Benjamin O. Tayo

Prediction of Loan Status Using Monte Carlo Simulation

Abstract

Using the data set provided, we built a simple model using Monte Carlo simulation for predicting the fraction of loans that will default after the 3-year duration of the loan. Our model revealed a 95% confidence interval of 14.8% $\pm$ 0.2% for Monte-Carlo simulation of N = 1000 replicated copies of the data set.

Keywords: Loan status, loan origination, loan charge-off, Monte Carlo simulation, predictive analytics

Introduction: Predicting the status of a loan is an important problem is risk assessment. A bank or financial organization has to be able to estimate the risk involved before granting a loan to a customer. Data Science and predictive analytics play an important role in building models that can be used to predict the probability of loan default. In this project, we are provided with a data set loan_timing.csv containing 50000 data points. Each data point represents a loan, and two features are provided as follows:

- The column with header "days since origination" indicates the number of days that elapsed between origination and the date when the data was collected.

- For loans that charged off before the data was collected, the column with header "days from origination to charge-off" indicates the number of days that elapsed between origination and charge-off. For all other loans, this column is blank.

Project Objective: The goal of this project is to use techniques of data science to estimate what fraction of these loans will have charged off by the time all of their 3-year terms are finished.

Exploratory Data Analysis: The data set was important in R and calculations were performed using R. We plot the following figures:

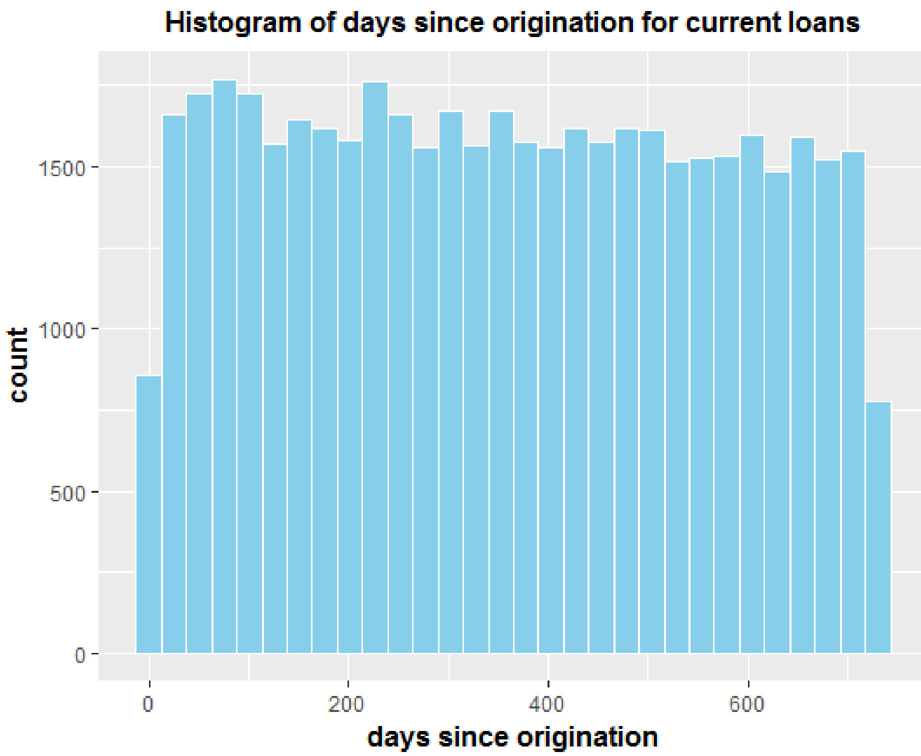


Figure 1: Histogram of days since origination for current loans.

