**Credit Risk Analysis**

Credit risk is primarily the loss that occurs due to failure of borrower or lender to abide by the terms and conditions of any financial contract i.e., the failure to make the required payments on loans due. Thus, in order to assess a variety of risks related to the borrower, the bank looks up on several factors related to creditor.

The borrower credit risk is evaluated by considering the factors like financial position of the borrower, its past performance on loan credibility, capital adequacy, credit score, debt-to-income ratio and collateral.

Reducing loan losses and ensuring that capital reserves appropriately reflect the risk profile is by implementing an integrated, quantitative credit risk solution. Thus, these solutions should include better model management that spans the entire modelling life cycle, real time scoring and limits monitoring, robust stress testing capabilities, data visualisation capabilities and business intelligence tools that get important information into the hands of those who need it and when they need it. Also, following up with the incidents and audit actions help overcome the issue of credit risk in the banking sector.

**Objectives**

This project resorts to looking up for the effect of variables like person income, age, home ownership on loan amount and default rate. In order to check the various cause and effect relationship, we have applied the knowledge of distribution theory and theories of statistics. We have used Excel and R programme for better visualisation and accuracy of models.

As credit risk is inherent in lending, various measures can be taken to minimize the risk.

Factors leading to credit risk:

* Credit concentration towards specific sector or individual leads to credit failure.
* Credit issuing process: It includes the flaws in the banks credit granting and monitoring processes. Thus, proper evaluation of credit worthiness of any borrower as credit history, capacity to repay, capital, loan conditions and collateral. In the absence of the above information, the credit worthiness of the borrower cannot be evaluated accurately. So, banks need to maintain sufficient caution while lending.
* Subjective decision making: Instances where credit is given on the basis of approval given by senior management leads to huge credit failures, Thus, there are cases when loans are granted to related parties with no credit evaluations being done and accordingly the risk of default also increases.
* Inadequate monitoring: Long term lending is always given against secured assets. However, the value of asset may deteriorate over the time. Thus, it is important to not only monitor the performance of the borrower but also the value of assets.
* Cyclical performance: All the industries go through a depression and a boom period. During the boom period the evaluations may result in good credit worthiness of the borrower. Also, the cyclical performance of the industry must be taken into the account in order to arrive at the results of credit evaluations more accurately.

Through our analysis we aim on checking the following in our dataset:

* If there is a significant difference between the mean loan amount given to people having different home ownership.
* If there is a significant difference between the loan amount for purpose like debt consolidation, education, home improvement, medical, personal and venture.
* If there is a significant difference between the means of loan amount given to people having different home ownership (Factor A) and the means of loan amount given to people taking loans for different intents (Factor B) when considered together (assuming no interaction)
* If there is an association between the purpose of loan and their home ownership.
* If there are chances of a loan default or not on the basis of logistic regression model and predict the possibility of default in future.

**Data**

Credit Risk Dataset is been taken from Kaggle data source. The data focuses on the information pertaining to the borrower and its business dealings along with collateral values. Information related to annual income, home ownership, employment length, loan intent, loan graded and loan amount is available in our data. (Link: https://www.kaggle.com/laotse/credit-risk-dataset). Below are the following variables in the dataset.

(The variables information is given as at the time of loan taken)

1. person\_age: Age of the borrower.

2. person\_income: Annual income of the borrower.

3. person\_home\_ownership: Home ownership type of the borrower.

4. person\_emp\_length: Amount of time in years the borrower is employed.

5. loan\_intent: Purpose of the loan.

6. loan\_grade: Classification system that involves assigning a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest.

7. loan\_amnt: Dimension of the loan taken.

8. loan\_int\_rate: Interest paid for the loan.

9. loan\_status: Dummy variable where 1 is default, 0 is not default.

10. loan\_percent\_income: Ratio between the loan taken and the annual income.

11. cb\_person\_default\_on\_file: Answers whether the person has defaulted before or not.

12. cb\_person\_cred\_hist\_length: Number of years of personal history since the first loan taken by that person.

**Exploratory Data Analysis:**

We have drawn some visual conclusions on our data sets by using the pivot table:

From the below chart, we observe that 20% of the total loan amount is been issued for Education purposes, 18% for Medical, 17% for Venture and another 17% for Personal, 16% for Debt consolidation and 12% for Home Improvement.

*Figure 1: Pivot Chart of Loan Amount and Loan Intent.*

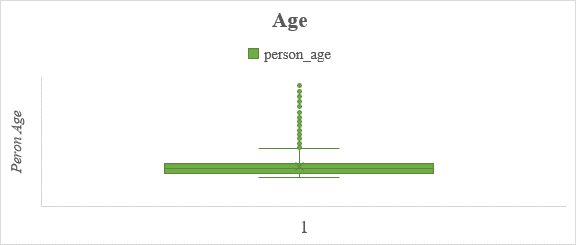
From the below chart, we observe that individuals with home ownership as mortgage and rent are issued the highest loan amount for various loan intents such as Debt Consolidation, Education, Home improvement, Medical, Personal and Venture.

