STAT 473: MACHINE LEARNING

**CREDIT RISK ANALYSIS**

horizontal line



# 

# 

# 

**CALIFORNIA STATE UNIVERSITY, LONG BEACH**

**BY: DANIEL GARCIA, NICHOLAS SHIERY, JOE SORIA**

# TABLE OF CONTENTS

[**Introduction** ……………………………………](#_arolcxe0i15c)……….**3**

[Data ………………………………………………..](#_la5jp5tnimjw) **3**

[Variable Information ……………………](#_yjpdsm1ku3xs)  **4**

[**Questions of Interest** ……………………..](#_ysb63d3o9ce6)**4**

[**Analysis**  ……………………………………………. **5**](#_n635qy5up76w)

Question 1 ……………………………………… **5**

Question 2 …………………………………… **7**

Question 3 ………………………………… **9**

Question 4 ………………………………. **12** **Conclusion**  ……………………………………. **13**

**References** ……………………………………  **14**

**Appendix**  ……………………………………… **15**

# 

# 

# 

# 

# 

# Introduction

The purpose of this credit risk analysis is to evaluate the creditworthiness of a borrower and assess the risk associated with offering a loan to them. In today's economy, lending institutions are exposed to a variety of risks, including credit risk, which can result in significant financial losses. Therefore, it is critical for lenders to conduct a thorough credit risk analysis before extending credit to any borrower.

In conducting this analysis, we have employed various machine learning techniques to analyze German financial data from the UC Irvine Machine learning laboratory. The results of this analysis are aimed to inform decision makers on whether or not to offer a loan based on specific features of candidates.

The following report is structured to present our findings and recommendations in a clear and concise manner. The central questions that guide this report are:

1. Which classification model best fits the data? Determined by (Accuracy, Specificity, etc.)?
2. What attributes are significant in determining whether an applicant is a good or bad credit risk?
3. Which type of applicants should the bank accept based on your findings?
4. Which type of applicants should the bank reject based on your findings?

As we answer these questions, we hope that our analysis can help banks lead to more informed decision making when evaluating a borrower's risk profile.

## Data

The original data set we used was taken from UCI Machine Learning repository and was titled “German Credit Data” [1]. The specific data set we used for this analysis was a modified version of this original data set. We got this data set from Kaggle [2], an online community of data scientists and machine learning practitioners who find and publish data sets. The original data set contained 20 variables and 1000 observations. The modified data set, which we used, had only 11 variables and 1000 observations.

The variables included in our data set are: ID, Age, Sex, Job, Housing, Saving account, Checking account, Credit amount, Duration, Purpose, and Risk.

## Variable Information

1. Age (Numeric)
2. Sex (Character: male, female)
3. Job (Numeric: 0- unskilled and non-resident, 1- unskilled and resident, 2- skilled, 3- highly skilled)
4. Housing (Character: own, rent, or free)
5. Saving account (Character: NA, little, moderate, quite rich, rich)
6. Checking account (Character: NA, little, moderate, rich)
7. Credit amount (numeric: in Deutsch Mark(German currency))
8. Duration (numeric: in months)
9. Purpose(Character: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/other)
10. Risk (Character: good, bad)

Before beginning our analysis, we noticed that the “Checking accounts” and “Saving accounts” columns had 394 and 183 “NA” values, respectively. We made the decision to change these “NA” values to “none”. Furthermore, we want to note that 409/478 applicants who had “NA” values in Saving accounts or Checking accounts were identified as good risk clients.

# Questions of Interest

1. What attributes are significant in determining whether an applicant is a good or bad credit risk?
2. Which classification model is best? determined by (Accuracy, Specificity, AUC.)?
3. Which type of applicants should the bank accept?
4. Which type of applicants should the bank reject?