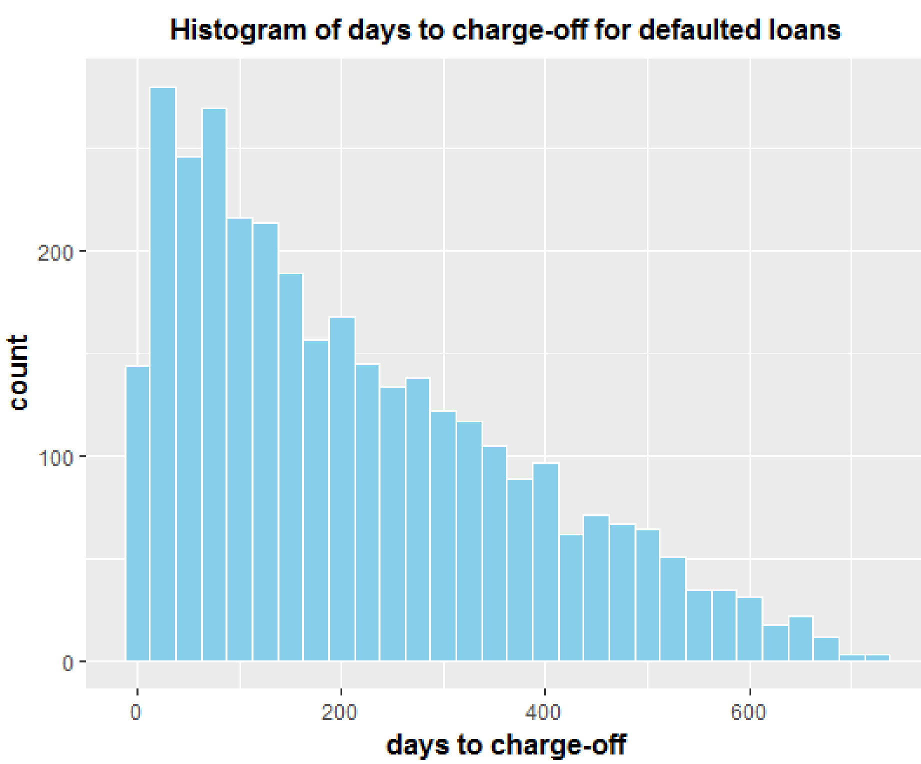


Figure 2: Histogram of days to charge-off for defaulted loans.

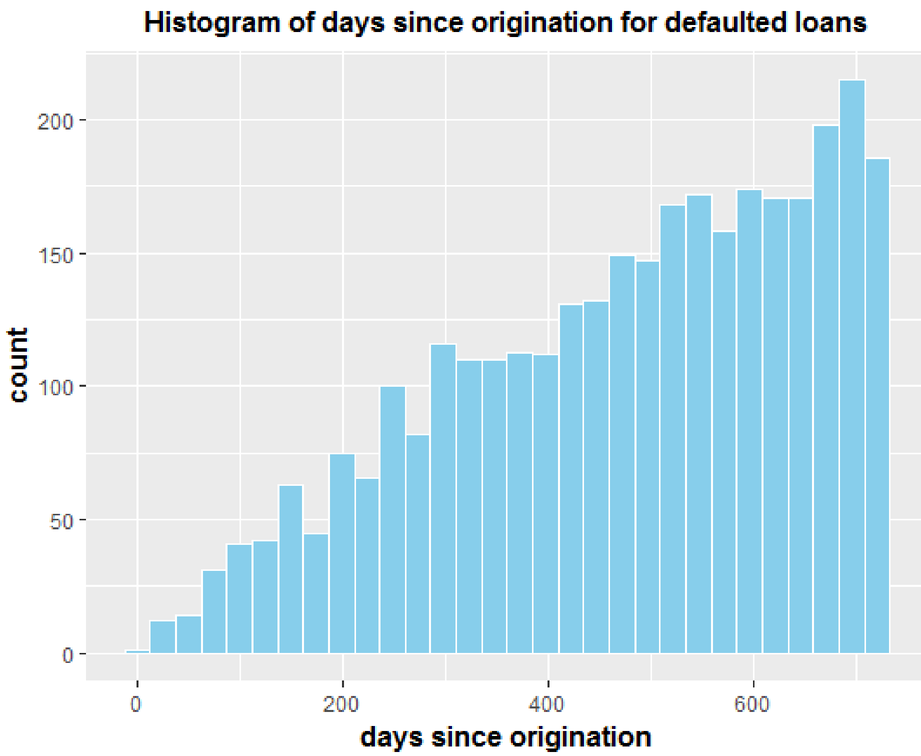


Figure 3: Histogram of days to charge-off for defaulted loans.

