Hasan Asif

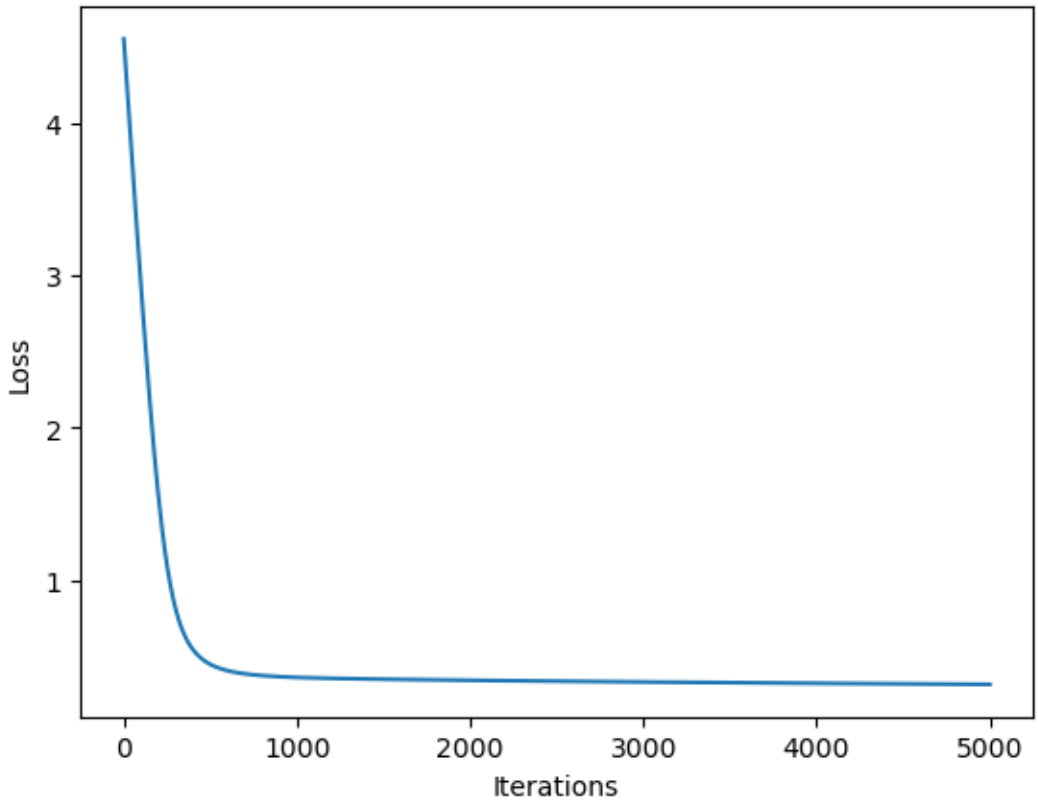
Hussain Vakharwala

CSE-574 Assignment 1 Report

PART 1:

1. **Provide your best accuracy.**
   Our best accuracy is 87%

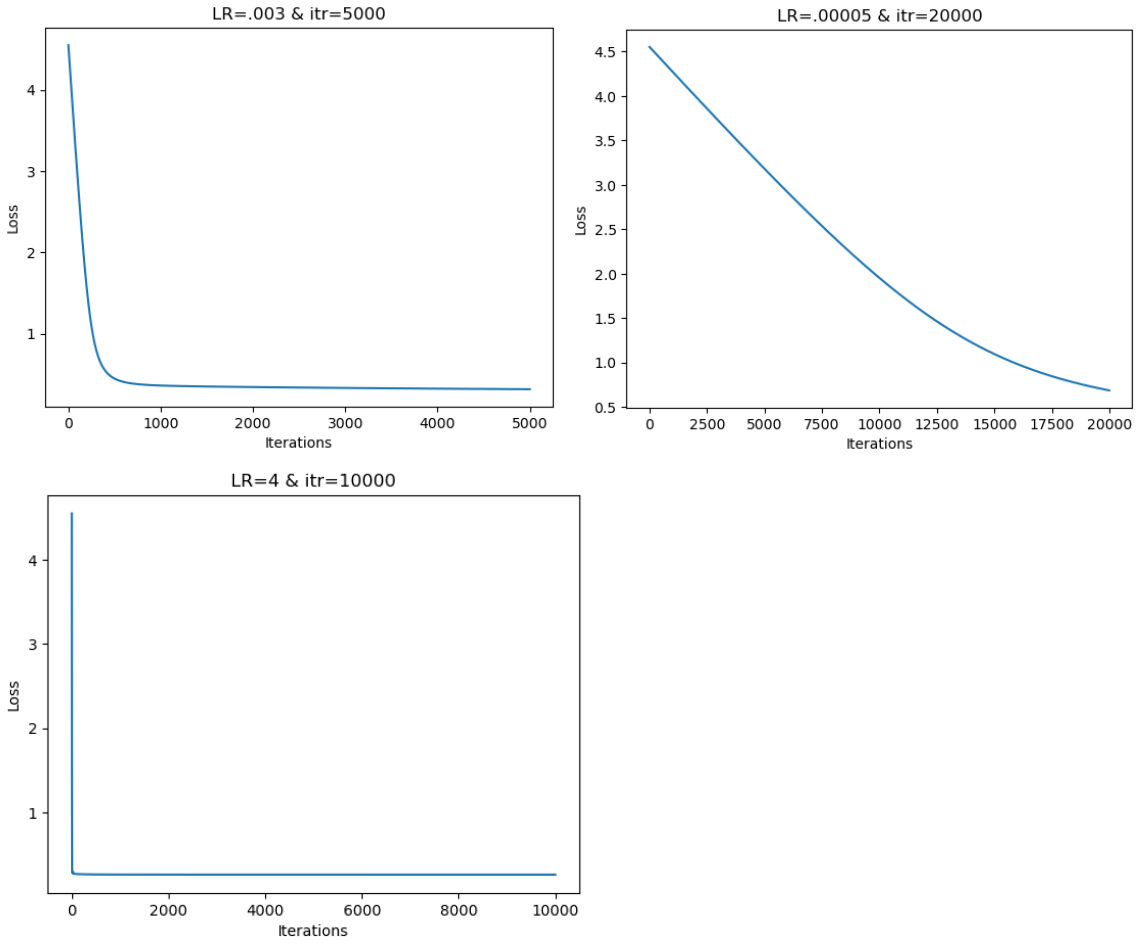
2. **Include loss graph and provide a short analysis of the results**.



From the above Loss graph it is shown that model converge quite model. For the first 500 iteration loss reduced exponentially and beyond that reduction is loss is quite low. So we can say model reached to the global minima.

3. **Explain how hyperparameters influence the accuracy of the model. Provide at least 3 different setups with learning rate and #iterations and discuss the results along with plotting of graphs. Ex: For 3 different learning rates, you can plot the graph to discuss impact and loss over the iterations.**

For this part we have 2 hyper-parameters Learning rate and iterations.



As learning rate is too low(.00005) model learns quite slow and even after running 20000 iteration it did not converge, similarly if learning rate is too large(4) loss reduces at exponential rate. For .003 loss and 5000 iteration model converge perfectly.

4. Discuss the benefits/drawbacks of using a Logistic Regression model.

Drawback:

- Logistic regression has a tendency to over-fit due to less number of records
- Prone to higher error due to multi collinearity
- Good to use only on linearly separated data
- As in the Penguin dataset due to many categorical features logistic regression struggle to find relation to the input feature.

Benefits:

- Logistic reg is simple and easy to understand and implement
- Computationally efficient on large datasets
- Less sensitive to the outlier as compared to other linear model

PART 2:

1. **Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?**

   We used the flight dataset, in which **"price"** is the Target variable and other are the independent variable. It has 300153 rows and 12 columns. Dataset is about the price of particular flight based on 11 feature(airline, flight, source_city, departure_time, stops, arrival_time, destination_city, class, duration, days_left) Price is the continuous column and many feature are categorical in nature such as Flight, arrival time, departure time, arrival city and many more.