*Figure 2: Pivot Chart of Loan Amount by Home Ownership and Loan Intent.*

From the below graph, we observe that the more of the loans are given for loan grade B. In addition, the highest loan amount is towards education followed by venture and medical purpose.

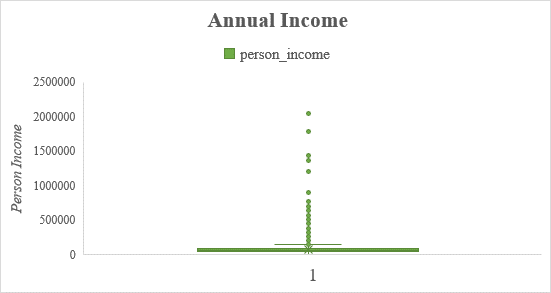
*Figure 3: Pivot chart of loan Amount by Loan Intent and Loan Grade.*

From the below boxplot, we observe that the distribution of age is right-skewed. This implies that most of the values lie on the right side of the scale. In other words, considering age, people belonging to the age group of 20-30 years populate the dataset.



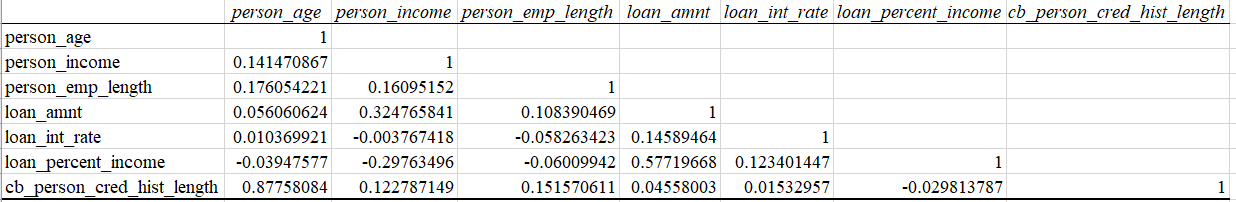
*Figure 4: Boxplot of Person Age*

From the below boxplot, we observe that the distribution of Income is right-skewed. This implies that most of the values lie on the right side of the scale, which tells us that the distribution of income is not balanced. Hence, we infer the presence of economic inequality in the data.



*Figure 5: Boxplot of Person Income*

Following are the correlation coefficients between the variables:



We observe that,

There exists a strong positive correlation between:

* Person age and credit history length with correlation coefficient equal to 0.87
* Person income and loan amount with correlation coefficient equal to 0.57

There exists a negative correlation between:

* Person employment length and loan interest rate with correlation coefficient equal to

-0.05

* Person income and loan interest rate with correlation coefficient equal to -0.003

The descriptive statistics obtained about the various variables are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| *person\_age* |  | *person\_income* |  |
|  |  |  |  |
| Mean | 27.71214026 | Mean | 66426.50559 |
| Standard Error | 0.036475335 | Standard Error | 304.636127 |
| Median | 26 | Median | 55900 |
| Mode | 23 | Mode | 60000 |
| Standard Deviation | 6.171988842 | Standard Deviation | 51547.45735 |
| Sample Variance | 38.09344627 | Sample Variance | 2657140359 |
| Kurtosis | 5.573872544 | Kurtosis | 202.6677262 |
| Skewness | 1.920264673 | Skewness | 9.049088933 |
| Range | 64 | Range | 2035784 |
| Minimum | 20 | Minimum | 4000 |
| Maximum | 84 | Maximum | 2039784 |
| Sum | 793454 | Sum | 1901923708 |
| Count | 28632 | Count | 28632 |
| Confidence Level(95.0%) | 0.071493365 | Confidence Level(95.0%) | 597.1010795 |

|  |  |  |  |
| --- | --- | --- | --- |
| *person\_emp\_length* |  | *loan\_amnt* |  |
|  |  |  |  |
| Mean | 4.780315731 | Mean | 9655.331447 |
| Standard Error | 0.023849761 | Standard Error | 37.39614307 |
| Median | 4 | Median | 8000 |
| Mode | 0 | Mode | 10000 |
| Standard Deviation | 4.035616407 | Standard Deviation | 6327.798706 |
| Sample Variance | 16.28619978 | Sample Variance | 40041036.46 |
| Kurtosis | 2.466914171 | Kurtosis | 1.347416683 |
| Skewness | 1.254613706 | Skewness | 1.17379551 |
| Range | 41 | Range | 34500 |
| Minimum | 0 | Minimum | 500 |
| Maximum | 41 | Maximum | 35000 |
| Sum | 136870 | Sum | 276451450 |
| Count | 28632 | Count | 28632 |
| Confidence Level(95.0%) | 0.046746649 | Confidence Level(95.0%) | 73.29819223 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *loan\_int\_rate* |  | | *loan\_percent\_income* | |  |
|  |  | |  | |  |
| Mean | 11.03970138 | | Mean | | 0.169489033 |
| Standard Error | 0.019085224 | | Standard Error | | 0.000628575 |
| Median | 10.99 | | Median | | 0.15 |
| Mode | 10.99 | | Mode | | 0.1 |
| Standard Deviation | 3.229409436 | | Standard Deviation | | 0.106361195 |
| Sample Variance | 10.4290853 | | Sample Variance | | 0.011312704 |
| Kurtosis | -0.661259386 | | Kurtosis | | 1.288448997 |
| Skewness | 0.204111828 | | Skewness | | 1.077014658 |
| Range | 17.8 | | Range | | 0.83 |
| Minimum | 5.42 | | Minimum | | 0 |
| Maximum | 23.22 | | Maximum | | 0.83 |
| Sum | 316088.73 | | Sum | | 4852.81 |
| Count | 28632 | | Count | | 28632 |
| Confidence Level(95.0%) | 0.037407934 | | Confidence Level(95.0%) | | 0.001232037 |
| *cb\_person\_cred\_hist\_length* | |  | |
|  | |  | |
| Mean | | 5.793552668 | |
| Standard Error | | 0.023858723 | |
| Median | | 4 | |
| Mode | | 3 | |
| Standard Deviation | | 4.03713288 | |
| Sample Variance | | 16.29844189 | |
| Kurtosis | | 3.72308196 | |
| Skewness | | 1.661837479 | |
| Range | | 28 | |
| Minimum | | 2 | |
| Maximum | | 30 | |
| Sum | | 165881 | |
| Count | | 28632 | |
| Confidence Level(95.0%) | | 0.046764215 | |

