TAHA SALMAN
260721174

COMP 551 Assignment 1

Part 1 – Sampling:

Q1)

```
random_float = random() //generate a random number between 0 and 1

if (random_float < 0.2)
      return "Movies";
else if (random_float < 0.6)
      return "COMP-551";
else if (random_float < 0.7)
      return "Playing";
else
      return "Studying"
```

Q2)

When sampling the routine for 100 days we get the following results:

Activities	Fraction of Days Spent
Comp 551	41/100
Movies	17/100
Playing	9/100
Studying	33/100

When sampling the routine for 1000 days we get the following results:

Activities	Fraction of Days Spent
Comp 551	417/1000
Movies	205/1000
Playing	79/1000
Studying	299/1000

The fractions are generally getting closer and closer to the multinomial distribution as the number of days (samples) increases

TAHA SALMAN
260721174

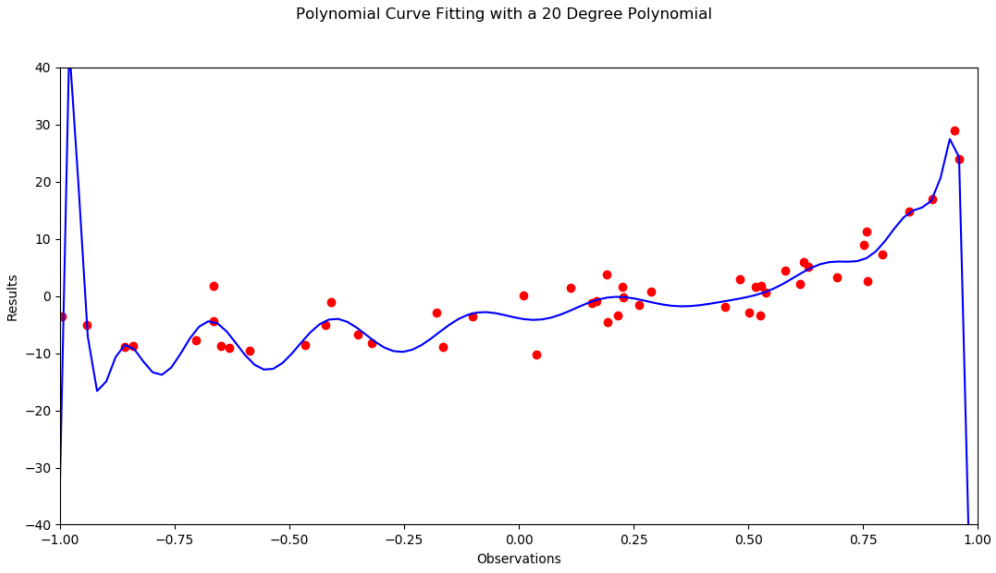
COMP 551 Assignment 1

Part 2 – Model Selection

Q1)

a) The Mean Squared Error (M.S.E.) for the training data is 6.475
The Mean Squared Error for the validation data is 1418.514

b) (Please run the code in the jupyter notebook to be able to zoom in and out of the graph)
Best Fit 20 Degree Polynomial (Blue line) against the Actual Data (Red Dots):