Figure 1 shows histogram of active loans, which are uniformly distributed over the days since origination.

From Figure 2, we see that proportion of loans that charged off decreases with increasing days from origination to charge-off. This shows that younger loans have a higher probability of defaulting. It also shows that 100% of loans defaulted within 2 years from date of origination.

Figure 3 shows the distribution of defaulted loans as a function of days since origination to the time when data about loan status was collected. The defaulted loans contain a large proportion (71%) of loans that are one year and older. These loans are less likely to default compared to younger loans.

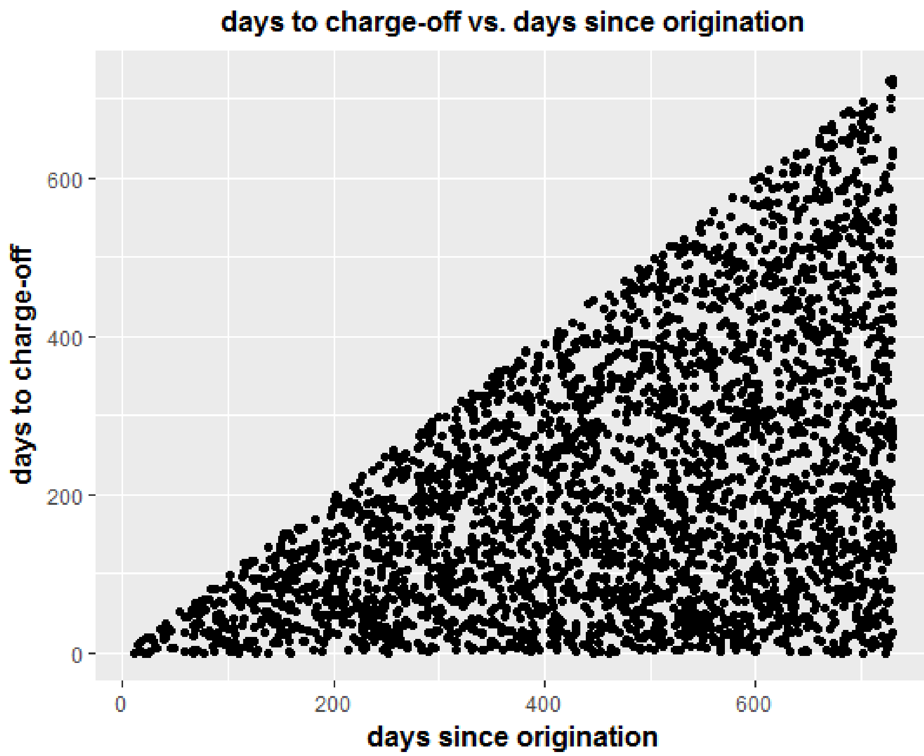
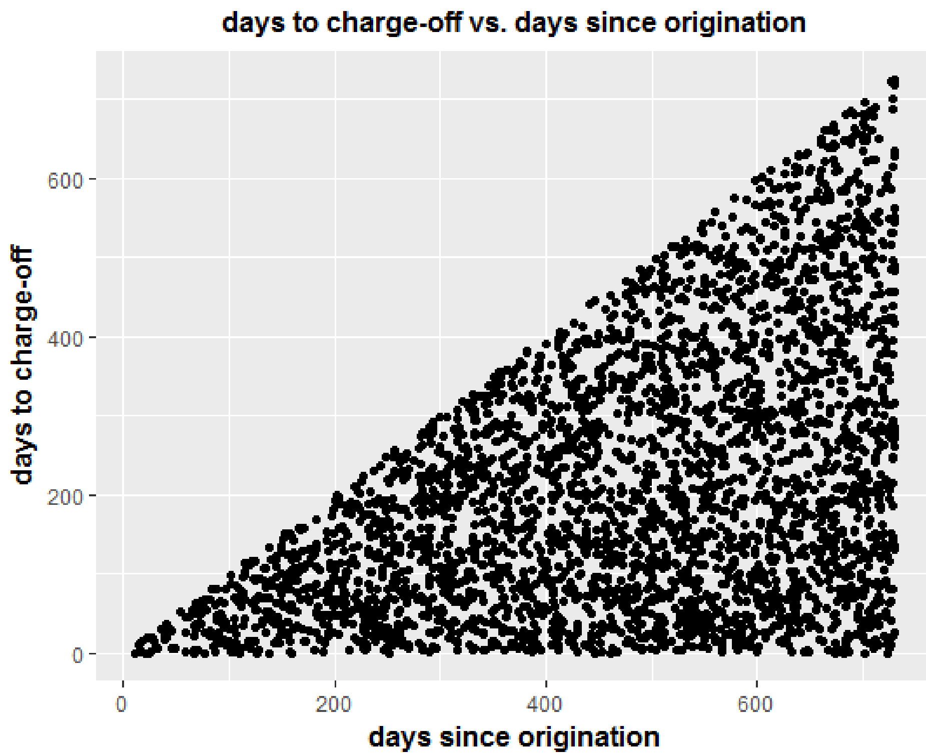


