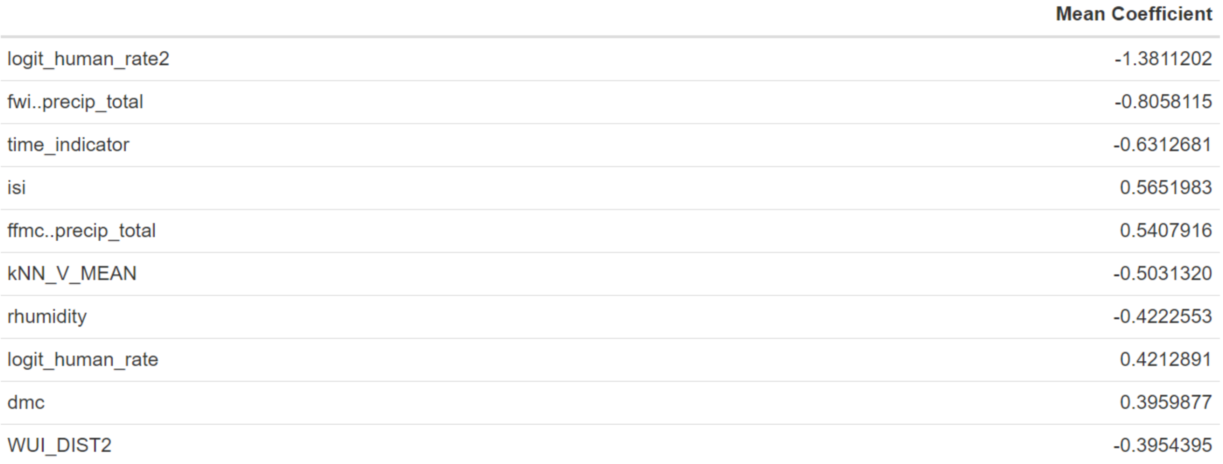
BINF 6970

Assignment 5

Jacob Hambly

Table 1. The top 10 features from the bagged LASSO model based upon the mean magnitude of the coefficients.



	Mean Coefficient
logit_human_rate2	-1.3811202
fwi..precip_total	-0.8058115
time_indicator	-0.6312681
isi	0.5651983
ffmc..precip_total	0.5407916
kNN_V_MEAN	-0.5031320
rhumidity	-0.4222553
logit_human_rate	0.4212891
dmc	0.3959877
WUI_DIST2	-0.3954395

Table 1 shows the 10 features identified from LASSO after bootstrapping 100 times on the ‘forest_fires.csv’ data and averaging the coefficients of the 100 LASSO models. The feature “logit_human_rate2” was identified as the most influential feature.

