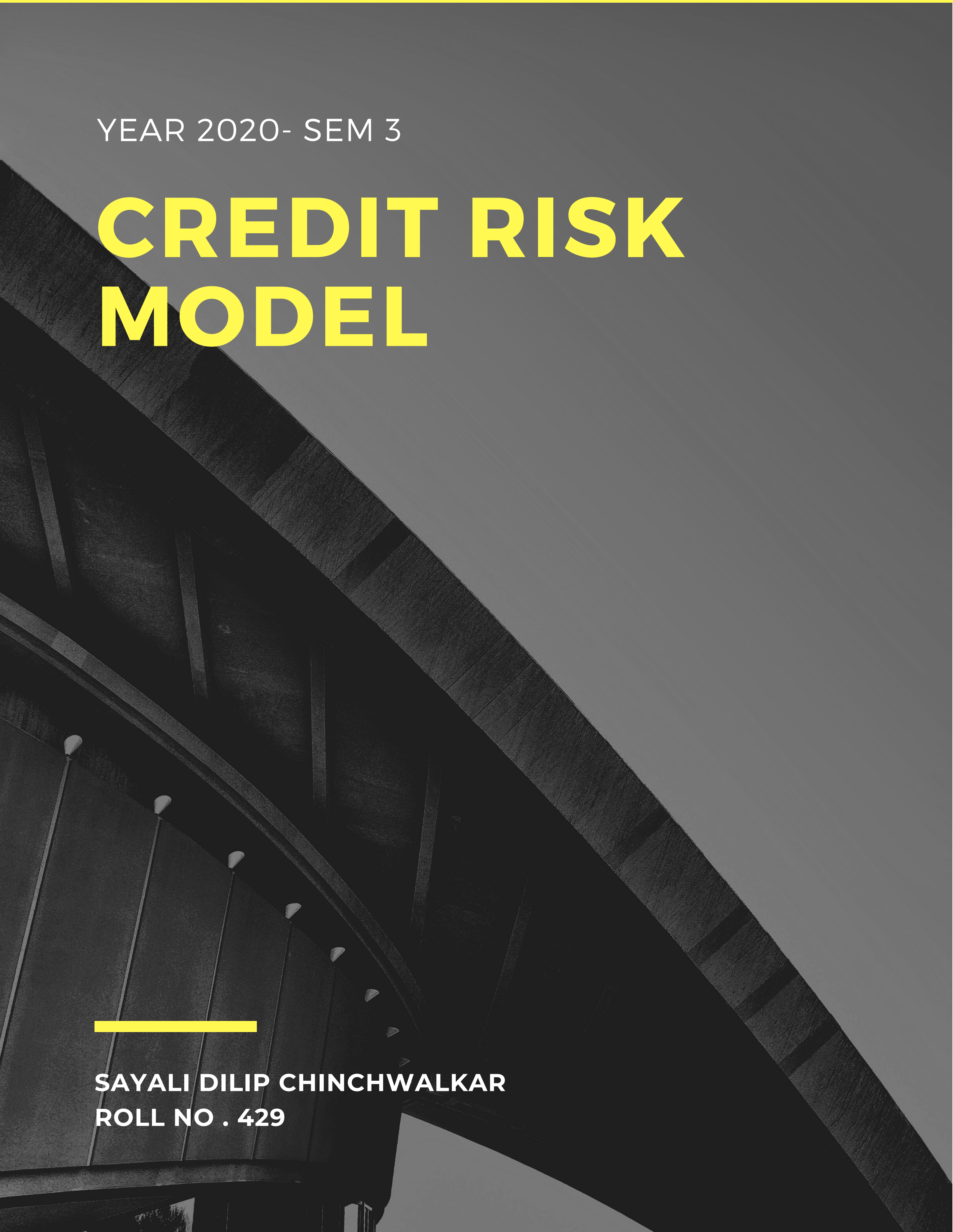
YEAR 2020- SEM 3

CREDIT RISK MODEL



SAYALI DILIP CHINCHWALKAR
ROLL NO . 429

TABLE OF CONTENT

01. Background

02. Objective

03. Summary

04. Data collection & EDA

05. Data cleaning

06. logistic regression model

07. Statistical testing

08. Matrix and ROC Curve

09. Cross-validation

10. Decision tree

11. Application of credit scoring model

DISCLAIMER

As per the new requirement, all banks are expected to adopt either a standardised approach or an internal model approach. the credit risk board committee of the bank has chosen an internal model approach for credit risk assessment.
This model would then be used by the bank to make the credit decision. to check the creditworthiness of the applicant for a loan, loan criteria, policy, rule. The models provide information on the level of a borrower's credit risk at any particular time.
If the lender fails to detect the credit risk in advance, it exposes them to the risk of default and loss of funds.
Lenders rely on the validation provided by credit risk analysis models to make key lending decisions on whether or not to extend credit to the borrower and the credit to be charged.
In an efficient market system, banks charge a high-interest rate for high-risk loans as a way of compensating for the high risk of default.

