Feature Selection and Classification for Simulated Data Set

Name : Jayanth Sivakumar

Kaggle Team Name : RandomForest

Challenges

- The dataset is very big (n = 600,p = 40000)
- Loading the dataset is very time consuming
  - Train Set = 187.241s
  - Test Set = 291.925s
- It takes **~ 8 minutes** to load both the datasets
- **Parallel Computing** helps overcome these challenges and improves the running time
- Using the R package's ***bigmemory***, It has the support for loading very huge datasets
- It loads the dataset as ***big.matrix*** type
- Loading both the datasets using *bigmemory* has been substantially reduced to **1.4 Minutes**
- It is very fast and efficient.

Variable Selection

- R package's ***bigstep*** that supports parallel computing, performs **modified stepwise selection procedure**
- In the first step the likelihood ratio tests between two regression models are performed:
  - with only the intercept
  - with the intercept and every single variable from the matrix X .
- P-values are calculated and variables with $p > minpv$ (threshold) are **excluded** from the model selection procedure.
- In the second step (**multi-forward**) we start with the null model and add variables which decrease *crit.multif* (in order from the smallest p-value).

Variable Selection

- The step is finished after we add *maxf* variables or none of remaining variables improve *crit.multif*.
- Then the classical **backward selection** is performed (with crit).
- When there is no variables to remove, the last step, the classical **stepwise procedure**, is performed (with crit).

Variable Selection

- **SelectModel** is the function from ***bigstep*** which does the variable selection
- It takes the matrix of type **big.matrix** and normal matrix type
- Dataset was loaded as **big.matrix** because of the high dimensionality
- **fitLogistic** option in the function fits the logistic model and calculate the log likelihood.
- It returns the names of the variable in the final model.
- Finally, the datasets have been filtered with these variables in both training and test set.
- Using these filtered datasets, the models have been created.

Model Creation and Assessment

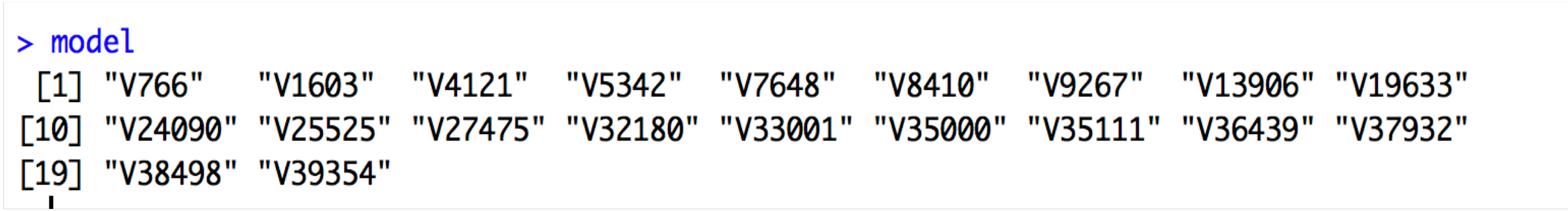
- Before variable selection, Lasso for classification was trained on the whole dataset.
- It took a long time and the **accuracy was 75%** in Kaggle.
- After selecting the variables using *bigstep*, the accuracy **improved to 96%**
- After the variable selection, the training and test set was filtered with these selected from **selectModel**
- The original dataset is split into training set and validation set (70 – 30) using this filtered dataset
- Split dataset is used for training on **Lasso using repeated CV**.

Model Creation and Assessment

- For the best performing model, the error on the validation set for each repeats was saved
- Based on the less validation error, the respective tuning parameters and the model was selected
- Apart from training the filtered dataset on Lasso, several other algorithms were used
- SVM was used to fit on the filtered dataset and it performed equally well like Lasso