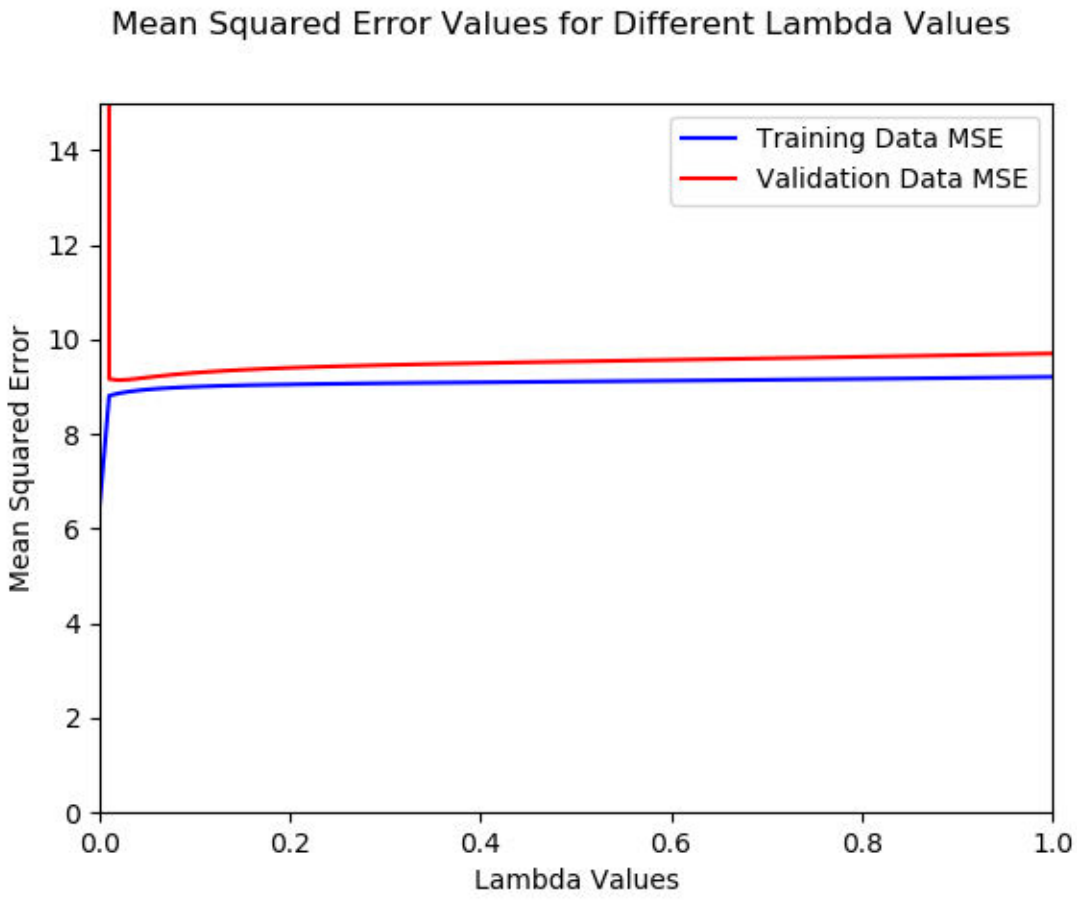


c) The model is overfitting because there is a huge mean squared error for the validation data compared to the training data as can been seen in part (a).

Q2)

a) (Please run the code in the jupyter notebook to be able to zoom in and out of the graph)