**Methodology:**

**ANOVA**

According to Prof. R. A. Fisher, Analysis of Variances (ANOVA) is the separation of total variance into two groups of causes, namely variance due to assignable causes (factors) and variance due to chance causes. ANOVA is an extension of t test. ANOVA is used to test the statistical difference between two or more means by drawing inferences about the means by analysing their variances.

**a) One Way ANOVA**

One Way ANOVA consists of a single factor (independent variable) with multiple levels or groups (generally more than two). Multiple observations are taken at each level. One Way ANOVA is used to test if there is a significant statistical difference between the means of different groups. In One Way ANOVA, we compare the variation within the levels with the variation across (between) levels.

**Assumptions of One Way ANOVA**

1) Population from which the observations are drawn must approximately follow Normal distribution.

2) The observations are drawn independently.

3) The variances of populations must be equal.

**Mathematical Model**

Consider that there are ‘k’ groups and ni observations are considered in the ith level

Let N be the total number of observations

Let yij be the jth observation from taken from the ith level. Therefore, the mathematical model is given by

yij = m + ai + ϵij ; i = 1,2,…,k; j = 1,2,…,ni

where m is the general mean effect, ai is the mean effect due to ith level and ϵij is the random error effect due to chance

**Hypothesis**

H0: There is no statistically significant difference between the means of k groups. a1=a2=…=ak

H1: There is a significant difference between the means of atleast 2 groups ai ≠ aj ; i ≠ j

**ANOVA Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Sum of Squares** | **DoF** | **Mean Square** | **Fcal** |
| **Factor** | SSF=j | k - 1 | *MSF* = *SSF*/ (k - 1) | *MSF/MSE* |
| **Residual** | SSE=∑∑(yij−)2 | N – k | *MSE* = *SSE*/ (N - k) |  |
| **Corr. Total** | SST=∑∑(yij−)2 | N - 1 |  |  |

**Decision Criteria**

If Fcal > Fk-1,n-k,α , we reject the null hypothesis at α% level of significance  
**Note:** One Way ANOVA only shows whether the means of different groups are not equal. It does not tell us the specific pair of groups whose means differ from each other. If we want to find out those groups, we can carry out pairwise t test with all possible pairs of groups.

**b) Two Way ANOVA (without interaction)**

It is an extension of One-Way ANOVA and Two Way ANOVA consists of 2 factors (independent variables) with ‘m’ and ‘n’ levels (groups) respectively. It is used to test if there is a significant difference in the means of different levels in each of the 2 factors when the 2 factors are considered together. Here, we are conducting this test without taking into consideration the interaction effect of the 2 factors.

**Assumptions of One-Way ANOVA**

1) Population from which the observations are drawn must approximately follow Normal distribution.

2) The observations are drawn independently.

3) The variances of populations must be equal.

4) The groups of each factor must have sample size.

**Mathematical Model**

Consider 2 factors F1 and F2 with ‘m’ and ‘n’ levels respectively.

Let N be the total number of observations

Let yij be theobservation from taken from the ith level of factor 1 and jth level of factor 2. Therefore, the mathematical model is given by

yij = m + ai + bj + ϵij ; i = 1,2,…,n; j = 1,2,…,m

where m is the general mean effect, ai is the mean effect due to ith level of factor 1, bj is the mean effect due to the jth level of factor 2 and ϵij is the random error effect due to chance

**Hypothesis**

H0A: There is no statistically significant difference between the means of m groups of factor 1 a1=a2=…=ak

H1A: There is a significant difference between the means of at least 2 groups of factor 1

ai ≠ aj ; i ≠ j

H0A: There is no statistically significant difference between the means of n groups of factor 2.

b1= b2 =…=bk

H1B: There is a significant difference between the means of a tleast 2 groups of factor 2 bi ≠ bj ; i ≠ j

