Random Forest

Learning Objectives

- What is Ensemble Modeling?
- What is Bagging?
- Random Forest Algorithm
- Out of Bag Error Rate
- Finding Optimal Number of Trees
- Finding Optimal Number of Variables to Select

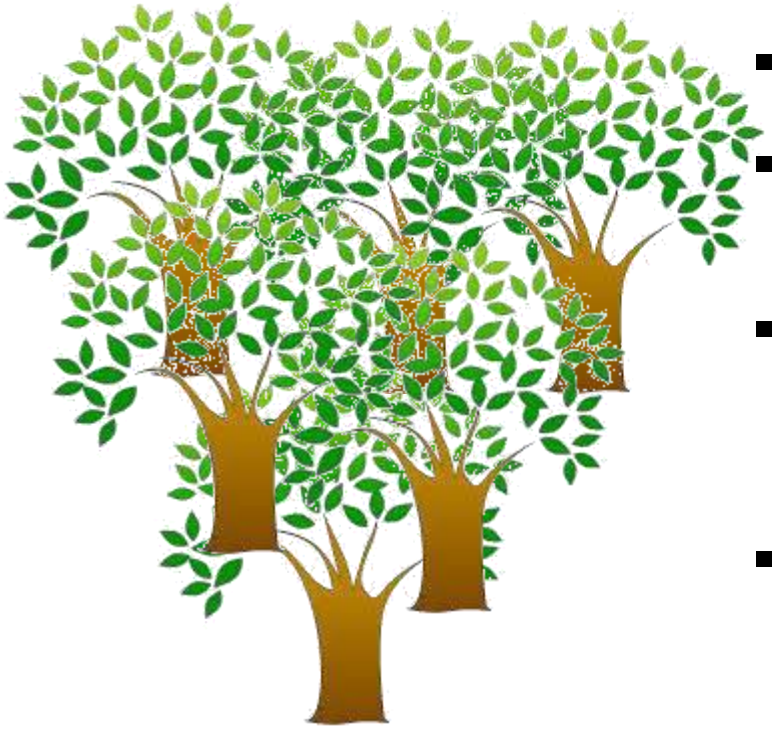
Some Concepts

- **Ensemble** : use of ***multiple learning algorithms*** to obtain better ***predictive performance*** than could be obtained from any of the constituent learning algorithms

- **Bootstrap aggregating**, also called **bagging:** Given a standard training set *D* of size *n*, bagging generates *m* new training sets D_i, each of size *n′*, by sampling from D uniformly with replacement. By sampling with replacement, some observations may be repeated in each D_i. The kind of sample is called Bootstrap. The *m* models are fitted using the above *m* bootstrap samples and combined (aggregated) by averaging the output (for regression) or voting (for classification).

