further information or context for us to do so. Finally, we transformed the month sold feature to instead describe the season sold, in order to simplify the identification of trends throughout later analysis, with 4 factor levels instead of 12. We believe this was appropriate due to common literature surrounding house prices often referring to seasons rather than months [1].

Logistic Regression Model

Our first classification technique was fitting a logistic regression model using all the predictor features within our dataset, assuming that the residuals followed a binomial distribution. This model assumes that there is a linear relationship between the logit function of the classification outcome and each of the predictor features [2], and in its simplest form can be written:

$$\log \frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} = \theta^T x_i,$$

where each $y_i \in \{0, 1\}$ denotes the classification outcome (house selling for below or above average price), and θ^T is the vector of coefficients for each feature x_i.

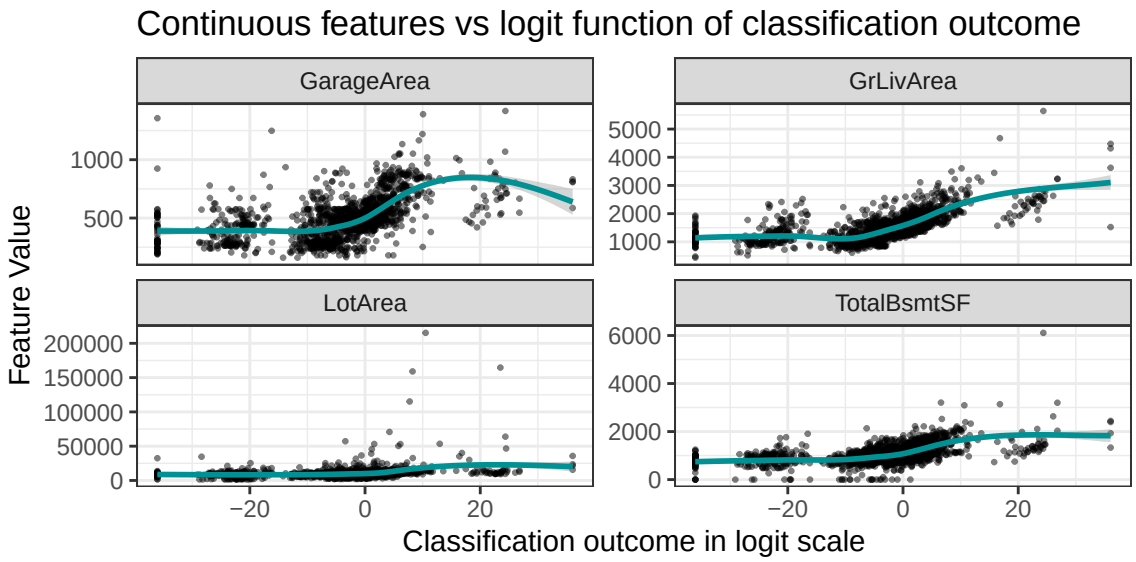

Figure 2: Assessing the linearity assumptions between continuous features and the logit function of the classification outcome.

Before analysing the results, we used Figure 2 to check the linearity assumption between continuous features and the log odds of the classification outcome, defined by the function: $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$ where p_i is the probability of the outcome. All continuous features except garage area appear to be fairly linearly associated with the classification outcome. For the purpose of this report we did not transform garage area, but noted that to improve the validity of this model in the future, we would investigate the use of a spline or polynomial transformation for this feature [2]. Through further assessment of variance inflation factors, we discovered a potentially problematic amount of collinearity among the number of kitchens and other features. We noted this as an improvement point to increase the performance of the model if required in the future.

To highlight one key finding, as lot area increases by one unit, the probability that the house price is above average is 2.7 $\times 10^{-5}$ times as large at the probability that the house price is below average when all other features are held constant. Figure 3 displays the sigmoid-shaped logistic relationship between lot area and average house price, for each season. In agreement with wider research regarding house prices [1], our model suggests that houses sold in spring and summer have higher probabilities of being above average price than being below, compared to winter and autumn. However, the extent to which this relationship is true may not be as significant as it appears in Figure 3, due to the aforementioned properties with extremely large lot areas being sold in summer.

