



VAISHNAV U

PGP-DSBA ONLINE

12/02/2023

PREDICTIVE MODELLING

PROJECT

BUSINESS REPORT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table of Contents | |  | | |
| Content | | | |
| Problem-1 | | | |
|  | | Summary | | 8 |
|  | | Introduction | | 8 |
|  | | Data Description & EDA | | 8 |
| 1.1 | | ****Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.**** | | 10 |
| 1.2 | | ****Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.**** | | 23 |
| 1.3 | | ****Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**** | | 25 |
| 1.4 | | ****Inference: Basis on these predictions, what are the business insights and recommendations.**** | | 35 |
| Problem-2 | | | |
|  | | Summary | | 37 |
|  | | Introduction | | 37 |
|  | | Data Description | | 37 |
| 2.1 | | Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis. | | 40 |
| 2.2 | | Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and build multiple models with different predictors. | | 46 |
| 2.3 | | Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized. | | 54 |
| 2.4 | | Inference: Basis on these predictions, what are the insights and recommendations. | | 60 |

|  |  |
| --- | --- |
| List of tables | |
|  | Table 1.1 | Sample dataset (Problem-1) | 9 |
|  | Table 1.2 | Descriptive statistics of data | 11 |
|  | Table 1.3 | Skewed data table | 18 |
|  | Table 1.4 | Correlation table | 22 |
|  | Table 1.5 | Outlier proportion table | 24 |
|  | Table 1.6 | Encoded data set | 25 |
|  | Table 1.7 | LINEAR REGRESSION USING SCIKIT LEARN - BASE MODEL | 25 |
|  | Table 1.8 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BASE MODEL- OLS TABLE | 26 |
|  | Table 1.9 | LINEAR REGRESSION USING STATS MODEL TEST DATA- BASE MODEL- OLS TABLE | 27 |
|  | Table 1.10 | VIF of the predictors | 27 |
|  | Table 1.11 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BASE MODEL-REMOVING COLUMNS | 28 |
|  | Table 1.12 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BASE MODEL-REMOVING COLUMNS- OLS TABLE | 28 |
|  | Table 1.13 | VIF of the predictors | 29 |
|  | Table 1.14 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BASE MODEL-REMOVING COLUMNS- OLS TABLE | 30 |
|  | Table 1.15 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BEST MODEL- REMOVING COLUMNS- OLS TABLE | 31 |
|  | Table 1.16 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BEST MODEL- OLS TABLE | 32 |
|  | Table 1.17 | LINEAR REGRESSION USING STATS MODEL TEST DATA- BEST MODEL- OLS TABLE | 33 |
|  | Table 1.18 | LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BEST MODEL | 34 |
|  | Table 1.19 | LINEAR REGRESSION USING SCIKIT LEARN- BEST MODEL | 34 |
|  | Table 2.1 | Sample dataset (Problem-2) | 38 |
|  | Table 2.2 | Descriptive statistics of data | 39 |
|  | Table 2.3 | Correlation table | 43 |
|  | Table 2.4 | Outlier percentage table | 44 |
|  | Table 2.5 | Encoded data set | 46 |
|  | Table 2.6 | Train data first five rows | 46 |
|  | Table 2.7 | Test data first five rows: | 47 |
|  | Table 2.8 | Target variable train data first five rows: | 47 |
|  | Table 2.9 | Target variable test data first five rows: | 47 |
|  | Table 2.10 | LOGISTIC REGRESSION USING STATS MODEL- BASE MODEL | 48 |
|  | Table 2.11 | VIF of the predictors | 48 |
|  | Table 2.12 | LOGISTIC REGRESSION USING STATS MODEL- BASE MODEL-REMOVING COLUMNS- LOGIT TABLE | 49 |
|  | Table 2.13 | LOGISTIC REGRESSION USING STATS MODEL- BASE MODEL-REMOVING COLUMNS- LOGIT TABLE | 50 |
|  | Table 2.14 | VIF of the predictors after removing the column | 51 |
|  | Table 2.15 | LOGISTIC REGRESSION USING STATS MODEL- BASE MODEL-REMOVING COLUMNS- LOGIT TABLE | 51 |
|  | Table 2.16 | LOGISTIC REGRESSION USING STATS MODEL- BEST MODEL | 52 |
|  | Table 2.17 | VIF of the predictors after removing the columns | 53 |
|  | Table 2.18 | Train data confusion matrix - Base model | 54 |
|  | Table 2.19 | Train data classification report -- Base model | 54 |
|  | Table 2.20 | Test data confusion matrix - Base model | 55 |
|  | Table 2.21 | Test data classification report - Base model | 56 |
|  | Table 2.22 | Train data confusion matrix - Best model | 57 |
|  | Table 2.23 | Train data classification report -- Best model | 57 |
|  | Table 2.24 | Test data confusion matrix - Best model | 58 |
|  | Table 2.25 | Test data classification report - Best model | 59 |

|  |  |
| --- | --- |
| List of figures | |
|  | Fig 1.1 | Box plot-Univariate analysis | 16 |
|  | Fig 1.2 | Histogram -Univariate analysis | 17 |
|  | Fig 1.3 | Bar plot -Bivariate analysis | 19 |
|  | Fig 1.4 | Pair plot | 20 |
|  | Fig 1.5 | Heat map | 21 |
|  | Fig 2.1 | Box plot-Univariate analysis | 40 |
|  | Fig 2.2 | Histogram -Univariate analysis | 41 |
|  | Fig 2.3 | Bivariate & Multi-variate analysis | 41 |
|  | Fig 2.4 | Pair plot | 42 |
|  | Fig 2.5 | Heat map | 43 |
|  | Fig 2.6 | Density plot | 44 |
|  | Fig 2.7 | Density plot | 45 |

|  |  |  |
| --- | --- | --- |
| ****Comp-activ data****  Summary | | |