https://en.wikipedia.org/wiki/Bootstrap_aggregating

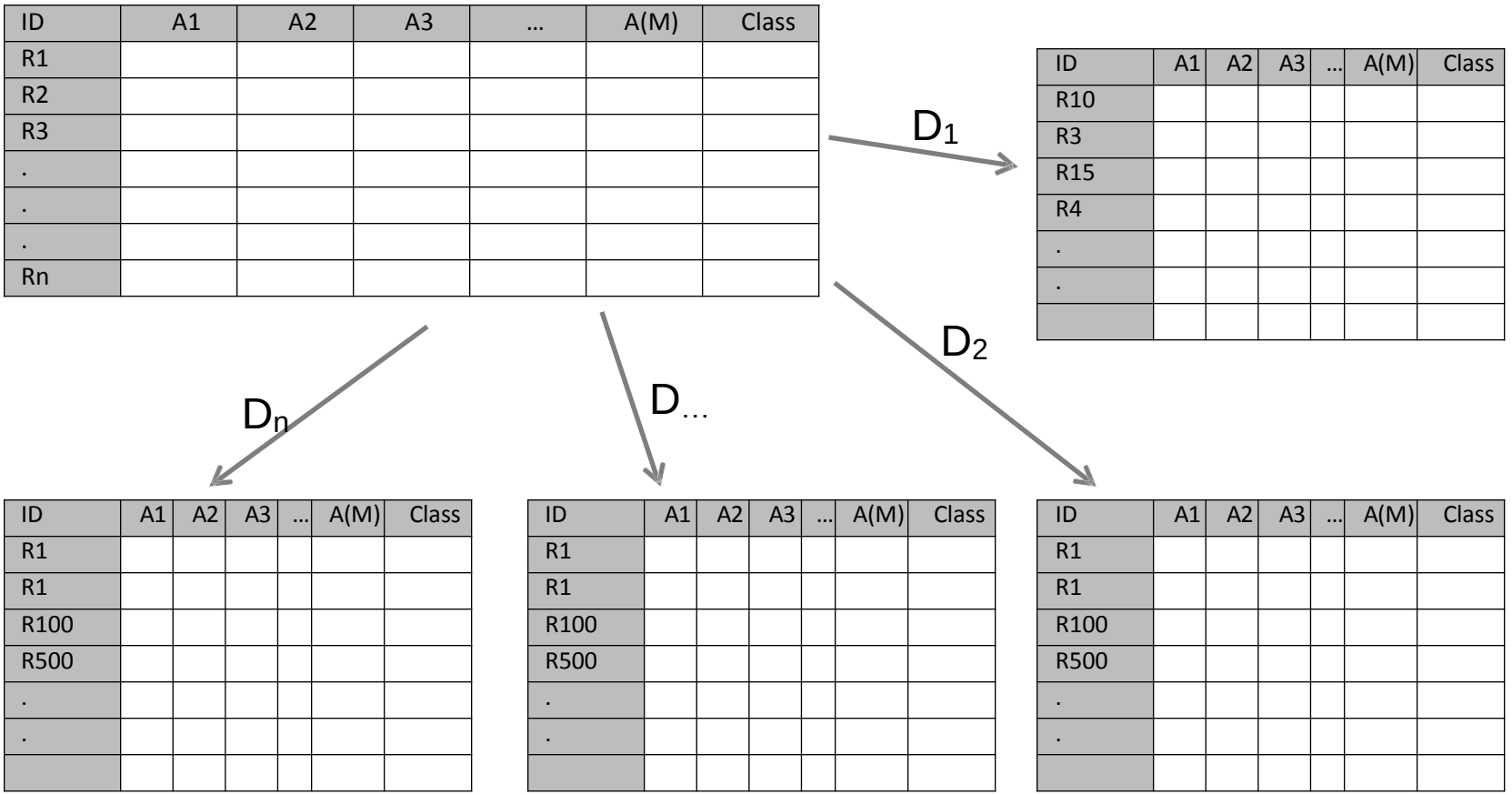
Random Forest



- Ensemble Technique
- Involves constructing multitude of decision trees at training time
- Prediction is based on mode for classification tree and mean for regression tree
- Help reduce over-fitting
  - Note: there is possibility of high over-fitting at individual tree level but averaging removes the bias