# Analysis

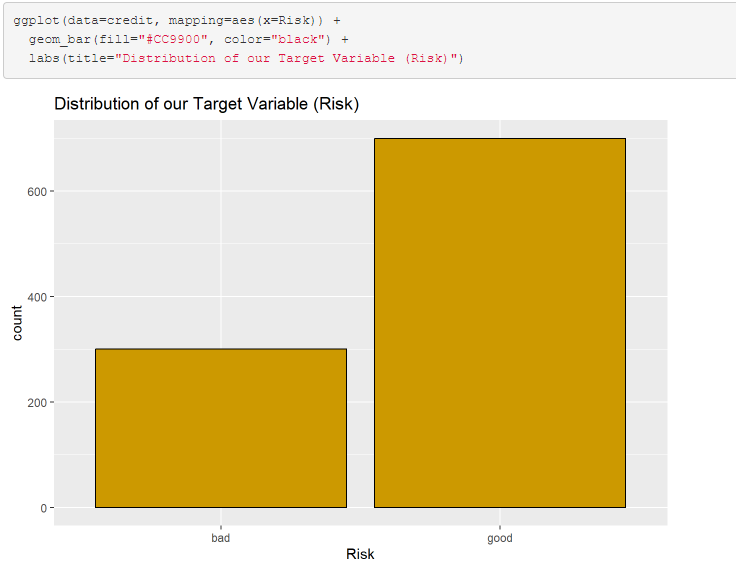
This section covers the details of the methods we used to answer our questions of interest, followed by an interpretation of the results for each method. Since each question uses different techniques to gain insight, we divide this section by each question.

1. **QUESTION 1**

This question is answered using machine learning models.

**EDA**

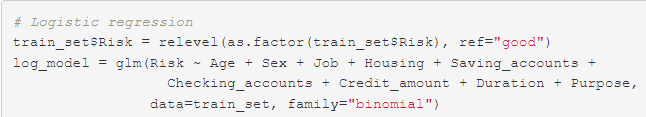
To answer this question we first note that our response variable (Risk) is categorical, taking on values of “good” or “bad”. It is also important to note that when we perform exploratory data analysis on our response variable, we find that 30% of our responses identified clients as “bad” risk while 70% of our clients were identified as “good” risk. The following bar plot shows these results:

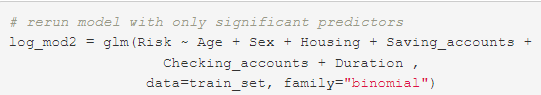


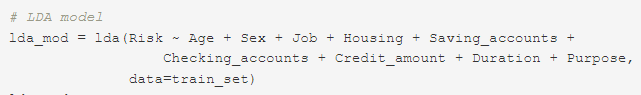
It is important to take these results into consideration when deciding on which evaluation criteria to judge our models.

**Modeling**

The five classification models used in this analysis are: Logistic regression, Linear Discriminant Analysis, Random Forest, and Bagging. Below, we show the code used for each model, interpret the results, and create a confusion matrix that will later be used to determine which is the best model for our analysis.

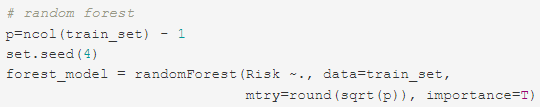
1. Logistic regression

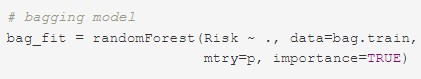
First, we ran the logistic regression model with all predictors. Then, after determining significant predictors using α = 0.05, we ran the model again with only significant predictors: Age, Sex, Housing, Saving accounts, Checking accounts, Duration.

1. Linear Discriminant Analysis

When we ran the LDA model, we were able to obtain the prior probabilities: 

1. Random Forest

Our random forest model had an Out-of-bag estimate of error rate of 26.75%.

1. Bagging

Our bagging model had an Out-of-bag estimate of error rate of 26%.

**Accuracy:**

**False Positive Rate:**

**Evaluation criteria:**

Which model is best? This depends on our bank's priorities. If the objective of our bank is to maximize profit, then accuracy will be the best evaluation criterion. However, if our bank is concerned primarily with minimizing the amount of loans offered to bad risk clients, then the False Positive Rate will be our evaluation criterion.

