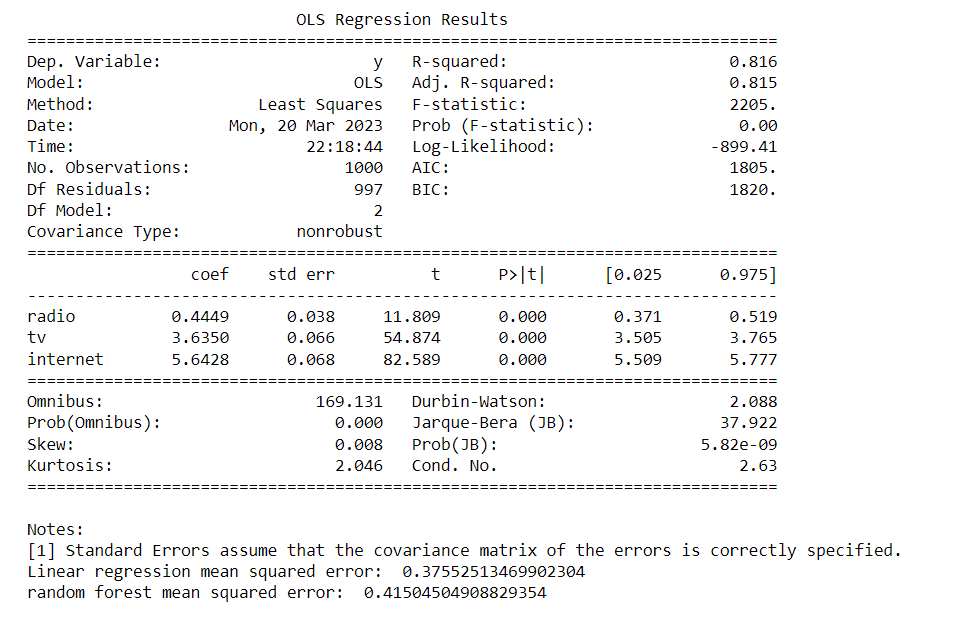
**Assignment 1**

**Name:** Danita Anubhuti Prakash

**StudentID:** s4745511

1.

After fitting the linear regression, the folowing is the result:



From the Coefficients of the Linear regression model we can understand that all the coefficients obatined are positive and statistically significant. Radio- 0.449 has the least contribution and Internet- 3.635 has the highest contribution. Where as TV- 5.6428 is in the middle of both of them. Here, it can be noted that Internet most effective for advertisement.

After making predictions on test set by using Linear model and RF, the mean squared error obtained is:



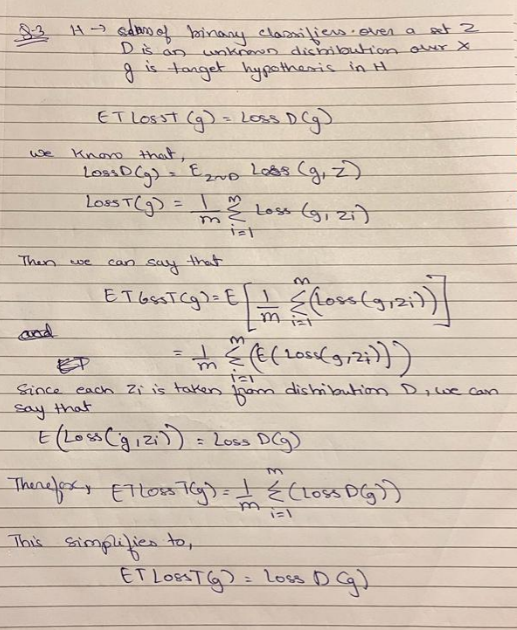
2. For the stated problem the hypothesis is H = {f(x, θ); θ ∈ Θ}. Such that we observe X1, X2, …Xn ~ F.

Here we can state that θ is the vector of parameters that take values in the parameter space Θ. Here, θ consists of standard deviation and mean (which takes a normal distribution). Where Θ (parameter space) is set of all possible values that mean and standard deviation can take.

Such that it can be written as:



3.





4.

(a) After fitting the model, the following are the results

|  |  |  |
| --- | --- | --- |
| Model | B1 | B2 |
| Model1 | 2.2 | 0 |
| Model1 | 0 | 0.7333 |

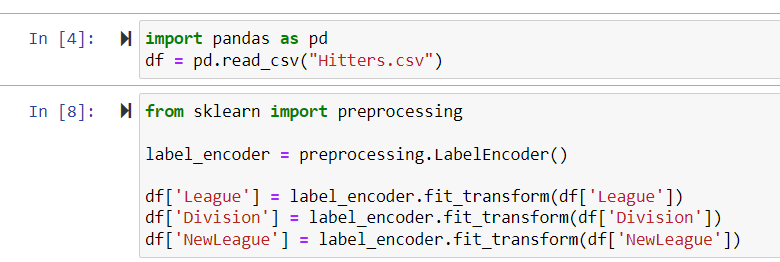
(b) The errors for the equation is:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Squared Error Loss | Absolute Error Loss | L1.5 Loss |
| Model1 | 0.559 | 0.64 | 0.584 |
| Model2 | 2.173 | 1.0 | 1.409 |

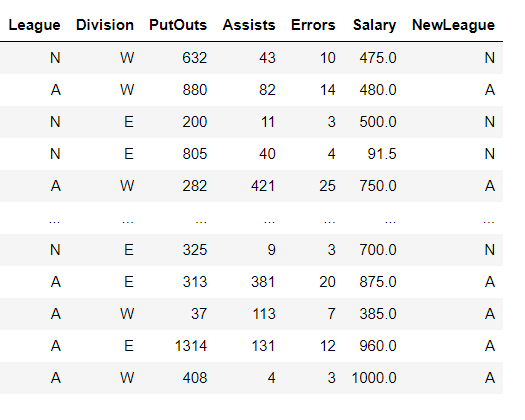
(c) According to the results obatined from the two models, it can be observed that Model1 assumes that y is a function of an error term. Here, Model2 has assumes that y is linear function of x. Specifically, the squared error loss of Model 1 is 0.559 compared to 2.173 for Model 2. The absolute error loss of Model 1 is 0.64 compared to 1.0 for Model 2, and the L1.5 loss of Model 1 is 0.584 compared to 1.409 for Model 2. This suggests that Model 1 performs better than Model 2 in terms of these evaluation metrics. Therefore, Model 1 is a more suitable fit for the input data compared to Model 2 based on these evaluation metrics.