| The comp-activ databases is a collection of a computer systems activity measures .  The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs. | | |
| Introduction | | |
| The data consists of 8192 rows and 22 columns. The objective is to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes. | | |
| Data description | | |
| 1 | Iread | Reads (transfers per second ) between system memory and user memory |
| 2 | Iwrite | writes (transfers per second) between system memory and user memory |
| 3 | scall | Number of system calls of all types per second |
| 4 | sread | Number of system read calls per second . |
| 5 | swrite | Number of system write calls per second . |
| 6 | fork | Number of system fork calls per second. |
| 7 | exec | Number of system exec calls per second. |
| 8 | rchar | Number of characters transferred per second by system read calls |
| 9 | wchar | Number of characters transfreed per second by system write calls |
| 10 | pgout | Number of page out requests per second |
| 11 | ppgout | Number of pages, paged out per second |
| 12 | pgfree | Number of pages per second placed on the free list. |
| 13 | pgscan | Number of pages checked if they can be freed per second |
| 14 | atch | Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second |
| 15 | pgin | Number of page-in requests per second |
| 16 | ppgin | Number of pages paged in per second |
| 17 | pflt | Number of page faults caused by protection errors (copy-on-writes). |
| 18 | vflt | Number of page faults caused by address translation . |
| 19 | runqsz | Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run. |
| 20 | freemem | Number of memory pages available to user processes |
| 21 | freewap | Number of disk blocks available for page swapping. |
| 22 | usr | Portion of time (%) that cpus run in user mode |
| Sample of dataset | | |
| Table 1.1 | | |
|  | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Problem-1 | | | | | | | |
| 1.1 | | | | Read the data and do exploratory data analysis (3 pts). Describe the data briefly. Interpret the inferences for each (3 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct. | | | |
|  | | | | Exploratory data analysis | | | |
| 1 | | | | Iread | 8192 non-null | int64 | |
| 2 | | | | Iwrite | 8192 non-null | int64 | |
| 3 | | | | scall | 8192 non-null | int64 | |
| 4 | | | | sread | 8192 non-null | int64 | |
| 5 | | | | swrite | 8192 non-null | int64 | |
| 6 | | | | fork | 8192 non-null | float 64 | |
| 7 | | | | exec | 8192 non-null | float 64 | |
| 8 | | | | rchar | 8088 non-null | float 64 | |
| 9 | | | | wchar | 8177 non-null | float 64 | |
| 10 | | | | pgout | 8192 non-null | float 64 | |
| 11 | | | | ppgout | 8192 non-null | float 64 | |
| 12 | | | | pgfree | 8192 non-null | float 64 | |
| 13 | | | | pgscan | 8192 non-null | float 64 | |
| 14 | | | | atch | 8192 non-null | float 64 | |
| 15 | | | | pgin | 8192 non-null | float 64 | |
| 16 | | | | ppgin | 8192 non-null | float 64 | |
| 17 | | | | pflt | 8192 non-null | float 64 | |
| 18 | | | | vflt | 8192 non-null | float 64 | |
| 19 | | | | runqsz | 8192 non-null | object | |
| 20 | | | | freemem | 8192 non-null | int64 | |
| 21 | | | | freewap | 8192 non-null | int64 | |
| 22 | | | | usr | 8192 non-null | int64 | |
| Descriptive statistics of data | | | | | | | |
| Table 1.2 | | | | | | | |
|  | | | | | | | |
| Five number summary of columns | | | | | | | |
| Five number summary of - ppgin  Minimum: 0.0  25%: 0.6  50% or Median: 3.8  75%: 13.8  Maximum: 292.61  IQR: 13.200000000000001  ----------------------------------------------------------------------------  Five number summary of - pflt  Minimum: 0.0  25%: 25.0  50% or Median: 63.8  75%: 159.6  Maximum: 899.8  IQR: 134.6  ----------------------------------------------------------------------------  Five number summary of - vflt  Minimum: 0.2  25%: 45.4  50% or Median: 120.4  75%: 251.8  Maximum: 1365.0  IQR: 206.4  ----------------------------------------------------------------------------  Five number summary of - freemem  Minimum: 55.0  25%: 231.0  50% or Median: 579.0  75%: 2002.25  Maximum: 12027.0  IQR: 1771.25  ----------------------------------------------------------------------------  Five number summary of - freeswap  Minimum: 2.0  25%: 1042623.5  50% or Median: 1289289.5  75%: 1730379.5  Maximum: 2243187.0  IQR: 687756.0  ----------------------------------------------------------------------------  Five number summary of - usr  Minimum: 0.0  25%: 81.0  50% or Median: 89.0  75%: 94.0  Maximum: 99.0  IQR: 13.0  ---------------------------------------------------------------------------- | | | | | | | |
| Five number summary of - exec  Minimum: 0.0  25%: 0.2  50% or Median: 1.2  75%: 2.8  Maximum: 59.56  IQR: 2.5999999999999996  ----------------------------------------------------------------------------  Five number summary of - rchar  Minimum: 278.0  25%: 34091.5  50% or Median: 125473.5  75%: 267828.75  Maximum: 2526649.0  IQR: nan  ----------------------------------------------------------------------------  Five number summary of - wchar  Minimum: 1498.0  25%: 22916.0  50% or Median: 46619.0  75%: 106101.0  Maximum: 1801623.0  IQR: nan  ----------------------------------------------------------------------------  Five number summary of - pgout  Minimum: 0.0  25%: 0.0  50% or Median: 0.0  75%: 2.4  Maximum: 81.44  IQR: 2.4  ----------------------------------------------------------------------------  Five number summary of - ppgout  Minimum: 0.0  25%: 0.0  50% or Median: 0.0  75%: 4.2  Maximum: 184.2  IQR: 4.2 | | | | | | | |
| ----------------------------------------------------------------------------  Five number summary of - pgfree  Minimum: 0.0  25%: 0.0  50% or Median: 0.0  75%: 5.0  Maximum: 523.0  IQR: 5.0  ----------------------------------------------------------------------------  Five number summary of - pgscan  Minimum: 0.0  25%: 0.0  50% or Median: 0.0  75%: 0.0  Maximum: 1237.0  IQR: 0.0  ----------------------------------------------------------------------------  Five number summary of - atch  Minimum: 0.0  25%: 0.0  50% or Median: 0.0  75%: 0.6  Maximum: 211.58  IQR: 0.6  ----------------------------------------------------------------------------  Five number summary of - scall  Minimum: 109.0  25%: 1012.0  50% or Median: 2051.5  75%: 3317.25  Maximum: 12493.0  IQR: 2305.25  ----------------------------------------------------------------------------  Five number summary of - sread  Minimum: 6.0  25%: 86.0  50% or Median: 166.0  75%: 279.0  Maximum: 5318.0  IQR: 193.0  ----------------------------------------------------------------------------  Five number summary of - fork  Minimum: 0.0  25%: 0.4  50% or Median: 0.8  75%: 2.2  Maximum: 20.12  IQR: 1.8000000000000003  ---------------------------------------------------------------------------- | | | | | | | |
| Five number summary of - pgin  Minimum: 0.0  25%: 0.6  50% or Median: 2.8  75%: 9.765  Maximum: 141.2  IQR: 9.165000000000001  ----------------------------------------------------------------------------  Five number summary of - lread  Minimum: 0.0  25%: 2.0  50% or Median: 7.0  75%: 20.0  Maximum: 1845.0  IQR: 18.0  ----------------------------------------------------------------------------  Five number summary of - lwrite  Minimum: 0.0  25%: 0.0  50% or Median: 1.0  75%: 10.0  Maximum: 575.0  IQR: 10.0  ----------------------------------------------------------------------------  Five number summary of - swrite  Minimum: 7.0  25%: 63.0  50% or Median: 117.0  75%: 185.0  Maximum: 5456.0  IQR: 122.0  ---------------------------------------------------------------------------- | | | | | | | |
| Univariate analysis | | | | | | | |
| IMG_256  Fig 1.1 | | | | | | | |
| IMG_256  Fig 1.2 | | | | | | | |
| All the column have outliers. Also most of the columns seems to be not normal distribution.  We will check the skewness of the columns provided. | | | | | | | |
| Table 1.3 | | | | | | | |
| If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.  If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.  If the skewness is less than -1 or greater than 1, the data are highly skewed.  From that skewness table it can be observed that most of the columns are right skewed. | | | | | | | |
| Bivariate analysis | | | | | | | |
| IMG_256  Fig 1.3 | | | | | | | |
| IMG_256  Fig 1.4 | | | | | | | |
| Fig 1.4 shows pair plot between all the numerical columns.  We will construct a heat map to show the correlation and collinearity between columns. | | | | | | | |
| IMG_256  Fig 1.5 | | | | | | | |
| Table 1.4 | | | | | | | |
| Table 1.4 shows correlation between the variables.Heat map, pair plot, and correlation table show that some of the columns are highly correlated.0.7 to 0.9 are considered a high correlation, 0.5 to 0.7 is moderately correlated, and below 0.3 indicate variables that have a low correlation. | | | | | | | |
|  | | | | | | | |
| 1.2 | | Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there. | | | | | | |
| Null values are present in the data set. There are 104 null values in rchar and 15 null values in wchar. We choose to replace the null values with median of each column.  There are so many values that equals zero in the data set . These values represent number of pages/characters transfreed / read / write per second for different attributes . So we choose not to treat them as they are significant for the data set. Dropping the values can adversely affect the data.  Number of zeros in lread = 675  Number of zeros in lwrite = 2684  Number of zeros in scall = 0  Number of zeros in sread = 0  Number of zeros in swrite = 0  Number of zeros in fork = 21  Number of zeros in exec = 21  Number of zeros in rchar = 0  Number of zeros in wchar = 0  Number of zeros in pgout = 4878  Number of zeros in ppgout = 4878  Number of zeros in pgfree = 4869  Number of zeros in pgscan = 6448  Number of zeros in atch = 4575  Number of zeros in pgin = 1220  Number of zeros in ppgin = 1220  Number of zeros in pflt = 3  Number of zeros in vflt = 0  Number of zeros in runqsz = 0  Number of zeros in freemem = 0  Number of zeros in freeswap = 0  Number of zeros in usr = 283 | | | | | | | | |
| Table 1.5 | | | | | | | | |
| The outlier proportions in dataset are as shown above.  Encoding the categorical column can result in the creation of new feature. The categorical column to be encoded is ‘runqsz’ .This column has only two levels , after encoding the column is changed to binary with feature name ‘runqsz\_Not\_CPU\_Bound’  There are no duplicated values in the data set.  We do not have to create new feature columns.Also some new columns will be created during the encoding. | | | | | | | | |
| It is clear from the outlier proportion table that all the columns have outliers. We choose not to treat the outliers because an industry expert can only evaluate the relevance of the outlier values. Replacing the outlier values without prior knowledge in the industry either by capping or imputing with median can change the out come of predictions. | | | | | | | | |
| 1.3 | Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. 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 Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning. | | | | | | | |
| Encoded data set    Table 1.6 | | | | | | | | |
| LINEAR REGRESSION USING SCIKIT LEARN - BASE MODEL | | | | | | | | |
| Base model intercept = 44.63798454  Base model co-efficient = [-1.98873283e-02, 4.79371418e-03, 1.00815134e-03,  -4.70609776e-04, -2.04096786e-03, -1.72224970e+00,  -8.96269805e-02, -4.06160177e-06, -1.16391775e-05,  -1.73903805e-01, 9.89391175e-02, -7.03391933e-02,  8.62333138e-03, -7.85628241e-02, 9.12669349e-02,  -5.93811492e-02, -4.15086898e-02, 2.22839067e-02,  -1.61716713e-03, 3.21923595e-05, 7.79079271e+00] | | | | | | | | |
| |  |  |  | | --- | --- | --- | | SCIKIT LEARN | BASE MODEL  TRAIN DATA | BASE MODEL  TEST DATA | | R2 | 0.6428 | 0.6312 | | Adj\_R2 | 0.6415 | 0.6280 | | RMSE | 10.81 | 11.59 |     Table 1.7 | | | | | | | | |
| LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BASE MODEL | | | | | | | | |
| Table 1.8 | | | | | | | | |
| Table 1.8 shows OLS regression result .Here R2 is 0.643 and adjusted R2 is 0.642 | | | | | | | | |
| LINEAR REGRESSION USING STATS MODEL TEST DATA- BASE MODEL | | | | | | | | |
| Table 1.9 | | | | | | | | |
| VIF of the predictors | | | | | | | | |
| Table 1.10 | | | | | | | | |
| Variance inflation factor measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression.The default VIF cut-off value is 5, only variables with a VIF less than 5 will be included in the model.  We will remove all the columns with high VIF(>5) to remove multicollinearity from model. | | | | | | | | |
| |  |  |  | | --- | --- | --- | | STATS MODEL | BASE MODEL  TRAIN DATA | BASE MODEL  TEST DATA | | R2 | 0.643 | 0.641 | | Adj\_R2 | 0.642 | 0.638 | | RMSE | 10.81 | 11.59 |   Table 1.11 | | | | | | | | |
| Creating different models by removing columns having high VIF :  First we will remove ‘sread’ .OLS table after removing sread is shown below. | | | | | | | | |
| Table 1.12 | | | | | | | | |
| From the table above we can see the minute change in adjusted R2 value.So we will drop this from the original train set .  Similarly the following columns are dropped one by one from the train set to reduce multi-collinearity.   1. fork 2. ppgout 3. pgfree 4. pgin 5. vflt | | | | | | | | |
| VIF of the predictors after removing the above mentioned columns: | | | | | | | | |
| Table 1.13 | | | | | | | | |
| From the table above we can see that no columns have VIF value greater than 5.  Now we have to remove the column that does not contribute for the prediction .p\_value in the OLS regression table helps to find the unwanted column.So we will remove all columns that have p\_value >0.05 .  The OLS regression table after removal of the columns having high VIF. | | | | | | | | |
|  | | | | | | | | |
| Table 1.14 | | | | | | | | |
| Now we will remove all the column that have p\_value > 0.05 . | | | | | | | | |
| First we will drop ‘Iwrite’ from the train set. | | | | | | | | |
| Table 1.15 | | | | | | | | |
| Similarly we will remove columns ‘pgscan’ and ‘ppgin’ from the train set.  OLS table after removal of the above mentioned column is as shown below. | | | | | | | | |
| LINEAR REGRESSION USING STATS MODEL TRAIN DATA- BEST MODEL | | | | | | | | |
| Table 1.16 | | | | | | | | |
|  | | | | | | | | |
| LINEAR REGRESSION USING STATS MODEL TEST DATA- BEST MODEL | | | | | | | | |
| To create the best model using test data we have to drop the following columns.   1. sread 2. fork 3. ppgout 4. pgfree 5. pgin 6. Vflt 7. Iwrite 8. pgscan 9. ppgin | | | | | | | | |
| Table 1.17 | | | | | | | | |
| |  |  |  | | --- | --- | --- | | STATS MODEL | BEST MODEL  TRAIN DATA | BEST MODEL  TEST DATA | | R2 | 0.637 | 0.632 | | Adj\_R2 | 0.636 | 0.630 | | RMSE | 10.90 | 11.58 |   Table 1.18 | | | | | | | | |
| LINEAR REGRESSION USING SCIKIT LEARN- BEST MODEL | | | | | | | | |
| Base model intercept = 40.72539792  Base model co-efficient = [ -1.53513810e-02, 1.40338271e-03, -2.41589875e-03,  -1.22430902e-01, -1.58531549e-06, -9.19198353e-06,  -1.54498183e-01, 5.50322353e-02, -4.16839155e-02,  -1.77197280e-03, 3.46498091e-05, 8.35931600e+00]   |  |  |  | | --- | --- | --- | | SCIKIT LEARN | BEST MODEL  TRAIN DATA | BEST MODEL  TEST DATA | | R2 | 0.637 | 0.624 | | Adj\_R2 | 0.624 | 0.630 | | RMSE | 10.90 | 11.58 |   Table 1.19 | | | | | | | | |
| 1.4 | | | Inference: Basis on these predictions, what are the insights and recommendations.  Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. | | | | | |
| Here we have used both scikit learn and stats model library to create the model. User CPU time is time spent on the processor running your program's code (or code in libraries).We have build a model which predicts the different attributes that affects the user CPU time.  The provided dataset consists of 22 columns . All these columns may not equally contribute to the CPU user time. Also while creating a linear regression model the effect of multi collinearity should be taken care. Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Although multicollinearity does not affect the regression estimates, it makes them vague, imprecise, and unreliable. Thus, it can be hard to determine how the independent variables influence the dependent variable individually. This inflates the standard errors of some or all of the regression coefficients.  The initial step of creating a model is to remove multicollinearity from the model. The following columns are removed from the dataset to remove in order the multicollinearity.   1. sread 2. fork 3. ppgout 4. pgfree 5. pgin 6. vflt | | | | | | | | |
| Now the p\_value should be checked. A low P-value (< 0.05) means that the coefficient is likely not to equal zero. A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable (User time). So the features with high p-value are removed from the model. The following columns have high p-value and are removed from the model. | | | | | | | | |
| 1. lwrite 2. pgscan 3. ppgin | | | | | | | | |
| Below show is the linear equation created from the best available model. | | | | | | | | |
| usr = 46.46160190866891 + -0.01731562635082673 \* ( lread ) +  0.0012228813935545313 \* ( scall ) +  -0.003787461868611286 \* ( swrite ) +  -0.2079981036627988 \* ( exec ) +  -3.884513749041014e-06 \* ( rchar ) +  -1.2038322893223802e-05 \* ( wchar ) +  -0.12701651876594405 \* ( pgout ) +  -0.05446972426887098 \* ( atch ) +  -0.0401047952702351 \* ( pflt ) +  -0.0016062898330734354 \* ( freemem ) +  3.1263613790683524e-05 \* ( freeswap ) +  7.715468801199788 \* ( runqsz\_Not\_CPU\_Bound) | | | | | | |  | |
| From the linear equation we can conclude that :  lread column reduces the portion of time cpu runs on user mode.If Iread increases by a unit the target variable reduces by 0.017. Variable scall increases the user , ie. number of system calls of all types per second increases the portion of time cpu runs on user mode.Similarly columns swrite,exec,rchar,wchar,pgout,atch,pflt and freemem reduces the potion of time cpu runs on user mode. Number of system exec calls per second (exec) is the most important value that reduces the portion of time cpu runs on user mode than other columns mentioned. Freewap and runqsz\_Not\_CPU\_Bound are the columns that increases the target variable. runqsz\_Not\_CPU\_Bound is the most important variable in terms of increase in time cpu runs on user model.  The columns we have selected here is the ones which is significant to the target variable in terms of p\_value and removed multi-collinearity from data set. The actual columns may differ with respect to the real world scenarios where we might have to consider the columns we have removed from the dataset. Only a domain expert could identify the importance of the columns. | | | | | | | | |

|  |  |  |
| --- | --- | --- |
| ****Contraceptive\_method\_dataset****  Summary | | |
| You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics. | | |
| Introduction | | |
| The data consists of 1473 rows and 10columns . Univariate ,bivariate analysis of the data will be conducted here based on the various factors. Then the model is created which predict the they use a contraceptive method of choice based on their demographic and socio-economic characteristics. Model will be created using logistic regression method and their performance will be compared. | | |
| Data description | | |
| 1 | Wife\_age | Age of wife |
| 2 | Wife\_ education | Education status of wife (4 levels) |
| 3 | Husband\_education | Education status of husband (4 levels) |
| 4 | No\_of\_children\_born | Number of children ever born |
| 5 | Wife\_religion | Wife's religion ( Non-Scientology, Scientology) |
| 6 | Wife\_Working | Wife's now working(Yes,No) |
| 7 | Husband\_Occupation | Husband's occupation(Categorical 1,2,3,4) |
| 8 | Standard\_of\_living\_index | Standard-of-living index(Ordinal from very low to high) |
| 9 | Media\_exposure | Media exposure (Good, Not good) |
| 10 | Contraceptive\_method\_used | Contraceptive method used (Yes,No) |

|  |  |  |  |
| --- | --- | --- | --- |
| Sample of dataset | | | |
| Table 2.1 | | | |
| Problem-2 | | | |
| Exploratory Data Analysis | | | |
|  | Column | Non-Null Count | Data type |
| 1 | Wife\_age | 1402 non-null | float64 |
| 2 | Wife\_ education | 1473 non-null | object |
| 3 | Husband\_education | 1473 non-null | object |
| 4 | No\_of\_children\_born | 1452 non-null | float64 |
| 5 | Wife\_religion | 1473 non-null | object |
| 6 | Wife\_Working | 1473 non-null | object |
| 7 | Husband\_Occupation | 1473 non-null | int64 |
| 8 | Standard\_of\_living\_index | 1473 non-null | object |
| 9 | Media\_exposure | 1473 non-null | object |
| 10 | Contraceptive\_method\_used | 1473 non-null | object |
| Descriptive statistics of data: | | | |
| Table 2.2 | | | |
|  | | | |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| 2.1 | Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis. | | |
| Univariate analysis | | | |
| IMG_256 | | IMG_256 | |
| Fig 2.1 | | | |
| These are the box plots for numerical columns in the dataset. It is evident from the box plot that there are outliers in the column number of children born. | | | |
| IMG_256 | | | IMG_256 |
| Fig 2.2 | | | |
| The inference from box plot can be proven true by visualizing histogram for the numerical columns.Here age column seems to be normally distributed and other columns are right skewed. | | | |
| Bivariate & Multi-variate analysis | | | |
| IMG_256 Fig 2.3 | | | |
| Count plot for the categorical columns are shown above. Most of wife’s are educated tertiary and the number of uneducated husbands are less than wife. Scientology is the religion most wife’s believes in also most of the then are not working. It is also evident from the count plot that most of the wife’s does not prefer the using of contraceptive method. | | | |
| Pairplot of all the numerical columns in dataset: | | | |
| IMG_256 Fig 2.4 | | | |
| Heat map to show correlation of all the numerical columns in dataset: | | | |
| IMG_256 Fig 2.5 | | | |
| The heat map of the dataset shows that there is only moderate correlation between columns. 0.7 to 0.9 is considered a high correlation, 0.5 to 0.7 is moderately correlated and below 0.3 to 0.5 indicate variables that have a low correlation. | | | |
| Correlation table : | | | |
| Table 2.3 | | | |

|  |
| --- |
| Outlier percentage in numerical columns : |
| Table 2.4 |
| 3% of the data in number of children born columns are outliers.  The 3% outlier can be removed by capping the values.We have defined outliers by lower and upper limit ,if the data points falls below lower limit and above upper limit it is considered as an outlier.By capping we will assign the upper or lower limit outlier values. |
| There are 71 null values in Wife’s age column and 21 null values in number of children born column.  To impute the null values we will check the density plot of these columns with mean and median shown as axial vertical line. |
| IMG_256  Fig 2.6 |
| IMG_256  Fig 2.7 |
| We will impute the null values with median values of each column. |
| While checking for duplicated columns it can be observed that 90 rows in the dataset are duplicated.  The duplicated rows should be removed from the dataset .  After removal of these duplicated the dataset shrinks to 1383 rows. |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| 2.2 | | | Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and build multiple models with different predictors. |
| Converting object to categorical or code: | | | |
| Table 2.5 | | | |
| All the categorical columns are encoded in the above table shows.  Standard of living index is a ordinal categorical variable so it is ranked as 1 : Very low, 2: Low,  3: High , 4: Very high.  All other categorical columns are encoded using get\_dummies . One dummy variable should be dropped to avoid being trapped by the dummy variable. Variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity. | | | |
| Train\_test\_split | | | |
| Train data first five rows: | | | |
| Table 2.6 | | | |
| Test data first five rows: | | | |
| Table 2.7 | | | |
| Target variable train data first five rows: | | | |
| Table 2.8 | | | |
| Target variable test data first five rows: | | | |
| Table 2.9 | | | |
|  | | | |
| LOGISTIC REGRESSION USING STATS MODEL- BASE MODEL | | | |
| Table 2.10 | | | |
| VIF of the predictors: | | | |
| Table 2.11  From the VIF table it can be observed that wife age has the highest value of VIF followed by husband education and number of children born.  We will remove all the columns that have VIF value more than 5 in the order given in table.  In linear regression we check R2 and Adjusted R2 values here we will check pseudo-R2 to find the important independent variables that affect the target variable.  pseudo-R2 of base model is 0.1461.  Wife’s age will be removed at first | | | |