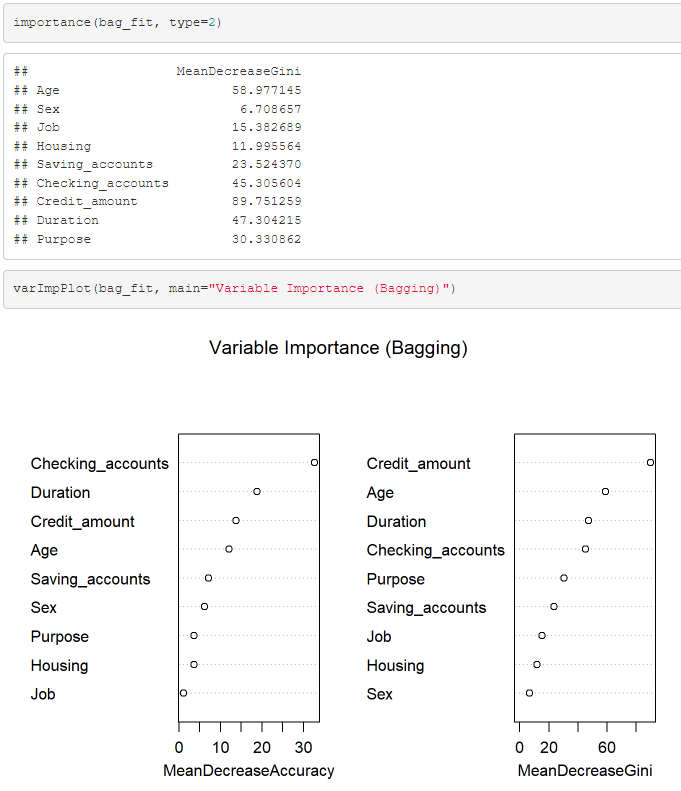
The accuracy results for the bagging model is 73.5% and the accuracy for the LDA model is 71%, giving a 2.5% difference. The False Positive Rates for the bagging model and the LDA model are 53.73% and 52.24%, respectively, which is about a 1.5% difference. Taking these results into consideration, we choose the bagging model as the best model since we will gain more income on the high accuracy.

1. **QUESTION 2**

We use two of our models to answer this question: Bagging model and the Logistic regression model.

When we ran the logistic regression model on our data, we found that the significant predictors at the 5% level of significance include: Age, Sex, Owner of a House, no saving accounts, Checking account, and Duration. We know that these predictors are significant by the p-value being less than 0.05.

The bagging model further validates the results of the logistic regression model by identifying the most important variables in the analysis. The following is the code and output for retrieving the important variables identified by the bagging model:



From this model we see that the four most important variables in terms of the mean decrease in accuracy and the mean decrease in Gini index are: Age, Duration, Credit amounts, and checking accounts.

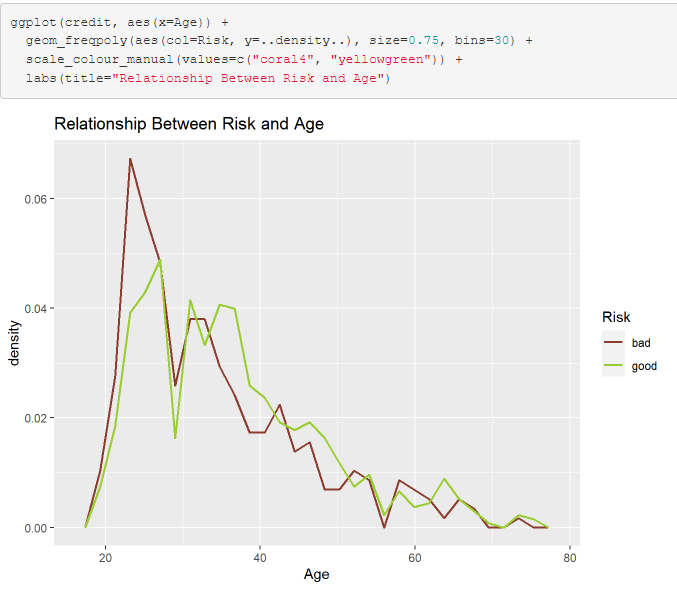
If we take the most important variables from both models into account, we find that age, duration, and checking accounts are the three most important variables for determining an applicant’s risk.

