LENDING CLUB LOAN DEFAULTER PREDICTION

COURSE: BUAN 6356.003

Submitted to: Prof. Shujing Sun

Created by: Group 11
Aditi Sable, Bhargava Reddy Gade,
Devansh Chopra, Priya Bhasin,
Rohan Sharma, Venkata Sai Susmitha

Business Context:

For this project, we will be exploring publicly available data from Lendingclub.com. The lending club connects people who need money (borrowers) with people who have money (investors). An investor would want to invest in people who showed a profile of having a high probability of paying back the loan.

Problem Statement:

If the customer meets the credit underwriting criteria of LendingClub.com, loan is approved. Approving the loan does not guarantee that the lender will pay it back. To solve this problem, we will be using lending data from 2007 to 2010 and try to classify and predict whether the borrower paid back their loan in full or not.

Once the model is ready, Lending club can use it to predict whether the customer will pay back the loan or not based on historical trends used to train the model and then decide to approve the loan. Thus, lending club will not solely depend on its lending criteria and will have strong additional evidence to make decisions.

Data Source:

We will use lending data from 2007 to 2010 and try to classify and predict whether the borrower paid back their loan in full.

Entire 1 GB dataset from 2007 - 2018:
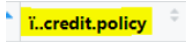https://www.kaggle.com/datasets/wordsforthewise/lending-club

Challenge and solution: We will be using subset of dataset as 1 GB dataset is difficult to work with causing long wait time to load and make even small changes on our system: https://cdn-stage.fedweb.org/fed-2/13/LoanStats3a.csv
Industry standard system can be used to deal with huge data to decrease processing times.

Data Preparation:

Better cleaned and prepared is the data, better will be the results. Below are the steps and challenges faced during data preparation (We did few steps in the .csv file itself):

- Removing columns containing a lot of Null/NA values. Luckily, our dataset doesn't contain Null/NA values for features we need to train our model.
- Removing features not required.
- Analyzing dataset carefully to see which features can be created as factor variables.
- Used fileEncoding="UTF-8-BOM" while importing data to eliminate special characters embedded with column name during import as shown: ï..credit.policy
- Removing extra characters and symbols like "$" to make data numerical.
- Making categorical features a factors.

Here are what the columns represent:

- credit. policy: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
- Purpose: The purpose of the loan (takes value,"credit_card","debt_consoliation",
- " educational","major_purchase","small_business", and "all_other").
- int.rate: The interest rate of the loan, as a proportion (a rate of 11%would be stored as 0.11). Borrowers judged by LendingClub.com to be riskier are assigned higher interest rates.
- installment: The monthly installments owed by the borrower if the loan is funded.
- log.annual.inc: The natural log of the self-reported annual income of the borrower.
- dti: The debt-to-income ratio of the borrower (Amount of debt divided by annual income).
- Fico: The FICO credit score of the borrower.
- days.with.cr.line: The number of days the borrower has had a credit line.
- revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
- revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
- Inq.last.6mnths: The borrower's number of inquiries by creditors in the last 6 months.
- delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
- Pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

Exploratory Data Analysis (EDA):
It is very important to go thoroughly and anlyse our data before fitting a model.