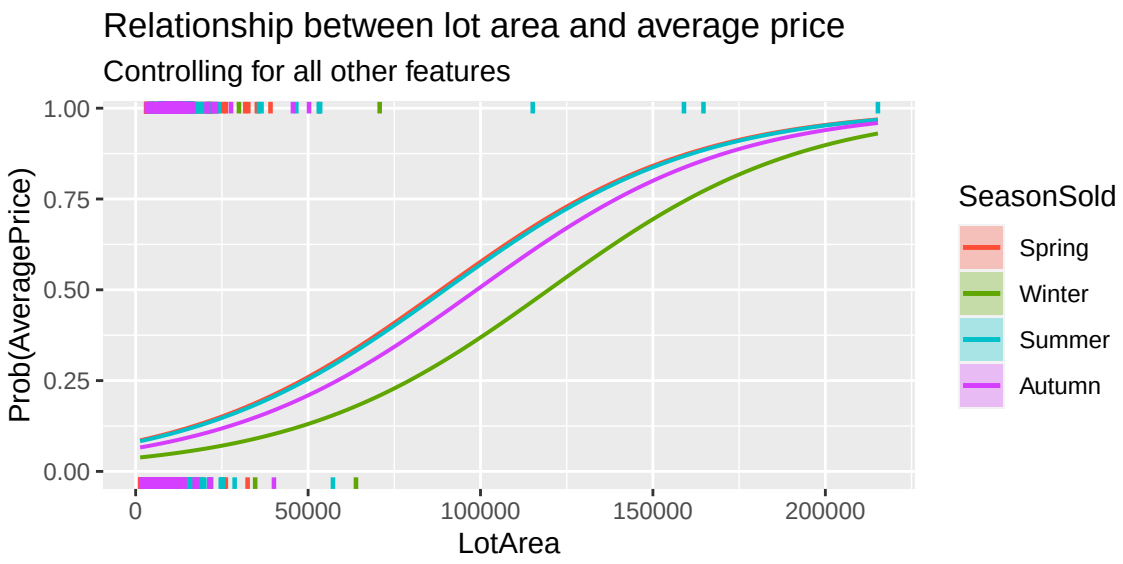

Figure 3: Relationship between lot area and average price, after fitting the logistic regression model.

Assessment of model

To gain an initial understanding of the accuracy of this model, we compared the proportion of correct and incorrect predictions of average house price. Our model is 94.485% accurate. However, this result is likely to be an over estimate of accuracy due to the introduction of an over fitting bias, because the model was both trained and tested on the same set of observations. To remedy this, we performed 10-fold cross validation and computed the accuracy rate for each fold as a proportion of the test data correctly classified. The accuracy of the model ranges from 69.12% to 93.48%, and the averaged accuracy rate across each fold is approximately 88.55%, which is below our initial projection of accuracy, but still suggests that roughly 88.55% of the time, the model will correctly predict whether a house price is above or below average.

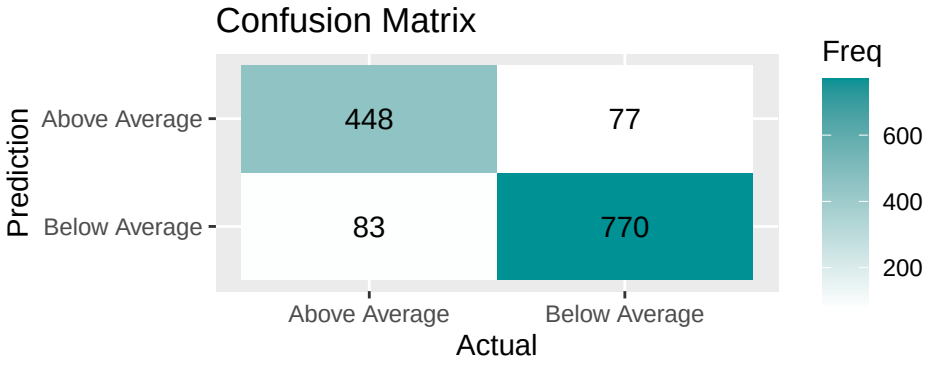

Figure 4: Plot displaying the counts of incorrect and correct predictions.