2. **Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

   For the numerical features following are the main statistic

	duration	days_left	price
mean	12.22102	26.00475	20889.6605
std	7.191997	13.561	22697.7674
min	0.83	1	1105
25%	6.83	15	4783
50%	11.25	26	7425
75%	16.17	38	42521
max	49.83	49	123071

   Dataset contains no null values.

3. **Provide at least 5 visualization graphs with a brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.**

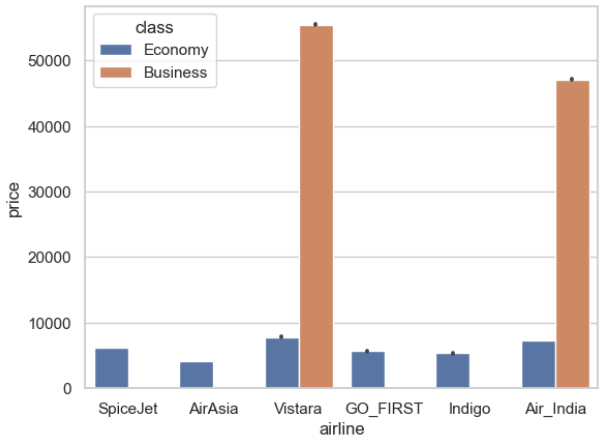


Figure 1: It shows the relation between the Price and Airline based on the Classes. For Vistara and AirIndia business class is available and which is almost 5 times expensive the economy class.

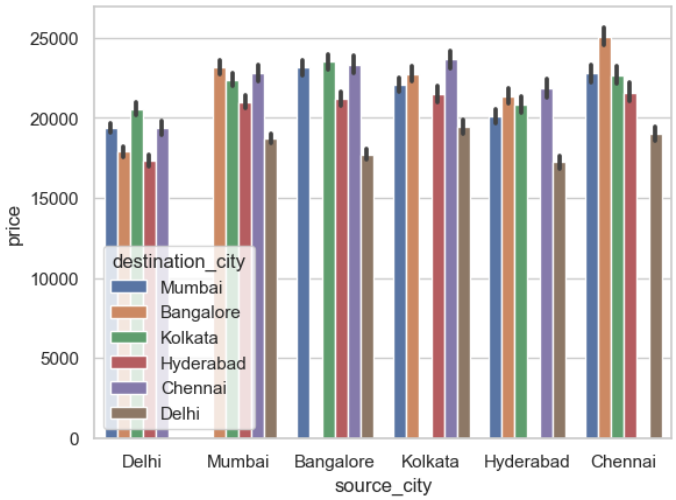


Figure 2: It shows price of the flight from source to destination city, highest is the price from Chennai to banglore and the cheapest flight are from Delhi.

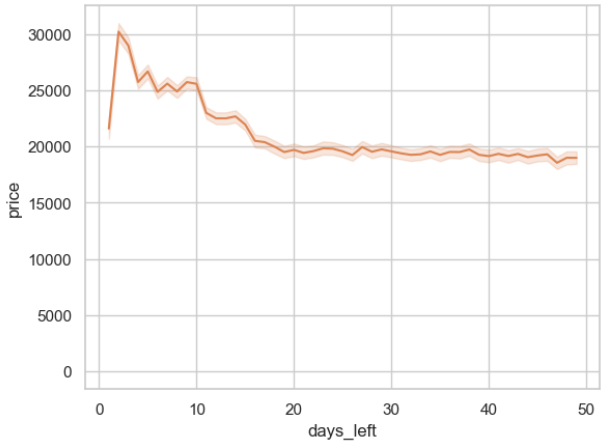


Figure 3: It represents as the number of days increases the price reduces but after approx. 20 days there is no substantial reduction on the price, it almost remains constant.

