Probability of Default Model

Advanced Data Mining and Analytics

Submitted By Group_7

Akintoye Timothy Olugbenga

(oakintoy@kent.edu**)**

Vamsi Krishna Darapaneni

(vdarapan@kent.edu)

Alshehri Saleh (salsheh4@kent.edu**)**

Contents

Project Report - Group 7

Group Project Contribution Table	
Project Team Member:	**Contribution:**
Akintoye Timothy Olugbenga	***Code, Report, and Presentation***
Vamsi Krishna Darapaneni	***Head Coding***
Alshehri Saleh	***Had not done anything assigned to him in the Group Project***

1. Introduction (Project Description)

In the U.S, consumer borrowing is at a record $14 trillion "https://www.reuters.com/article/us-usa-fed-household-debt/u-s-household-debt-tops-14-trillion-and-reaches-new-record-idUSKBN20521Z" (,2020) . While lenders may see pockets of additional growth here and there, the credit economy is long overdue for some cooling, or even a reversal. If we should we say when and how the economy turns, it will certainly be the banks and lenders who tightened up their credit portfolios the fastest and most likely that we are most heading for a financial failure like 2008.

That doesn't have to mean abandoning higher-risk borrowers and rushing to an all-prime portfolio. Lenders can now use complex machine learning models that rely on hundreds or thousands of times more variables to assess risk. There are better Ai ML tools, that can analyze overall portfolio risk with a finer lens and use those insights, for example, adjust price on incoming borrowers to optimize yield heading into an economic downturn.

During a downswing, lenders should want as much granularity as they can get to reduce defaults and build more resilience into their portfolios.

Every industry is evaluating options and adopting ways to create value in our technology-driven world. In modern days, artificial intelligence and machine learning are revolutionizing the financial world and changing industries' experience, including banking, for the better. The effects of the new technologies are vast, though most banks are still in the early stages of adopting evolving AI technologies.

Artificial intelligence (AI) digitalizes banks and helps them meet the competition posed by FinTech players. About 32% of financial service providers are already utilizing AI technologies, such as predictive analytics, and voice recognition, just to name a couple, according to a joint research conducted by the National Business Research Institute and Narrative Science. Machine Learning and Artificial Intelligence strengthen banks' competitiveness through different methods, for instance, developing a better understanding of customer's behaviors and demands, predicting future outcomes and trends, cognitive process automation, robotic automation of processes, fraud prevention, and reducing errors and risk, among others. Ecobank utilizes machine learning to analyze "big data" to prevent fraud and monitor potential threats to customers. Even the United States Postal Service performs character recognition of handwritten characters by utilizing an algorithm and a computer vision system. Moreover, machine learning is used in banking to automatically reach lending decisions and to create a process that is used to approve or deny a candidate for a loan. The same methodology is used when making credit decisions.

When a customer applies for a loan or credit card, a machine learning algorithm evaluates her or his request against a database of thousands of other customers that have previously applied for a similar loan. The machine learning algorithm evaluates several variables to generate an aggregate model of past loan outcomes. The main data points are the loan amount, applicant income, and co-applicant income, if applicable. Comparing the data points with the data points of thousands of prior loanees, machine learning can create a predictive risk score.

The bank sets a base risk score threshold, with applicants falling below this threshold being accepted for their desired loans. J.P. Morgan Chase found that nearly 80% of loan servicing errors are due to contract interpretation miscalculations. They commence utilizing machine learning to analyze documents and extract critical data points and clauses. What previously required thousands of man hours manually reviewing and transcribing documents can now be completed in seconds. (JP Morgan)

In this project, we will be running the underwriting department of a bank to determine who would be approved or denied for a loan.

```
## Spliting into Training and Test

```{r}
train_data <- cbind(new_data2, target)

training <- createDataPartition(
  train_data$loss,
  p = 0.8,
  list = F,
  times = 1
)

train_data1 <- train_data[training,]
test_data1 <- train_data[-training,]

training_x <- train_data1 %>% dplyr::select(-c("loss")) %>% data.matrix()
testing_x <- test_data1  %>% dplyr::select(-c("loss")) %>% data.matrix()

training_y <- train_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
testing_y <- test_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
training_x
```
```

We will anticipate and incorporate both the default and the severity of the losses that would result. Our goal is to build a bridge between traditional banking, where we are looking to decline the total use of consumption of economic capital, to an asset-management perspective, where we aim to optimize the risk to the financial investor.

We will analyze three scenarios that are based on real world financial investment systems with slightly varying condition parameters.