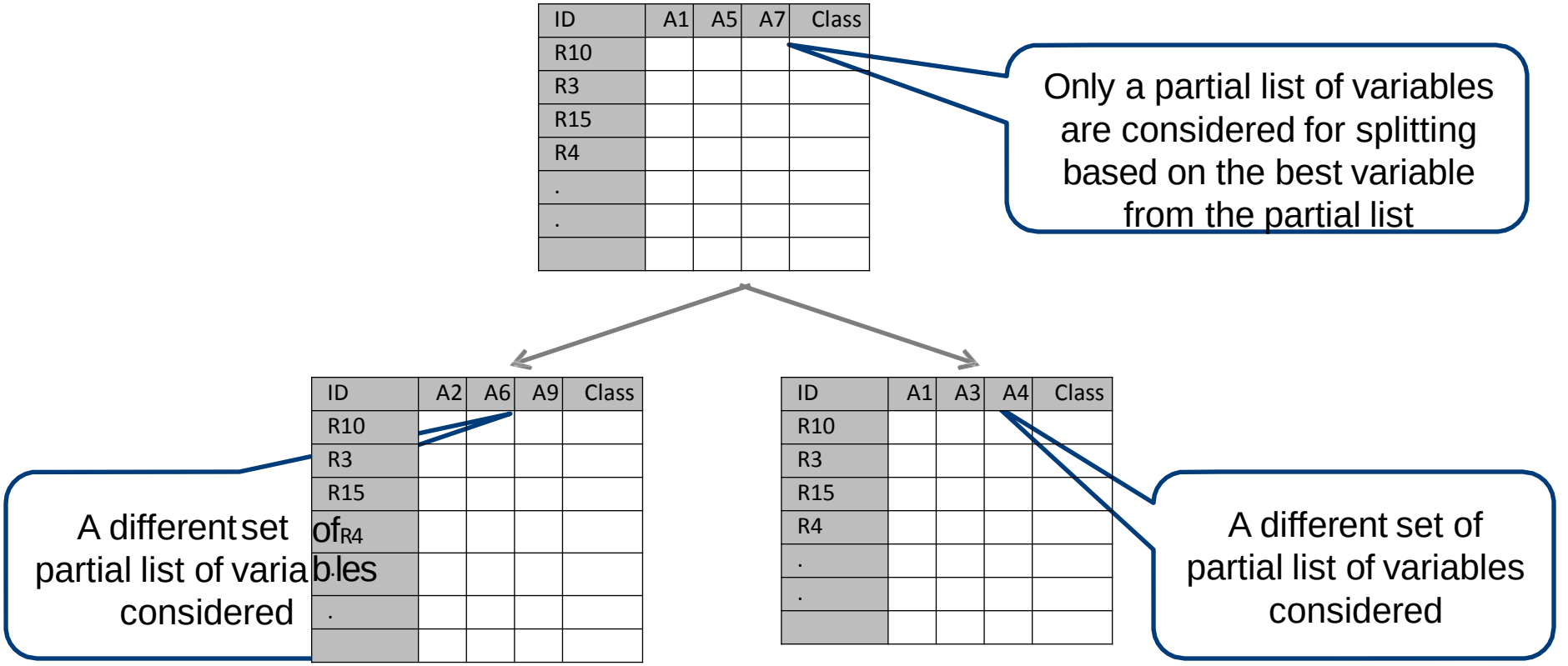
RF Algorithm

- Step 1: Random Sampling with replacement



RF Algorithm… contd

- Step 2: Building the tree for each sample with only partial set of ‘m’ variable being considered at each node
- m << M where M is total number of predictor variables



RF Algorithm… contd

Step 3: Classifying

- Based on ‘n’ samples… ‘n’ tree are built
- Each records is classified based on the n tree
- Final class for each record is decided based on voting

Note: We do not have the pruning step in RF

Some original papers on RF proved that the RF error rate depends on two factors

1. The *correlation* between any two trees in the forest. Increasing the correlation increases the forest error rate.
2. The *strength* of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.
3. Reducing m reduces both the correlation and the strength. Increasing it increases both. Somewhere in between is an "optimal" range of m - usually quite wide

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Building Random Forest in Python

Building the model using RandomForest

importing the data

data <- pd.read_csv("datafile/data.csv")

data.head()