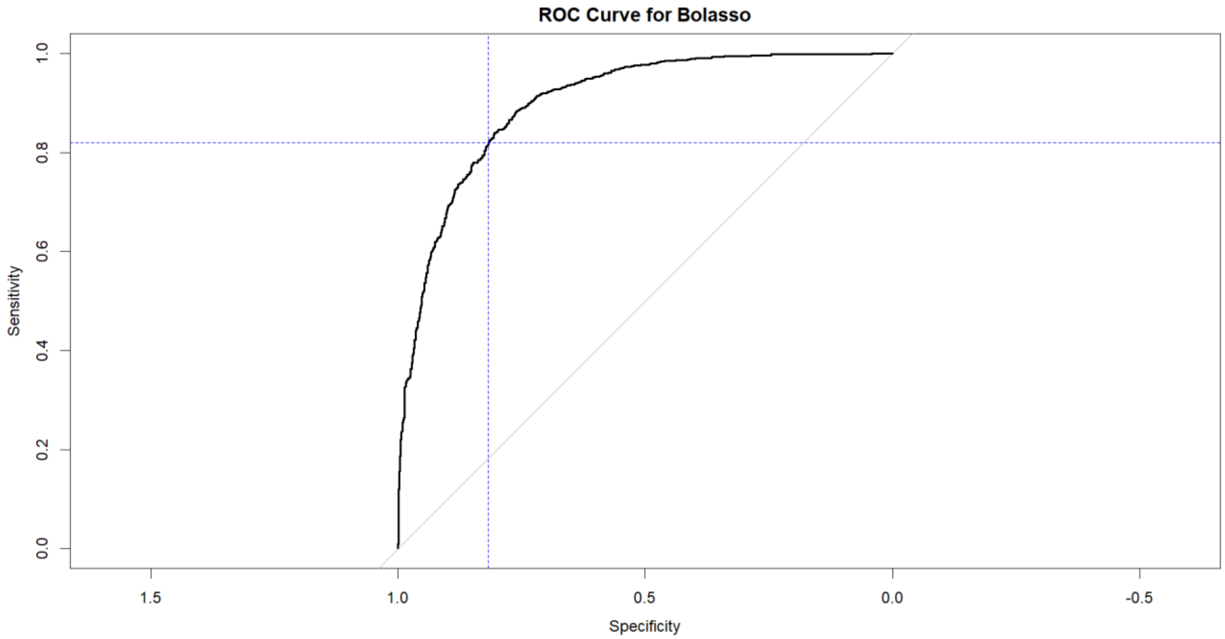


Figure 1. ROC curve generated from the test set predictions using bagging of 100 bootstrapped LASSO models. The sensitivity and specificity for the optimal threshold (point on the curve with the closest Euclidean distance to the point (1,0) on the y-axis) are shown as blue dotted lines.

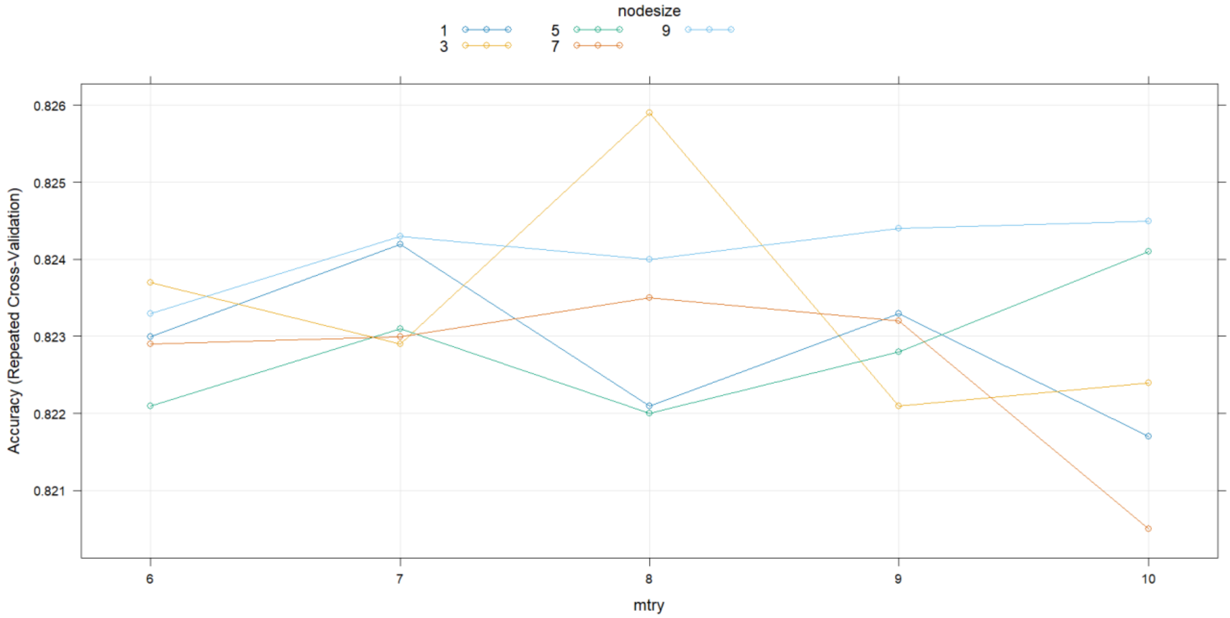


Figure 2. Optimization of a Random Forest model using a grid search over 5 values of mtry and 5 values of node size for a total of 25 models. The x-axis represents the 'mtry' values and the y-axis represents the validation score during 5-fold cross validation. The colour of the lines represents the node size. The model with the highest mean validation accuracy was the model with a mtry of 8, and a node size of 3.

