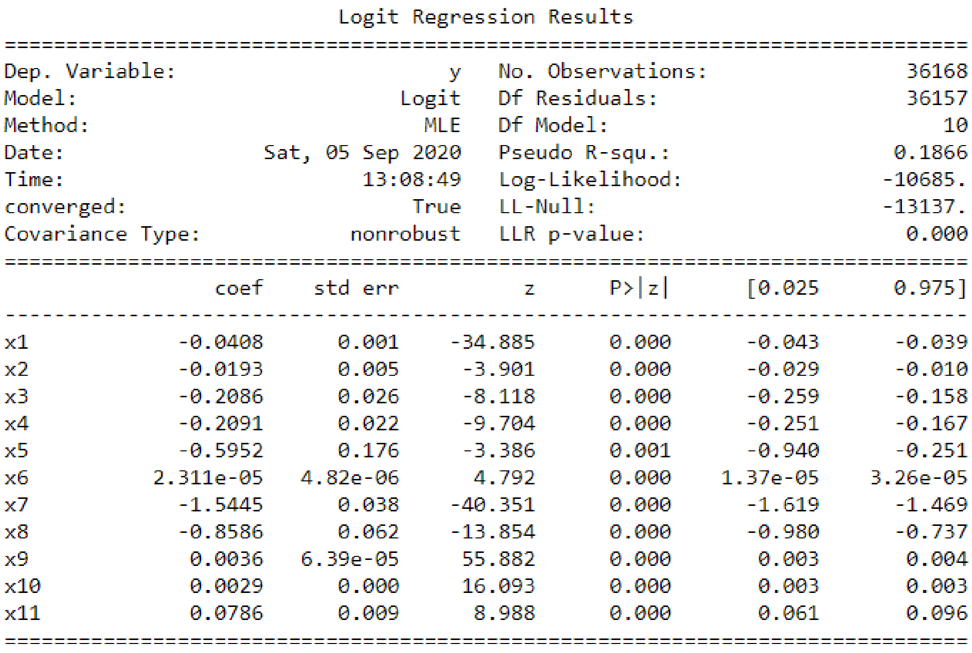
Fraud & Risk Analytics Assignment

1. Significance of the output

Logistic Regression



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                36168
Model:                          Logit   Df Residuals:                    36157
Method:                           MLE   Df Model:                           10
Date:                Sat, 05 Sep 2020   Pseudo R-squ.:                  0.1866
Time:                        13:08:49   Log-Likelihood:                -10685.
converged:                       True   LL-Null:                       -13137.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0408      0.001    -34.885      0.000      -0.043      -0.039
x2            -0.0193      0.005     -3.901      0.000      -0.029      -0.010
x3            -0.2086      0.026     -8.118      0.000      -0.259      -0.158
x4            -0.2091      0.022     -9.704      0.000      -0.251      -0.167
x5            -0.5952      0.176     -3.386      0.001      -0.940      -0.251
x6          2.311e-05   4.82e-06      4.792      0.000    1.37e-05    3.26e-05
x7            -1.5445      0.038    -40.351      0.000      -1.619      -1.469
x8            -0.8586      0.062    -13.854      0.000      -0.980      -0.737
x9             0.0036   6.39e-05     55.882      0.000       0.003       0.004
x10            0.0029      0.000     16.093      0.000       0.003       0.003
x11            0.0786      0.009      8.988      0.000       0.061       0.096
==============================================================================
```

This Model Summary includes two segments. The first segment provides model fit statistics and the second segment provides model coefficients, their significance and 95% confidence interval values.

I . Model fit Statistic

1. **Dep. Variable** – It is the dependent variable in the table.
2. **No .of Observations** – it is the total no. of observations in table.
3. **Model** – it shows which model is used .
4. **Df Residuals** - the df(Residual) is the sample size minus the number of parameters being estimated, so it becomes df(Residual) = n - (k+1), so here 36168 – 11 = 36157.
5. **Method** – this model is fitted using Maximum likelihood estimation method i.e. the parameter estimates are those values which maximize the likelihood of the data which have been observed.
6. **Date** – the date on which this output is generated.

7. **Pseudo r square** – when we use OLS regression method , we use R Square but when Logistics regression is used we use Pseudo R square and it measures the goodness of the model.

   Formula = 1 –(LL/LL-NULL) which is 18% and very less and hence our model is less reliable.
8. **Time** – when this output is generated.
9. **LL- Null-** is the value when model fitted using only intercept value and its value is-13137 and it can be interpreted using log likelihood where all predictors are considered.
10. **Log - likelihood** –Likelihood Ratio test (often termed as LR test) is a goodness of fit test used to compare between two models; the null model and the final model. The test revealed that when the model fitted with only intercept (null model) then the log-likelihood was -13137, which significantly improved when fitted with all independent variables (Log-Likelihood = -10685). Fit improvement is also significant (p-value <0.05).