a. Checking correlation among feature (We will use only 1 out of 2 highly correlated feature to reduce redundancy of information when training our model):

```
> cor(loan[, c("int.rate", "log.annual.inc", "dti", "fico", "days.with.cr.line")])
                    int.rate log.annual.inc         dti        fico days.with.cr.line
int.rate          1.00000000     0.03900259  0.22000563 -0.71482077       -0.12402216
log.annual.inc    0.03900259     1.00000000 -0.10970543  0.08619377        0.22097830
dti               0.22000563    -0.10970543  1.00000000 -0.24119099        0.06010112
fico             -0.71482077     0.08619377 -0.24119099  1.00000000        0.26387975
days.with.cr.line -0.12402216     0.22097830  0.06010112  0.26387975        1.00000000
```
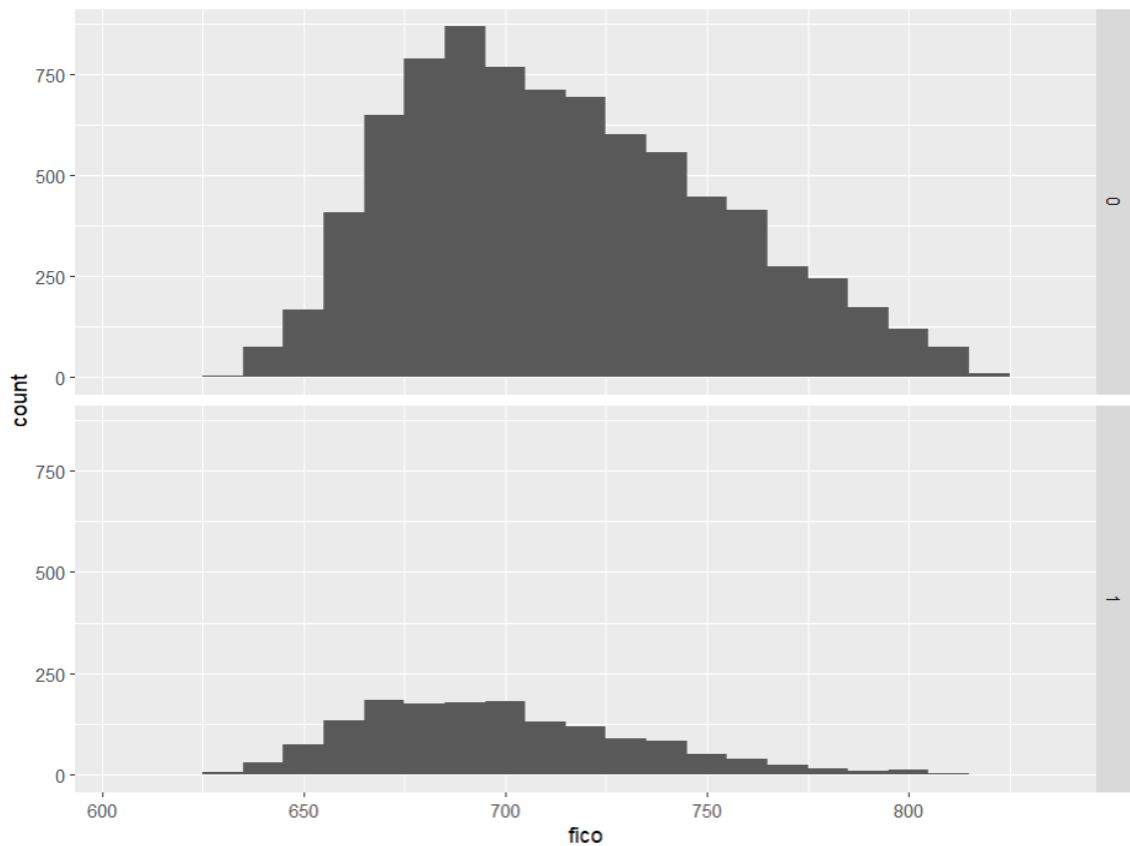
b. Majority of people made loan repayment. (0 stands to loan repaid)
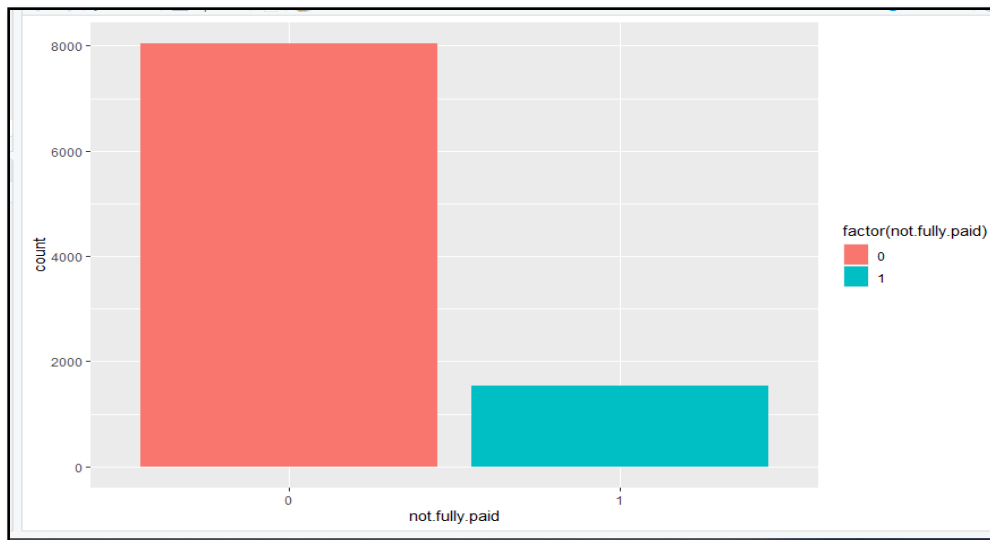
c. Distribution of Fico score: data is normally distributed.
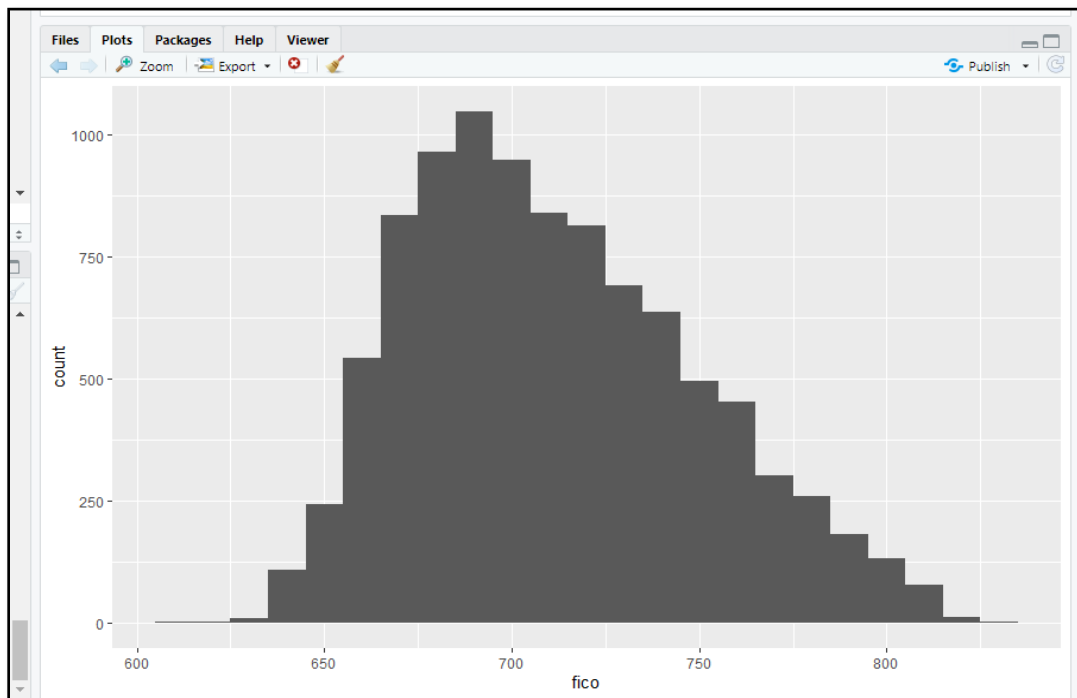
d. Fico score versus Loan Repayment (0 stands for loan repaid):
People with low credit (<700) or fico score are mostly the ones who did not pay back the loan.

