EMAIL SPAM CLASSIFIER: Performance Evaluation of Classification Models with Resampling Techniques

Kunjkumar Modi

Kunjkumar.modi@ontariotechu.net

University of Ontario Institute of Technology,

Oshawa, Ontario, Canada

Abstract

This report evaluates the performance of various classification algorithms (Logistic Regression, Naïve Bayes, SVM, Decision Tree, and Random Forest) on a email spam classification task. To address class imbalance, the models were tested with different resampling techniques (Random Oversampling, Random Undersampling, and SMOTE). After cross-validation, Random Forest with Random Oversampling achieved the best balance between accuracy and recall. The findings suggest that Random Forest with Random Oversampling is the most effective model, particularly for tasks requiring minimized false negatives.

1. Introduction

This report presents an analysis of various classification algorithms applied to a dataset using different sampling techniques to handle class imbalance. The goal is to evaluate the performance of the models using accuracy, precision, recall, and F1-score.

2. Dataset Information

A. Descriptive Statistics

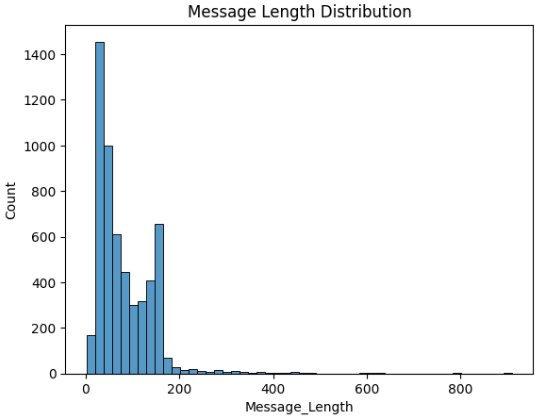
The dataset consists of 5,572 messages classified into two categories: spam and ham (non-spam). The most frequent message in the dataset is "Sorry, I'll call later," appearing 30 times. The total number of unique messages is 5,157.

B. Class Distribution

The dataset is imbalanced, with 4,825 messages classified as ham and 747 as spam. This imbalance necessitates the use of resampling techniques to improve model performance.

C. Message Length Statistics

Message length varies significantly, with a minimum length of 2 characters and a maximum of 910 characters. The average message length is approximately 80 characters.



3. Methodology

The classification models tested include:

- Logistic Regression
- Naïve Bayes
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

To handle class imbalance, three sampling techniques were applied:

- Synthetic Minority Over-sampling Technique (SMOTE)
- Random Oversampling
- Random Undersampling ()

A cross-validation approach (5-fold) was used to obtain reliable performance metrics.

4. Resampling Methods

- **SMOTE (Synthetic Minority Over-sampling Technique)**
  - Shape of Resampled Data: After applying SMOTE, the shape of the training set becomes (7718, 1000), which means there are 7718 samples with 1000 features each.
  - Class Distribution: The class distribution in the resampled data is balanced, with 3859 samples in each class (1 and 0).

- **Random Over-sampling**
  - Shape of Resampled Data: The shape of the training set after random over-sampling is (7718, 1000), indicating 7718 samples with 1000 features each.
  - Class Distribution: The class distribution is balanced with 3859 samples from each class (0 and 1).
- **Random Under-sampling**
  - Shape of Resampled Data: The shape of the training set after random under-sampling becomes (1196, 1000), which means there are 1196 samples with 1000 features each.
  - Class Distribution: The class distribution is balanced, with 598 samples from each class (0 and 1).

5. SPOT-CHECK CLASSIFICATION RESULTS

SMOTE Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9578	0.7898	0.9329	0.8554
Naïve Bayes	0.9390	0.7056	0.9329	0.8035
SVM	0.9659	0.8323	0.9329	0.8797
Decision Tree	0.9516	0.8417	0.7852	0.8125
Random Forest	0.9740	0.9688	0.8322	0.8953

Random Oversampling Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9740	0.8947	0.9128	0.9037
Naïve Bayes	0.9372	0.6927	0.9530	0.8023
SVM	0.9794	0.9145	0.9329	0.9236
Decision Tree	0.9489	0.8433	0.7584	0.7986
Random Forest	0.9731	0.9760	0.8188	0.8905