##iImport RandomForest Classifier

From sklearn.ensemble import RandomForestClassifier

Calling syntax to build the RandomForest

RandomForestClassifier(bootstrap = True, criterion = 'gini',

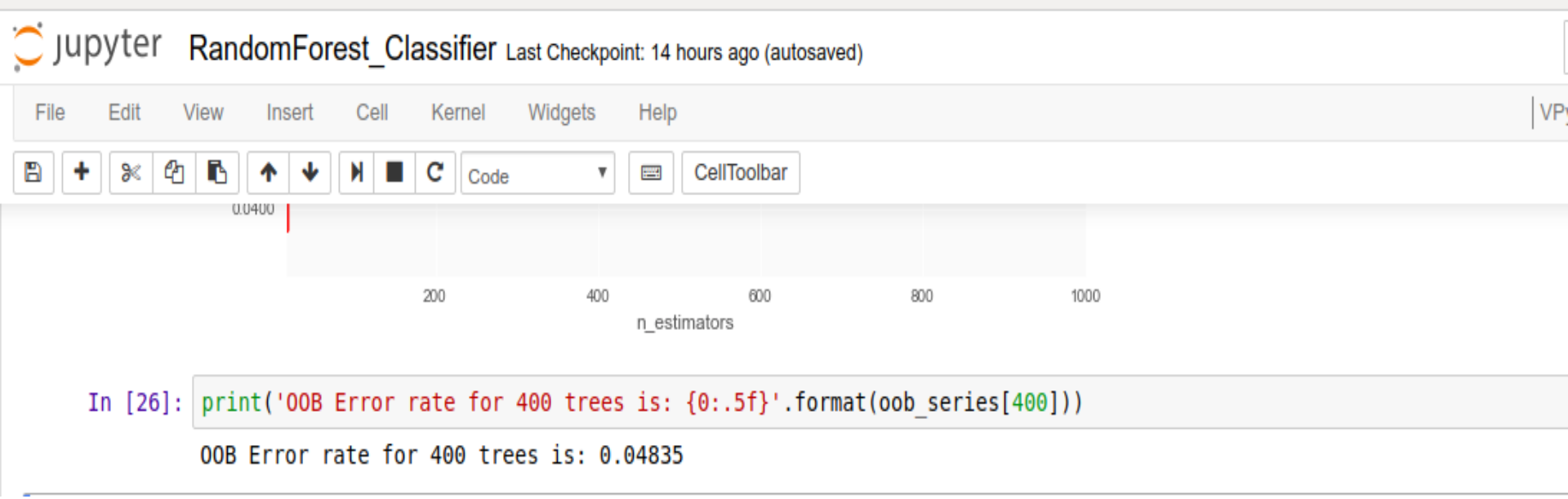
N_estimator=100, ## number of trees to be built

Max_features = auto, ## number of variables randomly sampled as candidate at each split

Max_leaf_node = 10, ## minimum number of records in terminal node

Min_samples_split=TRUE ## minimum samples for the split to occur

OOB Estimate of error rate



<u>OOB Error Rate Computation Steps</u>

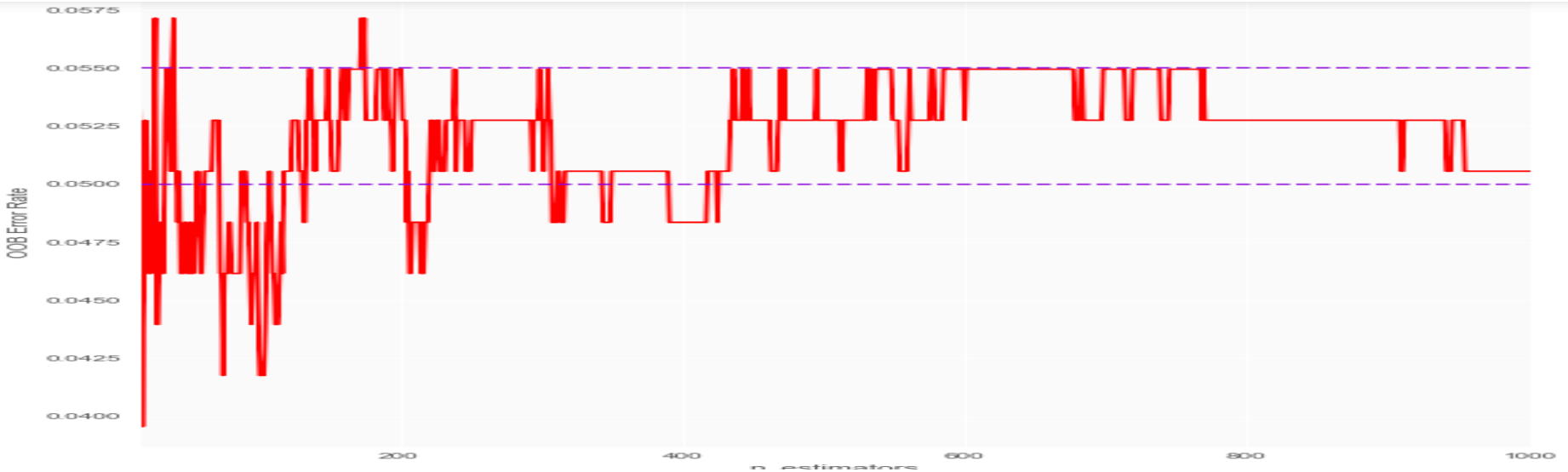
- Sample left out (out-of-bag) in K^{th} tree is classified using the K^{th} tree
- Assume j cases are mis-classified
- Proportion of time that j is not equal to true class averaged over all cases is the oob estimate of error rate