Figure 4:Plot of days to charge-off vs. days since origination for defaulted loans.

Model Selection: Our data set has only 2 features or predictors, and surfers from the problem of prevalence: 93% of the loans have an active status, while 7% has a default status. Use of Linear Regression for predicting fraction of loans that will have charged off after the 3 years loan duration produces a model that is biased towards the active loans.

Figure 3 indicates that relationship between days to charge-off and days since origination for defaulted loans can be simulated using Monte Carlo (MC) simulation. We therefore choose MC simulation as our model for predictive proportion of loans that will default.

Model Calculations: We generate a MC simulation of the defaulted loans, and compare with the original, as shown in figures 4 and 5:



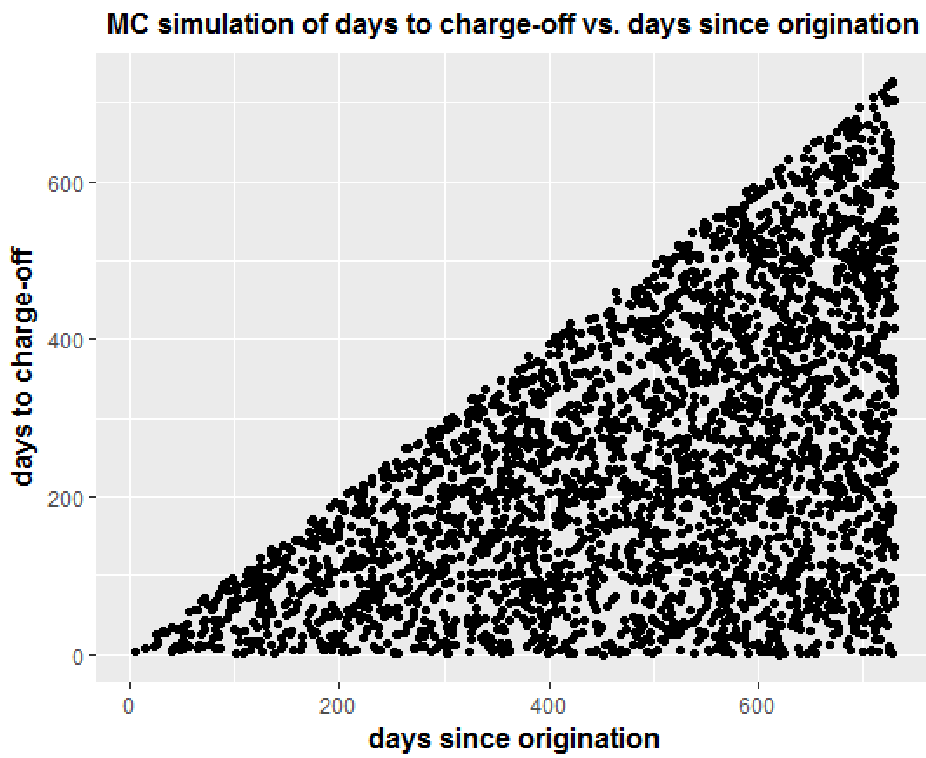


Figure 5: Original and MC simulation of days to charge-off vs. days since origination.

Because there is randomness associated with the charge-off of a loan, we see that MC simulation provides a good approximation for distribution of defaulted loans.