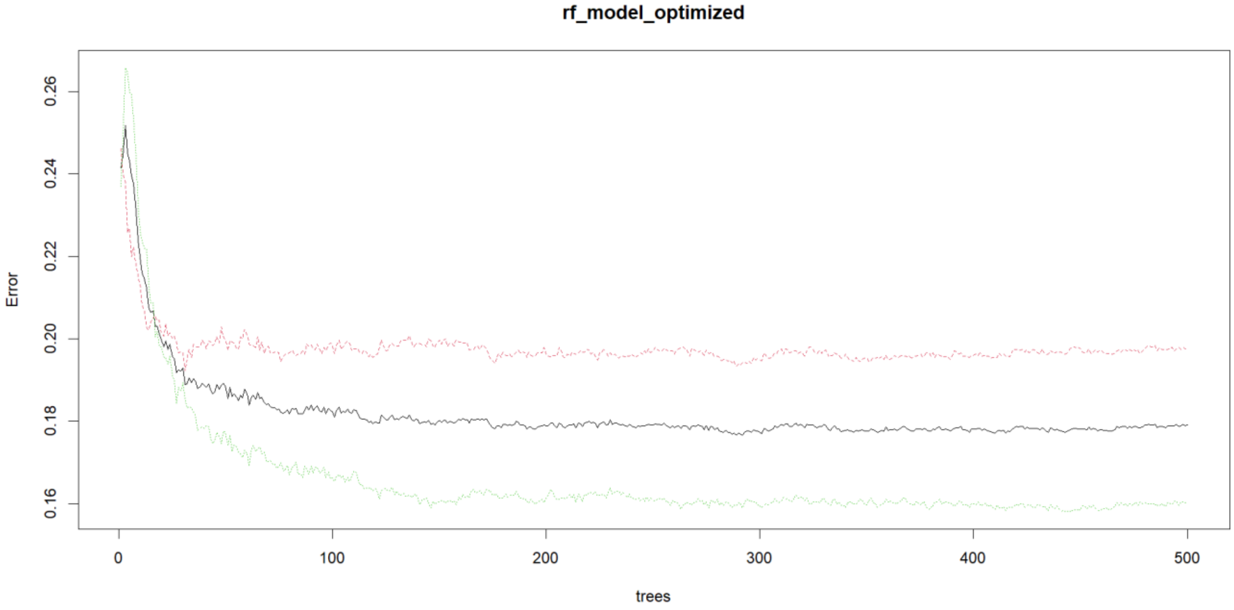


Figure 3. The out of bag error (OOB) and the number of trees in the random forest model. The black line represents the mean OOB. The red line represents OOB associated with 'human caused (1)' and the green line represents OOB associated with 'not human caused (0)'. 200 trees were determined to be an optimal amount due to the leveling off in the OOB errors.