Selected Variables

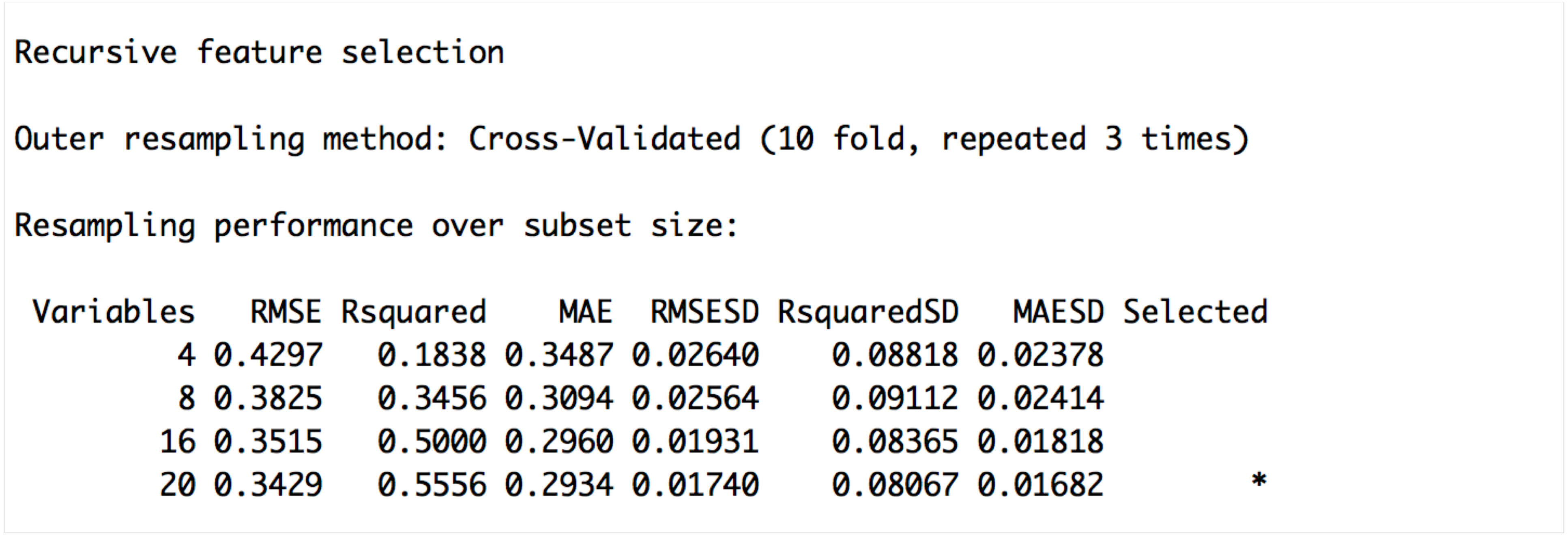
- Only **20 variables** were used to find the best performing model
- Using Lasso, the filtered dataset with these 20 variables gave the highest accuracy on the validation set as well as on Kaggle Entry.
- The Variables are



```
> model
 [1] "V766"   "V1603"  "V4121"  "V5342"  "V7648"  "V8410"  "V9267"  "V13906" "V19633"
[10] "V24090" "V25525" "V27475" "V32180" "V33001" "V35000" "V35111" "V36439" "V37932"
[19] "V38498" "V39354"
```

Variable Importance

- The plot shows the variable importance,
- To check if the variables selected outperformed, **Recursive Feature Selection** has been used
- As you can see the selected model from RFS is with 20 variables.
- All these 20 variables have very less P-value.
- On the other hand, the correlation between these variables and the outcome is high showing that these variables are the strong predictors.

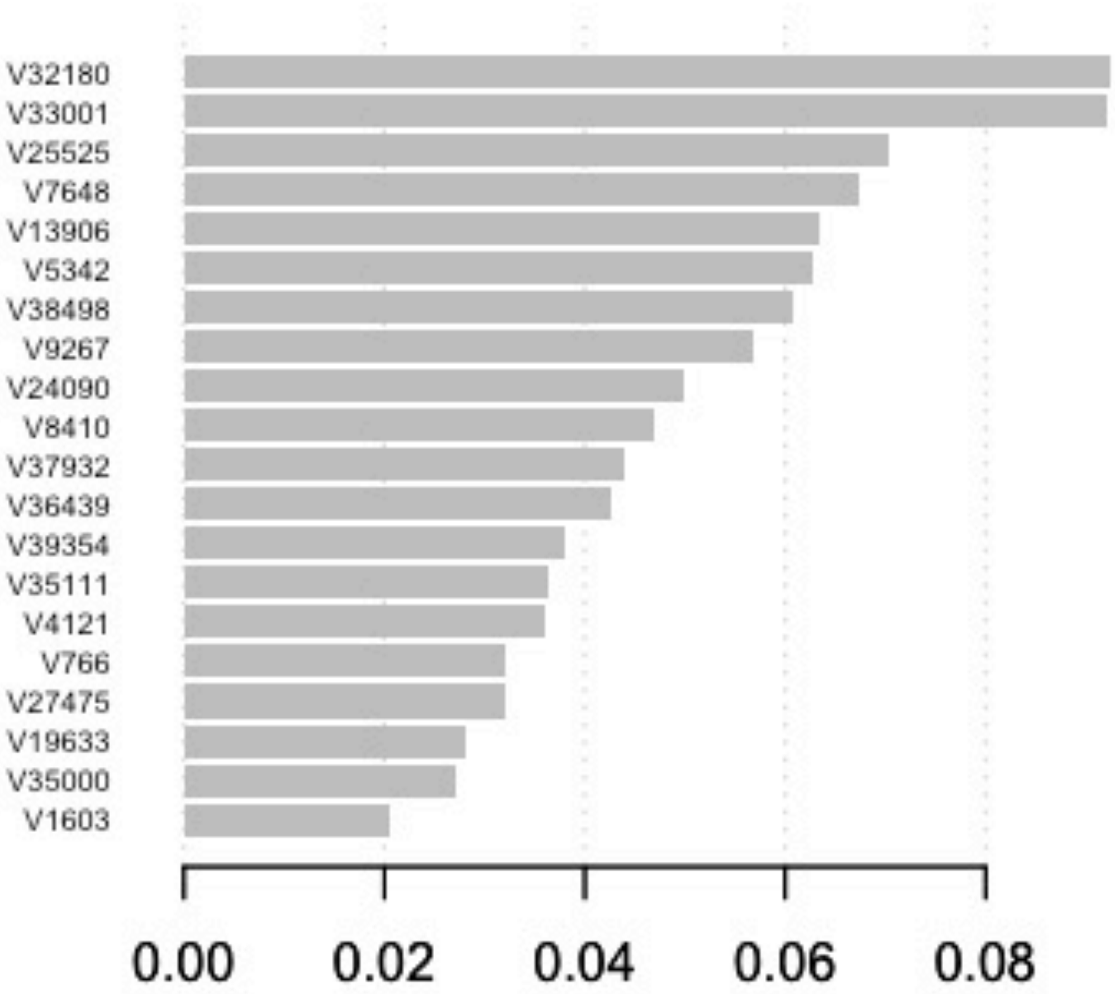


```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 3 times)

Resampling performance over subset size:

 Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         4 0.4297   0.1838 0.3487 0.02640    0.08818 0.02378         
         8 0.3825   0.3456 0.3094 0.02564    0.09112 0.02414         
        16 0.3515   0.5000 0.2960 0.01931    0.08365 0.01818         
        20 0.3429   0.5556 0.2934 0.01740    0.08067 0.01682        *
```


V32180
V33001
V25525
V7648
V13906
V5342
V38498
V9267
V24090
V8410
V37932
V36439
V39354
V35111
V4121
V766
V27475
V19633
V35000
V1603
0.00
0.02
0.04
0.06
0.08

Learners

Learners	Variable Selection?	Training Error	Validation Error	Kaggle Test Error
Lasso	Yes	0	0.0333333	0.02143
SVM*	Yes	0	0.0333333	0.025
Logistic Regression*	Yes	0	0.02	0.06667
Biglasso	No	0	0.1833333	0.18472
Elastic Net	No	0	0.2	0.29286

*Without variable selection – stack overflow error

Thank you!