Scenarios

```
## Final Descision on Customer Loan Approval in Scenario 1

```{r}
sc_1 %>%
  mutate(loan_approval = ifelse(Oppurtunity >= Risk, 1, 0)) %>%
  select(loan_approval) %>%
  write.csv(file = "C:/Users/krish/Desktop/Subjects/Spr-20/Adv Data Mining/Group Project/G7_S1.csv", row.names = F)
```
```

Total Requested Loan where Oppurtunity >= Risk in Scenario 1

```{r}
sc_1 %>%
  mutate(loan_approval = ifelse(Oppurtunity >= Risk, 1, 0)) %>%
  mutate_at(c("loan_approval"), as.factor) %>%
  filter(loan_approval == 1) %>%
  select(requested_loan) %>%
  transmute(total_requested_loan = format(sum(.), big.mark = ",", scientific = F)) %>%
  head(1)
```

	total_requested_loan <chr>
1	268,562,496

1 row

Choosing the Model based on Rsquared Metric

```{r}
matrix <- coef(glmnet_model$finalModel, s = glmnet_model$bestTune$lambda)

rsqr <- data.frame(
  name = matrix@Dimnames[[1]][matrix@i + 1],
  coefficient = matrix@x
  ) %>%
  filter(name !="(Intercept)") %>%
  arrange(-abs(coefficient)) %>%
  use_series("name") %>%
  as.character()

rsqr
```

```
  [1] "f596" "f599" "f404" "f378" "f228" "f13"  "f67"  "f598" "f386" "f260" "f283" "f399" "f653" "f9"   "f288" "f281" "f392" "f768" "f514" "f629"
 [21] "f654" "f739" "f382" "f395" "f297" "f637" "f344" "f776" "f556" "f59"  "f634" "f361" "f261" "f655" "f70"  "f300" "f75"  "f413" "f291" "f17"
 [41] "f243" "f131" "f253" "f68"  "f589" "f397" "f29"  "f31"  "f588" "f159" "f66"  "f80"  "f330" "f219" "f111" "f198" "f630" "f775" "f740" "f153"
 [61] "f130" "f57"  "f262" "f289" "f252" "f140" "f305" "f293" "f45"  "f91"  "f590" "f383" "f145" "f54"  "f322" "f613" "f522" "f523" "f737" "f663"
 [81] "f640" "f248" "f109" "f428" "f223" "f83"  "f673" "f648" "f443" "f677" "f211" "f479" "f71"  "f434" "f526" "f144" "f619" "f636" "f546" "f638"
[101] "f290" "f524" "f212" "f180" "f93"  "f82"  "f249" "f450" "f94"  "f621" "f331" "f199" "f89"  "f593" "f422" "f61"  "f384" "f148" "f1"   "f189"
[121] "f273" "f680" "f600" "f385" "f340" "f423" "f220" "f669" "f436" "f279" "f32"  "f631" "f142" "f41"  "f516" "f682" "f286" "f73"  "f446" "f132"
[141] "f618" "f208" "f171" "f146" "f609" "f188" "f341" "f441" "f168" "f313" "f240" "f566" "f110" "f536" "f332" "f134" "f396" "f715" "f612" "f114"
[161] "f359" "f659" "f169" "f306" "f716" "f232" "f533" "f190" "f143" "f622" "f104" "f733" "f101" "f296" "f406" "f149" "f150" "f329" "f512" "f666"
[181] "f309" "f339" "f158" "f161" "f44"  "f239" "f201" "f84"  "f294" "f310" "f311" "f295" "f312"
```

Evaluating Customers in Scenario 1

CAPITAL <- $1.4B

```{r}
sc_1 <- test_scenario1_2 %>%
  select(id) %>%
  bind_cols(requested_loan_1_2, PD, LGD) %>%
  rename(LGD = x) %>%
  mutate(
    Oppurtunity = requested_loan * 0.0432 * 5 * (1 - PD),
    Risk = requested_loan * PD *  LGD
  )

summary(sc_1)
```

```
      id            requested_loan         PD                 LGD             Oppurtunity           Risk
 Length:25471       Min.   :10002   Min.   :0.0002264   Min.   :-31.303   Min.   :    0   Min.   :-593327
 Class :character   1st Qu.:29943   1st Qu.:0.0419769   1st Qu.:  2.901   1st Qu.: 5814   1st Qu.:   7517
 Mode  :character   Median :49554   Median :0.0746819   Median :  7.462   Median : 9672   Median :  18086
                    Mean   :49852   Mean   :0.0920020   Mean   :  8.593   Mean   : 9781   Mean   :  18623
                    3rd Qu.:69974   3rd Qu.:0.1229939   3rd Qu.: 12.961   3rd Qu.:13697   3rd Qu.:  32298
                    Max.   :89995   Max.   :1.0000000   Max.   : 72.161   Max.   :19321   Max.   : 648426
```

```
## Evaluating Customers in Scenario 1

CAPITAL <- $1.4B
```