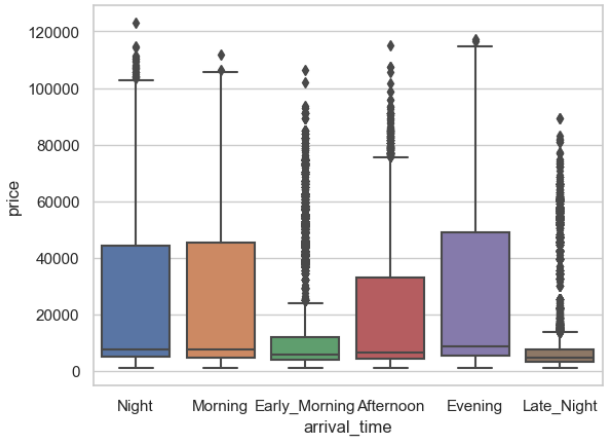


Figure 4: it depicts the box plot between the price and the arrival time. It shows high outlier present on the early morning flight price, price is lowest for late night price.

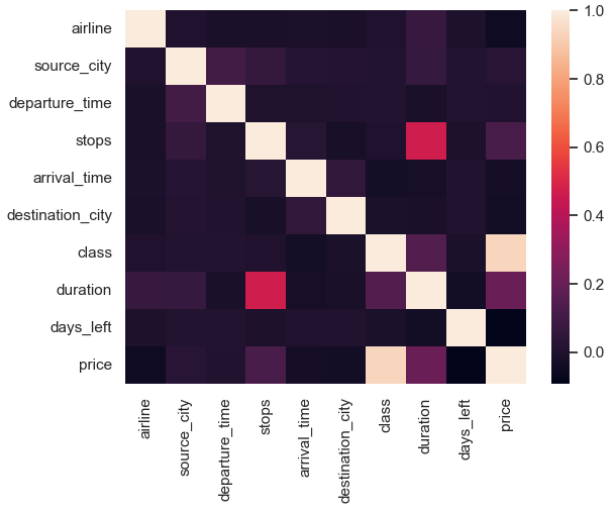
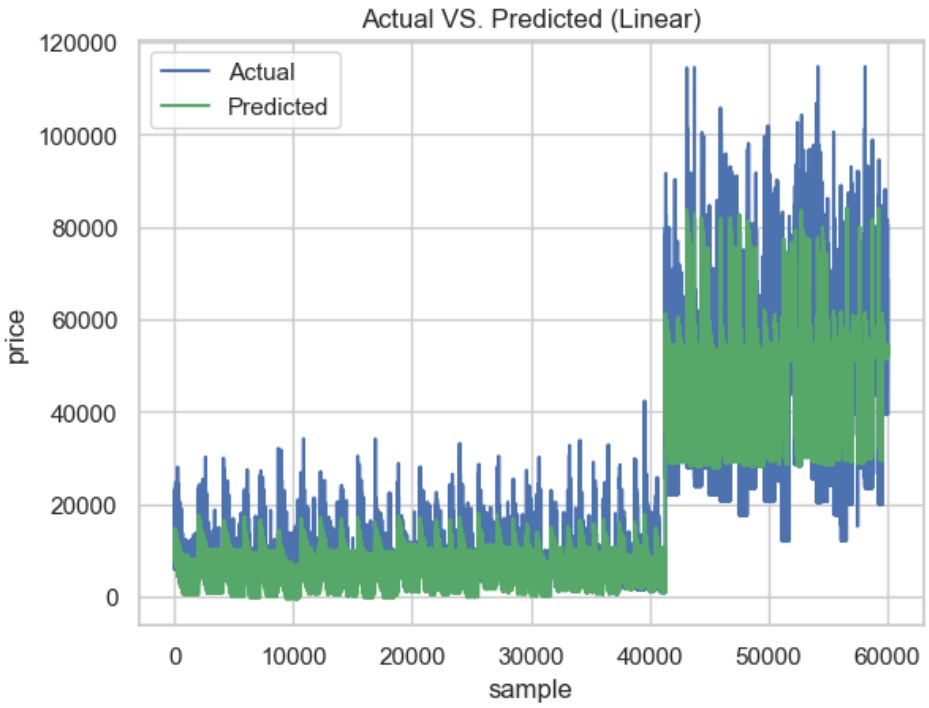


Figure 5: it depicts the correlation between the price and the feature. Depicts high relation between stops and duration. Price is highly related to the duration, class, stops.

