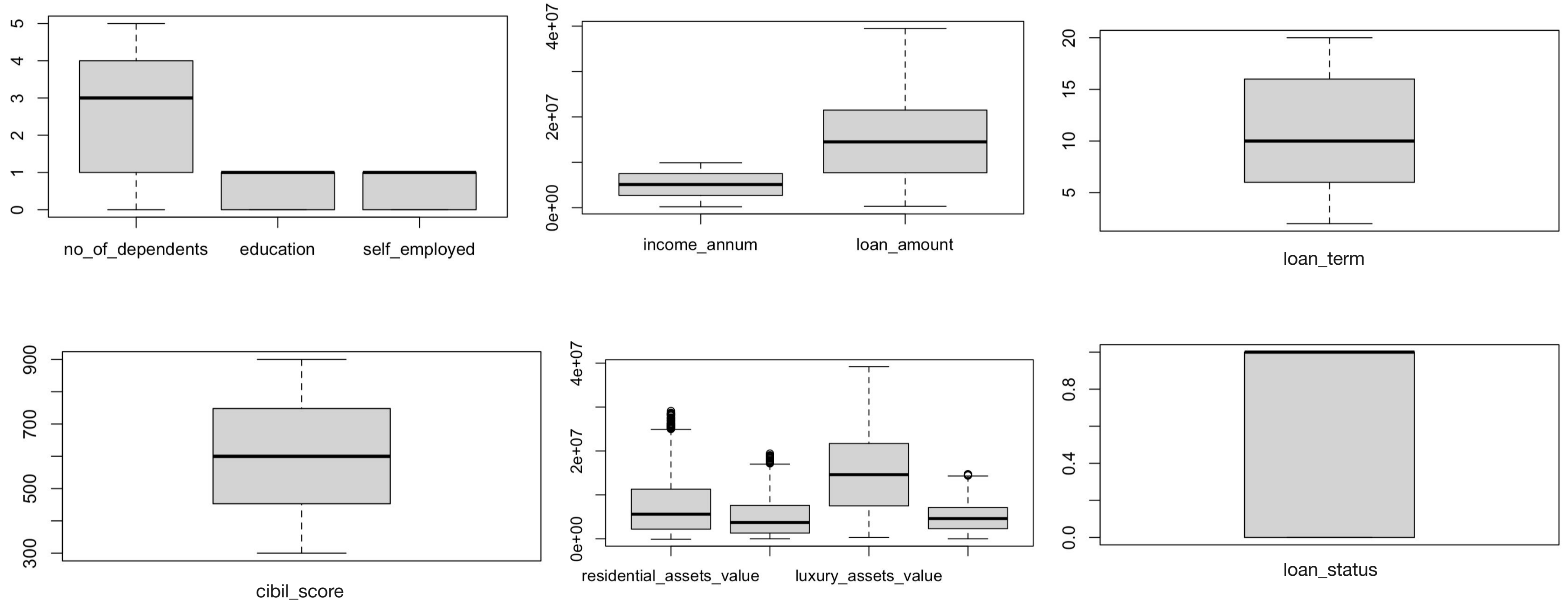
Predicting Loan Approval Using Random Forests