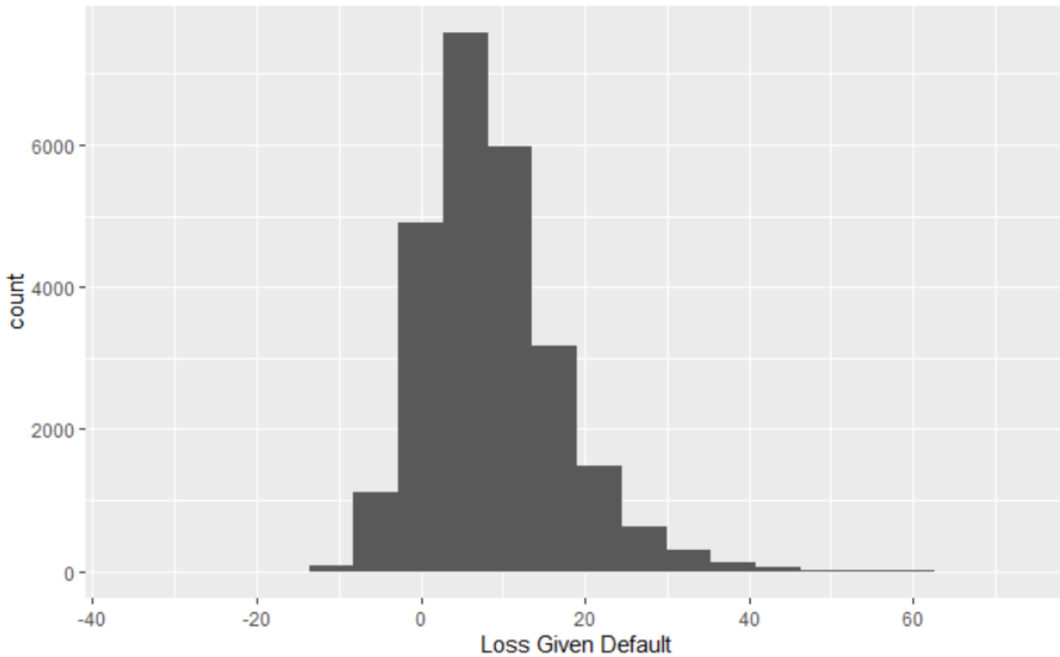
```{r}
sc_1 <- test_scenario1_2 %>%
  select(id) %>%
  bind_cols(requested_loan_1_2, PD, LGD) %>%
  rename(LGD = x) %>%
  mutate(
    Oppurtunity = requested_loan * 0.0432 * 5 * (1 - PD),
    Risk = requested_loan * PD *  LGD
  )

summary(sc_1)
```

```
       id            requested_loan        PD                  LGD              Oppurtunity          Risk        
 Length:25471       Min.   :10002    Min.   :0.0002264   Min.   :-31.303   Min.   :    0    Min.   :-593327  
 Class :character   1st Qu.:29943    1st Qu.:0.0419769   1st Qu.:  2.901   1st Qu.: 5814    1st Qu.:   7517  
 Mode  :character   Median :49554    Median :0.0746819   Median :  7.462   Median : 9672    Median :  18086  
                    Mean   :49852    Mean   :0.0920020   Mean   :  8.593   Mean   : 9781    Mean   :  18623  
                    3rd Qu.:69974    3rd Qu.:0.1229939   3rd Qu.: 12.961   3rd Qu.:13697    3rd Qu.:  32298  
                    Max.   :89995    Max.   :1.0000000   Max.   : 72.161   Max.   :19321    Max.   : 648426  
```

The key purpose of this project is to examine and carefully evaluate these scenarios, aiming to make

calculated predictions that will minimize the risk of any financial investment, and maximize the total profit returns to financial investors based on the clients that are deemed fit to be offered a loan. We will access the risk by evaluating the individual's probability of defaulting, including their potential loss given a default.

```{r}
ggplot(as.data.frame(lgd), aes(x = lgd)) +
  geom_histogram(bins = 20) +
  xlab("Loss Given Default")
```



A potential client should offer an opportunity that is likely to result in profit based on the loan's given interest rate. The bank will be then able to access the expected gain or loss when considering potential clients in their decision process.

