Project Phase - II

Date : 06 April 2023

EMAIL SPAM DETECTION

Comparing Best ML Models

Presented To

Prof Kunal Kumar

Presented By

Sunidhi Pandey
Shantanu Singh Chandel

Machiine Learning Project

Agenda

Basic Details of the entire Content of the Project.

Objective of the Project

Description of the Project

Requirements

Advantages and Disadvantages

Methodology

Comparison for Best ML model

Future Scope

Project Outcome & Conclusion

Objective

To preface the Goal of the entire proejct to work.

- Understand the Dataset.
- Build classification models to predict whether or not the email is spam.
- Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithms.

Description

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography, etc.

Our collection of spam e-mails came from our postmaster and individuals who had filed spam. Our collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

The dataset, taken from the UCI ML repository, contains about 4600 emails labelled as spam or ham.

Requirements

We have a basic need of Hardware requirement as well as Software requirement.

1	Operating System - 64 bit
2	Dataset file in .csv extension
3	IDE with Libraries
4	Web Browser
5	i3 Processor with 2 GB RAM (Minimum)
6	Input Devices & Screen Resolution (1280*1024)

Advantages & Disadvantages

ADVANTAGES –

❖ Spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria.

❖ Good Efficiency

❖ Greater accuracy comparison using different ML Algorithms.

DISADVANTAGES –

❖ This project in not 100% accurate.

❖ It is possible to make mistake.

❖ Work differently in different environment.

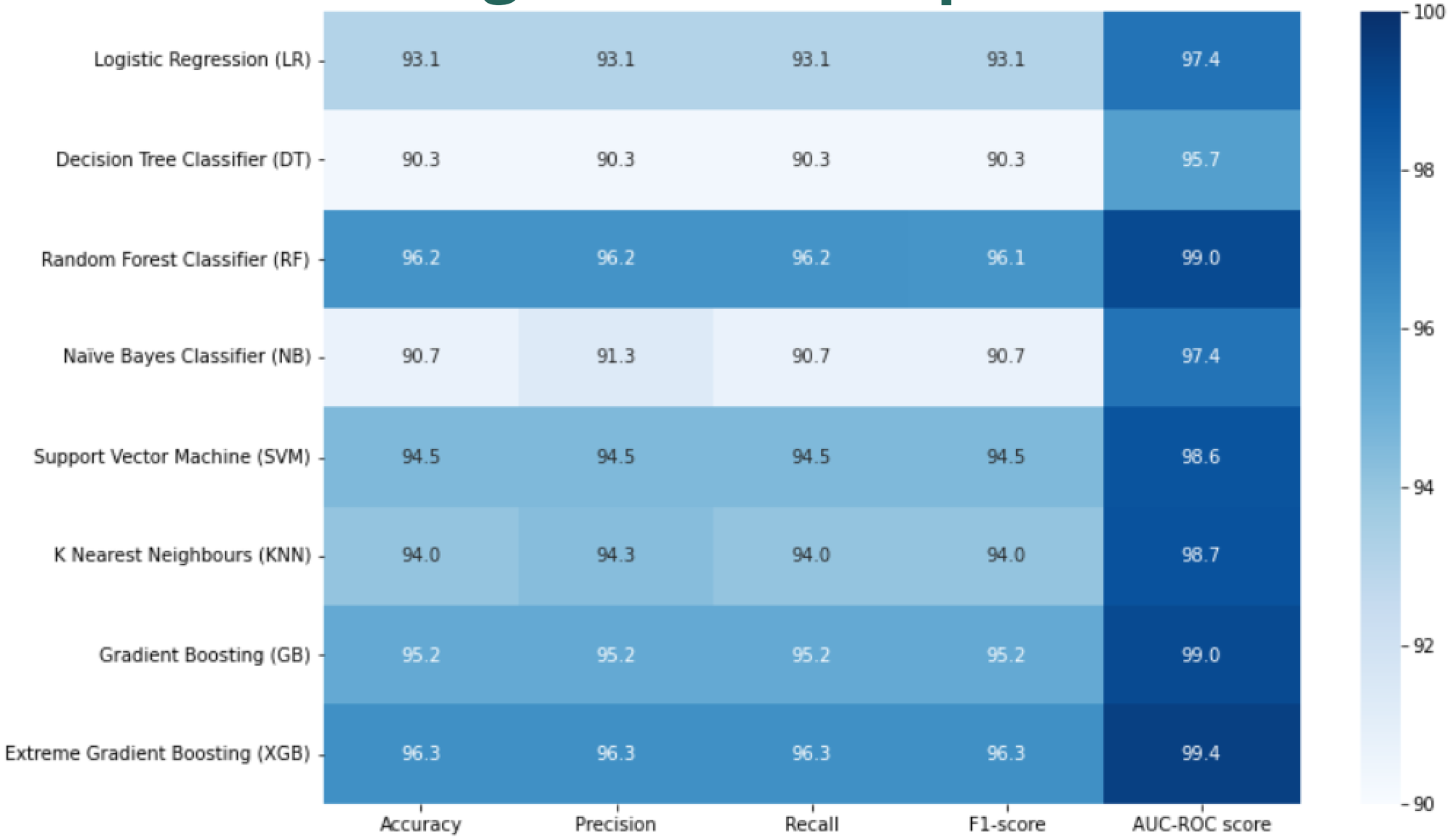
Methodology

Stractegic Plan of Action:

1. Data Exploration
2. Exploratory Data Analysis (EDA)
3. Data Pre-processing
4. Data Manipulation
5. Feature Selection/Extraction
6. Predictive Modelling
7. Project Outcomes & Conclusion

LR	Logistic Regression	**SVM**	Support Vector Machine
DT	Decision Tree	**KNN**	K Nearest Neighbours
RF	Random Forest Classifier	**GB**	Gradient Boosting
NB	Naive Bayes Classifier	**XGB**	Extreme Gradient Boosting

ML Algorithms Comparison



Future Scope

- Review spam detection is essential since it can ensure justice for the sellers and retain the trust of the buyer on the online stores.
- The algorithms developed so far have not been able to remove the requirement of manual checking of the reviews. Hence there is scope for complete automation of spam detection systems with maximum efficiency.
- With growing popularity of online stores, the competition also increases. The spammers get smarter day by day and spam reviews become untraceable.
- It is necessary to identify the spamming techniques in order to produce counter algorithms.

Project Outcomes & Conclusions

- The Dataset was quiet small totalling around 4600 samples & after preprocessing 14.6% of the datasamples were dropped.
- The samples were slightly imbalanced after processing, hence SMOTE Technique was applied on the data to balance the classes, adding 16.7% more samples to the dataset.
- Visualising the distribution of data & their relationships, helped us to get some insights on the relationship between the feature-set.
- Feature Selection/Elimininations was carried out and appropriate features were shortlisted.
- Testing multiple algorithms with fine-tuning hyperparamters gave us some understanding on the model performance for various algorithms on this specific dataset.
- The Random Forest Classifier & XG-Boost performed exceptionally well on the current dataset, considering Precision Score as the key-metric.
- Yet it wise to also consider simpler model like Logistic Regression as it is more generalisable & is computationally less expensive, but comes at the cost of slight misclassifications.


Thank you!