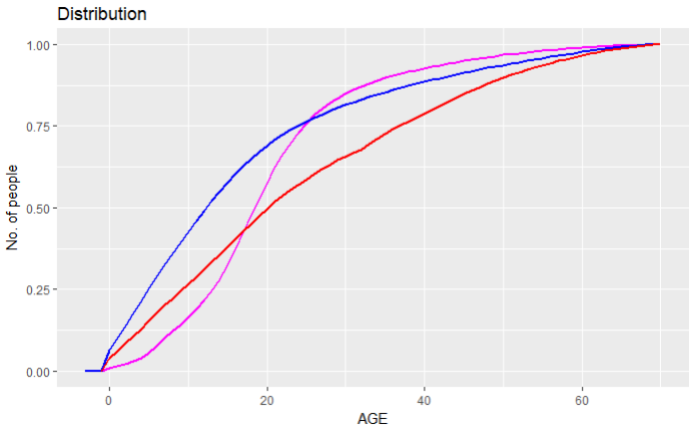
Name : Rohit Garla Net ID: rg3365
Name : Mengrui Zhang Net ID:mz2258

Assignment 4(TEAM 4)

Introduction

The assignment covers up the topic of selecting a model which can predict based on the mortgage data provided. Entropy was considered to be the best attribute to compare these models.

Filtering the Data based on Age



From the graph we can observe that cumulative distribution of No. of people are different for different group. This is also an intrinsic motivation that our AGE parameter may be valid, since it does have difference patterns for different groups. We can do this analysis for other variables later.

Logistic Regression Model

For the assignment, we have used multinomial logistic regression, since we are dealing with dependent variable which has three categories.

The underline technique will be same as the logistic regression for binary classification until calculating the probabilities for each target. Once the probabilities were calculated.

Splitting the Regressor

- Splitting the Regressor into two parts namely. One which equals age if age<=20,while takes up the value 20 when age>20.
- The other Regressor takes up the value of age when age >20 and 20 when age <20.
- Then combining these models and calculating the entropy of the resulting model.

Random Forest Model

As a comparison for our logistic regression, we also built a random forest model. Random forests are ensembles of decision trees. Random forests are one of the most successful machine learning models for classification and regression. They combine many decision trees to reduce the risk of overfitting.

Results

Logistic Regression after combining both the regressor.

```
Entropy, 0.18914404008366115, 0.22701063313803096, 0.1866856017456471, 0.3051048331195982
```

Random Forest with features

```
Entropy, 0.1897972146213277, 0.19287326248368547, 0.193334537487614, 0.21460650006233023
```

Random Forest with features2

```
Entropy, 0.1894288747986628, 0.19249134338696336, 0.19314775361291484, 0.20909734935873797
```

Models after dropping the column Season

```
========= Entropy =========
Entropy, 0.1929848886307612, 0.1922622485380742, 0.1936120371641474, 0.19349052995964636
Entropy, 0.1897972146213277, 0.1892316679011291, 0.193334537487614, 0.1931591573319771
Entropy, 0.18977974294204664, 0.1893843070715667, 0.19269467211442673, 0.19248612722407496
```

Conclusion

Based on our models, we may find that the input volume may extremely affect our result of entropy. It seems that, this should have some skewness of our real group volume . From the results we came to the conclusion splitting the model based on age and then recombining them caused the model to over fit the data. We also came to the conclusion that the Random Forest model performed better when compared to the Logistic Regression Model. Random Forest is able to discover more complex dependencies at the cost of more time for fitting. If it's established, that our variable of interest has the linear dependency from the predictors, we may will get similar results with both models. Therefore, due to huge computational complexity of RF compared to regression's it might have performed better.