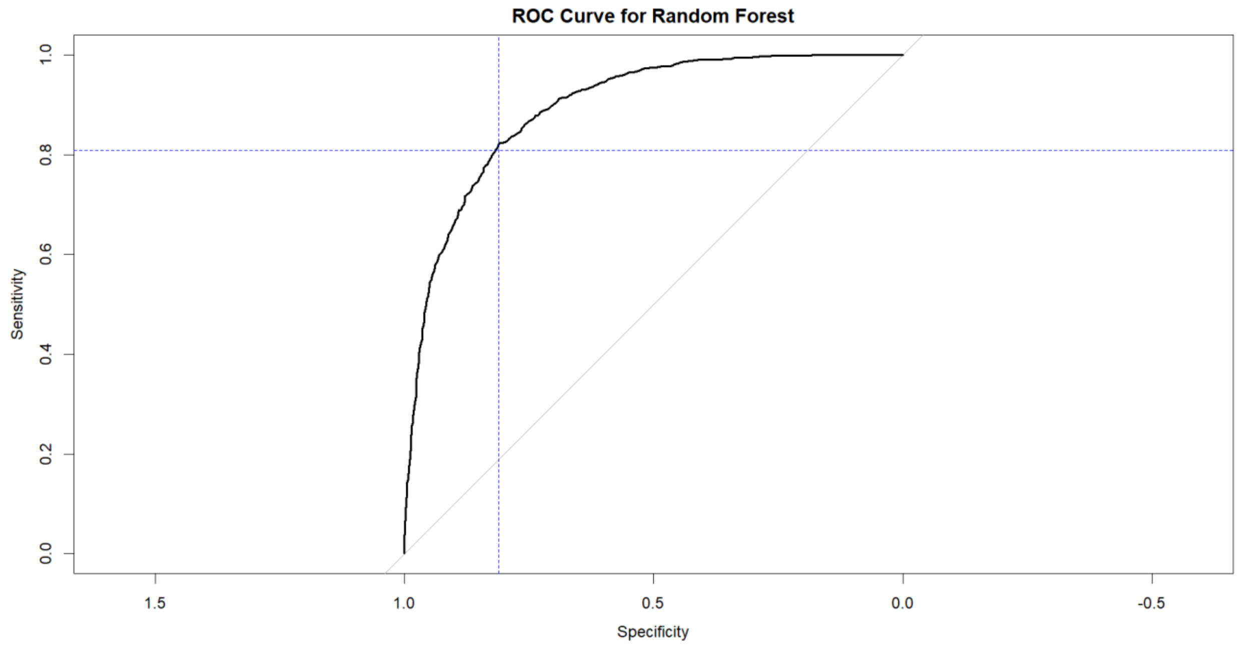
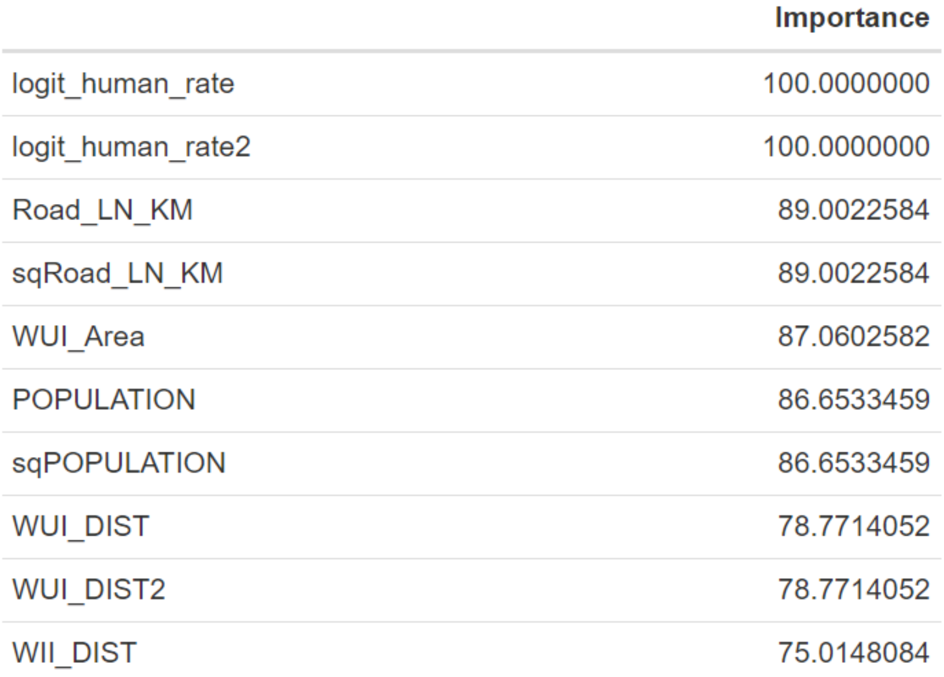


Figure 4. ROC curve generated from the test set predictions using the optimized Random Forest model (mtry = 8, node size = 3, tree size = 200). The sensitivity and specificity for the optimal threshold (point on the curve with the closest Euclidean distance to the point (1,0) on the y-axis) are shown as blue dotted lines.

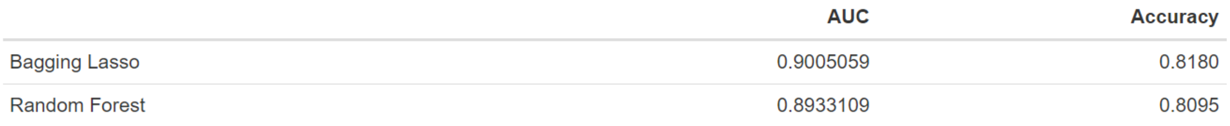
Table 2. The top 10 features from the optimized Random Forest model (mtry = 8, node size = 3, tree size = 200) based on how much each predictor contributes to the reduction in accuracy when that predictor is permuted randomly.



	Importance
logit_human_rate	100.0000000
logit_human_rate2	100.0000000
Road_LN_KM	89.0022584
sqRoad_LN_KM	89.0022584
WUI_Area	87.0602582
POPULATION	86.6533459
sqPOPULATION	86.6533459
WUI_DIST	78.7714052
WUI_DIST2	78.7714052
WII_DIST	75.0148084

For both the Random Forest and the bagged Lasso model, ‘logit_human_rate2’ is seen as the most important feature. However, compared to the bagged Lasso model, the Random Forest places more importance on ecumene features. 3 of the top 10 features are ecumene features in the bagged LASSO model where each of the top features are ecumene features in the Random Forest model. These ecumene features are most likely related to occurrence of person-caused forest fires in Canada. Therefore, the Random Forest model could capture the relationship between person-caused forest fires compared to the bagged LASSO model.