Predictions: Since we have demonstrated that the defaulted loans with charge-off and days since origination in the first 2 years (i.e. 0 to 730 days) can be approximated using a MC simulation, we can predict the fraction of loans that will charged off by the time all of their 3-year terms are finished using MC simulation.

The total number of charged off loans in our data set is 3,305. This means that are 46,695 loans that are currently active. Of these active loans, a certain proportion will default over the 3-year period. To estimate the total fraction of defaulted loans, we simulated defaulted loans with charge-off and days since origination covering entire duration of loan (i.e. 0 to 1095 days), then by appropriate scaling, we computed the fraction of loans that will have charged off after the 3-year term i.e., 1095 days.

By creating 1000 random trials, we obtained the following distribution for the fraction of defaulted loans 3-year term:

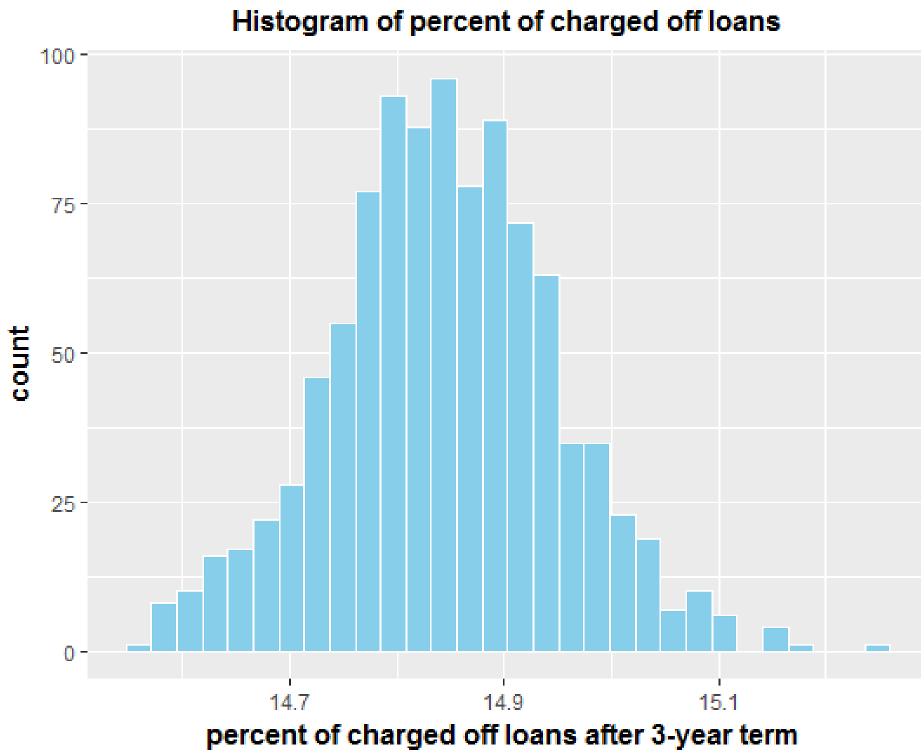


Figure 6: Histogram for fraction of charged off loans after 3-year term using N = 1000 samples.