| S  Table 2.12 | | | |
| It is clear that wife’s age column has reduced the pseudo R2 value by 0.05  Similarly we have removed number of children born column and the pseudo R2 value changed to 0.056 .So an overall reduction of 0.09 can be observed. This proves that both these columns contribute heavily in the prediction of contraceptive method usage and cannot be removed from the dataset.  We will now remove husband’s occupation column . | | | |
| Table 2.13 | | | |
| VIF of the predictors after removing the column mentioned above : | | | |
| Table 2.14 | | | |
| Husband\_education\_Tertiary and Standard\_of\_living\_index columns are removed in order to reduce multicollinearity.After the removal of these columns the logit regression table will be as shown below. | | | |
| Table 2.15 | | | |
| All the columns that can be removed due to collinearity are removed ,except wife age and number of child born.These column affect the target variable more than other removed columns.  Now we will remove those columns that have p-value greater than 0.05.   1. Wife\_Working\_Yes 2. Wife\_education\_Uneducated 3. Husband\_education\_Secondary 4. Husband\_education\_Uneducated   These columns are removed in order .  Logit regression table after removal of these columns are shown below. | | | |
| LOGISTIC REGRESSION USING STATS MODEL- BEST MODEL | | | |
| Table 2.16 | | | |
| There is only a slight reduction in the psuedo R2 value also it can be observed that no columns have higher p\_value. So we can conclude that this is the final and best model available. | | | |
| VIF of the predictors after removing the columns : | | | |
| Table 2.17 | | | |
| LOGISTIC REGRESSION USING SCIKIT LEARN - BASE MODEL | | | |
| Base model intercept = 0.58201826  Base model co-efficient = [-0.10390837, 0.60073882, 0.16911357, 0.32585482, -0.53292681,  -0.00342944, -0.35733293, 0.50497762, 1.38811975, -0.26457862,  0.4685209 , 0.0577456 , -0.66051907] | | | |
| LOGISTIC REGRESSION USING SCIKIT LEARN - BEST MODEL | | | |
| Best model intercept = 1.74527352  Best model co-efficient = [ -0.10018808, 0.58896143, -0.63908663,  -0.71943599, 0.59363862,1.39816272]  Here we have removed columns such as 'Husband\_Occupation' , 'Husband\_education\_Tertiary',' Standard\_of\_living\_index' , 'Wife\_Working\_Yes', 'Wife\_education\_Uneducated' , 'Husband\_education\_Secondary', 'Husband\_education\_Uneducated' from the data set to create best model. | | | |
| 2.3 | | Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized. | |
| Base mode Train data :Accuracy score, Confusion matrix ,roc\_auc\_score,f1 score, precision and recall | | | |
| Accuracy score : 0.71 | | | |
| Confusion matrix :   |  |  |  | | --- | --- | --- | |  | PREDICTION | | | ACTUAL | 0 (PREDICTED NEGATIVE) | 1 (PREDICTED POSITIVE) | | 0 (ACTUAL NEGATIVE) | 235(TN) | 178(FP) | | 1 (ACTUAL POSTIVE) | 99(FN) | 456(TP) |   Table 2.18 | | | |
| Classification report : | | | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | precision | | Recall | | F1-score | | support | | 0 | 0.70 | 0.57 | | 0.63 | | 413 | | | 1 | 0.72 | 0.82 | | 0.77 | | 555 | | | accuracy |  |  | | 0.71 | | 968 | | | macro avg | 0.71 | 0.70 | | 0.70 | | 968 | | | weighted avg | 0.71 | 0.71 | | 0.71 | | 968 | |   Table 2.19 | | | |
| ROC\_AUC\_SCORE : 0.764 | | | |
| ROC\_CURVE PLOT : | | | |
| IMG_256  Fig 2.20 | | | |
| Base mode Test data :Accuracy score, Confusion matrix ,roc\_auc\_score,f1 score, precision and recall | | | |
| Accuracy score : 0.78111 | | | |
| Confusion matrix :   |  |  |  | | --- | --- | --- | |  | PREDICTION | | | ACTUAL | 0 (PREDICTED NEGATIVE,NOT CLAIMED) | 1 (PREDICTED POSITIVE,CLAIMED) | | 0 (ACTUAL NEGATIVE,NOT CLAIMED) | 101(TN) | 99(FP) | | 1 (ACTUAL POSTIVE,CLAIMED) | 53(FN) | 162(TP) |   Table 2.20 | | | |
| Classification report : | | | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | precision | | Recall | | F1-score | | support | | 0 | 0.66 | 0.51 | | 0.57 | | 605 | | | 1 | 0.62 | 0.75 | | 0.68 | | 215 | | | accuracy |  |  | | 0.63 | | 415 | | | macro avg | 0.64 | 0.63 | | 0.63 | | 415 | | | weighted avg | 0.64 | 0.63 | | 0.63 | | 415 | |   Table 2.21 | | | |
| ROC\_AUC\_SCORE : 0.681 | | | |
| ROC\_CURVE PLOT : | | | |
| IMG_256  Fig 2.21 | | | |
| Accuracy is the number of correct predictions made divided by the total number of predictions made.Train data has 71% and test data has 78% accuracy score.Lesser the false predictions more the accuracy. An accuracy measure of anything between 70%-90% is not only ideal, it's realistic.  Confusion matrix is a 2x2 tabular structure reflecting the performance of the model in four blocks.True positive(TP) and True negative(TN) are the correct predictions.False positive(FP) and False negative(FN) are the incorrect predictions.Lesser the false predictions more the accuracy.  There are four ways to check if the predictions are right or wrong:  TN / True Negative: the case was negative and predicted negative.  TP / True Positive: the case was positive and predicted positive.  FN / False Negative: the case was positive but predicted negative.  FP / False Positive: the case was negative but predicted positive.  A Classification report is used to measure the quality of predictions. True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report | | | |