OOB Error Rate … contd

- OOB Estimate of Error Rate is dependent on two key factors
  - n_estimators
  - Max_features

ax.set_facecolor('#fafafa')

oob_series.plot(kind 'line, color 'red)

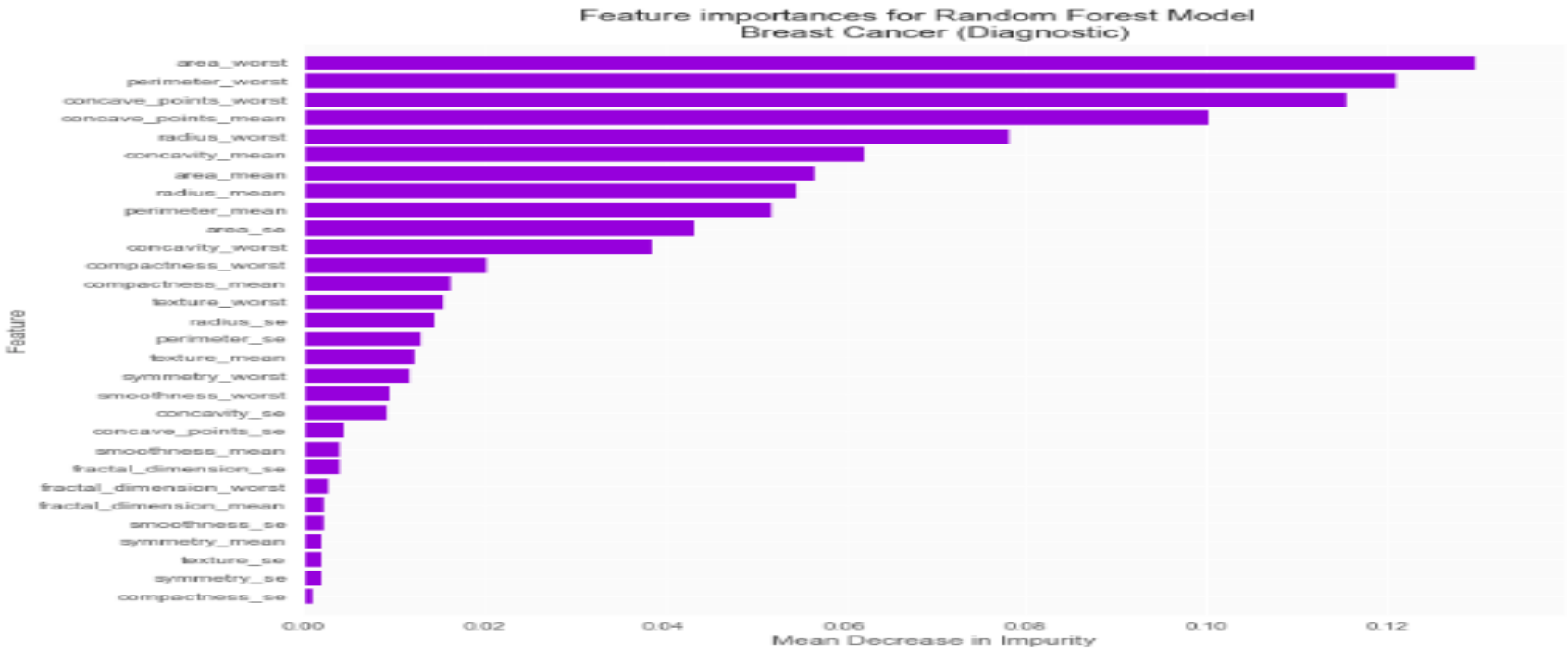


Variable Importance

List the importance of the variables.

var_imp_rf = variable_importance(fit_rf)

Importances_rf = var_imp_rf['importance']



