Contents

[Introduction: 3](#_Toc8488918)

[Data Preparation: 4](#_Toc8488919)

[ Handling missing data 4](#_Toc8488920)

[ Standardizing the data 5](#_Toc8488921)

[ Data Scaling 5](#_Toc8488922)

[ Splitting of data 5](#_Toc8488923)

[Predictors: 6](#_Toc8488924)

[a) Linear Regression: 6](#_Toc8488925)

[b) Ridge regression: 7](#_Toc8488926)

[c) Lasso regression: 8](#_Toc8488927)

[Evaluation 9](#_Toc8488928)

[a) Linear Regression 9](#_Toc8488929)

[b) Lasso Regression: 10](#_Toc8488930)

[c) Ridge Regression: 10](#_Toc8488931)

[Conclusion 11](#_Toc8488932)

[1. Key Difference Predictors: 11](#_Toc8488933)

[2. Regression Results 12](#_Toc8488934)

[References: 12](#_Toc8488935)

**Table of Figures.**

[Figure 1: Feature comparison 4](#_Toc8488936)

[Figure 2: Formula for regulating the information 5](#_Toc8488937)

[Figure 3: Splitting training and test data 5](#_Toc8488938)

[Figure 4: Correlation matrix 6](#_Toc8488939)

[Figure 5: OveralQual vs SalePrice 7](#_Toc8488940)

[Figure 6: GrLivArea vs SalePrice 7](#_Toc8488941)

[Figure 7: Coefficient estimate formula 8](#_Toc8488942)

[Figure 8: Lasso L1 penalty 8](#_Toc8488943)

[Figure 9: Probability plot 9](#_Toc8488944)

[Figure 10:Optimal number of features 10](#_Toc8488945)

[Figure 11: Different alpha value 10](#_Toc8488946)

[Figure 12: Coefficient estimate for B using ridge regression 11](#_Toc8488947)

[Figure 13: Feature graph 11](#_Toc8488948)

[Figure 14: Coefficient in the model 12](#_Toc8488949)

**Property price prediction using machine learning classifiers**

# Introduction:

The given problem is a supervised learning problem in which we use three regression predictors namely Linear, Ridge and Lassos regression to predict the sale of house prices. We have been provided with a dataset which has several features based on which we can predict the sale price of the house.

Regression is a supervised learning method which allows us to examine the relationship between two or more variables of interest. Regression predictive modelling is the task of approximating a mapping function from input variables to a continuous output variable. A constant output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes. For example, a house may be predicted to sell for a specific dollar value. We can use a regression model to determine its selling price.

Regression is performed on a dataset which is already classified, wherein each item in a dataset is assigned to a category. In the dataset that we have been provided, these categories are represented in columns. There are 79 columns or features of the house with which we predict the sale price of the home.

We performed three types of regression on the dataset provided which are as follows:

a) Linear regression (multiple regression) – which we used to predict the value of one dependent (SalePrice) variable with multiple interdependent variables (GrLivArea and OverallQual). The Linear regression model was found to be the best performing model for determining the sale price value of the house with a rise of 0.0178 and r-square value of 0.88824.

b) Ridge regression – which we used to alleviate multi-collinearity amongst the predictor variables (Alley, MiscFeatures, fireplace, Fence, PoolQc, frequency, MSSubClass, MSZoning, LotFrontage, LotArea, Street, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, YearBuilt, YearRemodAdd, LotFrontage). On performing the Ridge regression model on the given dataset, we obtained a rmse value of ‘0.14063682725654697.’

c) Lasso regression – which we used to optimize our prediction by shrinking the coefficients to 0. ’Condition2\_PosN’ had the most significant negative factor and ‘MSZoning\_RH’ had the highest positive impact out of the 284 variables that we derived. On performing the Lasso regression model on the given dataset, we obtained an rmse value of ‘0.02104436592084776’.

We have split the dataset into 75% training data and 25% testing data for Linear regression, 70% training data and 30% testing data for Ridge and Lasso regression.