Project Deliverables:

The following items needs to be delivered

1. **Project report (25 marks):**

This is your end of project delivery document. It is a document that summarizes different aspects of the project work. It includes the following sections:

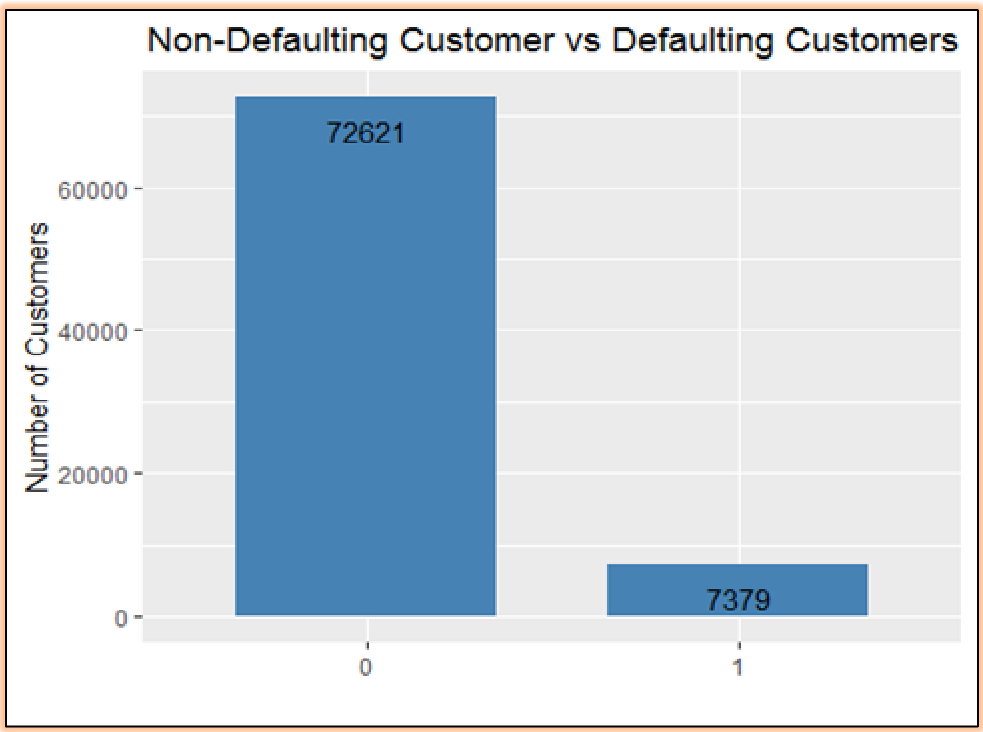
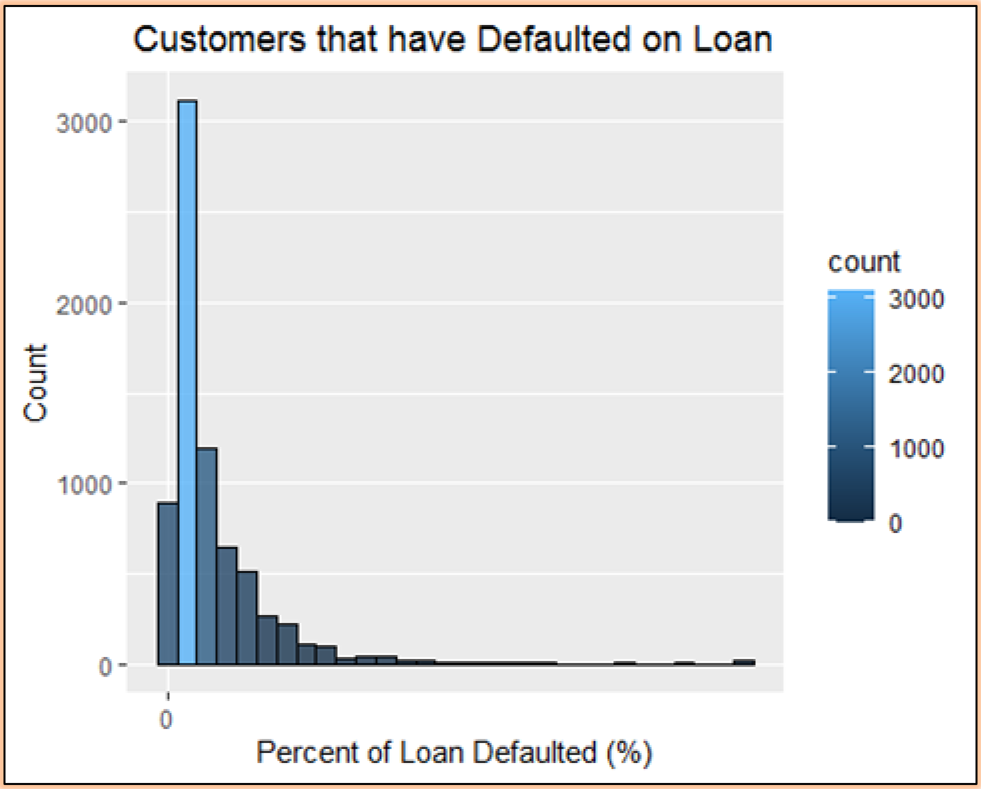
1. Project Goal
2. Overview of data, including data exploration analysis
3. Details of your modeling strategy (i.e. what technique and why)
4. Estimation of model's performance
5. Insights and conclusions