1. **QUESTION 3**

To answer this question, we consider our most important predictors identified in question 2 and perform exploratory data analysis on each variable. However, not all variables identified as the most important showed obvious visual differences amongst good risk and bad risk clients. The goal in answering this question is to visually convince our audience that certain variables have a favorable proportion of good risk clients to bad risk clients based on values of the variables.

Age

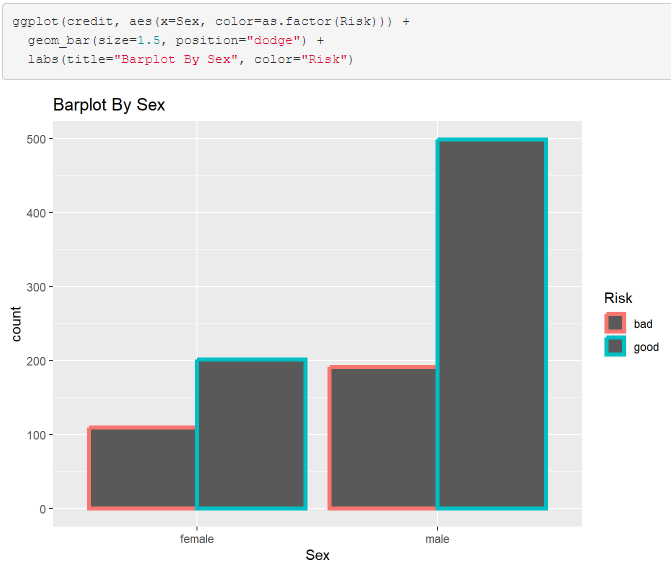
First, let's look at a frequency plot of age separated by good risk clients and bad risk clients.



This frequency plot shows that in the age range of approximately 30-52, the amount of good risk clients exceeds the amount of bad risk clients. In fact, this age range represents the biggest gap of bad risk clients to good risk clients by far.

Sex

The next variable of interest will be sex. We will show a bar plot to demonstrate the proportion of good risk clients to bad risk clients by sex.

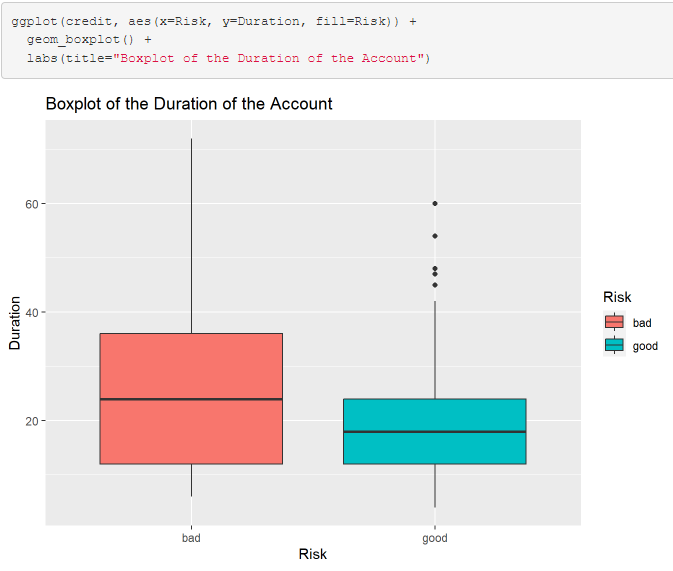


This bar plot shows that the proportion of good risk clients to bad risk clients is more than double. To be precise, the amount of male good risk clients is 499 and the amount of male bad risk clients is 191. Meaning that the proportion size is 499/191=2.612. Hence, we want to target males. The following code shows these results:



Duration

Duration of a loan is the final variable of interest for answering this question. We knew from the logistic regression model that longer duration accounts tended to promote bad risk clients. The boxplot validates the numerical results:



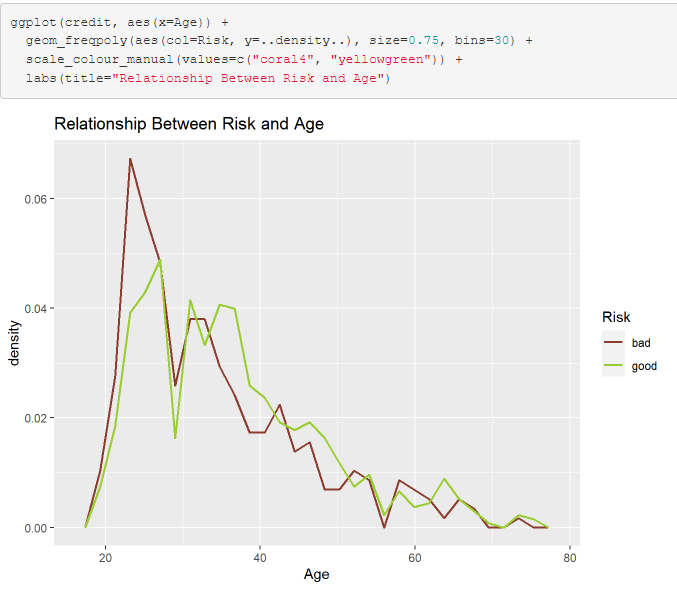
From this boxplot, we see that good risk clients tend to be associated with shorter terms of loans, typically below 20 months.

1. **QUESTION 4**

The two main variables identified for this question were age and purpose. Exploratory data analysis was used to visually highlight the proportion of bad loans to good loans amongst groups in these particular variables.

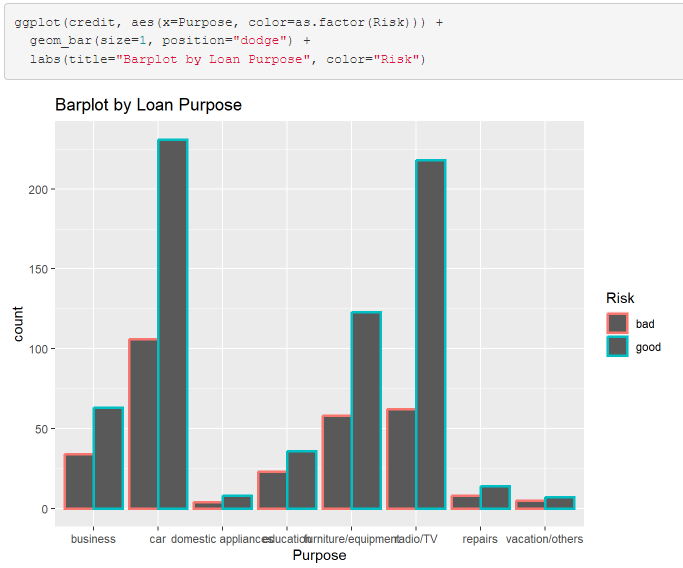
Age

Age showed a significant amount of bad risk loans compared to good risk loans for those below 30-years-old. In fact, this age range represents the biggest gap of bad risk clients to good risk clients by far. Once again, here is the code and frequency plot:



**Purpose**

When we look at specific groups within the purpose of the loan of applicants, we see the proportion of bad risk clients to good risk clients at an alarming rate in vacation/other, business, repairs, and education purposes. These results are demonstrated in the following bar plot:



# Conclusion

Our results suggest that the best model for classifying applicants as good risk or bad risk is the bagging model. We found the two best models to be the bagging model and the LDA model. The bagging model is best because it has an accuracy of 73.5%, which is the highest of all models. This means that the bagging model will allow us to identify the maximal amount of good risk clients who we can offer loans to while minimizing the maximal amount of bad risk clients who we will decline loans to. Although the LDA model has the lowest false positive rate, the difference in the false positive rate between the bagging model and the LDA model was only about 1.5%. Whereas the difference in accuracy between the bagging model and the LDA model was 2.5%. Making the bagging model the best overall model.