OBJECTIVE

Pinpointing the amount of risk that comes with each loan is a difficult task. Some of the factors that go into the complex credit risk calculation include the probability of default, the amount of exposure at the time of default, how much the loan is expected to be worth at the time of default, and the overall loss if there is a default. to predict the likelihood of default, lenders leverage historical data to guess how a consumer will behave in the future. analyst create a model that will identify the creditworthiness of loan applicants. determine the probability of default of a potential borrower, to quantify the level of risk if the borrower made any default. The rise of analytics and Big Data have helped enhance the process of credit risk measurement. By leveraging data, there is less guesswork and more science behind the ability to predict whether someone will default on any given loan.

SUMMARY

To produce an internal credit scoring model, we have been given historical data of the loan applicants. Data collection, exploratory data analysis, data cleaning was the priority. the logistic regression model has been created to check the significance of the variable. in this model, we have chosen those variables which have passed the underlying assumption. LR TEST, pseudo r2, Hosmer Lemeshow test, Somer's D test were tested to confirm the goodness of fit of the model, i.e selected variables are added value to the model are not insignificant. the threshold for the matrix was considered as 0.4, which is giving the accuracy of 74% for the training and testing data set. ROC, k fold cross-validation has been used to estimate the accuracy of the model. decision tree methods have also been performed to check the best fit model and to analyse which model is giving the highest accuracy

DATA COLLECTION & EDA

Data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information. The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

Exploratory data analysis (EDA) have been used for the initial analysis and findings done with data sets, maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models

```
> str(insurance)
'data.frame':	6000 obs. of  19 variables:
 $ cust_id             : int  6252029 5110070 2846491 9264318 9412980 6111903 1613014 7940321 1673336 2336197
 $ acc_no              : num  6.25e+11 5.11e+11 2.85e+11 9.26e+11 9.41e+11 ...
 $ checking_balance    : chr  "< 0 USD" "1 - 200 USD" "< 0 USD" "1 - 200 USD" ...
 $ months_loan_duration: int  12 36 11 15 10 14 24 18 24 30 ...
 $ credit_history      : chr  "good" "good" "critical" "good" ...
 $ purpose             : chr  "car" "car" "car" "renovations" ...
 $ amount..USD.        : int  1274 12389 3939 1308 1924 3973 6615 2124 11938 2406 ...
 $ savings_balance     : chr  "< 100 USD" "unknown" "< 100 USD" "< 100 USD" ...
 $ employment_duration : chr  "< 1 year" "1 - 4 years" "1 - 4 years" "> 7 years" ...
 $ percent_of_income   : int  3 1 1 4 1 1 2 4 2 4 ...
 $ years_at_residence  : int  1 4 2 4 4 4 NA 4 3 NA ...
 $ age                 : int  37 37 40 38 38 22 75 24 39 23 ...
 $ other_credit        : chr  "none" "none" "none" "none" ...
 $ housing             : chr  "own" "other" "own" "own" ...
 $ existing_loans_count: int  1 1 2 2 1 1 2 2 2 1 ...
 $ job                 : chr  "unskilled" "skilled" "unskilled" "unskilled" ...
 $ dependants          : int  1 1 2 1 1 1 1 1 2 1 ...
 $ phone               : chr  "no" "yes" "no" "no" ...
 $ default             : chr  "yes" "yes" "no" "no" ...
```

The given data set contains character and numeric variables, for decided purpose class type of some variables has changed. "yes" have been considered as "1' and "NO" as "0"

DATA CLEANING

The OUTLIER:- An outlier can cause serious problems in statistical analyses. Given data set contains typing error in the purpose variable as **"car0"**, considering it as an outlier after its comparison with other purposes, we have replaced it with **"car".** interpretation of statistics derived from data sets that include outliers may be misleading

```
> data.frame(variable)
              variable
1                  car
2          renovations
3 furniture/appliances
4                 car0
5             business
6            education
```

