COMP6208 Team Mai$on Data Exploration

Sameen Islam *si1u19@soton.ac.uk* Ian Simpson *ijs1c20@soton.ac.uk* Zuzanna Skorniewska *zas1g17@soton.ac.uk*

***Abstract*—This report outlines COMP6208 Team Mai$on's exploration of the Ames, Iowa Housing Dataset with a view to next stage of the project, modelling and prediction of house prices.**

I. Introduction

HOUSE price prediction is the aim of the project, within the dataset of 1,465 observations and 80 variables on houses sold in Ames, Iowa between 2006 and 2010 (hereafter known as the 'Ames dataset'). Data exploration is the focus of this report, which lays the foundation for modelling and prediction which will be the subject of a later report.

This report is laid out roughly in the order in which the group tried to logically approach the subject. First, the data itself was examined. Types of variable were understood, empirical distributions examined, and quality issues (such as data not matching the data provider's schema) identified and resolved, which led to a quality-assured dataset with practical encodings. Secondly, statistical and machine learning techniques were leveraged to understand feature importance which should be of direct aid in selecting (and eliminating) variables for modelling.

II. Data

A. Overview

Data on the properties of all houses sold between 2006-2010 from the Ames City Assessor's Office was collated by De Cock [1] in 2011. Kaggle simplified the data in 2017 for use in their House Prices competition [2], by removing 33 'overly technical' features and by retaining only one record for the hundred or so houses sold more than once during the observation period. This resulted in 80 features on 2,930 observations with a random 50% 'training' and 'test' split. However as $SalePrice$ labels are not provided on the 'test' set, since the aim of Kaggle's competition is to submit predictions of this label, it was determined to be incompatible with the group's aims and so Kaggle's 'training' set was adopted as the whole - this we refer to as the 'Ames dataset'. Thus, data exploration was performed on this 1465×80 dataset, and this data will be split into training/test by the group for the modelling report.

Plain text descriptions of labels are available at [2].

B. Empirical Distribution

Histograms and kernel density estimates were generated for each feature in order to gain insights into their distribution. For example, house prices in the Ames dataset appear to tend to a lognormal distribution, as shown in Fig. 1. Approximately half of the observations lie in the range \$100-200k, and so models trained on this data will have better performance in predicting houses' value in this price range and may not be so accurate for value prediction of house prices in the tails, such as these that are valued at more than \$300k. Whilst our test dataset ought to be homogeneous in distribution, insufficient sample size for extreme values and model bias towards the mean will be watch-outs.

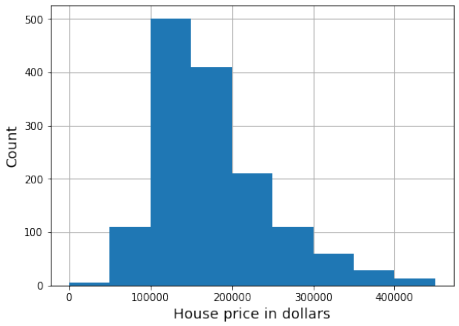

Fig. 1. The distribution of house prices in the Ames dataset is skewed with a long right-tail. Sample size in the tails, and bias towards the central mass, are examples of issues that will be faced when modelling.

C. Feature Encoding

Non-numeric features needed to be encoded numerically for exploration and modelling, and this was an important activity since 47 out of 80 features required some encoding.

34 features were converted to ordinal. For some features this was trivial: $GarageCond$ had values Poor/Fair/Average/Good/Excellent which were mapped to $0-4$. For other features, deeper inspection and consideration was required: $Electrical$ had values Breaker/FuseA/FuseF/FuesP/Mix which, after short research on US electrical standards led to a $3/2/1/0/2$ mapping.

13 features were determined to be categorical, since no ordinality could be identified. For example, $Neighbourhood$ identifies the area of the city in which the property is located. Such features were dummy encoded, and this was responsible for an increase in dimensionality from 80 to 194. Whilst some algorithms such as random forest handle dummy features adequately, these features had to be excluded from most other algorithms, and this highlighted to the group the shortcomings of categorical features and how desirable it is to encode them ordinally where possible.

