Project Report

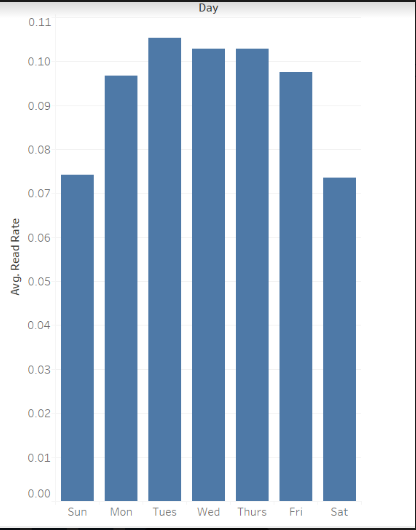
Preprocessing:

* Removed “from\_domain\_hash”, “Domain\_extension” and “from\_domain\_hash” columns from the training dataset as they have many levels in them
* Identified that only one row has missing values and deleted it
* Identified number of outliers in each column
* Checked if any column has a zero or near Zero variance
* Performed multicollinearity test and checked correlation in independent variables
* Created dummy variables for column “day”

I chose to perform 5 fold cross validation with R2 as evaluation metric to compare the different models. Cross validation was performed to reduce the variance of resulting metrics.

EDA:

* Plotting a scatter plot for campaign\_size and unique\_user\_cnt shows that both are highly correlated.
* There are around more than 1% of campaigns where unique user is only 1.
* From the analysis it is clear that a campaign which starts on Saturday and Sunday have less read\_rate compared to weekday as users tend not to check their email on weekends. Tuesday have a high read\_rate in the weekends.



Modeling:

***Linear Regression***: Tried to fit the model and evaluated the R2 with cross validation. Got an adjusted R2 of 0.715. Plotted the residual plots to check if errors are randomly distributed. Also checked the assumption of constant variance (Homoscedasticity) using ncvTest function. Performed stepwise regression and did not have substantial improvement.

Performed Lasso regression by tuning the alpha value. With alpha of 0.001 we get R2 of 0.69 which is less than linear regression which mean we need to consider all the independent variables in regression to get the better result.

***Random Forest:*** Performed parameter tuning on random forest and identified 600 to be best fit for n\_estimators. Got R2 of 0.784 which is a huge improvement over the linear regression result. Avg\_are most important variables that we got from the feature importance given by model

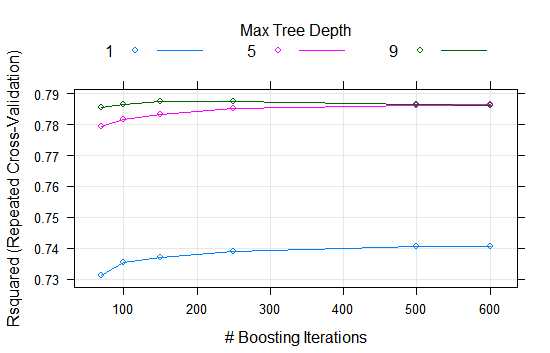
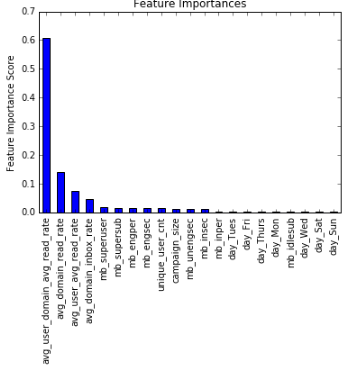


Figure1.1 a) Feature Importance of Random Forest

b) R2 for each of the combination of parameters for GBM

***GradientBoostingMachines:*** This is an ensemble technique which learns from the errors of previous model and is executed linearly. Hence this take lot of time to execute. There are many parameters that can be tuned which control the over fitting and complexity of model. Performed GridSearchCV to find out optimal parameters and got R2 of 0.793 which is a slight improvement over random forest.

Also tried out Xgboost which is a variant of GBM and which runs in parallel unlike GBM. Got R2 of 0.795 which was slight improvement of GBM.

Conclusion:

Based on the analysis we suggest that campaign should be started on Tuesday to get the highest read rate. From random forest model feature importance we can say that Avg\_user\_domain\_avg\_read\_rate and avg\_domain\_read\_rate plays important role in predicting the read\_rate.

Provided with some more domain knowledge we can compute the actual business value for each of the model. If we choose to move this model to production, I would like to go with Random Forest model compromising on the metrics as it would be faster compared to GBM and Xgboost. As each of the tree is independent we can run this model parallel and hence would be best suit for production.