Based on our calculations, the 95% confidence interval for the fraction of loans that will have charged off after the 3-year loan duration is accordingly 14.8% $\pm$ 0.2%.

Conclusions: We have presented a simple model based on the MC simulation for predicting the fraction of loans that will default at the end of the 3-year loan duration period. Different models could be used such as logistic regression, decision tree, etc. I would like to try these different approaches to see if the results are comparable to the MC simulation results.

Appendix: R Code for Performing Data Analysis

```
#R Code for Predicting Loan Status

# author: Benjamin O. Tayo

# Date: 11/22/2018

#Import Necessary Libarries
library(readr)
library(tidyverse)
library(broom)
library(caret)

# IMPORTATION OF DATASET
df<-read_csv("loan_timing.csv",na="NA")
names(df)=c("origination","chargeoff")
#partition data set into two: default (charged off ) and current
index<-which(!(df$chargeoff=="NA"))
default<-df%>%slice(index)
current<-df%>%slice(-index)

#EXPLORATORY DATA ANALYSIS

# Figure 1: Histogram of days to charge-off for defaulted loans

default%>%ggplot(aes(chargeoff))+geom_histogram(color="white",fill="skyblue")+
  xlab('days to charge-off')+ylab('count')+
  ggtitle("Histogram of days to charge-off for defaulted loans")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )

# Figure 2: Histogram of days since origination for defaulted loans

default%>%ggplot(aes(origination))+geom_histogram(color="white",fill="skyblue")+
  xlab('days since origination')+ylab('count')+
  ggtitle("Histogram of days since origination for defaulted loans")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )
```

```
# Figure 3: Plot of days to charge-off vs. days since origination for defaulted loans

default%>%ggplot(aes(origination,chargeoff))+geom_point()+
  xlab('days since origination')+ylab('days to charge-off')+
  ggtitle("days to charge-off vs. days since origination")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )

# Figure 4: Histogram of days since origination for active loans

current%>%ggplot(aes(origination))+geom_histogram(color="white",fill="skyblue")+
  xlab('days since origination')+ylab('count')+
  ggtitle("Histogram of days since origination for current loans")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )

# Monte Carlo Simulation of Defaulted Loans
set.seed(2)
N <- 3*365 # loan duration in days
df_MC<-data.frame(u=round(runif(15500,0,N)),v=round(runif(15500,0,N)))
df_MC<-df_MC%>%filter(v<=u)
df_MC<-df_MC%>%filter(u<=730 & v<=730) #select loans within first 2 years

df_MC[1:nrow(default),]%>%ggplot(aes(u,v))+geom_point()+
  xlab('days since origination')+ylab('days to charge-off')+
  ggtitle("MC simulation of days to charge-off vs. days since origination")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )


# Predicting fraction of these loans will have charged off by the time all of their 3-year terms are finished.
set.seed(2)
```

```
B<-1000
fraction<-replicate(B, {
  df2<-data.frame(u=round(runif(50000,0,N)),v=round(runif(50000,0,N)))
  df2<-df2%>%filter(v<=u)
  b2<-(df2%>%filter(u<=730 & v<=730))
  total<-(nrow(df2)/nrow(b2))*nrow(default)
  100.0*(total/50000.0)
})

mean(fraction)

fdf<-data.frame(fraction=fraction)

fdf%>%ggplot(aes(fraction))+geom_histogram(color="white",fill="skyblue")+
  xlab('fraction of charged off loans after 3-year term')+ylab('count')+
  ggtitle("Histogram of total fraction of charged off loans")+
  theme(
    plot.title = element_text(color="black", size=12, hjust=0.5, face="bold"),
    axis.title.x = element_text(color="black", size=12, face="bold"),
    axis.title.y = element_text(color="black", size=12, face="bold"),
    legend.title = element_blank()
  )

# Calculate Confidence Interval of Percentage of Defualted Loans after 3-year term

mean<-mean(fraction)
sd<-sd(fraction)
confidence_interval<-c(mean-2*sd, mean+2*sd)
confidence_interval
```