D. Missing Values

6,965 values ($\approx 5\%$ of all values) were blank. Sometimes this purposefully meant the lack of the feature in question, e.g. no fireplace; but other times this was due to missing data. Filling in blanks was thus performed on a per feature basis. For ordinal features, the mode occurrence was used to fill blanks, and for continuous features, the mean was used.

III. Feature Importance

A. Rationale

Methods generally increasing in sophistication were applied to understand feature importance in the dataset, a summary of which is provided in Table 1.

B. Method: Pearson Correlation Coefficient

The Pearson correlation coefficient [3] provides a measure of the strength of linear association between two variables, yielding a coefficient $\in [-1, 1]$ such that the square of the coefficient measures the proportion of variance accounted for, and the sign indicates the direction of the relationship [4]. Many insights were gained, of which a selection is discussed.

14 features showed at least moderate correlation ($|corr| > 0.5$) with $SalePrice$, the highest being $OverallQual$ (0.79), $GrLivArea$ (0.71), and $ExterQual$ (0.68).

To understand the relationships within the whole dataset, and not just with $SalePrice$, the statistic was also computed for every combination of non-categorical variables. However, analysis of the resultant correlation matrix was challenging due to the large number of variables. In order to simplify matters, only variables which had at least one correlation of magnitude > 0.5 were retained, resulting in a smaller and more interpretible matrix. An excerpt is shown in Figure 2, and inspection of the rightmost column indicates that $OverallQual$, $YearBuilt$, $YearRemodAdd$, $ExterQual$, and $BsmtQual$ all show moderate-strong correlations with $SalePrice$. The central area also shows that all of these features also correlate with each other. These results make intuitive sense: exterior quality and basement quality assess the quality of major components of a property; older properties tend to be of lower quality, perhaps at least due in part to the opportunity for deterioration over the passage of time. $YearRemodAdd$ denotes the year a property was remodelled, however it is given the same value as the construction date if no remodelling has occurred. Since 764 (52%) of properties have not been remodelled, high correlation between these two variables is to be expected.

Translating these insights back to the ultimate task of predicting $SalePrice$, this analysis showed that whilst $ExterQual$ has the third highest correlation with $SalePrice$, it may be a weaker predictor in the presence of $ExterQual$.

C. Method: Linear Regression (Exhaustive Search and Lasso)

Multiple linear regression is typically used as a predictive tool, but by adding and removing terms it can cast light on the significance of one variable in the presence of others and thus can be used for exploration.

An exhaustive search was conducted whereby all possible model combinations of 5 or fewer terms (plus a constant term) were fit, and the R^2 recorded. Whilst for prediction, such a search has low statistical power (without any validation step), it was deemed to be suitable for exploration as feature importance is being measured in aggregate. Figure 3 shows the top 100 models by R^2. The highest performing model utilises $OverallQual$, $ExterQual$, $BsmtFinSF1$, $GrLivArea$, and $GarageCars$. Frequency of feature occurrence in the top 100 models was chosen a feature importance metric, the results of which are also shown in Figure 3.

Lasso regression [5] applies L1 regularisation to linear regression to achieve a sparse model. This was far quicker to run and agreed with exhaustive search on three top 5 features.

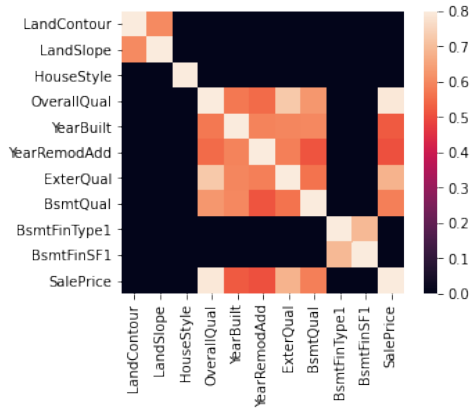

Fig. 2. Example: understanding Pearson correlation by focusing on a subset of variables, and showing only pairs with $|corr| > 0.5$

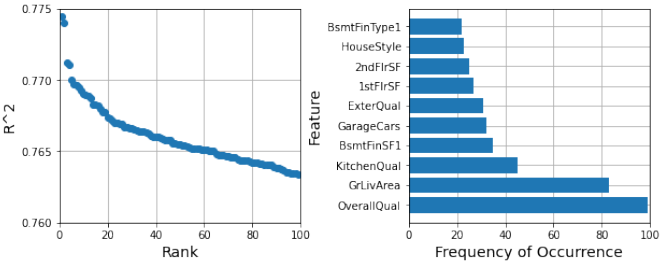

Fig. 3. Linear Reg. search: top 100 models with 5 or fewer terms by R^2. Distribution by R^2 illustrates the methodology. Frequency of feature occurrence in the top 100 models was used as a measure of feature importance.