Variable Importance

- Random Forest computes two measures of Variable Importance
  - Mean Decrease in Accuracy
  - Mean Decrease in Gini

- Mean Decrease in Accuracy is based on permutation
  - Randomly permute values of a variable for which importance is to be computed in the OOB sample
  - Compute the Error Rate with permuted values
  - Compute decrease in OOB Error rate (Permuted - Not permuted)
  - Average the decrease over all the trees

- Mean Decrease in Gini is computed as "**total decrease in node impurities from splitting on the variable, averaged over all trees**"

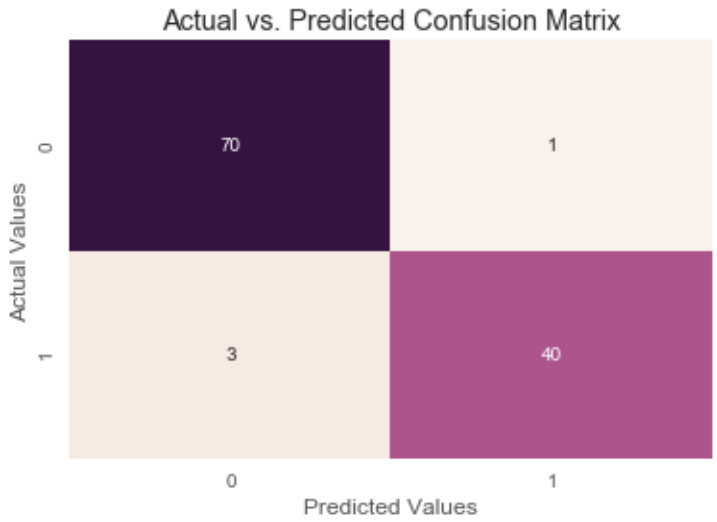
Finding optimal values using GridSearchCV

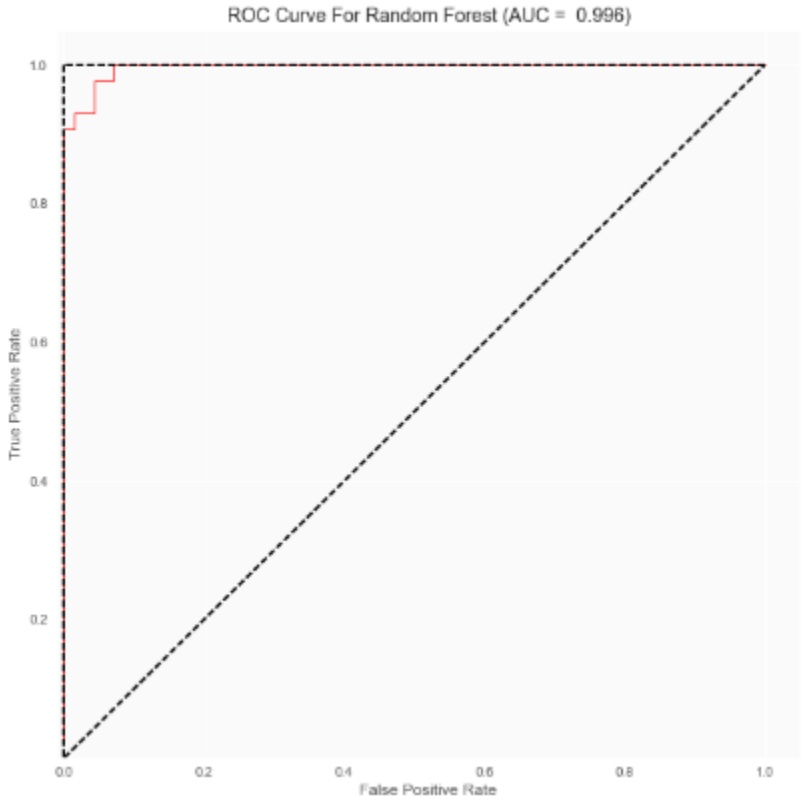
```
np.random.seed(42)
start = time.time()

param_dist = {'max_depth': [2, 3, 4],
              'bootstrap': [True, False],
              'max_features': ['auto', 'sqrt', 'log2', None],
              'criterion': ['gini', 'entropy']}

cv_rf = GridSearchCV(fit_rf, cv = 10,
                     param_grid=param_dist,
                     n_jobs = 3)

cv_rf.fit(training_set, class_set)
print('Best Parameters using grid search: \n', cv_rf.best_params_)
end = time.time()
print('Time taken in grid search: {0: .2f}'.format(end - start))
```

```
Best Parameters using grid search:
 {'bootstrap': True, 'criterion': 'gini', 'max_depth': 3, 'max_features': 'log2'}
Time taken in grid search:  6.18
```