Group 8: Caitlyn Blair, Shruti Hardasani, Rainna Sena

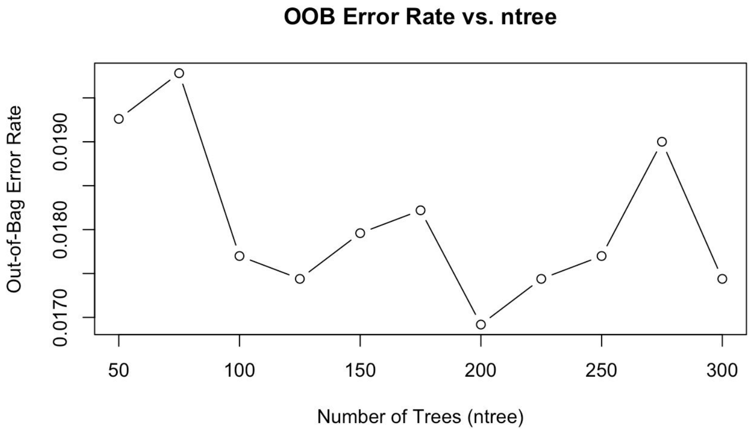
```
> summary(loandata)
    loan_id     no_of_dependents   education      self_employed     income_annum
 Min.   :   1   Min.   :0.000    Min.   :0.0000   Min.   :0.0000   Min.   : 200000
 1st Qu.:1068   1st Qu.:1.000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2700000
 Median :2135   Median :3.000    Median :1.0000   Median :1.0000   Median :5100000
 Mean   :2135   Mean   :2.499    Mean   :0.5022   Mean   :0.5036   Mean   :5059124
 3rd Qu.:3202   3rd Qu.:4.000    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:7500000
 Max.   :4269   Max.   :5.000    Max.   :1.0000   Max.   :1.0000   Max.   :9900000
  loan_amount          loan_term      cibil_score    residential_assets_value
 Min.   :  300000   Min.   : 2.0   Min.   :300.0   Min.   : -100000
 1st Qu.: 7700000   1st Qu.: 6.0   1st Qu.:453.0   1st Qu.: 2200000
 Median :14500000   Median :10.0   Median :600.0   Median : 5600000
 Mean   :15133450   Mean   :10.9   Mean   :599.9   Mean   : 7472617
 3rd Qu.:21500000   3rd Qu.:16.0   3rd Qu.:748.0   3rd Qu.:11300000
 Max.   :39500000   Max.   :20.0   Max.   :900.0   Max.   :29100000
 commercial_assets_value luxury_assets_value bank_asset_value    loan_status
 Min.   :       0        Min.   :  300000    Min.   :       0   Min.   :0.0000
 1st Qu.: 1300000        1st Qu.: 7500000    1st Qu.: 2300000   1st Qu.:0.0000
 Median : 3700000        Median :14600000    Median : 4600000   Median :1.0000
 Mean   : 4973155        Mean   :15126306    Mean   : 4976692   Mean   :0.6222
 3rd Qu.: 7600000        3rd Qu.:21700000    3rd Qu.: 7100000   3rd Qu.:1.0000
 Max.   :19400000        Max.   :39200000    Max.   :14700000   Max.   :1.0000
```

- Our dataset is the "Loan Approval Dataset" from Kaggle
- Loan status is our dependent variable
- We will exclude loan ID
- Most outcomes have very large means. Self employed has the lowest with 0.50 and residential assets has the highest of 74,726,617.

Boxplots of all Independent Variables



Random Forest



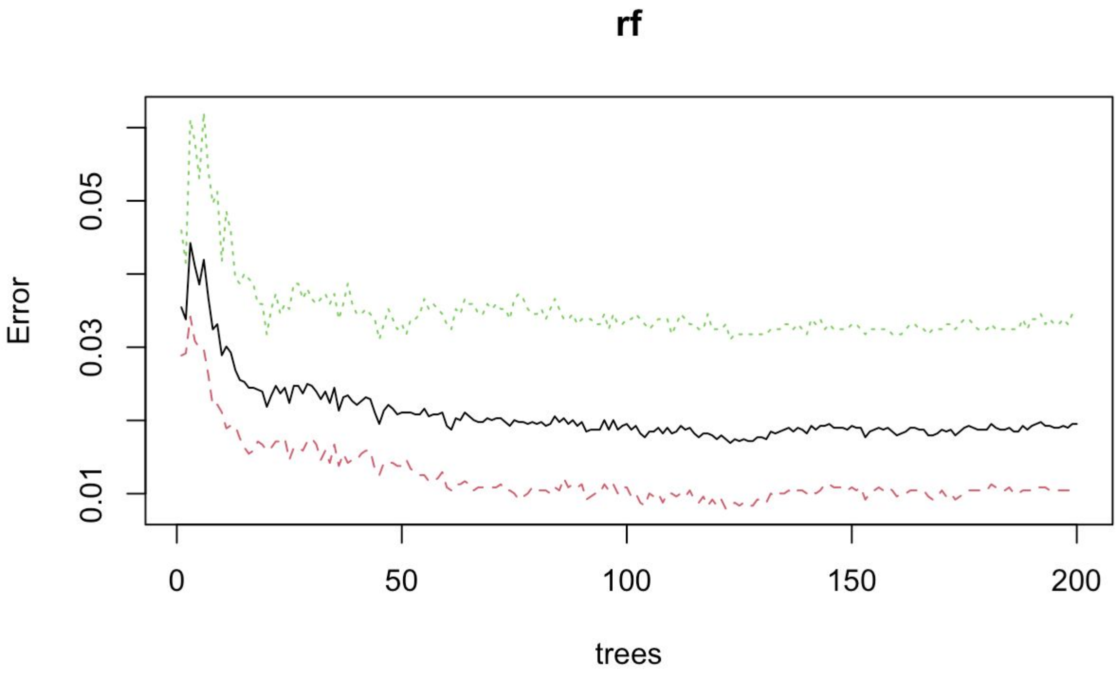

```
Call:
 randomForest(formula = loan_status ~ . - loan_id, data = loan_train,
ntree = 200, mtry = 3, oob.prox = TRUE)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 3

        OOB estimate of  error rate: 1.95%
Confusion matrix:
           Approved  Rejected class.error
 Approved      2369        25  0.01044277
 Rejected        50      1398  0.03453039
```

