SCHOOL OF GEOGRAPHY **UNIVERSITY OF LEEDS**	UNIVERSITY OF LEEDS

COURSEWORK COVERSHEET

Student ID number	2	0	1	6	6	9	0	9	0
Module code	GEOG5917M								
Module title	Big Data and Consumer Analytics								
Assignment title	Analysis and Prediction of Liverpool Home Prices								
Marker	Roger Beecham								
Declared word count	2333								

By submitting the work to which this sheet is attached you confirm your compliance with the University's definition of Academic Integrity as: "a commitment to good study practices and shared values which ensures that my work is a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine". Double-check that your referencing and use of quotations is consistent with this commitment.

You also confirm that your declared word count accurately reflects the number of words in your submission, excluding the overall title, bibliography/reference list, text/numbers in tables

and figures (although table and figure captions are included in the word count).

Contents

Analysis and Prediction of Liverpool Home Prices

1 Introduction:

Liverpool, renowned for its rich cultural heritage, vibrant economy, and diverse communities, has witnessed significant growth and transformation in its housing sector over the years. The report's objective is to develop a model that can predict Liverpool homes prices.The given dataset `assign_data.Rdata`[1] has been observed and the aims of the project is to:

- **Explore** the given dataset ensuring its quality and eligibility to build the right model.
- Apply the appropriate **modeling** technique to the dataset and evaluate the model using required metrics.
- Look for areas of **improvement** where the model may be strengthened or for any more information in the dataset that can help the model.

Since elastic net regression provides a reliable method of handling complicated datasets with multicollinearity and high dimensionality, it has been utilised for modeling [1]. This study will go into deeper detail about the additional advantages and application of this technique.

2 Exploratory Data Analysis:

The Liverpool homes dataset has around `2211` records with `10` variables. The detailed description of the dataset is provided in the `Table 1`. The dataset will be examined in terms of its outliers, quality, correlation and distribution. From `Table 1`, we can understand that the dataset has multiple independent variables for model usage. The geometry parameter has been removed as it doesn't add value to the model.

[1] The given dataset is in .Rdata format and the modeling is done using R studio.

Variable name	Description
Price	Mean Price of home in the area.(£1000s).
Beds	Number of Bedrooms.
`gs_area`	Local Green Space Area(in %).
`u_16`	% of people aged below 16.
`u_25`	% of people aged between 17 to 24.
`u_45`	% of people aged between 25 to 44.
`u_65`	% of people aged between 45 to 64.
`o_65`	% of people aged above 65.
unmplyd	% of people who are unemployed.
geometry	Area co-ordinates.

Table 1: Liverpool Homes Dataset Description.

2.1 Data Distribution:

The Price is distributed normally, and a large number of outliers that could upset the model are found in the `Figure 1` below.

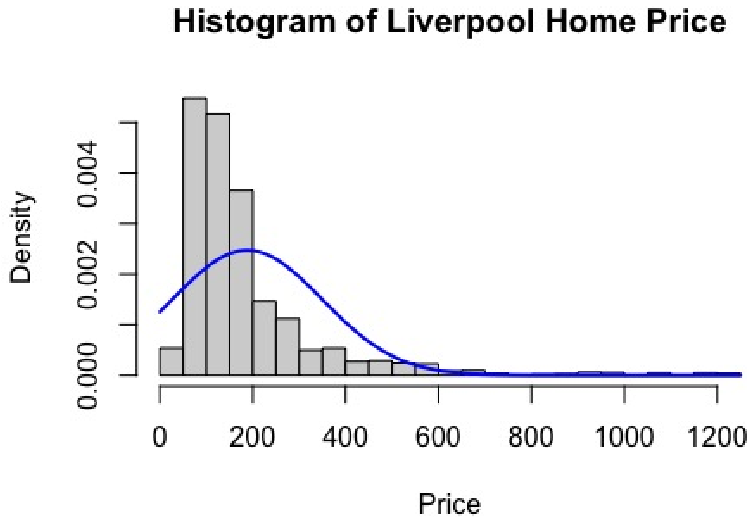


Figure 1: Histogram plotted along with the Distribution curve.