# Data Preparation:

* Handling missing data**:** In the dataset provided, there are few missing features. It is found that there are 33 missing values for the features offered in the dataset. We have handled the missing data for the three regression models as follows:

1. Linear regression: In Linear regression model, we can process the missing data in a couple of ways, such as replacing the data with the highest correlated feature or by simply dropping the element if another active function is available for prediction. In our case, we have lowered the missing data as we have other stable features such as “GarageBand”.
2. Ridge regression: There are null values present in the training data and test data. From figure 2.1, we can observe an invalid feature ‘electrical’ and almost null values for elements, ‘MasvrnArea’ and ‘MasVnrType’. Features such as Alley, MiscFeatures, fireplace, Fence, PoolQc and LotFrontage had significant benefits. Therefore we select these features to predict the sale price of the house.

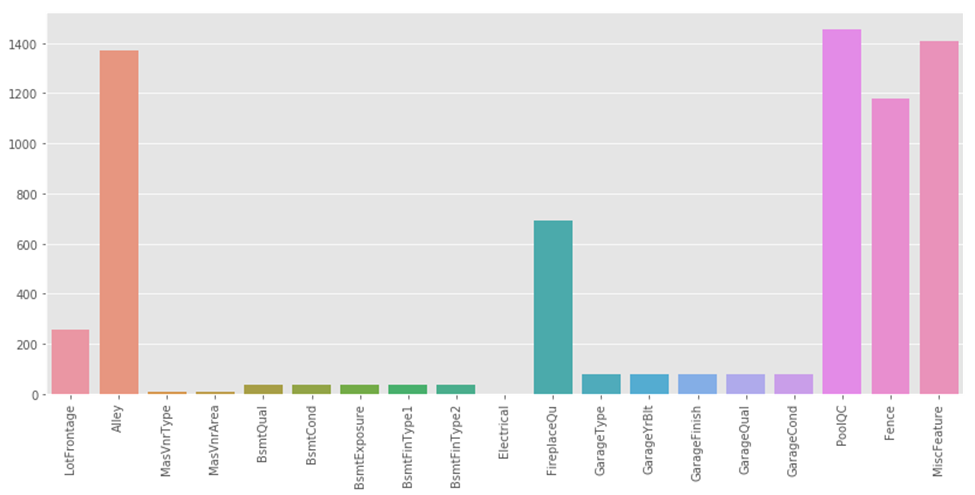


Figure : Feature comparison

1. Lasso regression: If we look at the features carefully, we can find redundant or highly correlated features. We can use just one of the correlated features to avoid redundancy. For example, ‘Garagecond’ and ‘GarageYrBlt’ are correlated features. We can use only ‘GarageCond’ as the variable for prediction.

* Standardizing the data**:** To normalize the data, we should always be aware of outliers because it can severely affect our models. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Standardizing of data is very important because most of machine learning models will converge much faster if the features are on the same scale.

We want to establish a threshold that defines observation as an outlier. To do so, we standardized our data by converting data values to have a mean of 0 and a standard deviation of 1. The formula for standardizing the data is as follows:

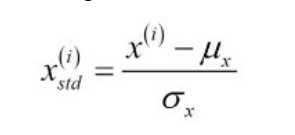
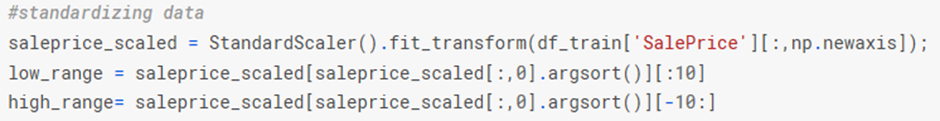


Figure : Formula for regulating the information

We compute the mean and standard deviation of a particular feature and apply it in the above formula for each of our observations.

We used the following function in the program to standardize the data:



* Data Scaling**:** Data scaling is also known as feature scaling or data normalization. It is a method used to regulate the range of independent variables or features of data. The pre-processed data might contain attributes with a mixture of scales (values) for various quantities. Data scaling is a process where we rescale the benefits of some features so that we get a standard normal distribution with a mean of 0 and a standard deviation of 1.
* Splitting of data**:** The next step is to break the dataset into training data and test data. As mentioned earlier, we have divided the dataset into 75% training data and 25% testing data for Linear regression, 70% training data and 30% testing data for Ridge and Lasso regression. Splitting of data is necessary so that we can estimate the predictive power of our model by predicting the unseen data (test data).

