**Business Report on Prediction of portion of time that CPU run in user mode from Sun Sparcstation**

## Read the data and performing the basic analysis of the data

|  |  |
| --- | --- |
| Data Field | Data Type |
| lread - Reads (transfers per second ) between system memory and user memory | int64 |
| lwrite - writes (transfers per second) between system memory and user memory | int64 |
| scall - Number of system calls of all types per second | int64 |
| sread - Number of system read calls per second . | int64 |
| swrite - Number of system write calls per second . | int64 |
| fork - Number of system fork calls per second. | float64 |
| exec - Number of system exec calls per second. | float64 |
| rchar - Number of characters transferred per second by system read calls | float64 |
| wchar - Number of characters transfreed per second by system write calls | float64 |
| pgout - Number of page out requests per second | float64 |
| ppgout - Number of pages, paged out per second | float64 |
| pgfree - Number of pages per second placed on the free list | float64 |
| pgscan - Number of pages checked if they can be freed per second | float64 |
| atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second | float64 |
| pgin - Number of page-in requests per second | float64 |
| ppgin - Number of pages paged in per second | float64 |
| pflt - Number of page faults caused by protection errors (copy-on-writes). | float64 |
| vflt - Number of page faults caused by address translation | float64 |
| runqsz - Process run queue size | object |
| freemem - Number of memory pages available to user processes | int64 |
| freeswap - Number of disk blocks available for page swapping | int64 |
| usr - Portion of time (%) that cpus run in user mode | int64 |

**There are few Non null values missing in the ‘rchar’ and ‘wchar’ in the above-mentioned dataset. Since it is of the float datatype.**

**There are no duplicate values present in the dataset.**

**There some outliers present in the dataset.**

**Table, Excel

Description automatically generated*With the Outliers***

**It is mandatory to treat the outlier because it will improve the accuracy of the dataset.**

**Graphical user interface, application, table, Excel

Description automatically generated**

***After treating outlier***

Perform checks for significant variables using appropriate method from stats model. Create multiple models and check the performance of Predictions on Train and Test sets using R square, RMSE & Adj R square. Compare these models and select the best one with appropriate reasoning**.**

**The Significant checks are performed in the model**

**R Square on training data = 0.6259**

**R square on testing data = 0.6171**

**RMSE on training data = 11.112**

**RMSE on testing data = 11.80**

**We are using the Linear Regression using statsmodels(OLS).**

Business Insight from the data

**The final linear Regression equation is**

**usr= (52.68) \* const + (-0.08) \* lread + (0.05) \* lwrite + (0.0) \* scall + (0.0) \* sread + (-0.01) \* swrite + (-0.91) \* fork + (-0.08) \* exec + (-0.0) \* rchar + (-0.0) \* wchar + (-0.65) \* pgout + (-0.04) \* ppgout + (0.15) \* pgfree + (-0.0) \* pgscan + (1.19) \* atch + (0.3) \* pgin + (-0.17) \* ppgin + (-0.05) \* pflt + (0.02) \* vflt + (-7.28) \* runqsz + (-0.0) \* freemem + (0.0) \* freeswap**

**Th above is the Linear regression equation of the sample dataset.**

**The Above are the one of the most and least significant variable in the dataset.**

**There are some positive co-eff such as atch(Number of page attaches) as there increase in 1.3868 time that cpu run in user mode with respect to the 1 number of page attached.**

**There are also some negative co-eff such as pgout, runqsr, fork and so on.**

**Insights:**

**The variable atch(Number of page attaches)(1.3868) has the major impact. The More the number of pages attached goes up the portion of the time that CPU run in user mode.**

**The variable pgin(Number of page-in requests per second)(0.2834) has increasing impact. The more number of the number of pages attached goes up the portion of the time that cpu run in user mode.**

**The variable runqsz(Process run queue size)(-7.29) has the negative impact. There is a decrease in the portion of the time that CPU run in user mode with respect to the Process run queue size.**

**the variable fork(Number of system fork calls per second)(-0.89)has the negative impact. There is a decrease in the portion of the time that CPU run in user mode with respect to the number of system fork calls per second.**

**Business Report on Contraceptive prevalence Survey of Indonesia Ministry of Health**

## 

## Read the data and performing the basic analysis of the data

**The data is having 1473 rows and 10 Columns**

|  |  |
| --- | --- |
| **Data Field** | **Data Type** |
| Wife's age | float64 |
| Wife's education | object |
| Husband's education | object |
| Number of children ever born | float64 |
| Wife's religion | object |
| Wife's now working? | object |
| Husband's occupation | int64 |
| Standard-of-living index | object |
| Media exposure | object |
| Contraceptive method used | object |

**There are few data missing in the ‘Wife age’ and ‘No of children born’. But since these are having float datatype. The mean value of the above columns can be replaced in the missing data place.**

**There are 80 Duplicate rows in the whole dataset. We can drop the duplicates as it will not impact the dataset.**

**There are few outliers in the No of Children born. We are going to ignore those outliers as it will not impact the data if we remove it.**

Apply the Logistic Regression, LDA and CART and the business insight from the various models.

Logistic regression:

The AUC Score for training data: 0.722

Chart, line chart

Description automatically generated

The AUC Score for testing data: 0.722

Chart, line chart

Description automatically generated

Confusion matrix for the training dataset:

Chart, treemap chart

Description automatically generated

Confusion matrix for testing dataset:

Chart, treemap chart

Description automatically generated

**CONCLUSION**

**Note:**

**Precison: tells us how many predictions are actually positive out of all the total positive predicted.**

**Recall: how many observations of positive class are actually predicted as positive.**

**Inferences:**

**For No contraceptive method (Label 0 ):**

**Precision (62%) – 62% of married not use contrceptive method out of all the married women do not use contraceptive method.**

**Recall (45%) – Out of all the married women who do not use contraceptive methods, 45% of womens have been predicted correctly .**

**For contraceptive methods (Label 1 ):**

**Precision (64%) – 64% of women predicted are actually having contraceptive out of all the womens predicted to have contraceptive method.**

**Recall (78%) – Out of all the married actually having contraceptive , 78% of women have been predicted correctly .**

**Overall accuracy of the model – 64 % of total predictions are correct**

**Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification**

Linear Discriminate Analysis:

AUC for the Training Data: 0.691

AUC for the Test data: 0.640

Chart, line chart

Description automatically generated

**Conclusion:**

The model accuracy on the training as well as the test set is about 64%, which is roughly the same proportion as the class 0 and class 1 observations in the dataset. This model is affected by a class imbalance problem. Since we only have 1473 observations, if re-build the same LDA model with more number of data points, an even better model could be built.

Please do go ahead ad try to build other classification algorithms learnt so far in the course and compare the performances of those.

**Inference:**

Inferences using the default value 0.5 for cut-off for test data

For {Customer who did not Churn (Label 0 )}:

Precision (64%) – 64% of the married women not use Contraceptive method, out of all the married women.

Recall (43%) – Out of all the women who will not use contraceptive method, 43% of women are not using the contraceptive method .

For {Customer who did Churn (Label 1 )}:

Precision (64%) – 64% of women who used contraceptive method are correctly predicted ,out of all the married women who used contraceptive method that are predicted .

Recall (81%) – Out of all the married women who used contraceptive methods , 81% of married women who did used contraceptive methods have been predicted

correctly .

Overall accuracy of the model – 85 % of total predictions are correct

Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the model is a good model for classification

CART Method:

AUC Score for training data: 50%

Chart, line chart

Description automatically generated

AUC Score for Testing data: 50%

Chart, line chart

Description automatically generated

**Conclusion**

**Accuracy on the Training Data: 56%**

**Accuracy on the Test Data: 54%**

**AUC on the Training Data: 50%**

**AUC on the Test: 50%**

**Accuracy, AUC, Precision and Recall for test data is almost in line with training data. This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.**

Business conclusion:

**Since the model only has 1473 entries the accuracy is lesser. Maybe if we train and test the model with the higher dataset. It will provide a much better result**.