# Model Building – Bulletin Board Post Recommendation Classification

***Data & Model Summary***

|  |  |
| --- | --- |
| Data Dimensions | 16423 entries and 15 features |
| Output Variable | Recommended (Type = Classification) |
| Models Tested | SVM Logistic Regression Naïve Bayes KNN Decision Trees  Random Forest  Default Model |
| Feature Selection via | *stepAIC (for Logistic Regression)*  [used exhaustive feature selection – can be justifiable as number of features are less]  *rfe,rfeControl (for random forest)*  *Decision Tree does feature selection itself* |

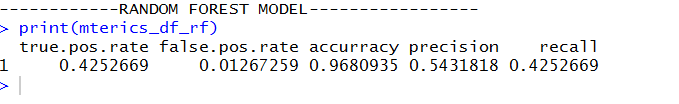
|  |
| --- |
| **Comparison of All the classifiers (when all features are used as predictors)**  ------------NAIVE BAYES-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.6850534** 0.05724734 **0.9339341** 0.2977572 0.6850534  ------------SUPPORT VECTOR MACHINES-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.3309609 0.006493916 **0.9708336** 0.6435986 0.3309609  ------------LOGISTIC REGRESSION-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.2651246 0.007124393 **0.9679717** 0.5687023 0.2651246  ------------KNN -----------------  true.pos.rate false.pos.rate accurracy precision recall  0.07829181 0.005863439 **0.9627961** 0.3211679 0.07829181  ------------DEFAULT MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  1 1 **0.9657797** 0.9657797 1  ------------DECISION TREE MODEL-----------------    true.pos.rate false.pos.rate accurracy precision recall  0.3825623 0.0109703 **0.9682762** 0.5526992 0.3825623  ------------RANDOM FOREST MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.4572954** 0.007944014 **0.9737563** 0.6710183 0.4572954 |

What is interesting to note here is the difference in the true positive rate. Naïve Bayes Classifier gives us the highest true positive rate

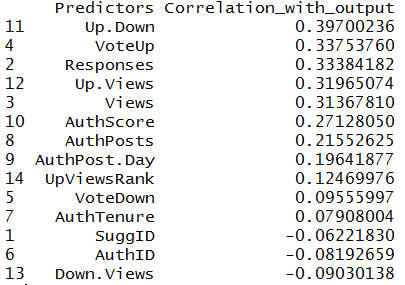
|  |
| --- |
| **After removing SuggID and AuthID (Because these are random and don’t add any value**  **in predicting the classification)**  ------------NAIVE BAYES-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.6886121 0.05686905 0.9344212 0.3002327 0.6886121  ------------Support Vector Machines-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.2900356 0.006746107 0.9691896 0.6037037 0.2900356  ------------LOGISTIC REGRESSION-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.2455516 0.007124393 0.967302 0.5498008 0.2455516  ------------KNN -----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.2491103** 0.01027678 0.9643792 0.4620462 0.2491103  ------------DEFAULT MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  1 1 0.9657797 0.9657797 1  ------------DECISION TREE MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.405694 0.01235735 0.9677282 0.5377358 0.405694  ------------RANDOM FOREST MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  0.4733096 0.009520207 0.9727821 0.6378897 0.4733096 |
| **Original Model with only 8 variables ( without any derived variables)**  ------------NAIVE BAYES-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.4715302** 0.04325074 0.9401449 0.278654 0.4715302  ------------Support Vector Machines-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.3398577** 0.006620011 0.9710163 0.6452703 0.3398577  ------------LOGISTIC REGRESSION-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.1690391** 0.006304773 0.9654752 0.4871795 0.1690391  ------------KNN -----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.2544484** 0.01008764 0.9647446 0.4719472 0.2544484  ------------DEFAULT MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  1 1 0.9657797 0.9657797 1  ------------DECISION TREE MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.3024911** 0.009457159 0.9669975 0.53125 0.3024911  ------------RANDOM FOREST MODEL-----------------  true.pos.rate false.pos.rate accurracy precision recall  **0.4217082**  0.008700586 0.9718078 0.632 0.4217082 |
| |  | | --- | |  | |

As we see the ***true positive rate*** has significantly improved by the introduction of the new variables**.**

**Random Forest with top 5 features**



**Important features only based on correlation**



**Important points to note**

* For logistic regression using the stepAIC function for exhaustive search gave all the features
* For random Forest the top 5 features (VoteUp, Views, Up.Down, Down.Views, Responses) gave the same true positive rate as did all the features, it means all the other features are not adding value to our classification model