D. Method: Non-negative Matrix Factorisation

Non-negative Matrix Factorization (NMF) decomposes a matrix, V, into WH whereby rows of H are non-negative factors of V [6]. For application of NMF to the data, matrix V was constructed using non-categorical features and excluding target variable, $SalePrice$. To aid subsequent identification of important features, V was column-wise standardised, truncated at $\pm 5 s.d.$ and then 5 added to all entries in order to make V non-negative.

NMF decomposition of V of rank 8 yielded a 8×66 matrix of factors, H and a 1095×8 matrix of weights, W. The correlation of columns of W against the target variable, $SalePrice$, was calculated. The strongest factor had a negative correlation with $SalePrice$ of -0.42. Inspection of the elements of the factor showed it gives strongest weighting to $HouseStyle$, $2ndFlrSF$, and $HalfBath$. The second strongest factor had a positive correlation with $SalePrice$ of 0.36 and gave strongest weighting to $BsmtFinSF1$, $BsmtFinType1$, and $BsmtFullBath$. Although the results were meaningful, they were not significant, and are documented for completeness.

E. Method: Principal Component Analysis

Principal Component Analysis (PCA) [7] is a popular tool used for feature analysis by means of a set of principal components – directions contributing to the most of data's variance. For our analysis, PCA was applied to a trimmed dataset with only non-categorical features. The first 3 principal components were found to explain 47.9%, 21.8% and 14% of the total data variance, thus accounting altogether for 83.7% of the total data variance. The loading plot in figure 4 shows top 10 features that influence the first two principal components the most. In general, none of the most influential features show to be inversely correlated. In particular, we can see that *BsmtFinType1* strongly influences PC2, whereas *YearRemodAdd* strongly influences PC1. Moreover, sets of features relating to common houses' properties (e.g. a basement or a garage) such as (*BsmtFinType1*, *BsmtExposure*) or (*GarageFinish*, *GarageYrBlt*, *GarageType*) can be seen to be strongly collinear (denoted by similar vectors' direction).

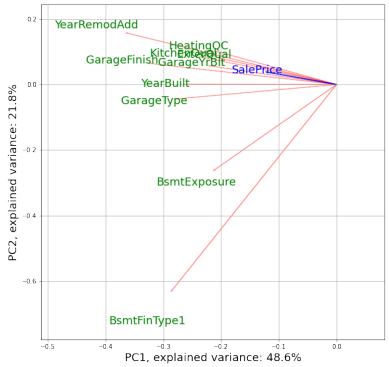

Fig. 4. PCA loading plot showing top 10 features that influence first two principal components the most.

The vector corresponding to the target feature (SalePrice denoted in blue in figure 4) implies collinearity between features relating to garage's specifications, *YearRemodAdd* and *YearBuilt*. We can hence expect these features to be relevant in predicting houses' prices.

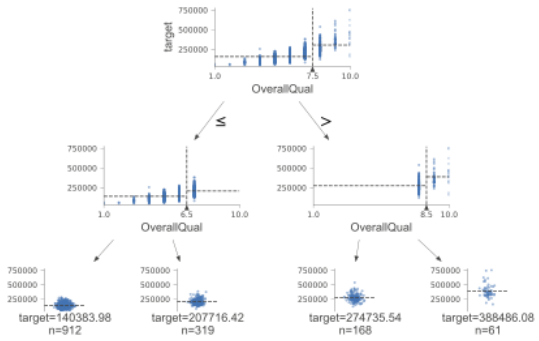

Fig. 5. Decision tree with $d_m = 2$ showing *OverallQual* is the most important feature for determining *SalePrice*. The graphs in each node of the tree show splitting of subsets (input features on x-axis and target feature on y-axis). The dashed line in the leaves show the target prediction of this model.

