CREDIT DEFAULT RISK ASSESSMENT

NATE TALAMPAS

WHAT IS CREDIT DEFAULT RISK?

- Credit default risk is the risk a lender takes that a borrower will not make the required payments on a debt obligation.
- Earlier credit and risk management analysis would be conducted by analyzing the borrower's credentials and capabilities, which was more prone to error.
- Machine learning algorithms are more efficient in performing credit risk assessments with better precision and at faster speeds.

RESEARCH QUESTIONS

What variables are most significant in predicting credit default risk?

How do different machine learning algorithms perform in predicting credit risk?

DETERMINING SIGNIFICANT PREDICTORS

- To determine which predictors are significant, we perform a logistic regression. All regression coefficients with a p-value less than 0.05 will be statistically significant.
- Significant predictors include loan status, annual income, home ownership, employment length, loan intent, loan grade, loan amount, interest rate, and percent income.

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -4.121e+00  1.991e-01 -20.698  < 2e-16 ***
person_age                  -1.282e-02  6.239e-03  -2.055  0.03992 *  
person_income                9.197e-07  3.179e-07   2.893  0.00382 ** 
person_home_ownershipOTHER   4.285e-01  3.011e-01   1.423  0.15477    
person_home_ownershipOWN    -1.790e+00  1.128e-01 -15.865  < 2e-16 ***
person_home_ownershipRENT    8.282e-01  4.291e-02  19.300  < 2e-16 ***
person_emp_length           -1.414e-02  5.004e-03  -2.826  0.00472 ** 
loan_intentEDUCATION        -8.728e-01  6.106e-02 -14.295  < 2e-16 ***
loan_intentHOMEIMPROVEMENT   5.032e-02  6.779e-02   0.742  0.45792    
loan_intentMEDICAL          -1.555e-01  5.769e-02  -2.696  0.00702 ** 
loan_intentPERSONAL         -6.398e-01  6.240e-02 -10.253  < 2e-16 ***
loan_intentVENTURE          -1.140e+00  6.653e-02 -17.140  < 2e-16 ***
loan_gradeB                  1.095e-01  8.277e-02   1.322  0.18604    
loan_gradeC                  2.405e-01  1.246e-01   1.931  0.05347 .  
loan_gradeD                  2.332e+00  1.565e-01  14.903  < 2e-16 ***
loan_gradeE                  2.500e+00  1.974e-01  12.663  < 2e-16 ***
loan_gradeF                  2.798e+00  2.729e-01  10.256  < 2e-16 ***
loan_gradeG                  6.342e+00  1.054e+00   6.014 1.81e-09 ***
loan_amnt                   -1.011e-04  4.276e-06 -23.633  < 2e-16 ***
loan_int_rate                8.816e-02  1.812e-02   4.865 1.15e-06 ***
loan_percent_income          1.316e+01  2.511e-01  52.396  < 2e-16 ***
cb_person_default_on_fileY   2.152e-02  5.312e-02   0.405  0.68534    
cb_person_cred_hist_length   1.140e-02  9.449e-03   1.207  0.22750    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SUMMARY OF THE DATA SET

- The potential predictors to predict the outcome of whether a person will default include age, annual income, home ownership, employment length (in years), loan intent, loan grade, loan amount, interest rate, percent income, historical default, and credit history length.
- The dataset contains 32,581 observations and 12 variables.

SUMMARY STATISTICS OF DATASET

```
   person_age     person_income      person_home_ownership person_emp_length loan_intent         loan_grade          loan_amnt      loan_int_rate  
 Min.   :20.00   Min.   :   4000   Length:28632          Min.   : 0.00     Length:28632       Length:28632       Min.   :  500   Min.   : 5.42  
 1st Qu.:23.00   1st Qu.:  39456   Class :character      1st Qu.: 2.00     Class :character   Class :character   1st Qu.: 5000   1st Qu.: 7.90  
 Median :26.00   Median :  55900   Mode  :character      Median : 4.00     Mode  :character   Mode  :character   Median : 8000   Median :10.99  
 Mean   :27.71   Mean   :  66427                         Mean   : 4.78                                           Mean   : 9655   Mean   :11.04  
 3rd Qu.:30.00   3rd Qu.:  80000                         3rd Qu.: 7.00                                           3rd Qu.:12500   3rd Qu.:13.48  
 Max.   :84.00   Max.   :2039784                         Max.   :41.00                                           Max.   :35000   Max.   :23.22  
  loan_status      loan_percent_income cb_person_default_on_file cb_person_cred_hist_length
 Min.   :0.0000   Min.   :0.0000      Length:28632              Min.   : 2.000            
 1st Qu.:0.0000   1st Qu.:0.0900      Class :character          1st Qu.: 3.000            
 Median :0.0000   Median :0.1500      Mode  :character          Median : 4.000            
 Mean   :0.2166   Mean   :0.1695                                Mean   : 5.794            
 3rd Qu.:0.0000   3rd Qu.:0.2300                                3rd Qu.: 8.000            
 Max.   :1.0000   Max.   :0.8300                                Max.   :30.000            
```