e. Fico versus Credit policy:
It is very clear that people above fico 655+ only qualify for a loan from lending club. Moreover, it can also be seen that a score above 655 will not guarantee you a loan. There can be multiple factors deciding this (Example: person is blacklisted by govt. agencies). In exceptional circumstances loan may be given below fico score 655.
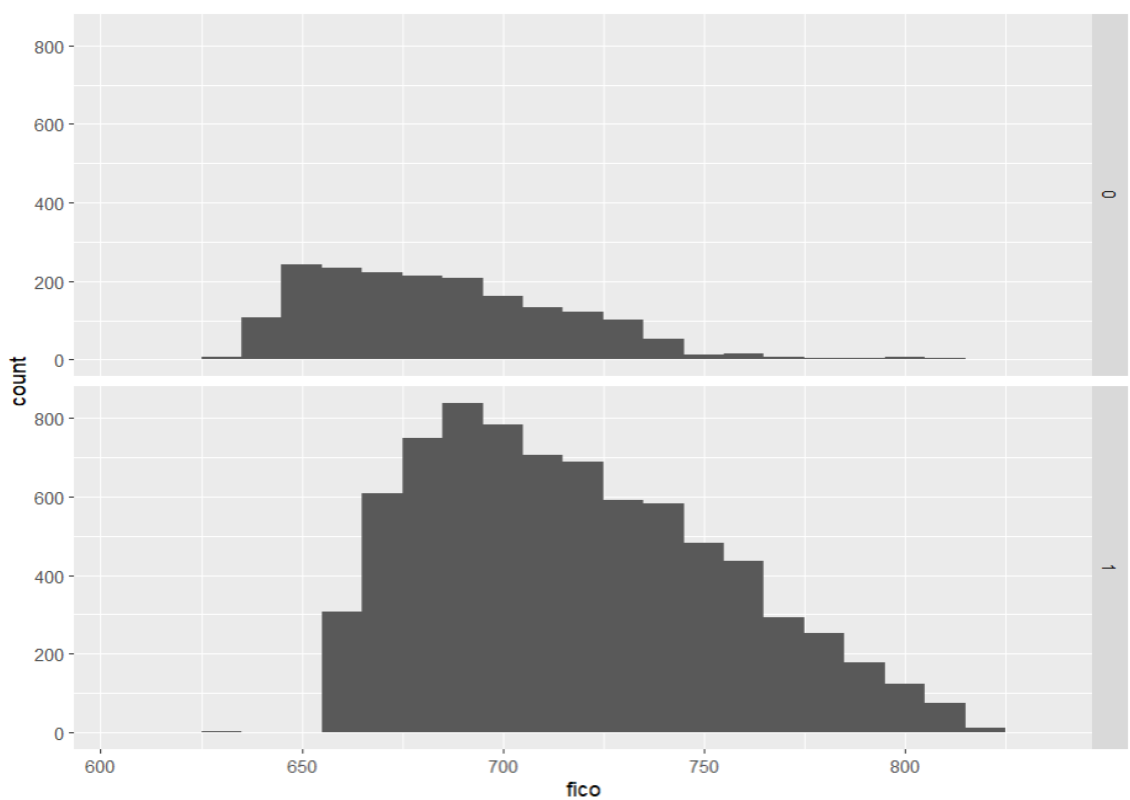
f. Interest Rate Distribution:
    Normal Distribution curve of interest rate and seems centered around 13% , right skewed plot.
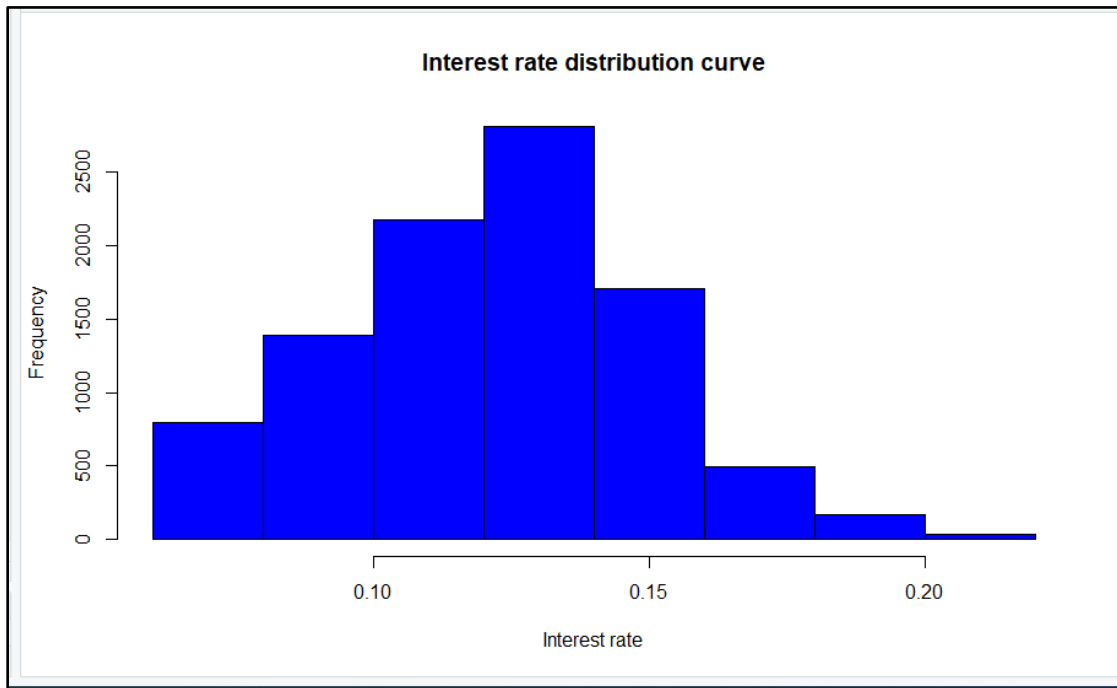
Model Preparation:

We will run two machine learning models for loan repayment prediction (decision tree and logistic regression model). After comparing their performance, we will select the best between the two for making predictions.

a. **Decision Tree:** Decision trees are popular machine learning models used for classification and regression. As the name suggests, the model returns output in a tree-like structure with a series of nodes and leaves. 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. Each node represents a test on input, and each leave is a class assignment. Decision trees are so popular because they can handle missing values and outliers and provide easily understandable rules.

b. **Logistic Regression:** A logistic regression model is a powerful machine learning model that extends the idea of linear regression to predict the binary outcome based on historical data.

R Code Used:

a. Decision Tree:

```r
# now we will split the data into testing and training data sets
# we will first randomly select 2/3 of the rows
set.seed(123) # for reproducible results
train <- sample(1:nrow(loan_data), nrow(loan_data)*(2/3)) # replace=FALSE by default

# Use the train index set to split the dataset
#  churn.train for building the model, churn.test for testing the model
loan.train <- loan_data[train,]    # 6385 rows
loan.test <- loan_data[-train,]    # the other 3193 rows

# Classification Tree with rpart
# install.packages('rpart')
library(rpart)
fit <- rpart(not.fully.paid ~ ., # formula, all predictors will be considered in splitting
             data=loan.train, # dataframe used
             #method="anova",
             method="class",
             control=rpart.control(xval=10, minsplit=100, cp=0), # xval: num of cross validatio
             parms=list(split="gini")) # criterial for splitting: gini default, entropy if set

library(rpart.plot)
rpart.plot(fit, type = 1, extra = 1, main="Classification Tree for Loan Repayment Prediction")

loan.pred <- predict(fit,loan.test, type="class")
loan.actual <- loan.test$not.fully.paid
length(loan.pred)
length(loan.actual)
table(loan.pred, loan.actual)

##### Model Evaluation #####
#Confusion matrix. Please use caret packageset.seed(123)
confusionMatrix(loan.pred, loan.actual)
#Accuracy is  83.4%
```
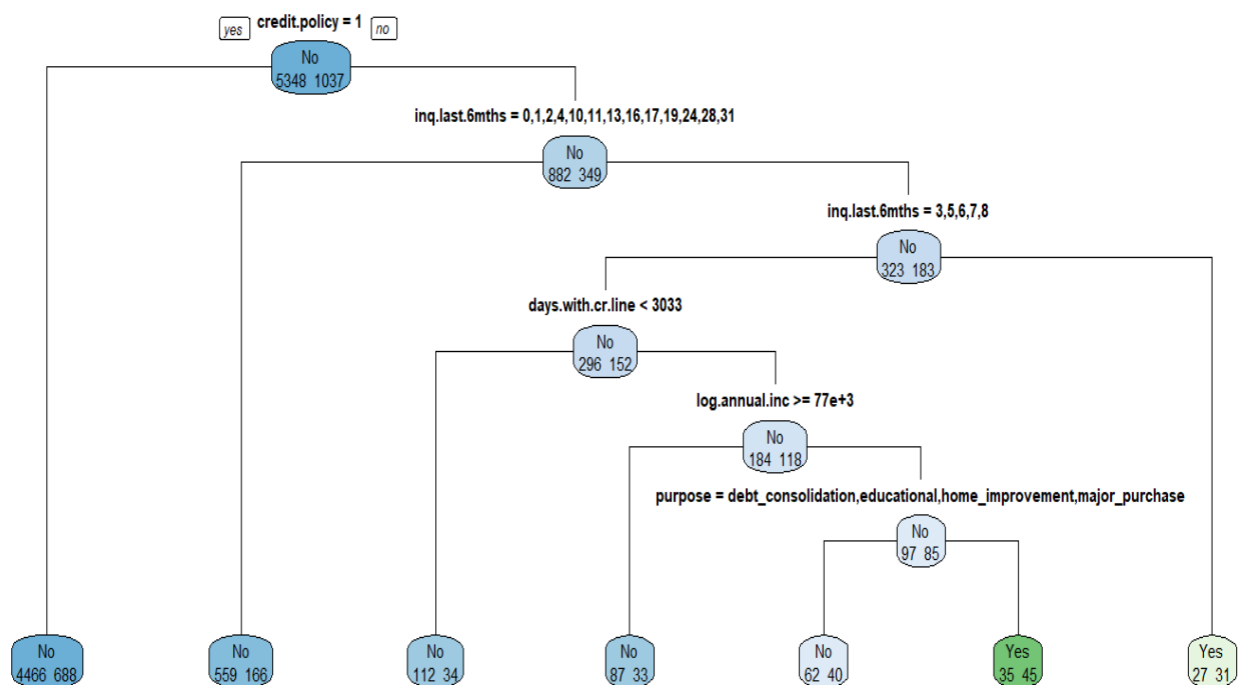
Classification Tree for Loan Repayment Prediction

