Statistical Learning: Practical Data Analysis Assignment

# Tanya Ambastha – 20200024 Vaishnavi Milind Gadekar-20200073

## MIS41120: Statistical Learning - Summer Trimester 2020/21

Declaration of Authorship: We “**Tanya Ambastha and Vaishnavi Milind Gadekar**” declare that all material in this assessment is our own work except where there is clear acknowledgment and appropriate reference to the work of others.

Date of Submission: 16th July 2021



Work Distribution:

**Tanya:** Worked on the Multiple Regression section for coding and report creation, and helped to identify an appropriate freely accessible dataset.

**Vaishnavi:** Worked on the SVM aspect for coding and report creation, and helped identify an appropriate freely accessible dataset.

**1. Introduction**

In this report we are demonstrating our practical skills of data analysis, model selection, prediction, model interpretability and usage of statistical tools which we learnt in the statistical learning course. We are implementing two mainstream supervised learning models, Multiple Regression and Support Vector machines on two datasets, namely, Boston data set and Cars dataset. Boston dataset is a standard dataset used for statistical and machine learning analyses. Cars dataset is the realistic dataset we have chosen for our analysis. To counter the overfitting issue, we have applied three regularization methods (ridge regression, lasso and elastic net) to our models. We have analyzed and compared the model performances, efficiency, accuracy and interpretability.

**2. Dataset Selection**

As we are fond of automobiles, we decided to find a dataset around this. This freely available and realistic dataset is based on various features of the cars of a Chinese automobile company and the prices of those cars. Basically, our aim is to predict car prices (continuous response variable) based on the independent variables which describe various features/characteristics of a cars We will also understand that which independent variables are highly correlated to the target feature ‘price’ and by knowing this, businesses can strategize their profit and marketing techniques.

**3. Statistical tool chosen: R**

We have chosen R as it is the best tool for data and statistical analysis. It has a wide range of libraries which supports machine learning operations, plots, graphs and other data analysis operations. The R tutorials introduced us to the basics of R in statistical learning and hence we were motivated to use it for our project.

**4. Data Preparation and Exploratory Data Analysis**

* 1. Boston Dataset

The Boston dataset contained no missing values. There are 14 columns and 507 rows and we did not remove any row or column as the data was already clean. We considered the column ‘medv’ (Median value of owner-occupied homes in thousands of dollars) as target/response variable. A correlation matrix was plotted, figure 1, to understand which input variables had the highest impact on the medv target variable. So, it can be observed that the input variable, ‘rm’ (average number of rooms per dwelling) is highly correlated to our target variable.