Figure 4 displays the counts of correct and incorrect predictions, showing how 9.09% of below average house prices were incorrectly predicted as being above average, compared to 15.63% of above average house prices being incorrectly predicted as below average. This suggests that the model might be better at correctly predicting below average house prices than above average house prices for this particular dataset.

Furthermore, whilst Figure 4 gives a good indication of performance, counting the number of correct classifications

ignores the probabilistic element of prediction. Considering this, we also calculated the Brier score for each fold and took an average of these scores after cross validation. The Brier score measures the accuracy of probabilistic predictions and can be calculated $\frac{1}{n}\sum_1^n (p_i - y_i)^2$, where y_i denotes the true value of a house being above or below average (1 or 0), and p_i denotes the probability assigned to $p(y_i = 1|x_i)$ [3]. For a set of predictions, a lower Brier score, which can take values between 0 and 1, indicates better predictive performance. We see that the average Brier score is 0.12, indicating good performance prediction.

Naive Bayes Model

We fitted a Naive Bayes model to the data, using the features lot area, year built, number of full bathrooms, number of bedrooms, and house style as predictors, as they are all likely to impact the sale price of a house.

In choosing these features, we first investigated the importance of each feature for accuracy within the previous logistic regression model. We fitted the model a further 29 times, leaving one feature out each time. The estimated accuracy of the model in each case is displayed in Figure 5, although it is important to note that these values were calculated before 10-fold cross validation, due to limited computational power. Whilst not a definitive indication of feature importance, it can be suggested that estimated model accuracy decreases when leaving out features such as the number of bathrooms. We included both this feature and the number of bedrooms, since market research also suggests that the number of bedrooms and bathrooms in a house play a crucial role in determining its sale price [4].

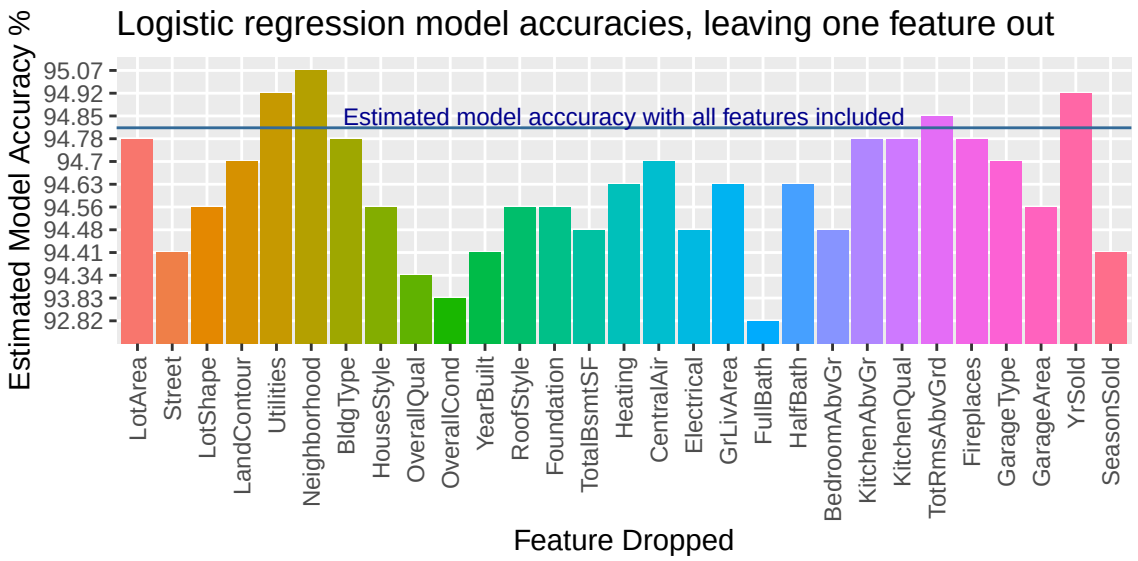

Figure 5: Estimated model accuracy when fitting the logistic regression model and leaving out one feature at a time.