HOW WE BUILD MODELS

- Determine if data is imbalanced
- Convert character values into factor data type
- Split data into training and test sets

```r
# converting categorical values into factor values
factor_names = c("person_home_ownership", "loan_intent", "loan_grade")
df1 = df1 |>
mutate_at(factor_names, factor)

# checking if data is balanced
table(df1$loan_status)/nrow(df1)

# The data is unbalanced. 78.34% of individuals did not default.
```

```r
# splitting into training set and test set
set.seed(123)
n = nrow(df1)
prop = 0.5
train_id = sample(1:n, size = round(n*prop), replace = FALSE)
test_id = (1:n)[-which(1:n %in% train_id)]

train_set = df1[train_id, ]
test_set = df1[test_id, ]
```

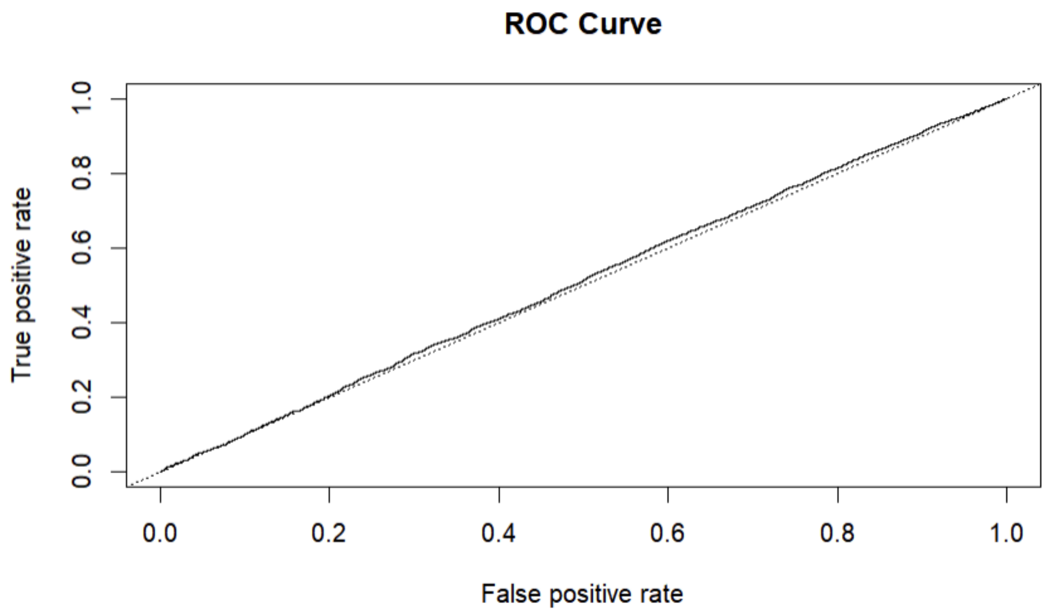
LOGISTIC REGRESSION

- We fit a logistic regression.

```
# fitting logistic model
logi_reg = glm(loan_status ~ ., family = "binomial", data = train_set)
summary(logi_reg)

# creating confusion matrix
logi_pred = ifelse(predict(logi_reg, data = test_set, type = "response") >
0.5, 1, 0)
tb_log = table(predict_status = logi_pred,
                true_status = test_set$loan_status)
tb_log

# computing accuracy
logi_acc = ((9453 + 500) / (9453 + 2586 + 1779 + 500))*100
cat("Accuracy:", logi_acc)

# ROC curve; computing AUC
library(ROCR)
logi_pred = predict(logi_reg, data = test_set)
pred = prediction(logi_pred, test_set$loan_status)
perf = performance(pred, "tpr", "fpr")
plot(perf, main = "ROC Curve")
abline(0, 1, lty=3)

auc = as.numeric(performance(pred, "auc")@y.values)
cat("\nAUC:", auc)
```

LOGISTIC REGRESSION

```
              true_status
predict_status    0    1
             0 9492 2512
             1 1766  546
Accuracy: 69.5139
```




```
AUC: 0.5030266
```

Accuracy is 0.6951

AUC value is 0.5030

LINEAR DISCRIMINANT ANALYSIS

- We fit a LDA model using the training set.