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Sum of Squares** | **DoF** | **Mean Square** | **Fcal** |
| **Factor 1** | SSF1=j | m - 1 | *MSF1* = *SSF1*/ (m - 1) | *F1 = MSF1/MSE* |
| **Factor 2** | SSF2=i | n – 1 | *MSF2* = *SSF2*/ (n - 1) | *F2 = MSF2/MSE* |
| **Residual** | SSE=∑∑(yij−)2 | (m-1) (n-1) | *MSE* = *SSE*/ ((m-1) (n-1)) |  |
| **Total** | SST=∑∑(yij−)2 | mn - 1 |  |  |

**ANOVA Table**

**Decision Criteria**

If F1 > Fm-1,(m-1)(n-1),α , we reject the null hypothesis H0A at α% level of significance  
If F2 > Fn-1,(m-1)(n-1),α , we reject the null hypothesis H0B at α% level of significance

**Chi Square Test of Association/ Pearson’s Chi Square test of Association**

This test is also called the text of independence. This test is used to answer the question whether 2 categorical variables (attributes) have statistically significant relationship with each other. That is, this test can be used to test if the 2 categorical variables are independent of each other or not.

**Assumptions**

1) The observations must be frequency (count)

2) The categories of the attributes should be mutually exclusive

3) Independent groups should be taken for study.

4) The data should be of nominal or ordinal type

5) The observed frequency must exceed 5 for all the classes.

**Hypothesis**

H0: The two variables are independent

H1: The two variables are not independent

**r X c Contingency Table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Attribute B | B1 | B2 |  | Bj |  | Bc | Total |
| Attribute A |  |  |  |  |  |  |  |  |
| A1 |  | O11 | O12 | .. | O1j | .. | O1c | (A1) |
| A2 |  | .. | .. | .. | .. | .. | .. | .. |
| .. |  | .. | .. | .. | .. | .. | .. | .. |
| Ai |  | .. | .. | .. | Oij | .. | .. | (Ai) |
| .. |  | .. | .. | .. | .. | .. | .. | .. |
| Ar |  | .. | .. | .. | .. | .. | Orc | (Ar) |
| Total |  | (B1) | .. | .. | (Bj) | .. | (Bc) | N |

where Oij is the observed frequency of class ij

Expected frequency of class ij is given as

Eij = ((Ai) x (Bj)) / N

**Test Statistic**

**Decision Criteria**

If X2 > X2 ­(r-1)(c-1), α , we reject the null hypothesis at α% level of significance

**Logistic Regression**

It is a transformation of linear regression. It is used for a classification problem in which the dependent variable is a categorical variable of binary type (takes only 2 possible values). There is no restriction on the type of independent variables. Logistic regression is used to model the probability that an observation for the dependent value takes any one out of the two possible values.

It involves the following criteria:

* let p(x) be a linear function of x. Every increment of a component of x would add or subtract so much to the probability. The conceptual problem here is that p must be between 0 and 1, and linear functions are unbounded. Moreover, in many situations we empirically see “diminishing returns” — changing p by the same amount requires a bigger change in x when p is already large (or small) than when p is close to 1/2. Linear models can’t do this.
* The log p(x) be a linear function of x, so that changing an input variable multiplies the probability by a fixed amount. The problem is that logarithms are unbounded in only one direction, and linear functions are not.
* Finally, the easiest modification of log p which has an unbounded range is the logistic (or logit) transformation, log p /1−p. We can make this a linear function of x without fear of nonsensical results. (The results could still happen to be wrong, but they’re not guaranteed to be wrong.)

Model logistic regression model is that

log p(x)/1 − p(x)= β0 + x · β

Solving for p, this gives:

p(x; b, w) =eβ0+x·β/(1 + eβ0+x·β) = 1/(1 + e−(β0+x·β))

Notice that the over-all specification is a lot easier to grasp in terms of the transformed

probability that in terms of the untransformed probability.1

**Output and Analysis:**

We have applied the following concepts of distribution theory and other statistical techniques using Microsoft Excel and R programming:

1. **One Way ANOVA**

To check if the loan amount provided is the same for people having different home ownership.

**Hypothesis:**

H0:  There is no significant differences between the loan amount for home ownership i.e., mortgage, other, own and rent. (µ1=µ2=µ3=µ4)

H1: There is significant differences between the loan amount for atleast 1 pair of different home ownership. (µi≠µj); i ≠ j

**Anova: Single Factor**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
| Mortgage | 11799 | 125496550 | 10636.2022 | 46188156.6 |
| Other | 94 | 1046125 | 11128.9894 | 36226146.5 |
| Own | 2192 | 20014600 | 9130.74818 | 39454223.6 |
| Rent | 11799 | 102828425 | 8715.01187 | 33815938.5 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| Source of Variation | SS | df | MS | F | P-value | F crit |
| Between Groups | 22580969375 | 3 | 7526989792 | 188.447522 | 6.8116E-121 | 2.605252 |
| Within Groups | 1.0337E+12 | 25880 | 39942100.1 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 1.05628E+12 | 25883 |  |  |  |  |

**Decision Criteria:** Reject H0, if F critical is less than F calculated

**Conclusion:**

As, F critical value is less than F calculated value, so **we reject H0** and conclude that there is a significant difference between the loan amount of atleast 1 pair of home ownership types

1. **One Way ANOVA:** To check if loan amount provided is the same for people taking loan for different intents.