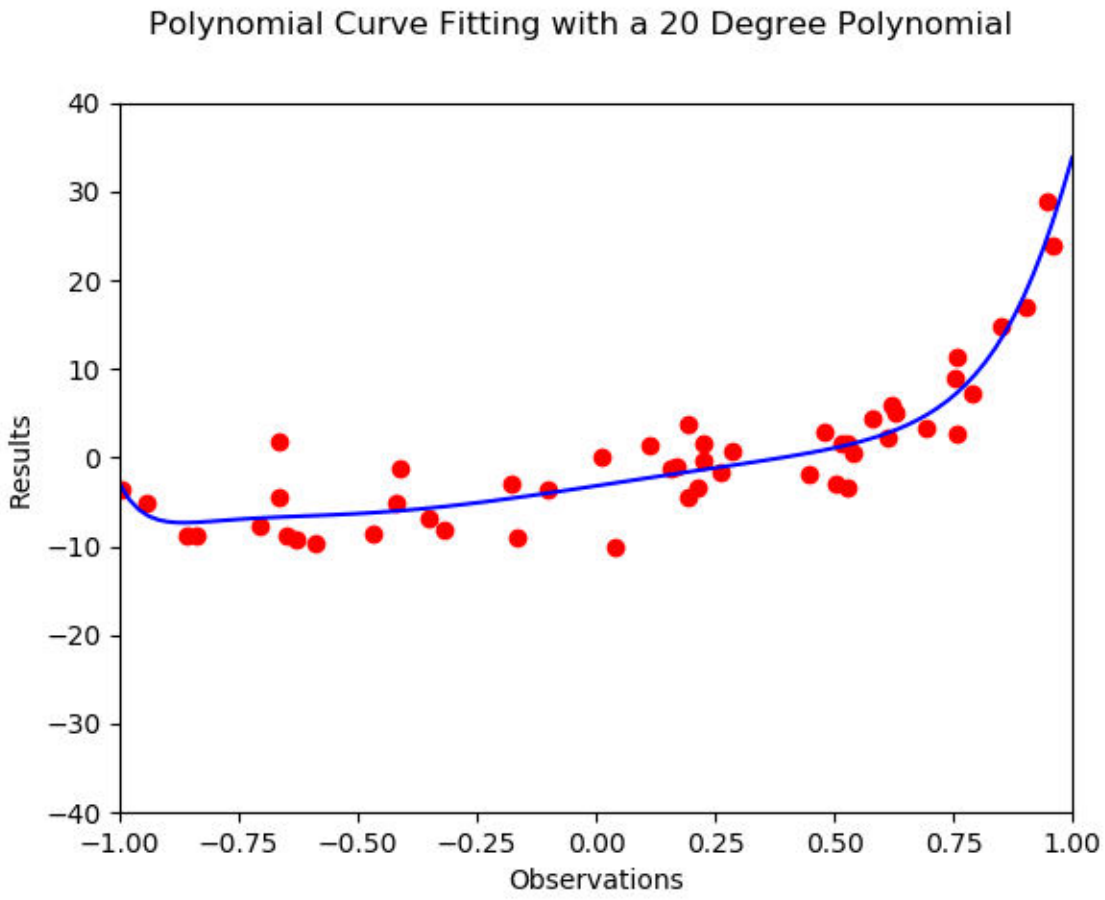
COMP 551 Assignment 1



b) The best value for λ is 0.02 as it gives the smallest M.S.E. of 9.135098784694396 (for validation data) for the lambdas tested (100 values between 0 and 1). Using this value of λ for the polynomial generated with the test data gives us a M.S.E. of 10.730218400927388.

c) (Please run the code in the jupyter notebook to be able to zoom in and out of the graph)
Best Fit 20 Degree Polynomial (Blue line) with Lambda 0.02 against the Actual Data (Red Dots):

COMP 551 Assignment 1



d) The model is neither overfitting nor underfitting. It seems to get the basic trend of the data captured in the polynomial plotted. It might maybe need more data to find the best fit. There are outliers to the line but overall it fits relatively well with the data.

Q3) I think the degree of the source polynomial is much less than 20 as only a small value of lambda with regularization led to a significantly smaller M.S.E. Judging from the shape of the polynomial plotted with regularization, I would say that the degree of source polynomial is probably 8 or 9.

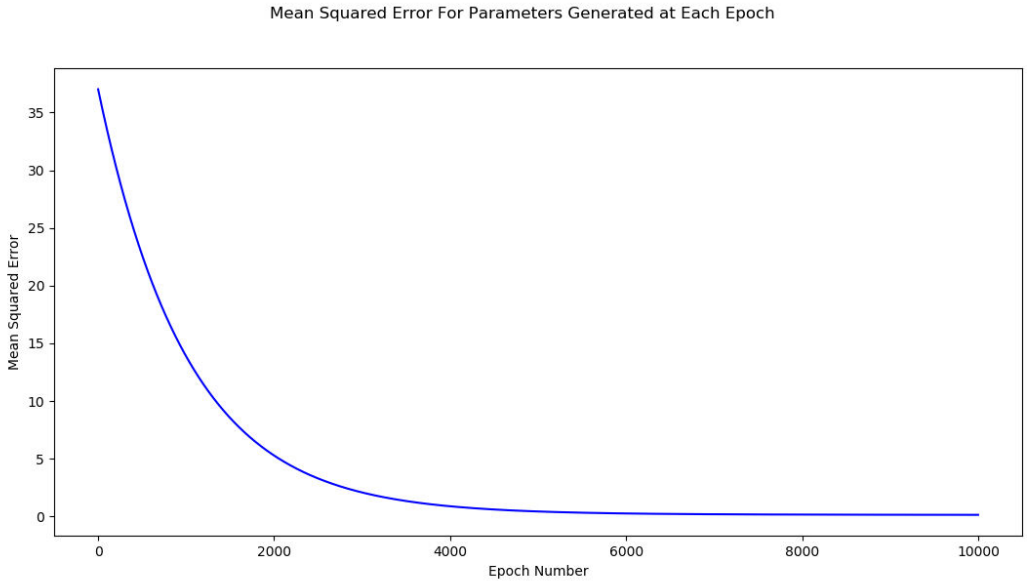
TAHA SALMAN
260721174

COMP 551 Assignment 1

Part 3 – Gradient Descent for Regression

Q1)