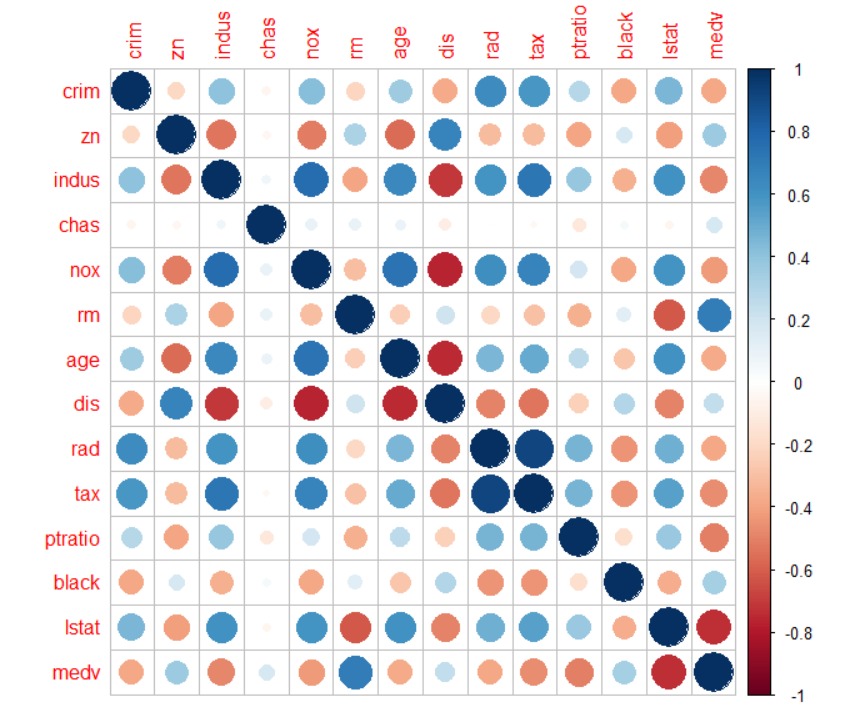


Figure 1: Correlation matrix of Boston dataset

* 1. Cars Dataset

This dataset also did not contain any missing values. It has 26 columns and 206 rows. We removed the first column, that is, ‘car\_ID’ as this was just a serial number type column and unrelated to other parameters. We considered the column ‘price’ (car prices) as target/response variable. Further we removed three columns which were categorical variables and had a lot of categories which were not relevant to our analysis had very low correlation with our target variable, namely, ’CarName’, ‘enginetype’ and ‘fuelsystem’. It can be observed from figure 2, that ‘enginesize’ is one of the highly correlated input variables with our target variables. So, for our analysis there are 21 input variables and 1 target variable.

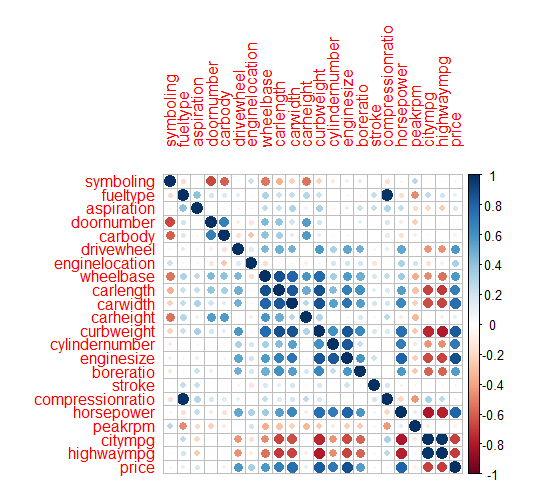


Figure 2: Correlation matrix of Cars dataset

**5. Multiple Regression method**

This method is used to model a relationship between two or more predictor/explanatory variables and a response/target variable by fitting a linear equation to the data. This one of the easiest to understand method of machine learning and is widely used.

5.1 Boston Housing Dataset

5.1.1 Model Establishment and. output

In order to perform multiple linear regression, we have split the data into 80% training data and 20% test data. We set the seed to 100 so that we get the same results each time we run the code and have deterministic datasets. The libraries used were MASS and CARET were used which are the best packages for classification and training in R. Summary function was used to evaluate which for input variables H0 can be rejected by investigating the p values. The summary function even gave the error standard error and t-values for further analysis.

The question 15b (James et al., 2017), asks us the correlation coefficients for which to reject H0: βj = 0. Going by the norms, when the p-value for input/predictor variable is below 5%, H0 can be rejected and the respective variable can be classified as statistically significant. Thus, for the following predictors, null hypothesis can be rejected: a) zn (proportion of residential land zoned for lots over 25,000 sq.ft), dis (weighted distances to five Boston employment centres), rad (index of accessibility to radial highways) and black (proportion of blacks per town) with respective p-values of 0.017025, 0.000502, 6.46e-11 and 0.040702.

The model performance and accuracy R2, adjusted R2 and RMSE were calculated by user defined functions and the output is given in table 1.

|  |  |  |
| --- | --- | --- |
|  | RMSE | R^2 |
| Train | 4.34 | 0.73 |
| Test | 5.84 | 0.73 |

Table 1: Performance results of Boston dataset- multiple linear regression

Result analysis: Looking at the r squared value, it seems the model fits perfectly and there is no underfitting or overfitting issue.