F. Method: Decision Tree

Formally, let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be a dataset such that each $x_i \in \mathcal{X}$ is the i-th example with a corresponding target feature y_i. A decision tree model [8] recursively produces subset $S_i \in \mathcal{X}$ according to some feature threshold f_α corresponding to a binary decision based on a feature that produces the lowest target variance, σ_y^2. This gives rise to a tree whose maximum depth is a hyperparameter $d_m \in \mathbb{N}$. Empirically, we found decision trees are the most interpretable when $d_m \leq 3$.

The basic idea behind using this model to perform feature importance is that features selected to determine f_α when $d \to 0$ will maximally reduce σ_y^2, therefore, meaning their relative importance is high. It is interesting to observe that a model of $d_m = 3$ produces a model depicted by figure 5, where we find all relatively important features relate to the size of a property. This corresponds well with human intuition, which tells us that generally, a more expensive property should be 'well-polished' and large. Conversely, a cheaper property may lack glamour and be relatively small. An analysis of feature importance in this way produced three features with the highest level of relative importance: *OverallQual*, *GrLivArea* and *2ndFlrSF*.

G. Method: Random Forest

Random forest [9] is an ensemble learning technique which uses many 'weak learners' together to create a stronger model. By averaging the prediction of individual decision tree regressors, we can further lower target variance. A collection of decision trees are thus named forest. Each tree is constructed from the training set using bootstrap sampling (sampling without replacement). The precision of these important features hinges on model quality; a model with poor accuracy cannot be trusted. Since we have no access to the test set in this phase of the project, we use an out-of-bag (OOB) score to *estimate* model accuracy on test set. The feature importances extracted for this task had an OOB score $= 86.5\%$ which indicates a relatively accurate model.

H. Method: Random Forest - Permutation Feature Importance

It is known that the feature importance computed with random forest is prone to importance bias on features that are numerical or has high cardinality. Furthermore, feature importance values given for a model learnt from a training set can be different from test set due to capacity to overfit. Therefore, we apply permutation importance, which considers the reduction in mean squared error (MSE) when a particular feature is not available. Indeed a model cannot be trained without a feature being present, so the feature is kept, but values are replaced with noise that originate from the same distribution as original feature values, which is obtained by shuffling the values from a particular feature.

I. Method: Shapley Values

Using Game Theory, SHAP [10] provides with us a quantitative explanation behind each prediction made. More specifically, it gives us a quantity by which each feature contributes

TABLE I
TOP 5 FEATURES FOR PREDICTING SALEPRICE, BY MODEL

Feature	Pearson Corr.	LR (Search)	LR (Lasso)	Decision Tree	Random Forest	Perm. Imp.	GBM	XGB
OverallQual	1	1	2	1	1	1	1	1
GrLivArea	2	2		2	2	2	2	
ExterQual	3		5					3
KitchenQual	4	3	4					
GarageCars	5	5	3		5		3	2
TotalBsmtSF					3	3	5	5
BsmtQual			1					4
BsmtFinSF1		4				4	4	
2ndFlrSF				3	4	5		

to making a prediction, which we use as an indicator of its relative importance. Figure 6 provides a summary of important features extracted with this method where we find low *GrLivArea* and low *TotalBsmtSF* are important features for reducing property price. On the other hand, high values of *OverallQual* and *1stFlrSF* are important features for driving property prices up.

J. Method: Gradient Boosting

As is the case with random forest, gradient boosting [11] is an ensemble learning method. Being a boosting technique, it expands the model by adding weak learners sequentially with an objective to minimise the chosen differentiable loss function. For comparison and generality, two types of gradient boosting models were used: the general gradient boosting model, provided by sklearn optimised on least squares loss function, and XGBoost. XGBoost differs from common gradient boosting method as it relies on 1st and 2nd order derivatives of the loss function and for its advanced use of regularisation. For both cases, we searched for the 10 most relevant features in predicting houses' prices. Results for the top 5 features are shown in column 7 and column 8 in Table 1.

In general, we found that both XGBoost and Gradient Boosting take *OverallQual* of the house as the primary measure in predicting houses prices. However, Gradient Boosting was observed to rely on this single feature rather heavily, attributing it nearly with 0.5 out of the total unity of importance measure. On the other hand, XGBoost distinguished 3 features that it found roughly semi-important: *OverallQual*, *GarageCars* and *ExterQual*. *GarageCars* can be also found in the top 3 of most important according to Gradient Boosting, however *ExterQual* is not even among the 10 most important features. Overall both of these models agree on 7 out of 10 most important features. In summary, this insight shows us that *OverallQual* is probably one of the most relevant features for predicting houses' prices. Besides this feature, we may expect *GarageCars*, *ExterQual* and *GrLivArea* to contribute rather significantly to a prediction of houses' prices.

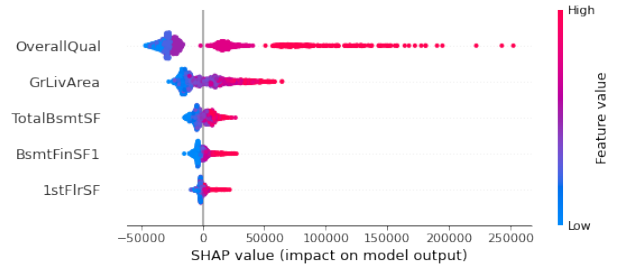

Fig. 6. Features ranked according to SHAP values. Red points have high feature values and blue points have low. A feature with high SHAP contributes to a higher SalePrice while a lower value have the opposite effect.

K. Comparison of Results

Table 1 shows the top 5 features identified by model (except PCA and NMF which give more general insights). The eight models yield a total of nine unique such features, indicating some consensus. $OverallQual$ was in top 5 for all models, and $GrLivArea$ was in top 5 for all but two models, and are likely to be very important in modelling. Taking the top 10 features, the models start to diversify in their views, and there are 28 unique features. Ensembling of these models is likely to be of interest in the modelling report.

IV. CONCLUSION

This report outlines exploration of the Ames dataset. Firstly, the data was encoded, blanks were filled, and the empirical distribution of features plotted and examined. Secondly, *Pearson Correlation Coefficient, Linear Regression (Search and Lasso), Decision Tree, Random Forest, Shapley* and *Gradient Boosting* were used to identify relevant features, from which a set of 8 generally important features was identified, within which $OverallQual$ and $GrLivArea$ are very important. An set of an additional 22 relevant features was also identified. Most of the model types used for data exploration will be able to be used for the prediction task, and the models may even be able to be combined through ensembling.

REFERENCES

[1] D.De Cock, *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*, in Journal of Statistics Education, Vol 19, No 3, 2011. https://doi.org/10.1080/10691898.2011.11889627
[2] Kaggle (URL) https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[3] K Pearson, *Regression, Heredity, and Panmixia*, in Philosophical Transactions of the Royal Society of London, 1896 https://doi.org/10.1098/rsta.1896.0007
[4] Armstrong, RA. *Should Pearson's correlation coefficient be avoided?* Ophthalmic Physiol Opt 2019; 39: 316– 327. https://doi.org/10.1111/opo.12636
[5] Tibshirani, R. *"Regression Shrinkage and Selection via the Lasso."* Journal of the Royal Statistical Society. Series B, Vol 58, No 1, 1996 https://www.jstor.org/stable/2346178
[6] D. Lee & H. Seung *Learning the Parts of Objects by Non-negative Matrix Factorization*; Nature Vol 401, Oct 1999 https://doi.org/10.1038/44565
[7] K Pearson *"On lines and planes of closest fit to systems of points in space"*, London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901 https://doi.org/10.1080/14786440109462720
[8] Kamiński B, Jakubczyk M, Szufel P. *"A framework for sensitivity analysis of decision trees."* CEJORS 2018;26(1):135-159. https://doi.org/10.1007/s10100-017-0479-6
[9] Tin Kam Ho *"Random decision forests"* Proceedings of 3rd ICDAR, pp. 278-282 vol.1 https://doi.org/10.1109/ICDAR.1995.598994
[10] A Roth, *"The Shapley Value"* https://doi.org/10.1017/CBO9780511528446.002
[11] C. Li *"A Gentle Introduction to Gradient Boosting"* College of Computer and Information Science Northeastern University. http://www.chengli.io/tutorials/gradient_boosting.pdf