b. Logistic Regression:

```
##### Splitting Data set #####
set.seed(123) # for reproducible results
train <- sample(1:nrow(loan), nrow(loan)*(2/3)) # replace=FALSE by default

# Use the train index set to split the dataset
#  train_data for building the model
#  test_data for testing the model
train_data  <- loan[train,]    # 6385 rows
test_data   <- loan[-train,]   # the other 3193 rows

##### Building logistic regression model #####
#using glm() function or "Generalized Linear Model" to perform logistic regression
#first we are using all the features  to train model and later use our knowledge to use specific features

log.reg <- glm(not.fully.paid ~ ., data = train_data, family = "binomial")
summary(log.reg)

#Tuning our model. This is where domain knowledge comes into existence.
#Clearly Credit.policy is not of much use here
#Interest rate is highly correlated to fico so eliminating it also. This is what we highlighted in our initial steps du
log.reg.rev <- glm(not.fully.paid ~ purpose + installment + log.annual.inc + fico + revol.bal +  pub.rec, data = train_
summary(log.reg.rev)

#making Predictions using test data. The feature to be predicted is at index 14. Note: Index starts at 1 and not 0 as i
predict_loan <- predict(log.reg.rev,test_data[-14], type = "response")
#Notice that we get continuous values and not just 0 or 1
head(predict_loan, n = 30)

#as we want the dependent variable (not.fully.paid) take values binary values: 0 or 1.
#we can  make it to 0 or 1 values by assigning any values below 0.5 as 0, and above 0.5 as 1 (this can be anything else
binary_predict <- as.factor(ifelse(predict_loan > 0.5, 1, 0)) # Fine tuned but 0.5 is giving highest accuracy
head(binary_predict, n = 30)
```

Models Comparison and Selection:

We will go ahead with choosing **Logistic Regression** over Decision Tree as it has better accuracy and AUC (Higher the AUC, the better the model performance). Also, based on other parameters, logistic regression dominates over decision tree

Model	Accuracy	AUC
Logistic Regression	85 %	0.65
Decision Tree	83 %	0.41

ROC is a probability curve and AUC (Area under the curve) represents the degree or measure of separability. Clearly Logistic Regression should be adopted based on below results.

