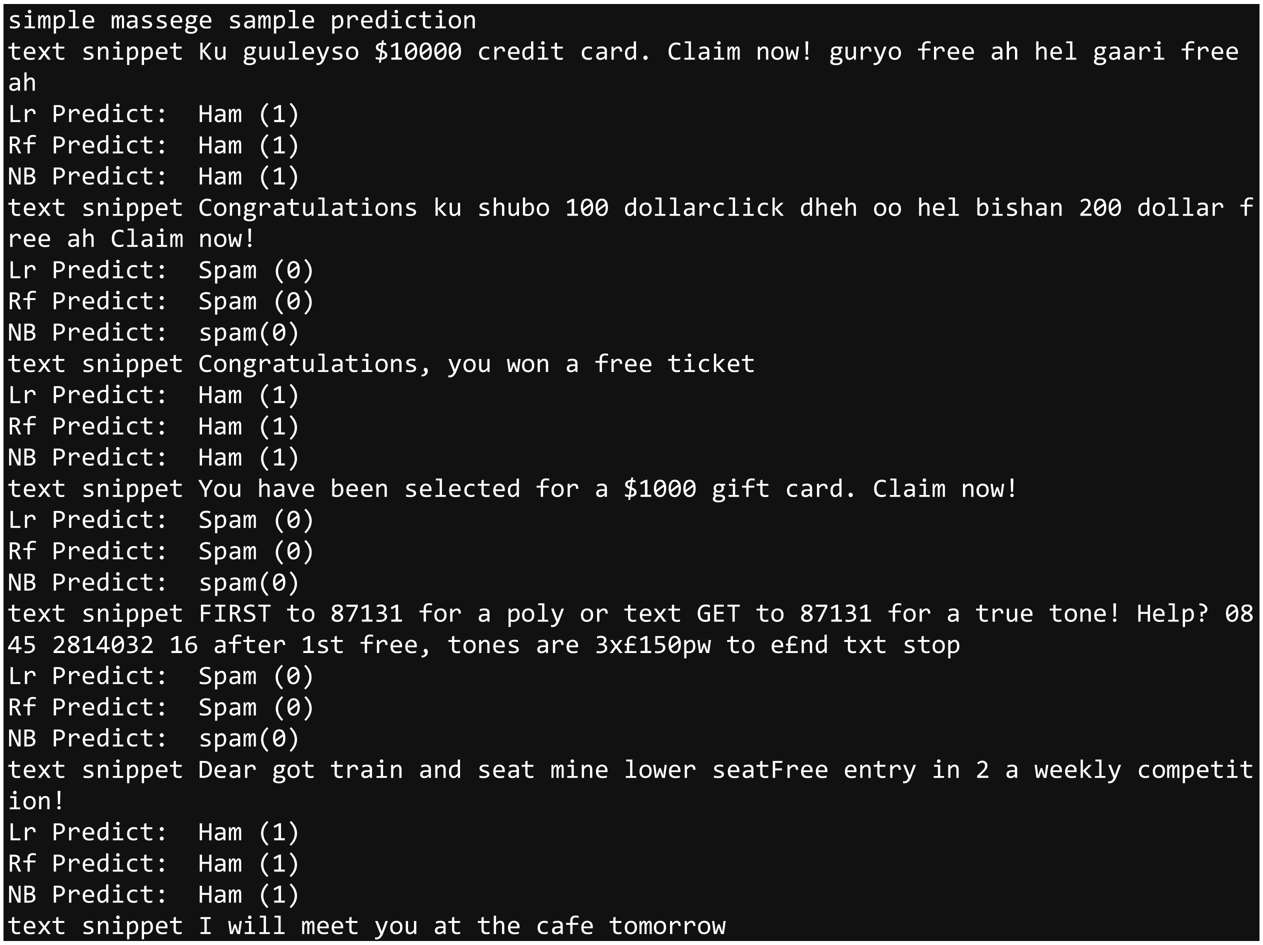
1. What did you implement?

1. i implemented Linear Regression (LR) and Random Forest Regressor (RF) Naive Bayes (NB) to predict detect spam
2. I loaded a cleaned dataset
3. The dataset was split into 80% training and 20% testing
4. I trained 3 models on the training data
5. I evaluated the models using metrics: Accuracy score, precision score, recall score, f1_score, confusion matrix

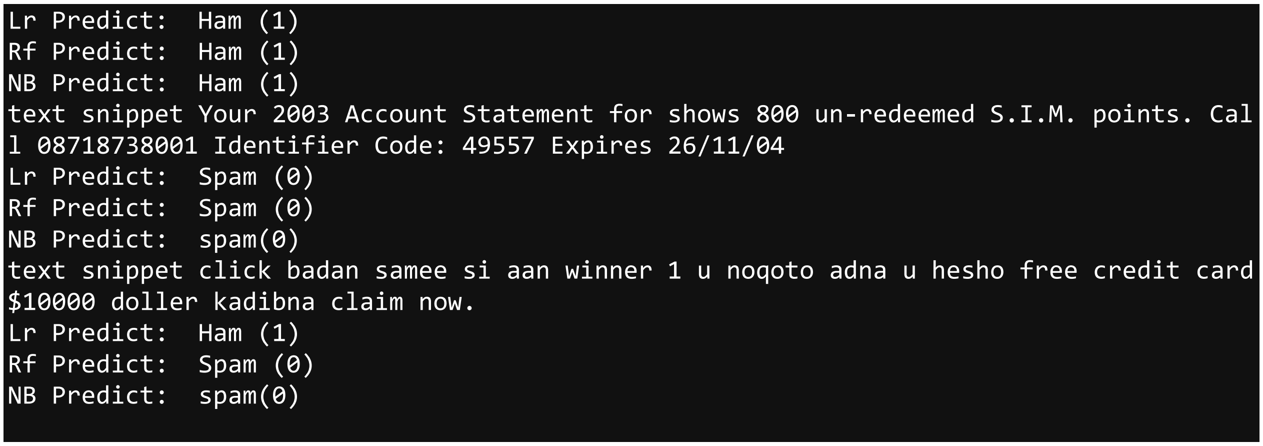
2. Comparison of Models (Sanity Check)

1: Compare the results of the 3 sanity check messages

waxan ku tijaabiyeye models ka 10 sanity check messages 7 spam ah iyo 4 Ham ah



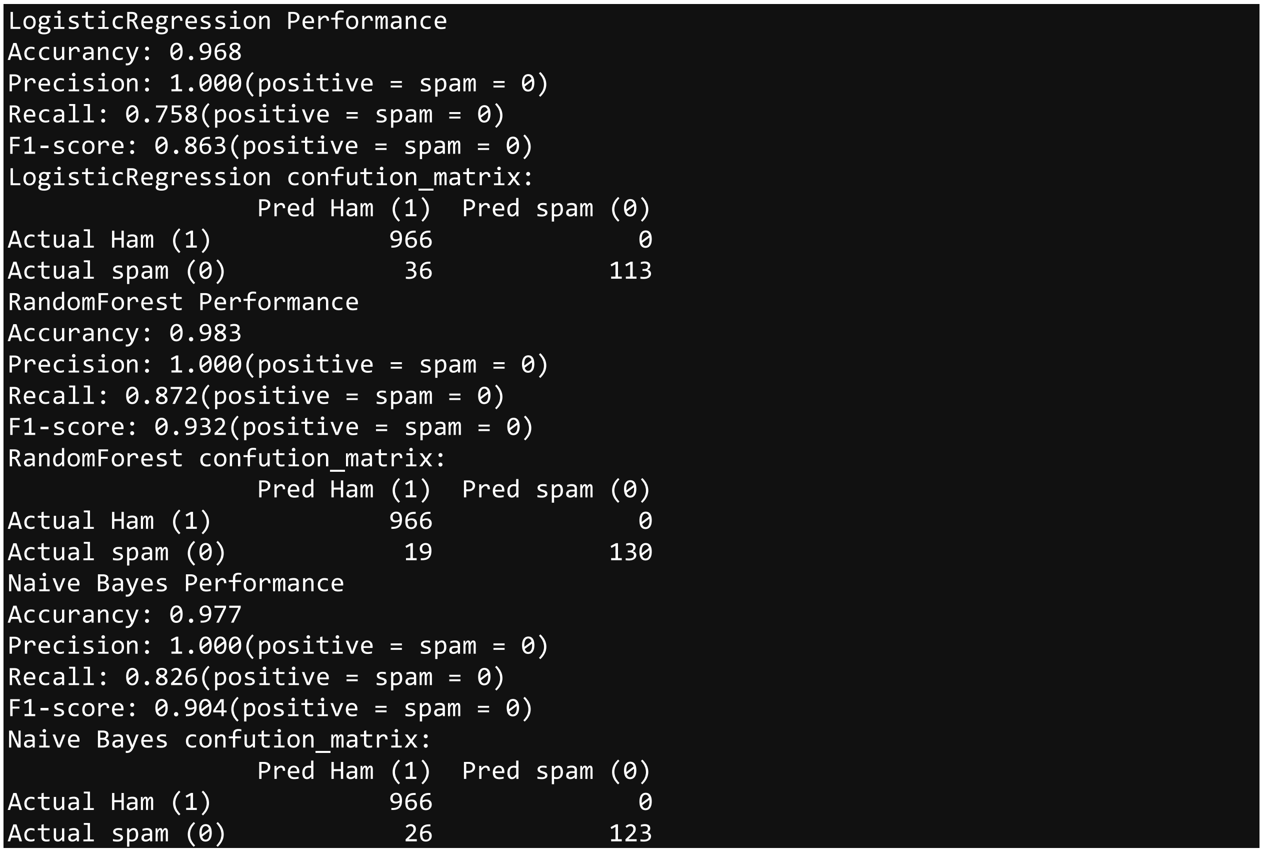
```
simple massege sample prediction
text snippet Ku guuleyso $10000 credit card. Claim now! guryo free ah hel gaari free
ah
Lr Predict:  Ham (1)
Rf Predict:  Ham (1)
NB Predict:  Ham (1)
text snippet Congratulations ku shubo 100 dollarclick dheh oo hel bishan 200 dollar f
ree ah Claim now!
Lr Predict:  Spam (0)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
text snippet Congratulations, you won a free ticket
Lr Predict:  Ham (1)
Rf Predict:  Ham (1)
NB Predict:  Ham (1)
text snippet You have been selected for a $1000 gift card. Claim now!
Lr Predict:  Spam (0)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
text snippet FIRST to 87131 for a poly or text GET to 87131 for a true tone! Help? 08
45 2814032 16 after 1st free, tones are 3x£150pw to e£nd txt stop
Lr Predict:  Spam (0)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
text snippet Dear got train and seat mine lower seatFree entry in 2 a weekly competit
ion!
Lr Predict:  Ham (1)
Rf Predict:  Ham (1)
NB Predict:  Ham (1)
text snippet I will meet you at the cafe tomorrow
```



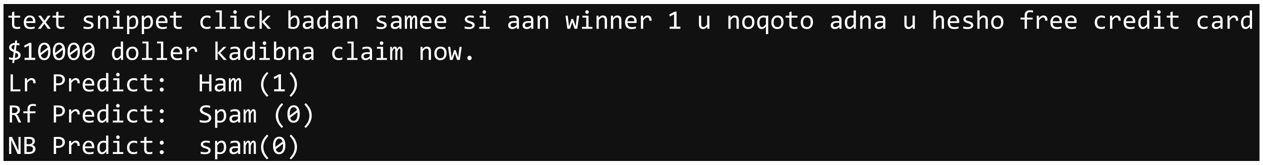
```
Lr Predict:  Ham (1)
Rf Predict:  Ham (1)
NB Predict:  Ham (1)
text snippet Your 2003 Account Statement for shows 800 un-redeemed S.I.M. points. Cal
l 08718738001 Identifier Code: 49557 Expires 26/11/04
Lr Predict:  Spam (0)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
text snippet click badan samee si aan winner 1 u noqoto adna u hesho free credit card
$10000 doller kadibna claim now.
Lr Predict:  Ham (1)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
```

2: Did all models agree? If not, which one gave more realistic predictions?

Yes



```
LogisticRegression Performance
Accurancy: 0.968
Precision: 1.000(positive = spam = 0)
Recall: 0.758(positive = spam = 0)
F1-score: 0.863(positive = spam = 0)
LogisticRegression confution_matrix:
                 Pred Ham (1)  Pred spam (0)
Actual Ham (1)            966              0
Actual spam (0)            36            113
RandomForest Performance
Accurancy: 0.983
Precision: 1.000(positive = spam = 0)
Recall: 0.872(positive = spam = 0)
F1-score: 0.932(positive = spam = 0)
RandomForest confution_matrix:
                 Pred Ham (1)  Pred spam (0)
Actual Ham (1)            966              0
Actual spam (0)            19            130
Naive Bayes Performance
Accurancy: 0.977
Precision: 1.000(positive = spam = 0)
Recall: 0.826(positive = spam = 0)
F1-score: 0.904(positive = spam = 0)
Naive Bayes confution_matrix:
                 Pred Ham (1)  Pred spam (0)
Actual Ham (1)            966              0
Actual spam (0)            26            123
```



```
text snippet click badan samee si aan winner 1 u noqoto adna u hesho free credit card
$10000 doller kadibna claim now.
Lr Predict:  Ham (1)
Rf Predict:  Spam (0)
NB Predict:  spam(0)
```

Hadda waxa saxsan rf iyo nb ga

3: Understanding NaiveBayes:

What is Naive Bayes?

Naive Bayes is a simple and efficient machine learning classification algorithm based on Bayes' Theorem, used to predict the category of a data point using probability, It assumes that all features in the data are independent of each other given the class, simplification, Naive Bayes performs well such as spam filtering, document categorization, and sentiment analysis

Why is it often used in spam detection?

Naive Bayes is often used in spam detection because it efficiently handles high-dimensional text data and works well with categorical input variables like word frequencies, It classifies emails as spam or not spam by calculating the probability that an email belongs to each category based on its content, assuming words appear independently

What are its advantages and limitations?

Advantages of Naive Bayes

1. Simplicity and efficiency: It has few parameters, which makes it fast to train and predict
2. Works well with high-dimensional data: Especially useful in text classification where many features (words) are present
3. Requires less training data: Compared to other complex models
4. Good performance in many domains: Despite the simplistic assumption of feature independence
5. Easily updated: probabilities update with new data, useful for continuously evolving datasets like emails

Limitations of Naive Bayes

1. Strong independence assumption: It assumes features are conditionally independent given the class, which is rarely true in reality and may reduce accuracy
2. Poor performance: When features are strongly dependent, it may not capture these relationships well.
3. Zero probability problem: If a feature never appears in training for a class, it assigns zero probability
4. May not perform well on complex data: For example, in image recognition, where dependencies between features are critical

4: Metrics Discussion:

Which model had better = Accuracy, Precision, Recall, F1-Score, and Confusion Matrix?

Naive Bayes

What does the Confusion Matrix tell you about false positives and false negatives?

100% and 82%

Naive Bayes Performance

Accurancy: 0.977

Precision: 1.000(positive = spam = 0)

Recall: 0.826(positive = spam = 0)

F1-score: 0.904(positive = spam = 0)

Your Findings:

Summarize in 1–2 paragraphs which model you prefer for price prediction models and why?

For spam detection, a model with balanced precision and recall high F1-score is often recommended, as it minimizes both false positives legitimate emails marked as spam and false negatives spam emails that bypass the filter

Ultimately, a recommendation would lean towards the model with the best F1-score and a confusion matrix reflecting acceptable false positive and false negative rates based on the use case prioritie spam filters typically prioritize minimizing false negatives to block spam, but not at the cost of excessive false positives

By: Mohamed Mohamud