Table 3. The area under the ROC curve and accuracy associated with the bagging of the 100 bootstrapped LASSO models and the optimized Random Forest model.



	AUC	Accuracy
Bagging Lasso	0.9005059	0.8180
Random Forest	0.8933109	0.8095

Based on Table 3, we can see similar AUC and accuracy for the bagged LASSO model and Random Forest with the Bagged LASSO model performing slightly better. Both bagged algorithms seem to be comparable in terms of performance on the test set.

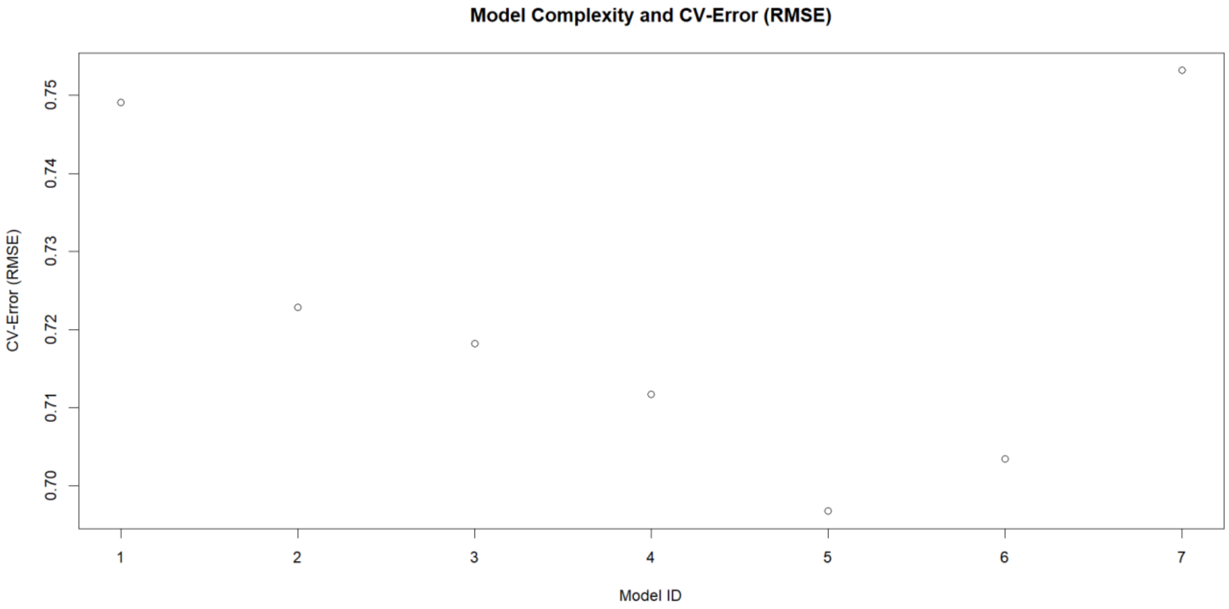


Figure 5. The mean cross validation error (RMSE) and model ID. Model complexity increases from 1 to 7.