**Hypothesis:**

**H0**:  There is no significant differences between the loan amount for purposes as debt consolidation, education, home improvement, medical, personal and venture (µ1=µ2=µ3=µ4=µ5=µ6)

**H1**: There is significant differences between the loan amount for atleast 1 pair of loan intents. (µi≠µj); i ≠ j

**Anova: Single Factor**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
| DEBTCONSOLIDATION | 4565 | 44119925 | 9664.824754 | 40754064.11 |
| EDUCATION | 4565 | 41928700 | 9184.819277 | 36919266.1 |
| HOMEIMPROVEMENT | 3199 | 33316775 | 10414.7468 | 44370923.61 |
| MEDICAL | 4565 | 42110050 | 9224.545455 | 37713746.37 |
| PERSONAL | 4565 | 43691425 | 9570.958379 | 39514256.78 |
| VENTURE | 4565 | 44085575 | 9657.30011 | 39255710.97 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 3579124443 | 5 | 715824888.6 | 18.11650875 | 5.45629E-18 | 2.214442961 |
| Within Groups | 1.02803E+12 | 26018 | 39512297.79 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 1.03161E+12 | 26023 |  |  |  |  |

**Decision Criteria:** Reject H0, if F critical is less than F calculated

**Conclusion:**

As, F critical value is less than F calculated value, so **we reject H0** and conclude that there is a significant difference between the loan amount for atleast 2 purposes

1. **Two Way ANOVA**

To check if there is a significant difference between the means of loan amount given to people having different home ownership (Factor A) and the means of loan amount given to people taking loans for different intents (Factor B) when considered together (assuming no interaction)

**Hypothesis:**

H01: There is no significant differences between the loan amount for home ownership i.e., mortgage, other, own and rent. µ11=µ21=µ31=µ41

H11 There is a significant difference between the loan amount for atleast 1 pair of different home ownership. (µi1≠µj1); i ≠ j

H02: There is no significant differences between the loan amount for purposes as debt consolidation, education, home improvement, medical, personal and venture. µ12=µ22=µ32=µ42=µ52=µ62

H12: There is significant differences between the loan amount for atleast 1 pair of loan intent. (µi2≠µj2); i ≠ j

Anova: Two-Factor Without Replication

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *SUMMARY* | *Count* | *Sum* | *Average* | *Variance* |
| MORTGAGE | 6 | 125481550 | 20913591.67 | 6.26496E+12 |
| OTHER | 6 | 1046125 | 174354.1667 | 2756152604 |
| OWN | 6 | 20014600 | 3335766.667 | 3.00907E+12 |
| RENT | 6 | 129909175 | 21651529.17 | 1.89015E+13 |
|  |  |  |  |  |
| DEBTCONSOLIDATION | 4 | 44119925 | 11029981.25 | 1.52514E+14 |
| EDUCATION | 4 | 54282825 | 13570706.25 | 1.79509E+14 |
| HOMEIMPROVEMENT | 4 | 33281775 | 8320443.75 | 6.66204E+13 |
| MEDICAL | 4 | 49458575 | 12364643.75 | 1.54363E+14 |
| PERSONAL | 4 | 47099325 | 11774831.25 | 1.33093E+14 |
| VENTURE | 4 | 48209025 | 12052256.25 | 1.13058E+14 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Rows | 2.31956E+15 | 3 | 7.73186E+14 | 148.8530419 | 2.20517E-11 | 3.287382105 |
| Columns | 6.29773E+13 | 5 | 1.25955E+13 | 2.424867387 | 0.084220396 | 2.901294536 |
| Error | 7.79143E+13 | 15 | 5.19429E+12 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 2.46045E+15 | 23 |  |  |  |  |

**Decision Criteria:** Reject H0, if F critical is less than F calculated

**Conclusion:**

As, F critical value is lesser than F calculated value, so we reject H01 and conclude that there is a significant difference between the loan amount provided to atleast 1 pair of home ownership types

As, F critical value is greater than F calculated value, so we do not reject H02 and conclude that there is no significant difference between the loan amount for purposes as debt consolidation, education, home improvement, medical, personal and venture.

1. **Chi Square test for association of attributes**

To check if there is an association between the reasons why people take loan and person’s home ownership.

**H0:** There is no association between the reason people take loan and the type of home they own.

**H1:** There is association between the reason people take loan and the type of home they own.

**R Code:**

*library(readxl)*

*data<-read\_excel("C:\\Users\\user\\Downloads\\Creditrisk\_Excel project\_Team9.xlsx")*

Above code is to import the dataset.

