Project 2 Part 1 Report: Binary Classification of properties.

s1811292

2023-04-28

Executive Summary

Given a data set regarding properties in a city in the USA, with the information collected between 2006 and 2010 inclusive. We saught to determine if a statistical model could accuratley classify properties into two groups based on there sale price. With one group representing properties that sold for above the average price in the city, and the other group properties that sold for below average. We had information such as the size of the property, number of bedrooms, the year it was built and the neighborhood it was in. We subsequently analysed the data to determine trends or extract useful information from the data, that would be useful in predicting whether a property would sell for above or below average. We found that it is relatively easy to classify properties into the aformentioned groups. Even with relatively small amounts of information such as just the size, neighborhood and year built it models could achieve over 80% accuracy in classification. We could conlduce that the principle of building models to complete classificaiton of properties was valid and could easily be done; however we also cautioned against the expansion of such models to multiple cities without evaluating the whether assumptions we made are accurate in new cities.

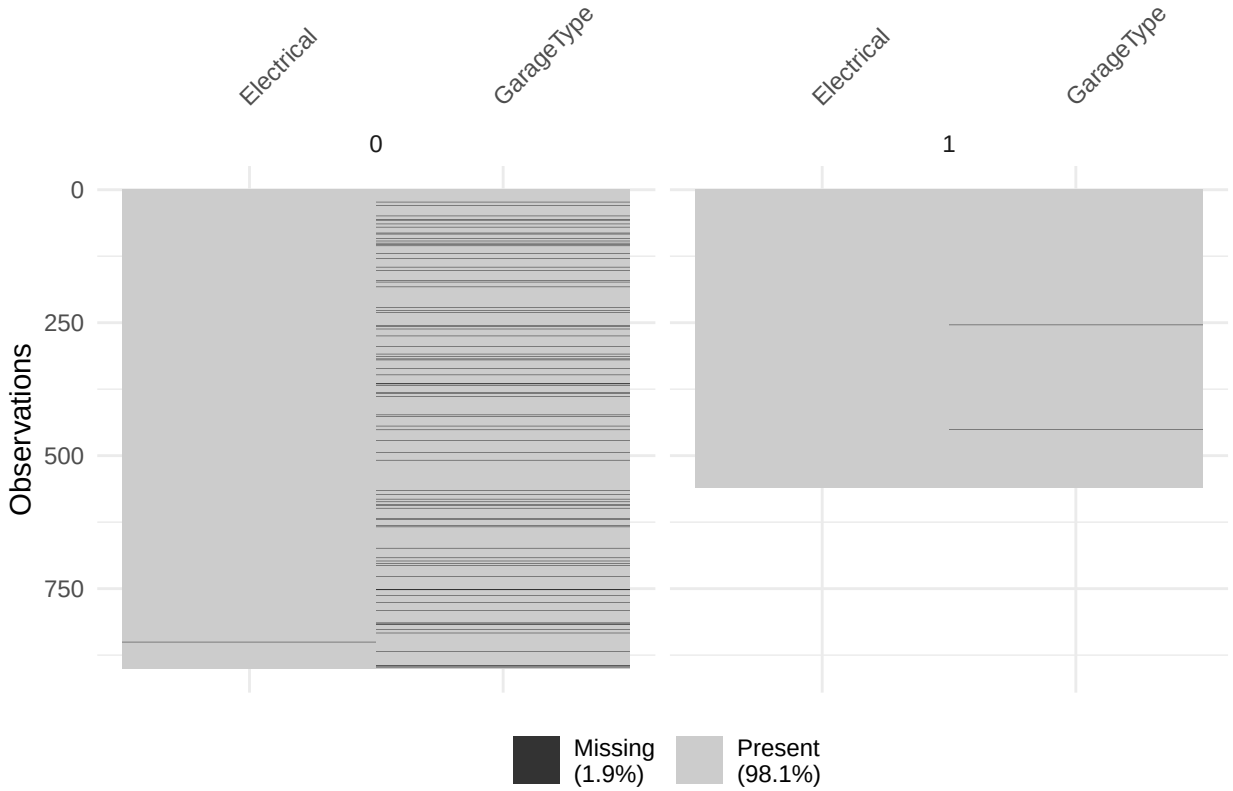
Introduction

We have been provided a data set concerning properties sold in a city in the USA over a 5 year period, from 2006 to 2010 inclusive. We have been asked to build a model to classify properties into two groups. Properties which sold for above the average price, and properties which sold for below average price. We shall use the mean prices from our data set to define the average price for this task. The goal is to produce a model that can accurately classify properties. To evaluate performance we shall use 10-fold cross validation. There are two models wich we shall evaluate, a Logistic Regression model using all the features in the data set; second a Naive Bayes Classifier, using 5 self selected representative features. One broad assumption we shall make is that the dataset is representative of cities of similar demographics and other variables, i.e. that it is not unique to this city only. Additionally we shall assume that market conditions in the housing market are stable across the 5 years, if conditions where unstable sale prices and distribution of the classes may vary year to year.

Methodology

The data contains 1460 records, with 31 variables or features, including one for an Id for each record and the sale price. So we have 29 features to make predictions from. As sale price is currently recorded on a contionus scale of its actualy sold price, we first had to convert this to discrete values to represeent two groups; using 0 to represent it sold for below average, and 1 to represent it sold for above average. First we had to inspect and clean the data. We had 82 missing values. However these were only confined to 2 variables. these were, Electrical, GarageType.