You can include snapshots of your R code and the outputs in the report (recommended).

```
## Recoding Target Variable

```{r}
target <- loan_training_data %>%
  select(loss) %>%
  mutate_at(c("loss"), ~ifelse(. > 0, 1, 0)) %>%
  mutate_at(c("loss"), as.factor)

summary(target)
```
```

```
 loss
 0:72621
 1: 7379
```

You have to submit a single document per group in PDF format. The first page of the document should include a table with a list of the names of the group participants and a very brief summary of contribution

```
## Recoding Target Variable

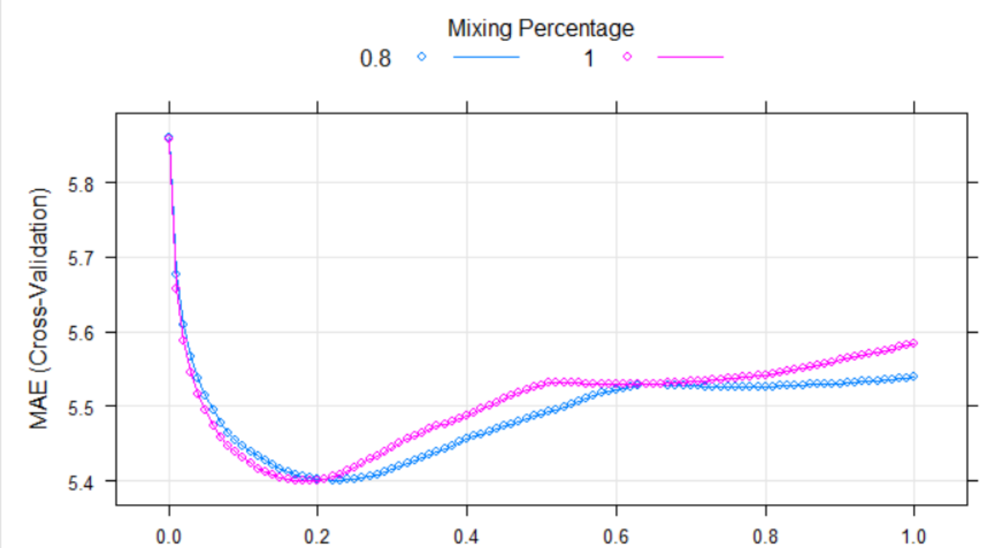
```{r}
target <- loan_training_data %>%
  select(loss) %>%
  mutate_at(c("loss"), ~ifelse(. > 0, 1, 0)) %>%
  mutate_at(c("loss"), as.factor)

summary(target)
```

  loss
  0:72621
  1: 7379
```

of each team member.

2. Overview of Data & Exploration Analysis

In the first scenario, the bank has $1.4 billion to distribute among 25,471 applicants. The primary goal is to offer loan funds only to customers who can generate high expected returns with a low probability of default, preferably with minimal loss in the event of default. To make a statistically wise decision, our predictive model must be accurate in forecasting customers who should be approved or denied for a loan.

In the second scenario, the bank has only $450M loan for customers. Due to the lower amount of total loan capital, the clients must be more strictly selected to maximize gain. Therefore, the goal is to select the top clients who will most likely generate the largest profitable return. By using a proper algorithm we should be able to predict the ideal loanees out of the 25,471 applicants who have lower default rates and higher expected returns.

Finally in the last scenario, the bank has the same $1.4 billion capital as scenario one. However, our model has to consider different interest rates since each customer proposes their own interest rates. It then must predict different expected gain and loss for each customer based on the varying interest rates. Subsequently, the bank will decide on approval or rejection of an applicant's requested loan.

The banking dataset used for this project identifies customers by ID numbers and various other variables instead of their names. The columns are poorly labeled which causes difficulties in inspecting and incorporating both the default and the severity of the losses that may result. The variable's titles do not provide any domain knowledge to help us identify the purpose of the columns. As such, we have to rely solely on data preparation and correlation analysis before moving to preparation of the dataset and data modeling.