Measuring RF Model performance

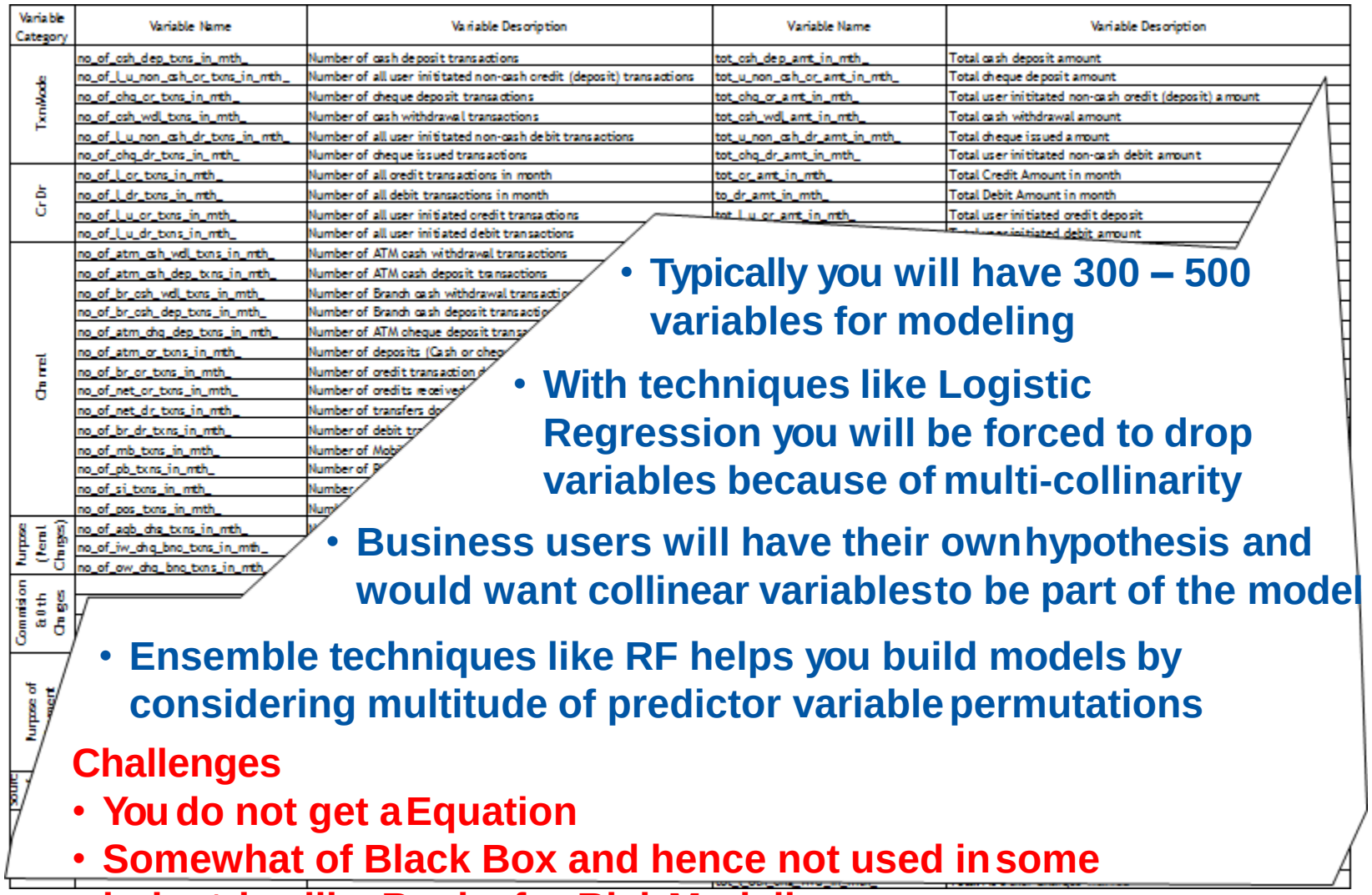
Syntax remains same as for the earlier model





Why I like RF technique?