Through the logistic regression model and the bagging model, we were able to identify the most important attributes that determine whether an applicant is a good or bad credit risk as age, duration, and checking accounts. We used this information to uncover patterns within these variables to identify which groups we found favorable as loan candidates and which groups we found alarming as loan candidates. Our ideal loan candidate will be a male in the age range of 30-52, for radio/TV or car purposes, with a short term duration on the account. Generally, we want to stay away from applicants that are 30-years-old or younger who want to take a loan out for vacation/other, business, repairs, or education purposes.

**References**

[1] UCIrvineMachineLearningRepositoryGermanCreditData. [https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+ data)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+)

[2] Kaggle. <https://www.kaggle.com/datasets/uciml/german-credit>

# Appendix

---

title: "Credit Risk"

author: "Daniel Garcia"

date: "2023-04-15"

output:

html\_document: default

pdf\_document: default

word\_document: default

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

#### libraries

```{r}

library(tidyverse)

library(tree)

library(skimr)

library(MASS)

library(ROCR)

library(randomForest)

library(gbm)

```

#### Data

```{r}

# importing data

credit = read\_csv("C:\\Users\\dgray\\OneDrive\\Stats\\Stat 473 Machine learning\\Data\\german\_credit\_data.csv")

# observing data

glimpse(credit)

```

```{r}

unique(credit$`Saving accounts`)

unique(credit$`Checking account`)

```

#### cleaning data

```{r}

colnames(credit)[c(1, 6, 7, 8)]= c("id", "Saving\_accounts", "Checking\_accounts", "Credit\_amount")

glimpse(credit)

```

```{r}

# checking for NA values

na.credit = credit[rowSums(is.na(credit))>0,]

na.credit

```

```{r}

sum(is.na(credit$Checking\_accounts))

sum(is.na(credit$Saving\_accounts))

```

```{r}

unique(credit$Checking\_accounts)

```

\

478 observations have NA values in at least 1 column. Without going through all observations, only savings accounts and checkings accounts columns are the only columns that produce NA's.

\

\* number of NA values in Saving\_accounts column: 183

\* number of NA values in Checking\_accounts column: 394

\

\* unique values in Saving accounts: NA, little, quite rich, rich, moderate.

\

\* unique values in Saving accounts: NA, little, rich, moderate.

\*\*lets replace all NA values with "none"\*\*

```{r}

# replace NA values with "none"

credit[is.na(credit)]="none"

glimpse(credit)

```

```{r}

df7=credit %>%

filter(Saving\_accounts=="none" | Checking\_accounts=="none")

nrow(df7)

df7 %>%

filter(Risk=="good") %>%

nrow()

```

\

\* 409/478 applicants who had NA values in Saving accounts or Checking accounts were identified as good risk clients

\

now that the data is clean, lets start Exploratory data analysis

```{r}

# create training set and testing set

n = nrow(credit)

set.seed(1)

Z = sample(n, size=n\*0.8)

train\_set = credit[Z,]

test\_set=credit[-Z,]

```

#### EDA

```{r}

skim(credit)

```

\* mean age is about 36

\* mean credit amount is about $3,270

\* mean duration is about 21 months

```{r}

sum(credit$Risk=="bad")/nrow(credit)

```

\*\*30% of our data has credit risk classified as bad, 70% is good\*\*

```{r}

ggplot(data=credit, mapping=aes(x=Risk)) +

geom\_bar(fill="#CC9900", color="black") +

labs(title="Distribution of our Target Variable (Risk)")

```

```{r}

numerical\_credit = credit %>%

dplyr::select(2, 8, 9)

GGally::ggpairs(numerical\_credit)

```

\* credit amount and duration have the highest correlations with respect to risk.

```{r, warning=FALSE}

ggplot(credit, aes(x=Sex, color=as.factor(Risk))) +

geom\_bar(size=1.5, position="dodge") +

labs(title="Barplot By Sex", color="Risk")

```

\

\* over 1/3 of loans to females result in bad loans.