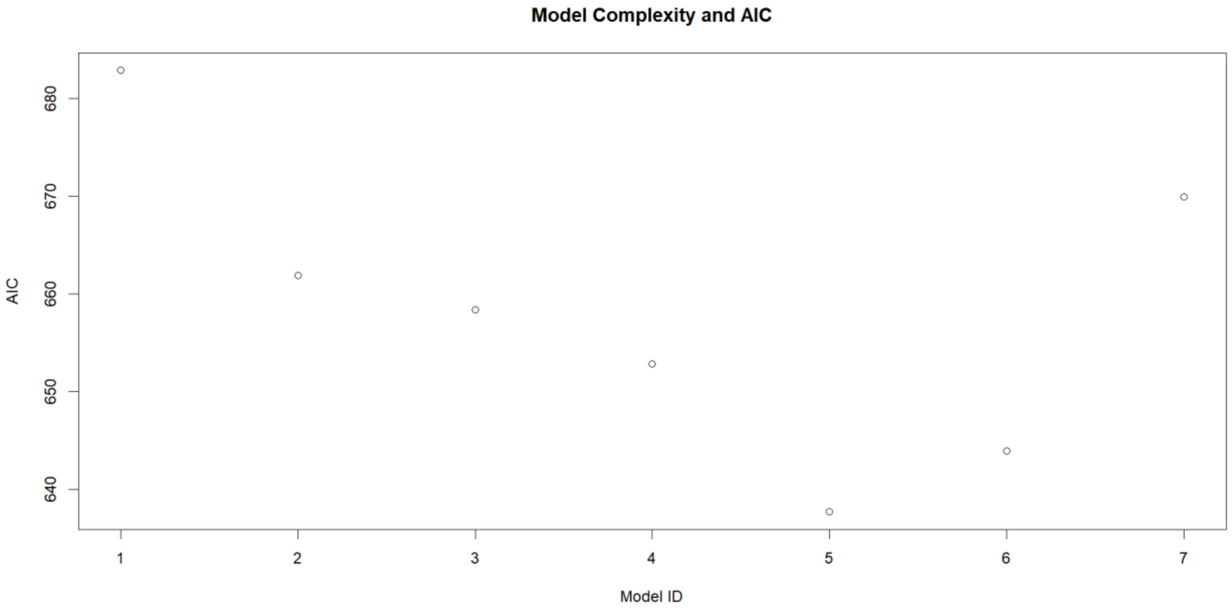


Figure 6. The Akaike Information Criterion (AIC) and model ID. Model complexity increases from 1 to 7.

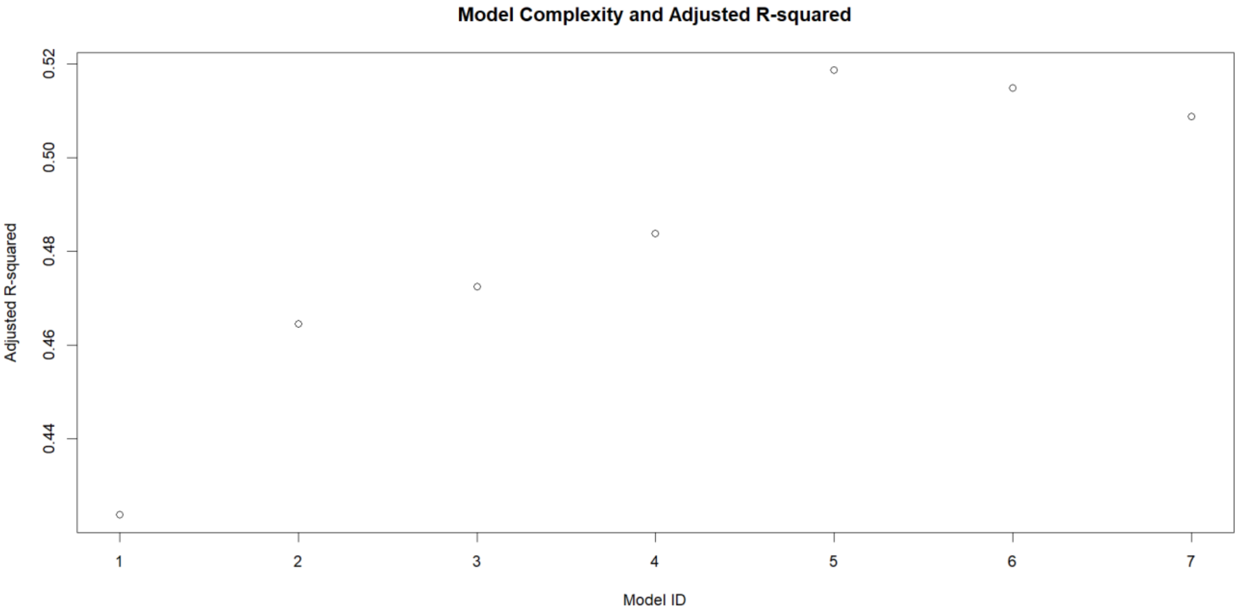


Figure 7. The adjusted r-squared value and model ID. Model complexity increases from 1 to 7.

Based on Figure 5, there is a notable downward trend in the cross-validation (CV) error across different model IDs, reaching its lowest point at model ID 5, followed by an increase at model ID 7. This suggests that model ID 5 achieved the lowest CV error and may represent the best fit for the data. This observation is further supported by Figures 6 and 7. Figure 6 illustrates a similar trend observed in Figure 5, where model ID 5 exhibits the lowest Akaike Information Criterion (AIC). A lower AIC value indicates a better-fitted model, considering both model performance and complexity. Additionally, figure 7 reinforces this trend by displaying the adjusted R-squared values for the models. Model ID 5 shows the highest adjusted R-squared, indicating that it explains a greater proportion of variance in the data compared to other models, while also considering model complexity.