- … very good technique to pacify Business Users



Variable Category	Variable Name	Variable Description	Variable Name	Variable Description
TxnMode	no_of_csh_dep_txns_in_mth_	Number of cash deposit transactions	tot_csh_dep_amt_in_mth_	Total cash deposit amount
	no_of_l_u_non_csh_cr_txns_in_mth_	Number of all user inititated non-cash credit (deposit) transactions	tot_u_non_csh_cr_amt_in_mth_	Total cheque deposit amount
	no_of_chq_cr_txns_in_mth_	Number of cheque deposit transactions	tot_chq_cr_amt_in_mth_	Total user inititated non-cash credit (deposit) amount
	no_of_csh_wdl_txns_in_mth_	Number of cash withdrawal transactions	tot_csh_wdl_amt_in_mth_	Total cash withdrawal amount
	no_of_l_u_non_csh_dr_txns_in_mth_	Number of all user inititated non-cash debit transactions	tot_u_non_csh_dr_amt_in_mth_	Total cheque issued amount
	no_of_chq_dr_txns_in_mth_	Number of cheque issued transactions	tot_chq_dr_amt_in_mth_	Total user inititated non-cash debit amount
Cr Dr	no_of_l_cr_txns_in_mth_	Number of all credit transactions in month	tot_cr_amt_in_mth_	Total Credit Amount in month
	no_of_l_dr_txns_in_mth_	Number of all debit transactions in month	to_dr_amt_in_mth_	Total Debit Amount in month
	no_of_l_u_cr_txns_in_mth_	Number of all user initiated credit transactions	tot_l_u_cr_amt_in_mth_	Total user initiated credit deposit
	no_of_l_u_dr_txns_in_mth_	Number of all user initiated debit transactions		
Channel	no_of_atm_csh_wdl_txns_in_mth_	Number of ATM cash withdrawal transactions		
	no_of_atm_csh_dep_txns_in_mth_	Number of ATM cash deposit transactions		
	no_of_br_csh_wdl_txns_in_mth_	Number of Branch cash withdrawal transactio		
	no_of_br_csh_dep_txns_in_mth_	Number of Branch cash deposit transactio		
	no_of_atm_chq_dep_txns_in_mth_	Number of ATM cheque deposit transa		
	no_of_atm_cr_txns_in_mth_	Number of deposits (Cash or cheq		
	no_of_br_cr_txns_in_mth_	Number of credit transaction		
	no_of_net_cr_txns_in_mth_	Number of credits receive		
	no_of_net_dr_txns_in_mth_	Number of transfers d		
	no_of_br_dr_txns_in_mth_	Number of debit tr		
	no_of_mb_txns_in_mth_	Number of Mobi		
	no_of_pb_txns_in_mth_	Number of P		
	no_of_si_txns_in_mth_	Number		
	no_of_pos_txns_in_mth_	Numb		
Purpose (Penal Charges)	no_of_aqb_chq_txns_in_mth_			
	no_of_iw_chq_bnc_txns_in_mth_			
	no_of_ow_chq_bnc_txns_in_mth_			
Commission & Other Charges				
Purpose of Loan				
Source				

- **Typically you will have 300 – 500 variables for modeling**
- **With techniques like Logistic Regression you will be forced to drop variables because of multi-collinarity**
- **Business users will have their own hypothesis and would want collinear variables to be part of the model**
- **Ensemble techniques like RF helps you build models by considering multitude of predictor variable permutations**

Challenges

- **You do not get a Equation**
- **Somewhat of Black Box and hence not used in some industries like Banks for Risk Modeling**

Questions?? … Thankyou

Contact
Kuls.utkarsh1205@gmail.com