5.

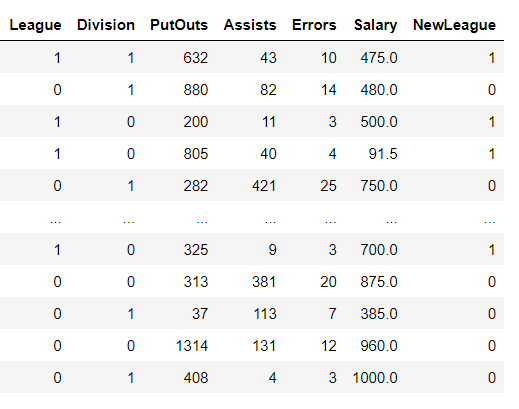
(a) Using jupyter, I loaded the dataset and changed the categorical values using Label Encoding



Example of the columns League, Division and NewLeague before and after Label Encoding for the rows: (0-4 and 258-262). Before doing Label Encoding:

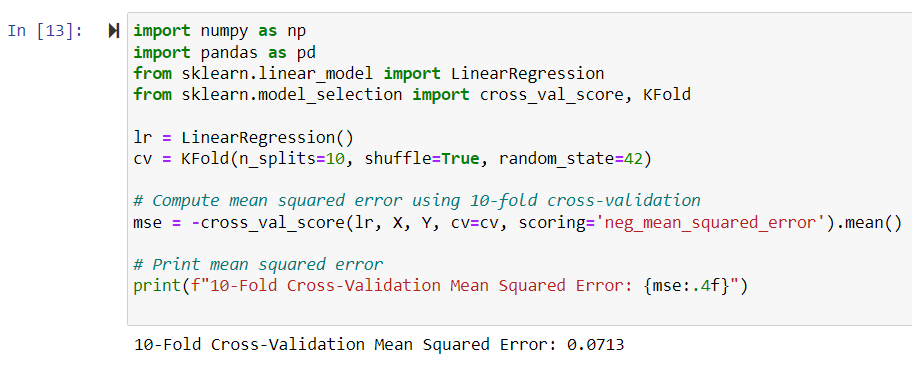


After doing Label Encoding

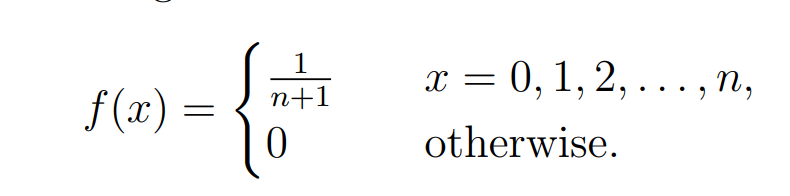


(b) Label encoding assigns numerical values while one hot encoding creates new binary variables for each category to convert categorical values. Label encoding assigns a number to each value, where as one hot encoding creates separate columns for each categorical value. I choose Label encoding over One Hot encoding as it maps each data in a category with a number label which belongs to that category inturn replacing the column with numerical values. This proves to be more efficient for our dataset.

(c) Algorithm for Fitting the Linear regression and reporting 10-fold cross validation. Here the 10 fold cross- validation mean squared error obtrained is 0.0713



6. The given descrete pmf(x) is :



Ans.

To generate a random number from the above function, the following steps need to followed.

1. Generate a random number between 0 and 1. Let the number be denoted as r.
2. Compute the cumulative distribution function F(x), that is

F(x) = Σ f(i), where i<=x

1. Find the smallest value of x for which F(x)>r. This x value is the random variable.

The time complexity of this method in terms of n is O(n) since we need to compute the CDF by summing up to ‘n’ terms which takes O(n) time. Hence, to find the smallest x can be done in O(n) time depending on how the CDF is computed.

Generation of Random Variable

One way is from the given pmf is to use the inverse transform sampling method.

For this, we need to calculate the given cumulative distribution function (CDF) as F(x) = x+1/n+1 and then generate a random variable U on [0,1]. The random variable X can be obtained by solving for X in the equation F(X)=U.

The time complexity of this method is O(1) because calculating the CDF and solving for X can be done in constant time, independent of the value of n.

7. For the given question, by writing the algorithm, The confidence for interval by using Crude Monte Carlo Algorithm is :

Confidenc interval= (0.2390524280383081,0.24094264219864833)

The difference between l and given f(x) is

Difference = 0.000303448054010141

Explanasion:

This code uses the Monte Carlo method to estimate the value of an integral of a given

function over the interval [0,1]. It generates random values of x and evaluates the function at those points, storing the results in an array. It then computes the mean and standard deviation of the array and uses these values to compute a confidence interval for the estimated value of the integral. The code also calculates the true value of the integral using integration from the python library sympy. Finally, it compares the estimated value obtained through

Monte Carlo simulation with the true value obtained through symbolic integration, computing the absolute difference between the two values.

The algorithm written is:

