Comparison of Logistic Regression and Random Forest on predicting email spam

By Imsa Zulfiqar

Background

Spam emails are emails sent to recipients without their knowledge or consent, which often contain marketing.

Problem Description

By classifying emails as spam, we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.

Approach

1. Predict spam emails using Logistic Regression and Random Forest
2. Compare the results of logistic regression with a previous study by Shahbaz Ahmad Khanday, and Suraiya Parveen.

Exploratory Data Analysis

- The dataset used is Spambase Data Set from UCI Machine Learning Repository.
- The original dataset consists of 4,601 rows and 58 columns; there are 57 continuous real attributes and 1 nominal class attribute (label). Email is spam if label=1, not spam if label=0.
- No missing values were found in the dataset.
- The dataset is slightly imbalanced with 2788 emails classified as not spam and 1813 emails classified as spam, as shown in Figure 1.
- The dataset includes no stop words.
- The dataset was checked to see which words most contribute to the email being considered spam. Figure 2 shows the top 15 words with the most frequent occurrence among which "you" and "your" contributed the most to the emails being classified as spam.
- Figure 3 shows the word cloud for an email specified as spam. The most percentage was of the words "length", "you", "your", "mail", "will" and "000".
- The Correlation heatmap shown in figure 4, the correlation matrix, investigates the correlation with the target variable "label". None of the features is highly correlated.

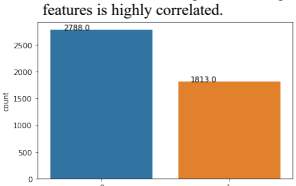

Figure 1. Class imbalance

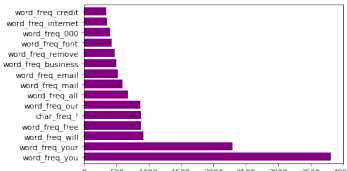

Figure 2. Top 15 words used in spam emails.

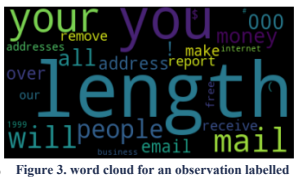

Figure 3. word cloud for an observation labelled as spam.

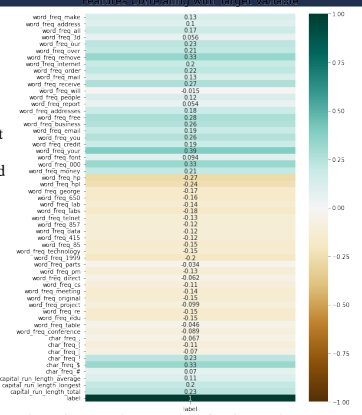

Figure 4. Pearson's correlation of variables with the target.

Why these models were chosen?

Logistic Regression

- By estimating probabilities with the help of a logistic function, logistic regression analyses the connection between the categorical dependent variable and one or more independent variables.
- Logistic regression is a supervised learning algorithm, and it performs some basic tests on the given distribution of data which involves finding and calculating some statistical domains like mean and standard deviation.

Advantages

- It makes no assumptions about distributions of classes in feature space.
- It can interpret model coefficients as indicators of feature importance.
- Easier to implement and more efficient to train.

Disadvantages

- The limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
- It is sensitive to outliers.

Random Forest

- Random forest is a supervised learning algorithm that relies on collecting various decision trees to arrive at any solution.
- It is an ensemble algorithm that considers the results of more than one algorithm of the same or different kinds of classification.

Advantages

- Random Forests are not influenced by outliers to a fair degree. It does this by binning the variables.
- Random Forests generally provide high accuracy and balance the bias-variance trade-off well. Since the model's principle is to average the results across the multiple decision trees it builds, it averages the variance as well.

Disadvantages

- Random forest is like a black box algorithm, you have very little control over what the model does.
- Not suitable for linear methods with a lot of sparse features

Hypothesis Statement

- Logistic Regression is likely to perform better than random forest since it makes no assumptions about distributions of classes in feature space.
- The RF is likely to spend longer testing time than Logistic regression.

Methodology

Baseline model

- Split data into a 70: 30 split for train and test data using hold out approach. The test data remained unseen to models until end.
- The model was trained using all predictors.
- Lasso regularization was used for logistic regression to avoid overfitting.
- For random forest, out-of-bag-Prediction was used as a hyperparameter which computes predicted responses using the trained bagger for out-of-bag observations in the training data

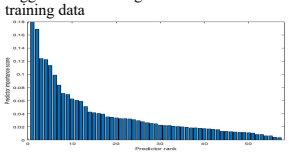

Figure 5. Ranking predictors

Optimised Model

- The dataset was split using 5 k-fold cv into training and testing sets.
- Feature selection was done by ranking features for classification using the minimum redundancy maximum relevance (MRMR) algorithm.
- Top 20 predictors were selected because after these, the trend didn't change much as shown in figure 5.
- The models were trained only on the predictors selected from MRMR algorithm.
- For Logistic regression, ridge regularization was used and a solver combination of "Stochastic gradient descent", and "Limited-memory BFGS" was used to balance optimization speed and accuracy.

Evaluation Metrics and Results

After training, the models were tested using the test dataset. The evaluation metrics used to evaluate the performance all the models included on the unseen dataset included:

- Confusion matrix
- F1-score
- Precision and Recall
- Accuracy
- OOB error
- ROC-curve

The best model was identified as the model which used feature selection and k-fold cv. This had the highest AUC, precision and recall and accuracy.

	Logistic Regression (base)	Random Forest (base)	Logistic regression Optimised	Random forest Optimised
Accuracy	0.7413	0.94457	0.99891	0.89565
ROC score	0.8096	0.97757	0.99839	0.94484
Precision	0.81775	0.95936	0.99818	0.96607
F1-score	0.79471	0.95515	1	0.91851
Recall	0.77293	0.95096	0.99909	0.8754

Table 1. Performance Evaluation of all models

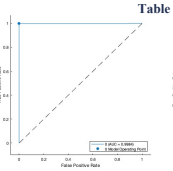

Figure 6. ROC LR (optimised)

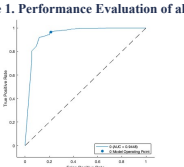

Figure 7. ROC RF (optimised)

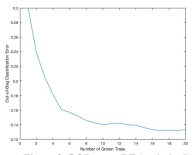

Figure 8. OOB-error RF (optimised)

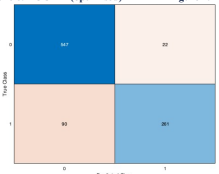

Figure 9. Confusion matrix RF (optimised)

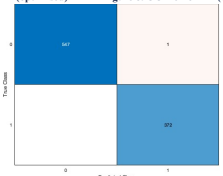

Figure 10. Confusion matrix LR (optimised)

Analysis and Critical Reflections

- The AUC score of the optimised model of Logistic regression was higher than that achieved by Shahbaz Ahmad Khanday, and Suraiya Parveen.
- Ridge regression was used in the optimised model because Collinear predictors, small numbers of predictors, and data with subgroups are all good candidates for Ridge regression and it creates a robust model whereas LASSO focuses more on creating a sparse model.
- For feature selection, the MRMR algorithm was used. Figure 5 shows the top 20 predictors. The drop in score between the first two important predictors is large, while the drops after the 20th predictor are relatively small. A drop in the importance score represents confidence in feature selection. Therefore, the large drop implies that the software is confident in selecting the most important predictor. The small drops indicate that the difference in predictor importance are not significant.
- After feature selection, we obtained the final trained logistic model and random forest, results can be seen in figures 6 and 7 showing the roc score.
- The optimised random forest model did not perform well even after the feature selection as compared to the base model which was trained on all the features.
- Performance of the optimised Logistic regression model improved significantly as compared to the base model. The performance improved was mainly because of using ridge regression with the solver combination of "Stochastic gradient descent", and "Limited-memory BFGS".
- Out-of-bag error for the random forest reduces as the number of trees grow as shown in figure 8.

Lessons Learned

- When interpreting, it's important to pay attention to imbalance labels because accuracy alone is not a good evaluation approach
- Attention should be paid for choosing the hyperparameters for random forests since there are a lot of them.

Future Work

- Applying the Synthetic Minority Oversampling Technique (SMOTE) to resolve the problem of imbalance labels.
- Applying chi-square tests feature selection to improve the performance of random forest.
- More predictors can be added to the dataset for a more accurate classification of emails as spam and training with other algorithms such as Naive Bayes can be done for prediction.

References

https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/
https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04
https://towardsdatascience.com/the-basics-logistic-regression-and-regularization-828b0d2d206c
https://osall.eu/pdf/10.4108/eai.27-2-2020.2303291
https://ico.org.uk/your-data-matters/online/spam-emails/
https://uk.mathworks.com/matlabcentral/answers/399062-how-to-set-tree-arguments-in-treebagger
https://uk.mathworks.com/matlabcentral/answers/273113-explanation-of-hyperparameter-tuning-procedure-for-regression-tree-ensembles
https://uk.mathworks.com/help/stats/treebagger.html
https://uk.mathworks.com/help/stats/fscmrmr.html
https://uk.mathworks.com/help/stats/feature-selection.html
https://uk.mathworks.com/help/stats/rocmetrics.html#mw_438f9e0d-3715-4861-bc88-0a126fe9540a
https://uk.mathworks.com/help/stats/confusionmat.html
https://uk.mathworks.com/help/stats/classificationlinear-class.html
https://uk.mathworks.com/help/stats/tune-random-forest-using-quantile-error-and-bayesian-optimization.html