OOB Error Rate VS Number of trees shows that 200 trees gives the lowest OOB

Running the model using 200 trees and an mtry of 3 because the square root of 10 is approximately 3

Plot of the Random Forests



- The plot of the random forests shows the error rates at each tree number

- Range from 0-200

Confusion Matrix

Training Set

Confusion Matrix and Statistics

	Reference	
Prediction	Approved	Rejected
Approved	2394	0
Rejected	0	1448

Accuracy : 1

Test Set

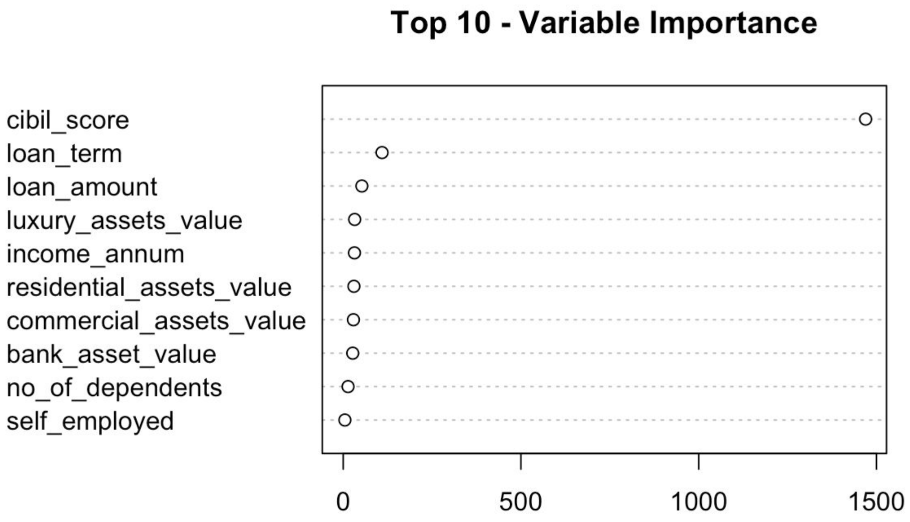
Confusion Matrix and Statistics

	Reference	
Prediction	Approved	Rejected
Approved	261	5
Rejected	1	160

Accuracy : 0.9859

Our model accuracy for our training set is 100% while our model accuracy for our test set is 98.59%, which proves to be a good model

Variable Importance



- Cibil score is the most important factor in determining whether a loan will be approved or not, valued at 1470
- The next most important is loan term at 109
- Based on the mean decrease impurity

Logistic Regression Coefficients

For the original model, we ran logistic regression with all variables besides loan ID because it was just a count. Their coefficients are shown below.

```
Call:  glm(formula = loan_status ~ . - loan_id, family = binomial, data = loan_train)

Coefficients:
             (Intercept)           no_of_dependents    education Not Graduate
               1.136e+01                  1.384e-02                 1.470e-01
       self_employed Yes               income_annum               loan_amount
              -1.110e-01                  5.825e-07                -1.410e-07
               loan_term                cibil_score  residential_assets_value
               1.496e-01                 -2.483e-02                -5.167e-09
 commercial_assets_value        luxury_assets_value           bank_asset_value
              -1.513e-08                 -2.423e-08                -5.718e-08
```

Forward Selection and AIC

The most efficient variables based on forward selection are cibil score, loan term, loan amount and annual income.

```
Step:  AIC=1893.77
loan_status ~ cibil_score + loan_term + loan_amount + income_annum
```

The AIC for the optimal model was 1724, which is lower than the AIC for model 1, making the optimal model the better model.

```
> AIC(model_1)
[1] 1730.85
> AIC(model_opt_train)
[1] 1724.181
```

Misclassification Matrix

Training Set

(171 + 156) / (2238 + 156 + 171 + 1277) =0.085

Error rate 8.5%

Truth	Predicted 0	Predicted 1
Approved	2238	156
Rejected	171	1277

Test set

(16 + 19) / (243 + 19 + 16 + 149) = 0.082

Error rate 8.2%

Truth	Predicted 0	Predicted 1
Approved	243	19
Rejected	16	149

Conclusion

- Based on our findings random forest is more accurate for this dataset based on accuracy, but both methods have high accuracy
- Cibil score is the most important factor for both random forest and logistic regression

Thank you. Any questions?