After the outliers are eliminated, duplicate records and completeness of the dataset are examined. Moreover the skewness in the Price parameter is noted.

2.2 Data Quality:

Luckily there are no missing values in any of the records given in the dataset. However many duplicate records has been found in the dataset and it has been cleaned to filter the unique records. Now around 2004 records are remaining in the dataset.

2.3 Correlation:

The correlation gram for the `assign_data.Rdata` is plotted in the `Figure 2` below to know how much the given variables affect the price.[2]

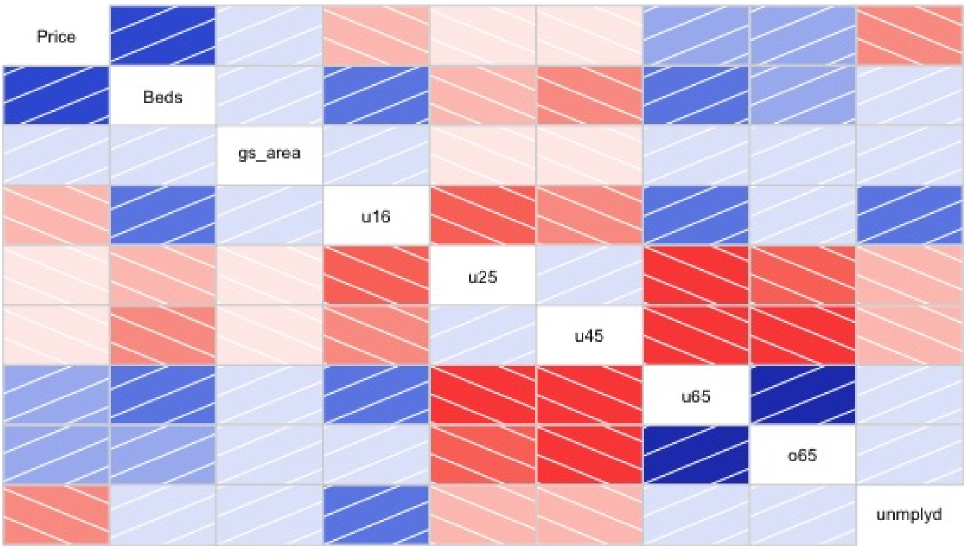


Figure 2: Correlation gram of Liverpool Homes Dataset.(Blue indicates positive link and red indicates negative link)

Price and Beds exhibit a strong positive link, as can be seen. Moreover, other parameters like `u_65`,`o_65`,`u_16` & `u_25` have high multicollinearity between them. Looks like `gs_area` has the lowest contribution to the Price among other parameters.

[2] The multicollinearity between the independent variables is one of the major reasons for choosing ENET Regression which will be further explained in the report.

2.4 Data Pre-processing:

Following every type of study, the dataset is prepared for modeling through pre-processing. The geometry parameter is dropped and the Price's outliers have been eliminated. When skewness in the distribution is discovered, the price parameter is subjected to a log transformation. The skewness in the dataset distribution can be corrected with the aid of log transformation. Now the dataset is subjected to min-max scaling using the `PreProc` argument in R. Moreover the dataset is now split into train and test dataset using `createDataPartition` function in R.

3 Modeling:

Since the output variable is **continuous**, linear regression is chosen as the right modeling choice. However, normal linear regression can't handle big dataset and input parameters with multicollinearity [2]. Hence the **elastic net regression** is chosen as the ideal modeling choice.

3.1 What is Elastic Net Regression?

Elastic Net Regression is a powerful machine learning technique used for modeling complex datasets. When normal linear regression struggles with overfitting or mulitcollinearity issues, elastic net regression can be very helpful in modeling [1]. It is a combination of ridge regresssion and lasso regression[2].

Ridge Regression is achieved by adding a penalty term to the least square estimate function. It helps the model to avoid overfitting by using the bias happened by the penalty term. The penalty term consists of squares of independent variable co-efficients[1].

$$LossFunction = \sum_{i=1}^{n}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|^2. \tag{1}$$

where,

- y is the dependent variable (the variable we are trying to predict),
- X_i is the independent variable,

- β is the slope of the regression line(co-efficient of independent variable),
- λ is the strength of penalty term.

Lasso Regression on the other hand is also achieved by adding a penalty term to the least square estimate function. Similar to ridge regression, lasso regression also helps the model to avoid overfitting by using the bias happened by the penalty term. However the penalty term consists of absolute value of independent variable co-efficients instead of squares[1].

$$LossFunction = \sum_{i=1}^{n}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda|\boldsymbol{\beta}|. \tag{2}$$

where,

- y is the dependent variable (the variable we are trying to predict),
- X_i is the independent variable,
- β is the slope of the regression line(co-efficient of independent variable),
- λ is the strength of penalty term.

By adding both the penalty terms from (1) and (2) with the least square estimate function, the loss function for the **elastic net regression** is derived.

$$LossFunction = \sum_{i=1}^{n}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda[(1-\alpha)|\boldsymbol{\beta}| + \alpha|\beta|^2]. \tag{3}$$

where,

- y is the dependent variable (the variable we are trying to predict),
- X_i is the independent variable,
- β is the slope of the regression line(co-efficient of independent variable),
- λ is the strength of penalty term.
- α is the weight of ridge and lasso penalties.

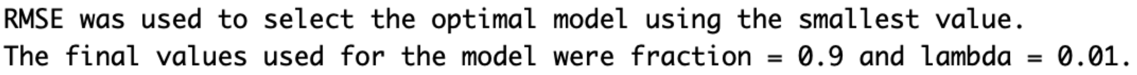
3.2 Advantages of Elastic Net Regression:

- Suitable for Big Data.
- Robustness.
- Reduces Overfitting.
- Flexible Tuning Parameter.

3.3 Applying Elastic Net Regression on Liverpool Homes Dataset:

Around 30% of the dataset is used for testing and 70% is used for training. To improve model selection, ten cross validations are used during the sampling process. For modeling, the caret and elastic net package is imported.

For the tuning grid, the penalty strength(λ) is set from 0 to 0.001 and the weight of the penalties(α) is set from 0.05 to 1. The metric is chosen as **Root Mean Squared Error**(RMSE) and the penalty is done using `expand.grid` function.



```
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were fraction = 0.9 and lambda = 0.01.
```

Figure 3: Tuning Result after testing multiple iterations.

From `Figure 3`, it is observed that α value $= 0.9$ and λ value $= 0.01$ serves the best model. The penalty strength of all possible combinations are compared with each other with respect to the RMSE. `Figure 4` illustrates that $\lambda = 0.01$ is the best fit for the model as the corresponding curve touches the lowest point in the graph which indirectly means low RMSE.

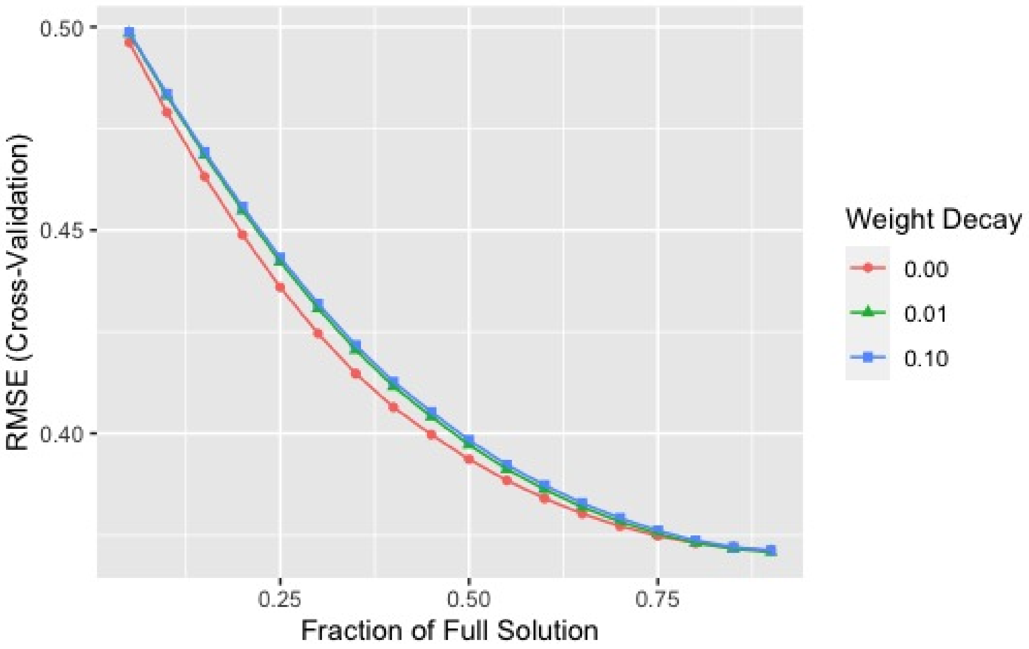
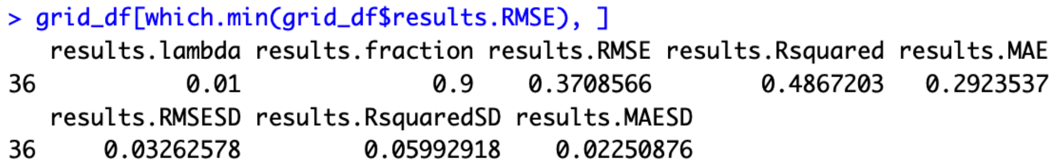


Figure 4: Penalty Strength(λ) Analysis

The metric for the tuning grid result is displayed in `Figure 5`.



```
> grid_df[which.min(grid_df$results.RMSE), ]
   results.lambda results.fraction results.RMSE results.Rsquared results.MAE
36           0.01              0.9    0.3708566        0.4867203   0.2923537
   results.RMSESD results.RsquaredSD results.MAESD
36     0.03262578          0.05992918    0.02250876
```

Figure 5: Metrics of Tuning Grid Results.

As it is already known that the ideal model's penalty strength(λ) and penalty weight(α) are 0.01 and 0.9 respectively, the RMSE value is displayed as 0.37 which should be generally low for a regression model. It measures the average deviation between the actual observed values and the predicted values produced by the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{4}$$

where,

- RMSE is the Root Mean Squared Error,

- y_i is the actual observed value,
- $\hat{y_i}$ is the predicted value,
- n is the number of observations.

Moreover the Mean Absolute Error is displayed as 0.29 which also should be low for an ideal model. It measures the average absolute difference between the actual observed values and the predicted values produced by the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y_i})|. \tag{5}$$

where,

- MAE is the Mean Absolute Error,
- y_i is the actual observed value,
- $\hat{y_i}$ is the predicted value,
- n is the number of observations.

Also the R squared value is displayed as 0.48 which indicates a good model if the value is in range between 0.4 to 0.7. Very high value might indicate overfitting.It is a statistical measure that represents the proportion of the variance in the dependent variable (output variable) that is explained by the independent variables (input variable) in a regression model.

$$R^2 = 1 - (SSR/SST). \tag{6}$$

where,

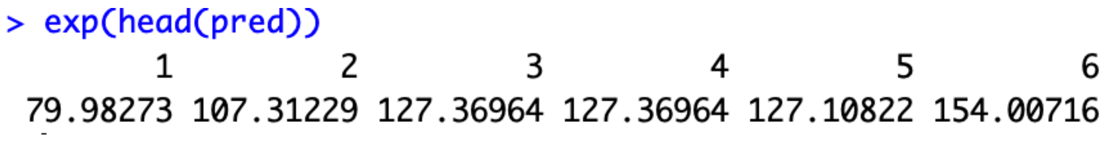
- SSR is the residual sum of the squares,
- SST is the total sum of squares

Now, the penalty strength and weight is considered for elastic net regression for final modeling. The dependent and independent variables are listed below in `Table 2`:

Independent Variables	Dependent Variable
Beds	log(Price)
`gs_area`	
`u_16`	
`u_25`	
`u_45`	
`u_65`	
`o_65`	
unmplyd	
geometry	

Table 2: Finalised Input and Output parameters for modeling.

Now the elastic net regression is applied on the train dataset for training the model. The modeling is done using the `enFit` command. The achieved tuning parameters are entered, the distribution is set to "gaussian" and the method is selected as "enet" which stands for elastic net regression. The prediction is done on the test dataset using the `predict()` function. The predicted Price values are transformed to exponential values as logarithmic Price values are used for modeling. The predicted price values are printed in the `Figure 6` below:



```
> exp(head(pred))
        1         2         3         4         5         6 
 79.98273 107.31229 127.36964 127.36964 127.10822 154.00716 
```

Figure 6: Price Values predicted by the ENET Model.

4 Final Model Evaluation:

The relation between the observed Price values and the actual Price values are plotted in the `Figure 7`. The `ggplot` function has been used for the plot. As we can see that the points form a collinearity which indicates a good sign for the model. The final model's metrics are extracted using the `postResample()` function.

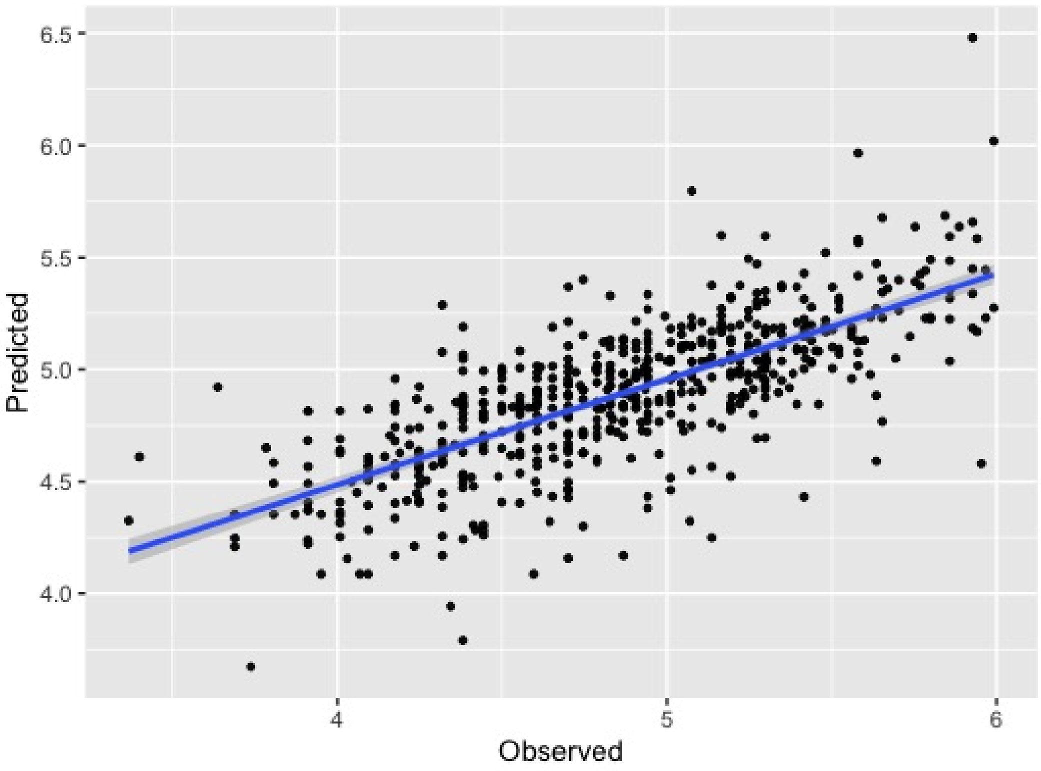
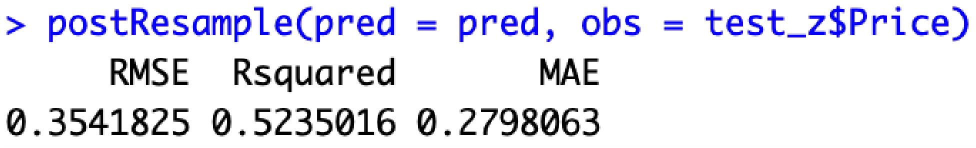


Figure 7: Observed vs Predicted Price Values

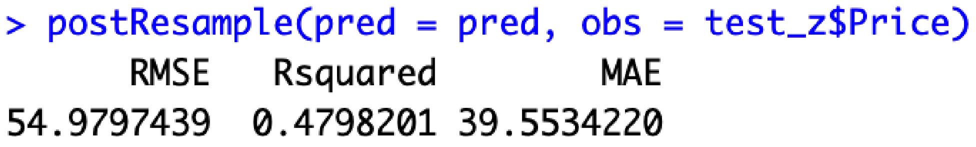
The `Figure 7` explains that the RMSE value is 0.3541825 , followed by R squared value as 0.5235 and the MAE as 0.2798063. As mentioned earlier, these metric values indicate a good model when they are in logarithmic values.



```
> postResample(pred = pred, obs = test_z$Price)
     RMSE  Rsquared       MAE
0.3541825 0.5235016 0.2798063
```

Figure 8: Final Model's Metrics(in Logarithmic)

Now the model is evaluated with real Price to calculate the precise robustness of the model.

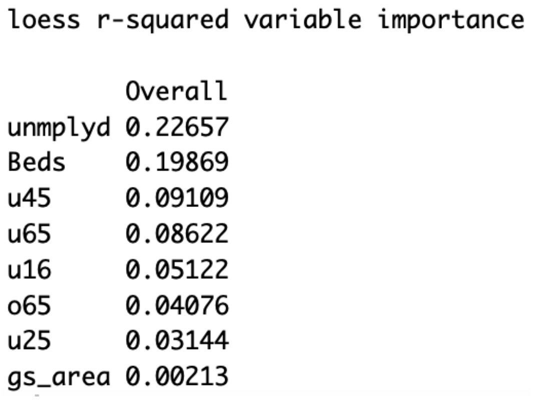


```
> postResample(pred = pred, obs = test_z$Price)
      RMSE   Rsquared        MAE
54.9797439  0.4798201 39.5534220
```

Figure 9: Final Model's Metrics(in Real Values)

As it is observed in `Figure 9` that the RMSE is 54.9797, followed by R squared value as 0.47 and the MAE as 39.5534. The mentioned metrics again prove that the model is a good fit for the data.

The importance of independent variables are calculated using the `VarImp()` function and the importance values are listed below in the `Figure 10`.



```
loess r-squared variable importance

         Overall
unmplyd 0.22657
Beds    0.19869
u45     0.09109
u65     0.08622
u16     0.05122
o65     0.04076
u25     0.03144
gs_area 0.00213
```

Figure 10: Importance Value of Input Variables

It has been observed that the unmplyd and Beds contribute highly to the model whereas `u_25` and `gs_area` has very less contribution.

5 Further Discussion:

5.1 Limitations:

- The Liverpool Dataset has 8 independent variables which might be pretty low for predicting the home price.
- The given dataset is assumed that the given information is precise and not manipulated. If there is any presence of manipulated data, it might have affected the model.
- Age of the homes might have been provided, which would have played a major role in predicting the Price.
- Median Income of the people around the area might have a crucial role too.
- Population of the respective area is a major miss in the dataset.

5.2 Area of Future Work:

The dataset is trained for other models such as kNN(K- nearest neighbour) , SVM(Support Vector Machine) and cox proportional model(TB). The variable importance for the corresponding models are plotted below in `Figure 11`. Further study can be progressed by comparing the models in every possible way.

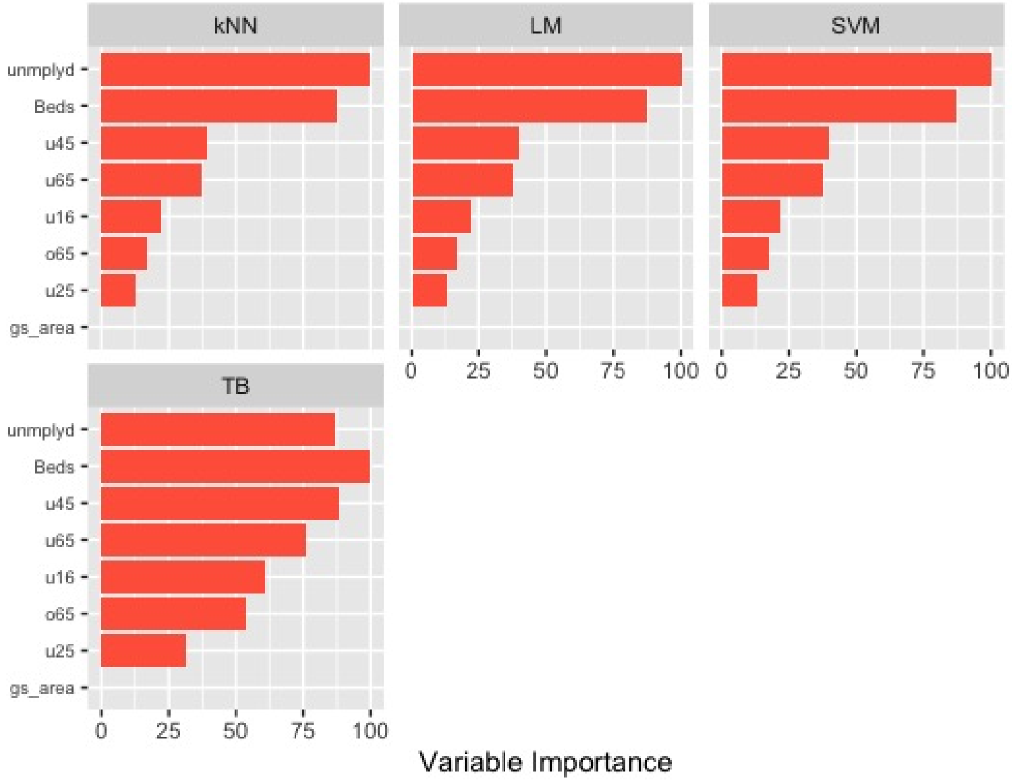


Figure 11: Variable Importance comparison with other models

6 Conclusion:

The elastic net regression provides a robust model by combining both lasso and ridge regression[1] . The model is applied on the dataset and the insights are extracted in the report.

References

[1] Cannon Giglio and Steven D. Brown. Using elastic net regression to perform spectrally relevant variable selection: Elastic net regression for variable selection. *Journal of chemometrics*, 32:e3034–, 2018.

[2] Hyemin Han and Kelsie J. Dawson. Applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood. *Journal of adolescence (London, England.)*, 93(1):20–27, 2021.