Structure of the Data

The dataset consists of 80,000 rows, each of which corresponds to a customer, and 762 columns, which corresponds to varying variables that describe a customer's profile. All variables are numeric, including the one that represents the "loss" percentage of the loan for each customer. Moreover, the dataset contains two identical variables which are a link to the identification of each customer.

We took an overview of the data structure to expose the types of variables that were present. In the data preparatory stage, we carried out an extensive amount of data cleaning, but there was no need for the restructuring of the variables. Furthermore, by observing the dataset and coding in R, we discovered that many data values were missing and the amount of missing data from variables ranged from 0% to 17.83%. Next, we examined the dataset for variables that had zero variance or near zero variance, as well as those that were highly correlated to each other.

Subsequently, we noticed that the data included a "loss" column, which indicated that customers with zero in the "loss" column had paid their loans off in full without defaulting. Whereas customers whose numerical values were greater than zero had defaulted in their loan’s payment at some points. Interestingly, the majority of the defaulting clients we found paid off around 75% of their loan before defaulting. In figure 1 we demonstrated the data between defaulting vs non-defaulting customers. The total number of customers who defaulted were insignificant compared to the customers who did not default on their loan. As a successful banking company, it is important to know why some of these customers defaulted, and to take advantage of the information for business gain. In the absence of specific reasons for default based on the provided variables in the dataset, we came to the conclusion that numerous factors could be responsible. Examples could be the loss of a job due to a pandemic like Covid-19, poor health, loss of a loved one, financial hardship due to unforeseen occurrences, etc. We cannot state precisely why some customers defaulted without additional domain knowledge.



Figure - Histogram of Default % on Defaulting Customers

By comparing defaulting and non-defaulting customers, we found that 72,621 customers paid their loans off in full, while 7,379 customers ended up defaulting on their loans at some point in time. These numbers gave us a historic default rate of 10.16% for the given data set. Figure 2 below shows a histogram of these percentages.



Figure 2 - Comparison of defaulting customer to non-defaulting customers

Each bar represents the percentage of customers who have defaulted on loan. The data shows a positively skewed histogram, which means that the percentage of loan defaults is skewed toward the right. The histogram shows us that the majority of customers defaulted during the final 25% of their loan.

3. Modeling Strategy Details

For our modeling strategy, we looked at numerous techniques as we contemplated on the appropriate technique. We separated the models into two parts. While we set one model for predicting the probability of default **(PD)**, we set the other model for predicting

```
## Spliting into Training and Test

```{r}
train_data <- cbind(new_data2, target)

training <- createDataPartition(
  train_data$loss,
  p = 0.8,
  list = F,
  times = 1
)

train_data1 <- train_data[training,]
test_data1 <- train_data[-training,]

training_x <- train_data1 %>% dplyr::select(-c("loss")) %>% data.matrix()
testing_x <- test_data1  %>% dplyr::select(-c("loss")) %>% data.matrix()

training_y <- train_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
testing_y <- test_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
training_x
```
```

```
## Recoding Target Variable

```{r}
target <- loan_training_data %>%
  select(loss) %>%
  mutate_at(c("loss"), ~ifelse(. > 0, 1, 0)) %>%
  mutate_at(c("loss"), as.factor)

summary(target)
```
```

```
 loss
 0:72621
 1: 7379
```

the expected loss given default (LGD).

```
```{r}
write.csv(lgd, file = "C:/Users/krish/Desktop/Subjects/Spr-20/Adv Data Mining/Group Project/LGD.csv", row.names = F)
```
```

The Probability of Default PD analysis is a method applied by larger institutions to estimate their expected loss.

```
glmnet <- train(
  x = training_x,
  y = training_y,
  method = "glmnet",
  metric = "MAE",
  tuneGrid = tune_grid,
  trControl = cross_validation
)

glmnet$results %>%
  select(alpha, lambda, Rsquared, MAE, RMSE) %>%
  right_join(glmnet$bestTune)

plot(glmnet)
```



A PD is normally measured by assessing past-due loans. It is allocated to each risk measure and represents as a percentage the likelihood of default. PD is calculated by running a migration analysis of similarly rated loans.

```
## Imputing NA values using Median

```{r}
model_imputation <- preProcess(data, method = "medianImpute")
new_data <- predict(model_imputation, data)

anyNA(new_data)

```
```

```
[1] FALSE
```

```
## Recoding Target Variable

```{r}
target <- loan_training_data %>%
  select(loss) %>%
  mutate_at(c("loss"), ~ifelse(. > 0, 1, 0)) %>%
  mutate_at(c("loss"), as.factor)

summary(target)
```
```

```
 loss
 0:72621
 1: 7379
```

median

The calculation is for a specific time frame and measures the percentage of loans that default. The PD is then assigned to the risk level, and each risk level has one PD percentage. LGD is a method used by the banking industry or segment. It measures the expected loss and is shown as a percentage. LGD represents the amount unrecovered by the lender after selling the underlying asset if a borrower defaults on a loan. An accurate LGD variable may be difficult to determine if portfolio losses differ from what was expected and an inaccurate LGD may be due to the segment being statistically small.