MISSING VALUE:- missing values occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. **insurance data contain 1320 missing values in the column of the year at residence & we have replaced those values with the mean value of the column.** Missing data can be handled similarly as censored data. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, the analysis may be biased.

```
colSums(is.na(insurance))
          cust_id               acc_no     checking_balance months_loan_duration
                0                    0                    0                    0
   credit_history              purpose         amount..USD.      savings_balance
                0                    0                    0                    0
employment_duration    percent_of_income   years_at_residence                  age
                0                    0                 1320                    0
     other_credit              housing existing_loans_count                  job
                0                    0                    0                    0
       dependants                phone              default
                0                    0                    0
```

FACTORS:- we have converted variable **default, phone, existing loan count, dependent** into class factors. no. fo financial dependency on loan applier, no. of phones person owns, default status, exting number of loan cant be in integer format. their value would be either 1 or 0. To create a factor variable we use the as. factor function

```
> prop.table(prop)

  0   1
0.7 0.3
```

The proportion of insurance data is about 70%-30%, stating that there are 70 % chances that loan applier will not make default and 30% chances that will make default

LOGISTIC REGRESSION

To create a **parsimonious model,** we have utilised the existing data i.e. split data into training and testing data set, training data have been used for the variable selection process and for the model. The test set is used only at the conclusion of these activities for estimating a final, unbiased assessment of the model's performance.

cust id, account no. , year at residence, housing, job, dependent these variable shows insignificancy in the model, as their p-value is greater than 0.05, so we have removed those variable in the model_2. also in model_2
Null deviance: 5131.3 on 4199 degrees of freedom
Residual deviance: 3998.3 on 4170 degrees of freedom, as lower the value of residual better the model able to predict the value of response variable.

MODEL_1

```
> model_1 =glm(default~.,data = train,family = binomial)
> summary(model_1)

Call:
glm(formula = default ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9026  -0.7775  -0.4047   0.8099   2.7578

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.678e-01  4.030e-01  -1.657 0.097494 .
cust_id                       1.299e-05  2.547e-04   0.051 0.959322
acc_no                       -1.302e-10  2.547e-09  -0.051 0.959237
checking_balance> 200 USD    -9.082e-01  1.753e-01  -5.180 2.22e-07 ***
checking_balance1 - 200 USD  -3.760e-01  9.930e-02  -3.786 0.000153 ***
checking_balanceunknown      -1.838e+00  1.103e-01 -16.660  < 2e-16 ***
months_loan_duration          2.768e-02  4.283e-03   6.463 1.03e-10 ***
credit_historygood            8.173e-01  1.254e-01   6.517 7.18e-11 ***
credit_historyperfect         1.329e+00  2.087e-01   6.368 1.92e-10 ***
credit_historypoor            6.176e-01  1.570e-01   3.934 8.34e-05 ***
credit_historyvery good       1.393e+00  2.054e-01   6.779 1.21e-11 ***
purposecar                    1.723e-01  1.487e-01   1.159 0.246595
purposeeducation              6.276e-01  2.044e-01   3.070 0.002139 **
purposefurniture/appliances  -2.402e-01  1.458e-01  -1.647 0.099520 .
purposerenovations            3.583e-01  2.774e-01   1.292 0.196415
amount..USD.                  8.552e-05  1.978e-05   4.324 1.53e-05 ***
savings_balance> 1000 USD    -1.203e+00  2.402e-01  -5.011 5.43e-07 ***
savings_balance100 - 500 USD -2.964e-01  1.352e-01  -2.193 0.028339 *
savings_balance500 - 1000 USD -3.796e-01  1.946e-01  -1.950 0.051139 .
savings_balanceunknown       -9.438e-01  1.216e-01  -7.764 8.24e-15 ***
employment_duration> 7 years -2.894e-01  1.347e-01  -2.149 0.031664 *
employment_duration1 - 4 years -3.162e-01  1.120e-01  -2.823 0.004753 **
employment_duration4 - 7 years -9.528e-01  1.399e-01  -6.810 9.79e-12 ***
employment_durationunemployed  5.713e-03  1.943e-01   0.029 0.976543
percent_of_income             2.744e-01  4.027e-02   6.815 9.42e-12 ***
years_at_residence           -4.407e-02  4.383e-02  -1.006 0.314611
age                          -1.538e-02  4.308e-03  -3.569 0.000358 ***
other_creditnone             -5.715e-01  1.111e-01  -5.143 2.70e-07 ***
other_creditstore            -3.111e-01  2.013e-01  -1.545 0.122296
housingown                   -1.415e-01  1.409e-01  -1.004 0.315187
housingrent                   2.889e-01  1.628e-01   1.774 0.075986 .
existing_loans_count2         4.266e-01  1.132e-01   3.769 0.000164 ***
existing_loans_count3        -8.082e-03  2.990e-01  -0.027 0.978436
existing_loans_count4         3.695e-01  5.279e-01   0.700 0.483940
jobskilled                    4.177e-02  1.319e-01   0.317 0.751416
jobunemployed                -1.873e-01  2.986e-01  -0.627 0.530522
jobunskilled                 -1.745e-02  1.608e-01  -0.108 0.913603
dependants2                   1.235e-01  1.128e-01   1.095 0.273551
phone1                       -2.405e-01  9.509e-02  -2.529 0.011440 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5131.3  on 4199  degrees of freedom
Residual deviance: 3974.9  on 4161  degrees of freedom
AIC: 4052.9

Number of Fisher Scoring iterations: 11
```