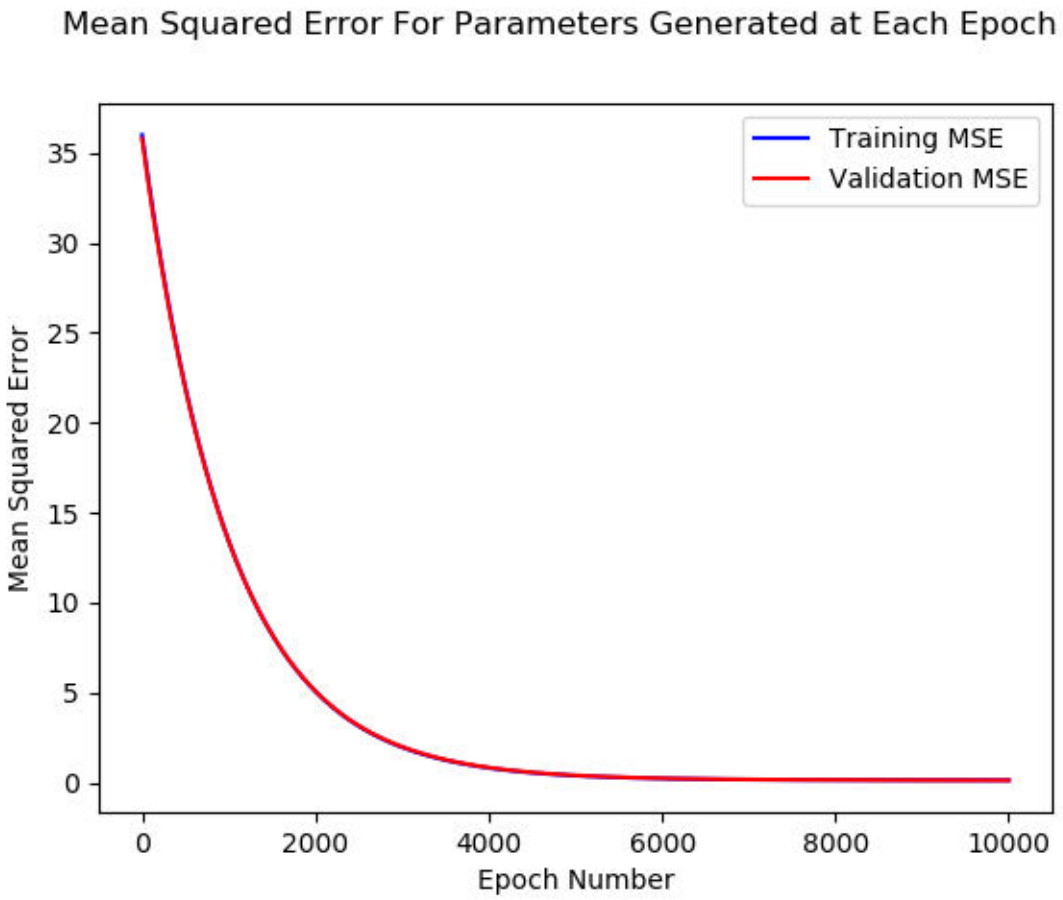
a) I calculated the MSE for 10000 epochs. It would be very tedious to report them all so I plotted them as can be seen on the jupyter notebook and below:



As can be seen in the first few epochs the Mean Squared Error is very high but after somewhere around the 5,000 epoch it starts converging to a value close to 0. The MSE for the last epoch (10,000) is 0.18564891951890658.

b) Below you will find the training and validation MSE plotted for every epoch. They are very similar and as a result overlap one another. Please try zooming into the curve on the jupyter notebook.

COMP 551 Assignment 1



Q2)

a) In the following table, you will find the MSE on Validation Data for Different Step Sizes. Please check the jupyter notebook code to explore other values.

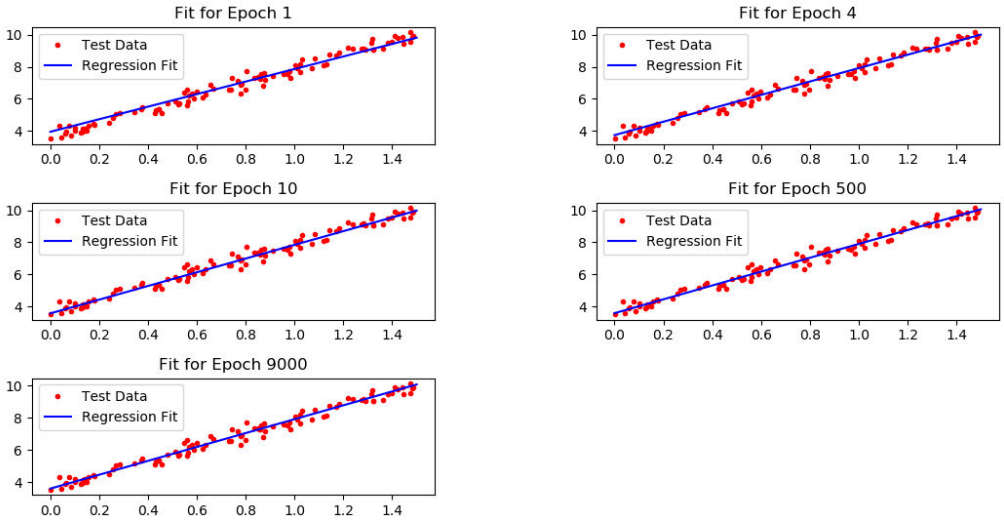
Step Size	**M.S.E.**
1.0	0.8773823108185421
0.1	0.07716799864100732
0.01	0.07378495100292111
1e-3	0.0741367223513358
1e-4	0.07407035186247066
1e-5	*0.0740703447845628*
1e-6	0.07407037780518864
1e-7	0.07407037785222964
1e-8	0.07407037786859148
1e-9	0.07407037786975364