The initial step for our model involved cleaning and reducing the data set into a more manageable size, known as feature selection. Specifically, we removed zero variance, near zero variance, and highly correlated variables. This process resulted in a reduction of variables from 740 to 246.

Before running through a regularization mode, we attributed the missing values by using a median imputation method to reduce the data set down to more critical variables before the model building process.

```
## Imputing NA values using Median

```{r}
model_imputation <- preProcess(data, method = "medianImpute")
new_data <- predict(model_imputation, data)

anyNA(new_data)

```
```

```
[1] FALSE
```

To select and regulize variables effectively and enhance the prediction accuracy, We intended to use

```
## Spliting into Training and Test

```{r}
train_data <- cbind(new_data2, target)

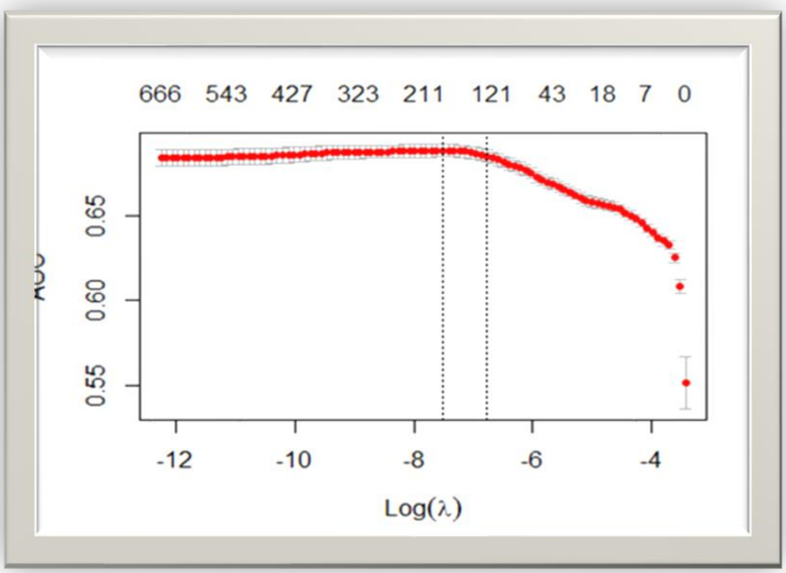
training <- createDataPartition(
  train_data$loss,
  p = 0.8,
  list = F,
  times = 1
)

train_data1 <- train_data[training,]
test_data1 <- train_data[-training,]

training_x <- train_data1 %>% dplyr::select(-c("loss")) %>% data.matrix()
testing_x <- test_data1  %>% dplyr::select(-c("loss")) %>% data.matrix()

training_y <- train_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
testing_y <- test_data1 %>% dplyr::select(c("loss")) %>% use_series("loss")
training_x
```
```

chose Lasso and Random Forest Models. The Lasso model returned a total of around 180 variables as important to the default target variable. The top 10 were selected for further analysis and remaining variables for Principal component analysis, PCA.



I

Figure 3 - Plot of the AUC values versus the log of the lambda values.

Figure 4 -

ROC determines the accuracy of a classification model at a user defined threshold value and evaluates the model's accuracy by using Area Under Curve (AUC). The area under the curve, AUC, represents the performance of the ROC curve. Higher the area and curve, better the model which seems to be the case in our graphic with 91% accuracy. ROC is plotted between True Positive Rate and False Positive Rate. In general, the aim of this model is to push the red curve toward the left corner close to one and to maximize the area under the curve. The grey line represents the ROC curve at 0.5 threshold where sensitivity is equal to specificity. Based on our data, we displayed the relation between sensitivity and specificity figure 5 as well.

