Short Essay Questions Part 2

Question No 1:

Part (a):

As we have collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We have to check the effect of profit, number of employees, industry on the CEO salary. Salary is a continuous variable and also quantitative in nature. So we can select a regression method in this scenario of the problem. The reason is that regression problems always works on the variables which are continuous in nature. CEO salary can be predicted by using the regression method to solve the problem.

Part (b):

Profit, Number of Employees and Industry are the variables which will affect the CEO salary. When Profit of a company increases then CEO is also rewarded in the form of increase in the salary. Number of Employees can also has a big effect on the salary of CEO; when number of employees increases then the payroll budget is divided into the number of employees which will decrease the salary cap of CEO. Industry can also has a big factor on the salary of a CEO, e.g. Automotive industry CEO will have greater salary than Agriculture industry CEO salary due to revenue difference of both industries.

Part (c):

We can use Multiple Linear Regression method which have general equation like this

$$Yi = \beta0 + \beta1xi1 + \beta2xi2 + ... + \beta pxip + \epsilon$$

where, for *i*=*n* observations:

Yi = dependent variable
xi = explanatory variables
ϐ0 = yintercept (constant term)
Bp = slope coefficients for each explanatory variable
ϵ = the model's error term (also known as the residuals)

Similarly we can write the selected equation as

CEO Salary = β0 + β1 * profit + β2 * number of employees + β3 * industry + ϵ

In this way we can calculate the CEO salary by using the above selected equation.

Question No 2:

Logistic Regression:

Logistic Regression is one of the most commonly used Supervised Machine Learning algorithms that is used to model a binary variable that takes only 2 values – 0 and 1. Logistic Regression is basically a classification algorithm. The objective of Logistic Regression algorithm is to develop a mathematical equation that can give us a score in the range of 0 to 1. This score gives us the probability of the variable taking the value 1.

Example:

A Logistic Regression model classifier can be used to identify whether a tumour is malignant or if it is benign. Malignant are cancerous cells while benign are noncancerous cells. Several medical imaging techniques are used to extract various features of tumour. For instance, the size of the tumour, the affected body area, radius, texture, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension etc. These features are then fed to a Logistic Regression classifier to identify if the tumour is malignant or if it is benign.

Reason to select Logistic Regression over Linear Regression:

Linear Regression is a supervised machine learning algorithm used for regression purposes on continuous quantitative variables. Linear Regression provide predicted values which are close to the target values. Linear Regression cannot help us to do classification task. We have to define threshold for Logistic Regression. By default the threshold value is 0.5 in Logistic Regression algorithm. It means that when the model gives prediction probability for class 0 greater than class 1, then our model gives output of class 1. When the model gives prediction probability for class 1 greater than class 0, then our model gives output of class 1. This is the reason that we use Logistic Regression over Linear Regression as it helps to classify the target.

Question No 3:

Forward Selection approach:

Forward selection approach is a variable selection method which begins with a model that contains no variables then starts adding the most significant variables one after the other until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model. We have to determine the most significant variable to add at each step and choose a stopping rule. The most significant variable can be chosen so that, when added to the model It has the smallest p-value, or It provides the highest increase in R2, or It provides the highest drop in model Residuals Sum of Squares compared to other predictors under consideration. The stopping rule is satisfied when all remaining variables to consider have a p-value larger than some specified threshold, if added to the model. When we reach this state, forward selection will terminate and return a model that only contains variables with p-values < threshold. The threshold can be A fixed value for instance 0.01, 0.3 or 0.6. Determined by Akaike Information Criterion or Determined by Bayesian information criterion.

Backward Selection approach:

Backward selection is a variable selection method which begins with a model that contains all variables under consideration then starts removing the least significant variables one after the other until a specified stopping rule is reached or until no variable is left in the model. We have to determine the least significant variable to remove at each step and Choose a stopping rule. The least significant variable is a variable that which has the highest p-value in the model or Its elimination from the model causes the lowest drop in R2 or Its elimination from the model causes the lowest increase in Residuals Sum of Squares compared to other predictors. The stopping rule is satisfied when all remaining variables in the model have a p-value smaller than some specified threshold. When we reach this state, backward elimination will terminate and return the current step's model. The threshold can be A fixed value for instance 0.01, 0.3 or 0.6. Determined by Akaike Information Criterion or Determined by Bayesian information criterion.

Question No 4:

Residual Standard Error:

The residual standard error is used to measure how well a regression model fits a dataset. it measures the standard deviation of the residuals in a regression model.

$$\textbf{Residual standard error} = \surd\Sigma(y - \hat{y})2/df$$

where:

y: The observed value

Ŷ: The predicted value

df: The degrees of freedom, calculated as the total number of observations – total number of model parameters.

The smaller the residual standard error, the better a regression model fits a dataset. Conversely, the higher the residual standard error, the worse a regression model fits a dataset. A regression model that has a small residual standard error will have data points that are closely packed around the fitted regression line

R-Squared:

Rsquared (R2) measures the proportion of the variance in the response variable that can be explained by the predictor variable in a regression model.

$$R2 = [(n\Sigma xy - (\Sigma x)(\Sigma y)) / (\surd n\Sigma x2-(\Sigma x)2 * \surd n\Sigma y2-(\Sigma y)2)]2$$

The actual calculation of R-squared requires several steps. This includes taking the data points (observations) of dependent and independent variables and finding the line of best fit, often from a regression model. From there you would calculate predicted values, subtract actual values and square the results. This yields a list of errors squared, which is then summed and equals the unexplained variance. To calculate the total variance, you would subtract the average actual value from each of the actual values, square the results and sum them. From there, divide the first sum of errors (explained variance) by the second sum (total variance), subtract the result from one, and you have the R-squared.

Preference:

Residual Standard Error is more preferable than R- Squared. The reason is that we can see the standard difference between the actual values and the predicted values by Residual Standard Error. R-Squared provide us the confidence about the dependent variable with respect to the independent variable but it cannot explain how the data points are near the model predicted line. By using Residual Standard Error we can judge the performance of our model easily.

Practical Exercise

Question No 4:

We have a multi-class classification problem with 8 classes. So binary logistic classifiers we will need to solve the problem using the proposed one-vs-one approach as follow:

$$Formula = (NumClasses * (NumClasses - 1)) / 2$$

$= (8 * (8 - 1)) / 2$
$= (8 * 7) / 2$
$= 56 / 2$
$= 28$

So we will need 28 binary logistic classifiers