MODEL_2

```
> summary(model_2)

Call:
glm(formula = default ~ . - cust_id - acc_no - years_at_residence -
    housing - job - dependants, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9307  -0.7900  -0.4088   0.8097   2.6658

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -8.565e-01  2.951e-01  -2.903 0.003698 **
checking_balance> 200 USD    -9.596e-01  1.734e-01  -5.535 3.11e-08 ***
checking_balance1 - 200 USD  -4.172e-01  9.805e-02  -4.255 2.09e-05 ***
checking_balanceunknown      -1.870e+00  1.095e-01 -17.075  < 2e-16 ***
months_loan_duration          2.712e-02  4.199e-03   6.458 1.06e-10 ***
credit_historygood            8.372e-01  1.249e-01   6.702 2.05e-11 ***
credit_historyperfect         1.387e+00  2.063e-01   6.720 1.82e-11 ***
credit_historypoor            6.372e-01  1.559e-01   4.088 4.35e-05 ***
credit_historyvery good       1.402e+00  2.016e-01   6.953 3.57e-12 ***
purposecar                    2.008e-01  1.462e-01   1.373 0.169644
purposeeducation              6.948e-01  2.009e-01   3.459 0.000541 ***
purposefurniture/appliances  -2.297e-01  1.439e-01  -1.596 0.110469
purposerenovations            3.002e-01  2.765e-01   1.086 0.277652
amount..USD.                  8.605e-05  1.911e-05   4.504 6.68e-06 ***
savings_balance> 1000 USD    -1.144e+00  2.356e-01  -4.853 1.21e-06 ***
savings_balance100 - 500 USD -2.520e-01  1.338e-01  -1.883 0.059737 .
savings_balance500 - 1000 USD -3.695e-01  1.931e-01  -1.914 0.055669 .
savings_balanceunknown       -9.297e-01  1.200e-01  -7.746 9.51e-15 ***
employment_duration> 7 years -3.143e-01  1.306e-01  -2.407 0.016079 *
employment_duration1 - 4 years -3.327e-01  1.108e-01  -3.003 0.002669 **
employment_duration4 - 7 years -9.575e-01  1.385e-01  -6.913 4.74e-12 ***
employment_durationunemployed -5.411e-02  1.753e-01  -0.309 0.757569
percent_of_income             2.660e-01  3.933e-02   6.763 1.36e-11 ***
age                          -1.697e-02  4.074e-03  -4.164 3.12e-05 ***
other_creditnone             -5.753e-01  1.101e-01  -5.227 1.72e-07 ***
other_creditstore            -3.920e-01  2.002e-01  -1.958 0.050261 .
existing_loans_count2         4.271e-01  1.123e-01   3.803 0.000143 ***
existing_loans_count3        -1.230e-02  2.924e-01  -0.042 0.966446
existing_loans_count4         5.355e-01  4.853e-01   1.103 0.269897
phone1                       -2.329e-01  8.776e-02  -2.654 0.007949 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5131.3  on 4199  degrees of freedom
Residual deviance: 3998.3  on 4170  degrees of freedom
AIC: 4058.3

Number of Fisher Scoring iterations: 5
```