The best step size seems to be 0.01 as it has the lowest M.S.E. and is relatively fast too.

COMP 551 Assignment 1

b) For step size 0.01, after 10,000 epochs, the M.S.E. with Test Data is 0.069575115103576

Q3) (Please check jupyter notebook for more details and to change the visualization examples)



Part 4 – Real Life Dataset

Q1)

a) Please check jupyter notebook and attached files for the data and code. Replacing missing values with the sample mean may or may not be a good solution. It depends on the type of data we are dealing with. Here, the first 5 features are non-predictive so there is no point in filling them in. Furthermore, if some variable is very skewed or is bimodal, the mean will be a bad substitute for missing data.

b) We could use the median or mode for each column rather than the mean. We could even replace the missing data with 0.

c) After analyzing the data, it becomes clear that the first 5 columns contain non-predictor variables, so there would be no point to try and fill their missing data. First we remove them and then fill all the missing values with median. This would have better performance compared to the mean, since for some variables the data could be very skewed or bimodal. As a result, the mean would return a value that may not be possible. The median at least returns a value that is valid for other examples. As a result, the

COMP 551 Assignment 1

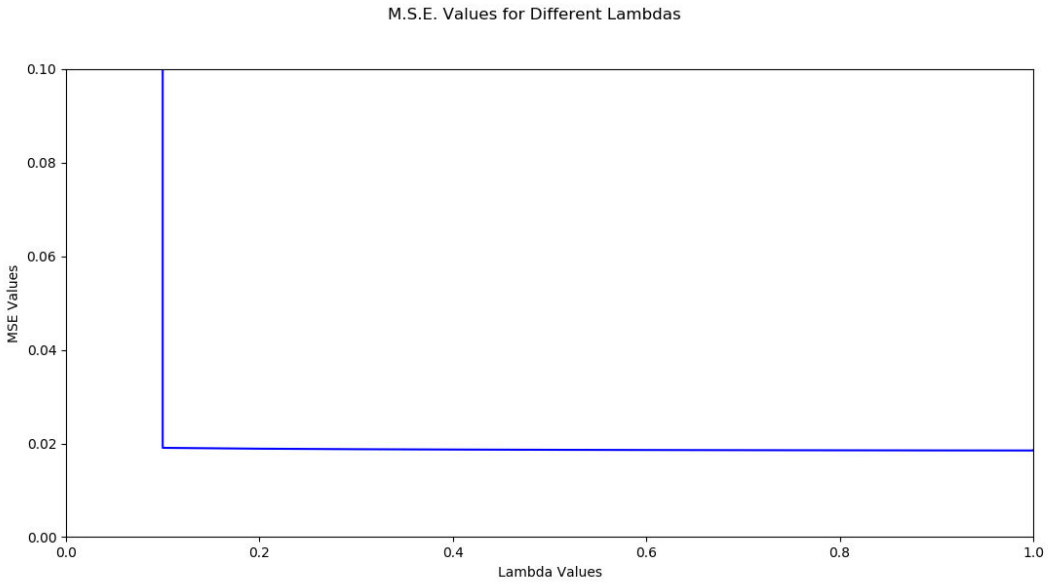
M.S.E. is smaller with median. Another good method (which I implemented first before reading part 3) is removing all features that have values missing for instances greater than a specified threshold (such as 20 percent of all total instances of data collected). Now for the features left, we simply remove all instances where we have missing data from the data set. As a result, the data will be significantly less biased than with the mean.

d) Please check the file 'Datasets/CrimeData/crime_data_updated_custom.csv' submitted with the code.

