# Summary and Further improvements

## Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Training Accuracy | Training Accuracy after Hyperparameter Tuning | Testing Accuracy |  | Model |
| Random Forest | Bag of Word | 96.35% | 96.74% | 96.40% |  | Model 1 |
|  | TF-IDF | 96.01% | 96.80% | 96.85% |  | Model 2 |
|  |  |  |  |  |  |  |
| Logistic Regression | Bag of Word | 96.91% | 97.02% | 98.20% |  | Model 3 |
|  | TF-IDF | 97.08% | 97.42% | 97.98% |  | Model 4 |

The above table shows the training accuracy, training accuracy after hyperparameter tuning and the testing accuracy of all four models. The greater a model's testing accuracy is, the better it can generalize to unknown data and generate better predictions and insights, resulting in more business value. After comparing the testing accuracy of the four models in the table above, it is determined that the model 3, the logistic regression model developed using Bag of Word values has the highest test score. As a result, it is selected as the final classification model for BBC news website.

## Further improvements

### Using Decision Tree Classifier

We have built four classification models using two Machine Learning classifiers: Random Forest Classifier and Logistic Regression Classifier. Since there are many other classification algorithms which we can try out, we can consider to further improve the models using Decision Tree classifier. Decision Tree classifier offers several advantages: it is simple to read and understand; it necessitates less data pre-processing from the user, such as the elimination of the need to normalize columns; It is ideal for variable selection and may be used for feature engineering, such as anticipating missing values; because of the non-parametric character of the algorithm, the decision tree makes no assumptions regarding distribution. *(Refer to figures below for the codes used for building decision tree classifier.)*

### Codes for building Decision Tree Classifier using Bag of Words

### 

### 

### Codes for building Decision Tree Classifier using TF-IDF values

### 