In addition to this, we included the year built, lot area and house style. Figure 6 illustrates that, in the given dataset, there appears to be a higher density of newly built houses being sold for above average price, compared to older houses. Further, the results from the logistic regression in Figure 3 have already suggested a relationship between classification outcome and lot area, when all other features are controlled. Finally, it is intuitive that house style may impact the sale price of a house. For example, a two story house would be costly to build than a one story house, and would therefore have a higher value and selling price [5].

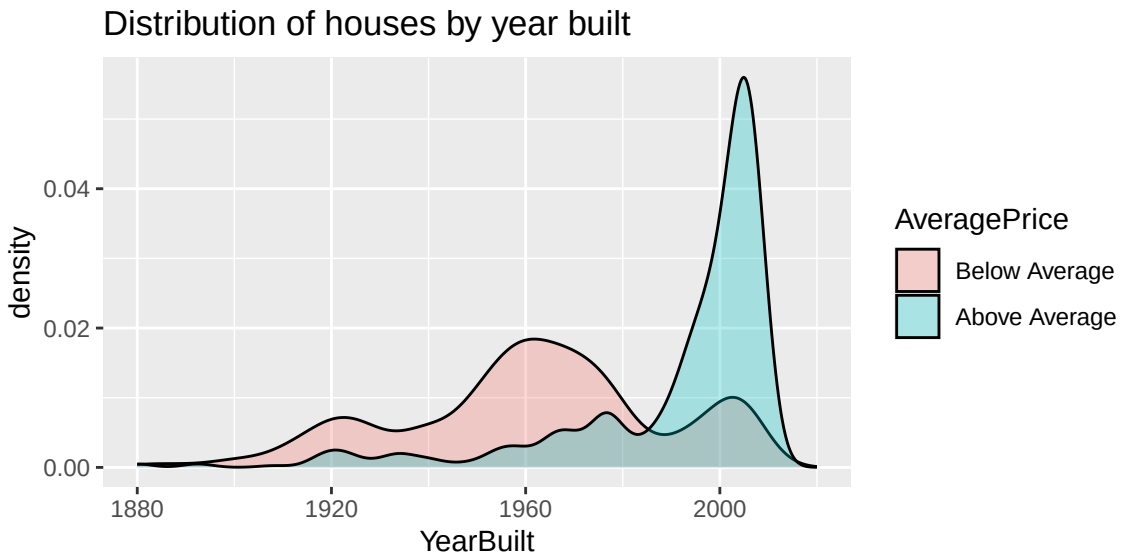

Figure 6: Density plot of houses sold below and above average price by year built.

We modelled these features independently, assuming no correlation between features. We treated the continuous features, $x_1 =$ lot area and $x_2 =$ year built, as Gaussian; the count features, $x_3 =$ number of full bathrooms and $x_4 =$ number of bedrooms, as Poisson; and the discrete feature, $x_5 =$ house style, as categorical (Multinoulli). We used the class conditional distributions of these features to predict the class, c, of the feature vector, $x = (x_1, x_2, x_3, x_4, x_5)$. The possible classes are being sold above, $c = 1$, or below, $c = 0$, average house price. The class conditionals were:

$$p(x \mid y = c) = \mathrm{N}(x_1 \mid \mu_{1,c}, \sigma^2_{1,c})\mathrm{N}(x_2 \mid \mu_{2,c}, \sigma^2_{2,c})\mathrm{Pois}(x_3 \mid \lambda_{3,c})\mathrm{Pois}(x_4 \mid \lambda_{4,c})\mathrm{Multinoulli}(x_5 \mid \theta_{i,c}), \quad c \in \{0, 1\}.$$

Each class c has 7 parameters, $(\mu_{1,c}, \sigma^2_{1,c}, \mu_{2,c}, \sigma^2_{2,c}, \lambda_{3,c}, \lambda_{4,c}, \theta_{i,c})$. We estimated the first six parameters, shown in Table 1, by computing the means and variances of their respective features in our dataset.

Table 1: Estimates of parameters of Normal and Poisson distributions.

	Below average	Above average
$\mu_{1,c}$	9090.19	13306.40
$\sigma^2_{1,c}$	22210933.77	227250079.71
$\mu_{2,c}$	1961.43	1991.62
$\sigma^2_{2,c}$	733.52	509.82
$\lambda_{3,c}$	1.35	1.96
$\lambda_{4,c}$	2.78	3.01

Since the house style feature is discrete, we estimated the parameter $\theta_{i,c}$ as the proportion of category i in class c in the given dataset. Table 2 shows the computed estimates for $\theta_{i,c} = \Pr(x_5 = i \mid y = c)$ for each possible category, i, in each class, c.

Table 2: Estimates of parameter for each category of Multinoulli distribution.

	Below average	Above average
2 Story	0.2239	0.4590
1 Story	0.5182	0.4648
1.5 Story	0.1501	0.0400
Split Foyer	0.0340	0.0038
Split Level	0.0621	0.0210
2.5 Story	0.0117	0.0114

These results suggest that as lot area and number of bedrooms and bathrooms of a house increase, the property has an increasingly higher probability of selling for above average price than for below. Figure 7 demonstrates a similar density distribution to what is seen in Figure 6. It suggests that the probability of a house selling for above average price is higher than for below for more recent builds. Table 2 highlights that houses which are 1 story, 1.5 story, split foyer, and split level are more likely to sell for below average price than above, while 2 story houses have a higher probability of selling for above average price. 2.5 story houses have similar probabilities of being sold either below or above average selling price, but we interpret this with caution, as this house style accounts for only 1.2% of the observations in the given dataset.

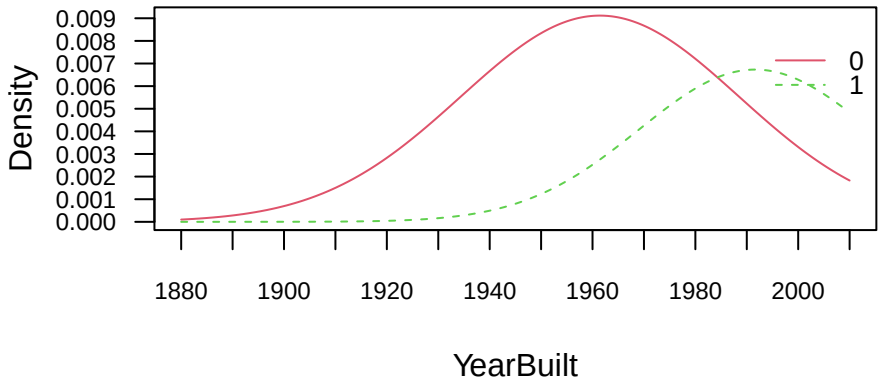

Figure 7: Density plot of houses sold below and above average price by year built based on the fitted Naive Bayes model.

Assessment of performance

We assessed the performance of the Naive Bayes model using 10-fold cross validation. We computed an accuracy rate for each fold as the proportion of the test data correctly classified, ranging from 68.35% to 76.81%. Taking the mean of these 10 rates, we found that our model has an overall accuracy rate of approximately 73.15%. These results indicate that our model performs relatively well at classifying whether a house should be sold below or above average price based on these predictors.