| Best mode Train data :Accuracy score, Confusion matrix ,roc\_auc\_score,f1 score, precision and recall | | | |
| Accuracy score : 0.71 | | | |
| Confusion matrix :   |  |  |  | | --- | --- | --- | |  | PREDICTION | | | ACTUAL | 0 (PREDICTED NEGATIVE) | 1 (PREDICTED POSITIVE) | | 0 (ACTUAL NEGATIVE) | 226(TN) | 187FP) | | 1 (ACTUAL POSTIVE) | 105(FN) | 450(TP) |   Table 2.22 | | | |
| Classification report : | | | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | precision | | Recall | | F1-score | | support | | 0 | 0.68 | 0.55 | | 0.61 | | 413 | | | 1 | 0.71 | 0.81 | | 0.76 | | 555 | | | accuracy |  |  | | 0.70 | | 968 | | | macro avg | 0.69 | 0.68 | | 0.68 | | 968 | | | weighted avg | 0.70 | 0.70 | | 0.69 | | 968 | |   Table 2.23 | | | |
| ROC\_AUC\_SCORE : 0.748 | | | |
| ROC\_CURVE PLOT : | | | |
| Fig 2.22 | | | |
| Best mode Test data :Accuracy score, Confusion matrix ,roc\_auc\_score,f1 score, precision and recall | | | |
| Accuracy score : 0.64 | | | |
| Confusion matrix :   |  |  |  | | --- | --- | --- | |  | PREDICTION | | | ACTUAL | 0 (PREDICTED NEGATIVE) | 1 (PREDICTED POSITIVE) | | 0 (ACTUAL NEGATIVE) | 96(TN) | 104(FP) | | 1 (ACTUAL POSTIVE) | 45(FN) | 170(TP) |   Table 2.24 | | | |
| Classification report : | | | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | precision | | Recall | | F1-score | | support | | 0 | 0.68 | 0.48 | | 0.56 | | 200 | | | 1 | 0.62 | 0.79 | | 0.70 | | 215 | | | accuracy |  |  | | 0.64 | | 415 | | | macro avg | 0.65 | 0.64 | | 0.63 | | 415 | | | weighted avg | 0.65 | 0.64 | | 0.63 | | 415 | |   Table 2.25 | | | |
| ROC\_AUC\_SCORE : 0.702 | | | |
| ROC\_CURVE PLOT : | | | |
| Fig 2.22 | | | |
| 2.4 | Inference: Basis on these predictions, what are the insights and recommendations.  Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. | | |
| We will choose the best model to predict the women’s use contraceptive method.  The target variable is binary 0 & 1. 1 indicate the women prefer to use contraceptive method.1 is denoted in confusion matrix as true positive.  Precision : The precision tells us the accuracy of positive predictions.Among the points identified as positive by the model,how many are really positive this is the objective of precision .High precision means not many True values were predicted as False.Here we need to check precision for the ‘1’ in confusion matrix.Train data shows there is 71% precision and test data shows 62%.  The recall, also named sensitivity, tells us the fraction of correctly identified positive predictions.  What fraction of the True predictions were actually True , recall finds this property. High recall means Predicted most True values correctly.Here 81% recall value is seen for train data and 79% for the test data.  The f1-score, or F measure, measures precision and recall at the same time by finding the harmonic mean of the two values.Per classification report we have 76% and 70% of f1-score in train and test data respectively.  AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between  0.6-0.7 and failed for AUC values between 0.5-0.6.Here we have AUC score of 0.74for train data and 0.70for test data which implies the values are fair.The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. | | | |
| From confusion matrix it is clear that there are more number of values in true positives(TP) for both train and test data.This indicate the models ability to predict actual one as one in other words it can be established that this model can be used to correctly predict the use of contraceptive method.Also recall value confirm that how many actual true data points are identified as True. | | | |
| Coefficients are easy to interpret in linear regression but not in logistic regression, as the estimates produced in the latter are not as intuitive. In logistic type regression, the logit transformation reveals the independent variable’s impact on the variation of the dependent variable’s natural logarithm of the odds.  **Log(odds) = -1.74527352 + -0.10018808( Wife\_age) + 0.58896143(No\_of\_children\_born) +**  **-0.63908663(Wife\_religion\_Scientology) + -0.71943599(Media\_exposure\_Not\_Exposed) + 0.59363862(Wife\_education\_Secondary)+ 1.39816272(Wife\_education\_Tertiary)**  If we substitute values for each variables we can find the probability of values belonging to 0 & 1 (ie use of contraceptive method)  **Odds = exp(log(odds)**  **Probability = odds/(1+odds)**  So using these formulas we find the probability for increase in value of each variable. | | | |