ASSUMPTION OF LOGISTIC REGRESSION

```
> vif(model_2)
    checking_balance> 200 USD    checking_balance1 - 200 USD        checking_balanceunknown
                       1.1296                         1.4327                         1.3896
          months_loan_duration              credit_historygood          credit_historyperfect
                       1.7740                         2.5124                         1.2623
           credit_historypoor        credit_historyvery good                     purposecar
                       1.4362                         1.5254                         3.2314
              purposeeducation       purposefurniture/appliances          purposerenovations
                       1.6117                         3.3657                         1.2833
                  amount..USD.       savings_balance> 1000 USD    savings_balance100 - 500 USD
                       2.1428                         1.0361                         1.1410
  savings_balance500 - 1000 USD        savings_balanceunknown   employment_duration> 7 years
                       1.0721                         1.1154                         2.0114
employment_duration1 - 4 years employment_duration4 - 7 years employment_durationunemployed
                       1.8126                         1.5612                         1.3870
               percent_of_income                            age              other_creditnone
                       1.2575                         1.3332                         1.3708
             other_creditstore            existing_loans_count2          existing_loans_count3
                       1.2956                         1.7715                         1.1491
          existing_loans_count4                         phone1
                       1.0564                         1.2020
>
```

Assumption of logistic regression:- Logistic regression assumes that there is no severe multicollinearity among the explanatory variables. There are No Extreme Outliers. There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable. we have used a **Variance inflation factor** (VIF)to measure the amount of multicollinearity between variables. the decided threshold was 2 . Variable value above 2 consider a multicollinear variable.variable purpose and amount USD showing the multicollinearity as their value is greater than 2

MODEL_3

```
> summary(model_3)

Call:
glm(formula = default ~ . - cust_id - acc_no - years_at_residence -
    housing - job - purpose - amount..USD. - dependants, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8040  -0.7816  -0.4214   0.8630   2.5721

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.782862   0.254148  -3.080  0.00207 **
checking_balance> 200 USD     -1.038996   0.171011  -6.076 1.24e-09 ***
checking_balance1 - 200 USD   -0.403722   0.095708  -4.218 2.46e-05 ***
checking_balanceunknown       -1.856245   0.107888 -17.205  < 2e-16 ***
months_loan_duration           0.038423   0.003311  11.603  < 2e-16 ***
credit_historygood             0.787478   0.123038   6.400 1.55e-10 ***
credit_historyperfect          1.408751   0.200066   7.041 1.90e-12 ***
credit_historypoor             0.621929   0.153722   4.046 5.21e-05 ***
credit_historyvery good        1.433623   0.201728   7.107 1.19e-12 ***
savings_balance> 1000 USD     -1.095693   0.233819  -4.686 2.79e-06 ***
savings_balance100 - 500 USD  -0.205144   0.132343  -1.550  0.12112
savings_balance500 - 1000 USD -0.475380   0.192078  -2.475  0.01333 *
savings_balanceunknown        -0.841709   0.117850  -7.142 9.18e-13 ***
employment_duration> 7 years  -0.313551   0.129131  -2.428  0.01518 *
employment_duration1 - 4 years -0.319753  0.109681  -2.915  0.00355 **
employment_duration4 - 7 years -0.931655  0.136825  -6.809 9.82e-12 ***
employment_durationunemployed  0.057956   0.171979   0.337  0.73612
percent_of_income              0.187453   0.035702   5.250 1.52e-07 ***
age                           -0.012184   0.003966  -3.072  0.00212 **
other_creditnone              -0.575209   0.108995  -5.277 1.31e-07 ***
other_creditstore             -0.490365   0.199018  -2.464  0.01374 *
existing_loans_count2          0.441327   0.110776   3.984 6.78e-05 ***
existing_loans_count3          0.019800   0.284595   0.070  0.94453
existing_loans_count4          0.459434   0.478861   0.959  0.33734
phone1                        -0.087298   0.082764  -1.055  0.29153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5131.3  on 4199  degrees of freedom
Residual deviance: 4067.3  on 4175  degrees of freedom
AIC: 4117.3

Number of Fisher Scoring iterations: 5
```

In model 3 variable **phone** was showing insignificancy toward the model, and also IN existing loan count variable two out of three value having their p-value greater than 0.05 but still, I have kept that variable .Because unpaid dues are always a concern for lenders, repayment patterns.

FINAL MODEL

AIC

Akaike information criterion

AIC test penalizes models which use more independent variables (parameters) as a way to avoid over-fitting. AIC of the final model is less than the previous model. Lower indicates a more parsimonious model, relative to a model fit with a higher AIC.

```
> anova(model_4,test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: default

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                  4199     5131.3              
checking_balance      3   592.91      4196     4538.4 < 2.2e-16 ***
months_loan_duration  1   150.30      4195     4388.1 < 2.2e-16 ***
credit_history        4   108.69      4191     4279.4 < 2.2e-16 ***
savings_balance       4    78.12      4187     4201.2 4.361e-16 ***
employment_duration   4    50.59      4183     4150.7 2.722e-10 ***
percent_of_income     1    27.11      4182     4123.5 1.920e-07 ***
age                   1     9.23      4181     4114.3 0.0023841 ** 
other_credit          2    28.47      4179     4085.8 6.568e-07 ***
existing_loans_count  3    17.44      4176     4068.4 0.0005743 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

VARIABLES

default in repayment

defaults mostly depend on the amount held in current bank count, loan tenure, credit history, the amount held in saving account employment duration, income, age, other credit, existing loan count