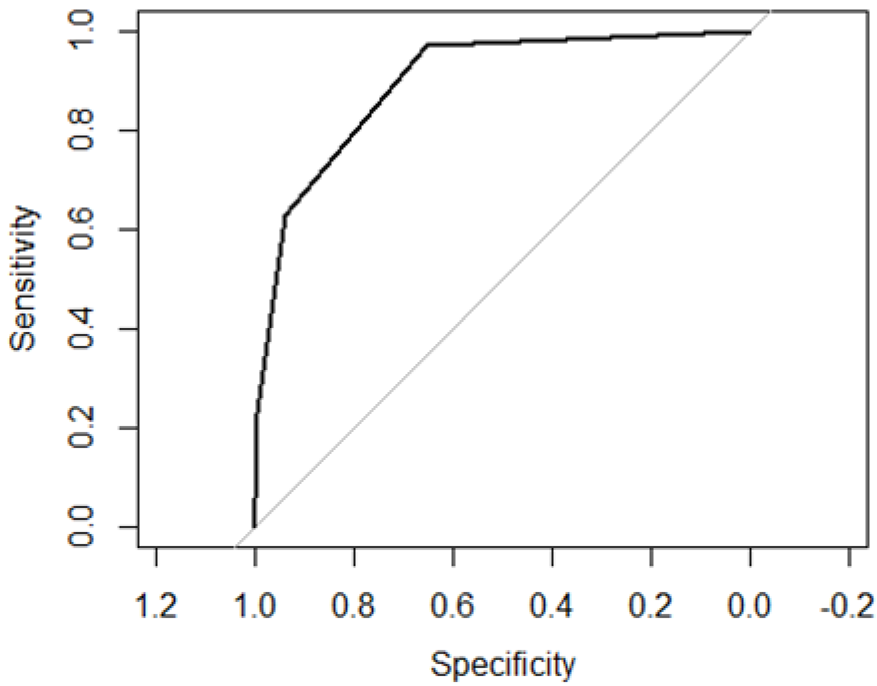


Figure 5 - The sensitivity and specificity correlation

Reference

1. Dr Razavi Kent State university - Bing. 2020. Dr Razavi Kent State university - Bing. [ONLINE] Available at: https://www.bing.com/search?q=Dr+razavi+Kent+State+university&cvid=f430d1b57edf48ffaaef6210d108a6fd&FORM=ANNTA1&PC=U531. [Accessed 02 May 2020]

2. Zest AI. 2020. Transform Your Underwriting with AI & Machine Learning | Zest AI. [ONLINE] Available at: https://www.zest.ai/article/how-ai-can-optimize-lender-yields. [Accessed 18 April 2020].

3. (https://www.reuters.com/article/us-usa-fed-household-debt/u-s-household-debt-tops-14-trillion-and-reaches-new-record-idUSKBN20521Z)

4. How AI can optimize your existing borrower pools . 2020. How AI can optimize your existing borrower pools . [ONLINE] Available at: https://www.bai.org/banking-strategies/article-detail/how-ai-can-optimize-your-existing-borrower-pools. [Accessed 02 May 2020].

5. Credit score ranges & what they mean | Chase. 2020. Credit score ranges & what they mean | Chase. [ONLINE] Available at: https://www.chase.com/personal/credit-cards/education/credit-score/credit-score-ranges-and-what-they-mean. [Accessed 02 May 2020].

6. TechRadar. 2020. How AI and machine learning are improving the banking experience | TechRadar. [ONLINE] Available at: https://www.techradar.com/news/how-ai-and-machine-learning-our-improving-the-banking-experience. [Accessed 28 April 2020].

7. Why banks need artificial intelligence - Wipro. 2020. Why banks need artificial intelligence - Wipro. [ONLINE] Available at: https://www.wipro.com/en-US/business-process/why-banks-need-artificial-intelligence/. [Accessed 28 April 2020].

8. TechRadar. 2020. How AI and machine learning are improving the banking experience | TechRadar. [ONLINE] Available at: https://www.techradar.com/news/how-ai-and-machine-learning-our-improving-the-banking-experience. [Accessed 28 April 2020].

9. NeuroChain. 2020. Machine Learning in Banking: How Banks Approve Automated Loans. [ONLINE] Available at: https://www.neurochaintech.io/machine-learning-in-banking-loan-decisions/. [Accessed 28 April 2020].

10. Stephanie Guttman. 2020. Machine Learning in Banking and Construction Loan Administration - Rabbet. [ONLINE] Available at: https://rabbet.com/blog/machine-learning-in-banking-and-construction-loan-administration/. [Accessed 28 April 2020].

11. Nathan Yau. 2020. How to Read and Use Histograms in R | FlowingData. [ONLINE] Available at: https://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/. [Accessed 28 April 2020].

12. HackerEarth Blog. 2020. Practical Guide to Logistic Regression Analysis in R | HackerEarth Blog. [ONLINE] Available at: https://www.hackerearth.com/blog/developers/practical-guide-logistic-regression-analysis-r/. [Accessed 28 April 2020].

13. Investopedia. 2020. Exposure at Default (EAD) Definition. [ONLINE] Available at: https://www.investopedia.com/terms/e/exposure_at_default.asp. [Accessed 28 April 2020].