*View(data)*

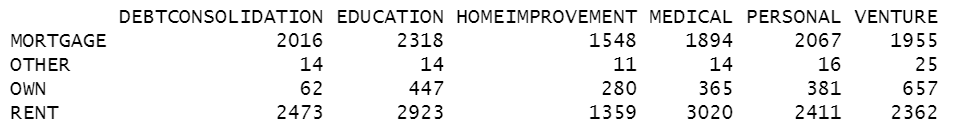
*tbl1 = table(data$person\_home\_ownership, data$loan\_intent)*

*tbl1*

Contingency table:

*install.packages("readxl")*

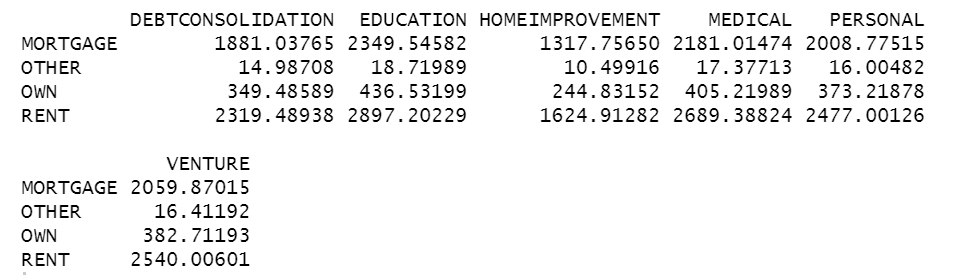
Output:



*chisq.test(tbl1)$expected*

Contingency table with expected frequencies:

Output:

 Test:

*chisq.test(tbl1)*

Output:

Pearson's Chi-squared test

data:  tbl1

X-squared = 652.87, df = 15, p-value < 2.2e-16

\

**Decision Criteria:** Reject H0 if p value < alpha (0.05)

**Conclusion:** The p value is **lesser than 2.2e-16**, which is less than 0.05. Hence, we reject H0

Therefore, there is an association between the reason why people take loan and the type of home they own.

1. **Logistic Regression**

**The dependent variable considered for the regression is loan\_status which shows whether a person has defaulted on the loan or not. Since, it is a binary variable (i.e. takes only 2 possible values), we decide to apply logistic regression.**

**Objective:** To identify the variables statistically affecting the risk of default (no repayment of loan) by a person and build a logistic regression model predict the possibility of risk of default for future loan applications  
Note: The dataset was cleaned and the missing observations were deleted as the total number of missing observations were not very large as compared to the total number of observations.

**Loading the dataset in R**

**R Code:**

CR <- read.csv(file.choose() , header = T)#dataset is stored in a variable names CR

**Converting the categorical variable to numerical variable for further analysis**

**R Code:**

CR$loan\_status=as.factor(CR$loan\_status)

CR$person\_home\_ownership=as.factor(CR$person\_home\_ownership)

CR$loan\_intent=as.factor(CR$loan\_intent)

CR$loan\_grade=as.factor(CR$loan\_grade)

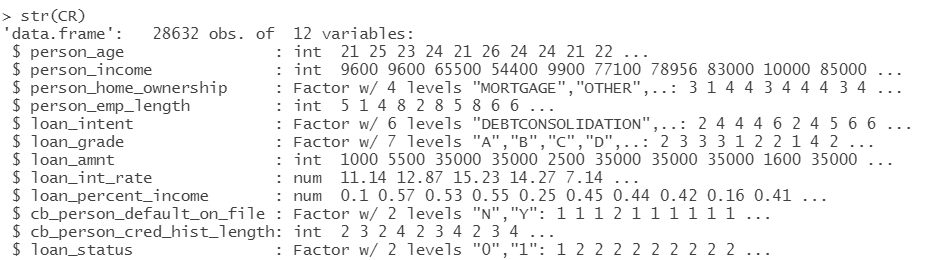
CR$cb\_person\_default\_on\_file=as.factor(CR$cb\_person\_default\_on\_file)

**Checking the structure of dataset**

**R Code:**

str(CR)

Output:



**Partitioning data into train data to build the model and test data to check the accuracy of the model**

**R Code:**

library(caret)

set.seed(123)

training.samples <- createDataPartition(CR$loan\_status, p = 0.7, list = FALSE)

train.data <- CR[training.samples, ]

test.data <- CR[-training.samples, ]

**Logistic Regression**

**Trial model with all the variables to check the statistical significance of each variable**

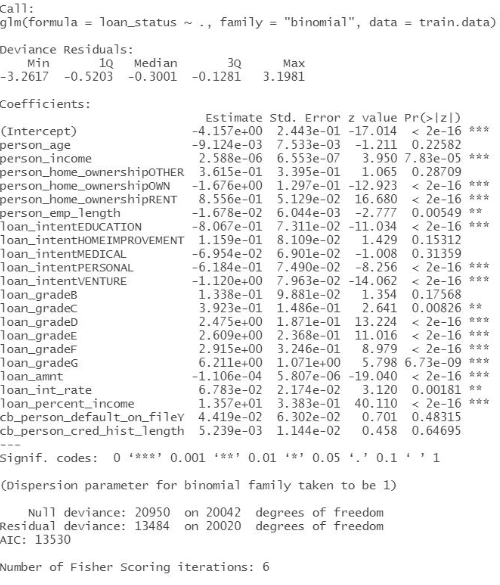
**R Code:**

library(ISLR)

logreg <- glm(loan\_status ~ ., data = train.data, family = "binomial")

summary(logreg)

Output:



**We see that the variables person\_age, cb\_person\_default\_on\_file, cb\_person\_cred\_hist\_length are statistically insignificant. Hence, we will not consider these variables in the next model. The variable loan\_percent\_income is the ratio of the variables, person\_income and loan\_amount. Hence, to avoid multicollinearity, we will exclude loan\_person\_income from our model.**

**New model with only the significant variables**

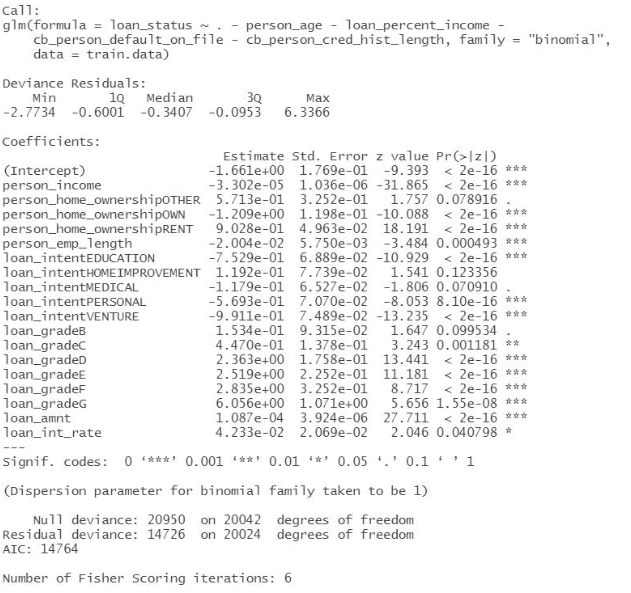
**R Code:**

logreg1 <- glm(loan\_status ~ . - person\_age - loan\_percent\_income - cb\_person\_default\_on\_file

- cb\_person\_cred\_hist\_length, data = train.data, family = "binomial")

summary(logreg1)

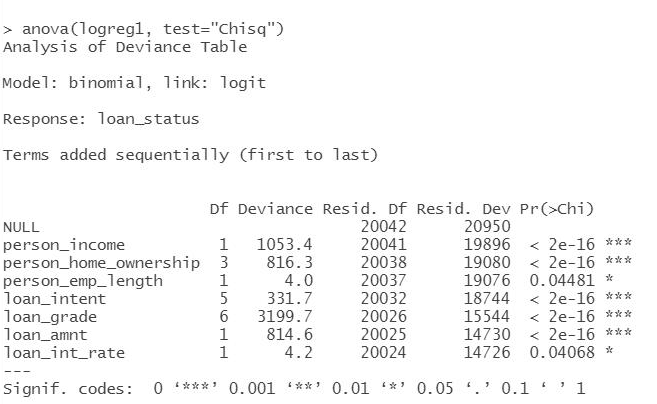
Output:



**ANOVA test to confirm the statistical significance of all the variables**

**R Code:**

anova(logreg1, test="Chisq")

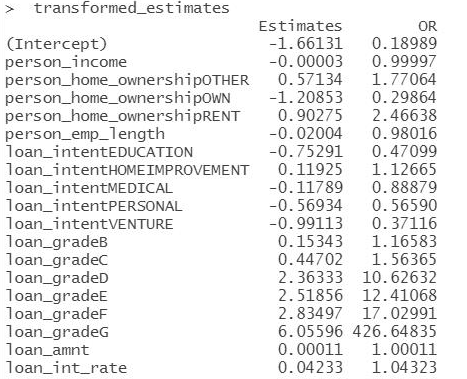


**Table to understand the estimates which are in the form of log odds and the odds ratio**

**R Codes:**

transformed\_estimates<-cbind(Estimates=round(coef(logreg1),5),OR=round(exp(coef(logreg1)),5))

Output:



The column of estimates is in the form of log odds. This means that when the interest rate increases by 1 unit, we can expect an increase of about 0.04233 in the log odds. It helps us to understand whether the effect of a predictor is positive or negative on the dependent variable. Log odds are simply the logarithmic value of the odds ratio and often difficult to interpret.

Hence, we created a column of odds ratio (OR) for the corresponding predictors. Odds ratio is the exponential value of the log odds. Odds ratio is defined as the probability of success divided by the probability of failure. Here, odds ratio can be interpreted as if the interest rate increases by 1 unit, the probability of default increases by 1.04323 or about 4.32%.

