Assignment-1

Drasti Shah

# Background

Home loans are a necessity these days and the to be eligible for one you need to satisfy a number of factors. This dataset outlines the various factors on which loan eligibility is dependent to automate the process. I will be using this dataset (source: [Loan eligibility dataset] (<https://www.kaggle.com/vikasukani/loan-eligible-dataset>)) to analyze what factors help predict the eligibility of an individual to get a loan. Firstly, we will check if gender plays any role in the loan eligibility process [1]. However, it is not the only parameter to be considered [1]. Secondly, [2]. We will check the impact of being married on loan eligibility. We then check the self - employed status in [3] because it is very important to guarantee the timely pay back of a loan. [4] puts light on the Applicant’s Income which is an important factor to considered to decide upon the loan amount to be sanctioned. Co - applicant Income is another factor to be considered since this is also a guarantor of timely and complete payment of loan amount [5]. The next parameter is the Loan amount for which the customer is eligible [6]. Consequently, the term after which the loan is due is also to be considered [7]. Credit History is one of the most important parameters to determine whether a person should be given the loan because it showcases the past entire record of the payments made by the consumer [7]. Property Area mentions the location of the house for which a particular loan amount is requested [8].Creating a model which predicts your eligibility for any given loan amount eases the process for both the customer and the bank. In Addition, it sets a criteria to become eligible for the loan too.

# Data Description

|  |  |  |
| --- | --- | --- |
| Attributes | Definition | Data\_Type |
| Loan\_ID | An identification number which helps in uniquely identifying to every loan | chr |
| Gender | Gender of every individual | Factor |
| Married | Marital status of the individual | Factor |
| Dependents | The number of dependents each individual has | chr |
| Education | Tells if individual is graduate or not | Factor |
| Self\_Employed | Gives information about the employment status | Factor |
| ApplicantIncome | Mentions the income of an applicant | int |
| CoapplicantIncome | Mentions the income of a co-applicant | num |
| LoanAmount | The amount of loan disbursed | int |
| Loan\_Amount\_Term | The term for which a loan amount is disbursed | int |
| Credit\_History | The credit history of the applicant | int |
| Property\_Area | The area in which the property is bought | Factor |
| Loan\_Status | Loan status showing whether the individual has recieved it or not | Factor |

The Loan eligibility dataset has 13 attributes. In this dataset, the Loan\_Status attribute is the dependent variable. The dataset has no missing values but has some incorrect data types. Categorical data with incorrect data types were also factorized using R’s factor function. The dependent variable Loan\_Status is a 1/2 value, which defines if an individual gets loan or not. The Loan\_ID states a unique ID to identify each individual’s loan. The Gender variable states whether the person is a male or female. The Married attribute provides information about the marital status of an individual. The dependents attribute mentions the number of dependents. The education attribute points out if the individual is graduate or not. The Self\_Employed shows the status of a person’s occupation. The Applicant Income tells you the income of an applicant. The Coapplicant Income mentions the income of any co-applicant. The Loan Amount tells you the total amount of loan which was applied for. The Loan\_Amount\_Term mentions the total days for which loan is sanctioned. The Credit\_History states whether the history is good or not. The Property\_Area mentions the location at which the property was bought (Urban, Semiurban, Rural). All of these variables will help in predicting the probability of the Loan\_Status of an individual.

# Method

# removing loan\_id column as it's irrelevant for the model  
loan\_data <- loan\_data[-1]

# splitting the dataset into training set and test set  
set.seed(123)  
split = sample.split(loan\_data$Loan\_Status, SplitRatio = 0.30)  
training\_set = subset(loan\_data, split == FALSE)  
test\_set = subset(loan\_data, split == TRUE)

# training the logistic regression model on the training\_set  
# . operator in the formula uses all the columns expect the is\_promoted to fit the logistic regression model  
classifier = glm(formula = Loan\_Status ~ .,  
 family = binomial,  
 data = training\_set)  
summary(classifier)

# training the logistic regression model on the training\_set  
# removed the gender & recruitment\_channel variable which turned out to be less significant for the model to predict value of is\_promoted.  
classifier\_2 = glm(formula = Loan\_Status ~ Gender + Married + Credit\_History + CoapplicantIncome+ Property\_Area + Loan\_Status,  
 family = binomial,  
 data = training\_set)  
summary(classifier\_2)