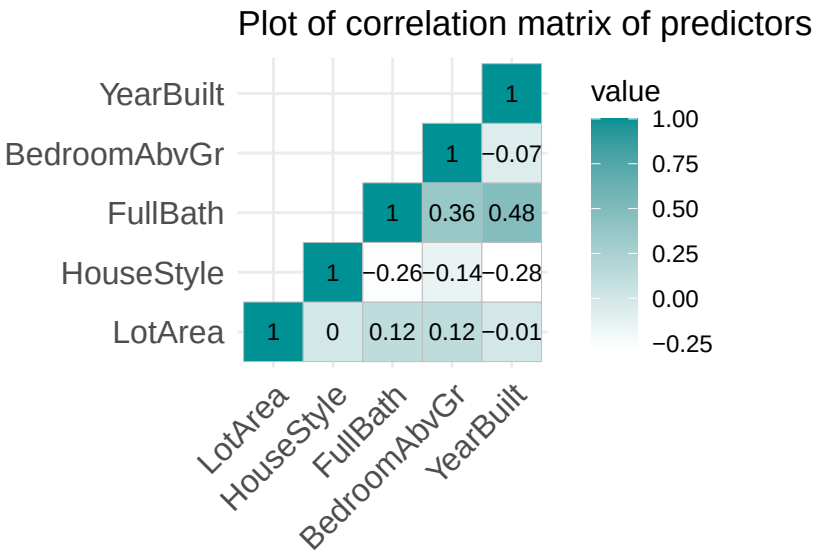

Figure 8: Plot of correlation matrix of predictors, showing that our assumption of independent features may not be valid due to moderate positive correlation between number of full bathrooms and both year built and number of bedrooms.

Figure 8 illustrates that there is moderate correlation between some of the predictor features, particularly between number of full bathrooms and both year built and number of bedrooms. As the Naive Bayes model assumes independence of features, this correlation may have contributed to a degree of inaccuracy within our model.

Conclusion

Our findings indicate that our models performed relatively well at classifying sales prices. Table 3 compares the average accuracy and Brier score of the logistic regression and Naive Bayes models.

Table 3: Summary of accuracy and scores after 10-fold cross validation.

Measure	Logistic Regression	Naive Bayes
Average Accuracy (%)	88.55	73.15
Average Brier Score	0.12	0.27

As expected, the logistic regression model performs better than the Naive Bayes model, as it uses all 29 available predictors. However, despite using only 5 predictors, the Naive Bayes model performs relatively well. It has an average prediction accuracy 15.4% below that of the logistic regression model. The difference in Brier score of the two models, 0.15, also indicates that the logistic regression model performs better.

In conclusion, our findings suggest that the models implemented could be used to predict whether a house will sell for above or below the average market price. The logistic regression model may have performed so well as it used a higher number of predictors, which allows for a more comprehensive and detailed understanding of the features that affect the sales price of a house. However, the Naive Bayes model still had a good predictive performance. Future research could be conducted to determine a combination of features to include which would result in an optimal balance between model complexity and prediction accuracy.

References

[1] The Advisory, *Definitve Guide: When is the Best Time to Sell Your House?* https://www.theadvisory.co.uk/house-selling/best-time-to-sell-house/#the-4-seasons-compared

[2] Statistical tools for high-throughput data analysis *Logistic Regression Assumptions and Diagnostics in R* http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

[3] Brier, Glenn W. (1950), *Verification of forecasts expressed in terms of probability,* Monthly Weather Review 78.1 https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[4] Property Road, *What Affects The Price Of A House?*, https://www.propertyroad.co.uk/what-affects-the-price-of-a-house/

[5] G.J. Gardner Homes, *Choosing your Home: Single Storey vs Double Storey Houses*, https://www.gjgardner.com.au/learn/choosing-your-home/single-storey-vs-double-storey-houses/#:~:text=When%20considering%20the%20cost%20of,single%20home%20of%20comparative%20size.