**Prediction on test data & Confusion Matrix to check model adequacy**

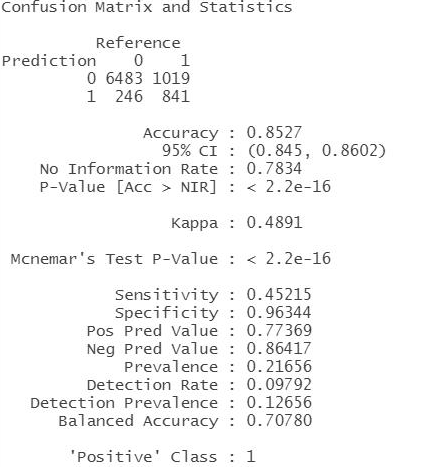
**R Codes:**

library(dplyr)

test.data$predicted.risk=predict(logreg1,newdata=test.data,type="response")

table(test.data$loan\_status, as.numeric(test.data$predicted.risk >= 0.5))

Output:



**ROC Curve**

**R Codes:**

library(ROCR)

pred = prediction(test.data$predicted.risk, test.data$loan\_status)

as.numeric(performance(pred, "auc")@y.values)

# Make predictions on training set

predictTrain = predict(logreg1, type="response")

# Prediction function

ROCRpred = prediction(predictTrain, train.data$loan\_status)

# Performance function

ROCRperf = performance(ROCRpred, "tpr", "fpr")

# Plot ROC curve

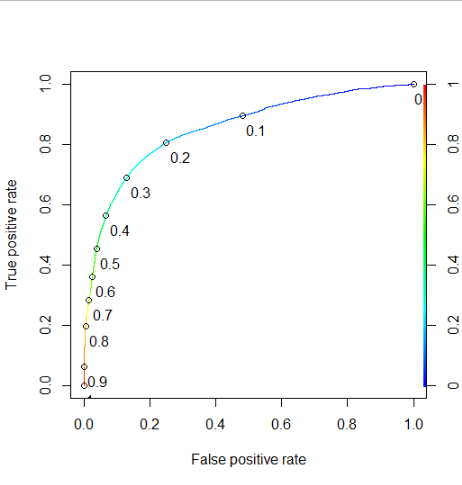
plot(ROCRperf)

# Add colors

plot(ROCRperf, colorize=TRUE)

# Add threshold labels

plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(1,0,by=-0.1), text.adj=c(-0.2,1.7))



Here, we see that if we consider the threshold probability as 0.35, the sensitivity improves drastically. To confirm this, we find the confusion matrix with the threshold probability as 0.35.

**R Codes:**

pred1 <- ifelse(predicted>=0.35,1,0)

pred1 <- as.factor(pred1)

testing=data.frame(test.data,pred1)

y\_act1 <- testing$loan\_status

mean(pred1 == y\_act)

caret::confusionMatrix(pred1, y\_act1, positive = "1")

**Output:**



As observed from the ROC curve, the sensitivity increases from 0.45 to 0.63 when we consider the threshold of 0.35. We also observe that there is no significant change in the overall accuracy.

**Conclusion:** We have identified the variables which, have a statistically significant effect on the dependent variable loan\_status and this model can help predict the possibility of default in future.

**Overall Conclusions**

The aim of our analysis was to understand the banking industry and the factors which lead to loan defaults.

Through our analysis we found that the loans are more granted to young people and the purpose of loan is usually education. It can be concluded that other purposes do not attract much of the loan amount. Hence there is a loan amount provided is significantly different for various purposes. Furthermore, loan amount provided is not same for people with different home ownership, this is somewhat intuitive. It should be noted that the credit policy of the bank seems to be sparing as it provides more loan to those who are living in the rented houses, even though they do not have assured collateral of property we notice that bank is willing to take the risk and provide high amount of loan.

As we performed logistic regression to predict if a loan defaults, we noticed that person home ownership, person employment length, loan intent, loan grade, loan amount, interest rate, person income are significant with respect to default status and age, default on file, credit history length, loan percent income are insignificant. It should be noted that loan grade, loan amount, interest rate have a positive effect on the dependent variable whereas home ownership, loan intent have a negative effect. Moreover, the sensitivity (true positive rate) changes drastically at threshold probability 0.35 i.e. we should consider that if the value of dependent variable is more than 0.35, the person is likely to default.

**Limitations:**

1. One of the techniques used in the analysis is Logistic Regression. However, we do not have complete understanding of the certain parts of this topic (like log odds, sensitivity, specificity etc.) due to which we might have not drawn conclusions completely.
2. We did not have complete information about the dataset. The time period when the data was collected is not known. Also, the data lacks information on currency, place and bank details.
3. As, we are not equipped with the techniques to work with outliers for a large dataset, we did not deal with outliers at this stage of our analysis.
4. Prediction and analysis could not be done precisely using Excel, so we have used R for drawing certain conclusions.

**Future Scope:**

* The model created using Logistic Regression has approximately 86% accuracy, which can be conveniently used to predict the loan status for other customers in future. Accordingly, these predictions can be used in order to reduce the number of defaulters.
* According to our data, we observed that most of the loans are provided to education sector. Thus, banks can decide upon their marketing strategies and lending rates on the kind of industry or the sector they want to target.
* Regular monitoring of the customer’s collateral value, business earnings, information on their family status, education level can lead to accurate predictions of the default status. Moreover, if the data is known for a specific bank, the columns such as loan grade and person’s credit history length will generate precise prediction of the defaulters.
* Banks help in providing capital funding to various sectors, which directly affects the growth of the economy. Therefore, if details about the country for which the data is taken into consideration is available, the model can be extrapolated to make inference about the economy of the country.

**REFERENCES:**

<https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-risk/>, <https://www.bankrate.com/glossary/c/credit-risks>, <https://www.wallstreetmojo.com/credit-risks-in-banks/>, http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf