**Solution**

The multiple regression and KNN regression are performed on the dataset to calculate the future purchase amount prediction of the person with the help of the independent variables age, gender, family size, membership, and discount card type.

Part A:

The multiple regression can be performed by using age, gender, family size, membership, and discount card type as explanatory variable (independent variable) and purchase amount as the response variable (dependent variable).

Summary Output

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.144459045 |
| R Square | 0.020868416 |
| Adjusted R Square | 0.010958177 |
| Standard Error | 293.5619902 |
| Observations | 500 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 5 | 907350.3632 | 181470.0726 | 2.105743004 | 0.063452775 |
| Residual | 494 | 42572249.18 | 86178.64206 |  |  |
| Total | 499 | 43479599.54 |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 1221.953906 | 52.36800324 | 23.33397935 | 9.88667E-82 | 1119.062419 | 1324.845393 | 1119.062419 | 1324.845393 |
| Age | 0.70502848 | 0.767803803 | 0.91824041 | 0.358941118 | -0.803535342 | 2.213592301 | -0.803535342 | 2.213592301 |
| Gender (Male is 1) | -9.855460093 | 26.32192921 | -0.374420128 | 0.708252437 | -61.57220069 | 41.86128051 | -61.57220069 | 41.86128051 |
| Family size | -5.697533096 | 7.649853356 | -0.744789845 | 0.456752942 | -20.72779469 | 9.332728495 | -20.72779469 | 9.332728495 |
| Membership (with membership is 1) | 3.003039574 | 26.34013627 | 0.114010024 | 0.909276167 | -48.74947385 | 54.755553 | -48.74947385 | 54.755553 |
| Discount card type (0 means no card and there are three discount cards | -35.65203522 | 12.10075291 | -2.946265864 | 0.003368271 | -59.42732518 | -11.87674527 | -59.42732518 | -11.87674527 |

Table 6

**R square:** From the output table 6, R square is 0.020, that means 0.2% of the variation in the purchase amounts can be explained by the independent variables (age, gender, family size, membership, and discount card type). Since the R squared value is less than 75%, indicating that the regression technique is not statistically significant with the explanatory variables of this dataset. However, R square is not enough to indicate if the model is accurate for the given dataset.

**Adjusted R:** It is another measure for accuracy of the model like R square which is between 0 and 1. Closer to 1 means better accuracy. The value for this model is 0.01 which is not good to present accurate prediction.

**Standard Error:** This is the average distance that the observed values fall from the regression line. In the output table 4, the observed values fall an average of 293.56 units from the regression line. It is an inflated standard error for regression co-efficient.

**Significance F:** 0.063 is the significance F for the solved model. As the value is more than 0.05, we can conclude that there is a no significance relationship between the purchase amounts and the explanatory variables (age, gender, family size, membership, and discount card type).

Model implementation and equation to predict the risk of diabetes using the independent variable of age, weight, and gender.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Regression equation y= m+b1x1+b2x2+b3x3+b4x4+b5x5 y= 1221.95+ 0.705\*age+ (-9.855)\*gender + (-5.697)\*family size + 3\* membership+ (-35.652)\* discount card type | | | | | | |
|  |
|  |
| Model Implementation | | | | | | |  |
| x1 | x2 | x3 | x4 | x5 | y |  |  |
| 20 | 1 | 4 | 0 | 3 | 1096.452778 |  |  |
| 50 | 1 | 4 | 1 | 0 | 1227.562777 |  |  |

Table 7

Using the estimated regression equation, we can calculate the purchase amount of the person with the help of the independent variables age, gender, family size, membership, and discount card type.

This model is not appropriate for precise prediction of the purchase based on the explanatory variables as the model indicators like R square, standard error and significance F is not up to the benchmark values as explained above.

Residual output of the prediction:

Note: The observation below is only for the first 10 rows of the dataset.

|  |  |  |
| --- | --- | --- |
| *Observation* | *Predicted Purchase amount ($)* | *Residuals* |
| 1 | 1154.708091 | -317.7080907 |
| 2 | 1135.69827 | 397.3017297 |
| 3 | 1181.205766 | 395.7942344 |
| 4 | 1112.965969 | -57.96596874 |
| 5 | 1162.592953 | 116.4070468 |
| 6 | 1226.545274 | 435.4547261 |
| 7 | 1223.496465 | -465.4964645 |
| 8 | 1152.722554 | -137.7225544 |
| 9 | 1203.191198 | 73.80880238 |
| 10 | 1199.787299 | -314.7872992 |

Part B:

Performing a KNN regression model will help in simplifying model and regression equation. We can divide the data into training and test data set to help the model to perform better and unbiased prediction. KNN regression is very simple for implementing and new dataset can be added which will not impact the accuracy of the model later. This technique is a non-parametric, which means that it does not make assumptions on the underlying data distribution. This model is recommended when there is no prior knowledge of the distribution of the data. This will help in making the prediction of the purchase from the e-commerce website based on the explanatory variables of the dataset.

**Solution 6:**

Part A:

The logistic regression equation for the given output is

Part B

Step 1: Convert the actual and predicted dataset in binary variables using if condition of excel

Step 2: Create a Pivot table of the actual and predicted value. Add the actual value in the row and predicted value in the column and count of actual column in the value section.

This will give us a confusion matrix.

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | Predicted |  |  |
|  | 0 | 1 | Grand Total |
| 0 | TN=97 | FP=199 | 296 |
| 1 | FN=463 | TP=560 | 1023 |
| Grand Total | 560 | 759 | 1319 |

Table 8

Step 3: Accuracy is true actual and positive/ total values

= TP+TN/ (TP+TN+FP+FN)

= (97+560)/ (97+560+463+560) = 0.498

= 49.8% is the accuracy of the model

Step 4: Error = 1-Accuracy

= 0.502

= 50.2%

Step 5: Sensitivity is the True positive/ Actual true

=TP/ (TP+FN)

= 560/ (560+463) = 0.547

= 54.7% sensitivity

Step 6: Specificity is True negative/ Actual False

= TN/ (TN+FP)

= 97/ (97+199) = 0.327

= 32.7% Specificity