\* around 1/4 of all loans to males result in bad loans.

```{r}

credit %>%

filter(Sex=="male" & Risk=="good") %>%

nrow()

```

```{r}

credit %>%

filter(Sex=="male" & Risk=="bad") %>%

nrow()

499/191

```

```{r}

ggplot(credit, aes(x=Purpose, color=as.factor(Risk))) +

geom\_bar(size=1, position="dodge") +

labs(title="Barplot by Loan Purpose", color="Risk")

```

\

\* although a lot of bad loans come from Cars, so too do a lot of good loans.

\* Radio/TV purposes is second but the proportion size isnt that bad.

\* Seems like nearly 1/3 of all loans for furnutire/equipment are bad loans.

```{r, warning=FALSE}

ggplot(credit, aes(x=Age)) +

geom\_freqpoly(aes(col=Risk, y=..density..), size=0.75, bins=30) +

scale\_colour\_manual(values=c("coral4", "yellowgreen")) +

labs(title="Relationship Between Risk and Age")

```

\

\* More bad credit risks typically come from younger ages (around 18-33)

\* Good credit risks typically come from middle aged groups (around 35-52)

\* No obvious visual insights for groups aged over 52

```{r}

ggplot(credit, aes(x=Risk, y=Duration, fill=Risk)) +

geom\_boxplot() +

labs(title="Boxplot of the Duration of the Account")

```

\

\* Duration on the account influences the Risk associated with the account.

\* Short term loans are typically better

```{r}

ggplot(data=credit, mapping=aes(x=Purpose, y=Duration)) +

geom\_boxplot(fill='#A4A4A4', color="black") +

facet\_wrap(~Risk, nrow=2) +

labs(title="Boxplot of Credit Amount by Loan Purpose", y="Duration")

```

\

\* Longer duration accounts are typically requested for vacation/other purposes and business purposes

### Modelling

#### types of models:

1. Logistic regression

2. LDA

3. Random forest

4. Bagging

5. Boosting

##### 1.

\*\*note: since reference is "good", we are predicting "bad" risk\*\*

```{r}

# Logistic regression

train\_set$Risk = relevel(as.factor(train\_set$Risk), ref="good")

log\_model = glm(Risk ~ Age + Sex + Job + Housing + Saving\_accounts +

Checking\_accounts + Credit\_amount + Duration + Purpose,

data=train\_set, family="binomial")

summary(log\_model)

```

```{r}

# rerun model with only significant predictors

log\_mod2 = glm(Risk ~ Age + Sex + Housing + Saving\_accounts +

Checking\_accounts + Duration ,

data=train\_set, family="binomial")

summary(log\_mod2)

```

\*INTERPRETATION:\*

```{r}

log\_pred = predict(log\_mod2, newdata = test\_set, type = "response")

log\_prob = ifelse(log\_pred < 0.314, "good", "bad")

log\_conf\_mat= table(predict\_status=log\_prob,

true\_status=test\_set$Risk)

log\_conf\_mat

```

#### 2.

```{r}

# LDA model

lda\_mod = lda(Risk ~ Age + Sex + Job + Housing + Saving\_accounts +

Checking\_accounts + Credit\_amount + Duration + Purpose,

data=train\_set)

lda\_mod

```

\* prior probabilities:

$\hat{\pi}\_{good} = 0.70875$ , $\hat{\pi}\_{bad} = 0.29125$

```{r}

lda\_pred\_class = predict(lda\_mod, newdata=test\_set)$class

lda\_conf\_mat = table(predict\_status= lda\_pred\_class,

true\_status= test\_set$Risk)

lda\_conf\_mat

