QUESTION 1A

I am using the holdout method to split the dataset and evaluate the model

Holdout is an easy method to split the data and isn't as computationally expensive as K-Fold CV

QUESTION 2

BINARY

Number of Mistakes = **25**

Predictive Error = 25 / 1800 = **1.39%**

Optimum Epoch = **3**

Most Positive 15 = compani, dollar, freebsd, market, ultim, will, your, pleas, deathtospamdeathtospamdeathtospam, minut, guarante, hour, click, sight, remov

Most Negative 15 = wrote, run, music, said, seem, date, latest, of, standard, but, i, it, network, too, copyright

COUNT

Number of Mistakes = **56**

Predictive Error = 56 / 1800 = **3.11%**

Optimum Epoch = **472**

Most Positive 15 = numberm, free, numberb, our, am, market, en, cnumber, remov, monei, compani, z, will, pa, dollarnumb

Most Negative 15 = ii, new, iii, but, spam, file, cnet, set, tech, pro, wrote, subscript, ie, some, can

TFIDF

Number of Mistakes = **51**

Predictive Error = 56 / 1800 = **2.81%**

Optimum Epoch = **202**

Most Positive 15 = bnumber, numberb, insur, compani, dollar, below, china, retail, our, year, call, dollarnumb, remov, numberc, numbera

Most Negative 15 = re, date, but, strip, write, subscript, shop, friend, network, wrote, url, juli, daili, she, canon

QUESTION 3

BINARY

Linear Regression — 23 mistakes

Gaussian Naive Bayes — 277 mistakes

Bernoulli Naive Bayes — 89 mistakes

Multinomial Naive Bayes — 31 mistakes

Complement Naive Bayes — 30 mistakes

COUNT

Linear Regression — 46 mistakes

Gaussian Naive Bayes — 121 mistakes

Bernoulli Naive Bayes — 89 mistakes

Multinomial Naive Bayes — 410 mistakes

Complement Naive Bayes — 407 mistakes

TDIDF

Linear Regression — 27 mistakes

Gaussian Naive Bayes — 246 mistakes

Bernoulli Naive Bayes — 98 mistakes

Multinomial Naive Bayes — 38 mistakes

Complement Naive Bayes — 35 mistakes