```
> AIC(model_3)
[1] 4117.293
> AIC(model_4)
[1] 4116.407
```

but why these variables ???

- A healthy **credit score** can directly impact the rate of interest offered to the loan applier.
- **existing loan counts** are not a problem but these will help to identify the unpaid dues, payment pattern, missing EMI.
- every lender sets minimum **income criteria** that should breach.
- **Employment status** is very important to consider to check does the applicant has a stable job, steady flow of income.
- **Age** criteria need to get considered as lenders is concerned with how many years borrower have left as a salaried or working profession or eligibility for loan during the early year of their career.
- **other credits** like a store or any other property are mostly security-based in which borrowers get the loan so if the property worth is higher then the bank offers them a higher loan amount so it's an important parameter to decide the loan amount.
- **loan tenure** is ideally considered as it affects the monthly instalment.
- **current account** and **saving account** can have a mild effect, as some banks do an enquiry about healthy deposit and withdrawal history

STATISTICAL TESTING

Goodness-of-fit tests are statistical tests aiming to determine whether a set of observed values match those expected under the applicable model. we have used the following methods to check the goodness of fit of logistic regression:- likelihood ratio test, Hosmer-Lemeshow tests, Classification tables, ROC curves, pseudoR2, Somer'd test. for statistical testing, considering the null hypothesis as an unknown parameter is equal to 0.

likelihood ratio test shows that the p-value is less than 0.05, so we reject the null hypothesis, conclude that model is a good fit.

```
Model 2: default ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  24 -2034.2
2   1 -2565.6 -23 1062.9  < 2.2e-16 ***
```

```
> pR2(model_4)
fitting null model for pseudo-r2
          llh        llhNull           G2      McFadden         r2ML         r2CU
-2034.2035361 -2565.6300686  1062.8530651     0.2071330     0.2235789     0.3170074
```

McFadden's pseudo R2, values from 0.2-0.4 indicate excellent model fit. for this model, the McFadden value is 0.2071, which shows that model is a good fit.

```
> hoslem.test(train$default,fitted(model_4))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  train$default, fitted(model_4)
X-squared = 4200, df = 8, p-value < 2.2e-16
```

The Hosmer-Lemeshow test (HL test) is a goodness of fit test for logistic regression, especially for risk prediction models. $p<0.05$, we reject the null hypothesis, model is a good fit. selected variable add value to the model.

```
> library(InformationValue)
> somersD(train$default,fitted(model_4))
[1] 0.5969528
```

Somers D to compare the predictive performance of models. Higher values indicate better predictive performance. Somers d give the 59.69% value. which state concordant pair are high than discordant pair, so the model has a decent predictive ability of 59%

CONFUSION MATRIX

A confusion matrix presents how a classification model becomes confused while making predictions. A good matrix (model) will have large values across the diagonal and small values off the diagonal. Measuring a confusion matrix provides better insight in particulars of is our classification model is getting correct and what types of errors it is creating.

```
> confusionMatrix(prediction,train$default)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2352  481
         1  588  779

               Accuracy : 0.7455
                 95% CI : (0.732, 0.7586)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 3.667e-11

                  Kappa : 0.4083

 Mcnemar's Test P-Value : 0.001187

            Sensitivity : 0.8000
            Specificity : 0.6183
         Pos Pred Value : 0.8302
         Neg Pred Value : 0.5699
             Prevalence : 0.7000
         Detection Rate : 0.5600
   Detection Prevalence : 0.6745
      Balanced Accuracy : 0.7091

