--- title: "Loan Prediction Project" author: "Aditi Teriar" date: "10/31/2020" output: pdf_document ---
```{r setup, include=FALSE}

# Reading into data frames and loading required packages

library(dplyr)

library(ggplot2)

library(rpart)

setwd("/home/ankit/loan_pred/")

train <- read.csv("train_u6lujuX_CVtuZ9i.csv",na.strings = c("","NaN"," "))

test <- read.csv("test_Y3wMUE5_7gLdaTN.csv",na.strings = c("","NaN"," "))

test$Loan_Status <- as.factor("NA")

#Combining training and test set

df.loan <- rbind(train[,2:13],test[,2:13])

#Missing values Summary

Variable <- colnames(df.loan)

NA_count <- sapply(df.loan, function(x) sum(is.na(x)))

miss_summ <- data.frame(Variable,NA_count,row.names = NULL)

miss_summ %>%

 arrange(desc(NA_count))

#Treatment of missing values

df.loan$Self_Employed[is.na(df.loan$Self_Employed)] = as.factor("No")

#Treatment of missing values in Loan Amount Term

df.loan$Loan_Amount_Term[is.na(df.loan$Loan_Amount_Term)] = 360

df.loan %>%

 group_by(Education,Self_Employed) %>%

 summarise(GroupMedian = mean(LoanAmount,na.rm = TRUE))

#imputing missing loan amount using sub categories

ind <- which(is.na(df.loan$LoanAmount))
```

```
df.loan[ind,]$LoanAmount[df.loan[ind,]$Education == "Graduate" & df.loan[ind,]$Self_Employed == "No"] <- 145.82

df.loan[ind,]$LoanAmount[df.loan[ind,]$Education == "Graduate" & df.loan[ind,]$Self_Employed == "Yes"] <- 174.24

df.loan[ind,]$LoanAmount[df.loan[ind,]$Education == "Not Graduate" & df.loan[ind,]$Self_Employed == "No"] <- 116.7

df.loan[ind,]$LoanAmount[df.loan[ind,]$Education == "Not Graduate" & df.loan[ind,]$Self_Employed == "Yes"] <- 131.56

#Credit History is a high impact variable

df.loan$Credit_History = as.character(df.loan$Credit_History)

df.loan$Credit_History[is.na(df.loan$Credit_History)] = "Not Available"

df.loan$Credit_History = as.factor(df.loan$Credit_History)

#Married Missing Values

df.loan$Married[is.na(df.loan$Married)] = as.factor("Yes")

#Gender Missing Values

df.loan$Gender[is.na(df.loan$Gender)] = as.factor("Male")

#Dependents Missing Values

df.loan$Dependents[is.na(df.loan$Dependents)] = as.factor("0")

cat("There are total", sum(is.na(df.loan)), "missing values in the dataset")

#Feature Engineering

df.loan$TotalIncome <- log(df.loan$ApplicantIncome + df.loan$CoapplicantIncome)

df.loan$TotalIncomeLoanRatio = log(((df.loan$ApplicantIncome + df.loan$CoapplicantIncome)/df.loan$LoanAmount)*(as.numeric(df.loan$Loan_Amount_Term)/360))

df.loan$LoanAmount <- log(df.loan$LoanAmount)

df.loan <- df.loan[,!(names(df.loan)) %in% c("ApplicantIncome","CoapplicantIncome")]

#Applying Logistic Regression Model

train_up<- df.loan[1:614,]

test <- df.loan[615:981,]

model <- glm(train_up$Loan_Status~.,family = binomial(link = 'logit'),data = train_up, maxit = 100)

summary(model)
```

#Fitting the Model

fitted_results <- predict(model, newdata=test, type="response")

fitted_results <- ifelse(fitted_results > 0.5,"Y","N")

test_up <- read.csv("test_Y3wMUE5_7gLdaTN.csv", stringsAsFactors = TRUE)

submit <- data.frame(Loan_ID = test_up$Loan_ID, Loan_Status = fitted_results)

write.csv(submit,"/home/ankit/loan_pred/1405_sub_1.csv",row.names = FALSE)

knitr::opts_chunk$set(echo = TRUE) ```