11. **Converge –** shows the error present in the model if it is true then no error and if false there is some error. But our model has true convergence means no error present.

II. Interpreting second part -model coefficients, their significance and 95% confidence interval values.

1. The coefficients table shows that all predictor variables are significant as all have p-value less than 0.05.
2. The coefficients interpretation is as follows –

   i. Each one-unit change in x2 will decrease the log odds of having by -0.019.

   ii. Each one-unit change in x3 will decrease the log odds of having by -0.020.

   iii. Each one-unit change in x4 will decrease the log odds of having by -0.020.

   iv. Each one-unit change in x5 will decrease the log odds of having by -0.59.

   v. Each one-unit change in x6 will decrease the log odds of having by -2.311e-05.

   vi. Each one-unit change in x7 will decrease the log odds of having by -1.54.

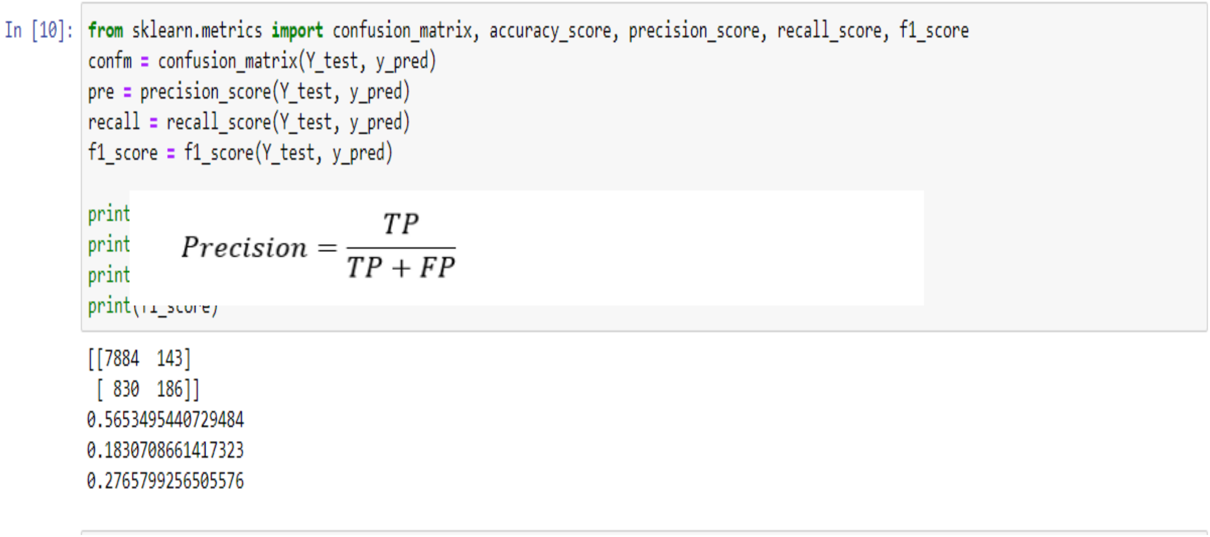
   Each one-unit change in x8 will decrease the log odds of having by -0.85.

   vii. Each one-unit change in x9 will increase the log odds of having by -0.0036.

   viii. Each one-unit change in x10 will decrease the log odds of having by -0.029.

   ix. Each one-unit change in x11 will decrease the log odds of having by -0.078

Confusion Matrix ,Precision score , Recall score and f1 score Interpretation



```
In [10]: from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score
         confm = confusion_matrix(Y_test, y_pred)
         pre = precision_score(Y_test, y_pred)
         recall = recall_score(Y_test, y_pred)
         f1_score = f1_score(Y_test, y_pred)

         print
         print
         print
         print(f1_score)
```

$$Precision = \frac{TP}{TP + FP}$$

```
[[7884  143]
 [ 830  186]]
0.5653495440729484
0.1830708661417323
0.2765799256505576
```

I. **Confusion Matrix** - A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The number of correct and incorrect predictions are summarized with count values and broken down by each class as follows -

- True Positive (TP): Outcome where the model correctly predicts the positive class, it is 186 in our model.
- True Negative (TN): Outcome where the model correctly predicts the negative class , it is 7884 in our model.
- False Positive (FP): Also called a type 1 error, an outcome where the model incorrectly predicts the positive class when it is actually negative , it is 143 in our model.
- False Negative (FN): Also called a type 2 error, an outcome where the model incorrectly predicts the negative class when it is actually positive , it is 830 in our model.

II. **Precision Score** - it answers the question “What proportion of positive identifications was actually correct?”

$$= \frac{186}{186+143}$$

= 0.565 , it means 56% of positives were actually positive from all positives predicted.

III. **Recall Score** - What proportion of actual positives was identified correctly?”

$$= \frac{186}{186+ 830}$$

= 0.183 , it means 18.3% total amount of relevant instances that were retrieved.

$$Recall = \frac{TP}{TP + FN}$$

IV. **F1 Score** – it is a measure of robustness and preciseness of your model.

$$F1\ score = \frac{2 * (precision * recall)}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

$$= \frac{(2*186)}{(2*186) + 143 + 830}$$

= 0.276 , it means 27% of our model is accurate.