Logistic Regression ROC

Decision Tree ROC
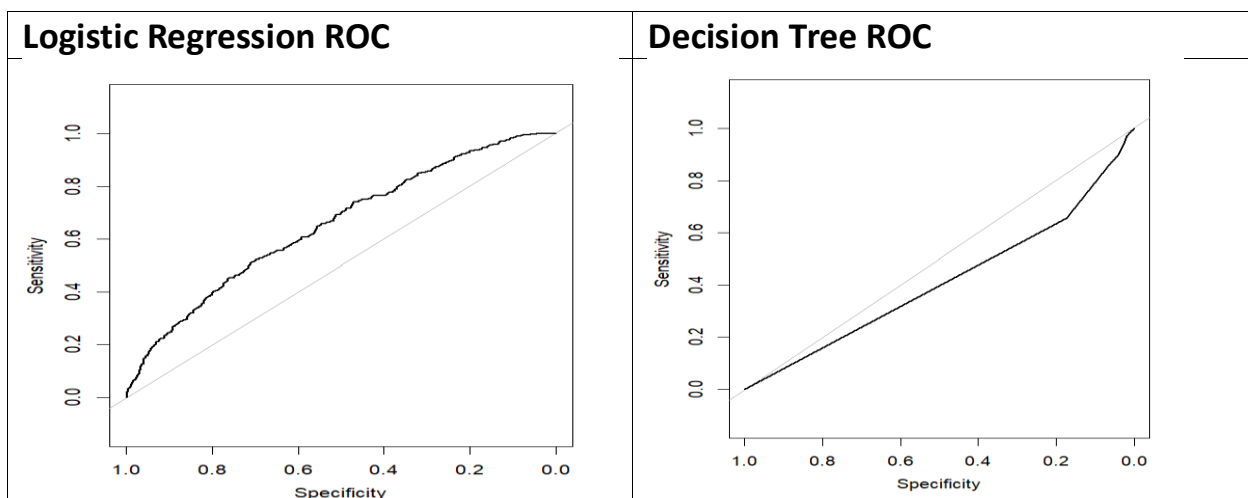
Logistic Regression Model Interpretation:

a. People who take loan to pay for credit card are less likely to replay back. These people are taking a loan to payback their bill which states there is very high probability they will not pay back the loan. Similar trend can e observed for debt consolidation.

b. Loan taken for educational purpose has high chances of being repaid. Investing in human capital (education) is the best investment.

c. Small businesses should be given loan as the repayment chances are highest (1.93 from below screenshot).

d. It is being noticeable that borrowers with derogatory public records category "pub.rec1" are likely to pay back. This category can be one which may be low in impact as compared to others and may be given less weight when approving loan.

```
> round(data.frame(summary(log.reg.rev)$coefficients, oddSR= exp(coef(log.reg.rev)) ),2)
                         Estimate Std..Error z.value Pr...z..    oddSR
(Intercept)                  7.75       0.74   10.49     0.00  2319.18
purposecredit_card          -0.60       0.13   -4.59     0.00     0.55
purposedebt_consolidation   -0.36       0.09   -3.93     0.00     0.70
purposeeducational           0.15       0.18    0.83     0.41     1.16
purposehome_improvement      0.01       0.16    0.09     0.93     1.01
purposemajor_purchase       -0.21       0.19   -1.11     0.27     0.81
purposesmall_business        0.66       0.14    4.81     0.00     1.93
installment                  0.00       0.00    4.29     0.00     1.00
log.annual.inc               0.00       0.00   -2.19     0.03     1.00
fico                        -0.01       0.00  -12.72     0.00     0.99
revol.bal                    0.00       0.00    2.72     0.01     1.00
pub.rec1                     0.51       0.13    3.89     0.00     1.66
pub.rec2                    -0.84       1.06   -0.80     0.43     0.43
pub.rec3                   -11.91     303.33   -0.04     0.97     0.00
pub.rec4                   -12.40     535.41   -0.02     0.98     0.00
pub.rec5                   -12.06     535.41   -0.02     0.98     0.00
```

Conclusion:

Based on our analysis we can say that people with high fico scores and salaries will pay back the loan and mostly these people fit into the lending club's criteria for approving the loan. At 84.53% accuracy, a simple model like logistic regression gives powerful results as compared to Decision Tree. We can make the model better by giving more data and applying multiple binary ML algorithms to get the best-fit algorithm. Lending Club should consider all the factors we discussed in model interpretation and during the analysis to approve a loan along with their present credit underwriting criteria. Our model can enhance the whole process and help company generate more revenue by making data driven decisions and targeting customers who have high probability of paying back the loan.