5.1.2 Regularization approaches

The results for using ridge regression via L2 regularisation as well as lasso regression with L1 regularisation and elastic net combining L1 and L2 are shown in table 2 below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Ridge Regression | Lasso | Elastic Net |
| RMSE | train | 4.281693 | 4.281097 | 4.291739 |
| test | 6.275801 | 6.280966 | 6.251675 |
| R squared | train | 0.7690863 | 0.7691297 | 0.7682803 |
| test | 0.6403773 | 0.6400858 | 0.6410981 |

Table 2: Performance results of Boston dataset- multiple linear regression with regularization

It is noted that as there was no issue of overfitting, regularization did not really help in improving the model accuracy as the initial regression model was a good fit.

5.2 Cars dataset

Similar to Boston data, Cars data was divided into 80% training and 20% test data and the formulas and functions were applied in similar manner. The model performance and accuracy R2, adjusted R2 and RMSE were calculated by user defined functions and the output is given in table 3.

|  |  |  |
| --- | --- | --- |
|  | RMSE | R^2 |
| Train | 66540.48 | 0.8801 |
| Test | 64713.11 | 0.8801 |

Table 3: Performance results of Cars dataset- multiple linear regression

Result analysis: Looking at the r squared value, it seems the model fits perfectly and there is no underfitting or overfitting issue.

5.2.1 Regularization approaches

The results for using ridge regression via L2 regularisation as well as lasso regression with L1 regularisation and elastic net combining L1 and L2 are shown in table 4 below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Ridge Regression | Lasso | Elastic Net |
| RMSE | train | 2539.504 | 2518.716 | 2539.819 |
| test | 3760.048 | 2546.128 | 3769.599 |
| R squared | train | 0.9046413 | 0.9041468 | 0.9046039 |
| test | 0.679267 | 0.6797647 | 0.6777185 |

Table 4: Performance results of Cars dataset- multiple linear regression with regularization

It is noted that as there was no issue of overfitting, regularization did not really help in improving the model accuracy as the initial regression model was a good fit.

**6. Support Vector Machine**

The dataset preparation remains same as the multiple linear regression model. SVM does not assume a probabilistic model, so there are no Standard-Errors or P-values to define ,accept and reject the null hypothesis. In order to perform SVM, we have split the data into 80% training data and 20% test data. We set the seed to 100 so that we get the same results each time we run the code and have deterministic datasets. The libraries used were e1071 and CARET were used which are the best packages for classification and training in R.

6.1 Boston Housing dataset

Table 5 shows the model performance and accuracy R2, adjusted R2 and RMSE were calculated by user defined functions and the output.

|  |  |  |
| --- | --- | --- |
|  | RMSE | R^2 |
| Train | 2.805503 | 0.871469 |
| Test | 4.785038 | 0.6626634 |

Table 5 : Performance results of Boston dataset- SVM

Result analysis: The model tends to overfit as the R2 test value is considered relatively low as compared to train value. Thus, the model overfits and regularisation would be needed to enhance the model.

6.1.1 Regularization approaches

The results for using ridge regression via L2 regularisation as well as lasso regression with L1 regularisation and elastic net combining L1 and L2 are shown in table 6 below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Ridge Regression | Lasso | Elastic Net |
| RMSE | train | 1.44 | 3.6597 | 5.89256 |
| test | 5.75 | 3.1673 | 3.1674 |

Table 6: Performance results of Boston dataset- multiple linear regression with regularization

6.3 Cars data set

|  |  |  |
| --- | --- | --- |
|  | RMSE | R^2 |
| Train | 1626.602 | 0.9514676 |
| Test | 3191.31 | 0.5500611 |

Table 7 : Performance results of Cars dataset- SVM

6.3.1 Regularization approaches

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Ridge Regression | Lasso | Elastic Net |
| RMSE | train | 3179.844 | 2346.341 | 2947.426 |
| test | 3853.048 | 2175.602 | 3824.542 |

Table 8: Performance results of Cars dataset- multiple linear regression with regularization

SparseSVM Accuracy

Lasso:

Train accuracy: 97%

Test accuracy: 93%

Elastic Net

Train accuracy: 97%

Test accuracy: 94%

**Conclusion:**

Multiple linear regression model is best for both data sets.

References:

Kumar, M., 2021. Car Price Prediction Multiple Linear Regression. [online] Kaggle.com. Available at: <<https://www.kaggle.com/hellbuoy/car-price-prediction>> [Accessed 16 July 2021].