Q2) The average MSE over these 5 Datasets is 0.019040236846756355. Please check jupyter notebook for more details

Q3)

a) The smallest MSE 0.02002176625519709 was found for λ = 1

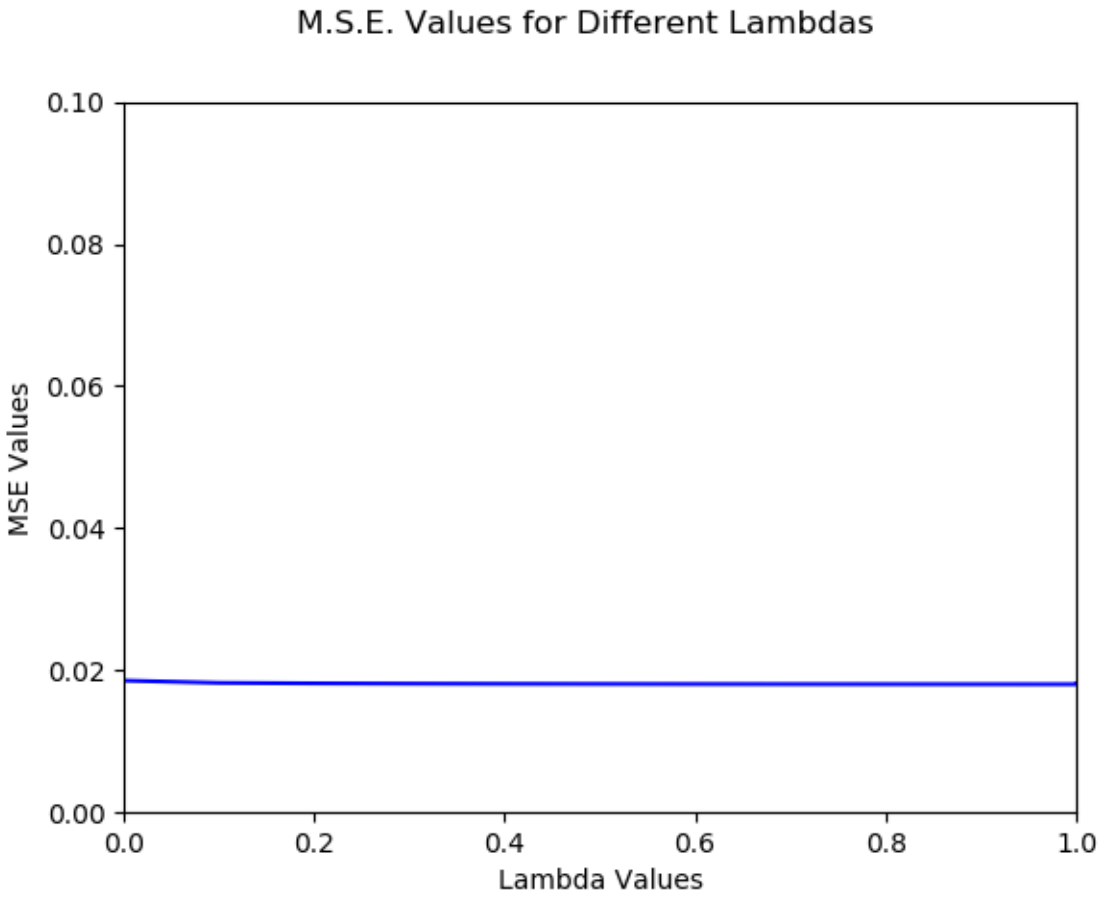


b) It can be seen that only a small proportion of the overall number features contain missing data. The rest of the features have barely any missing data. As a result, maybe we should take out all the missing data by removing features which have missing data for instances over a certain defined threshold. For example, if the threshold is defined to be 0.5, all features that have data missing for 0.5*(Total Instances of Data) or more should be removed. We will then be left with an almost complete dataset with very few

COMP 551 Assignment 1

data instances with missing values. For such instances, we can simply remove the entire data instance.

c) With a reduced set of features, the smallest M.S.E. was also for λ =1 at 0.01745486643383115.



d) From the above results, it seems that the performance of the model with a reduced set of features is slightly better than the model where the missing data is filled with medians. Further tests would be needed to find out the exact degree of improvement. However, the little improvement we see can perhaps be attributed to the reduction in bias of the dataset. When adding the median for missing values (or mean) we keep introducing bias to the dataset leading to an error due to this bias. When removing the features and also removing data instances where there is any missing data, we get a more complete dataset with less bias even though it may be smaller.