# comparing both the models  
anova(classifier, classifier\_2, test = 'LR')

# Creating the Confusion Matrix  
cm = table(test\_set[, 12], y\_pred)  
cm

# calculating the accuracy of the model  
n = sum(cm) # total number of instances in the confusion matrix  
diag = diag(cm) # number of correctly classified instances in the confusion matrix  
accuracy = sum(diag) / n   
accuracy\*100 # multiplying by 100 to get percentage value

At this stage, we began with data pre-processing. We removed the Loan\_id variable from the data frame, as it was irrelevant the model building. In the next step, we split the dataset into two parts: a training set and a test set. 70% of the data was stored in the training set, while the remaining 30% was stored in the test set. We chose the logistic regression model for creating a predictive model for the dataset. The model was built using the backward elimination approach; The base model was created by passing everything in the glm() function and, the results of the model were analyzed using the summary function. We found a few insignificant variables in the model (Dependents, Education, Self\_Employed, ApplicantIncome, LoanAmount, Loan\_Amount\_Term and Credit\_History). These variables were removed while creating the second model. However, when we used anova test to compare both the models, we got a p-value of 0.7865 for the second model, which is enough to prove that the base model was a better fit. So we went ahead by choosing the base model as the final model. We used the predict function to predict the probability for the test data. We also set the decision boundary as 0.5. So probability > 0.5 will be set to 1 else 0. We also created a confusion matrix, which will display a matrix of correct and incorrect (True positives, False positives, True negatives, and False negatives) predictions performed by the model. At last, we calculated the accuracy of the model.

# Result

* Model summary of the base model to know the significant variables

##   
## Call:  
## glm(formula = Loan\_Status ~ ., family = binomial, data = training\_set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4570 -0.3931 0.4750 0.6972 2.4319   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.951e+00 1.208e+00 -2.442 0.01462 \*   
## GenderMale 4.045e-01 3.993e-01 1.013 0.31105   
## MarriedYes 6.350e-01 3.475e-01 1.827 0.06765 .   
## Dependents1 -3.501e-01 4.150e-01 -0.844 0.39883   
## Dependents2 5.275e-01 4.658e-01 1.132 0.25744   
## Dependents3+ -2.883e-01 5.648e-01 -0.510 0.60972   
## EducationNot Graduate 4.502e-02 3.961e-01 0.114 0.90950   
## Self\_EmployedYes -2.155e-01 4.251e-01 -0.507 0.61218   
## ApplicantIncome -2.248e-05 4.468e-05 -0.503 0.61485   
## CoapplicantIncome -7.389e-05 5.138e-05 -1.438 0.15038   
## LoanAmount -5.881e-04 2.286e-03 -0.257 0.79701   
## Loan\_Amount\_Term -4.741e-04 2.678e-03 -0.177 0.85948   
## Credit\_History 3.832e+00 5.300e-01 7.231 4.78e-13 \*\*\*  
## Property\_AreaSemiurban 1.070e+00 3.915e-01 2.732 0.00629 \*\*   
## Property\_AreaUrban -2.020e-01 3.717e-01 -0.544 0.58673   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 415.78 on 335 degrees of freedom  
## Residual deviance: 296.10 on 321 degrees of freedom  
## AIC: 326.1  
##   
## Number of Fisher Scoring iterations: 5

* Model summary of the 2nd model

##   
## Call:  
## glm(formula = Loan\_Status ~ Gender + Married + Credit\_History +   
## CoapplicantIncome + Property\_Area + Loan\_Status, family = binomial,   
## data = training\_set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3019 -0.4048 0.4703 0.6934 2.4577   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.400e+00 6.526e-01 -5.210 1.89e-07 \*\*\*  
## GenderMale 4.299e-01 3.916e-01 1.098 0.27226   
## MarriedYes 6.672e-01 3.122e-01 2.137 0.03259 \*   
## Credit\_History 3.806e+00 5.218e-01 7.294 3.00e-13 \*\*\*  
## CoapplicantIncome -6.540e-05 4.706e-05 -1.389 0.16468   
## Property\_AreaSemiurban 1.072e+00 3.823e-01 2.805 0.00503 \*\*   
## Property\_AreaUrban -1.292e-01 3.490e-01 -0.370 0.71123   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 415.78 on 335 degrees of freedom  
## Residual deviance: 300.83 on 329 degrees of freedom  
## AIC: 314.83  
##   
## Number of Fisher Scoring iterations: 5