The graph below shows the missing data in the varibles by class, 0 representing sold below average prices, 1 representing sold above price.



There are multiple ways to deal with missing data. Considering Electrical, there is actually only 1 missing value. However electrical values are likely to depend on observed data e.g. year the house was built, overall quality etc. As a result this is likely missing at random (MAR), so we should use a regression imputation method, however for simplicity and due to the fact we have a large sample size we shall conduct complete case analysis and remove the record as it is only a single record. Garage type however is problematic, inspecting the values there is no value for no garage, and there are 81 missing values, there is a possibility that these missing values are actually representative of no garage. My approach was to recode these as a new value of "None", based on two facts: one that it is reasonable there would be properties with no garage in a representative sample in a city (our main assumption), two that the value of garage area is 0, and not missing, which supports this theory. We shall also encode our categorical variables as factors for analysis in R.

Now the data is cleaned, we shall examine the effect variables have on sale price, and how they differ in each class. For our logistic regression model this is was not as important, as we were going to use all the features, but it is still useful, so we can determine if there is any interaction between two variables. For the Naive bayes model this was more important was we are selecting just 5 features out of 29 to classify each property. So is important to select representative features. We are using 10 fold cross validation for evaluation, from this we shall inspect 2 values: the R^2 score and the training time, as a model is not really useful if it takes an impractical amount of time to train. I used R code to conduct the explortory data analysis on the data set, and other analysis. I then used python and sklearn, to build and evaluate the models, exporting the cleaned data set to python then exporting the evaluation results from python to R for discussion.

Analysis

The final data we had was practically the original data set, we only removed one observation, and no variables. As we are using 10-fold cross validation, we have 10 different sets of prediction data, and 10 different sets

of training data. This method gives us a good representation of the accuracy of the type of model, for the data and selection of features. As we are using 10 folds, we get 10 different models, what we want ideally is good average scores across the 10 folds, as if some folds preform Significantly worse than the others, then clealry there is an issue with the training and validation set used in that folds, which could indicate there is something we haven accounted for or a trend in the data we have missed. One main assumptions we made was the data set was representative.

Table 1: Table displaying Number of properties in each class in each Neihborhood (Class 1 = sold above average, Class 0 = sold below average

Neighborhood	Total number or properties	Class 1	Class 0
Blmngtn	17	10	7
Blueste	2	0	2
BrDale	16	0	16
BrkSide	58	6	52
ClearCr	28	21	7
CollgCr	150	98	52
Crawfor	51	30	21
Edwards	100	9	91
Gilbert	79	41	38
IDOTRR	37	0	37
MeadowV	17	0	17
Mitchel	49	8	41
NAmes	225	21	204
NoRidge	41	41	0
NPkVill	9	0	9
NridgHt	77	73	4
NWAmes	73	39	34
OldTown	113	6	107
Sawyer	74	1	73
SawyerW	59	28	31
Somerst	86	63	23
StoneBr	25	23	2
SWISU	25	3	22
Timber	37	30	7
Veenker	11	9	2

With caution the assumption that the data set is representative, can be accepted as valid, as each neighborhood has generally the same distribution in the two classes. If one or multiple had a completely skewed distribution as in the class split was vastly different to the other neighborhoods, then this assumption would require further analysis. One could attempt to build models for each neighborhood and compare perfomances to the overall model, however some neighborhoods could not be modelled to a great degree of accuracy as they have very few values in one class, and logistic regression will fail to converge if there is not values in each class. For example the Blueste neighborhood has only 2 properties in class 1 compared to 1555 in class 0. From this table above we can also see that neighborhood is not a great value in determining whether a Property sold for above or below the average price.

The second assumption can be evaluated by viewing the year sold variable as a time series, and plotting the average sale price across the 5 years, what we want is a stable trend.

Table 2: Distribution of classes by month sold across 5 years

Month Sold (over 5 years aggregated	Number sold Above average	Number Sold Below Average
1	79	40461
2	75	34490
3	143	78443
4	186	103446
5	274	151645
6	346	176921
7	322	171634
8	176	90112
9	95	44754
10	121	67884
11	115	64415
12	87	38926

Table 3: Distribution of classes by year sold

Year Sold	Number sold Above average	Number Sold Below Average
2006	434	231986
2007	460	233378
2008	417	220188
2009	468	253351
2010	240	124228

From the table and graphs above we can see that market conditions stayed stabel across the 5 year period,

additionally that the distribution of properties sold is consistent between properties that sold above and properties thatg sold below the average price. Below We also examine the effect of quality and condition on the sales price class.

Table 4: Condition, Quality and Kitched Quality for each class

Class	Overall Quality (1-9	Overall Condition (1-9	Most Common Kitched
0	5.361513	5.691880	TA
1	7.285714	5.389286	TA

From these plots we can see that properties quality is broadly the same in each class. However Condition is not, almost all properties that sold above average have a condition of 5, whereas properties that sold below average had a larger distribution, with 5, 6, 7 being the most common in decreasing order. As a result condition may be useful to use for the Naive bayes classifier, as it appears to distinquish the classses, better than quality anyway. Interestingly Quality seems to be more important in determining if a property sold for above average. As the averga quality for properties sold below is just over 5, and for properties sold above it si just over 7.

Results

Our models both performed very well, with Logistic Regression having a slighlty higher average performance, maximum performance and minimum preformace. The 5 features we selected for the Naive Bayes model were: Lot Area, Overall Quality, Year Built, Neighborhood and Number of rooms above ground.

Table 5: Logistirc Regression Results

Fold	Training Time	Validation time	Precision	Accuracy
0	0.1329811	0.0011840	0.9259259	0.9315068
1	0.2352593	0.0010560	0.8750000	0.9041096
2	0.6329420	0.0010171	0.9272727	0.9383562
3	0.4145281	0.0009837	0.8596491	0.8972603
4	0.1930518	0.0019510	0.8965517	0.9315068
5	0.0444469	0.0009899	0.9803922	0.9520548
6	0.9046462	0.0010679	0.8965517	0.9315068
7	0.3553784	0.0030637	0.8888889	0.9041096
8	0.8019741	0.0032468	0.8813559	0.9246575
9	0.8652239	0.0011361	0.8620690	0.9034483

Table 6: Naive Bayes

Fold	Training Time	Validation time	Precision	Accuracy
0	0.0043662	0.0032647	0.8600000	0.8630137
1	0.0038860	0.0029931	0.8727273	0.8972603
2	0.0036118	0.0018401	0.8225806	0.8904110
3	0.0024588	0.0016382	0.8333333	0.8630137
4	0.0020518	0.0017631	0.8363636	0.8698630
5	0.0022779	0.0018528	0.8813559	0.9246575
6	0.0024610	0.0018861	0.7758621	0.8356164
7	0.0021513	0.0014999	0.8627451	0.8698630
8	0.0019352	0.0014610	0.8196721	0.8835616
9	0.0019357	0.0014281	0.8148148	0.8482759

Discussion

As we showed in the analysis the assumptions we made were valid assumptions to make, and did not create significant bias in the results of the models. We can answer the question of can you accurately predict whether a property will sell for below average or above the average price; yes is the answer supprissing well actually, even a simple model with only 5 features (Naive Bayes) has strong and reliable preformance. The case could be made that despite have reduced performance compared to Logistic Regression, the Naive Bayes model is more reliable in its predictions. it has a lower variation in the precision scores. However it did have greater variation in its accuracy compared to Logistic Regression.

We can clearly observe that both the Logistic Regression model and the Naive Bayes models both preform very well, achieving precision and accuracy scores over 80% across all 10 folds. The precision scores are lower for both models. We can interpret this as overall the models classify around 90% correctly in to the right class. However there precision score is lower on average, which means its being over generous with which properties it classifiyes as sold above average, as it is classify around 80% of properties correclty as being sold above average.

In regards to training times, the Naive Bayes classifiers are far faster to train and validate across all 10 folds. This should be expceted at the algorithim is in general less computationally expensive, and we only used 5 features to train it; where as Logistic Regression has a higher compuational cost and we were working with 29 features.

In terms of model selection, you would have to side in favor of Logistic Regression; because its accuracy was always above 90%, and it did achieve precision of 98/ in one case. Further is validation time, i.e. how

long it took to predict results, was only a single order of magnitude larger, in general it was in the order of 0.001, compared to 0.0001 for Naive Bayes. In the context of dealing with large sums of money and property purchases, you would want prioritse precision and accuracy over training time.

Flaws

Furthermore, the reason for preference for Logistic Regression, as there is a variety of methods and variations of the algorithm we can use to refine and improve results, such as regularization and accounting for interactions between features. We may also seek to evaluate Atkin information criteria in a more in depth model selection to find a simpler model, which requires fewer features to predict. As it may not be cheap of easy to obtain the vast majority of information we had in the data set, and the vast majority of reatures would be either missing at random or missing not at random depending on how many variables we can collect observations for.

One potential flaw in the data we had was that we had an imbalance in the classes; as we had a larger number of properties that sold for below average, than sold for above average. Expanding on that point, we made and concluded it was a valid assumption that this data set was representative of a typical city, or more specifcally it was representative of cities with similar demographics and variables. We can be confident that our models will perform well in this city and in similar cities. We should not be confident, that our models would generalise to say all cities in a given coutnry such as the USA, as some cities are far greater division in terms of socioeconmic status. For example some cities may have outlier properties such as a few properties which sold for several million, while most properties in the city are in the scale of hundreds of thousands. Then the mean would be heavily skewed. In summary we would need to evaluate certain socioeconomic and demographic information about a city before we could confidently use our models.

Conclusion

We have conducted an effective analysis of the data we had avalible, being selective with which features we conducting in depth analysis of. We also valdiated assumptions we had. Moreover we successfully built and evaluated models which were accurate and good models to solve the task at hand of classification of properties as to whether they would sell for below or above the average price.