       'Positive' Class : 0
```

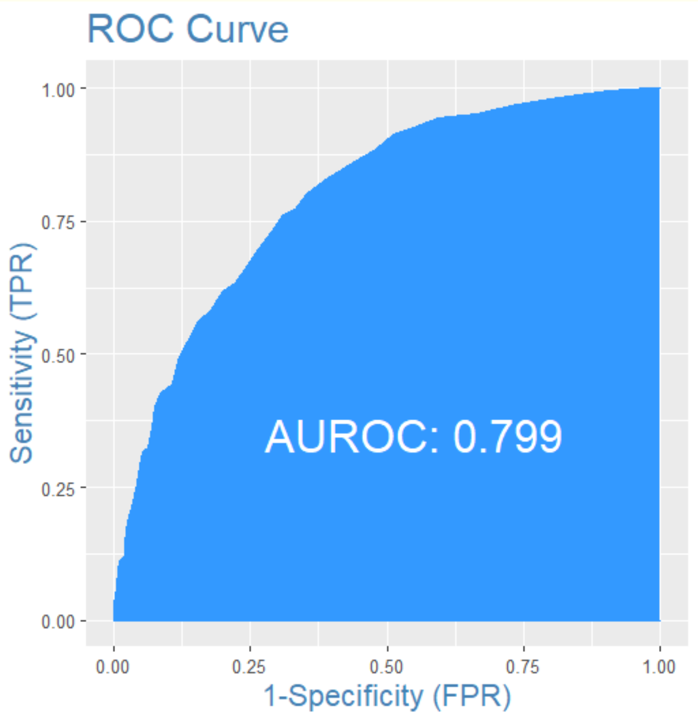
```
> confusionMatrix(pred,testing$default)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 990 191
         1 270 349

               Accuracy : 0.7439
                 95% CI : (0.7231, 0.7639)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 2.109e-05

                  Kappa : 0.4147

 Mcnemar's Test P-Value : 0.0002803

            Sensitivity : 0.7857
            Specificity : 0.6463
         Pos Pred Value : 0.8383
         Neg Pred Value : 0.5638
             Prevalence : 0.7000
         Detection Rate : 0.5500
   Detection Prevalence : 0.6561
      Balanced Accuracy : 0.7160

       'Positive' Class : 0
```

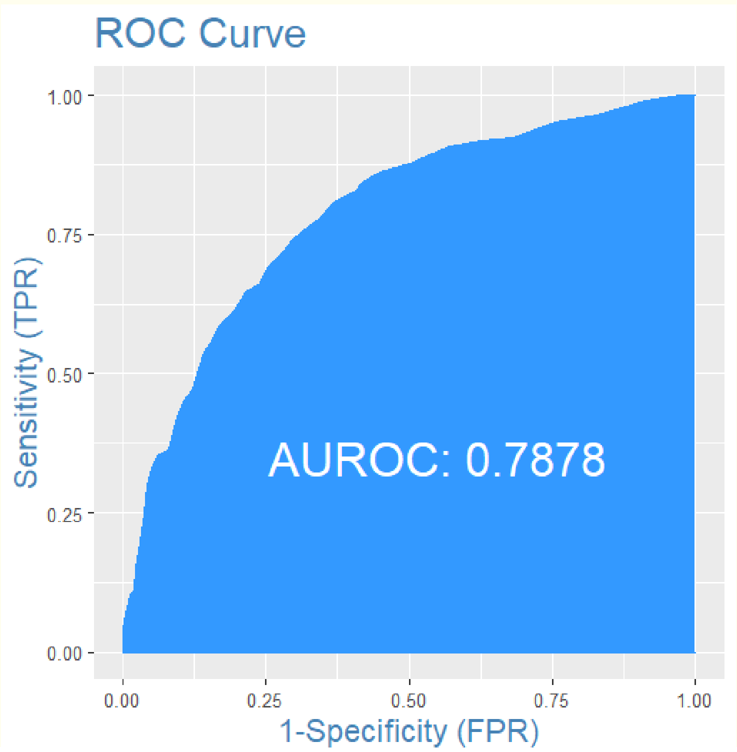
confusion matrix summarizes are the model's predictions. It gives us the number of correct predictions (True Positives and True Negatives) and the number of incorrect predictions. for confusion matric ideal threshold considered was **0.4.** with this threshold, the model gives a good sort accuracy and the highest number of true positives and negatives. **accuracy of the training and testing data set is 74%.**

ROC CURVE

A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate better performance.

TRAINING DATA SET



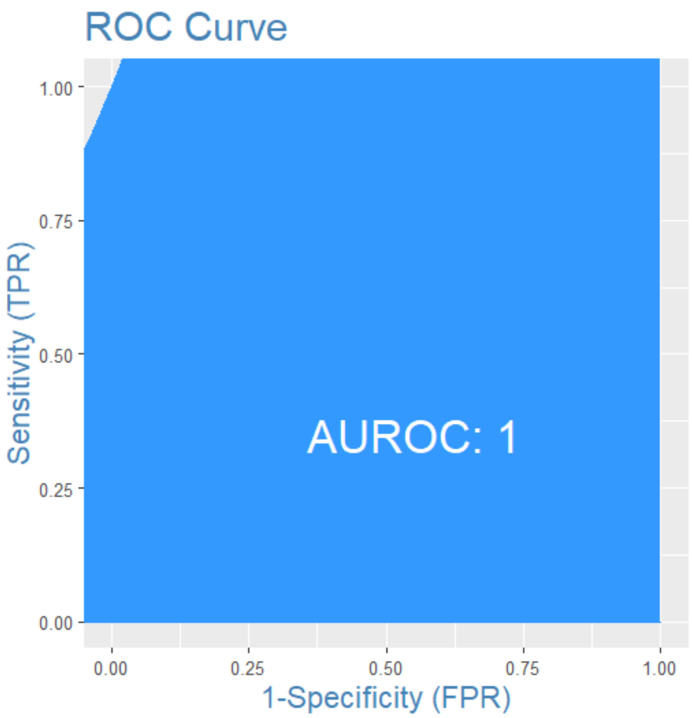


TESTING DATA SET

K-FOLD CROSS-VALIDATION

K FOLD CROSS VALIDATION gives better insight into data. in this, we have set cross-validation with 5 folds because most commonly it is set to 5 or 10. MODEL has an accuracy of 75%.

```
> confusionMatrix(as.factor(predict_class),testing$default)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1105  285
         1  155  255

               Accuracy : 0.7556
                 95% CI : (0.735, 0.7753)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 9.192e-08

                  Kappa : 0.375

 Mcnemar's Test P-Value : 7.756e-10

            Sensitivity : 0.8770
            Specificity : 0.4722
         Pos Pred Value : 0.7950
         Neg Pred Value : 0.6220
             Prevalence : 0.7000
         Detection Rate : 0.6139
   Detection Prevalence : 0.7722
      Balanced Accuracy : 0.6746

       'Positive' Class : 0
```

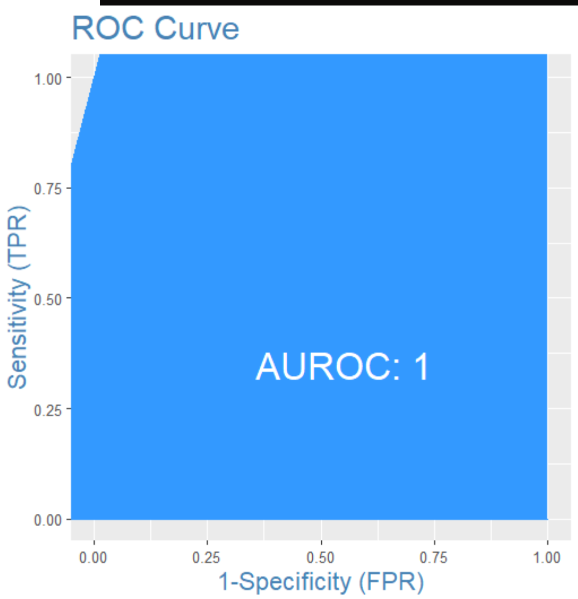


Receiver operating characteristic (ROC) metric to evaluate the quality of the output of a classifier using cross-validation.

DECISION TREE

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features problem of logistic regression being

```
> confusionMatrix(default,testing$default)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1124  274
         1  136  266

               Accuracy : 0.7722
                 95% CI : (0.7521, 0.7914)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 4.114e-12

                  Kappa : 0.415

 Mcnemar's Test P-Value : 1.324e-11

            Sensitivity : 0.8921
            Specificity : 0.4926
         Pos Pred Value : 0.8040
         Neg Pred Value : 0.6617
             Prevalence : 0.7000
         Detection Rate : 0.6244
   Detection Prevalence : 0.7767
      Balanced Accuracy : 0.6923

       'Positive' Class : 0
```



hard to interpret is much more serious than it first appears as compared to logistic regression decision tree gives better accuracy of 77 % .

APPLICATION

Credit evaluation is one of the most crucial processes in banks “credit management decisions”. Credit and risk analyst knowledge can help to identify the socioeconomic background of the applicant; requests are automatically processed through models of credit scoring assigning default probabilities based on some threshold that will be classified as “good” or “bad.” It will help the bank for the management of credit losses, for the evaluation of new loan programs, risk-based pricing that will lead to profit maximization. The bank can use this model to assess Individual customer scoring and enterprise scoring the risk of bankruptcy and insolvency can be examined with this model

THANK YOU

SAYALI DILIP CHINCHWALKAR
ROLL NO . 429