* Anova test results comparing the base model and the updated model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
| 321 | 296.1022 | NA | NA | NA |
| 329 | 300.8279 | -8 | -4.72565 | 0.7864561 |

* Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| N | 18 | 26 |
| Y | 3 | 97 |

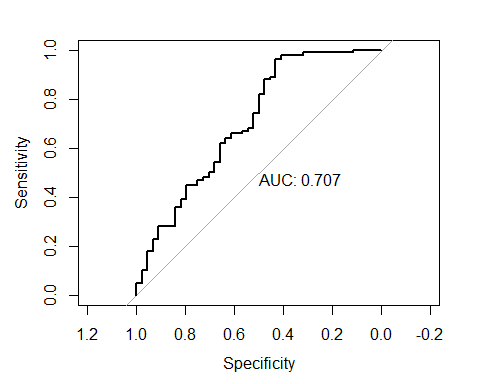
* Accuracy of the model

## [1] 79.86111

* Plotting the roc curve

## Setting levels: control = N, case = Y

## Setting direction: controls < cases



In the result section, we have provided the results of the method section. To start with, from the summary of the model, we can depict that Gender, Married, Credit\_History, CoapplicantIncome, Property\_Area are the significant variables impacting the eligibility of any loan. After that, the anova test result helped us in choosing the correct model. The confusion matrix table shows us the number of correct & incorrect predictions (True Positives = 97, True Negatives = 26, False Positives = 3, False Negatives = 18). We have also included the result of the accuracy of the model calculated in the method section. In the end, we have included a ROC curve. A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. We have got an AUC of 0.707. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. The value of AUC states that we have created a good model.

# Conclusion

In this study, we performed analysis on a real-world dataset. The dataset was based on Loan Eligibility data. It had a total of 13 attributes. We were supposed to find out the attributes which influenced the eligibility of an individual for a loan. We also made some hypotheses based on past researches and tried to examine those in our dataset. We started by exploring the dataset and managed to find some discrepancies in the data which we solved later. We also provided a table with a high-level data description, for the user to get an idea of the dataset. After this, we went on to the modeling part. We followed the conventional approach of splitting the dataset into training and test set. We implemented the logistic regression algorithm. The model was trained using the training data. We used the backward elimination approach and removed insignificant variables in the updated model. However, the anova test suggested that the base model was more efficient. Summary data of the model helped in concluding the hypothesis. Factors including Gender, Married, Credit\_History, CoapplicantIncome, Property\_Area turned out to be significant for getting a loan. Once the model was ready, we started predicting the value of Loan\_status for the test data. The probability with value > 0.5 were set to 1 else 0. The confusion matrix helped us in comparing the predicted values with the actual values; This way we can have an idea of how many correct/incorrect predictions were made by the model. We then visualized the ROC curve(receiver operating characteristic curve) which helped us in showing the performance of a classification model at all classification thresholds. We got an AUC of 0.707. The value of AUC suggested that we have created a good model.

# References

1. Holvoet, Nathalie. “Impact of microfinance programs on children’s education: Do the gender of the borrower and the delivery model matter?.” Journal of Microfinance/ESR Review 6.2 (2004): 3., <https://scholarsarchive.byu.edu/esr/vol6/iss2/3/>
2. Sooryamoorthy, R. “Microfinance and women in Kerala: Is marital status a determinant in savings and credit-use?.” Sociological bulletin 54.1 (2005): 59-77., <https://journals.sagepub.com/doi/abs/10.1177/0038022920050104?journalCode=soba>
3. Ghosh, Souvik. “„Housing Finance in India and Appraisal Process of Home Loans with Specific reference to Indian Overseas Bank‟.” International Journal of Science and Research 3.8 (2014): 129-135., <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.4065&rep=rep1&type=pdf>
4. Al-qerem, Ahmad, Ghazi Al-Naymat, and Mays Alhasan. “Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection.” 2019 International Arab Conference on Information Technology (ACIT). IEEE, 2019., <https://ieeexplore.ieee.org/abstract/document/8991084>
5. Kumar, Ashwani, Raman Dugyala, and Pronaya Bhattacharya. “Prediction of Loan Scoring Strategies Using Deep Learning Algorithm for Banking System.” Innovations in Information and Communication Technologies (IICT-2020). Springer, Cham, 2021. 115-121., <https://link.springer.com/chapter/10.1007/978-3-030-66218-9_13>