```
#A PDF file (maximum allowed size is 2MB) providing the findings and justifications for the following topics:
#Write a few lines about training dataset quality and any errors found in the training dataset
#Explain the data preprocessing steps.
#Explain and justify the model chosen by you for your prediction.
#Create .mat files from existing edited csv files. mat files are easier to use with octave
#Octave installation is required for things to work.
#======== Write Here============


####################################################
##Input Dataset
####################################################

#Checksum for data files
#Check Dataset for quality and errors found in training dataset. These steps can be performed using Excel.
#Check for blank cells and text fields
#Identify columns with blank cells and consider removing the entire feature if unnecesary
#Remove unrequired Columns
#Change some text attributes into numbers

#======== Write Here============
#Do this for both 'test_dataset.csv' and 'training_dataset.csv':
#Remove attributes user_id, mail_id, hacker_timezone
#Change names to numbers from mail_category_1 to 1 and so on using replace
#Change names to numbers from mail_type_1 to 1 and so on using replace
#Change all FALSE values to 0 and all TRUE values to 1 using replace

#Found some blank cells in mail_category  mail_type  hacker_timezone
#Set Blank mail_category to 19
#Remove mail_type since it has only one value.
#Remove hacker_timezone as it is unlikely to relate to our ouput.


#Do this for 'training_dataset.csv':
#Removed attributes click_time, clicked, open_time, unsubscribe_time

#Save edited 'training_dataset.csv' as 'accurate_input_data.csv'
486049 rows
#Save edited 'test_dataset.csv' as 'accurate_test_data.csv'
207425 rows

#==============================
```

#==============================

##

# Model Selection

##

High Level Steps in wrapper.m

Tried Using one vs all linear classification algorithm

Plotted Learning Curve with various input size. Taking Multiple data sets from a randomized data set and then averaging cost.

Didn't find any substantial improvement for using multi class classification vs single class.

So moved to linear classification with just one class 'opened'.

Plotted Learning Curve with varying input size. Taking Multiple data sets from a randomized data set and then averaging cost.

Found High Bias.

Hence tried polynomial features with normalisation.

Plotted Learning Curve with varying polynomial degrees. (Tried with Lamda 1,3,10). Taking Multiple data sets from a randomized data set and then averaging cost.

Found out polynomials are converging at d=1.

Mapped Lambda X Error in Cross Validation Curve

Found at 3rd degree polynomial, lambda=160 gives the least cross validation error for average case.

#Now train and test final algorithms and check their efficiency.

#Divide input set into 60% training set, 20% validation set and 20% test set

#Train using training set and check accuracy using test set

#then compute prediction for test set assignment

##

# Training Model

##

#Train the entire training data set on the chosen model

#In our case polynomial, normalised, regularised, Single Class Linear Classification