Random Undersampling Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9659	0.8627	0.8859	0.8742
Naïve Bayes	0.9274	0.6574	0.9530	0.7781
SVM	0.9507	0.7500	0.9463	0.8368
Decision Tree	0.9318	0.6995	0.8591	0.7711

Random Forest	0.9803	0.9568	0.8926	0.9236

6. Cross-Validation Results

Model Performance with SMOTE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9374	0.9633	0.9096	0.9356
Naïve Bayes	0.9044	0.8963	0.9230	0.9080
SVM	0.9852	0.9718	0.9995	0.9854
Decision Tree	0.9740	0.9744	0.9744	0.9745
Random Forest	0.9935	0.9938	0.9922	0.9928

Model Performance with Random Oversampling

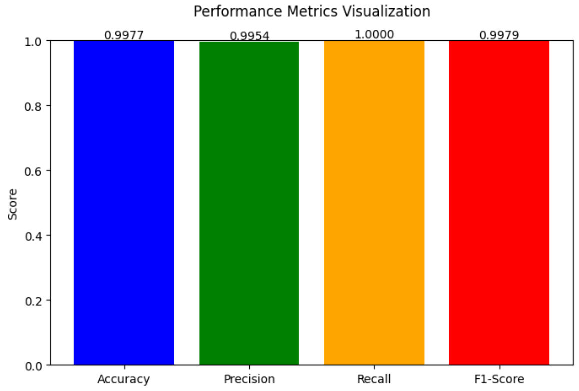
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9675	0.9832	0.9513	0.9669
Naïve Bayes	0.9167	0.9043	0.9404	0.9205
SVM	0.9885	0.9886	0.9883	0.9885
Decision Tree	0.9845	0.9713	0.9995	0.9862
Random Forest	0.9974	0.9969	1.0000	0.9978

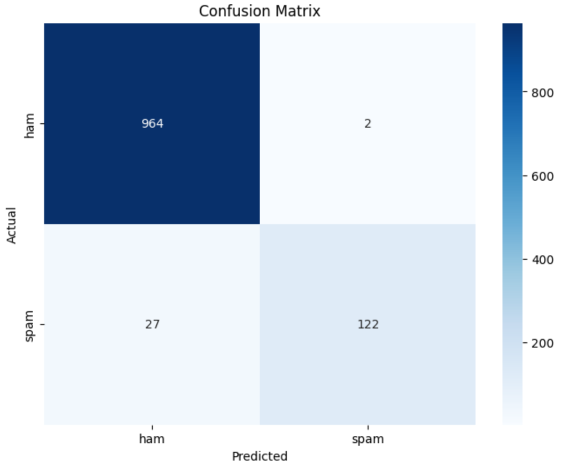
Model Performance with Random Undersampling

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9147	0.9720	0.8546	0.9091
Naïve Bayes	0.8436	0.8470	0.8546	0.8473
SVM	0.9532	0.9577	0.9481	0.9528
Decision Tree	0.9147	0.9238	0.8913	0.9154
Random Forest	0.9389	0.9832	0.8763	0.9312

7. Best Model:

For optimal results, **Random Forest with Random Oversampling** should be used due to its superior accuracy, precision, and recall balance. **SVM with SMOTE** is another strong choice for applications where minimizing false negatives is critical.





8. Conclusion and Recommendations

- Random Forest consistently outperforms other models, achieving the highest accuracy across all sampling methods.
- SVM performs exceptionally well in terms of recall, particularly with SMOTE.
- Logistic Regression provides reliable performance but is slightly outperformed by SVM and Random Forest.
- Naïve Bayes struggles the most, especially with Undersampling.
- Random Oversampling and SMOTE improve recall, which is crucial for fraud detection.

9. References

- (**Dataset Link**)Spam email classification. (2023, December 22). Kaggle. https://www.kaggle.com/datasets/ashfakyeafi/spam-email-classification
- Brownlee, J. (2016) Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End. Machine Learning Mastery, San Francisco.
- Tuanai. (2024, November 20). 💌 Email Spam Classification With Accuracy > 98% 🔰 . https://www.kaggle.com/code/tuanai/email-spam-classification-with-accuracy-98#4.-Split-the-data-set

10. YouTube Video Presentation Link

- https://www.youtube.com/watch?v=ANVBPNhau-A