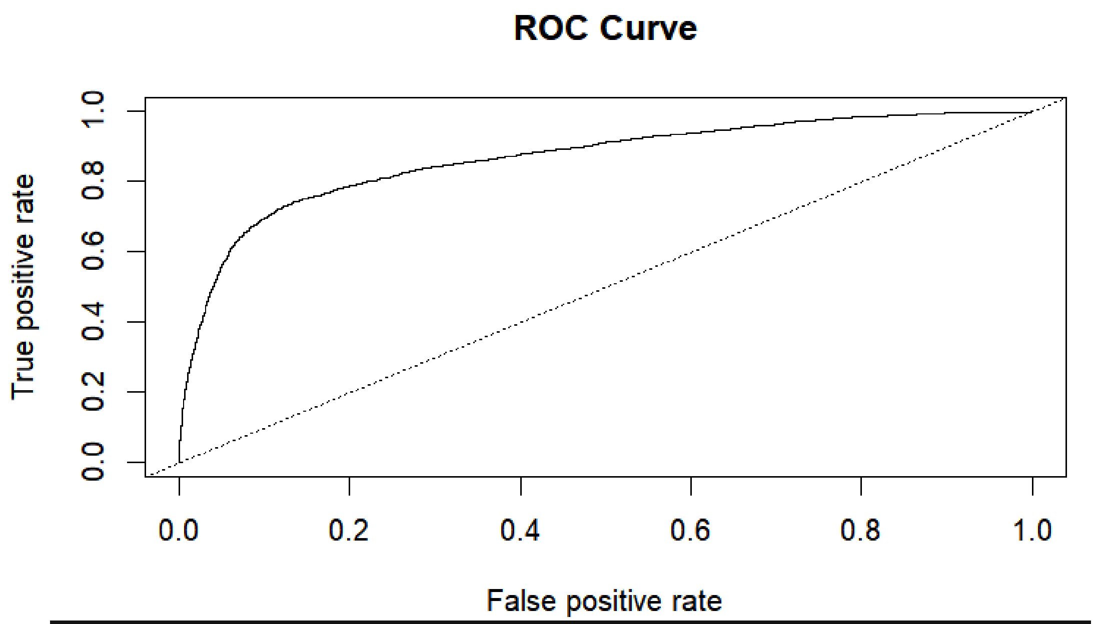
```r
# building a LDA model
library(MASS)
lda_fit<- lda(loan_status ~ ., data = train_set)
lda_fit

# creating confusion matrix
lda_pred = predict(lda_fit, test_set)$class
table(predict_status = lda_pred,
      true_status = test_set$loan_status)

# computing accuracy
acc1 = (10609 + 1801) / (10609 + 1257 + 649 + 1801) *100
cat("Accuracy:", acc1)

# computing
library(ROCR)
lda_pred = predict(lda_fit, test_set)
pred = prediction(lda_pred$posterior[,2], test_set$loan_status)
perf = performance(pred, "tpr", "fpr")
plot(perf, main = "ROC Curve")
abline(0, 1, lty=3)
```

LINEAR DISCRIMINANT ANALYSIS

```
              true_status
predict_status     0     1
             0 10609  1257
             1   649  1801
Accuracy: 86.68623
```

Accuracy is 0.8669

AUC value is 0.8669




```{r}
auc = as.numeric(performance(pred, "auc")@y.values)
auc
```

```
[1] 0.8669293
```

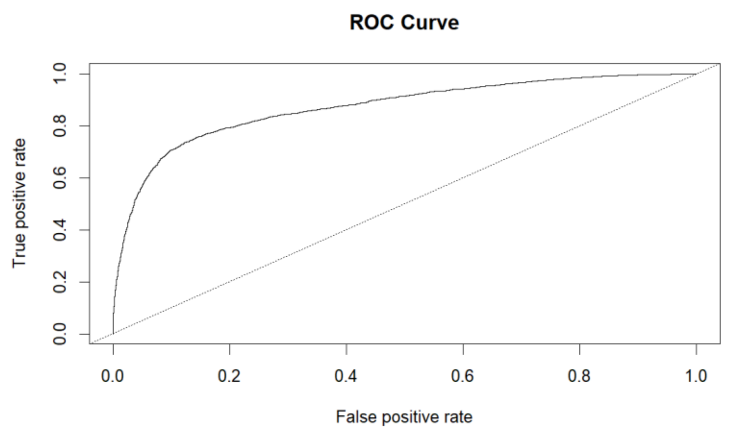
RIDGE REGRESSION

```r
library(glmnet)
xmat = model.matrix(loan_status ~ ., df1)[,-1]
y = df1$loan_status

for (i in 1:ncol(xmat)){
  xmat[,i] = scale(xmat[,i], center=FALSE)
}

mod.ridge = glmnet(xmat, y, alpha=0, family="binomial")

plot(mod.ridge, xvar="lambda", label=TRUE)

coefs.ridge = coef(mod.ridge)

set.seed(123)
cv.out = cv.glmnet(xmat, y, alpha=0, nfolds=5, family="binomial")
best.lambda = cv.out$lambda.min
best.lambda

test.std = model.matrix(loan_status ~ ., test_set)[,-1]

for (i in 1:ncol(test.std)){
  test.std[,i] = scale(test.std[,i], center=FALSE)
}
best.ridge = glmnet(xmat, y, alpha=0, lambda=best.lambda,
family="binomial")
```

```r
# computing accuracy
ridge.pred = predict(best.ridge, newx = test.std, type="response")
ridge.pred = ifelse(ridge.pred > 0.5, "Yes", "No")
cm.ridge = table(pred=ridge.pred, true=test_set$loan_status)
cm.ridge

ACC = (cm.ridge[1, 1] + cm.ridge[2, 2])/sum(cm.ridge)
cat("Accuracy:",ACC)

# computing AUC
ridge.prob = predict(best.ridge, newx=test.std, type="response")
ridge.pred = prediction(ridge.prob, test_set$loan_status)
ridge.perf = performance(ridge.pred, "tpr", "fpr")
plot(ridge.perf, main="ROC Curve")
abline(0,1,lty=3)

ridge.auc=as.numeric(performance(ridge.pred, "auc")@y.values)
cat("\nAUC:",ridge.auc)
```

RIDGE REGRESSION

```
        true
pred        0      1
   No   10821   1480
   Yes    437   1578
Accuracy: 0.8660939
AUC: 0.868914
```



Accuracy is 0.8661

AUC value is 0.8689

CLASSIFICATION TREE

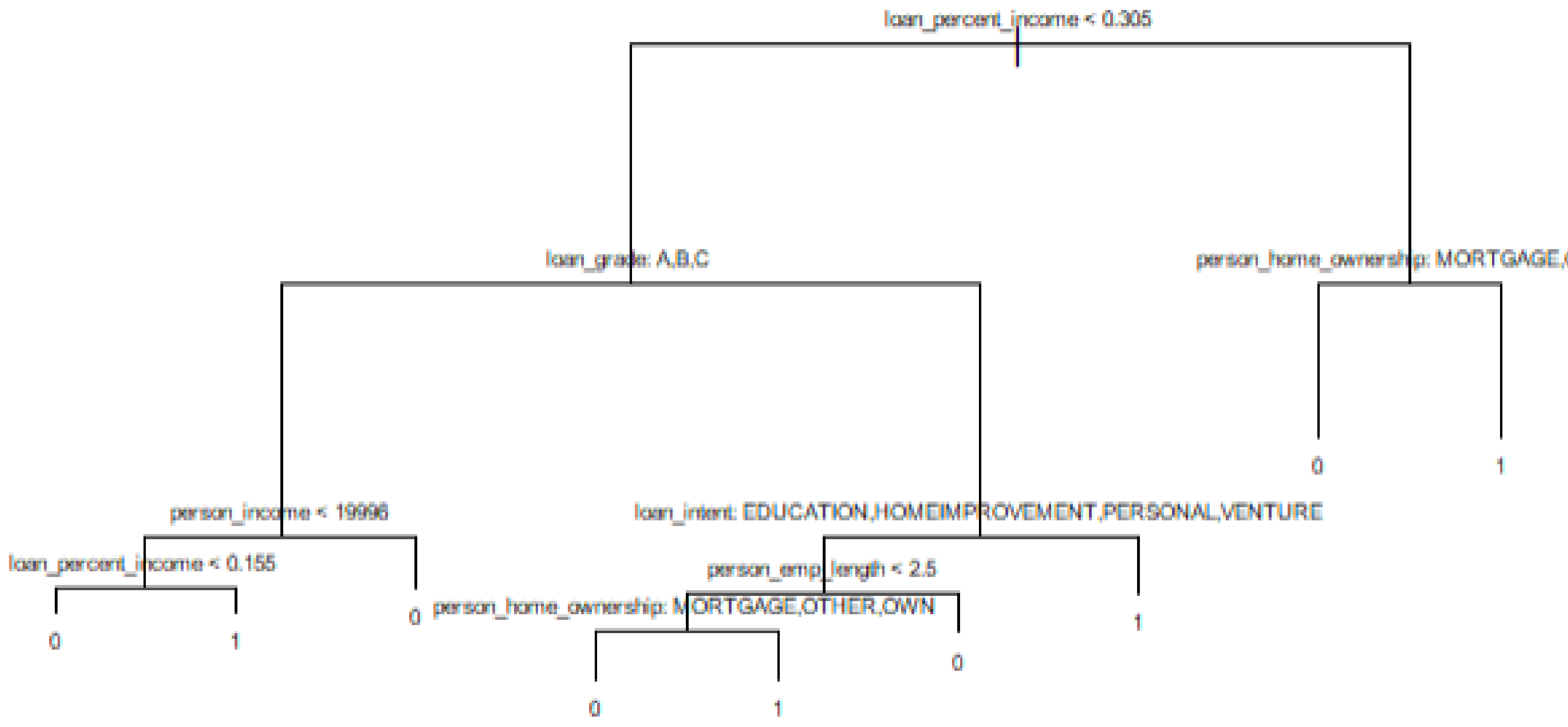
```
library(tree)

train_set$loan_status <- as.factor(train_set$loan_status)
# building the classification tree
mod.tree <- tree(loan_status ~ ., data = train_set)
cv.out = cv.tree(mod.tree)
cv.out$size[which.min(cv.out$dev)]
cv.out

plot(mod.tree)
text(mod.tree, pretty=0, cex=0.5)

tree.pred = predict(mod.tree, test_set, type="class")
cm.tree = table(pred = tree.pred, true=test_set$loan_status)
cm.tree

tree_acc = (cm.tree[1,1] + cm.tree[2, 2])/sum(cm.tree)
cat("Accuracy:", tree_acc)

tree.pred = prediction(as.numeric(tree.pred),
as.numeric(test_set$loan_status))
tree.perf = performance(tree.pred, "tpr", "fpr")
plot(tree.perf, main="ROC Curve")
abline(0,1,lty=3)

tree.auc = as.numeric(performance(tree.pred, "auc")@y.values)
cat("\nAUC:",tree.auc)
```

```
> cv.out
$size
[1] 9 8 6 5 4 3 1

$dev
[1]  7361.057  7625.395  8374.738  9223.638  9356.498 10724.083 15077.568

$k
[1]      -Inf  210.2959  380.8347  466.7247  512.3568 1368.3568 2185.1979

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

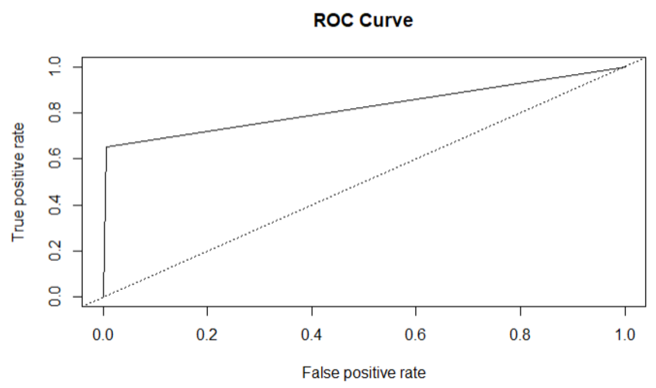
```
> cv.out$size[which.min(cv.out$dev)]
[1] 9
```

CLASSIFICATION TREE

- You are considered not a risk if:
  - Your loan is more than 30.5% of your annual income, and you either mortgage or own a house
  - Your loan has a grade of A, B, or C, with the intent on the graph, and have been employed for more than 2.5 years
- You are considered a risk if:
  - Your income is less than 19900 and your loan is more than 15.5% of your income
  - You have been employed for less than 2.5 years and you rent



CLASSIFICATION TREE




```
      true
pred      0     1
   0 11199  1061
   1    59  1997
Accuracy: 0.9217659
AUC: 0.8239002
```

Accuracy is 0.9218

AUC value is 0.8239

SUMMARY OF MAIN RESULTS

	Logistic Regression	Linear Discriminant Analysis	Ridge Regression	Classification Tree
Accuracy	69.51%	86.69%	86.61%	**92.18%**
ROC/AUC	0.5030	0.8669	**0.8689**	0.8239

While Classification Tree has the best accuracy, the Ridge Regression has the highest AUC value. If the lender has low-risk clients or low loan amounts, use Classification Tree. If the lender has high-risk clients or high loan amounts, then use Ridge Regression.

CHALLENGES AND POSSIBLE FUTURE WORK

We could not fit a K-nearest-neighbor algorithm because there were too many ties.

For future work, we could use unsupervised learning methods like neural networks.