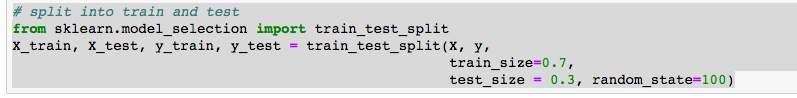


Figure : Splitting training and test data

We used the above code to split 70% of our data to the training set (x\_train and y\_train) and the remaining 30% to the test set.

Predictors:

1. Linear Regression: Since we have more than one independent variable/features, we used the formula for multiple linear regression, which is

**Y = α0 + α1X1 + α2X2 + α3X3 + … + αnXn**

where Y is the dependent variable,  
X1, X2, X3… are independent variables  
and α is the coefficient of regression.

From the values of α, we can easily understand which features or independent variables are playing a vital role in predicting the value of Y. The α value can be determined from a correlation matrix which is shown below.

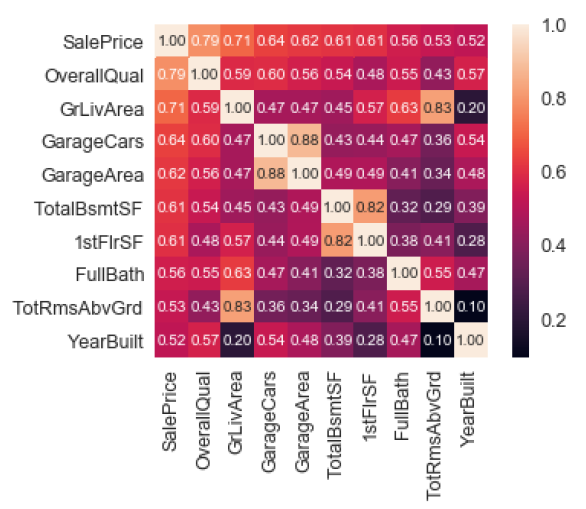


Figure : Correlation matrix

From the above picture, the variables most correlated with 'SalePrice' are as follows:

* 'OverallQual' and 'GrLivArea' are strongly correlated with 'SalePrice'.
* 'GarageCars' and 'GarageArea' are also some of the highly correlated variables. However, 'GarageCars' and 'GarageArea' have redundant features. Therefore, we only consider one of these variables in our analysis (we can keep 'GarageCars' since its correlation with 'SalePrice' is higher than that of ‘GarageArea’).
* 'TotalBsmtSF' and '1stFloor' are also highly correlated with one another. We use 'TotalBsmtSF' in our analysis.
* It seems that 'YearBuilt' is slightly correlated with 'SalePrice'.

Since 'OverallQual' and 'GrLivArea' play a significant role in determining the sale price, we plot the graph with these two features.

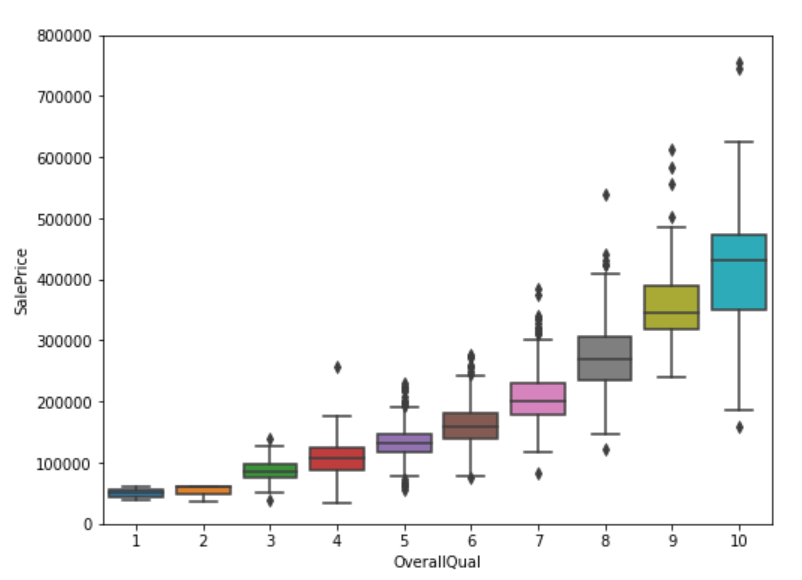


Figure : OveralQual vs SalePrice

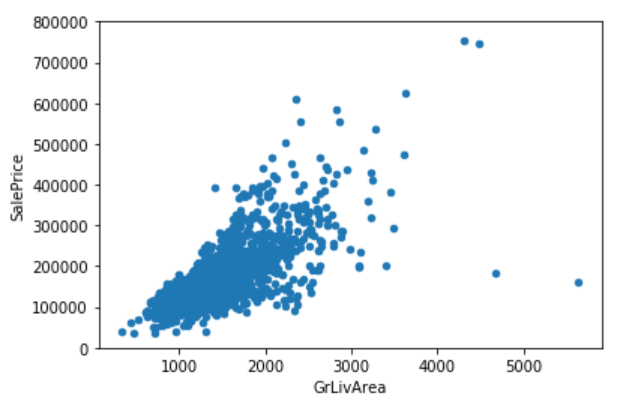


Figure : GrLivArea vs SalePrice

We conclude that sale price is strongly correlated with ‘GrLivArea’ and ‘OverallQual’, which means that where the overall quality and the living area is good, the amount of houses is at the highest.

1. Ridge regression: The value of lambda which is the coefficient of regression is initially set to 0. Since the Ridge regression from ‘learn’ utilizes alpha to imply lambda, we have substituted that value here. We replaced the alpha value from the range 0.00001 to 2. When the alpha value increases, the line turns out to be progressively less fitting it, and it moves towards a catchline. Therefore when the alpha value is 0.00001, the track is the best fitting. We utilized the ‘RidgeCV’ bundle from ‘learn’. This enabled us to create several alpha values and feed them in the ridge model. Based on the ‘RidgeCV’s scoring parameter of neg\_mean\_squared\_error, we can determine which alpha we should use after the model has finished running. Return ridges. Alpha provides the best estimation of alpha for our model.

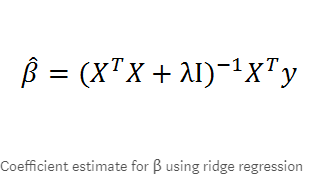
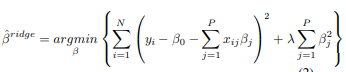
 

Figure : Coefficient estimate formula

Here λ ≥ 0 is an unpredictable parameter that controls the level of regularization. Plainly, λ = 10 compares the loads of β which is near 0 while not approaching to 0. The minimization estimation of βˆridge was found to be 0.328.

1. Lasso regression: In the dataset, we have 79 which describes the various aspects of a residential house and out of these, 36 are numerical, and 43 are categorical variables. There is 1460 example provided in training set and 1459 models in the test set. From the figure ….. we can see the correlation between the 79 variables. The most exciting thing to get our attention is the four coloured squares, the ’2ndFlrSF’,’1stFlrSF’,’ WoodDeckSF’ and ’GarageArea’ variables. They are so highly correlated which means that they give almost the same information in predicting the sale price of the house. Therefore during data pre-processing, we keep just one of these variables.

Lasso is also a shrinkage method like a ridge, but with a little difference, it uses the L1 penalty like the following:

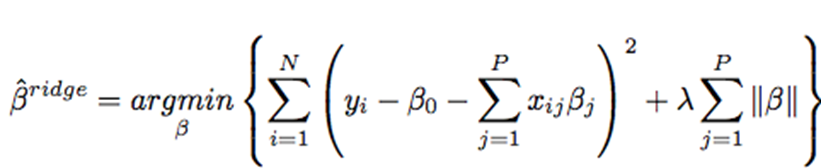


Figure : Lasso L1 penalty

Lasso regression uses the total sum of the coefficients as the penalty item. Suppose that we have two variables in our model, the substantial green area is the constraint regions (|β1| + |β2| ≤ t and β 2 1 + β 2 2 ≤ t 2) and the red ellipses are the contours of the least squares error functions. In our dataset, we have 79 variables in total; then after the transformation of the categorical variables, we get 284 variables. We used lasso regression on these variables and it selects 113 variables among the 284 variables. By using the cross-validation with fold 5, the β value chosen was 0.0001. The minimization of the error was found to be βˆridge = 0.1169. Then we displayed the coefficients of the ten most positively correlated variables and ten most negatively correlated variables in the below figure. The variable ’Condition2\_PosN’ has the most significant negative factor. The next variable having a high impact is the ‘Exterior1st\_BrkComm’ variable. For the positive aspect, the most influential variable was found to be ‘MSZoning\_RH’, which means that when the house is located in the residential zone, the house price will increase.

# Evaluation

The principal methodology of the report is to model the house price as dependent variables using independent variables. We seek to segregate the impact or contribution of each independent variables in price variation. The methods involved are a) Identification of house price determinants b) Develop models according to the regression technique methods. The model development is observed for the existence of multicollinearity and other problems that might affect the stability of the models developed.

1. Linear Regression**:**

In linear regression, the ratio of the dependent and independent variable should be proportionally similar. After plotting, we can see from the below chart that our rate of independent and dependent variables are also proportionally identical. The probability accuracy is also very high as we can see it almost follows the theoretical quantiles (the red line or the best fit line).

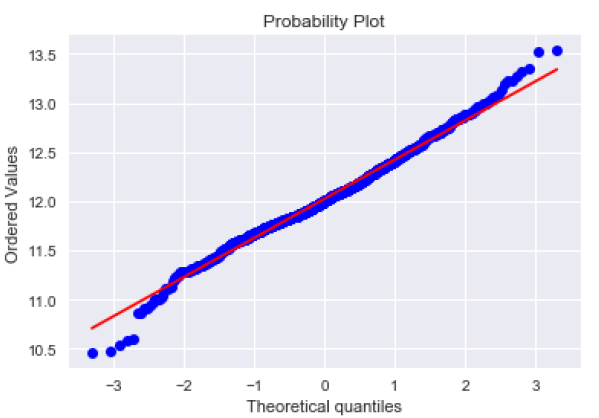


Figure : Probability plot

1. Lasso Regression:

According to figure, lasso regression was tested with the different alpha value. This graph represents the test and train Negative Mean Absolute Error, which shows how the different alpha value changes the accuracy of the result.

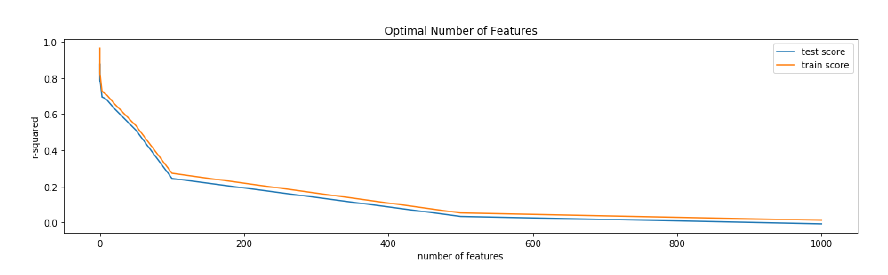
Also, from the value of the mean error, it can be noted that it is close to the actual cost of the calculation. 

Figure :Optimal number of features

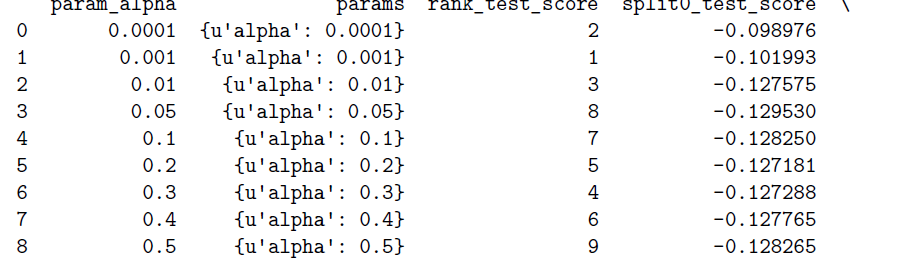


Figure : Different alpha value

By using different alpha value, we are able to generate the test score. It will have significant impact on the penalties of the regression. The r-squared esteem is a proportion of how close the information is to the fitted relapse line. It takes an incentive somewhere in the range of 0 and 1, one implying that the information clarifies the majority of the difference in the objective. Higher r-squared esteem implies a superior fit. This means that our features explain approximately 89 percent of the variance in our target variable. If our anticipated qualities were indistinguishable to the real attributes, this chart would be the straight line y=x because each anticipated esteem x would be equivalent to each genuine esteem y.

## Ridge Regression:

According to our analysis, ridge regression is ideal within the given dataset as there are a number of predictors, all with non-zero coefficients. Ridge regression can carry out well with several predictors every having little outcome. This prevents the coefficients of regression models with many correlative predictors from being poorly determined and exhibiting high variance. Ridge regression shrinks the coefficients of correlative predictors equally towards zero. So, for example, given k alike predictors, we would get coefficient consistent with /kth the extent that anybody predictor would get if match one by one. Ridge regression thus doesn’t impose coefficients to disappear, so we cannot opt for a model which solely depends om the prophetic set of variables. The ridge regression estimator deals the regression problem in Eqn. (1) using penalized least squares:

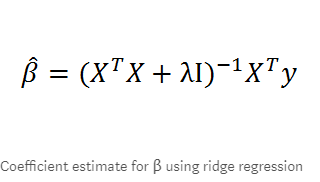


Figure : Coefficient estimate for B using ridge regression

# Conclusion

According to our analysis, followings statements conclude the overall prediction of the report for the sales of a house price of the provided dataset. Based on the regression techniques, we can observe that the predictors vary by the different parameters processed from the dataset.

## Key Difference Predictors:

Ridge regression provides the best accuracy when we compete with the feature group. The best features used are GrliveArea, LotArea, GarageArea etc. The accuracy is highest when the alpha value of the model is 0.001 and when it is based on the RidgeCV’s scoring parameter of negative mean squared error.

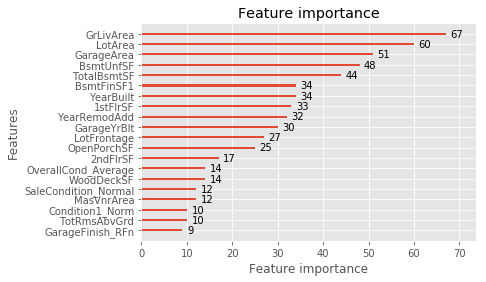


Figure : Feature graph

Lasso on the other hand, gave insights concerning chosen options, which are useful in understanding the correlations of house options and its sale costs. According to our analysis, coefficients of residence area, the material of the roof and neighbourhood have the most considerable real statistical significance in predicting a house’s sale price whereas a factor of condition2 Position Number gives most negative statistical significance (cross-validation with the fold 5).

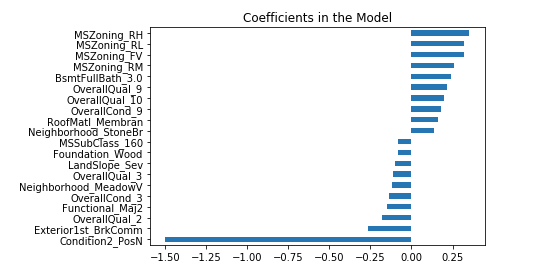


Figure 14: Coefficient in the model

Both Ridge and Lasso regression provided a high level of accuracy output with few tuning parameters.

1. Regression Results:

|  |  |  |
| --- | --- | --- |
| Regression | RMSE | R square value |
| Linear Regression | 0.0178 | 0.88824 |
| Lasso Regression | 0.02104 | 0.86964 |
| Ridge Regression | 0.11730934552048036 | 0.8325321552 |

As seen in the above table, the best-performing model is Linear Regression, with rmse of 0.0178 and R square value value of 0.88824. The most common issue that may be seen while running the dataset is that all the models appear to possess some fluctuation in them.

# References:

Bharathi , A.N, Yuvaraj, D.R.N & Dhivya, B. 2019. Predicting housing prices using advanced regression techniques. International Journal of Advance Research, Ideas and Innovations in Technology. 5(1), pp. 370-374.

Gao, G., Bao, Z., Cao, J., Qin, A.K., Sellis, T. and Wu, Z., 2019. Location-Centered House Price Prediction: A Multi-Task Learning Approach. arXiv preprint arXiv:1901.01774.

Pasha, G.R. and Shah, M.A., 2004. Application of ridge regression to multicollinear data. Journal of research (Science), 15(1), pp.97-106.

Kwon, S., Han, S. and Lee, S., 2013. A small review and further studies on the LASSO. Journal of the Korean Data and Information Science Society, 24(5), pp.1077-1088.