```

#### 3.

```{r}

# random forest

p=ncol(train\_set) - 1

set.seed(4)

forest\_model = randomForest(Risk ~., data=train\_set,

mtry=round(sqrt(p)), importance=T)

```

```{r}

rf.pred = predict(forest\_model, test\_set, type="class")

forest\_conf\_mat = table(pred=rf.pred,

true= test\_set$Risk)

forest\_conf\_mat

```

##### 4.

```{r}

# boosting model

bag.df = credit %>%

mutate(Sex = factor(Sex),

Housing= factor(Housing),

Saving\_accounts = factor(Saving\_accounts),

Checking\_accounts = factor(Checking\_accounts),

Purpose = factor(Purpose),

Risk= factor(ifelse(Risk=="good", 1, 0)))

bag.df = bag.df %>%

dplyr::select(-id)

# create training set and testing set

n = nrow(bag.df)

set.seed(1)

Z = sample(n, size=n\*0.8)

bag.train = bag.df[Z,]

bag.test=bag.df[-Z,]

```

```{r}

# bagging model

bag\_fit = randomForest(Risk ~ ., data=bag.train,

mtry=p, importance=TRUE)

bag\_fit

```

```{r}

importance(bag\_fit, type=2)

```

```{r}

varImpPlot(bag\_fit, main="Variable Importance (Bagging)")

```

```{r}

yhat.test\_bag = predict(bag\_fit, bag.test, type="class")

bag\_conf\_mat= table(pred=yhat.test\_bag,

true= bag.test$Risk)

bag\_conf\_mat

```

#### What is our Goal for a model?

\*\*if the goal of our analysis is to minimize lending to people who are a bad credit risk and maximize lending to people who are a good credit risk, then the optimal solution would be to pick the model with the highest overall accuracy. Which model results in the highest overall accuracy?\*\*

#### Accuracy

```{r}

# accuracy for logistic model

log\_acc = sum(diag(log\_conf\_mat))/sum(log\_conf\_mat)

log\_acc

```

```{r}

# accuracy for LDA model

lda\_acc = (lda\_conf\_mat[1,2] + lda\_conf\_mat[2,1])/sum(lda\_conf\_mat)

lda\_acc

```

```{r}

# accuracy for random forest model

forest\_acc = (forest\_conf\_mat[1,2] + forest\_conf\_mat[2,1])/sum(forest\_conf\_mat)

forest\_acc

```

```{r}

bag\_acc = sum(diag(bag\_conf\_mat))/sum(bag\_conf\_mat)

bag\_acc

```

```{r}

data.frame(cbind(log\_acc, lda\_acc, forest\_acc, bag\_acc))

```

#### Conclusion

\*\*Bagging model is the best model to maximize overall profit\*\*

\*\*If the goal of our analysis is to prevent loss, then we might be concerned with the false positive rate. In other words, by picking the model with the lowest false positive rate we minimize the risk associated with giving loans to parties who are a bad risk.\*\*

#### False Positive rate

```{r}

lda\_FPR = 1 - lda\_conf\_mat[2,1]/sum(lda\_conf\_mat[2,1], lda\_conf\_mat[1,1])

lda\_FPR

```

```{r}

logistic\_FPR = 1 - log\_conf\_mat[2,1]/(log\_conf\_mat[1,1]+log\_conf\_mat[2,1])

logistic\_FPR

```

```{r}

forest\_FPR = 1 - forest\_conf\_mat[2,1]/sum(forest\_conf\_mat[2,1], forest\_conf\_mat[1,1])

forest\_FPR

```

```{r}

bag\_FPR = 1-bag\_conf\_mat[1,1]/sum(bag\_conf\_mat[2,1], bag\_conf\_mat[1,1])

```

```{r}

data.frame(cbind(lda\_FPR, logistic\_FPR, forest\_FPR, bag\_FPR))

```