4. Provide your loss value

Our loss value for the Linear Regression is .0012

5. Show the plot comparing the predictions vs the actual test data



6. Discuss the benefits/drawbacks of using OLS estimate for computing the weights

Benefits:

- OLS is easy to implement and computationally very efficient as compared to the other methods for weight finding.
- handle nonlinear relationships between the target and input variables.

Drawbacks:

- It performs very poor on non-normalized data
- Sensitive to outliers
- If the relationship is non-linear, OLS may not provide an accurate representation of the data.

7. Discuss the benefits/drawbacks of using a Linear Regression model.

Benefits:

- Linear Regression is easy to implement and computationally very efficient as compared to the other linear model
- L1 and L2 regularizations available for reducing over fitting

Drawbacks:

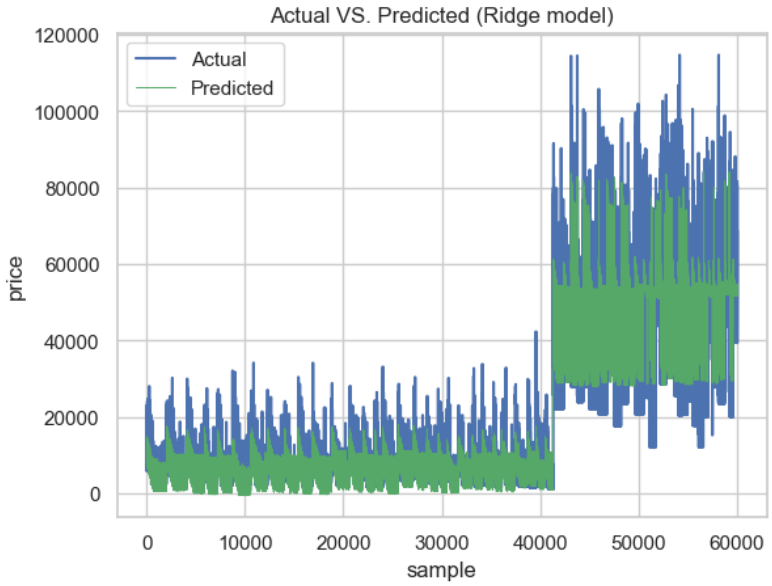
- Tends to over-fit a lot
- Assumption of linear relationship and non-collinearity between the features
- Limited to linear relationship for best results
- Prone to outliers

PART 3:

1. **Provide your loss value.**

   Best loss value is .0019 for lambda equal 1e-6

2. **Show the plot comparing the predictions vs the actual test data.**



3. **Discuss the difference between Linear and Ridge regressions. What is the main motivation for using l2 regularization?**

   Both Ridge and Linear regression use for Regression task. Linear regression fit the data assuming there is linear relationship and fits straight line on the training data as accurate to minimize the loss, where as Ridge regression is just its regularized form while fitting the line many times linear regression overfits so to overcome this problem we use Ridge regression.
   The main motivation for using Ridge regression, on flight price dataset is to avoid overfitting. Linear Regression fits the training data too closely, leading to poor performance on the test data. In our case for linear regression train and test MSE are almost equals whereas using ridge the train is higher(**.09)**and test MSE is quite lower**(.0012).**

4. Discuss the benefits/drawbacks of using a Ridge Regression model

Benefits:

- Reduces the overfitting by achieving low test mse mean model perform better on the unseen data
- In training it reduces the multicollinearity present in the dataset

Drawbacks:

- Not effective for no-linear dataset
- Need to find the optimal lambda value

Contribution:

Team Member	Assignment Part	Contribution(%)
Hasan Asif	**Part 1,2,3,Bonus**	**50%**
Hussain Vakharwala	**Part 1,2,3,Bonus**	**50%**